TESIS DOCTORAL
2013

APORTACIONES DESDE LA PSICOLOGÍA COGNITIVA Y LA INTELIGENCIA ARTIFICIAL AL DIAGNÓSTICO DE LA ENFERMEDAD DE ALZHEIMER

José María Guerrero Triviño

Máster Universitario de IA avanzada

Ingeniero en Informática

UNED

Departamento de Inteligencia Artificial

E.T.S.I Informática

Director Rafael Martínez Tomás

Codirectora Herminia Peraita Adrados

DEPARTAMENTO DE INTELIGENCIA ARTIFICIAL

E.T.S.I Informática

UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

APORTACIONES DESDE LA PSICOLOGÍA COGNITIVA Y LA INTELIGENCIA ARTIFICIAL AL DIAGNÓSTICO DE LA ENFERMEDAD DE ALZHEIMER

José María Guerrero Triviño

Máster Universitario de IA avanzada

Ingeniero en Informática por la UNED

Director Rafael Martínez Tomás

Codirectora Herminia Peraita Adrados

A mi esposa y a mis hijos.

Las grandes almas son como las nubes, recogen para luego repartir.

(Kalidasa, poeta indio)

Agradecimientos.

Dedico este trabajo a mi familia, que sin su comprensión y paciencia no hubiese sido posible llevar a cabo esta tesis doctoral.

Agradezco a Rafael Martínez Tomás, director de la tesis, los comentarios para perfeccionar esta tesis doctoral, su apoyo, su esfuerzo por difundir las conclusiones de esta investigación y que me propusiera llevar a cabo la investigación de esta tesis doctoral a partir de las conclusiones de la investigación de Peraita y Grasso [1].

Agradezco a Herminia Peraita Adrados, codirectora de la tesis, la precisión de sus comentarios, las aclaraciones sobre su investigación, que es la base de esta tesis doctoral, su empeño y esfuerzo por ampliar el corpus lingüístico de definiciones orales con más casos.

Resumen.

Esta tesis doctoral tiene por objeto colaborar al diagnóstico de la EA en fase leve y moderada desde el campo de la Psicología Cognitiva y la IA. Para ello, se ha elaborado un método de diagnóstico que pretende ser complementario a las técnicas habituales, que permite identificar el deterioro cognitivo, de la memoria semántica declarativa, compatible con el estado más o menos leve de la EA. Es decir, estas alteraciones cognitivas se manifiestan de manera muy sutil en las primeras fases de la evolución de la enfermedad, permitiendo por tanto, su diagnóstico temprano.

La enfermedad de Alzheimer (EA) es una de las principales causas de demencia entre los ancianos; la prevalencia de esta enfermedad se incrementa con la edad, siendo responsable de entre el 50% y el 70% de los casos de demencia en la UE [2]. Esta enfermedad produce graves consecuencias en el paciente y, en su entorno familiar y social, sin mencionar el elevado coste económico que supone para las administraciones públicas, tanto en el ámbito sanitario, como en la atención socio-sanitaria.

El deterioro cognitivo leve (DCL) es un precursor de la enfermedad de Alzheimer y otras enfermedades neurodegenerativas (ENs). Una de las manifestaciones del DCL es el deterioro de la memoria semántica en sus aspectos declarativos. Diversas investigaciones indican que la EA y otras ENs causan trastornos en la memoria semántica, provocando errores en la denominación y descripción de objetos [3], es decir, entre otros síntomas, causa un deterioro semántico de categorías específicas referentes a categorías naturales y objetos básicos, y que además, se van degradando con el tiempo. Las investigaciones sugieren que el lóbulo temporal podría ser el responsable del deterioro de categorías específicas en los trastornos de la memoria semántica [4]. El deterioro semántico que cursa un déficit léxico-semántico-conceptual en la EA, es estudiado en el *Corpus Lingüístico de Definiciones Orales* elaborado por Peraita y Grasso [1], donde se analiza cómo se representan mentalmente determinadas categorías semánticas mediante modelos teóricos de rasgos o atributos semánticos, obtenidos a través de tareas lingüísticas explícitas. Por otro lado, en la literatura científica se señala un deterioro semántico diferencial entre las categorías de seres vivos y seres no vivos, y entre determinados tipos de rasgos. Esto se debe a un daño degenerativo del sistema nervioso central que provoca que algunas categorías semánticas se deterioren o se pierdan diferencialmente [1,5,6]. Este deterioro semántico

diferencial se ha podido evidenciar en algunos enfermos de EA y otras enfermedades neurodegenerativas.

En esta investigación se propone un método de diagnóstico de la EA complementario a las técnicas habituales, desde el campo de la psicología cognitiva y la inteligencia artificial (IA). A partir del análisis de los rasgos semánticos contenidos en definiciones orales de determinadas categorías semánticas, proporcionado por el corpus [1], se generan evidencias para una red bayesiana (BN) con las que se hallan patrones complejos de interacción entre la producción oral de rasgos semánticos y la EA. En esta tesis doctoral se ha diseñado un software que utiliza principalmente redes Bayesianas discretas e hibridas, combinándose estas BNs con otras técnicas de IA como los árboles de decisión o el análisis de clúster. El modelo cualitativo de estas redes Bayesianas (BNs) se ha diseñado a partir del conocimiento del dominio y el modelo cuantitativo se ha aprendido con algoritmos de aprendizaje automático, diseñados específicamente para esta investigación. En esta investigación se han desarrollado varios modelos de BN para las que se han utilizado distintas técnicas de modelado, las cuales representan condiciones y descubrimientos de la EA obtenidas de la literatura científica [1,7,8]. Con este software se han realizado numerosos experimentos con los que se han conseguido unos resultados excelentes, siendo especialmente relevante el resultado obtenido con la BN que modela de forma explícita el deterioro semántico diferencial entre los dominios semánticos de seres vivos y de seres no vivos. En otro experimento se ha constatado una influencia informativa entre la edad y nivel educativo, y la producción oral de rasgos semánticos. También es interesante la comparativa del resultado entre, una BN que modela la segmentación de las producción oral de atributos en los once bloques conceptuales que propone el corpus lingüístico de Peraita y Graso [1], y una BN que no segmenta la producción oral de rasgos semánticos en estos bloques conceptuales. Del mismo modo, se han conseguido unos resultados excelentes en los experimentos con redes Bayesianas hibridas, las cuales utilizan variables latentes para realizar construcciones hipotéticas de modelos causales basándose en hallazgos sobre la EA extraídos de la literatura científica.

Los resultados conseguidos en esta tesis doctoral son prometedores, creemos sinceramente que el método de diagnóstico que proponemos podría llegar a ser de aplicación clínica y no quedarse únicamente en el plano experimental. Este método de diagnóstico podría complementar, en la clínica diaria, a las técnicas habituales de

diagnóstico; sería un método muy barato, ecológico y accesible a poblaciones más extensas, en comparación con los métodos de diagnóstico basados en la evaluación clínica.

Confiamos en que el método de diagnóstico que proponemos sea sólo el principio de un proyecto de investigación más ambicioso, que permita hallar una evidencia convergente/discriminante del diagnóstico de la EA y otras enfermedades neurodegenerativas; explotando al máximo las posibilidades de la psicología cognitiva y la IA, y extendiendo la metodología a otros campos como la neuropsicología cognitiva o la neurología.

Abstract.

This doctoral thesis takes as an objective to contribute to the diagnostic of Alzheimer's disease (AD) during phase mild and moderate from the artificial intelligence and cognitive psychology field. For it, has been developed a diagnostic method that tries to be complementary to the usual techniques, which allows us to identify the cognitive impairment, of the declarative semantic memory, compatible with the state more or less mild of the AD. Namely, these cognitive disorders manifest in the early stages of the evolution of the disease, allowing therefore, its early diagnosis.

AD is a major cause of dementia among the elderly, with a prevalence which increases with age. In European Union, this disease is the major cause of dementia, it is the cause between the 50% and 70% of the cases. This disease causes serious consequences in the patients, their families and social environment, not to mention the high economic costs for public administration, in both the Health and Social-Health assistances.

Mild cognitive impairment (MCI) is a precursor of Alzheimer's disease and other neurodegenerative diseases. One manifestation of the MCI is the semantic memory impairment in its declarative aspects. Various researches indicate that the AD and other neurodegenerative diseases, causes disorder in semantic memory causing errors in the designation and description of objects [3]. The disorder of the semantic memory is characterized by a deterioration of specific categories concerning to basic natural categories and objects, and furthermore, will degrade over time. Researchers suggest that the temporal lobe may be responsible for the deterioration of specific categories in disorders of semantic memory [4]. The semantic deterioration that process a lexical-semantic-conceptual deficit in the AD, is studied in the linguistic corpus of oral definitions by Peraita and Grasso [1], which discusses how certain semantic categories are represented mentally by theoretical models of features or semantic attributes, obtained through explicit linguistic tasks. On the other hand, in the scientific literature is noted a semantic differential impairment between categories of living beings and artifact, and between certain types of features. This is due to a degenerative damage of the central nervous system which causes some semantic categories deterioration or gets lost differentially [1,5,6]. Early diagnosis of AD is very important to achieve more effective pharmacological treatments and cognitive therapies.

A diagnostic method of the MCI caused by AD, is proposed in this thesis, complementary to standard diagnostic techniques from the field of cognitive psychology and artificial intelligence (AI). From the analysis of semantic features contained in oral definitions of certain semantic categories, provided by corpus [1], the evidences for probabilistic networks are provided through which complex patterns of interaction between the EA and the oral production of semantic features are found.

Software designed in this thesis used Bayesians networks discrete and hybrid, although also other AI techniques are used as decision tree or cluster analysis. BNs qualitative model has been designed based on the knowledge of the domain and quantitative model has been learned by machine learning, which has been designed specifically for this investigation. With this software there have been numerous experiments that have achieved excellent results, being especially relevant the result obtained with the BN that models that explicitly represents the differential semantic impairment between living beings and artifact, which same AD and other neurodegenerative disease patient can present. On the other hand it has been found a functional influence between the age and educational level, and the production of semantic features. In another experiment has been found a functional influence between the age and educational level, and the production of semantic features. Also interesting is the result obtained in the experiment that compares a BN that models the segmentation of the oral production of linguistic attributes in the eleven conceptual blocks underlying to all representation of knowledge proposed by Peraita and Grasso corpus linguistic [1], versus a BN that only models the total production of semantic features, without any segmentation. In the same way, in experiments with hybrid BNs have been obtained excellent results. These BNs use latent variables by constructs hypothetical causal models, which are based on AD findings obtained from the scientific literature.

The results achieved in this thesis are promising; this diagnostic method could give rise to clinical application and not only experimental as in this thesis. This method might complement, in the daily clinic, to the usual techniques of diagnosis; it would be a method more cheap, ecological and accessible to more extensive populations, than the diagnostic methods based on clinical evaluation**.**

I trust that the method of diagnostic proposed in this thesis, it is only the beginning of an ambitious research project, which allows to find a convergent evidence and

discriminant in the AD diagnostic and other neurodegenerative diseases, exploiting the full potential of cognitive psychology and AI, and extending the methodology to other fields such as cognitive neuropsychology or neuroscience.

Índice

V Apéndices

Lista de Abreviaturas.

BN	Bayesian Network
CLG BN	Conditional Linear Gaussian Bayesian Networks
DC	Deterioro Cognitivo/Deterioro Cognitivo del contenido semántico
DS	Deterioro Semántico de categorías y dominios específicos
DSD	Deterioro Semántico en sus aspectos declarativos
DI	Diagrama de Influencia
DLSC	Déficit léxico-semántico-conceptual
EA	Enfermedad de Alzheimer
EN	Enfermedad Neurodegenerativa
fMRI	Functional magnetic resonance image
IA	Inteligencia Artificial
IU	Interfaz de Usuario
MMSE	Test Minimental
MR	Magnetic resonance
PET	Positron emission tomography
SNV	Seres no vivos/artefactos
SV	Seres vivos / categorías naturales
TPC	Tabla de probabilidades condicionales

Definiciones previas.

Dominios semánticos	SV y SNV
Categorías semánticas SV	Manzana, perro y pino.
Categorías semánticas SNV	Coche, silla y pantalón
Rasgos o atributos semánticos contenidos en las definiciones orales de cada categoría semántica	Taxonómicos, tipos, parte-todo, funcional, evaluativo, lugar y hábitat, comportamiento, causa/genera, procedimental, ciclo vital y otros

Lista de Tablas.

Lista de Figuras.

PARTE

Preliminares

I

Capítulo

Introducción

1

Esta investigación parte de un corpus lingüístico de definiciones orales libres con restricción temporal elaborado por Peraita y Grasso[1] [1], y en el que se analiza la producción de rasgos de las definiciones orales de seis categorías semánticas referentes a categorías naturales y a objetos básicos [1,8]. En este análisis se hallan diferencias cuantitativas en la producción oral de rasgos semánticos entre personas cognitivamente sanas y enfermos de EA. Para más detalles ver http://www.fbbva.es/TLFU/dat/DT3_2010_corus_linguistico_peraita_web.pdf y www.uned.es/investigacion-corpuslinguistico.

Estas definiciones orales, una vez analizadas, interpretadas y segmentadas en los once bloques conceptuales que propone el corpus lingüístico de Peraita y Grasso [1], constituyen las evidencias para una red Bayesiana (BN). Las BNs son la principal técnica de IA utilizada en esta tesis doctoral, aunque no es la única, ya que estas se combinan con otras técnicas de IA. Además, estas BNs se complementan con algoritmos de aprendizaje automático de su modelo cuantitativo y con otras mejoras que se detallan a lo largo de esta tesis doctoral. Las BNs han permitido combinar el conocimiento a priori del experto (modelo causal de la BN) y el conocimiento aprendido desde un conjunto de casos resueltos (modelo cuantitativo o parámetros probabilísticos de la BN). Las BNs han permitido expresar de forma gráfica las interacciones causales, lo cual ha facilitado la comunicación multidisciplinar. Por otro lado, las distintas técnicas de modelado de las BNs han de facilitado el análisis e interpretación de los resultados, permitiendo llegar a conclusiones interesantes para el diagnóstico de esta enfermedad.

Nos centramos exclusivamente en la EA debido a que el corpus lingüístico [1] utilizado en esta tesis doctoral se ha elaborado a partir de respuestas verbales en una tarea de producción, obtenidas de participantes cognitivamente sanos y enfermos de EA. A partir de estas definiciones orales se han hallado, con técnicas de IA, diferencias cuantitativas y patrones complejos en la producción oral de rasgos semánticos, permitiendo identificar el DCL en enfermos de EA.

---

1 Desde los años 80 hasta la actualidad existen muchos corpus lingüísticos que estudian las disociaciones categoriales, pero nuestro trabajo se centra en el corpus de Peraita y Grasso [1].

La estructura del capítulo es la siguiente. En la sección 1.1 se hace una introducción al método de diagnóstico. En la sección 1.2 se detalla la motivación de esta investigación. En la sección 1.3 se resumen los objetivos de este trabajo. En la sección 1.4 se presenta de forma concisa el estado actual de los métodos habituales de diagnóstico de la EA, así como las mejoras introducidas en las técnicas de IA utilizadas en esta tesis doctoral. En el último apartado se detalla la organización de la estructura de la tesis.

1.1 Introducción al método de diagnóstico.

El método de diagnóstico que se ha diseñado en esta tesis, analiza la alteración cognitiva que afecta a la memoria semántica en sus aspectos declarativos en un estadio muy temprano de la evolución de la enfermedad. El diagnóstico temprano de la EA es muy importante para conseguir una mayor eficacia en los tratamientos farmacológicos y en las terapias cognitivas.

La asociación entre unidades de información memorísticas de la memoria declarativa y sus representaciones mentales, es un concepto fundamental en la psicología cognitiva. La memoria declarativa (de la que forma parte la memoria semántica) está formada por unidades de información memorísticas. Cada unidad de información guarda un conjunto de relaciones definidas con otras unidades de información memorísticas (por ejemplo, «sirve para _», o «esto es un _»), de tal forma que pueden mapearse en forma de una red semántica [1,9]. A partir de esta red semántica se han diseñado varios modelos de BNs que permiten calcular la probabilidad de que determinadas unidades de información, así como sus relaciones con otras unidades información, se hayan deteriorado. Como se ha indicado anteriormente, el método de diagnóstico que proponemos utiliza el corpus lingüístico de definiciones orales libres con restricción temporal de Peraita y Grasso [1]. En este corpus lingüístico se propone la segmentación de la producción oral de atributos en once bloques conceptuales subyacentes a toda representación del conocimiento; el modelado en las BNs de dicha segmentación, mejora la eficacia de las mismas. Es decir, si se recuenta la producción oral de rasgos semánticos o atributos para todas las definiciones orales, el clasificador bayesiano muestra un peor rendimiento que segmentando la producción oral de atributos en los bloques conceptuales propuestos en [1] (ver capítulo 3 para más detalle).

El corpus lingüístico de Peraita y Grasso [1] se publicó en el año 2010 y fue a posteriori cuando se inició la investigación llevada a cabo en esta tesis doctoral. Este desfase entre ambas investigaciones presenta algunos inconvenientes, como por ejemplo, no es posible realizar otras exploraciones complementarias de la memoria semántica u otras pruebas neuropsicológicas a los participantes del corpus [1]. Los enfermos de EA fueron diagnosticados como tal por neurólogos, los cuales siguieron los criterios NINCDS-ADRDA. Los criterios Alzheimer NINCDS-ADRDA fueron propuestos en 1984 por el *National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association* y es el más utilizado en el diagnóstico de la EA. Estos criterios requieren, antes del diagnóstico clínico de la posible o probable EA, la presencia del DC y una sospecha confirmada del

síndrome de la demencia mediante varios test neuropsicológicos. Este criterio ha demostrado una buena fiabilidad y validez, pero actualmente existen métodos de diagnóstico basados en biomarcadores [10,11,12] que son capaces de diagnosticar la EA con una precisión muy alta, incluso pueden discriminar la EA de otras ENs. Sin embargo, estos métodos de diagnóstico tienen un elevado coste y aún no están accesibles a la clínica diaria.

El corpus lingüístico [1] se ha elaborado con un muestreo incidental, donde se seleccionaron deliberadamente los individuos qué forman parte de la muestra. El proceso de selección de los individuos fue muy costoso, ya que se necesitó la colaboración de 4 o 5 departamentos de neurología de los hospitales de la CAM (Hospital Universitario 12 de Octubre, Hospital La Paz, Hospital Fundación Jiménez Díaz y Hospital de la Princesa) y el acceso a sus bases de datos. A partir del corpus lingüístico de definiciones orales [1,13], las técnicas de IA pueden identificar posibles déficits léxico-semánticos-conceptuales de categorías específicas y pueden aprender los parámetros de las BNs. Del mismo modo, el corpus proporciona un marco teórico y metodológico para el análisis e interpretación de las definiciones orales que constituyen las evidencias para las BNs propuestas en esta investigación.

En la literatura científica y estudios epidemiológicos se indica que la causa más común de demencia en la Unión Europea es la EA (alrededor del 50-70% de los casos), otra de la causa es la demencia multiinfarto (alrededor del 30% de los casos de demencia), la enfermedad de *Pick*, demencia de cuerpos de *Lewy* y otros [14]. El método de diagnóstico que proponemos podría extenderse a otras enfermedades neurodegenerativas (ENs) que causen DC, aunque todavía falta determinar la especificidad de nuestro método de diagnóstico respecto a otras demencias de tipo no-EA, ya que sólo se han experimentado con enfermos de EA por la dificultad en la obtención de nuevos casos.

Por otro lado, nuestro trabajo no se ha limitado a aplicar unas técnicas estándar de la IA, sino que se han introducido una serie de innovaciones con las que se han obtenido unos resultados mejorados. Estas mejoras están dirigidas a optimizar el rendimiento del clasificador bayesiano cuando la EA es incipiente, que es justo cuando la aplicación de algoritmos de aprendizaje automático propios de la minería de datos produce más falsos positivos y falsos negativos, en comparación con el método de diagnóstico propuesto en esta tesis doctoral. Con este fin, se han creado varios modelos de BNs aplicando distintas técnicas de modelado, se han diseñado distintos algoritmos de aprendizaje automático del modelo cuantitativo y distintas estrategias en la discretización de los atributos numéricos. En concreto se han creado tres modelos de BNs discretas, un modelo de CLG BN y un modelo de BN híbrida, el cual utiliza un algoritmo de inferencia aproximada. Con estos modelos de BN se persiguen los siguientes objetivos:

- Modelar en la BN el deterioro semántico diferencial que presentan algunos enfermos de EA, entre los dominios de seres vivos (SV) y de seres no vivos (SNV) [1,6].

- Comparar el rendimiento de una BN que utiliza en el proceso de inferencia razonamiento abductivo, versus otra BN que utiliza en el proceso de inferencia razonamiento deductivo.
- Analizar las mejoras de rendimiento que se consiguen en los clasificadores, al segmentar la producción oral de atributos lingüísticos en los bloques conceptuales que propone el corpus [1].
- Crear un algoritmo de aprendizaje automático que sólo requiere para el aprendizaje automático del modelo cuantitativo una muestra de sujetos sanos. Esta característica es importante porque permite abaratar los costes del trabajo de campo y consecuentemente, posibilita la obtención de una muestra más amplia, diversa y más representativa de la población. No obstante, para la validación del método de diagnóstico, es necesaria una pequeña muestra de sujetos enfermos de EA.

Todo el software desarrollado en esta tesis se ha construido con software OpenSource y en entorno Web. Entre las funcionalidades que abarcan se distinguen: gestión de las prevalencias, gestión de la base de casos, cálculo automático del modelo cuantitativo de las distintas BNs –opcionalmente se podrían configurar las tablas de probabilidades condicionales (TPCs) con probabilidades subjetivas—, diagnóstico a partir de las definiciones orales, análisis de resultados, evaluación de la eficiencia de los modelos de BN y por último, se implementa un *framework* para la optimización de la TPC de la variable ***EA*** basado en algoritmos de estrategias evolutivas.

Los resultados obtenidos en los experimentos se han contrastado con el diagnóstico proporcionado por neurólogos, quienes utilizaron los criterios NINCDS-ADRDA. También se han aplicado diversas técnicas de minería de datos al corpus lingüístico con objeto de comparar la eficacia del método propuesto en esta tesis, respecto a otras técnicas de minería de datos.

Algunas aportaciones o aspectos novedosos, se han llevado a los congresos:

- III Congreso Internacional de Lingüística de Corpus (CILC2011) en la ciudad de Valencia (http://www.upv.es/contenidos/CILC2011).
- International Work-Conference on the Interplay Between Natural and Artificial Computation, en la isla de La Palma en las Islas Canarias (http://www.iwinac.org/iwinac2011/)
- Póster de investigación en el congreso Alzheimer Internacional 2011 celebrado en Madrid los días 22 y 23 de Septiembre. Seleccionado como trabajo relevante para una exposición oral en el área sociosanitaria (http://www.alzheimerinternacional2011.org)
- AI-HEALTH celebrado en San Cristóbal de la Laguna, Tenerife (Canary Islands, Spain) (http://perseo.inf.um.es/~aike/aihealth11)

Los diferentes planteamientos de la tesis se han publicado en:

- Foundations on Natural and Artificial Computation. Lecture Notes in Computer Science, 2011, Volume 6686/2011, 419-430, DOI: 10.1007/978-3-642-21344-1_44 (http://www.springerlink.com/content/t052vrn627047m55) [15]
- Actas del 3 Congreso Internacional de Lingüística de Corpus. Tecnologías de la Información y las Comunicaciones: Presente y Futuro en el Análisis de Corpus. ISBN: 9788469462256. Pág. 731-740 [16]. (http://www.upv.es/pls/obib/sic_publ.FichPublica?P_ARM=6032)
- AI-HEALTH. Asociación Española de la Inteligencia Artificial. (http://erevista.aepia.org/index.php/aia/article/viewFile/935/758)

1.2 Motivación.

La motivación de esta tesis parte de la necesidad que existe en la clínica diaria de un sistema de diagnóstico de la EA barato y precoz, y la suposición que desde el Corpus de definiciones orales [1] y los modelos bayesianos, se puede llegar a un software capaz de identificarla. Las BNs nos permiten modelar distintas hipótesis de investigación con las que pretendemos lograr una mayor precisión en el diagnóstico temprano de esta enfermedad. El diagnóstico precoz de esta enfermedad es muy importante para conseguir una mayor eficacia de los tratamientos farmacológicos, así como para entender la situación en la que se encuentran los pacientes y actuar en consecuencia. Actualmente el diagnóstico de la EA se realiza mediante los criterios clínicos NINCDS-ADRA y aunque se trata de un método de diagnóstico muy eficaz, requiere la presencia de un deterioro cognitivo o una sospecha confirmada del síndrome de la demencia.

El método de diagnóstico que proponemos en esta tesis doctoral permite identificar el deterioro de la memoria semántica (DS) en sus aspectos declarativos, siendo esta evidencia suficiente para inferir la presencia o la ausencia del DCL. Por otro lado, la EA provoca errores en la denominación y descripción de objetos, es decir, la EA afecta a la memoria semántica y por consiguiente, podemos predecir la posibilidad de padecer la EA a partir de la presencia de un aspecto del DC –concretamente el deterioro de la memoria semántica. Además, el DCL puede representar un estado pródromo para la EA, lo cual permite el diagnóstico en las primeras fases de la enfermedad. El diagnóstico del deterioro semántico de categorías específicas se infiere con redes probabilistas, más concretamente con BNs. La implantación de nuestro método de diagnóstico en la clínica diaria podría suponer un gran ahorro para las administraciones públicas. Además, al tratarse de un método de diagnóstico muy barato, podría utilizarse como un método preventivo y con una capacidad de aplicación a poblaciones muy extensas.

La investigación realizada ha proporcionado resultados prometedores como se pueden comprobar en la sección de experimentación del método. Se ha utilizado únicamente como instrumento metodológico el corpus de definiciones orales de Peraita y Grasso [1], pero sería posible tener en cuenta nuevas variables –síntomas y factores de riesgo— para hacer un diagnóstico más preciso y determinista.

1.3 Objetivos.

El objetivo de esta tesis doctoral, como se ha indicado anteriormente, es el diagnóstico de la EA en fase leve a partir del deterioro de la memoria semántica en sus aspectos declarativos y para ello, se ha diseñado un método de diagnóstico fiable, sencillo y económico. El deterioro de la memoria semántica provoca errores en la denominación y descripción de objetos básicos que permite a nuestro método de diagnóstico identificar los déficits léxico-semánticos-conceptuales contenidos en las definiciones orales con restricción temporal de determinadas categorías naturales y objetos básicos.

En los modelos causales de las BNs hay que considerar que la edad y el nivel educativo, pueden influir en la producción oral de rasgos semánticos. En algunas investigaciones como en [9], se indica que las unidades de información memorísticas de la memoria declarativa disminuyen en función del tiempo que hace que fue creado y aumentan en función del número de veces que son evocadas desde la memoria, es decir, un escaso nivel educativo puede hacer que disminuyan determinadas funciones cognitivas relativas a la memoria semántica. Por otro lado, otras investigaciones apuntan a que determinadas capacidades del cerebro como la memoria, el razonamiento y la comprensión (función cognitiva), pueden empezar a deteriorarse con la edad [17].

Por otro lado, algunas investigaciones apuntan a que la EA puede causar deterioros de categorías especificas generalizados, cuando la enfermedad está avanzada, y causar deterioros semánticos irregulares o focalizados, cuando la enfermedad es incipiente o incluso prodrómica [1,6].

En los modelos causales propuestos en esta tesis doctoral se utilizan la variable ***EA***, la cual está relacionada, entre otras, con la información de contexto: edad, sexo y nivel educativo. No ha sido posible en esta investigación disponer de estudio epidemiológico específico, por lo que se ha recurrido a la literatura científica [18] para obtener la prevalencia de la enfermedad estratificada por estos factores de contexto y poder así, calcular la TPC de la variable ***EA***.

En esta tesis se han realizado una serie de tareas para permitir identificar de forma precoz, la alteración cognitiva que afecta a la memoria semántica en sus aspectos declarativos; estas tareas son:

- Analizar, utilizando para ello distintas técnicas de modelado en las BNs (sección 5.5 y 5.4), el deterioro semántico diferencial entre los dominios SV y SNV. La BN que modela en su estructura estas relaciones causales mejora el rendimiento cuando la enfermedad es incipiente, respecto a otra BN donde no se tiene en cuenta el deterioro diferencial. Esta demostración es muy importante para el diagnóstico precoz de la enfermedad y para discernir la EA de otras demencias de tipo no-EA que no causen este deterioro focalizado.
- Analizar la influencia de la edad y el nivel educativo en la producción oral de rasgos semánticos. Se ha diseñado una estrategia que permite tener en cuenta estos factores, sin tener que añadir más de 100 enlaces informativos en la estructura de la BN. Es muy importante determinar si el déficit en la producción

oral de rasgos semánticos se puede deber a otros factores que nada tienen que ver con las ENs, como por ejemplo, la propia vejez o a un escaso nivel educativo.

- Optimización de la TPC de la variable ***EA***. Esta técnica permite reducir costes en el desarrollo del trabajo de campo, ya que permite recabar más casos sin tener en cuenta la aleatoriedad de la muestra, que de otra manera daría lugar a la aparición del sesgo. Cada individuo de la población representa los parámetros de la TPC que se pretenden optimizar. La función *fitness* del algoritmo utiliza, como métrica de adecuación a la solución, el AUC, el cual se calcula a partir de las probabilidades a posteriori de un subconjunto de casos del corpus lingüístico [1].
- Diseñar e implementar una BN híbrida que permita tener en cuenta investigaciones relevantes del campo de la psicología cognitiva [7]. Esta BN utiliza para la inferencia de las variables latentes los siguientes coeficientes: coeficientes de correlación de *Pearson*, coeficientes de regresión, ratios de ganancia de información, distancia euclídea modificada o ponderación de atributos en función de la producción oral de rasgos semánticos.
- Análisis comparativo de una BN que modela el conocimiento del dominio subyacente al corpus lingüístico de definiciones orales propuesto por Peraita y Grasso [1], versus una BN que sólo tiene en cuenta el recuento de rasgos semánticos de todas las definiciones orales.
- Desarrollo de un sistema software basado en web que, aunque no es necesario para estudiar y analizar el método de diagnóstico, si lo hemos considerado muy importante para conseguir financiación privada. Dadas las dificultades que existen en estos momentos para la inversión en investigación, hemos creído imprescindible que tenga proyección comercial y por ello, hemos considerado muy importante enmarcarlo dentro de los proyectos TIC para e-Salud. Los objetivos básicos para los que necesitamos la financiación privada son:
    - Diseñar una variante de nuestro método de diagnóstico que permita la adquisición de evidencias de forma automática o semiautomática. En nuestro método diagnóstico las transcripciones, los análisis e interpretaciones de las definiciones orales, se realizan de forma manual.
    - Añadir nuevas variables o factores de contexto al modelo causal con objeto de poder discriminar la EA de otras ENs.
    - Disponer de más casos para validar el método de diagnóstico.
    - Incorporar al sistema nuevas funcionalidades, como el análisis de decisión o las terapias cognitivas.

En el capítulo 12 se realizan a una serie de propuestas para futuras investigaciones desde los campos de la neuropsicología cognitiva, neurología y atención sociosanitaria.

1.4 Estructura de la tesis doctoral.

Esta tesis se ha organizado en cuatro partes, doce capítulos y dos apéndices. La primera parte ***Preliminares***, se compone de esta introducción, una descripción del estado de la disciplina y las aportaciones de esta tesis al estado del conocimiento.

La segunda parte ***Descripción de la propuesta*** se compone de los capítulos 3, 4, 5, 6 y 7. El capítulo 3 describe la metodología desde la perspectiva de la psicología cognitiva y de la IA. El capítulo 4 desarrolla un análisis estadístico de todas variables utilizadas en los modelos causales propuestos en esta investigación. El capítulo 5 describe en detalle las distintas técnicas de modelado utilizadas en las BNs discretas, así como las distintas estrategias de discretización y los distintos algoritmos de aprendizaje automático del modelo cuantitativo. El capítulo 6 describe en detalle el diseño de los modelos cualitativos y cuantitativos de las BNs híbridas, y las distintas técnicas de inferencia. El capítulo 7 describe el algoritmo de estrategias evolutivas que se ha utilizado para optimizar la TPC de la variable ***EA***.

La tercera parte ***Experimentos*** se compone de los capítulos 8, 9, 10 y 11. En el capítulo 8 se realiza una batería de experimentos con las BNs Discretas. En el capítulo 9 se realiza una batería de experimentos con las BNs híbridas. En el capítulo 10 se evalúan las mejores estrategias para la optimización de la TPC de la variable ***EA,*** teniendo en cuenta la eficacia y el coste computacional. En el capítulo 10 también se valida la eficacia del algoritmo de optimización y se analiza el problema del sobreajuste. En el capítulo 11 se aplican otros algoritmos de minería de datos sobre el corpus de definiciones orales [1] y el resultado de estos algoritmos se compara con los resultados de nuestro diagnóstico.

La cuarta parte ***Conclusiones*** se compone el capítulo 12 *Conclusiones y trabajos futuros,* donde se detallan las conclusiones de la tesis y se proponen nuevas vías de investigación.

En el apéndice A ***Interfaz de Usuario*** se describe brevemente la interfaz web de usuario que consideramos un aspecto novedoso de esta tesis doctoral. En el apéndice A se describen en amplitud las funciones de la interfaz de usuario.

En el apéndice B ***Experimentos complementarios*** se realizan una serie de experimentos para afianzar los resultados y conclusiones de esta tesis doctoral.

Capítulo

Estado del conocimiento y nuestras aportaciones al mismo

2

Actualmente existen numerosas investigaciones sobre la EA y por ello en este capítulo se analiza el estado del arte en el diagnóstico de esta enfermedad. Así mismo, se describen las aportaciones novedosas de esta tesis a las técnicas de IA para conseguir un buen método de diagnóstico.

La organización del capítulo es la siguiente. En la sección 2.1 se describe el estado del arte del diagnóstico de la EA. En la sección 2.2 se justifica cada una de las técnicas de IA utilizadas en esta tesis doctoral, siendo en los capítulos 5, 6 y 7, dónde se describen detalladamente el método de diagnóstico que proponemos.

2.1 Técnicas habituales empleadas para el diagnóstico de la EA.

Actualmente, las técnicas habituales de diagnóstico de la EA se basan, entre otras, en la evaluación clínica y en el juicio médico. También se utilizan marcadores biológicos que pueden diagnosticar la EA con bastante precisión y pueden discriminar en un porcentaje alto de casos, la EA de otras ENs [19], aunque estos sistemas no están accesibles a todo el mundo, pues son costosos y no están disponibles en la clínica diaria.

Técnicas de diagnóstico de la EA.

Normalmente la evaluación clínica [20] se suele complementar con información adicional obtenida con imágenes del cerebro PET, las cuales tienden a ser exploraciones caras. También se suelen realizar exploraciones complementarias analizando líquido extraído de la región inferior de la médula espinal [10]. La evaluación clínica consume mucho tiempo y es muy costosa, y a pesar de ello, se diagnóstica correctamente un gran número de casos en estadios tempranos de la enfermedad. La evaluación clínica normalmente requiere considerar la historia médica general y neurológica, así como un

examen neuropsicológico, neurológico, neuroimagen, conducta, analítica general, etc. (ver detalles en [21]).

Los marcadores biológicos o biomarcadores, podrían reemplazar estos procedimientos clínicos. Los biomarcadores se basan en la cuantificación de proteínas asociadas con características histopatológicas de la EA. Según el artículo *Biochemical Diagnosis of Alzheimer Disease by Measuring the Cerebrospinal Fluid Ratio of Phosphorylated tau Protein to β-Amyloid Peptide$_{42}$* [19], este método es bastante exacto, incluso es eficiente discriminando la EA de otras ENs (sensibilidad, 86%; especificidad 97%). Los datos de la investigación [19] los obtienen examinando 100 pacientes no hospitalizados. El fluido cerebroespinal se obtiene mediante punción lumbar y dicha punción se realiza como máximo dentro de la semana siguiente a la del examen neuropsicológico. En la investigación referida también se estudia la correlación entre el nivel del fluido cerebroespinal de phospho-tau y Aβ_{42}, y el test minimental (MMSE), concluyendo que existe una fuerte correlación. Cabe mencionar que en nuestra investigación también examinamos la correlación existente entre los enfermos de EA, los déficits léxico-semánticos-conceptuales que presentan estos enfermos y el test MMSE. El test MMSE fue publicado en 1975 por Marshal F. Folstein, Susan Folstein y Paul R. McHugh. El test MMSE consiste en un cuestionario de 30 puntos que se usan para detectar posibles alteraciones cognitivas. Estas preguntas tienen que ver con la orientación, la memoria, la atención, el cálculo, el lenguaje, la escritura y el dibujo.

Existen investigaciones recientes en las que se hallan nuevos marcadores biológicos que pueden ayudar a identificar que personas con DCL, podrían llegar a desarrollar la EA. Un ejemplo de estas investigaciones es [20,10] en la que se miden 1os niveles de 190 proteínas en sangre de 600 participantes en el estudio. Los participantes incluyen personas sanas, personas diagnosticadas como enfermos de EA o personas diagnosticadas con DCL. De los 190 niveles de proteínas, 17 fueron significativamente distintas en personas con DCL o con la EA. Estos marcadores se volvieron a comprobar con datos de 566 participantes, y sólo cuatro marcadores seguían siendo significativamente distintos: apolipoproteína E, B-type péptido natriurético, proteína C-reactiva y polipéptido pancreático. Los niveles de estas cuatro proteínas en sangre están correlacionadas con los niveles en los mismos pacientes de la proteína β-Amyloid Peptide$_{42}$. En la investigación [20,10] falta determinar la especificidad, ya que sólo fueron incluidos en el estudio un pequeño número de pacientes con DCL no EA [20,10]. Cabe destacar que los biomarcadores aún no están presentes en la clínica diaria.

También existen notables avances en el diagnóstico de la EA y en general el DCL basándose en imágenes. Debido al hecho de que la proteína Aβ es una de las principales características patológicas, se hace imprescindible la detección y cuantificación de estas placas de proteínas en el cerebro de los pacientes de EA. Las nuevas imágenes PET trazador de Ligandos[2] de las proteínas amiloide, ofrecen la posibilidad de obtener medidas de la proteína Aβ fibrilar y estudiar en el tiempo, el curso de Aβ en el cerebro

[2] http://es.wikipedia.org/wiki/Ligando

[11]. En el primer estudio en el que se realizan imágenes de la proteína Aβ, utilizan PET con Pitssburgh Compound-B (PIB) y se utilizan 16 pacientes diagnosticados con EA moderado y 9 controles. Comparando los controles con los pacientes de EA, estos mostraron una retención notable de PIB en áreas de la corteza asociativa[3] o áreas de asociación, y se encuentran grandes cantidades depositadas de Aβ. En los pacientes de EA la retención PIB fue incrementada más prominentemente en la corteza frontal. También fueron observados grandes incrementos de Aβ en la corteza parietal, temporal, y occipital, y en el cuerpo estriado. Se estudiaron tres personas jóvenes de 21 años y seis personas mayores sanas, y mostraron una baja retención de PIB en áreas corticales. Del mismo modo, no se hallaron grupos significativos de Aβ entre personas jóvenes y personas mayores sanas. Los resultados de este estudio sugieren que las imágenes PET con este nuevo trazador PIB, pueden proporcionar información cuantitativa sobre los depósitos Aβ en sujetos vivos [12].

Aplicación de la IA en la medicina.

La IA ha participado en investigaciones recientes en el diagnóstico de la EA con técnicas de análisis de las imágenes. Algunas de estas investigaciones se basan en campos de investigación como la biología o el diagnóstico basado en *functional magnetic resonance imagen* (fMRI). Desde el campo de la biología se hace referencia al artículo [22], en el que se emplean clasificadores Bayesianos y selección de atributos multivariados para inducir dependencias probabilísticas que podrían correlacionar o desvelar relaciones biológicas. En esta investigación [22] se han encontrado hallazgos interesantes, como la desactivación de los genes DEC1 y BTRC, involucrados en la regulación del reloj molecular que controla el ciclo circadiano del cuerpo. Otra investigación en la que también se utilizan BN con objeto de asistir en el diagnóstico basado en fMRI se ha publicado en [23]. El método de diagnóstico [23] consiste en 5 etapas: a) preprocesamiento de los datos de la fMRI para eliminar aquellos no relacionado con la tarea de variabilidad, b) modelar la forma en la que la respuesta depende del estímulo, c) extracción de atributos de los datos del fMRI, d) clasificación usando el algoritmo *Random Forests*. En [23] consiguen una especificidad y sensibilidad del 96%.

Dentro del mismo campo de investigación, en [24], se propone una BN en la que se combinan datos procedentes de resonancia magnética (MR) y variables que representan funciones clínicas/cognitivas, para ayudar al diagnóstico temprano del DCL. Esas imágenes MR fueron segmentadas y registradas a MNI Template[4], y con unos cálculos automáticos del volumen de la estructura se transformaron a un valor discreto. Esos volúmenes incluyen el hipocampo, tálamo, perirhinal cortex, etc. Con los datos de esas imágenes y otras medidas clínicas, utilizaron un método de búsqueda heurística para

3 Cualquiera de las extensiones de la corteza cerebral que no son sensorial o motor en el sentido habitual, sino que se asocian con etapas avanzadas de procesamiento de la información sensorial, la integración multisensorial, o la integración sensomotora.
4 MNI Template, 1998 (The standard template of SPM and International Consortium for Brain Mapping)

construir el modelo de BN y evaluar la fiabilidad. En esta investigación utilizaron 25 participantes con DCL. Los test realizados indican que el DCL es dependiente principalmente del hipocampo, tálamo y corteza entorhinal.

En la literatura científica se proponen otros métodos de diagnóstico experimentales, que no se basan en el análisis de imágenes, que permiten evaluar el estado cognitivo de los pacientes o incluso las deficiencias motoras causado por un derrame cerebral, a partir de la monitorización de actividades cotidianas y en especial la actividad de vestirse, que proporciona importantes indicadores [25]. En [26] relacionan la tortuosidad en las trayectorias del movimiento –movimientos irregulares— de ancianos con DC. Para ello se ayudaron de una red de sensores de banda ultra-ancha que utilizan transpondedores inalámbricos, con los que midieron la locomoción en 14 ancianos durante un día, con una precisión de 14 cm, cuando estos atraviesan zonas comunes mientras realizan actividades cotidianas como ir al comedor.

En [27] se utiliza una BN con el fin de mejorar el diagnóstico del DCL. Para ello han desarrollado un sistema experto para dirigir la predicción del DCL y la inferencia. En esta investigación descubren que los principales factores de influencia en el DCL son: *ANT* (Attentive Networks Test), *STM* (short-time memory test) y el nivel educativo.

Existen otros casos de éxito en el campo de la medicina donde se emplearon razonadores probabilísticos, como por ejemplo *Diaval*, que fue desarrollado en España por Francisco J. Díez, de la Universidad Nacional de Educación a Distancia (UNED). *Diaval* es un sistema experto para el diagnóstico de enfermedades cardíacas [28]. *Diaval* considera principalmente la información eco cardiográfica, aunque tiene en cuenta otra información como síntomas y signos, hallazgos electrocardiográficos, etc. Al igual que en la investigación que se presenta en esta tesis doctoral [29], *Diaval* cuenta con una interfaz gráfica que permite al usuario introducir información de forma ordenada. Son numerosas las aplicaciones en la medicina de BN y diagramas de influencia; se puede consultar un resumen de algunos de los trabajos más importantes en http://www.cisiad.uned.es/papers/medicina.php de Francisco J. Diez [30].

2.2 Aportaciones novedosas al estado del conocimiento.

El software que planteamos (y desarrollamos) es pionero en la aplicación de técnicas *Soft Computing* en el diagnóstico de la EA, y se puede extender a otras ENs. Las aportaciones novedosas desde la perspectiva de la IA para el diagnóstico temprano (experimental) de la EA son:

- Los modelos cualitativos de las BNs se han diseñado y construido a partir del conocimiento del dominio, representándose condiciones y descubrimientos recientes de la EA, obtenidas de la literatura científica [1,7,8]. Estos hallazgos científicos no se representan en los datos de corpus, entre otros motivos por el número reducido de casos analizados, lo cual hace muy difícil que los algoritmos de aprendizaje automático de la estructura de las BNs puedan descubrir enlaces causales que representen dichos hallazgos. Un ejemplo de ello

es el deterioro semántico diferencial que presentan algunos enfermos de EA cuando la enfermedad es incipiente. Por consiguiente, las BNs diseñadas en esta tesis, modelan mucho mejor la realidad del problema y hechos demostrados científicamente.

- En las BN discretas:
  - Se ha implementado un algoritmo para la discretización de atributos continuos que busca distintos centroides mediante análisis de clúster (k-Means++) [31], segmentando los conglomerados de atributos por edad y/o nivel educativo. Esta mejora en el método de discretización permite establecer relaciones informativas entre la edad, el nivel educativo y, todas las variables que componen el corpus lingüístico, sin necesidad de establecer cientos de enlaces causales en la estructura de la BN.
  - Se ha diseñado un algoritmo de aprendizaje automático [32] específico para el método de diagnóstico que se propone en esta tesis. Este algoritmo utiliza distintas técnicas de aprendizaje automático en función de los tipos de variables y las relaciones causales entre las variables. Por ejemplo, para construir la (TPC) de la variable ***EA***, teniendo en cuenta las relaciones informativas –información de contexto—, se utiliza *Naive Bayes*. Para aprender los parámetros de las TPCs de otras variables se utilizan otras técnicas de aprendizaje automático, descritas en los capítulos 5 y 6. Nuestro algoritmo de aprendizaje automático considera las variables intermedias –variables latentes— y lo hemos diseñado basándonos en investigaciones relevantes del campo de la psicología cognitiva, como por ejemplo [7], en las que se analiza el beneficio de las variables latentes. Las variables latentes son construcciones hipotéticas para un mejor entendimiento de determinadas teorías.
  - Debido a la limitación en cuanto al número de casos del corpus (81 casos), se añaden una serie de mejoras al algoritmo de aprendizaje automático de los parámetros, para calcular las probabilidades condicionales con correctores de Laplace y combinación de información de segmentos adyacentes, de aquellos segmentos de la muestra poco representativos de la población
  - Para propagar las evidencias por la BNs discreta e inferir las probabilidades a posteriori se utiliza *Elvira* [33]. Se desarrolla un mecanismo de sincronización de los parámetros de las TPCs de *Elvira*, con la estructura de datos propios del software que proponemos.
  - Dado que *Elvira* no dispone de métricas de rendimiento basadas en curvas ROC, se implementa un algoritmo para generar estas métricas. El objetivo de las métricas de rendimiento es comparar la eficacia de los distintos modelos de BNs y la eficacia de las distintas estrategias de discretización.

- En las BN híbridas.
    - Al igual que en las BN Discretas, se ha diseñado un algoritmo de aprendizaje automático de los parámetros.
    - Se han diseñado dos tipos de BN híbridas: CLG BN y BN híbridas con métodos de inferencia aproximada. Dicho algoritmo de inferencia aplica distintos métodos de inferencia en función de los tipos de relaciones causales y los tipos de variables (ver detalles en el capítulo 5).
    - Tanto en las CLG BN como en las BN con inferencia aproximada, se han diseñado distintos métodos de inferencia en los cuales intervienen: técnicas de IA (como arboles de decisión J48), métodos estadísticos y métodos matemáticos.
    - En las CLG BN se ha implementado un método de aprendizaje automático del modelo cuantitativo y un método de inferencia en el que se combinan hasta cuatro técnicas de inferencia distintas.
        - Inferencia para variables discretas.
        - Inferencia para variables discretas cuyos padres son variables continuas. En este caso se construye dinámicamente un árbol de probabilidad con el algoritmo J48.
        - Inferencia de variables continuas cuyos padres son variables continuas. Estos dos métodos de inferencia son novedosos y se pueden consultar el detalle de su implementación en la sección (6.3)
            - Inferencia por CLG BN.
            - Inferencia por método aproximado.
        - Inferencia de variables continuas que no tienen padres.
- Los algoritmos de aprendizaje automático del modelo cuantitativo tienen en cuenta estudios epidemiológicos –ajenos a esta investigación— y el conocimiento del dominio. Por tanto, el grado de correlación o incertidumbre aprendidas del corpus representará mejor la realidad del problema.
- Se utiliza una técnica de modelado que mejora sensiblemente la clasificación de los enfermos de EA en fase leve, teniendo en cuenta el deterioro semántico diferencial.
- Con las técnicas de optimización diseñadas e implementadas en esta tesis doctoral (en las que se utilizan algoritmos meméticos) se han conseguido unos resultados prometedores.

Todas estas aportaciones han supuesto las siguientes mejoras por el método de diagnóstico:

- Hay que considerar que cuando la enfermedad se encuentra en estado avanzado el daño cerebral es omnipresente y por tanto, en todas las definiciones orales se evidencian déficits léxico-semánticos-conceptuales. Pero cuando la enfermedad

es incipiente, el diagnóstico, a partir de la identificación del deterioro semántico de categorías específicas, se convierte en un proceso muy complejo y es en este estadio cuando nuestro método de diagnóstico es más eficaz que otros algoritmos de aprendizaje automático.

- Se añaden, a los modelos causales de las BNs, variables intermedias o latentes que permite realizar construcciones hipotéticas de determinadas teorías de investigación. Con esta técnica se han conseguido muy buenos resultados, además de conseguir un modelado del problema más adecuado a la realidad del problema. Desde la perspectiva de la IA, es técnica ha permitido representar las correlaciones positivas que existen entre los rasgos semánticos, sin crear enlaces causales y evitando los ciclos en las BNs.
- El corpus lingüístico de definiciones orales [1] lo constituyen 81 casos. Los algoritmos de aprendizaje automático son ineficaces con un número reducido de casos y con un gran número de variables (66 procedentes del corpus y 3 variables de contexto). Sin embargo, se han conseguido unos resultados excepcionales con nuestro método de diagnóstico, por un lado, debido al modelado de determinados descubrimientos científicos y por otro lado, gracias al diseño un algoritmo de aprendizaje automático del modelo cuantitativo que tiene en cuenta investigaciones recientes sobre la EA.
- La segmentación de atributos en los bloques conceptuales que propone [1], mejora significativamente el rendimiento de los clasificadores bayesianos. Se podrá comprobar en el capítulo 9 que todas estas mejoras innovadoras de las técnicas de IA consiguen una mejora del rendimiento respecto a otros algoritmos de IA aplicada. Si se elimina la segmentación de atributos en los bloques conceptuales que propone [1] y se hace un recuento de la producción oral de rasgos semánticos para cada categoría semántica, se consiguen resultados aceptables con algunos algoritmos de minería de datos cuando el DC está avanzado. Es muy importante detectar el deterioro semántico focalizado que se da cuando la enfermedad es incipiente y esta segmentación de las definiciones orales permite un diagnóstico más eficaz cuando la enfermedad se encuentra en sus primeros estadios.

Todas estas mejoras han contribuido a conseguir un método de diagnóstico eficaz, con un número muy reducido de errores; barato, ya que el diagnóstico es por software, y podría estar disponible en la WWW, ya que lo hemos diseñado en entorno web. El coste/utilidad del método de diagnóstico que proponemos es muy ventajoso. El coste de del método de diagnóstico es muy bajo, tanto desde la perspectiva económica como desde el riesgo o efectos secundarios a los que se somete al paciente, y la utilidad es muy alta, ya que este método es muy accesible a grandes poblaciones de personas sin que suponga un incremento en el coste de implantación en la clínica diaria. Consideramos que nuestro método de diagnóstico podría ser de gran utilidad en atención primaria o en el marco de los sistemas e-Salud.

PARTE

II

Descripción de la propuesta

Capítulo

Descripción metodológica

3

Antes de profundizar en el método de diagnóstico desde el campo de la IA, se realizará una introducción a la investigación llevada a cabo por Peraita y Grasso en la que se elaboró el corpus lingüístico de definiciones orales [1].

La EA causa una atrofia cerebral progresiva, bilateral y difusa, que comienza en regiones mesiales temporales para afectar luego al neocórtex, sobre todo al temporal, parietal y frontal [34]. Las investigaciones sugieren que el lóbulo temporal podría ser el responsable del deterioro de categorías específicas en los trastornos de la memoria semántica [4]. El corpus lingüístico de Peraita y Grasso [1] realiza una exploración de la memoria semántica en sus aspectos declarativos para tratar poner de relieve el DC. Desde la IA, se descubren patrones de interacción entre la EA y la producción oral de rasgos semánticos de definiciones orales referentes a determinadas categorías naturales y objetos básicos, que permiten discernir mediante software, las personas cognitivamente sanas de las personas que muestran un DCL.

El capítulo se divide en dos secciones. En la sección 3.1 se describe el método de diagnóstico. En la sección 3.2 se describen, en amplitud más que en profundidad, las técnicas de IA utilizadas, las mejoras introducidas a las técnicas de IA y la justificación del uso de cada una de las técnicas de IA.

3.1 Corpus lingüístico de definiciones orales.

La EA causa, entre otros, un trastorno de la memoria semántica en sus aspectos declarativos, provocando errores en la denominación y descripción de objetos [3] que se van agravando con el tiempo. El trastorno de la memoria semántica es una de las manifestaciones del DCL, la cual se considera un precursor de la EA. Puesto que esta alteración cognitiva se manifiesta de forma precoz con la EA, se podría diagnosticar en sus primeros estadios. Es del máximo interés la detección temprana de esta enfermedad y actualmente puede hacerse en contextos de investigación exclusivamente [11,19,20]; es a partir de técnicas de diagnóstico basadas en biomarcadores (neuroimagen, líquido cefalorraquídeo y pruebas genéticas) que han avanzado enormemente, hasta el punto que ya se puede diagnosticar la EA sin margen de error, y sin esperar al análisis anatomopatológico post-mortem. Pero, desgraciadamente, estos sistemas no llegan a todo el mundo, pues son costosos y no están disponibles en la clínica diaria. Pasará aún

un tiempo para que estas pruebas sean de rutina. Por lo tanto, en el momento actual se siguen aplicando criterios clínicos y neuropsicológicos.

El corpus lingüístico [1] se ha obtenido a partir de una muestra de participantes constituida por personas sanas y enfermos de EA, procedentes de Hospitales de la Comunidad Autónoma de Madrid. Han participado 81 participantes, de los cuales 42 son ancianos cognitivamente sanos y 39 padecen la EA. Los participantes tenían una edad comprendida entre 60 y 90 años, y se les ha solicitado una serie de definiciones orales con restricción temporal de determinadas categorías naturales y de objetos para el estudio de patologías en relación con el DC del contenido semántico que produce esta enfermedad. En la Tabla 1 se muestran las características demográficas que se han considerado en el corpus lingüístico, siendo estas: edad, nivel educativo y sexo.

Tabla 1.- Características demográficas de los sujetos.

	Edad	Nivel educativo	Sexo	Nº Sujetos
Sanos	más de 85	Secundarios	Hombre	1
		Primarios	Mujer	1
	80-84	Universitarios	Hombre	3
		Secundarios	Mujer	1
		Primarios	Hombre	1
		Primarios	Mujer	1
	75-79	Universitarios	Hombre	1
		Universitarios	Mujer	1
		Secundarios	Hombre	2
		Secundarios	Mujer	3
		Primarios	Mujer	3
	70-74	Universitarios	Hombre	1
		Universitarios	Mujer	2
		Secundarios	Hombre	1
		Primarios	Hombre	1
		Primarios	Mujer	1
	65-69	Universitarios	Hombre	2
		Universitarios	Mujer	1
		Secundarios	Hombre	1
		Secundarios	Mujer	1
		Primarios	Hombre	1
		Primarios	Mujer	2
	menos de 65	Universitarios	Hombre	5
		Secundarios	Hombre	2
		Secundarios	Mujer	2
		Primarios	Hombre	1

	Edad	Nivel educativo	Sexo	Nº Sujetos
EA	más de 85	Primarios	Mujer	2
	80-84	Primarios	Hombre	4
			Mujer	2
	75-79	Primarios	Hombre	1
			Mujer	2
	70-74	Universitarios	Hombre	1
		Secundarios	Hombre	1
			Mujer	1
		Primarios	Hombre	7
			Mujer	6
	65-69	Secundarios	Hombre	1
			Mujer	1
		Primarios	Mujer	3
	menos de 65	Secundarios	Hombre	1
		Primarios	Hombre	3
			Mujer	3

La tarea empírica mediante la que se obtuvieron las definiciones de categorías semánticas (tanto de seres vivientes como de no vivientes), y a partir de las cuales se elaboró el corpus, es una tarea de producción verbal libre de atributos con restricción temporal. La investigación sobre el deterioro de rasgos semánticos de categorías naturales y de objetos, en el marco más amplio del deterioro de la memoria semántica y la representación del conocimiento, ha constatado diferencias cuantitativas y cualitativas en la producción de rasgos entre personas cognitivamente sanas y enfermos de EA. Una vez obtenidas las definiciones verbales orales, se transcribieron, depuraron y analizaron según un protocolo de análisis de rasgos o atributos semánticos [35,1]. En este análisis se distinguen dos tipos de variables: unas cuantitativas (las frecuencias de producción de rasgos, para cada categoría, cada dominio, etc.) y otras cualitativas (los diferentes tipos de rasgos según el modelo). Un fragmento de [1], donde se detalla el método para analizar e interpretar las definiciones orales es: "*El modelo citado, a partir del cual se analizan e interpretan las definiciones, propone once bloques conceptuales básicos, considerados como componentes conceptuales que subyacen a la organización y representación de categorías de objetos. Cada uno ellos se distingue por una etiqueta léxica identificativa (bloque o componente funcional, clasificatorio o taxonómico, evaluativo, destinatario, etc.) y una estructura gramatical, a manera de enunciado verbal, con la cual, por lo general, se los introduce lingüísticamente ('sirve para…', 'es un…', 'es…', 'es para…', etc.). Los componentes conceptuales se refieren tanto a la categoría genérica de inclusión (por ejemplo: 'la silla es un mueble') –**componentes taxonómico**—, como a las partes que la forman o configuran (por ejemplo: 'la silla tiene respaldo, asiento y patas') –**componente parte o todo**—, a la función o uso (por ejemplo: 'sirve para sentarse') –**componente funcional**—, al lugar/hábitat donde suele*

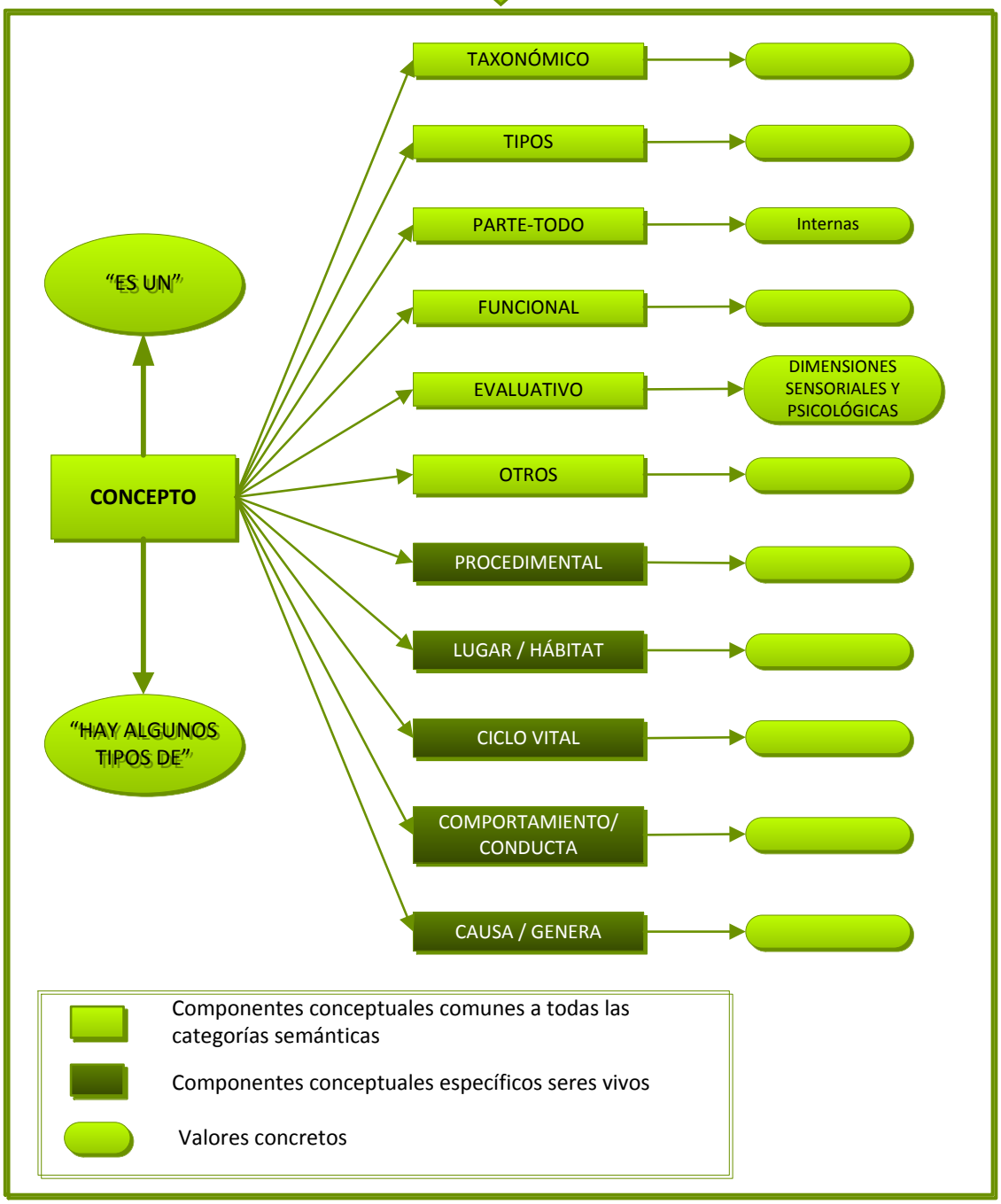
encontrarse (por ejemplo: 'se encuentra en las distintas habitaciones de la casa'), a las dimensiones de la evaluación tanto físicas (perceptuales: forma, color, tamaño, textura) como sociales y afectivas (bondad, simpatía) ***–componente evaluativo—****, como a los tipos o ejemplares que pertenecen a la misma (por ejemplo: 'hay sillas de cocina, de despacho, de bar, etc.'), al agente que las produce o genera (por ejemplo: 'la fábrica, el carpintero')* ***–componente causal—****y al procedimiento de uso* ***–componente procedimental****".* Posteriormente se realiza un análisis cuantitativo, donde se obtienen las frecuencias de cada uno de los tipos de rasgos, en función de un modelo de tipos de rasgos previamente definidos y caracterizado desde el punto de vista semántico [1].

La Figura 1 describe el esquema teórico utilizado para la segmentación de las definiciones orales con restricción temporal en rasgos semánticos. La producción oral libre con restricción temporal de una serie de características de categorías semánticas referentes a categorías naturales y objetos básicos; se han grababado y se han transcrito para su posterior análisis desde un determinado marco teórico de la semántica cognitiva. Una vez segmentada la producción lingüística obtenida, se realiza un análisis cuantitativo por el que se obtienen las frecuencias de cada uno de los tipos de rasgos en función de un modelo de rasgos previamente definidos y caracterizado desde el punto de vista semántico. En la Figura 1 se toma como ejemplo una mujer de 66 años de edad, con estudios primarios y a la que se le ha diagnosticado la EA en grado moderado.

Perro: Ladran mucho, y te muerden si te dejas, como a mí me mordió una vez, y que más, ya está ¿Qué le voy a decir de los perros? yo que sé cuál como no le tengo ni nada, no lo sé, no me acuerdo de nada, así que mira como estoy.

Pino: Allí por mi tierra hay, hay muchos pinos, pero a ver qué le voy a contar (INTE) no sé, luego cogen la resina lo que sea, ya está.

Manzana: Que están muy buenas y se comen... (INTE)... ah yo no sé, porque las manzanas las conozco...(INTE)...pues que son amarillas, otras coloradas... (INTE)... que le voy a decir si ya le dicho lo que es....



Perro

***Actividad conductual*:** ladran mucho, te muerden si te dejas

Pino

***Hábitat*:** por mi tierra hay

***Genera*:** resina

Manzana

Funcional: se comen

***Evaluativo sensorial*:** amarillas, coloradas, están muy buenas

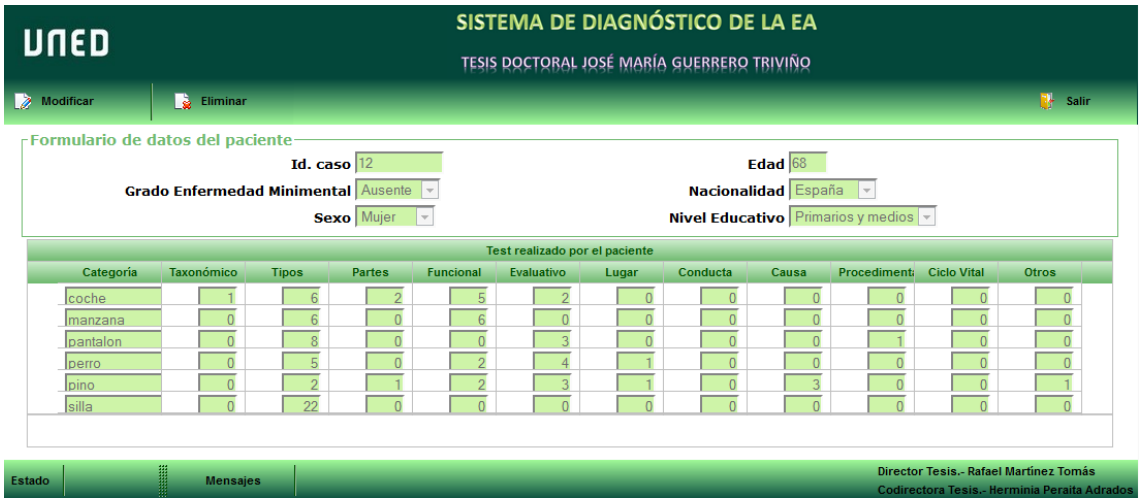


Figura 1.- Esquema seguido para la segmentación de la producción lingüística en rasgos semánticos.

Otro fragmento de [1] interesante para esta investigación es: "*Se ha podido evidenciar que existe un deterioro semántico diferencial entre las categorías de seres vivos y seres no vivos, en personas afectadas con determinadas patologías neurodegenerativas (EA, demencia semántica, demencia por cuerpos de Lewy, etc.), traumáticas (traumatismo craneal) e infecciosas (herpes por encefalitis). Las categorías semánticas se derivan de clasificaciones que se llevan a cabo en el mundo que nos rodea y que permiten tratar como equivalentes objetos que en sí son diferentes. Gracias a que nuestra memoria semántica se encuentra organizada en función de dichas categorías, podemos realizar una serie de funciones cognitivas importantes, tales como hacer inferencias, establecer relaciones entre ejemplares, atribuir propiedades a objetos que no conocemos, razonar; todo lo cual se basa en un principio de economía cognitiva. Las personas que sufren los déficit específicos de categoría, muestran una peor ejecución en tareas que afectan, total o parcialmente, el conocimiento del dominio categorial de los seres vivos*

mientras que el dominio de los objetos o artefactos –seres no vivos— está total o casi totalmente conservados. También existe un pequeño número de casos en el que se da el patrón contrario, hay un mayor deterioro del dominio de los objetos o artefactos, mientras que el dominio de los seres vivos está, en su mayor parte, preservado". Este hallazgo de la enfermedad se utiliza para modelar las BNs desde el conocimiento del dominio.

3.2 Técnicas de IA utilizadas.

En la medicina, como en otros dominios del mundo real, no siempre se dispone de un 100% de certeza en el acierto del diagnóstico, ni de toda la información necesaria para la consecución de los objetivos. Es posible que el diagnóstico se realice a partir de información incompleta, con incertidumbre e inexacta, por tanto, es necesario el uso de técnicas *Soft Computing*. De entre las técnicas *Soft Computing*, se ha optado por las BNs y las estrategias evolutivas, entre otras técnicas de IA.

Los objetivos de la aplicación de cada técnica de IA para el diagnóstico del deterioro semántico en sus aspectos declarativos (DSD), se pueden resumir en:

- ***Anális de clúster (k-Means++). Esta técnica se ha utilizado creando una jerarquía multinivel de conglomerados de atributos:*** Con esta técnica se determina cuándo un paciente presenta un posible déficit léxico-semántico-conceptual en una determinada categoría semántica y/o rasgo semántico. Para ello, se analizan con *k-Means++* el recuento de la producción oral de rasgos semánticos segmentando por edad y/o nivel educativo. El algoritmo asigna una etiqueta clasificativa en función de la distancia euclídea respecto al centroide encontrado con el algoritmo *k-Means++*. En cada segmento de conglomerado de atributos se buscan dos centroides para representar: el posible déficit léxico-semántico-conceptual (déficit presente) y la posible producción normal de atributos (déficit ausente). El objetivo principal de la jerarquía multinivel de conglomerados de atributos es establecer relaciones informativas entre determinadas variables de contexto, como la edad y el nivel educativo, y las variables del corpus. La razón de esta mejora responde a la importancia que tiene determinar si los déficits de la producción oral de rasgos semánticos pueden deberse a otros factores que nada tienen que ver con las ENs, como por ejemplo, la propia vejez o un escaso nivel educativo. Se ha optado por análisis de clúster porque ha permitido combinar métodos matemáticos con *K-Means++,* en aquellos segmentos de la muestra poco representativos de la población (ver detalles en el capítulo 5).
- ***BNs discretas e híbridas***: Son las principales técnicas de IA de esta tesis. Por la BN se propagan las evidencias –recuento de la producción oral de rasgos semánticos una vez analizadas e interpretadas— para inferir las probabilidades a posteriori de sufrir un deterioro cognitivo compatible con la EA y de estar

cognitivamente sano. En el ámbito de las BNs se ha desarrollado de forma especifica para esta tesis:

- ***Modelo cualitativo de la BN.*** Se construye desde el conocimiento del dominio y a partir del corpus lingüístico. En este modelo causal se pueden distinguir los tipos de variables siguientes: variables obtenidas a partir del corpus lingüístico, variables latentes o intermedias, variables de contexto y variables de interés.
- ***Modelo cuantitativo de la BN.*** En esta tesis se implementan algoritmos de aprendizaje automático para el modelo cuantitativo de la BN. Su función es calcular todas las TPCs de la BN en cuestión, la cual consta de 76 variables. El propósito del algoritmo de aprendizaje automático es encontrar el grado de asociación o correlación entre el déficit léxico-semántico-conceptual del contenido semántico y la EA.

- ***Arboles de decisión (C4.***5). Se utilizan en las BNs híbridas para generar árboles de probabilidad en las TPCs de determinadas variables discretas. *C4.5* también se utiliza para generar métricas de rendimiento en los experimentos. Los árboles de decisión se transforman en expresiones condicionales que se evalúan en tiempo de ejecución con un evaluador de expresiones. La evaluación dinámica de expresiones permite cambiar las TPCs de las BNs durante el proceso de inferencia de las BNs híbridas.
- ***Estrategias evolutivas.*** Se utiliza para optimizar las TPCs de determinadas variables. Se ha desarrollado en esta tesis un *framework* de estrategias evolutivas, escrito 100% en Java y adaptado específicamente a nuestro método de diagnóstico. El coste computacional las estrategias evolutivas es muy elevado y por esa razón se han utilizado únicamente en la optimización de la TPC de la variable ***EA***.

3.2.1 Justificación de las técnicas.

A continuación se enumeran la justificación del uso de cada una de las técnicas de IA utilizadas en nuestro método de diagnóstico.

BN

Las BNs son la principal técnica de IA utilizada en esta tesis doctoral [36]. Una BN consiste en dos componentes fundamentales: estructura y parámetros. La estructura de la red probabilista es la parte cualitativa de la red, es decir, define las relaciones causales, funcionales e informativas, identificadas desde el conocimiento del dominio. Por otro lado, los parámetros son las probabilidades condicionales y utilidades, y constituyen la parte cuantitativa de la red. Las razones por las se opta por las BNs se pueden resumir en:

- Una BN es una representación compacta e intuitiva de una relación causal entre entidades de un problema de un dominio determinado, donde las entidades son representadas como variables discretas sobre un conjunto finito de posibles valores mutuamente exclusivo y exhaustivo. Una variable es exhaustiva si todos los posibles valores de la variable están representados en su espacio de estados. Una variable es exclusiva si no existen dos pares de valores que representen el mismo estado para la misma variable. Los tipos de variables que se emplean en esta tesis doctoral son:

  - Variables de contexto: factores de riesgo y factores de protección –edad, el nivel educativo y sexo.
  - Variables objetivos: ***DSD*** y ***EA***. Los posibles estados de estas variables son: *ausente y presente*.
  - Variables que representan los déficits léxico-semánticos-conceptuales de determinadas categorías semánticas y/o rasgos semánticos. La EA pueden producir una gran cantidad de alteraciones cognitivas, pero en esta investigación nos centramos en el deterioro semántico de categorías específicas.
- La inferencia de las redes probabilistas se fundamenta en una base teórica del cálculo de probabilidades y de la teoría de la decisión. Por lo tanto, proporcionan un método matemático coherente para derivar conclusiones bajo incertidumbre, donde múltiples fuentes de información están implicadas en complejos patrones de interacción.
- Las BNs se pueden extender con diagramas de influencia, los cuales permiten tomar decisiones de forma normativa. Es decir, podemos maximizar la utilidad esperada en la recomendación de distintos tipos de terapias cognitivas o incluso tratamientos farmacológicos. Las preferencias son representadas como utilidades sobre escalas numéricas. Otra ventaja de los diagramas de influencia es que permiten combinar de forma explícita y sistemática, las opiniones de diferentes expertos y los datos experimentales, tales como los datos de estudios publicados en la literatura médica [37].
- Las BNs se basan en un lenguaje gráfico y es una potente herramienta para expresar las interacciones causales, al mismo tiempo que se expresan las relaciones de dependencia e independencia entre las entidades del dominio del problema. Esto permite disponer de un lenguaje compacto, que además, proporciona un medio excelente e intuitivo para la comunicación de ideas entre el ingeniero del conocimiento y el experto del dominio.
- La fuerza de las relaciones probabilísticas son representadas por probabilidades condicionales. Téngase en cuenta que las relaciones causales entre variables son rara vez deterministas, en el sentido de que si una causa está presente entonces el efecto puede ser concluido con certeza. Normalmente las relaciones causales entre variables suelen ir acompañadas de un factor de incertidumbre, un grado de correlación o de asociación, que se pueden expresar fácilmente en las BNs.
- Una vez definido el modelo cualitativo y cuantitativo no es necesario, para poder realizar inferencias, disponer de evoluciones temporales, es decir, su estructura y parámetros van a permanecer invariante en el proceso de inferencia.
- Otra aportación muy importante de las BNs es que dispone de mecanismos para producir una explicación de cómo la red ha inferido un diagnóstico.

CLG BN.

Una CLG BN $N = (\mathcal{X}, \mathcal{G}, \mathcal{P}, \mathcal{F})$ consiste en un grafo dirigido acíclico $\mathcal{G} = (V, E)$, un conjunto de distribuciones de probabilidades condicionales $\mathcal{P}$ y un conjunto de

funciones de densidad $\mathcal{F}$. Hay una función de densidad por cada variable continua. Las variables $\mathcal{X}$ se dividen en un conjunto de variables discretas y un conjunto de variables continuas. Cada nodo de $\mathcal{G}$ representa o bien, una variable aleatoria discreta en las que toma un conjunto finito de estados mutuamente exclusivos y exhaustivos, o bien, una variable aleatoria con distribución Gaussiana [36].

La CLG BN que se usa en esta tesis doctoral combina un método de inferencia probabilístico propio de las BNs, con métodos estadísticos y matemáticos. Esta BN utiliza en el proceso de inferencia conclusiones de investigaciones relevantes del campo de psicología cognitiva [7].

BN híbridas con inferencia aproximada.

En [32] se desarrolla un algoritmo de inferencia aproximada con objeto de simplificar el coste computacional del algoritmo de inferencia. La inferencia aproximada se basa por un lado, en distribuciones de probabilidad gaussiana y por otro lado, en probabilidades condicionales aproximadas basadas en simulaciones estadísticas. Al igual que en las CLG BN, el método de inferencia aproximada permite incluir algunas conclusiones de investigaciones relevantes del campo de la psicología cognitiva [7], al modelo cualitativo y cuantitativo de las BNs.

k-Means++.

k-Means es una técnica de clustering ampliamente usada que persigue buscar la distancia euclídea de promedio de los puntos de un mismo clúster. Aunque no ofrece garantías de precisión, su sencillez y velocidad lo hace un algoritmo muy atractivo. Las técnicas de clustering es una de las técnicas de aprendizaje automático y geometría computacional. El objetivo es seleccionar k centros tal que minimice la suma de las distancias euclídea entre cada punto y su centroide. El algoritmo Lloyd, normalmente referido como *k-Means*, comienza con k centros arbitrarios, básicamente elegidos con una distribución aleatoria uniforme entre el conjunto de puntos. Cada punto es asignado a su centroide más cercano y cada centro es recalculado como el centro del conjunto de puntos asignado a ese clúster. Esos dos pasos (asignación y recalculo del centro) son repetidos hasta que el proceso se estabiliza. Ya que hay como máximo k_n posibles clústeres, el proceso siempre termina. Desafortunadamente, la velocidad y simplicidad tiene un precio en la exactitud, existen casos en los que el algoritmo genera de forma arbitraria clústeres malos. Por esa razón se usa *k-Means*++ en lugar de *k-Means*, ya que *k-Means*++ propone inicializar *k-Means* eligiendo al azar centroides con probabilidades muy específicas. Concretamente, elige un punto p como centro con probabilidad proporcional a la contribución del potencial global. Es decir, el algoritmo *k-Means* comienza con un número arbitrario de centros de clúster, y *k-Means*++ propone una forma específica de elegir esos centros de clúster que hace que *k-Means*++ funcione mejor que *k-Means* [31]. Es muy importante para la aplicación de esta técnica, disponer de una muestra estratificada por cada segmento de edad y nivel educativo.

Árboles de decisión. Algoritmo C4.5.

El algoritmo *C4.5* fue desarrollado por JR Quinlan en 1993, como una extensión (mejora) del algoritmo *ID3* que fue desarrollado en 1986. El algoritmo *C4.5* es un algoritmo usado para generar árboles de decisión en problemas de clasificación. *C4.5* construye árboles de decisión desde un conjunto de datos de entrenamiento, usando el concepto de entropía de la información. Cada nodo del árbol *C4.5* elige un atributo de los datos que divida con más eficacia el conjunto de ejemplos en tantos subconjuntos como clases existan. Las características del algoritmo se pueden resumir en:

- Permite trabajar con valores continuos separando los posibles resultados en 2 ramas Vi<=N y Vi>N.
- Cuando todos los ejemplos en la lista pertenecen a la misma clase, simplemente crea un nodo hoja para el árbol de decisión, indicando la clase a la que pertenece cada instancia.
- Utiliza el método *divide y vencerás* para generar el árbol de decisión.
- Se basa en la utilización del criterio de proporción de ganancia. De esta manera se consigue evitar que las variables con mayor número de posibles valores salgan beneficiadas en la selección.
- Es recursivo.
- Para la poda de los árboles adopta la estrategia *post-prunning*, la cual consiste en generar el árbol completo y posteriormente plantear si debe podar para mejorar el rendimiento y de paso, obtener un árbol más corto.

Se ha optado por los árboles de decisión porque es una técnica ampliamente usada en problemas de clasificación supervisados. Además, genera reglas que pueden ser evaluadas dinámicamente con un evaluador de expresiones en tiempo de ejecución. Esto permite modificar en tiempo de ejecución los árboles de probabilidades y TPCs de las BNs híbridas.

Técnica de optimización: estrategias evolutivas.

Con esta técnica de IA se pueden optimizar un número determinado de TPCs de las BNs. Las TPCs son susceptibles de optimizarse, ya que pequeñas variaciones en sus valores dan lugar a cambios significativos en el proceso de inferencia de la BN. Queda fuera del alcance de esta tesis doctoral la optimización de todos los parámetros de la BN.

La computación evolutiva se inspira en el proceso natural de evolución de las especies y la metáfora fundamental en la que se basa, es que asocia el poder de la evolución natural a una forma particular de resolver problemas dentro de un contexto estocástico basado en pruebas y error. Las estrategias evolutivas son una técnica muy popular para la resolución de problemas complejos. Se han aplicado con éxito a un amplio número de dominios distintos y su uso se está incrementando continuamente. Hay que tener en cuenta que la dinámica de los algoritmos evolutivos son bastante difíciles de analizar, ya que son técnicas estocásticas, de alta dimensionalidad y en las cuales existe una

dependencia de la población para alcanzar un resultado óptimo. Además, el operador crossover hace del sistema un sistema no lineal y fuertemente interactivo [38,39].

Capítulo

Análisis Estadístico

4

En este capítulo no se pretende buscar una división entre las técnicas de aprendizaje automático de la IA y la estadística, ya que existe una continuidad entre ambas disciplinas. No obstante, una posible diferencia entre el aprendizaje automático y la estadística la podemos encontrar en [40], en donde se indica que la estadística está más relacionada con las pruebas de hipótesis, mientras que el aprendizaje automático está más relacionado con la formulación de procesos de generalización, como búsqueda de posibles hipótesis. La mayoría de los algoritmos de aprendizaje automático usan test estadísticos para producir unos resultados fiables cuando construyen reglas, árboles, etc., y también, para corregir los modelos que están sobreentrenados en la medida en que dependen fuertemente de un determinado conjunto de ejemplos. Los test estadísticos se usan para validar los modelos de aprendizaje automático de la IA y evaluar los algoritmos.

El objetivo de este capítulo es describir de forma numérica la información recogida en el corpus. También, en este análisis estadístico, se pone de manifiesto la diferencia cuantitativa en la producción oral de rasgos semánticos entre las personas enfermas de EA y las personas cognitivamente sanas, al mismo tiempo que se comprueba la dificultad para hallar hipótesis estadísticas para identificar el deterioro semántico, fundamentalmente por la gran dispersión en la producción oral de rasgos semánticos – tanto en la muestra de personas sanas como en la de enfermos de EA. Se podrá comprobar en este capítulo cómo existe una tendencia a producir menos rasgos semánticos las personas de mayor edad o de un escaso nivel educativo. Otro objetivo de este capítulo es analizar el nivel predictivo o el grado de correlación entre las variables del corpus lingüístico [1] y la EA. Por otro lado, todos los parámetros estadísticos calculados en este capítulo tienen un papel importante en las BN Híbridas, ya que estos se han utilizado para calcular las distribuciones de probabilidad gaussiana y en ecuaciones lineales estructurales o, dicho de otro modo, para construir relaciones estructurales con variables latentes (ver detalles en el capítulo 6).

El capítulo está organizado de la siguiente forma. En la sección 4.1 se identifican las variables del corpus. En la sección 4.2 se calculan las medias y las desviaciones típicas por cada segmento de la muestra considerado en esta investigación. En la sección 4.3 se calculan distintas medidas de correlación, asociación y ratios, entre las variables de

corpus y su valor esperado, entendiéndose como valor esperado el diagnóstico proporcionado por los neurólogos siguiendo los criterios NINCDS-ADRDA.

4.1 Identificación de variables del modelo.

En todos los modelos de BNs que proponemos se utilizan las mismas variables. Tal y como se ha indicado anteriormente, las evidencias para la BN se generan a partir del recuento de la producción oral de rasgos semánticos. En las BNs discretas todas las variables son exhaustivas[5] y mutuamente exclusivas[6]. En la Tabla 2 se pueden distinguir 4 grupos [36] de variables:

- ✓ Variables de información de contexto o factores de riesgo: Es la información que está presente antes de que ocurra el problema y que además, tiene una influencia causal, funcional o informativa sobre el problema. En este grupo de variables consideramos: la edad, el sexo y el nivel educativo.
- ✓ Variables de información que representan los síntomas. Son aquellas que indican si el sujeto cursa un déficit léxico-semántico-conceptual. Estas variables se informan con el recuento de los rasgos semánticos. Los rasgos semánticos comunes a todas las categorías semánticas son: *taxonómico*, *tipos*, *parte-todo*, *funcional*, *evaluativo*, *procedimental* y *otros*. Los rasgos semánticos no comunes a todas las categorías semánticas son: *lugar/hábitat*, *ciclo vital, comportamiento/conducta, procedimental* y *causa/genera*. La taxonomía de atributos que sirve de marco teórico y metodológico de evaluación para esta prueba, puede verse detalladamente en [13,1].
- ✓ Variables intermedias. Son variables no observables directamente, sus probabilidades a posteriori no son de interés inmediato, pero juegan un papel importante para lograr una correcta dependencia e independencia condicional de las propiedades y por tanto, una inferencia eficiente. En los modelos de BNs que se proponen, se incorporan variables intermedias para representar el déficit léxico-semántico-conceptual en los dominios semánticos SV y SNV. También se han creado variables intermedias para representar el déficit léxico-semántico-conceptual para las distintas categorías semánticas (***manzana***, ***perro***, ***pino***, ***coche***, ***silla*** y ***pantalón***).
- ✓ Variables de interés o hipótesis. Son las variables cuyas probabilidades a posteriori son de interés inmediato. En los modelos causales propuestos en esta tesis, se utilizan dos variables de interés o hipótesis, que son: ***DSD*** y ***EA***.

---

5 Una variable es exhaustiva si todos los posibles valores de la variable deben representar todo su espacio de estado, es decir, una variable siempre representa un estado.
6 Una variable es mutuamente exclusiva, si ningún par de valores del conjunto puede excluir a otros valores del mismo conjunto, es decir, una variable sólo puede estar en un estado en un momento determinado.

En la Tabla 2 se representan todas las variables utilizadas para el método de diagnóstico.

Tabla 2.- Identificación de variables del corpus, variables latentes, factores de riesgo y de protección.

VARIABLES UTILIZADAS EN LAS REDES BAYESIANA		
Tipo Variable	**Variable**	**Descripción**
Variables de Información de contexto. Factores de Riesgo.	Edad	La edad es el principal factor de riesgo no modificable. Esta variable es determinista y los valores que puede contener son: ➢ De 0 a 64 años: ➢ De 65 a 69 años. ➢ De 70 a 74 años. ➢ De 75 a 79 años. ➢ De 80 a 84 años. ➢ Más de 85 años. Estos intervalos de edad se han tomado del estudio epidemiológico [18] sobre la demencia y la EA.
	Nivel Educativo	Más que un factor de riesgo se puede considerar como un factor de protección. Las personas con mayor nivel educativo, es decir, con mayor número de años de escolarización, tienen un menor riesgo de padecer la EA. Esta variable es determinista y los valores que puede contener son: ➢ Primarios. ➢ Secundarios. ➢ Universitarios.
	Sexo	El sexo se puede considerar otro factor de riesgo, aunque no está demasiado claro que exista una relación directa entre la EA y el sexo. La EA afecta sensiblemente a más mujeres que a hombres, pero hay que tener en cuenta que la esperanza de vida de las mujeres es superior a la de los hombres.
Variable de interés o hipótesis	EA	Es una de las variables objetivo y puede tomar los siguientes estados: ➢ Ausente. ➢ Presente

VARIABLES UTILIZADAS EN LAS REDES BAYESIANA		
Tipo Variable	**Variable**	**Descripción**
	DSD (deterioro semántico en sus aspectos declarativos)	Es una variable objetivo y es la variable más importante de nuestro método de diagnóstico. Los valores que puede tomar son: ➢ Ausente: El paciente no muestra ningún tipo de deterioro semántico. ➢ Presente: El paciente muestra un deterioro semántico leve o moderado. Se estudia el deterioro en el conocimiento de determinadas categorías semánticas naturales y de objetos, y de los rasgos que supuestamente las configuran, representan y organizan. Es decir, el deterioro cognitivo en el conocimiento léxico-semántico-conceptual.
Variables Intermedias o latentes	Producción oral de rasgos semánticos en las categorías de objeto: Manzana, Perro, Pino, Coche, Silla, Pantalón.	Estas variables representan el grado del déficit léxico-semántico-conceptual para cada una de las categorías semánticas. Para las BNs discretas estas variables se discretizan en los siguientes estados: ➢ Ausente: Indica que la producción oral de rasgos semánticos para la categoría semántica es normal. ➢ Presente: El paciente muestra un déficit léxico-semántico-conceptual en la producción oral de rasgos semánticos en dicha categoría semántica.

VARIABLES UTILIZADAS EN LAS REDES BAYESIANA		
Tipo Variable	**Variable**	**Descripción**
Variables de Información de contexto (síntomas).	Taxonómicos, Tipos, Parte-todo, Funcional, Evaluativo, Lugar y Hábitat, Comportamiento, Causa/Genera, Procedimental, Ciclo Vital, Otros	Estas variables representan el posible déficit léxico-semántico-conceptual en cada rasgo semántico. Los estados que puede contener estas variables son: ➢ Ausente. ➢ Presente.
Variables Intermedias	SV	Variable intermedia que representa el posible déficit léxico-semántico-conceptual en las entidades biológicas. Con esta variable también se pretende estudiar el DS diferencial que presentan algunos enfermos de EA entre los dominios semánticos, SV y SNV. Se mide por la producción oral de rasgos semánticos que produce el participante, para cada uno de las categorías semánticas referente a categorías naturales. Los valores que puede contener estas variables son: ➢ Ausente. ➢ Presente.
	SNV	Esta variable es similar a la anterior, pero representa las categorías semánticas no biológicas.

4.2 Promedios y medidas de dispersión.

Es interesante calcular la tendencia central diferenciando la muestra por tramos de edad, por nivel educativo y, entre personas sanas y personas enfermos de EA. El promedio utilizado es la media aritmética y sólo se utiliza como un medio para dar información simplificada y con carácter orientativo, en ningún caso para dar pronósticos. En este capítulo no se ha realizado una diferencia de medias porque el trabajo se centra en la IA. La media y la desviación típica son parámetros muy importantes para el modelo

cuantitativo de las BNs híbridas. Cabe destacar que para el aprendizaje automático del modelo cuantitativo de este tipo de BN, sólo se utiliza una muestra constituida por sujetos sanos (para más detalle consultar capítulo 6).

Tal y como se indicó en el capítulo 1, el corpus [1] se ha elaborado con un muestreo incidental, y para su elaboración ha sido necesario la colaboración de varios departamentos de neurología de varios hospitales de la Comunidad Autónoma de Madrid.

Algunas investigaciones [9,17] señalan que la edad y el nivel educativo pueden influir en la producción oral de rasgos semánticos. Se puede comprobar en este análisis estadístico, que las personas de mayor edad tienden producir menos rasgos semánticos; al igual que las personas con menos años de escolarización. Estas observaciones se deben modelar en las BNs, pero para ello es necesario crear más de 120 enlaces causales entre la edad y el nivel educativo, y las variables del corpus; aumentando considerablemente la complejidad de los modelos cuantitativos de las BNs. Una solución innovadora que se propone en esta tesis doctoral consiste en aprovechar el proceso de discretización para modelar esta influencia informativa y para ello, se implementa un algoritmo que gestiona una jerarquía multinivel de conglomerados de atributos dónde se agrupan los casos en función de los segmentos por edad o nivel educativo a los que pertenezcan. En cada segmento se buscan dos centroides –por análisis de clúster— para representar la ausencia o presencia del déficit léxico-semántico-conceptual.

4.2.1 Medias y desviaciones típicas para la variable de interés EA y categorías semánticas.

En la Tabla 3 se muestran las medias aritméticas de la producción oral de rasgos semánticos para categorías semánticas: ***Manzana, Perro, Pino, Coche, Silla*** *y* ***Pantalón.*** Las variables ***SV*** y ***SNV*** representan la suma de las medias de sus respectivas categorías semánticas. La media aritmética se calcula, por un lado, para la muestra de sujetos sanos y, por otro lado, para la muestra de personas enfermas de EA.

Tabla 3.- Medias aritméticas y desviaciones típicas para las variables latentes y categorías semánticas.

Media y desviación típica Sin agrupamiento de conglomerados de atributos					
	Variables	$\bar{X}_{sanos}$	$\bar{X}_{EA}$	σ_{sanos}	$\sigma_{enfermos}$
Variables Latentes	SV	*42,9762*	*18,5897*	*13,7885*	*9,662*
	SNV	*39,0238*	*16,1538*	*11,7151*	*11,0847*
	Manzana	*14,2857*	*5,2821*	*5,6018*	*2,982*
	Perro	*15,2143*	*7,4615*	*7,7066*	*5,2957*
	Pino	*13,4762*	*5,8462*	*5,9356*	*3,6673*
	Coche	*13,881*	*5,8205*	*4,9396*	*4,0515*
	Silla	*12,7857*	*4,7949*	*5,7277*	*3,9013*
	Pantalón	*12,3571*	*5,5385*	*5,2117*	*4,1729*

En la Tabla 3 se pone de manifiesto la diferencia cuantitativa en la producción oral de rasgos semánticos entre las personas sanas y las personas enfermas de EA, siendo las personas sanas las que tienen un mayor promedio de producción oral de rasgos semánticos. Por tanto, esta medida de tendencia central está en línea con las investigaciones sobre la memoria semántica y representación del conocimiento que constata diferencias cuantitativas en la producción oral de rasgos entre personas sanas y enfermos de EA [1]. Del mismo modo, se pone de manifiesto en la Tabla 3 que existe una gran dispersión en la producción oral de rasgos semánticos, tanto en la muestra de sujetos cognitivamente sanos, como en la muestra de sujetos enfermos de EA. Un grado de incertidumbre tan elevado dificulta el diseño de un método de diagnóstico basado exclusivamente en técnicas estadísticas, ya que sería poco fiable.

4.2.2 Medias y desviaciones típicas por bloque conceptual de cada categoría semántica.

Como se ha indicado anteriormente, existen bloques conceptuales comunes a todas las categorías semánticas y otros que no son comunes, tal y como queda de manifiesto en la Tabla 4 y en la Tabla 5 (líneas marcadas en rojo). En la Tabla 4 se muestran las medias y las desviaciones típicas para los bloques conceptuales del dominio semántico ***SV***. Cabe señalar las dispersiones en la producción oral de rasgos semánticos en las categorías semánticas.

En la Tabla 5 se muestran las medias y las desviaciones típicas para los bloques conceptuales del dominio semántico ***SNV***. Al igual que con el dominio semántico ***SV***, existen variables que tienen mucha dispersión en el recuento de rasgos semánticos.

Tabla 4.- Medias aritméticas y desviaciones para los bloques conceptuales del dominio semántico SV.

Media y desviación típica SV Sin agrupamiento de conglomerados de atributos					
	Variables	$\bar{X}_{sanos}$	$\bar{X}_{EA}$	σ_{sanos}	$\sigma_{enfermos}$
Manzana	taxonómico	*0,7857*	*0,359*	*0,4704*	*0,486*
	tipos	*3,2619*	*1,1282*	*2,0489*	*1,9625*
	partes	*0,2857*	*0,3077*	*0,6358*	*0,7662*
	funcional	*2,3095*	*0,9231*	*2,5991*	*1,1329*
	evaluativo	*4,4048*	*1,7179*	*3,8195*	*1,9728*
	lugar/hábitat	*0,7143*	*0,1026*	*1,2932*	*0,5024*
	conducta	***0***	***0***	***0***	***0***
	causa	*0,0952*	*0*	*0,3702*	*0*
	procedimental	*0,7619*	*0,3333*	*1,3935*	*0,8983*
	ciclo vital	*0,5476*	*0,1282*	*1,0407*	*0,5221*
	otros	*1,119*	*0,2821*	*1,4517*	*0,7236*
Perro	taxonómico	*0,9048*	*0,2564*	*0,7905*	*0,4424*
	tipos	*4,3571*	*2,359*	*3,875*	*2,9154*
	partes	*0,3571*	*0,8974*	*0,9582*	*1,7591*
	funcional	*2,2619*	*1,1795*	*2,2638*	*1,6839*
	evaluativo	*4,1905*	*2*	*3,5902*	*2,1398*
	lugar/hábitat	*0,381*	*0,1026*	*0,6228*	*0,3835*
	conducta	*1*	*0,5385*	*1,546*	*1,1203*
	causa	***0***	***0***	***0***	***0***
	procedimental	*0,0714*	*0*	*0,4629*	*0*
	ciclo vital	*0,0714*	*0,0256*	*0,2607*	*0,1601*
	otros	*1,619*	*0,1026*	*1,6521*	*0,3074*
Pino	taxonómico	*0,9048*	*0,2564*	*0,6917*	*0,4424*
	tipos	*2,2381*	*0,7692*	*2,3144*	*1,5297*
	partes	*0,9762*	*0,6923*	*1,1788*	*1,2173*
	funcional	*2,5952*	*0,8974*	*2,7767*	*1,3138*
	evaluativo	*2,5952*	*1,0513*	*2,3691*	*1,05*
	lugar/hábitat	*1,2619*	*0,2564*	*1,5627*	*0,9095*
	conducta	***0***	***0***	***0***	***0***
	causa	*1,4048*	*1,4103*	*1,1699*	*1,4994*
	procedimental	*0,0476*	*0,0256*	*0,3086*	*0,1601*
	ciclo vital	*0,5238*	*0,2821*	*1,2733*	*0,7591*
	otros	*0,9286*	*0,2051*	*1,2953*	*0,5221*

Tabla 5.- Medias aritméticas y desviaciones para los tipos de rasgos del dominio semánticos SNV.

Media y desviación típica SNV Sin agrupamiento de conglomerados de atributos					
	Variables	$\bar{X}_{sanos}$	$\bar{X}_{EA}$	σ_{sanos}	$\sigma_{enfermos}$
Coche	taxonómico	*0,7381*	*0,1282*	*0,6648*	*0,3387*
	tipos	*4,1429*	*1,5385*	*3,3755*	*1,8757*
	partes	*2,8333*	*1,4103*	*3,8249*	*2,9084*
	funcional	*2,0952*	*1,2821*	*2,3144*	*1,9594*
	evaluativo	*1,6905*	*0,8205*	*1,7319*	*1,0227*
	lugar/hábitat	***0,119***	***0***	***0,5501***	***0***
	conducta	*0,0476*	*0*	*0,3086*	*0*
	causa	*0,0238*	*0*	*0,1543*	*0*
	procedimental	*0,1667*	*0,2564*	*0,5809*	*0,938*
	ciclo vital	***0***	***0***	***0***	***0***
	otros	*2,0238*	*0,3846*	*1,9937*	*0,8771*
Silla	taxonómico	*0,7143*	*0,0769*	*0,742*	*0,27*
	tipos	*5,0714*	*2,5641*	*4,1284*	*3,4701*
	partes	*1,1429*	*0,8718*	*1,6903*	*1,7042*
	funcional	*2,1905*	*1,1026*	*2,6617*	*1,6669*
	evaluativo	*1,7143*	*0,641*	*1,7431*	*1,0384*
	lugar/hábitat	***0,0476***	***0***	***0,2155***	***0***
	conducta	*0,0476*	*0*	*0,3086*	*0*
	causa	***0***	***0***	***0***	***0***
	procedimental	*0,119*	*0,1282*	*0,3952*	*0,3387*
	ciclo vital	***0***	***0***	***0***	***0***
	otros	*1,3095*	*0,1538*	*1,7035*	*0,4315*
Pantalón	taxonómico	*0,7381*	*0,2308*	*0,8851*	*0,4268*
	tipos	*6,1429*	*1,6923*	*5,4664*	*2,6473*
	partes	*1,1905*	*0,8205*	*1,5963*	*1,2747*
	funcional	*2,0714*	*1,1538*	*1,5364*	*1,2039*
	evaluativo	*1,6667*	*0,7179*	*2,0561*	*1,2128*
	lugar/hábitat	*0,5*	*0,0513*	*1,3837*	*0,3203*
	conducta	***0***	***0***	***0***	***0***
	causa	***0***	***0***	***0***	***0***
	procedimental	*0,1905*	*0,0513*	*0,6713*	*0,2235*
	ciclo vital	*0,0238*	*0*	*0,1543*	*0*
	otros	*0,2619*	*0,0769*	*0,5868*	*0,27*

De este análisis estadístico se obtienen las medias y desviaciones típicas que serán utilizadas en las BNs híbridas. Con este análisis estadístico no se pretende probar ninguna hipótesis, será con las técnicas de IA con las que se busquen patrones complejos en la producción oral de rasgos semánticos.

4.2.3 Medias de la producción oral de rasgos, segmentando por edad.

Existen distintos motivos que podrían explicar porque algunas personas realizan unas definiciones orales de objetos básicos pobres y que además, no están relacionadas con la EA u otras ENs. Las definiciones orales solicitadas a los sujetos son relativas a categorías naturales y objetos básicos, que cualquier persona conoce perfectamente y que debería poder describir sin dificultad alguna. Sin embargo, se ha podido comprobar que un bajo nivel educativo o la propia vejez cerebral, puede influir en la capacidad del individuo a la hora de realizar definiciones orales, aun siendo objetos básicos de uso cotidiano.

Este análisis estadístico tiene una mayor utilidad de aplicación sobre la muestra de sujetos sanos, porque una vez que la enfermedad está presente, el daño cerebral se vuelve tan omnipresente que el DC es significativo, independientemente de la edad o el nivel educativo.

Producción oral de rasgos semánticos del test oral de todas las categorías semánticas.

En la Figura 2 se muestran la suma de las medias aritméticas de la producción oral de rasgos semánticos de todas las categorías semánticas, segmentando el recuento de rasgos semánticos por tramos de edad y estado cognitivo. Se puede apreciar una tendencia a la baja, en la producción oral de rasgos semánticos, en los tramos de edad más avanzada, excepto para el tramo correspondiente a los mayores de 85 años, en el que sólo tenemos dos casos, tal y como se puede comprobar en la Tabla 1.

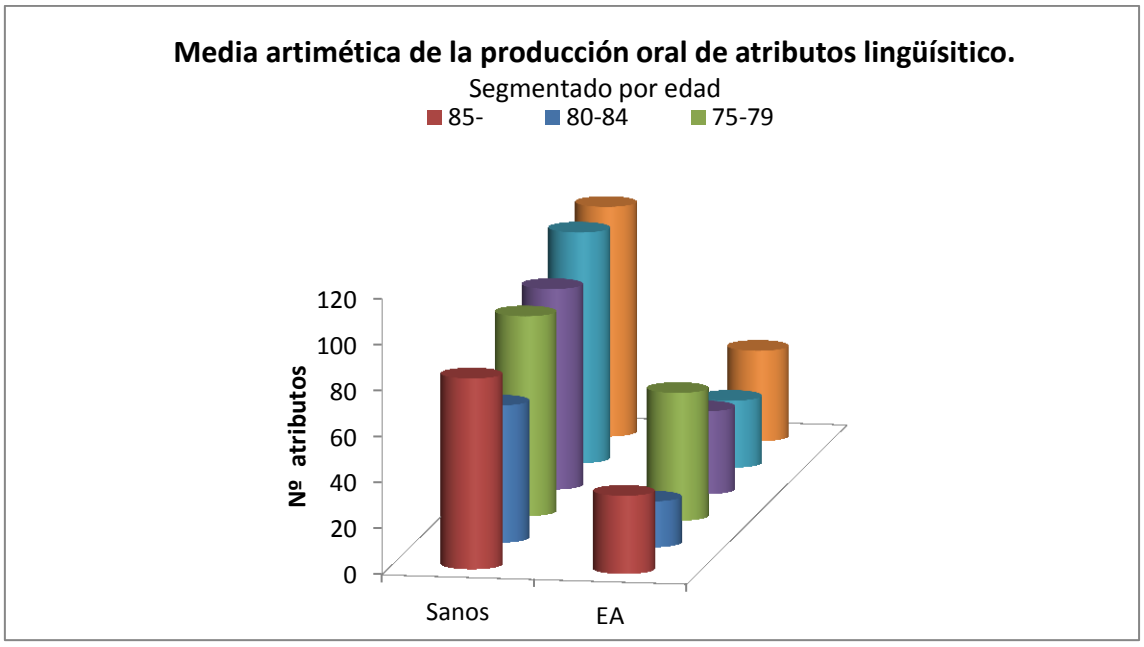


Figura 2.- Media aritmética de la producción oral de rasgos semánticos.

Producción oral de rasgos semánticos para cada dominio semántico.

En la Figura 3 se calculan las medias aritméticas de la producción oral de rasgos semánticos segmentando la muestra por tramos de edad, por los dominios semánticos (SV y SNV) y el estado cognitivo. En la Figura 3 se muestra una tendencia negativa en la producción oral de rasgos semánticos conforme aumenta la edad de los participantes.

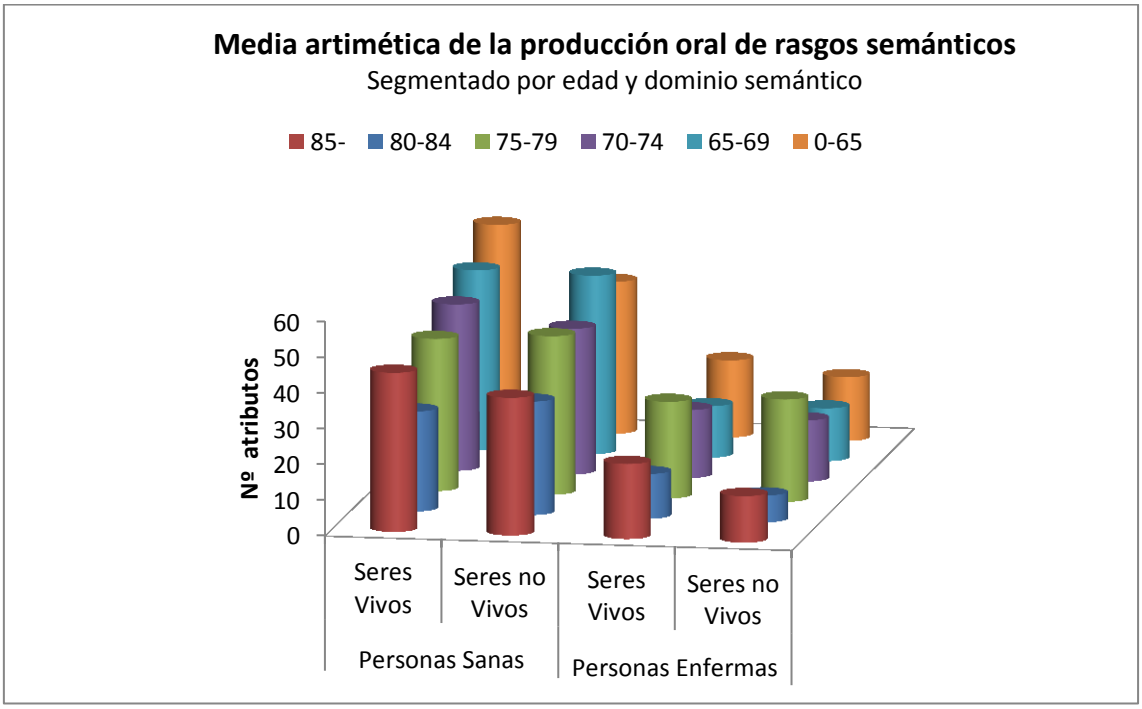


Figura 3.- Media aritmética de la producción oral de rasgos semánticos, segmentando por edad y dominio semántico.

Producción oral de rasgos semánticos para cada categoría semántica.

En la Figura 4 se calcula la media aritmética de la producción oral de rasgos semánticos segmentando la muestra por categoría semántica y edad. Algunos segmentos de edad, como por ejemplo, el comprendido entre 75 y 79 años de la categoría semántica ***perro***, existe una persona enferma de EA que produce 23 rasgos semánticos. Esto hace que se dispare la media, ya que en este segmento sólo existen tres participantes.

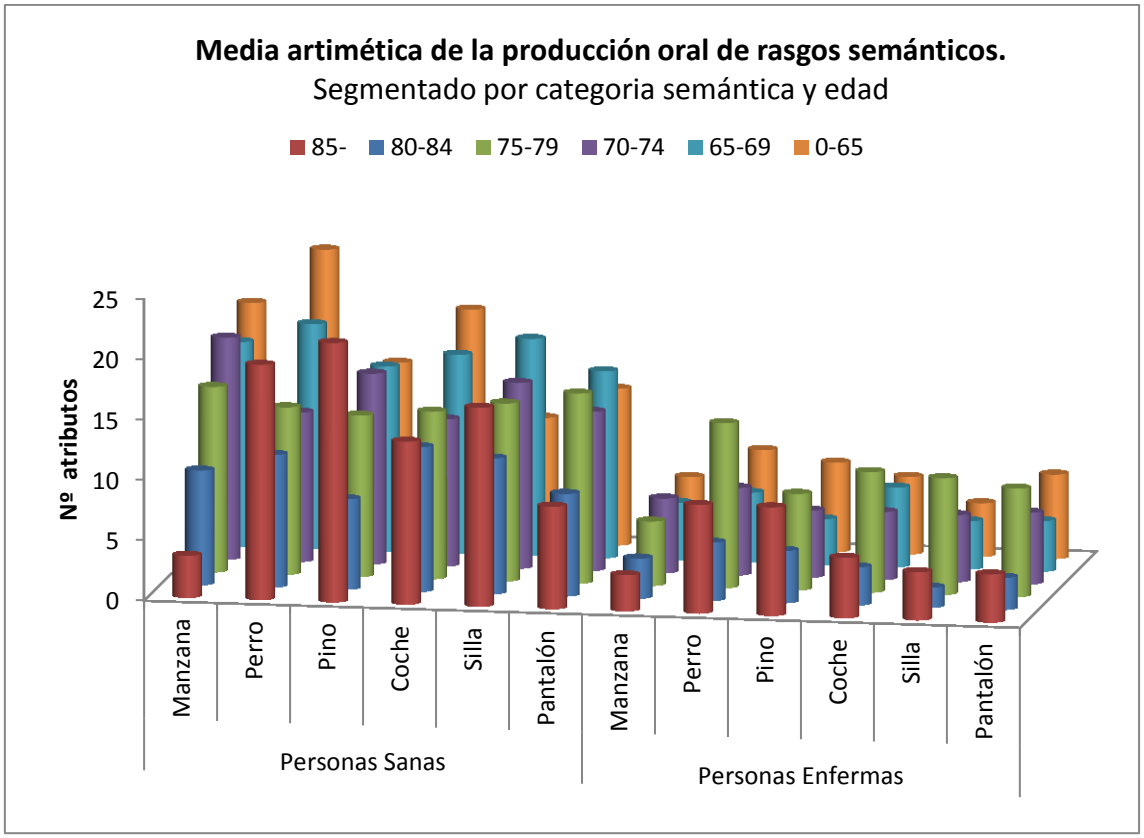


Figura 4.- Media aritmética de la producción oral de rasgos semánticos, segmentando por categoría semántica y edad.

En la Tabla 6 se muestran los valores cuantitativos de las medias aritméticas segmentados por edad.

Tabla 6.- Media aritmética de la producción oral de rasgos semánticos, segmentando por edad.

	Variables de interés						
	Segmentación del recuento de atributos por edad						
	Variables	85-	80-84	75-79	70-74	65-69	0-65
Personas Sanas	Seres Vivos	44,5	26,333	39,7	46,833	45,875	51,3
	Seres no Vivos	38,5	28,167	41,2	40,167	44,5	38,4
	Manzana	3,5	9	15	17,167	16	15,8
	Perro	19,5	9,8333	12,7	14,167	15,625	20,4
	Pino	21,5	7,5	12	15,5	14,25	15,1
	Coche	13,5	12,333	13,2	12,333	14,375	16,1
	Silla	16,5	8,8333	13,6	15,167	16,25	9,4
	Pantalón	8,5	7	14,4	12,667	13,875	12,9
	Variables de interés						
	Segmentación del recuento de atributos por edad						
	Variables	85-	80-84	75-79	70-74	65-69	0-65
Personas Enfermas	Seres Vivos	21	12,5	27	19	14,4	21,5714
	Seres no Vivos	13	7,5	28,6667	17,125	14,8	17,8571
	Manzana	3	3,3333	5,3333	6,1875	4,8	5,8571
	Perro	9	4,8333	13,6667	7,25	5,8	8,2857
	Pino	9	4,3333	8	5,5625	3,8	7,4286
	Coche	5	3,1667	10	5,625	6,6	6,4286
	Silla	4	1,6667	9,6667	5,5625	4	4,4286
	Pantalón	4	2,6667	9	5,9375	4,2	7

4.2.4 Medias de la producción oral de rasgos, segmentando por nivel educativo.

Esta sección analiza la producción oral de rasgos semánticos segmentando la muestra por nivel educativo y categoría semántica. Al igual que en la sección anterior, el análisis se realiza tanto para la muestra de sujetos sanos como para la muestra de sujetos enfermos de EA, aunque por las mismas razones que las expuestas en el apartado anterior, es de mayor interés la muestra de sujetos sanos.

Producción oral de rasgos semánticos del test oral para todas las categorías semánticas.

En la Figura 5 se muestra una tendencia al alza, en la producción oral de rasgos semánticos, en las personas con un mayor nivel educativo.

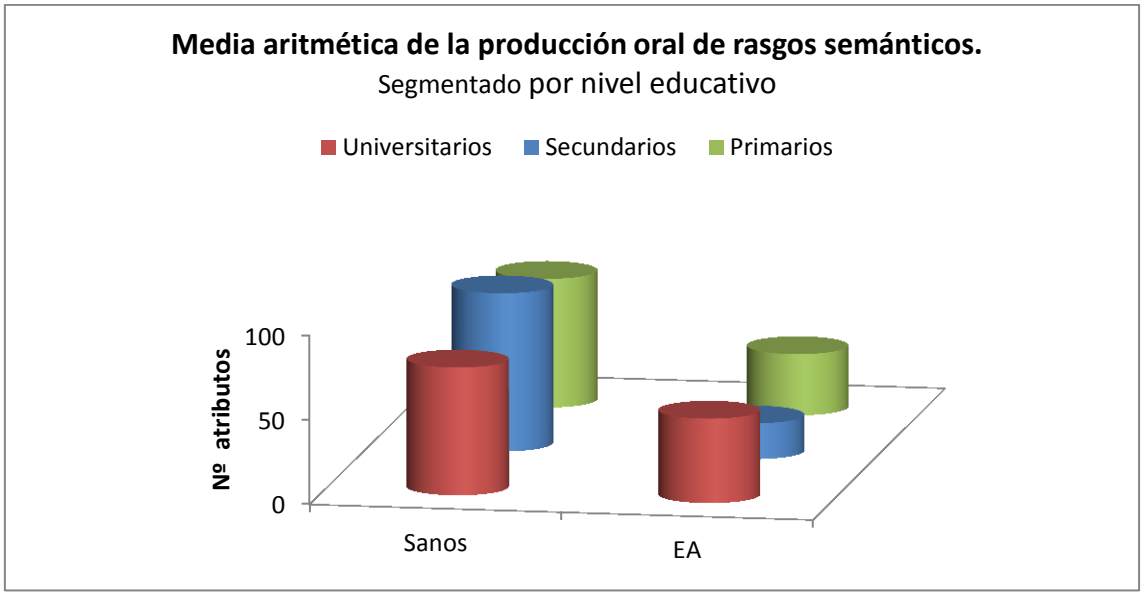


Figura 5.- Media aritmética de la producción oral de rasgos semánticos, segmentando por nivel educativo.

En la Figura 6 se muestra la media aritmética de la producción oral de rasgos semánticos segmentando la muestra por nivel educativo, enfermos de EA y personas cognitivamente sanas. Se muestra una tendencia a la baja, en la producción oral de rasgos semánticos, en las personas con sólo tienen estudios primarios.

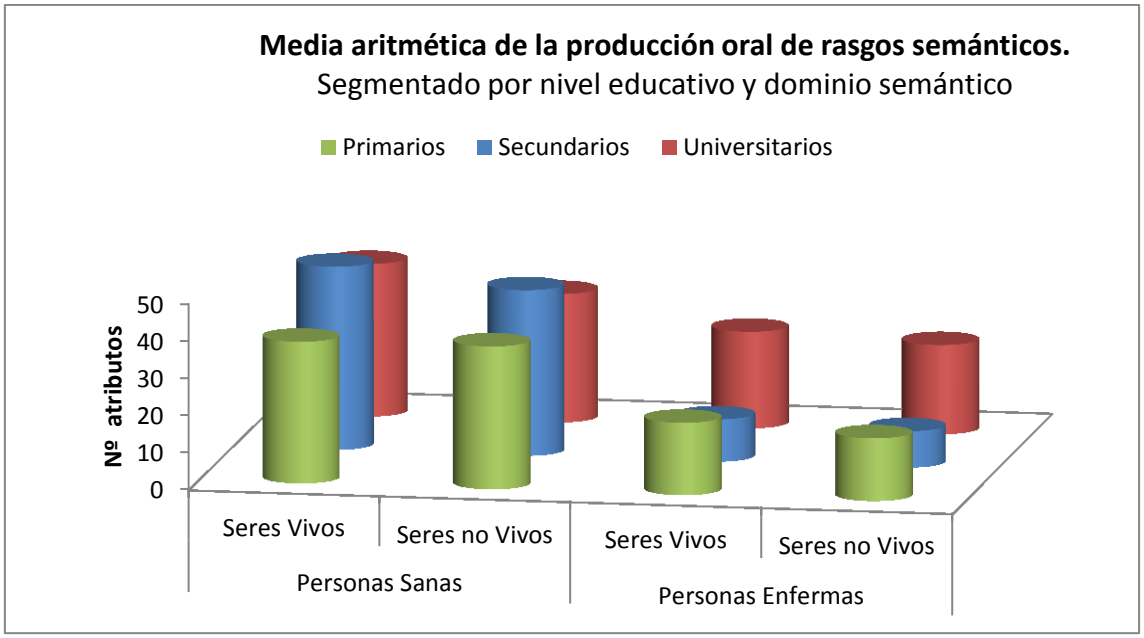


Figura 6.- Media aritmética de la producción oral de rasgos semánticos, segmentando por nivel educativo, dominio semántico, personas cognitivamente sanas y enfermas de EA.

En la Figura 7 se representan las medias aritméticas de la producción oral de rasgos semánticos, por cada categoría semántica, segmentando la muestra por nivel educativo y categoría semántica. Al igual que en la figura anterior, se pone de manifiesto que las personas que sólo tienen estudios primarios suelen producir menos atributos en todas las categorías semánticas.

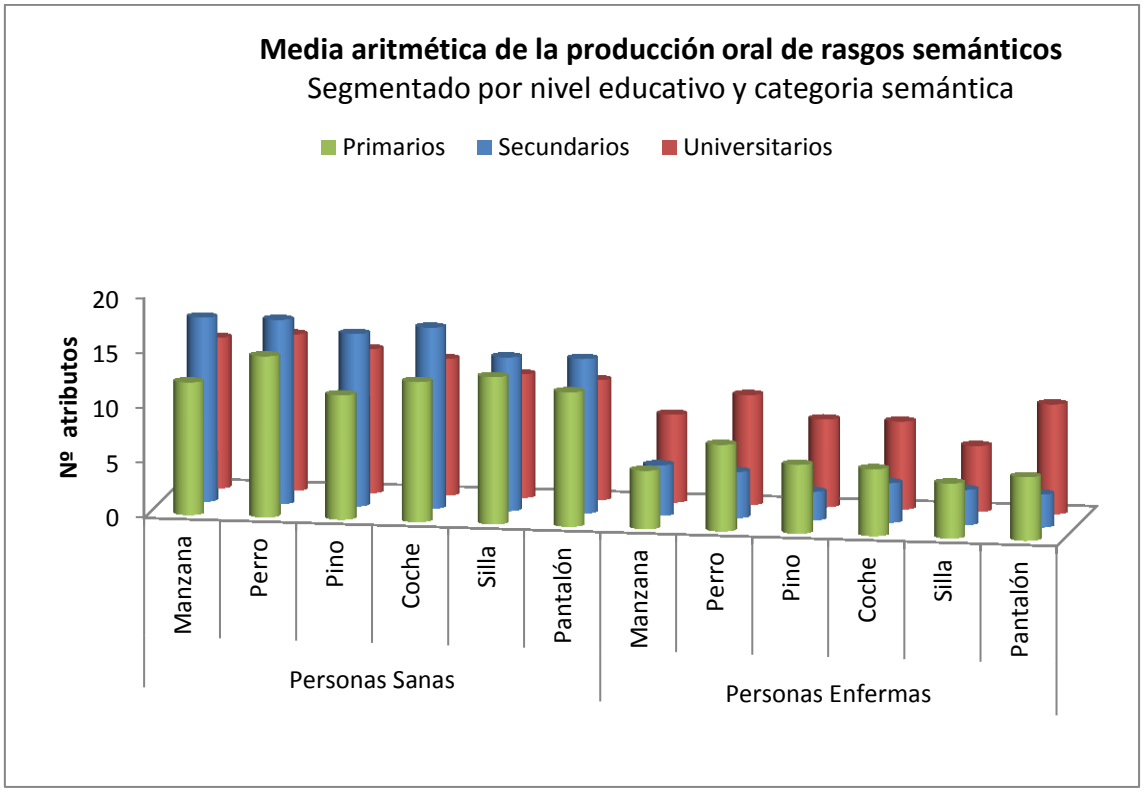


Figura 7.- Media aritmética de la producción oral de rasgos semánticos segmentados por nivel educativo (estudios primarios, secundarios y universitarios), categoría semántica, personas cognitivamente sanas y enfermas de EA.

En la Tabla 7 se muestran las medias aritméticas de la producción oral de rasgos semánticos por nivel educativo y categoría semántica.

Tabla 7.- Producción oral de rasgos semánticos segmentada por nivel educativo.

Personas Sanas	Media producción oral de rasgos semánticos Segmentación del recuento de atributos por nivel educativo			
	Variables	**Universitarios**	**Secundarios**	**Primarios**
	Seres Vivos	*41,125*	*49,2857*	*38,0833*
	Seres no Vivos	*34,625*	*44,5714*	*38,4167*
	Manzana	*13,75*	*16,7857*	*12,0833*
	Perro	*14,25*	*16,7857*	*14,6667*
	Pino	*13,125*	*15,7143*	*11,3333*
	Coche	*12,4375*	*16,5*	*12,75*
	Silla	*11,25*	*14*	*13,4167*
	Pantalón	*10,9375*	*14,0714*	*12,25*

Personas Enfermas	Media producción oral de rasgos semánticos Segmentación del recuento de atributos por nivel educativo			
	Variables	**Universitarios**	**Secundarios**	**Primarios**
	Seres Vivos	*26*	*11,4*	*19,4545*
	Seres no Vivos	*24*	*9,8*	*16,8788*
	Manzana	*8*	*4,6*	*5,303*
	Perro	*10*	*4,2*	*7,8788*
	Pino	*8*	*2,6*	*6,2727*
	Coche	*8*	*3,6*	*6,0909*
	Silla	*6*	*3,2*	*5*
	Pantalón	*10*	*3*	*5,7879*

Podemos concluir que las personas que sólo tienen estudios primarios realizan definiciones orales más pobre que las personas que han estudiado durante más años. Por tanto, es necesario que el método de diagnóstico propuesto en esta tesis doctoral tenga en cuenta tales circunstancias.

4.3 Coeficientes de correlación y asociación.

El objetivo de los coeficientes de correlación o asociación es ponderar el nivel de importancia de cada una de las variables del corpus, en función de su grado de asociación con la EA, es decir, se le da mayor peso a aquellas variables que ayudan a predecir mejor la EA. Con objeto de contrastar resultados se calculan varios métodos para la ponderación de la importancia de atributos en el proceso de inferencia. Los coeficientes utilizados en esta tesis doctoral son:

- **Coeficiente de correlación de *Pearson*.** Mide la intensidad de la asociación observada entre las variables del corpus y la EA.
- **Distancia euclídea modificada.** Una vez calculada la probabilidad de padecer un déficit léxico-semántico-conceptual en cada bloque conceptual con la función de distribución gaussiana, se calcula la distancia euclídea modificada respecto a su valor esperado. El valor esperado se informa con un 1, cuando el sujeto ha sido diagnosticado por los neurólogos como EA presente, y 0, cuando el sujeto está cognitivamente sano.
- **Ponderación por producción oral de rasgos semánticos.** Es el método más sencillo, calcula el peso de cada variable respecto a la producción oral de rasgos semánticos.
- **Coeficiente de regresión simple.** Estos coeficientes se calculan a partir de dos fórmulas de regresión simple, una para la muestra de sujetos sanos y la otra para muestra de sujetos enfermos de EA.
- **Ganancia de información.** Mide la efectividad de las variables en el diagnóstico de la EA.

Coeficiente de Pearson.

El coeficiente de correlación de Pearson es una medida de similitud entre dos muestras. El coeficiente de Pearson mide la relación entre dos variables aleatorias cuantitativas. A cada una de estas variables, antes de calcular el coeficiente de correlación, se les aplica una función de distribución gaussiana (ver detalles en el capítulo 6). Por otro lado, el grado de la enfermedad es una variable cualitativa y es necesaria su transformación a un valor cuantitativo. El coeficiente de correlación de Pearson proporciona a nuestro sistema de diagnóstico de un método estadístico multivariante que determina la contribución de varios factores –déficits léxico-semánticos-conceptuales de los rasgos semánticos— al déficits léxico-semánticos-conceptuales de las categorías naturales u objetos básicos. Las variables con un mayor grado de correlación con la enfermedad tendrán una mayor influencia en el diagnóstico que las variables con un coeficiente de correlación menor (ver detalles en el capítulo 6).

Hay que tener en cuenta que la EA causa un DC que afecta a la capacidad de producir rasgos semánticos en definiciones orales simples. No obstante, nos interesa conocer cuál es el grado de esta relación y, como puede aumentar o disminuir el grado de correlación en función del estadio de la enfermedad.

Estos coeficientes serán utilizados en el capítulo 6 para construir relaciones lineales con variables latentes. Cabe destacar que se han utilizado simplificadores en el aprendizaje automático del modelo cuantitativo de las BNs. Estos simplificadores se han combinado con un método de aprendizaje aproximado, ya que los parámetros de la TPC se aprenden combinando determinadas hipótesis y datos cuantitativos.

Distancia euclídea modificada.

La distancia euclídea es otro de los coeficientes que se ha utilizado para ponderar el peso de las variables en el proceso de inferencia. Al igual que en el cálculo del coeficiente de correlación de Pearson, para calcular la distancia euclídea se utiliza por un lado, una función de distribución gaussiana sobre cada variable y, por otro lado, el valor esperado en función de la presencia o ausencia de la enfermedad.

La distancia euclidiana es la medida más conocida de las distancias métricas. Puede calcularse por medio del teorema de Pitágoras, el cual define la relación (n-dimensional) entre puntos en un espacio euclidiano. La fórmula utilizada en esta tesis doctoral es:

$$d_E(X,Y) = \frac{\sqrt{(1-abs(y_1-x_1))^2 + (1-abs(y_2-x_2))^2 + \cdots + (1-abs(y_n-x_n))^2}}{n} \quad (4.1)$$

donde

> *x:* es la variable obtenida tras la aplicación de una función de distribución gaussiana. Originalmente esta variable parte del recuento de la producción oral de rasgos semánticos.

y: es el valor esperado que toma los valores 0, cuando la EA está ausente, y 1, cuando la EA está presente.

Esta fórmula contiene algunas modificaciones respecto a la fórmula original, en primer lugar, la distancia euclidiana no tiene límite superior, se incrementa conforme aumenta el número de variables. Su valor sólo depende de la escala de cada variable, por esa razón optamos por modificar la fórmula para obtener la distancia euclídea promedio, de tal forma que el valor que puede tomar la variable estaría comprendido en el rango [0,1]. En segundo lugar, se resta a 1 el valor absoluto de la distancia entre la función de distribución gaussiana y el valor esperado. La mejor forma de explicar esta modificación es mediante dos ejemplos:

- Supongamos una persona sana con una probabilidad de padecer la EA de 0,3, la fórmula seria: 1-(0-0,3) = 0,7. El valor esperado sería 0 y la probabilidad de estar cognitivamente sano es mejor cuanto más cerca se encuentra de 0. Como se debe dar un mayor peso a aquellas variables que más cerca están de 0 entonces la modificación consiste en restar dicha probabilidad a la unidad.
- Para una persona enferma con una probabilidad de padecer la EA de 0,7, la fórmula sería 1-(1-0,7)=0,7. Este caso es similar al anterior, salvo que la probabilidad de padecer la EA es mejor cuanto más cerca se encuentra de 1.

Ponderación de atributos.

Como se ha indicado anteriormente, este método asigna un mayor peso a aquellos rasgos y categorías semánticas en los que los participantes han producido un mayor número de rasgos semánticos. Por ejemplo, el peso que tendría la variable *taxonómico* de la categoría semántica ***manzana*** sería:

$$w_{taxonómico,manzana} = \frac{\sum_{i=1}^{N} f_i(taxonomico, manzana)}{N} \quad (4.2)$$

donde

$w_{taxonómico,manzana}$**:** Peso para la variable *taxonómico* de la categoría semántica ***manzana***.

$f(taxonomico, manzana)$: Función que representa el recuento de la producción oral de rasgos semánticos, por ejemplo para rasgo semántico *taxonómico* de la categoría semántica ***manzana***.

N*:* Es la producción oral de rasgos semánticos.

Ganancia de información.

En primer lugar, se define la entropía como una medida de pureza/contaminación de una colección arbitraria de elementos. El propósito del método de diagnóstico es detectar la presencia o ausencia de la EA en función de la presencia o ausencia del DSD. Se define la entropía relativa como:

$$Entropy(S) = -p_{\oplus} log_2 p_{\oplus} - p_{\ominus} log_2 p_{\ominus} \quad (4.3)$$

donde

S: Es el conjunto de casos del corpus.

$p_{\oplus}$: Es la proporción de enfermos de EA en S.

$p_{\ominus}$*:* Es la proporción de personas cognitivamente sanas en S.

Dada la entropía como medida de la contaminación de una colección de ejemplos de entrenamiento, se puede definir la ***ganancia de información*** como una medida de efectividad de un atributo en la clasificación de los datos de entrenamiento. La ganancia de información es simplemente la reducción esperada en la entropía causada por la partición de los ejemplos acorde a este atributo. De forma más precisa, la ganancia de información *Gain(S,A)* de un atributo *A* relativo a una colección de ejemplos *S* se define como:

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \qquad (4.4)$$

donde

v: Son los valores distintos que puede tomar el atributo *A*.

S_v: Es el subconjunto de ***S*** donde el atributo *A* toma el valor ***v***.

Cabe destacar que en este análisis utilizamos variables continuas como resultado de calcular la distribución de probabilidad normal o Gaussiana, al recuento de la producción oral rasgos semánticos de cada categoría semántica y rasgo semántico. Por el contrario, la clase a predecir es discreta, concretamente puede tomar dos valores la ausencia o presencia del DC. En este caso el algoritmo de Ganancia de información divide en múltiples intervalos el atributo continuo basándose en un umbral. Para este caso utiliza un ratio de ganancia de información [41] donde aparece el concepto *split information*:

$$SplitInformation(S,A) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|} \qquad (4.5)$$

donde

S_1 a S_c: son los *c* subconjuntos de las instancias resultantes de particionar *S* por el valor del atributo *c.*

El ratio de ganancia de información queda definido como:

$$Gain\ Ratio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)} \qquad (4.6)$$

Regresión simple.

Con la regresión simple se considera de forma independiente la muestra de sujetos sanos y la muestra de sujetos enfermos de EA. La regresión simple estima un modelo lineal con una variable independiente a partir de los mínimos cuadrados.

$$y = slope * x \qquad (4.7)$$

Las observaciones se generan con los pares (x, y) donde *x,* se calcula con una distribución normal de la producción oral de rasgos semánticos de la variable en cuestión, e *y,* es el valor esperado.

El objetivo de este algoritmo es obtener dos expresiones lineales, la primera a), para calcular la probabilidad predicha de estar cognitivamente sano; y la segunda b), para calcular la probabilidad predicha de estar cognitivamente enfermo de EA.

a) $P_{corregida}(sano) = \alpha * r_1 * P(sano)$

b) $P_{corregida}(ea) = \alpha * r_2 * P(sano)$

donde

r_1 y r_2: Usa el algoritmo de regresión simple definido en [42,43].

α: Factor de corrección.

$P_{corregida}$: Probabilidad corregida con el coeficiente de regresión para acercarse a su valor esperado.

P_{sano} y *P_{ea}:* Probabilidad calculada con una distribución de probabilidad gaussiana a partir del recuento de la producción oral de rasgos semánticos.

La regresión simple se utiliza en las BN híbridas para ponderar los pesos de las variables de la BN (ver detalles en el capítulo 6 y 9).

En la Tabla 8 y en la Tabla 9 se representan los coeficientes definidos anteriormente para cada variable del corpus. En el capítulo 6 se puede consultar el detalle de cómo se utilizan dichos coeficientes en el proceso de inferencia de dichas BNs híbridas.

Variables de Interés.

En la Tabla 8 se detallan los coeficientes de ponderación utilizados en las BN híbridas para las variables de interés y categorías semánticas. A título de ejemplo podemos observar cómo según la ganancia de información, Pearson y distancia euclídea; el objeto ***manzana*** es el que el mejor coeficiente de asociación o correlación con la EA, sin embargo la categoría semántica en la que se producen más atributos es ***perro***. Los coeficientes de regresión no se tienen en cuenta a la hora de determinar las variables más importantes para predecir la EA, ya que se tratan de coeficientes de regresión, no de asociación o correlación. Será en el apartado III donde se analizará en detalle los resultados de la utilización de estos coeficientes con las técnicas de IA.

Tabla 8.- Coeficientes para las variables de interés y categorías semánticas.

Variables de interés Sin agrupamiento de conglomerados de atributos						
Variables	**Pearson**	**Euclídea**	**Ponderación de Atributos**	**Ganancia de información**	**Regresión**	
					Sanos	**EA**
Seres Vivos	*0,6788*	*0,7651*	*0,5272*	*0,391*	*1,4949*	*1,0702*
Seres no Vivos	*0,6717*	*0,766*	*0,4728*	*0,449*	*1,5064*	*1,0647*
Manzana	*0,6923*	*0,7636*	*0,3186*	*0,388*	*1,496*	*1,0782*
Perro	*0,4872*	*0,6989*	*0,3676*	*0,344*	*1,5319*	*1,1786*
Pino	*0,6123*	*0,7302*	*0,3138*	*0,345*	*1,5004*	*1,1398*
Coche	*0,643*	*0,7473*	*0,357*	*0,415*	*1,556*	*1,0859*
Silla	*0,6024*	*0,7474*	*0,3191*	*0,267*	*1,4758*	*1,098*
Pantalón	*0,557*	*0,7241*	*0,3239*	*0,238*	*1,4972*	*1,1396*

En la Tabla 8 se puede comprobar que la categoría semántica que tiene un mayor grado de asociación con la EA, es la categoría semántica ***Manzana***; con un coeficiente de correlación de Pearson de 0,6923 y un ratio de la ganancia de información de 0,388. Por el contrario, la categoría semántica que peor grado de asociación tiene con la EA, es la categoría semántica ***Perro***; con un coeficiente de correlación de Pearson de 0,4872 y un ratio de la ganancia de información de 0,344.

Categoría semántica manzana.

En esta categoría semántica se pueden apreciar diferencias importantes en los coeficientes de correlación de *Pearson* y los coeficientes de la distancia euclídea. Una explicación al peor comportamiento de los coeficientes de correlación de *Pearson* puede ser que *Pearson* mide el grado de correlación lineal, pero no tiene en cuenta las relaciones no lineales. En la Tabla 9 se muestran los distintos coeficientes para la categoría semántica ***Manzana.***

Tabla 9.- Coeficientes para los rasgos semánticos de la categoría semántica manzana.

Manzana Sin agrupamiento de conglomerados de atributos						
Variables	**Pearson**	**Euclídea**	**Ponderación de Atributos**	**Ganancia de información**	**Regresión**	
					Sanos	**EA**
taxonómico	*0,4115*	*0,6967*	*0,0583*	*0,124*	*1,4781*	*1,1737*
tipos	*0,4695*	*0,721*	*0,2246*	*0,236*	*1,4553*	*1,1302*
partes	*0,0322*	*0,5598*	*0,0298*	*0*	*1,7457*	*1,5141*
funcional	*0,3235*	*0,6355*	*0,165*	*0,14*	*1,5101*	*1,3788*
evaluativo	*0,4158*	*0,6604*	*0,3127*	*0,134*	*1,5545*	*1,2834*
lugar/hábitat	*0,323*	*0,6067*	*0,0422*	*0,14*	*1,7101*	*1,4203*
conducta	*0*	*0*	*0*	*0*	*0*	*0*
causa	*0,189*	*0,536*	*0,005*	*0*	*2,0217*	*1,6625*
procedimental	*0,1801*	*0,5883*	*0,0558*	*0*	*1,6547*	*1,4818*
ciclo vital	*0,3043*	*0,5933*	*0,0347*	*0,116*	*1,791*	*1,4473*
otros	*0,3718*	*0,6407*	*0,072*	*0,191*	*1,5744*	*1,3405*

Categoría semántica perro.

En la Tabla 10 se muestran los distintos coeficientes para la categoría semántica ***Perro***. Según el ratio de ganancia de información, los rasgos semánticos que se desechan son: *partes, funcional, lugar/hábitat, conducta, causa, procedimental y ciclo vital*. Los rasgos semánticos que mejor predicen la EA son: *taxonómico, tipos, evaluativo y otros.*

Tabla 10.- Coeficientes para los rasgos semánticos de la categoría semántica perro.

Perro						
Sin agrupamiento de conglomerados de atributos						
Variables	Pearson	Euclídea	Ponderación de Atributos	Ganancia de información	Regresión	
					Sanos	EA
taxonómico	*0,4167*	*0,7151*	*0,04*	*0,126*	*1,4441*	*1,1853*
tipos	*0,3912*	*0,6779*	*0,3187*	*0,151*	*1,5641*	*1,263*
partes	*-0,1888*	*0,5209*	*0,0568*	*0*	*1,7852*	*1,5698*
funcional	*0,2437*	*0,6394*	*0,1445*	*0*	*1,5769*	*1,3379*
evaluativo	*0,456*	*0,6977*	*0,2813*	*0,209*	*1,6193*	*1,2082*
lugar/hábitat	*0,2852*	*0,632*	*0,0194*	*0*	*1,5918*	*1,3931*
conducta	*0,2575*	*0,6215*	*0,0748*	*0*	*1,6316*	*1,4127*
causa	*0*	*0*	*0*	*0*	*0*	*0*
procedimental	*0*	*0*	*0*	*0*	*0*	*0*
ciclo vital	*0,1577*	*0,5611*	*0,0052*	*0*	*1,9392*	*1,5913*
otros	*0,5151*	*0,6796*	*0,0594*	*0,331*	*1,5841*	*1,2952*

Categoría semántica pino.

En la Tabla 11 se muestran los distintos coeficientes para la categoría semántica ***Pino***. Según el ratio de ganancia de información, los rasgos semánticos que se desechan son: *partes, conducta, causa, procedimental, ciclo vital* y *otros*. Los rasgos semánticos que mejor predice la EA son: *taxonómico, tipos, funcional, evaluativo y lugar/hábitat.*

Tabla 11.- Coeficientes para los rasgos semánticos de la categoría semántica manzana pino.

Pino						
Sin agrupamiento de conglomerados de atributos						
Variables	Pearson	Euclídea	Ponderación de Atributos	Ganancia de información	Regresión	
					Sanos	EA
taxonómico	*0,5008*	*0,6935*	*0,0605*	*0,174*	*1,5584*	*1,1935*
tipos	*0,3665*	*0,6534*	*0,1562*	*0,139*	*1,5791*	*1,2778*
partes	*0,1657*	*0,6019*	*0,0856*	*0*	*1,6891*	*1,3614*
funcional	*0,3687*	*0,6396*	*0,1814*	*0,0995*	*1,6086*	*1,3269*
evaluativo	*0,4046*	*0,6498*	*0,1889*	*0,26*	*1,5524*	*1,3264*
lugar/hábitat	*0,4172*	*0,6514*	*0,0793*	*0,219*	*1,6233*	*1,2898*
conducta	*0*	*0*	*0*	*0*	*0*	*0*
causa	*0,0271*	*0,6016*	*0,1436*	*0*	*1,4945*	*1,3238*

Pino						
Sin agrupamiento de conglomerados de atributos						
Variables	Pearson	Euclídea	Ponderación de Atributos	Ganancia de información	Regresión	
					Sanos	EA
procedimental	*-0,0058*	*0,5075*	*0,0038*	*0*	*2,1356*	*1,7816*
ciclo vital	*0,1094*	*0,5552*	*0,0416*	*0*	*1,8228*	*1,5795*
otros	*0,3658*	*0,6282*	*0,0592*	*0,111*	*1,6494*	*1,3603*

Categoría semántica coche.

En la Tabla 12 se muestran los distintos coeficientes para la categoría semántica ***Coche***. Según el ratio de la ganancia de información, los rasgos semánticos que se desechan son: *partes, funcional, lugar/hábitat, conducta, causa, procedimental y ciclo vital*. Los rasgos semánticos que mejor predice la EA son: *tipos, taxonómico* y *otros*.

Tabla 12.- Coeficientes para los rasgos semánticos de la categoría semántica manzana coche.

Coche						
Sin agrupamiento de conglomerados de atributos						
Variables	Pearson	Euclídea	Ponderación de Atributos	Ganancia de información	Regresión	
					Sanos	EA
taxonómico	*0,5104*	*0,7036*	*0,0444*	*0,205*	*1,4833*	*1,1938*
tipos	*0,4334*	*0,68*	*0,2889*	*0,18*	*1,4687*	*1,2541*
partes	*0,2376*	*0,5982*	*0,2148*	*0*	*1,744*	*1,408*
funcional	*0,2065*	*0,6151*	*0,1704*	*0*	*1,5631*	*1,3838*
evaluativo	*0,2962*	*0,6279*	*0,1272*	*0*	*1,5863*	*1,359*
lugar/hábitat	*0,1533*	*0,524*	*0,0062*	*0*	*2,0948*	*1,7074*
conducta	*0,1077*	*0,5113*	*0,0025*	*0*	*2,1356*	*1,7815*
causa	*0,1077*	*0,5113*	*0,0012*	*0*	*2,1356*	*1,7815*
procedimental	*-0,0123*	*0,5278*	*0,021*	*0*	*1,974*	*1,6416*
ciclo vital	*0*	*0*	*0*	*0*	*0*	*0*
otros	*0,4742*	*0,6848*	*0,1235*	*0,164*	*1,5027*	*1,2356*

Categoría semántica silla.

En la Tabla 13 se muestran los distintos coeficientes para la categoría semántica ***Silla***. Según el ratio de ganancia de información, los rasgos semánticos que se desechan son: *partes, funcional, lugar/hábitat, conducta, causa, procedimental y ciclo vital*. Los rasgos semánticos que mejor predice la EA son: *taxonómico, tipos, evaluativo* y *otros*.

Tabla 13.- Coeficientes para los rasgos semánticos de la categoría semántica silla.

Silla						
Sin agrupamiento de conglomerados de atributos						
Variables	Pearson	Euclídea	Ponderación de Atributos	Ganancia de información	Regresión	
					Sanos	EA
taxonómico	*0,539*	*0,6848*	*0,0456*	*0,255*	*1,5851*	*1,2258*
tipos	*0,2939*	*0,657*	*0,4323*	*0,122*	*1,5152*	*1,2502*
partes	*0,113*	*0,5848*	*0,1133*	*0*	*1,621*	*1,4729*
funcional	*0,2432*	*0,6016*	*0,1865*	*0*	*1,7496*	*1,3879*
evaluativo	*0,3509*	*0,6497*	*0,134*	*0*	*1,5536*	*1,2986*
lugar/hábitat	*0,1533*	*0,5243*	*0,0028*	*0*	*2,1005*	*1,7023*
conducta	*0,1077*	*0,5113*	*0,0028*	*0*	*2,1356*	*1,7815*
causa	*0*	*0*	*0*	*0*	*0*	*0*
procedimental	*-0,0515*	*0,5266*	*0,0138*	*0*	*1,9522*	*1,622*
ciclo vital	*0*	*0*	*0*	*0*	*0*	*0*
Otros	*0,4852*	*0,6457*	*0,0843*	*0,189*	*1,6678*	*1,32*

Categoría semántica pantalón.

En la Tabla 14 se muestran los distintos coeficientes para la categoría semántica ***Pantalón***. Según el ratio de la ganancia de información, los rasgos que mejor predice la EA son: *taxonómico, y tipos.*

Tabla 14.- Coeficientes para los rasgos semánticos de la categoría semántica pantalón.

Pantalón						
Sin agrupamiento de conglomerados de atributos						
Variables	Pearson	Euclídea	Ponderación de Atributos	Ganancia de información	Regresión	
					Sanos	EA
taxonómico	*0,3635*	*0,6296*	*0,0544*	*0,0913*	*1,6739*	*1,3412*
tipos	*0,4835*	*0,6679*	*0,4408*	*0,194*	*1,6495*	*1,2408*
partes	*0,1207*	*0,5999*	*0,1116*	*0*	*1,5351*	*1,4271*
funcional	*0,3232*	*0,6556*	*0,1796*	*0*	*1,5003*	*1,2836*
evaluativo	*0,2782*	*0,6205*	*0,1333*	*0*	*1,6032*	*1,3782*
lugar/hábitat	*0,2369*	*0,5592*	*0,0313*	*0*	*1,9177*	*1,5668*
conducta	*0*	*0*	*0*	*0*	*0*	*0*
causa	*0*	*0*	*0*	*0*	*0*	*0*
procedimental	*0,0967*	*0,5348*	*0,0136*	*0*	*1,9923*	*1,6482*
ciclo vital	*0,1077*	*0,5113*	*0,0014*	*0*	*2,1356*	*1,7815*
otros	*0,1752*	*0,5769*	*0,019*	*0*	*1,7473*	*1,5036*

Con estas medidas de asociación o correlación se pueden medir la magnitud con la que se relacionan las variables del corpus con la EA. Estos coeficientes se utilizan en las BN híbridas con el propósito de intentar mejorar el diagnóstico cuando la enfermedad es incipiente. Se podrá comprobar en los experimentos que cuando la enfermedad está avanzada, el daño es tan omnipresente que se pueden detectar los déficits léxico-semánticos-conceptuales en todos los rasgos de todas las categorías semánticas, sin mayor dificultad. Pero cuando el sujeto presenta un deterioro selectivo, es necesario ponderar la importancia de cada variable del corpus para un diagnóstico más preciso. El deterioro selectivo puede proporcionar información muy valiosa para la detección de la EA en sus primeros estadios.

En la parte III, capítulo 9 se detalla cómo se utilizan estos coeficientes en la CLG BN y la BN con inferencia aproximada.

Capítulo

Modelado con BNs Discretas

5

Las BNs constituyen una de las principales técnicas de minería de datos y descubrimiento del conocimiento. Nacen de la intersección entre la IA, la estadística y la teoría de la probabilidad, y forman parte de la familia de los modelos gráficos probabilísticos. Son capaces de predecir el valor de las variables no observadas en las condiciones de hipótesis. Se pueden construir BNs con variables discretas, con variables continuas o combinando ambos tipos de variables. En esta tesis doctoral se emplean: BNs discretas, las cuales se explican en este capítulo, y BNs híbridas, las cuales se detallan en el capítulo 6.

El objetivo de este capítulo es describir las técnicas de modelado y los algoritmos de aprendizaje automático de los parámetros utilizados en tres BNs discretas, para comprobar (en el capítulo 8) su eficacia. La primera BN emplea razonamiento deductivo; la segunda BN razonamiento abductivo; y la tercera BN es similar a la segunda, pero modela de forma explícita el deterioro semántico diferencial entre los dominios SV y SNV.

La organización del capítulo es la siguiente. En la sección 5.1 se realiza una introducción a las técnicas de modelado utilizadas en las BNs discretas. En la 5.2 se detallan las distintas estrategias de discretización. En las secciones 5.3, 5.4 y 5.5 se detallan los modelos cualitativos y cuantitativos de las BNs discretas.

5.1 Introducción.

Es muy difícil o incluso imposible construir un modelo de BN que cubra todos los aspectos de un problema tan complejo como el diagnóstico de la EA. Los modelos de BNs planteados son por tanto, una aproximación al dominio del problema, en el sentido de que sólo examinan, entre otras alteraciones cognitiva, el posible deterioro de la memoria semántica en sus aspectos declarativos (causado por la EA). Si los modelos se construyen bajo circunstancias no consistentes con las condiciones de contexto, el resultado en general será poco fiable. En esta tesis doctoral se ha realizado una batería de experimentos con objeto de encontrar el modelo de BN que mejor se adecue al problema. También se experimenta con distintos algoritmos de aprendizaje y discretización con el mismo objetivo. En la Parte III: Experimentos, se muestran los resultados de los experimentos.

En el capítulo 11 se realiza una batería de experimentos con otros algoritmos de minería de datos con los que se consiguen unos resultados aceptables, pero no superan a los resultados obtenidos con nuestra propuesta. El diseño del modelo cualitativo de las BNs desde el conocimiento del dominio, el diseño e implementación de algoritmos de aprendizaje automático de los modelos cuantitativos de las distintas BNs, la incorporación de conclusiones de investigaciones relevantes del campo de la psicología cognitiva y la incorporación de nuevos descubrimiento sobre la EA, han permitido conseguir un método de diagnóstico fiable, eficaz y económico.

Todos los modelos que presentamos utilizan las mismas variables, pero con diferencias significativas en las relaciones causales, pudiéndose resumir estas diferencias en:

- El primer modelo de BN infiere la probabilidad de padecer la EA una vez conocido los déficits léxico-semánticos-conceptuales, es decir, se deduce la probabilidad de padecer la EA a partir del deterioro de la memoria semántica en sus aspectos declarativos. Para esta BN se crea el enlace causal ***EA*** ← ***DSD*** (razonamiento deductivo), siendo la TPC de la variable ***EA*** más compleja (144 parámetros) que la TPC de las otras BNs que se proponen (72 parámetros). Por esta razón se ha utilizado, en algunos experimentos, el estudio epidemiológico [18] y el simplificador *Naive Bayes,* para aprender estos parámetros.
- En el segundo modelo de BN es similar al primero, excepto por el enlace causal ***EA*** → ***DSD***, es decir, se busca la causa que mejor explica la aparición del deterioro de la memoria semántica (razonamiento abductivo). Este cambio en los enlaces causales implica un menor número de parámetros en la TPC de la variable ***EA***, es decir, este modelo simplifica el aprendizaje automático de los parámetros de la variable ***EA.***
- El tercer modelo es similar al segundo, pero modela explícitamente el deterioro semántico diferencial entre los dominios semánticos SV y SNV, es decir, se establecen los enlaces causales ***EA*** → $\boldsymbol{DSD_{SV}}$ y ***EA*** → $\boldsymbol{DSD_{SNV}}$. El objetivo de esta BN es contrastar los resultados con el modelo 2 de BN para analizar la eficacia de estos enlaces causales.

Existe otra alternativa de modelado, la cual se trata en el capítulos 6 (Modelado con una BN híbrida) y en la que utiliza inferencia por razonamiento deductivo entre las variables que representan los rasgos semánticos (variables predictoras) y las categorías semánticas (variables criterios); para ello se crea una red de ecuaciones lineales estructurales. Con las BNs discretas no se ha considerado esta estructura porque las TPCs de las variables que representan las categorías semánticas tendrían muchos parámetros y sería mucho más complejo aprender el modelo cuantitativo, por consiguiente se necesitaría un corpus de datos más amplio. Además, en el capítulo 11 se experimenta con una BN *Naive Bayes* y con otros algoritmos de aprendizaje automático.

Para cada BN se ha desarrollado un algoritmo específico de aprendizaje automático del modelo cuantitativo. A lo largo del capítulo se irán detallando las técnicas de modelado utilizadas en las distintas BNs, así como la formulación utilizada en los distintos algoritmos de aprendizaje automático del modelo cuantitativo.

5.2 Discretización de atributos numéricos por análisis de clúster. Algoritmo k-Means++.

Las BNs discretas utilizan exclusivamente variables discretas, sin embargo, los test orales con restricción temporal producen atributos numéricos, por tanto es necesario transformar estos atributos numéricos en atributos cualitativos. Estas estrategias de discretización son aplicables a todos los modelos de BN propuestos.

Del recuento de la producción oral de rasgos semánticos analizados e interpretados en el corpus lingüístico [1], se buscan dos centroides con el algoritmo de clustering *k-Means++*. Cada centroide define la ausencia o presencia del posible déficit léxico-semántico-conceptual.

Se ha seleccionado análisis de clúster, tal y como se indicó en la sección 3.2, por su sencillez, por su velocidad, por su eficacia en este problema y porque es una técnica que se ha utilizado en otras investigaciones del campo de la psicología [44,45,46]; aunque se hubiese sido factible utilizar otro algoritmo de discretización.

Como se ha indicado anteriormente, se utiliza una estructura jerárquica de conglomerados de atributos (recuento de la producción oral de rasgos semánticos) para crear una influencia informativa entre la edad y/o el nivel educativo, y las variables del corpus lingüístico. Es muy importante tener en cuenta estos factores de contexto a la hora de propagar las evidencias por la BN e inferir un diagnóstico, ya que es imprescindible determinar si los déficits en la producción oral de rasgos semánticos se pueden deber a otros factores que nada tienen que ver con la EA, como por ejemplo la propia vejez cerebral o un escaso nivel educativo.

La estrategia utilizada en la discretización de los atributos numéricos tiene un impacto importante en el proceso de aprendizaje automático del modelo cuantitativo, es decir, los parámetros de la BN varían significativamente en función de la estrategia de discretización utilizada. Para la búsqueda de los centroides, el algoritmo *k-Means++* utiliza la distancia euclídea utilizando la siguiente fórmula:

$$d(x, y) = \sqrt{(x - y)^2} \quad (5.1)$$

donde

x: es el centroide del clúster.

y: es el recuento de la producción oral de rasgos semánticos para una variable determinada.

La ausencia o presencia de los déficits léxico-semánticos-conceptuales constituyen las evidencias para la BN. Una vez discretizados los atributos numéricos, el algoritmo de aprendizaje automático del modelo cuantitativo establece el grado de asociación, correlación o fuerza de las relaciones, entre los posibles déficits léxico-semánticos-conceptuales y el deterioro de la memoria semántica causado por la EA. Por ejemplo, no

tiene la misma importancia la variable *taxonómico* de la categoría natural ***perro*** que la variable *tipo* de la misma categoría semántica, en el diagnóstico de la EA.

En esta tesis se han considerado tres estrategias de discretización. La primera estrategia, crea una jerarquía de conglomerados de atributos por categoría semántica y rasgo semántico. La segunda estrategia, crea una jerarquía de conglomerados de atributos por edad, categorías semánticas y rasgos semánticos. La tercera estrategia, crea una jerarquía de conglomerados de atributos por nivel educativo, categorías semánticas y rasgos semánticos. Se pudo comprobar en el Capítulo 4 (Análisis Estadístico) que existe una tendencia a producir menos rasgos semánticos las personas de mayor edad y las personas con un menor nivel educativo.

5.2.1 Discretización por objetos semánticos y rasgos.

En esta sección se discretizan las variables continuas creando para ello una jerarquía de conglomerado de atributos por cada categoría semántica y bloque conceptual, por ejemplo para la categoría semántica ***manzana*** se generan 22 clústeres, dos clúster por cada bloque conceptual. En la Figura 8 se representa un fragmento de la estructura jerárquica de los conglomerados de atributos. Los círculos dobles representan los centroides de cada clúster. Los círculos simples representan las categorías semánticas y rasgos semánticos a los que se les aplica la técnica de clustering.

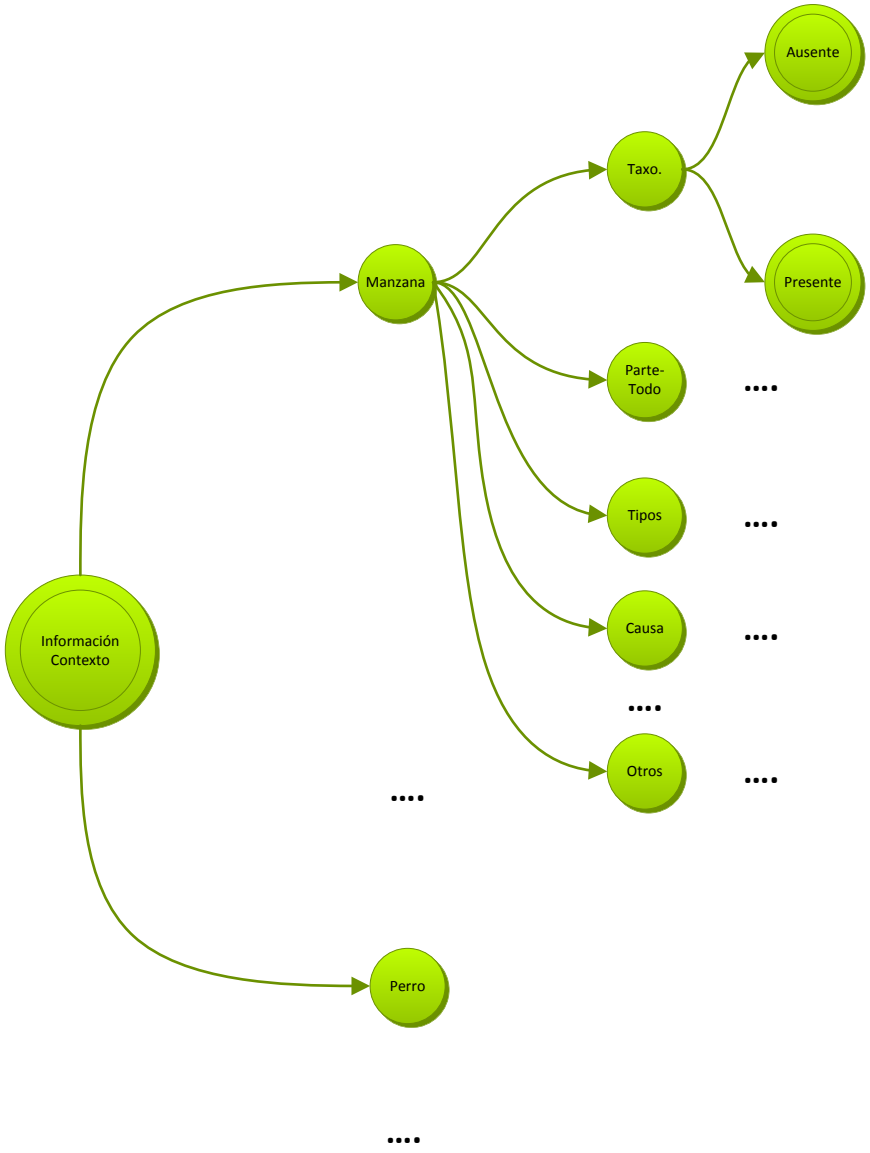


Figura 8.- Clústeres por categoría semántica y rasgo semántico.

El proceso de discretización se basa en la búsqueda del centroide más cercano a la del recuento de atributos del bloque semántico que se pretende discretizar. Una vez hallado el centroide más cercano, se le asigna una etiqueta identificativa que representa la clase a la que pertenece dicho centroide (déficits *ausente* o *presente*).

En la Tabla 15 se muestran los centroides encontrados para las variables de interés y las variables intermedias que representan las categorías semánticas. Cada centroide define la presencia o ausencia del déficit léxico-semántico-conceptual de categorías semánticas específicas. Por ejemplo, los participantes que produzcan aproximadamente 16 rasgos semánticos o atributos lingüísticos en el dominio semántico SV, *k-Means++* considerará que el participante presenta un déficit léxico-semántico-conceptual para este dominio semántico.

Tabla 15.- Resumen de centroides hallados por análisis de clúster de los DLSC (déficits léxico-semánticos-conceptuales) para la primera estrategia de discretización (por categoría semántica y rasgo semántico).

	Centroide producción lingüística normal	Centroide producción lingüística enfermos de EA
DLSC	89.0	29.0
$DLSC_{SV}$	45.0	16.0
$DLSC_{SNV}$	42.0	11.0
$DLSC_{Manzana}$	17.0	5.0
$DLSC_{Perro}$	22.0	6.0
$DLSC_{Pino}$	19.0	6.0
$DLSC_{Coche}$	18.0	5.0
$DLSC_{Silla}$	17.0	5.0
$DLSC_{Pantalón}$	16.0	5.0

La presencia del déficit léxico-semántico-conceptual señala la posible presencia del deterioro semántico. El algoritmo suma, por cada categoría semántica o dominio semántico, la producción oral de rasgos semánticos o atributos lingüísticos. Posteriormente se busca por cada variable dos centroides que definen la presencia o ausencia del posible déficits léxico-semánticos-conceptuales. Estos centroides se buscan con análisis de clúster, a partir de los conglomerados de atributos calculados con las formulas (5.2):

$$POA = \sum POA_{SV} + \sum POA_{SNV}$$

$$POA_{SV} = \sum_{\substack{Rasgos \\ Semánticos}} POA_{Manzana} + \sum_{\substack{Rasgos \\ Semánticos}} POA_{Perro} + \sum_{\substack{Bloques \\ Conceptuales}} POA_{Pino}$$

$$POA_{SNV} = \sum_{\substack{Rasgos \\ Semánticos}} POA_{Coche} + \sum_{\substack{Rasgos \\ Semánticos}} POA_{Silla} + \sum_{\substack{Rasgos \\ Semánticos}} POA_{Pantalón}$$

$$POA_{categoria\ semántica} = \sum POA_{taxonómico} + POA_{tipos} + POA_{partes} + \cdots \qquad (5.2)$$

donde

POA: Recuento de la producción oral de rasgos semánticos de todas las categorías semánticas.

POA_{SV}: Recuento de la producción oral de rasgos semánticos para las categorías semánticas relativas a categorías naturales.

POA_{SNV}: Recuento de la producción oral de rasgos semánticos para las categorías semánticas relativas a artefactos.

$POA_{categoría\ semántica}$: Recuento de la producción oral de rasgos semánticos para las categorías semánticas, ∈ {***manzana, perro, pino, silla, pantalón, coche***}.

POA_{rasgos}: Recuento de la producción oral de rasgos semánticos, ∈ {*taxonómico, tipos, parte-todo, funcional, evaluativo, lugar/hábitat, comportamiento, causa/genera, procedimental, ciclo vital, otros*}.

5.2.2 Discretización por edad, categorías semánticas y rasgos semánticos.

Este método de discretización utiliza análisis de clúster segmentando los conglomerados de atributos por edad (recuento de la producción oral de rasgos semánticos), categorías semánticas y rasgos semánticos. El procedimiento es similar al descrito en el apartado anterior, salvo que además se crea un conglomerado de atributos, por cada tramo de edad.

En la Figura 9 se representan los segmentos de conglomerados de atributos por cada tramo de edad.

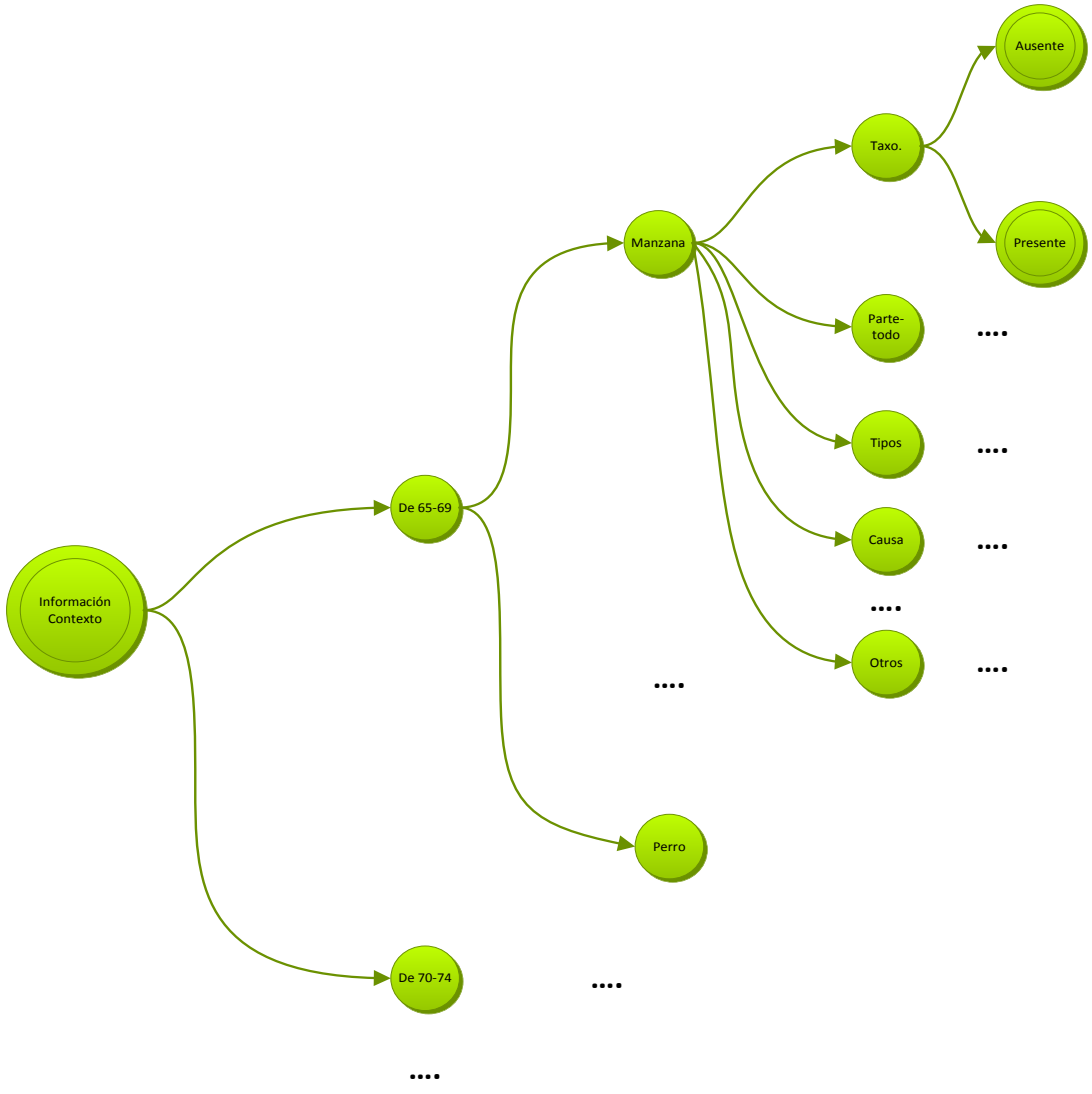


Figura 9.- Clustering por edad, categoría semántica y rasgo semántico.

En la Tabla 16 se muestran los centroides para las variables de interés y categorías semánticas por cada tramo de edad. Se puede comprobar como los centroides tienen a disminuir en función de la edad.

Tabla 16.- Resumen de centroides hallados por análisis de clúster de los DLSC para la segunda estrategia de discretización (por edad, categoría semántica y rasgo semántico).

	Edad	Centroide producción lingüística normal	Centroide producción lingüística enfermos de EA
DLSC	0-64	100.0	39.0
	65-69	100.0	29.0
	70-74	123.0	43.0
	75-79	87.0	43.0
	80-84	52.0	12.0
	85-	83.0	34.0
$DLSC_{SV}$	0-64	56.0	20.0
	65-69	56.0	20.0
	70-74	68.0	20.0
	75-79	54.0	33.0
	80-84	27.0	6.0
	85-	44.0	21.0
$DLSC_{SNV}$	0-64	39.0	13.0
	65-69	47.0	11.0
	70-74	43.0	12.0
	75-79	42.0	15.0
	80-84	41.0	9.0
	85-	38.0	13.0
$DLSC_{Manzana}$	0-64	18.0	5.0
	65-69	19.0	6.0
	70-74	20.0	6.0
	75-79	17.0	7.0
	80-84	10.0	2.0
	85-	5.0	0.0
$DLSC_{Perro}$	0-64	25.0	7.0
	65-69	22.0	7.0
	70-74	16.0	4.0
	75-79	20.0	10.0
	80-84	9.0	3.0
	85-	22.0	13.0
$DLSC_{Pino}$	0-64	12.0	4.0
	65-69	15.0	3.0
	70-74	22.0	5.0
	75-79	17.0	8.0
	80-84	9.0	2.0
	85-	21.0	9.0
$DLSC_{Coche}$	0-64	19.0	5.0
	65-69	15.0	3.0
	70-74	13.0	2.0
	75-79	15.0	9.0
	80-84	13.0	3.0
	85-	13.0	5.0

	Edad	Centroide producción lingüística normal	Centroide producción lingüística enfermos de EA
$DLSC_{Silla}$	0-64	10.0	2.0
	65-69	17.0	4.0
	70-74	18.0	5.0
	75-79	18.0	9.0
	80-84	22.0	3.0
	85-	16.0	4.0
$DLSC_{Pantalón}$	0-64	13.0	3.0
	65-69	16.0	4.0
	70-74	13.0	3.0
	75-79	17.0	9.0
	80-84	8.0	0.0
	85-	12.0	4.0

Al igual que en el método anterior, las variables intermedias se discretizan desde el recuento de la producción oral de rasgos semánticos, siguiendo las fórmulas (5.2).

Merece la pena analizar con k-Means++ que variables del corpus predicen mejor la EA y para ello, en la Tabla 17 se muestran las métricas de rendimiento. Con k-Means++ se infiere individualmente para cada dominio semántico y cada categoría semántica, la posibilidad de padecer un deterioro de la memoria semántica causada por la EA. Se puede comprobar que con la categoría semántica ***Manzana*** se consiguen mejores métricas de rendimiento que con la categoría semántica ***Perro,*** es decir, si sólo se les preguntará a los participantes que dijesen todo lo que sepan sobre una manzana, el método de diagnóstico propuesto en esta tesis sería más eficaz que si sólo se les preguntará por los perros.

Tabla 17.- Comparativa de la importancia predictiva de los dominios semánticos y categorías semánticas con análisis de clúster.

VARIABLES	FP RATE	TP RATE	PRECISIÓN	EXACTITUD
DS_{SV}	0,214	0,923	0,800	0,852
DS_{SNV}	0,119	0,846	0,868	0,864
$DS_{Manzana}$	0,262	0,923	0,766	0,827
DS_{Perro}	0,429	0,769	0,625	0,667
DS_{Pino}	0,333	0,949	0,725	0,802
DS_{Coche}	0,405	0,897	0,673	0,741
DS_{Silla}	0,214	0,872	0,791	0,827
$DS_{Pantalón}$	0,381	0,821	0,667	0,716

5.2.3 Discretización por nivel educativo, objetos semánticos y rasgos.

Este método de discretización utiliza análisis de clúster, segmentando los conglomerados de atributos por nivel educativo, categorías semánticas y rasgos semánticos. El procedimiento es similar al descrito en el apartado anterior, excepto que se crea un conglomerado de atributos lingüístico (recuento de la producción oral de rasgos semánticos) por cada nivel educativo.

En la Figura 10 se representan la estructura jerárquica de conglomerados de atributos por cada nivel educativo.

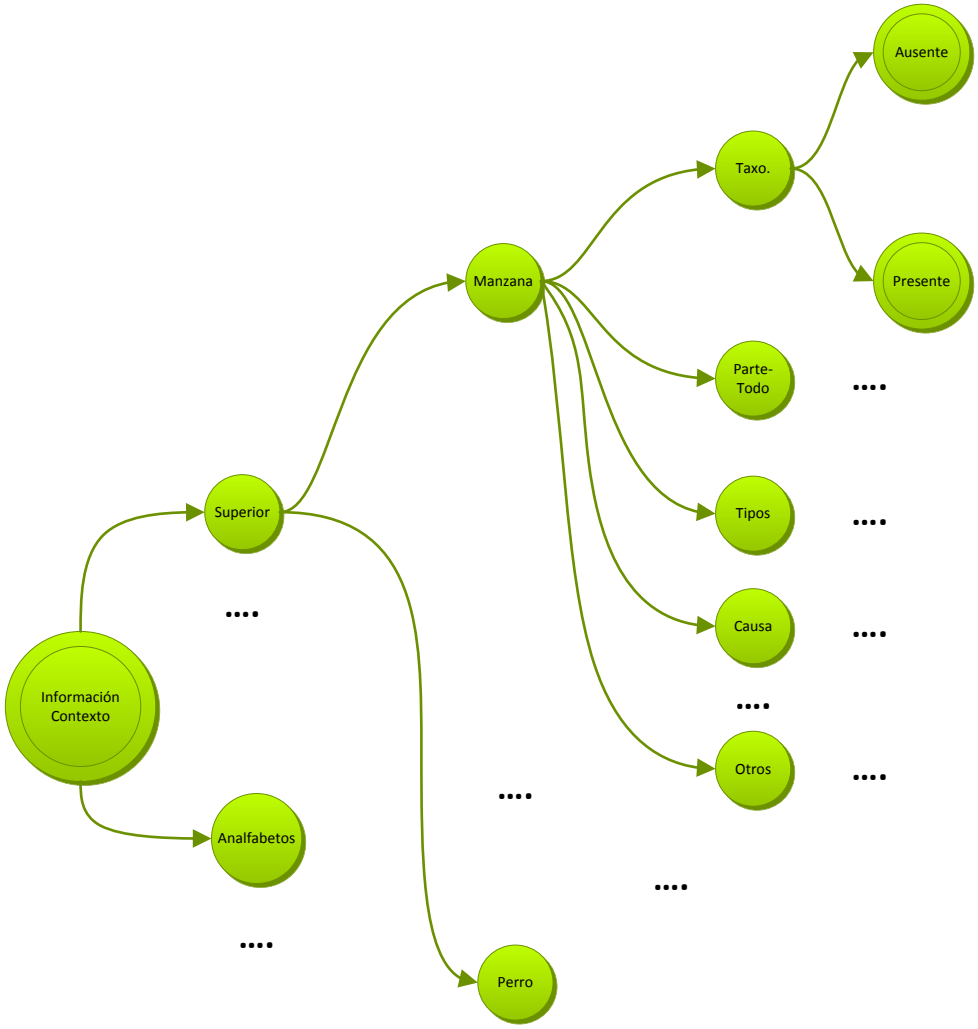


Figura 10.- Clustering por nivel educativo, categoría semántica y rasgo semántico.

En la Tabla 18 se muestran los centroides encontrados para las variables de interés y variables intermedias, segmentando los conglomerados de atributos por nivel educativo. En esta tabla se puede comprobar que las personas con más años de escolarización tienden a producir más rasgos semánticos.

Tabla 18.- Resumen de centroides hallados por análisis de clúster de los DLSC para la segunda estrategia de discretización (por nivel educativo, categoría semántica y rasgo semántico).

	Nivel Educativo	Centroide producción lingüística normal	Centroide producción lingüística enfermos de EA
DLSC	Primarios	76.0	28.0
	Secundarios	97.0	10.0
	Universitarios	111.0	57.0
$DLSC_{Coche}$	Primarios	14.0	4.0
	Secundarios	18.0	1.0
	Universitarios	16.0	9.0
$DLSC_{Manzana}$	Primarios	14.0	5.0
	Secundarios	16.0	4.0
	Universitarios	22.0	8.0
$DLSC_{Pantalón}$	Primarios	15.0	5.0
	Secundarios	19.0	6.0
	Universitarios	14.0	6.0
$DLSC_{Perro}$	Primarios	21.0	7.0
	Secundarios	23.0	6.0
	Universitarios	29.0	11.0
$DLSC_{Pino}$	Primarios	13.0	3.0
	Secundarios	17.0	4.0
	Universitarios	29.0	11.0
$DLSC_{Silla}$	Primarios	17.0	5.0
	Secundarios	14.0	3.0
	Universitarios	18.0	6.0
$DLSC_{SNV}$	Primarios	40.0	12.0
	Secundarios	46.0	3.0
	Universitarios	46.0	23.0
$DLSC_{SV}$	Primarios	37.0	15.0
	Secundarios	48.0	6.0
	Universitarios	62.0	27.0

En esta sección se han descrito en profundidad las distintas estrategias para la discretización de los atributos numéricos. Estas estrategias permiten de alguna manera, tener en cuenta la correlación existente entre la edad o nivel educativo, y la producción oral de rasgos semánticos.

5.3 Modelo 1: Inferencia por razonamiento deductivo.

Describimos en este apartado el primer modelo que presentamos de BN, destaquemos una vez más que todas las BNs utilizan las mismas variables. Los factores de riesgo no son causa de la enfermedad, por tanto implica que no se pueden utilizar modelos canónicos (OR/MAX) en el modelado de las BNs para reducir el número de parámetros.

5.3.1 Modelo cualitativo.

En el modelo de BN de la Figura 11, la probabilidad de padecer la EA se infiere a partir de sus síntomas, además, se combina el corpus lingüístico [1] con un estudio epidemiológico extraído de la literatura científica [18], en el aprendizaje automático de los parámetros.

En la Figura 11 se representa una estructura de BN que utiliza razonamiento deductivo en el proceso de inferencia. Aunque las variables que representan los rasgos semánticos se han nombrado de la misma forma en todas las categorías semánticas, son variables independientes.

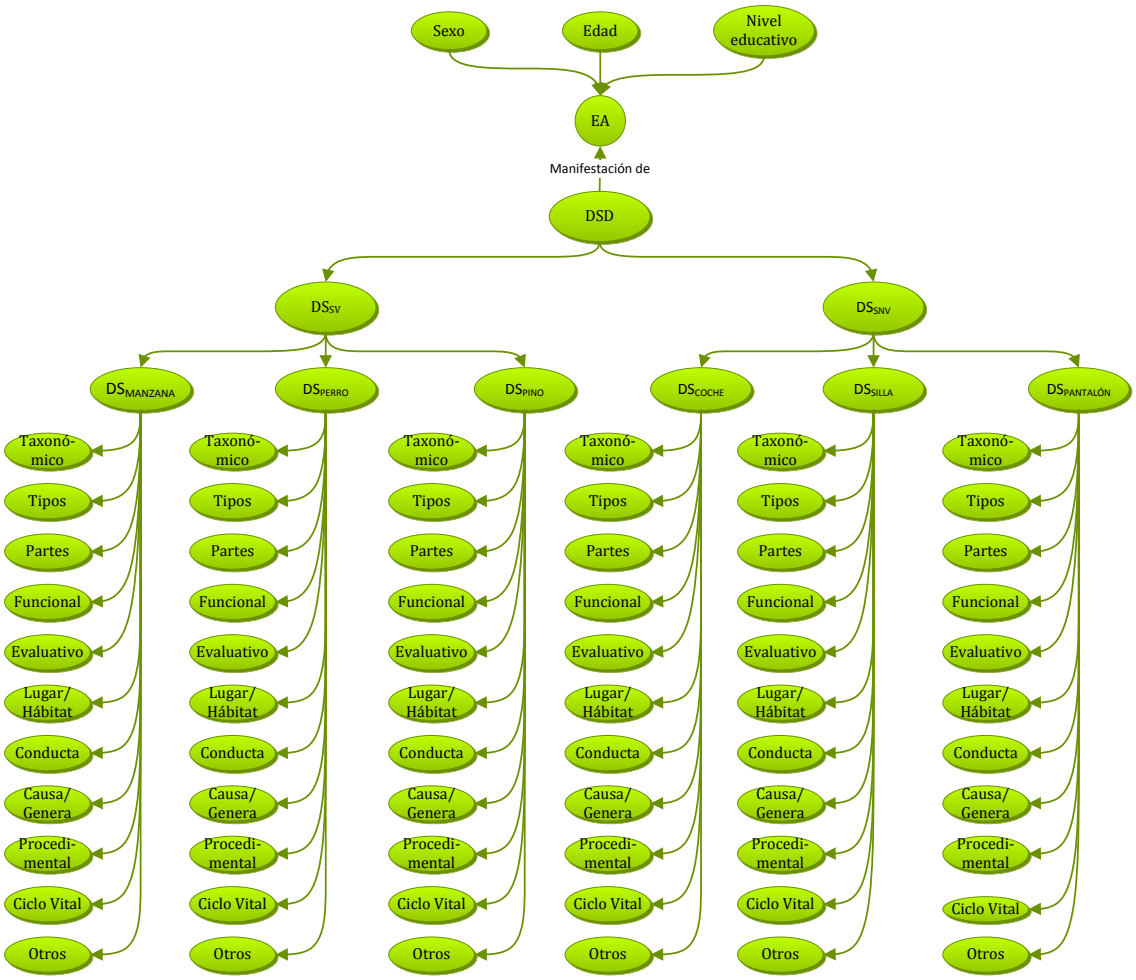


Figura 11.- Modelo 1 de BN. Inferencia para un razonamiento deductivo de la variable EA.

En la BN de la Figura 11 se establecen enlaces causales entre las variables de información de contexto o factores de riesgo, y la EA. Tal y como se indicó en el capítulo 1, se ha podido evidenciar (así lo demuestran los estudios epidemiológicos) que la edad, el sexo y el nivel educativo tienen algún tipo de correlación con la EA.

Por otro lado, en la BN de la Figura 11 se modela el enlace causal entre la *EA* y *DSD* –el deterioro de la memoria semántica es un aspecto DCL que es una entidad clínica previa a la demencia—, ya que tal y como se indicó en el capítulo 1, la causa más común de demencia en la Unión Europea es la EA (alrededor del 50-70% de los casos),

otra de la causa es la demencia multiinfarto (alrededor del 30% de los casos de demencia), la enfermedad de *Pick*, demencia de cuerpos de *Lewy* y otros [14]. En nuestras BNs sólo hemos tenido en cuenta la EA, aunque es factible aplicar nuestro método de diagnóstico a otras ENs no-EA. En la Figura 12 se representa un posible fragmento de un modelo causal de una BN alternativa, cuyo fin es predecir otras ENs.

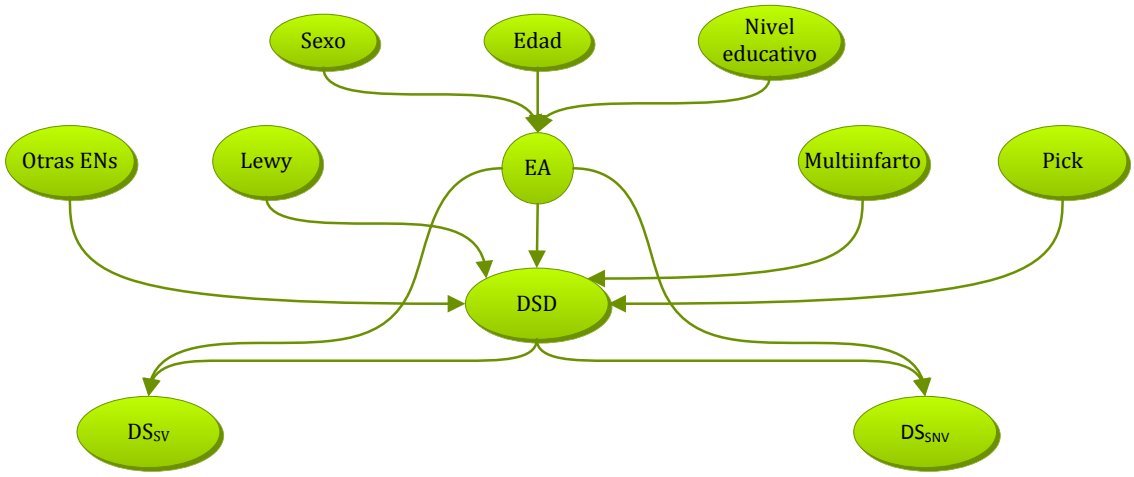


Figura 12.- Ejemplo de BN que predice otras ENs.

Con esta la relación causal ***EA*** ← ***DSD*** nos permite plantear la siguiente pregunta: dado un DSD ¿cuál es la probabilidad de que sea causado por la EA? Es decir, podemos conocer la probabilidad de que el paciente sufra la EA cuando presenta un déficit léxico-semántico-conceptual. Asimismo, los enlaces causales $\boldsymbol{DS_{SV} \rightarrow DSD}$ y $\boldsymbol{DS_{SNV} \rightarrow DSD}$ permiten conocer la probabilidad de que exista un déficit léxico-semántico-conceptual en el dominio semántico SV y/o en el dominio semántico SNV, y a partir del déficit en la producción oral de rasgos semánticos, se infiere la probabilidad de presentar un deterioro semántico y en consecuencia, la probabilidad de presentar el DCL.

Otra cuestión importante del modelo 1 de BN es el uso de las variables intermedias (***DSD***, $\boldsymbol{DS_{SV}}$, $\boldsymbol{DS_{SNV}}$, $\boldsymbol{DS_{manzana}}$, $\boldsymbol{DS_{perro}}$, $\boldsymbol{DS_{pino}}$, $\boldsymbol{DS_{coche}}$, $\boldsymbol{DS_{silla}}$ y $\boldsymbol{DS_{pantalón}}$) en el modelo causal. Por un lado, las variables $\boldsymbol{DS_{SV}}$ y $\boldsymbol{DS_{SNV}}$ tienen cierta importancia porque sirven para estudiar el deterioro semántico diferencial que presentan algunos enfermos de EA cuando la enfermedad es incipiente o de intensidad leve. Con la fuerza de las relaciones entre estas variables expresamos esta situación y además, nos puede ayudar a discriminar la EA de otras demencias de tipo no-EA. Por otro lado, también se deberían modelar en las BNS otras situaciones, como por ejemplo, las correlaciones positivas entre las variables de una misma categoría semántica y/o dominio semántico, es decir, si aumenta la producción oral de atributos en una categoría semántica determinada, aumenta la probabilidad de que el sujeto realice una mayor producción oral de rasgos semánticos en otras categorías semánticas del mismo dominio. Pero también podía existir una correlación negativa entre las categorías semánticas de distinto dominio semántico debido al deterioro cognitivo focalizado. Para modelar estas situaciones en las BNs son necesarios los ciclos, tal y como se representa en la Figura 13; y las BNs no

admiten ciclos. Esta es otra razón más, por las que se utilizan variables intermedias en el modelo de BN propuesto, tal y como se muestra en la Figura 13.

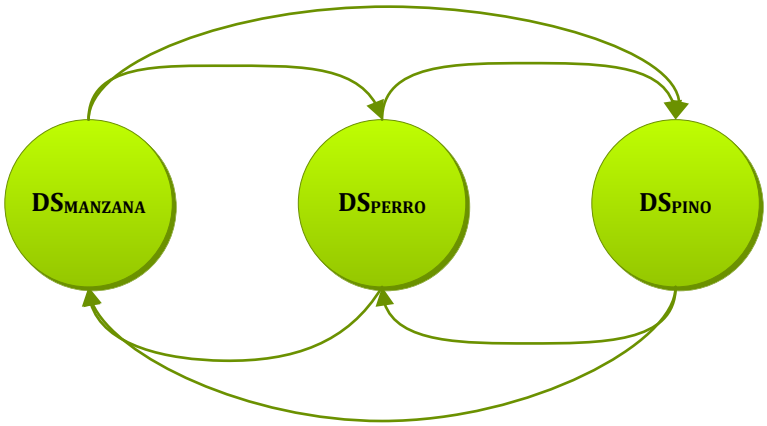


Figura 13.- BN con ciclos.

La Figura 14 representa la técnica de modelado utilizada para evitar los ciclos en la BN.

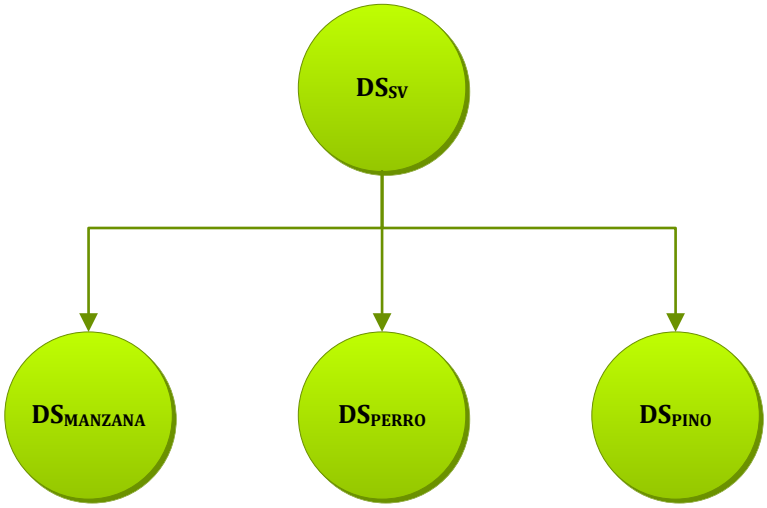


Figura 14.- BN sin ciclos.

Una vez analizado el modelo cualitativo de la BN con razonamiento deductivo, en la siguiente sección se describe en detalle el algoritmo de aprendizaje automático del modelo cuantitativo.

5.3.2 Modelo cuantitativo.

En esta sección se detallan las técnicas que se han utilizado para la construcción del modelo cuantitativo de cada una de las BNs.

En el modelo causal existen una serie de variables consideradas como factores de riesgo y factores de protección. Estas variables son deterministas y son: *sexo, edad* y *nivel educativo*. Existen artículos y estudios epidemiológicos [14] en los que se analizan otros factores de protección, como por ejemplo el nivel de ocupación. En [14] se realiza una

comparativa de la prevalencia de la EA según el nivel educativo y el nivel de ocupación, que es interesante considerar en futuras investigaciones.

Para desarrollar el modelo cuantitativo de la variable ***EA*** se ha recurrido al estudio epidemiológico [18]. Las principales razones por las que se utiliza este estudio [18] son:

- El modelo cuantitativo de la BN requiere que la muestra este formado por individuos elegidos al azar, para que sean representativos de la población.
- Todas las muestras deben tener un carácter aleatorio y es recomendable que no exista sesgo. Una de las motivaciones de la aparición del sesgo es porque la muestra no tenga carácter aleatorio. Esta clase de errores se conocen como *errores sistemáticos* y no permiten obtener conclusiones acerca del valor verdadero.

Se ha recurrido a este estudio [18] porque el corpus que usamos [1] se ha elaborado con un muestreo incidental dónde el investigador seleccionó deliberadamente un conjunto de casos proporcionados y diagnosticados por neurólogos. No se tuvo en cuenta la aleatoriedad de la muestra por la dificultad en la obtención de una muestra más amplia.

En la Tabla 19 se detalla el estudio epidemiológico sobre las prevalencia de la demencia y la prevalencia de la EA [18].

Tabla 19.- Estudio epidemiológico [18]. Factores de riesgo y protección.

	Prevalencia (%) de demencia por sexo, edad y nivel educativo.		Prevalencia (%) de EA por sexo, edad y nivel educativo.	
	Prevalencia	***Intervalo de Confianza***	***Prevalencia***	***Intervalo de Confianza***
Prevalencia Total	9,1	7,87-10,49	6,9	5,84-8,12
Sexo				
✓ **Mujeres**	11,8	9,93-13,87	9,4	7,70-11,28
✓ **Hombres**	5,6	4,17-7,38	3,7	2,58-5,25
Grupos de edad (años)				
✓ **65-69**	2	0,9-3	1,4	0,55-2,8
✓ **70-74**	5,2	3,57-7,4	3,3	2,01-5,14
✓ **75-79**	7,8	5,39-10,95	6,1	3,93-8,91
✓ **80-84**	13,5	9,57-18,41	9,2	5,9-13,43
✓ **≥ 85**	34,7	28,11-41,65	29,7	23,49-36,52
Estudios				
✓ **Primarios**	14,6	11,67-17,91	10,9	8,39-13,94
✓ **Secundarios**	7,3	5,91-8,82	5,7	4,52-7,13
✓ **Universitarios**	3,4	0,94-8,52	1,7	0,21-6,04

Para el cálculo de la TPC de la variable ***EA*** no es suficiente con el estudio epidemiológico, sino que es necesario diseñar e implementar un algoritmo de aprendizaje automático específico para el diagnóstico que pretendemos. En concreto, se necesitan todas las probabilidades condicionales del tipo (5.3), esto es, más de 70 valores (distintas combinaciones de los estados de cada variable):

$$P(EA \mid SEXO, EDAD, NIVEL\ EDUCATIVO) \qquad (5.3)$$

Del estudio epidemiológico se puede extraer las siguientes distribuciones de probabilidades conjuntas:

- $P(EA_{presente}, sexo)$
- $P(EA_{presente}, edad)$
- $P(EA_{presente}, nivel\ educativo)$

Es posible utilizar otros estudios epidemiológicos que tengan en cuenta las variables y estados que necesita el método de diagnóstico propuesto en esta tesis, pero hemos seleccionado [18]

Para calcular la TPC de la variable ***EA*** utilizamos el simplificador Bayesiano Ingenuo o *Naive Bayes*. Téngase en cuenta que para conocer todas las probabilidades a priori de los factores de riesgo y factores de protección en relación a la variable ***EA***, se necesita una gran cantidad de probabilidades a priori y es demasiado costoso elaborar un estudio epidemiológico adaptado a los requisitos de este método de diagnóstico.

El Método Bayesiano Ingenuo o *Naive Bayes* parte de la hipótesis que el diagnóstico es exclusivo (no puede haber dos de ellos a la vez) y exhaustivo (no hay otros diagnósticos posibles). En este modelo de BN se cumplen estas dos condiciones, ya que sólo pretendemos diagnosticar un DC compatible con la EA, y por tanto, podemos calcular esta TPC con la fórmula (5.4). Téngase en cuenta que cuando se menciona la exclusividad y exhaustividad en el diagnóstico, se refiere al únicamente al diagnóstico inferido por las BNs, es decir, nuestras BNs sólo diagnóstica el DSD causado por la EA, si pretendiéramos diagnosticar otras ENs con la misma BN, no sería posible utilizar el simplificar *Naive Bayes* para el cálculo de esta TPC.

$$P(EA|E,S,N,DCL) = \alpha * P(EA) * P(E \mid EA) * P(S \mid EA) * P(N \mid EA) * P(DSD \mid EA) \quad (5.4)$$

donde

α: Factor correctivo.

E: $Edad, \in \{0\ a\ 64, 65\ a\ 69, 70\ a\ 74, 75\ a\ 79, 80\ a\ 84, más\ de\ 85\}$

S: $Sexo, \in \{hombre, mujer\}$

N: $Nivel\ educativo, \in \{primarios, secundarios, universitarios\}$

DSD: $Deterioro\ semántico\ en\ sus\ aspectos\ declarativos, \in \{ausente, presente\}$

$$P(E \mid EA) = \frac{P(E, EA)}{P(EA)}$$

$$P(S \mid EA) = \frac{P(S, EA)}{P(EA)}$$

$$P(N \mid EA) = \frac{P(N, EA)}{P(EA)}$$

$$P(DSD \mid EA) = \frac{P(DSD, EA)}{P(EA)}$$

Para calcular la TPC de la variable ***DSD*** se emplea la siguiente fórmula:

$$P(DSD|DS_{SV}, DS_{SNV}) = \frac{N(DSD, DS_{SV}, DS_{SNV})}{N(DS_{SV}, DS_{SNV})} \quad (5.5)$$

donde

DSD: Deterioro semántico en sus aspectos declarativos, ∈{ausente, presente}

DS_{SV}: Deterioro semántico del dominio SV, ∈{ausente, presente}

DS_{SNV}: Deterioro semántico del dominio SNV, ∈{ausente, presente}

N(X): Recuento de casos que satisface las condiciones *X*.

Las TPCs para las variables DS_{SV} y DS_{SNV} se calculan con la fórmula (5.6):

$$P(x) = \frac{N(x)}{N} \qquad (5.6)$$

donde

x: déficit léxico-semánticos-conceptuales, $\in \{DS_{SV}, DS_{SNV}\}$

Se puede observar en la Figura 11 que existen muchas relaciones causales entre las variables intermedias que representan cada una de las categorías semánticas y las variables que representan los rasgos semánticos. Todas estas variables se han calculado según la formulación (5.7).

$$P(DS_{RS,CAT}|DS_{CAT}) = \frac{N(DS_{RS,CAT}, DS_{CAT})}{N(DS_{CAT})} \qquad (5.7)$$

donde

CAT: categorías semánticas, ∈ {*manzana, perro, pino, coche, silla pantalón*}

RS: rasgos semánticos, ∈ {*taxonómico, tipos, parte-todo, funcional, evaluativo, lugar/hábitat, comportamiento, causa/genera, procedimental, ciclo vital, otros*}

$DS_{RS,\,CAT}$: $Deterioro\ semántico\ de\ CAT\ y\ RS$, ∈ {*ausente, presente* }

5.4 Modelo 2: Inferencia por razonamiento abductivo.

La BN de la Figura 15 es similar a la anterior excepto por el enlace causal ***EA→DSD***. Con esta variación del modelo causal se simplifica la TPC de la variable ***EA*** y por tanto, su aprendizaje. Como se ha indicado anteriormente, con esta BN pretendemos analizar la diferencia en el rendimiento que se obtiene al simplificar el aprendizaje de los parámetros de la TPC de la variable ***EA.***

5.4.1 Modelo cualitativo.

El modelo 2 de BN es similar al anterior, salvo por la ausencia de los enlaces causales ***EA→*** $\boldsymbol{DS_{SV}}$ *y* ***EA→*** $\boldsymbol{DS_{SNV}}$

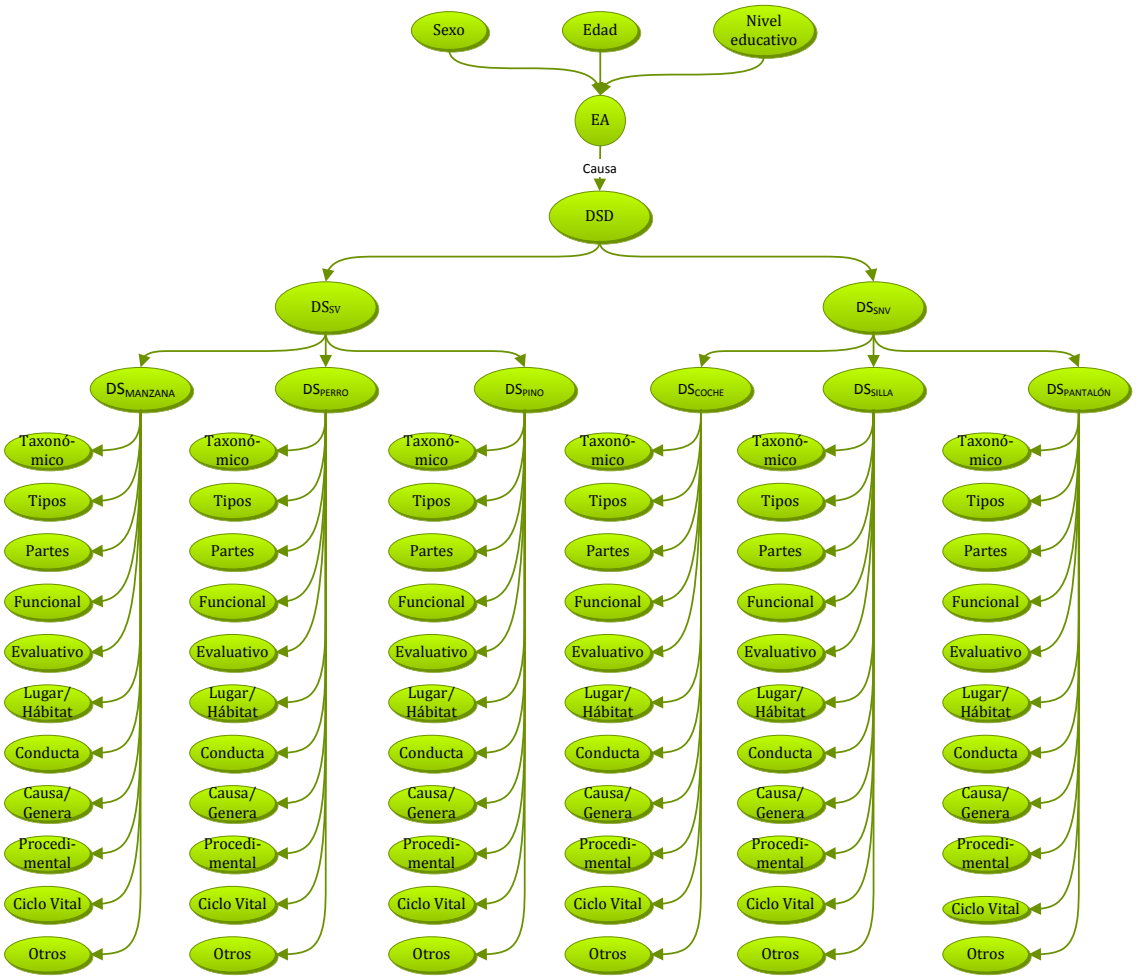


Figura 15.- Modelo 2 de BN. Inferencia por razonamiento abductivo.

5.4.2 Modelo cuantitativo.

En esta BN para el cálculo de la TPC de la variable ***EA*** se utiliza *Naive Bayes* con la siguiente fórmula:

$$P(EA|E,S,N) = \alpha * P(EA) * P(E \mid EA) * P(S \mid EA) * P(N \mid EA) \quad (5.8)$$

donde

α: Factor correctivo.

E: $Edad, \in \{0\ a\ 64, 65\ a\ 69, 70\ a\ 74, 75\ a\ 79, 80\ a\ 84, más\ de\ 85\}$

S: $Sexo \in \{hombre, mujer\}$

N: $Nivel\ educativo \in \{primarios, secundarios, universitarios\}$

$$P(E \mid EA) = \frac{P(E, EA)}{P(EA)}$$

$$P(S \mid EA) = \frac{P(S, EA)}{P(EA)}$$

$$P(N \mid EA) = \frac{P(N, EA)}{P(EA)}$$

En este modelo, la TPC de la variable ***DSD*** se calcula con la fórmula (5.9):

$$P(DSD \mid EA) = \frac{P(DSD\ , EA)}{P(EA)} \quad (5.9)$$

Las TPCs para las variables DS_{SV} y DS_{SNV} se calculan con las fórmulas (5.10) y (5.11):

$$P(DS_{SV} \mid DSD, EA) = \frac{N(DS_{SV}, DSD, EA)}{N(DSD, EA)} \quad (5.10)$$

donde

DS_{SV}: deterioro semántico en el dominio seres vivos, ∈ {ausente, presente}

$$P(DS_{SNV} \mid DSD, EA) = \frac{N(DS_{SNV,x1}, DSD, EA)}{N(DSD, EA)} \quad (5.11)$$

donde

DS_{SNV}: deterioro semántico en el dominio seres no vivos, ∈ {ausente, presente}.

Las TPCs de las variables que representan los rasgos semánticos se calculan con la fórmula (5.12):

$$P(RS_{CAT} \mid DS_{CAT}) = \frac{N(RS_{CAT}, DS_{CAT})}{N(DS_{CAT})} \quad (5.12)$$

Donde

CAT: categorías semánticas, ∈ {*manzana, perro, pino, coche, silla pantalón*}

RS: rasgos semánticos, ∈ {*taxonómico, tipos, parte-todo, funcional, evaluativo, lugar/hábitat, comportamiento, causa/genera, procedimental, ciclo vital, otros*}

$DS_{RS,\ CAT}$: $Deterioro\ semántico\ de\ CAT\ y\ RS\ ,\ \in \{ausente, presente\}$

5.5 Modelo 3: Inferencia por razonamiento abductivo y estudio del deterioro semántico diferencial entre los dominios SV y SNV.

El razonamiento abductivo parte de los síntomas y busca la hipótesis que mejor explica esos síntomas. Esa hipótesis se busca para ser la mejor explicación o la más probable. Esta BN parte de los síntomas, es decir, el DC que afecta a la capacidad de producción oral de rasgos semánticos y a partir de la cual, se infiere la probabilidad de que la EA sea la causa más probable de ese trastorno. A su vez el ***DSD*** se infiere partir de los déficits léxico-semánticos-conceptuales de categorías específicas. El objetivo de esta BN es comparar, mediante la inferencia de las BNs, la mejora en el rendimiento que se obtiene al modelar explícitamente el deterioro focalizado entre los dominios semánticos SV y SNV.

5.5.1 Modelo cualitativo.

El modelo 3 de BN es similar al de la sección anterior, salvo por los enlaces causales ***EA→DSD***, ***EA→DS$_{SV}$*** y ***EA→DS$_{SNV}$***. Estas diferencias en los enlaces causales parte de la siguiente premisa: dado un ***DSD*** ¿cuál es la probabilidad de que la EA sea la causa de la alteración cognitiva que afecta a la memoria semántica es sus aspectos declarativos? El método de diagnóstico que se plantea en esta tesis parte de la posibilidad de detectar los déficits léxico-semánticos-conceptuales a partir de las definiciones orales con restricción temporal. La aparición de estos déficits señala la presencia de un deterioro de la memoria semántica. Evidentemente, al cambiar el sentido de los enlaces causales, se requiere un algoritmo específico de aprendizaje automático de las TPCs de las variables afectadas, es decir, un algoritmo distinto al utilizado para el modelo cuantitativo de la BN de la sección 5.3.

Según la investigación llevada a cabo por Peraita y Grasso [1], la EA produce un mayor daño en las áreas temporolímbicas en las fases tempranas de la enfermedad, que puede manifestarse en un deterioro selectivo en el dominio semántico SV antes que en el dominio semántico SNV. No obstante, esto sólo es evidente en los estadios tempranos de la enfermedad, según va progresando la enfermedad, el daño se vuelve tan omnipresente que los errores ocurren en ambos dominios con igual frecuencia. La BN de la Figura 16 permite representar las conclusiones de esta investigación, tanto en su modelo cualitativo y como en su modelo cuantitativo.

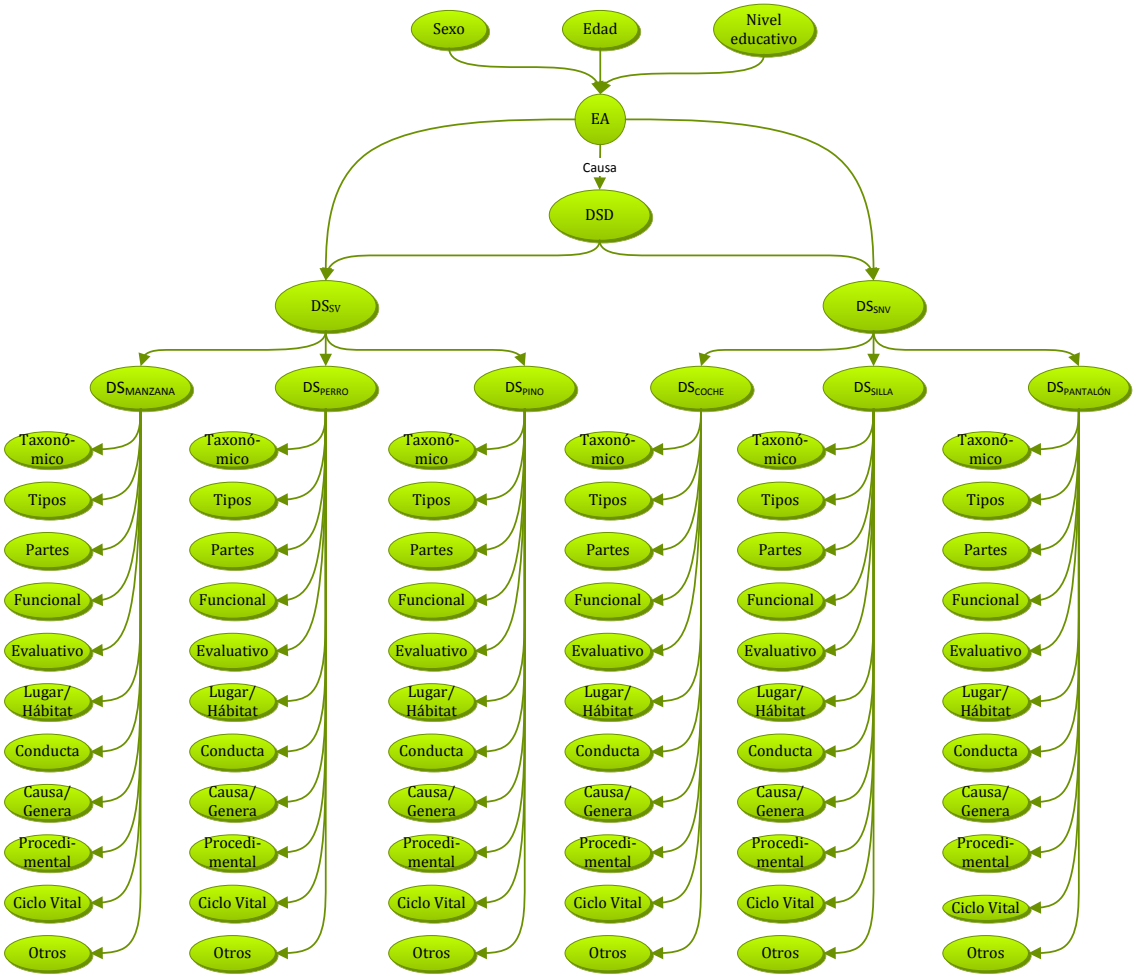


Figura 16.- Modelo 3 de BN. Inferencia por razonamiento abductivo y deterioro semántico diferencial entre SV y SNV.

5.5.2 Modelo cuantitativo.

El aprendizaje automático de los parámetros de esta BN es similar a la de la BN de la sección 5.3, pero con algunas diferencias que se detallan en esta sección.

La TPC de la variable ***EA*** se ha calculado con el simplificador *Naive Bayes* con la siguiente fórmula:

$$P(EA|E,S,N) = \alpha * P(EA) * P(E \mid EA) * P(S \mid EA) * P(N \mid EA) \quad (5.13)$$

donde

α: Factor correctivo.

E: $Edad, \in \{0\ a\ 64, 65\ a\ 69, 70\ a\ 74, 75\ a\ 79, 80\ a\ 84, más\ de\ 85\}$

S: $Sexo, \in \{hombre, mujer\}$

N: $Nivel\ educativo, \in \{primarios, secundarios, universitarios\}$

$$P(E \mid EA) = \frac{P(E, EA)}{P(EA)}$$

$$P(S\,|EA) = \frac{P(S, EA)}{P(EA)}$$

$$P(N \mid EA) = \frac{P(N, EA)}{P(EA)}$$

Para esta BN todos los cálculos se han realizado a partir del corpus lingüístico de definiciones orales [1].

La TPC para la variable ***DSD*** se ha calculado con la fórmula (5.14):

$$P(DSD|EA) = \frac{P(DSD, EA)}{P(EA)} \qquad (5.14)$$

Las TPCs para las variables DS_{SV} y DS_{SNV} para esta BN se calcula con las fórmulas (5.15) y (5.16):

$$P(DS_{SV}|DSD, EA) = \frac{N(DS_{SV}, DSD, EA)}{N(DSD, EA)} \qquad (5.15)$$

Y la TPC de la variable DS_{SNV} se ha calculado con la siguiente fórmula:

$$P(DS_{SNV}|DSD, EA) = \frac{N(DS_{SNV}, DSD, EA)}{N(DSD, EA)} \qquad (5.16)$$

El resto de cálculos son idénticos al modelo de la BN 2.

Capítulo

Modelado con BNs Híbridas

6

Hasta ahora se han considerado BNs exclusivamente con variables discretas, sin embargo existen razones por las que usar variables continuas, como por ejemplo, podría abaratar los costes del trabajo de campo, también nos han permitido realizar una inferencias con razonamiento deductivo, entre los rasgos semánticos (variables predictoras) y las categorías semánticas (variables criterios), a partir de una red de ecuaciones lineales estructurales, lo cual nos ha facilitado la realización de determinados experimentos que se detallan en el capítulo 9. En este capítulo se profundizará en el modelado –para el diagnóstico del DSD— de una BN híbrida, la cual usa variables discretas y continuas. Aportamos en esta tesis dos modelos de BNs hibridas, que comparten estructura aunque difieren en los mecanismos de inferencia: a) una *Conditional Linear Gaussian Bayesian Network,* también conocida como CLG BN, b) una BN híbrida con un método de inferencia aproximada. Las BNs híbridas se basan en relaciones lineales entre variables, dando lugar a distribuciones gaussianas multivariante cuyo propósito es describir variables no observadas descritas a partir de los rasgos semánticos.

La estructura de las BNs híbridas que presentamos en esta sección tiene parecido (lógicamente) con los anteriores modelos en cuanto al número de variables, pero cambia la dirección de los enlaces y por supuesto el tipo de las variables.

La organización del capítulo es la siguiente. En la sección 6.1 se realiza una descripción de las CLG BN. En la sección 6.2 se modela una BN Híbrida y se describe el algoritmo de aprendizaje automático para el modelo cuantitativo de esta BN. En la sección 6.3 se diseñan distintos métodos de inferencia. Con estos métodos de inferencia se consiguen mejorar los resultados respecto a otros algoritmos de minería de datos. En el primer método de inferencia, denominado CLG BN, se consideran un conjunto de variables continuas latentes que se relacionan linealmente con sus variables predictoras. En el segundo método de inferencia, denominado inferencia aproximada, las variables latentes se calculan a partir de la ponderación de la suma de las variables predictoras.

6.1 Introducción.

Una CLG BN $N = (\mathcal{X}, \mathcal{G}, \mathcal{P}, \mathcal{F})$ consiste en un conjunto de variables $\mathcal{X}$, un grafo dirigido acíclico $\mathcal{G} = (V, E)$, un conjunto de distribuciones de probabilidades

condicionales $\mathcal{P}$ y un conjunto de funciones de densidad $\mathcal{F}$. Hay una función de densidad por cada variable continua. Las variables $\mathcal{X}$ se dividen en un conjunto de variables discretas y un conjunto de variables continuas. Cada nodo de $\mathcal{G}$ representa o bien, una variable aleatoria discreta que toma un conjunto finito de estados mutuamente exclusivos y exhaustivos, o bien, una variable aleatoria con distribución Gaussiana [36].

La función de distribución de Gauss o distribución gaussiana [47] es una de las distribuciones de probabilidad que con más frecuencia aparecen en fenómenos reales. La importancia de esta distribución radica en que permite modelar numerosos fenómenos naturales, sociales y psicológicos.

Como se ha indicado anteriormente, dada la dificultad existente en la obtención de nuevos casos, se ha considerado interesante estudiar el comportamiento de las CLG BN, ya que es sumamente importante poder construir, con algoritmos de aprendizaje automático, el modelo cuantitativo de las BNs tan sólo con una muestra de personas cognitivamente sanas. Se parte de la siguiente premisa, si la EA produce una alteración cognitiva que afecta a tareas de producción oral de rasgos semánticos, bastaría con una muestra de sujetos sanos para calcular la media y la desviación típica del recuento de rasgos semánticos, ya que se da por hecho que las personas enfermas de EA realizarán definiciones orales más pobres en cuanto a la producción oral de rasgos. Cualquier variable continua con función de distribución de probabilidad Gaussiana puede ser especificada por su media y su desviación típica. En cualquier caso, es importante contar con una muestra estratificada por edad y/o nivel educativo que sea representativa de la población; además, es necesario disponer de una pequeña muestra de sujetos enfermos de EA para la validación del modelo.

En la CLG BN presentada en este capítulo existen variables continuas que dependen linealmente de todos sus padres, que a su vez son variables continuas. Por otro lado, existen variables discretas que son hijas de variables continuas, en este caso se construyen árboles de probabilidades con el algoritmo J48.

Es interesante utilizar variables intermedias o latentes ya que aunque estas no tienen un interés para el diagnóstico final, permiten explicar mejor el modelo, el resultado y lo más importante, estudiar el deterioro selectivo que presentan algunos enfermos de EA en fase leve. Por otro lado, existen otros estudios en el campo de la psicología en el que se utilizan modelos causales con variables latentes con un gran potencial en esta área [7]. Como se ha indicado en el capítulo 3, las definiciones orales solicitadas a los participantes de la muestra, se han segmentado en once rasgos semánticos y es posible definir un método para identificar los déficits léxico-semánticos-conceptuales mediante construcciones hipotéticas de determinadas teorías de investigación con variables latentes. Para construir estas hipótesis partimos principalmente de un problema estadístico, en el que tratamos de encontrar un modelo que se ajuste bien al problema y nos proporcione una medida cuantitativa de estos déficits. Por ejemplo, definimos el déficit léxico-semántico-conceptual de la categoría semántica ***manzana*** a partir de la relación lineal de los bloques conceptuales en los que se han segmentado las

definiciones orales. Estas relaciones lineales dan lugar a variables latentes o variables intermedias, cuyo fin es explicar y entender mejor la inferencia producida por la BN, al mismo tiempo que permiten realizar un mejor modelado del problema. Una vez definidos estos déficits léxico-semánticos-conceptuales, buscamos patrones en la aparición de estos déficits en enfermos de EA, ayudándonos para ello de las técnicas de IA descritas en el capítulo 3.

Una variable aleatoria continua de $\mathcal{X}$ tiene una función distribución normal μ y σ, tal que cada suceso de Y toma la probabilidad de que el sujeto esté cognitivamente sano:

$$P(X \in Y) = \int_{Y}^{\infty} p(x)dx \quad (6.1)$$

donde

$p(x)$: es la función de densidad de X con media μ y varianza σ^2 $\mathbb{N}(\mu, \sigma^2)$

$$p(x;\ \mu, \sigma^2) = \mathbb{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]} \quad (6.2)$$

Para las variables que representan los rasgos semánticos las probabilidades a posteriori se calculan a partir de la función de distribución Gaussiana o normal. Los coeficientes $\mu\ y\ \sigma^2$ se calculan por cada rasgo semántico a partir de la producción oral de rasgos semánticos de la muestra de sujetos sanos. Por ejemplo:

$$\mu_{taxonómico} = \frac{\sum_{i=1}^{n} f_i(taxonómico)}{n} \quad (6.3)$$

$$\sigma^2_{taxonómico} = \frac{\sum_{i=1}^{n} f_i(taxonómico) - \mu_{i(taxonómico)}}{n} \quad (6.4)$$

donde

n: Es el número de casos

Esta formulación se utiliza para las variables del corpus y las variables intermedias.

6.2 Modelado de una BN Híbrida.

6.2.1 Descripción del modelo.

Se ha modelado la BN híbrida usando las mismas variables que las BNs discretas del capítulo anterior. Sin embargo, se ha invertido el sentido de todos los enlaces causales que conectan las variables intermedias con las variables que representan los síntomas. En la Figura 17 se puede comprobar cómo las variables intermedias o latentes $DS_{manzana}$, DS_{perro}, DS_{pino}, DS_{coche}, DS_{silla} y $DS_{pantalón}$ son nodos hijos de los rasgos semánticos que lo componen. La inferencia de estas variables intermedias se realiza con razonamiento deductivo.

En la BN híbrida de la Figura 17 se puede observar cómo las variables que representan los rasgos semánticos o atributos de cada categoría semántica, son nodos padres de la variable intermedia o latente que representa dicha categoría semántica.

Se ha diseñado un algoritmo de inferencia para las variables intermedias o latentes continuas, que a su vez tienen *n* padres que también son variables continuas. Para la inferencia de las variables discretas se ha utilizado Elvira [33].

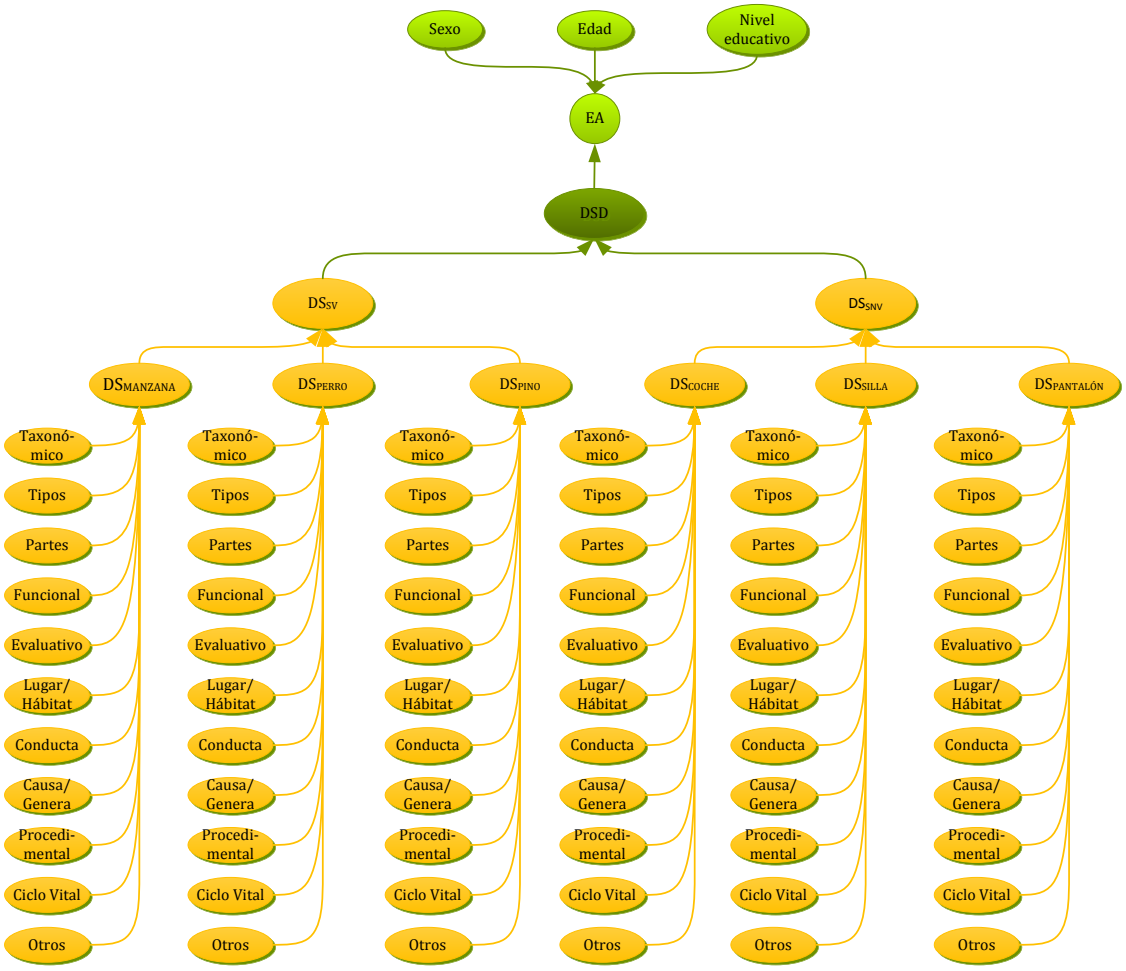


Figura 17.- BN híbrida - Razonamiento deductivo.

A diferencia de las BNs discretas, la BN híbrida utiliza distintos métodos de inferencia en función de cada tipo de variable. Los tipos de variables de la BN híbrida de la Figura 17 son:

- **Variables discretas** (nodo de color verde claro).
- **Variables discretas que utiliza arboles de probabilidad** (nodo de color verde oscuro).
- **Variables continuas** (color anaranjado).

Para los diferentes tipos de variables los métodos de inferencia diseñados e implementados en esta tesis doctoral son:

- **Variable discreta *EA***. Se calcula propagando las evidencias con Elvira [33]. En BN híbrida de la Figura 17 la probabilidad a posteriori de padecer la EA por la técnica de fuerza bruta sería:

$$P(ea_p|ne, sexo, edad) = \frac{P(ea_p, ne, sexo, edad)}{P(ne, sexo, edad)} \qquad (6.5)$$

$$P(ea_p, ne, sexo, edad) = P(ne) * P(sexo) * P(edad) * P(ea_p|ne, sexo, edad, dc_p) + P(ne) * P(sexo) * P(edad) * P(ea_p|ne, sexo, edad, dc_a)$$

$$P(ea_a, ne, sexo, edad) = P(ne) * P(sexo) * P(edad) * P(ea_a|ne, sexo, edad, dc_p) + P(ne) * P(sexo) * P(edad) * P(ea_a|ne, sexo, edad, dc_a)$$

$$P(ne, sexo, edad) = P(ea_p, ne, sexo, edad) + P(ea_a, ne, sexo, edad) \qquad (6.6)$$

donde

P(ne): Prevalencia de la EA por nivel educativo.

P(sexo): Prevalencia de la EA por sexo.

P(edad):Prevalencia de la EA por edad.

P(ea_p): Probabilidad de que la EA esté presente.

P(ea_a): Probabilidad de que la EA esté ausente.

DSD: En la TPC de esta variable, se crea un árbol de probabilidad de forma dinámica, una vez calculada la probabilidad a posteriori de padecer los déficits léxico-semánticos-conceptuales en los dominios semánticos SV y SNV.

*Variables continuas***:** Las variables continuas (color anaranjado) se calculan a partir de la producción oral de rasgos semánticos de las definiciones orales, una vez analizadas, interpretadas y posteriormente segmentadas en rasgos semánticos. En las BNs híbridas no se segmentan los casos ni por edad, ni por nivel educativo. Las variables continuas se calculan con la función de distribución gaussiana o normal:

$$P(X \in Y) = \int_Y^{\infty} p(x)dx$$

Es interesante el trabajo de Peter M. Bentler [7] en el que se utiliza un modelo causal con variables latentes. En este trabajo el bloque de construcción básico del modelo causal es una ecuación de regresión lineal.

En la ecuación (6.7) se especifica el efecto hipotetizado de ciertas variables (denominadas predictor) sobre otras variables (denominadas criterio o estándar).

$$DS_{man} = \alpha_1 taxonómico_{man} + \alpha_2 tipos_{man} + \cdots + \alpha_{11} otros_{man} \qquad (6.7)$$

donde:

$\alpha_1 .. \alpha_{11}$: Representan los coeficientes de regresión.

taxonómico, tipos...: Son las variables predictoras.

DS_{man}*:* Variable criterio o estándar.

man: Categoría semántica manzana.

Un posible diagrama *path* de este bloque de construcción, según el ratio de la ganancia de información, se representa en la Figura 18. Los cuadrados se usan para representar las variables que son medidas por la producción oral de rasgos semánticos. Los círculos dobles son las variables intermedias o criterios. Las influencias causales de los predictores sobre los criterios se representan con flechas unidireccionales. La fuerza del efecto de las predictoras sobre su criterio es indicada por el peso de cada flecha –grosor de la línea. En el modelo causal de la Figura 18 construimos una ecuación de regresión, denominada ecuación estructural, y los parámetros, parámetros estructurales. Implícitamente en cada ecuación están los parámetros asociados con las varianzas de las variables predictoras y con sus covarianzas.

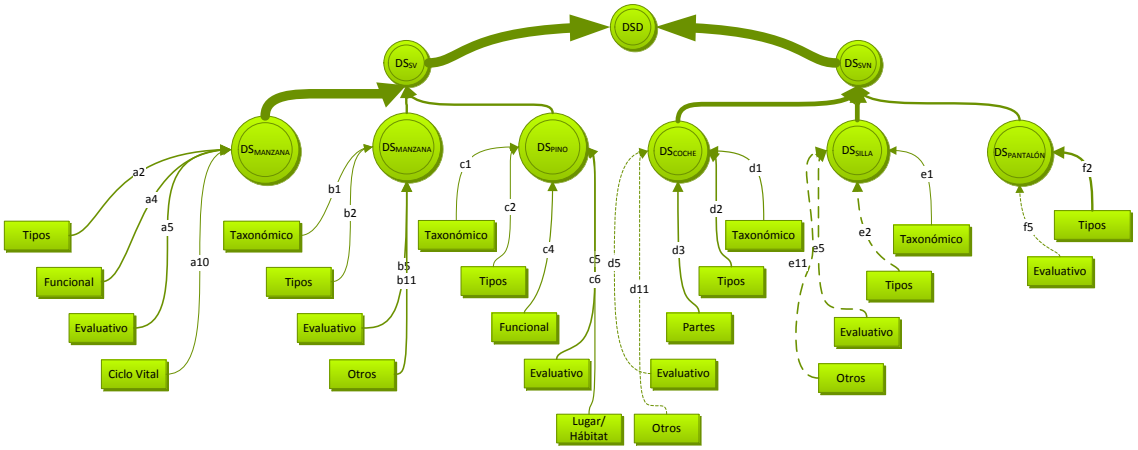


Figura 18.- Diagrama Path que representa la ecuación estructural según el coeficiente de ganancia de información.

En la Figura 18 no se han representado las variables cuyo ratio de la ganancia de información es un valor bajo, es decir, aquellas variables que son poco relevantes para la predicción de la EA. Utilizando el ratio de la ganancia de información en la construcción del diagrama *path* de la Figura 18, se han conseguido unos resultados excepcionales, como se puede comprobar en el capítulo 9.

6.2.2 Algoritmo de aprendizaje automático del modelo cuantitativo.

En esta sección se detalla el algoritmo de aprendizaje del modelo cuantitativo de la BN híbrida. El algoritmo de aprendizaje calcula la μ y la σ a partir de la muestra de sujetos sanos. Otra de las funciones de este algoritmo de aprendizaje automático es calcular las medidas de asociación o correlación siguientes: el coeficiente de correlación de Pearson, la distancia euclídea modificada, los coeficientes de regresión, el ratio de la ganancia de información y la ponderación de atributos. En el Algoritmo 1 se describe el proceso de aprendizaje automático del modelo cuantitativo de esta BN.

Algoritmo 1.- Algoritmo de aprendizaje automático

Definiciones previas.

Dominio semántico (DS): SV y SNV.

Categorías semánticas para DS_{SV}: manzana, perro y pino.

Categorías semánticas para DS_{SNV}: coche, silla y pantalón.

Rasgos por categoría semántica (rasgo): *taxonómicos, tipos, parte-todo, funcional, evaluativo, lugar y hábitat, comportamiento, causa/genera, procedimental, ciclo vital* y *otros*.

<Todos los cálculos se basan en la producción oral de rasgos semánticos de los ***n*** casos disponibles del corpus lingüístico de definiciones orales>

Detalle del Algoritmo.

Calcular_edia_stddev(rasgo){

Calculo por cada rasgo, según las siguientes fórmulas

$$\mu_{rasgo} = \frac{\sum_{i=1}^{n} rasgo_i}{n} \qquad \sigma_{rasgo} = \sqrt{\frac{\sum_{i=1}^{n}(rasgo_i - \mu_{rasgo})}{n}}$$

Almacenar resultados en hash p1

}

Calcular_distribuciones_de_probabilidad_gaussiana(rasgo){

$$P(EA_{presente}) = 1 - \int_{Y}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]} dx$$

Distribución normal de probabilidad de padecer la EA. La ecuación (6.1) define la probabilidad de estar cognitivamente sano. Esta fórmula transforma la probabilidad de estar cognitivamente sano por la probabilidad de sufrir un DSD, por esa razón se le resta 1.

Almacenar_Resultados en hash p2

}

Main () {

Calcular_media_stddev($\forall dominio, categoria\ y\ rasgo\ semántico$)

Calcular distribuciones de probabilidad ($\forall dominio, categoria\ y\ rasgo\ semántico$)

Calcular coeficientes de correlación y asociación[7](ver detalles capítulo 4) a partir de las distribuciones de probabilidad:

a.- Pearson.

b.- Distancia euclídea modificada.

c.- Pesos de atributos por producción lingüística.

[7] *Los coeficientes de correlación se calculan a partir de las distribuciones de probabilidad acumulada.*

d.- Regresión simple.

e.- Ganancia de información.

Almacenar resultados en hash r1.

Construir arboles de probabilidad para la variable DSD con el algoritmo J48.

Almacenar resultados en hash exp1.}

Todos los resultados se van almacenando en tablas hash con varios niveles de profundidad que se definen en función del tipo de segmentación de conglomerados de atributos (por edad o nivel educativo).

6.3 Inferencia.

En esta tesis doctoral se diseñan e implementan varios métodos de inferencia para la BN híbrida. Elvira no soporta los algoritmos que necesitamos y los hemos implementado. De forma general en las BNs híbridas, las variables continuas no son padres de variables discretas, sin embargo, se pueden construir las TPCs para estas variables, utilizando árboles de probabilidad. En las CLG BN tienen lugar cuatro métodos de inferencia distintos, los cuales se detallan a continuación. El Algoritmo 2 describe el proceso necesario para la inferencia de las variables utilizado para las CLG BN.

Algoritmo 2.- Inferencia por las CLG BN

Esquema general del Algoritmo.

Calcular_distribuciones_de_probabilidad_gaussiana(rasgo)

Calcular_probabilidad_variables_intermedias {

Ver apartado 6.3.2 y 6.3.3 para ver detalles

}

Propagar evidencias por las variables discretas {

DSD→Evaluar árbol de probabilidad y asignar probabilidad.

EA→Propagar evidencias variable discretas

Evaluar resultados

}

En la sección 6.3.1 se detalla cómo se construye el árbol de probabilidad y cómo se infiere la probabilidad por la BN híbrida. Los apartados 6.3.2 y 6.3.3 describen métodos de inferencia mutuamente exclusivos, es decir, no se aplican de forma simultánea en un mismo proceso de inferencia.

En el método de inferencia detallado en la sección 6.3.2, las variables continuas que son hijas de otras variables continuas (variables intermedias: *déficit léxico-semántico-conceptual* ***manzana***, ***perro***, ***pino***, ***coche***, ***silla***, ***pantalón***, ***SV*** y ***SNV)*** utilizan ecuaciones lineales estructurales, ya que se tratan de variables no observadas. El método

de inferencia descrito en la sección 6.3.3 se denomina *Inferencia aproximada* y es similar al descrito en la sección 6.3.2 excepto en el método de inferencia de las variables continuas que son hijas de otras variables continuas, es decir, las variables intermedias. En este caso las probabilidades a posteriori de las variables intermedias se calculan a partir de expresiones lineales de las probabilidades a posteriori de las variables predictoras, es decir, las expresiones lineales se construyen a partir de las probabilidades a posteriori de sus variables predictoras.

Hemos considerado muy importante la segmentación de los atributos en los once bloques conceptuales que propone el corpus lingüístico [1]. Sería posible establecer un método de inferencia que sólo tuviera en cuenta la producción oral de rasgos semánticos de todas las categorías semánticas en conjunto, pero el clasificador tiene un rendimiento inferior al clasificador que segmenta las definiciones orales en rasgos semánticos.

6.3.1 Inferencia con árboles de probabilidad.

En el modelo de BN híbrida de la Figura 17, existe una variable ***DSD*** cuyos padres son variables continuas. Un fragmento de esta BN se representa en la Figura 19.

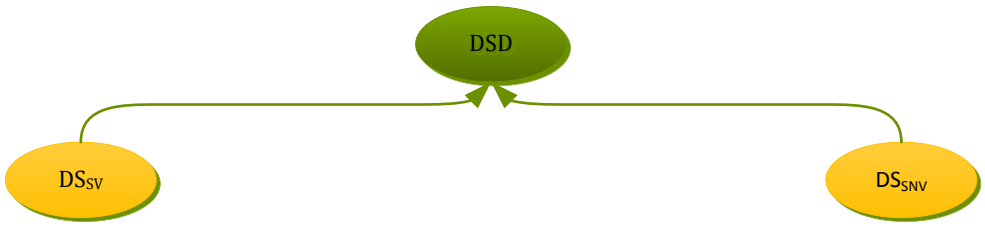


Figura 19.- Variable DSD de la CLG BN cuya inferencia se realiza mediante un árbol de probabilidad.

La TPC de esta variable se construye con un árbol de probabilidad que se crea con un algoritmo de aprendizaje automático del modelo cuantitativo. Una vez creado el árbol de probabilidad, se calculan el TPR (true positive rate) y TNR (true negative rate) para asignarle a cada rama del árbol la eficacia predictiva del nodo de decisión. El algoritmo de inferencia transforma el árbol de decisión en un conjunto de expresiones condicionales, asignando a cada expresión condicional un valor cuantitativo que representa la eficacia de la expresión en la predicción de la EA.

$$TPR = \frac{TP}{P} \qquad TNR = \frac{TN}{N} \qquad (6.8)$$

donde

TP: True positive.

TN: True negative.

TPR: True Positive Rate.

TNR: True Negative Rate.

En la Figura 20 se muestra un ejemplo de un árbol de probabilidad creado a partir de los experimentos.

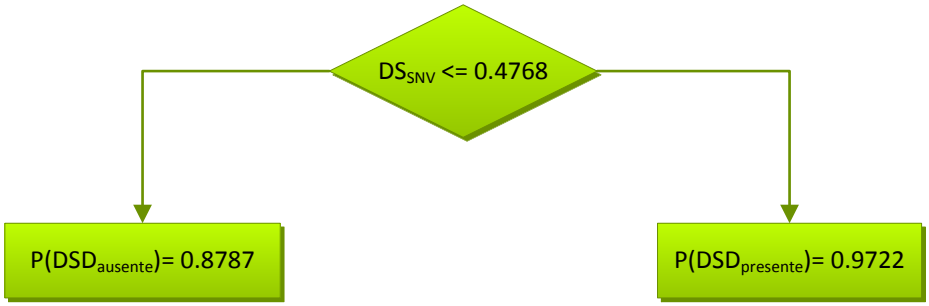


Figura 20.- Ejemplo de árbol de probabilidad creado para la variable DSD. Inferencia para variables continúas.

Los árboles de probabilidad se construyen con el algoritmo J48, que es la implementación de C4.5. El aprendizaje por árboles de decisión es un método por aproximación en el que la función a aprender es representada, como su nombre indica, por un árbol de decisión. Una vez construido el árbol se transforma en un conjunto de reglas *if-then* que constituye la base de conocimiento de un motor de evaluación de reglas en tiempo de ejecución. Los árboles de decisión es uno de los métodos de aprendizaje más populares de los algoritmos de inferencia inductiva y ha sido aplicado con éxito en una gran cantidad de tareas de aprendizaje para el diagnóstico médico.

El algoritmo C4.5 [48] es una extensión del algoritmo ID3 [41]. J48 es un *open source Java* de la implementación del algoritmo C4.5 y forma parte del *framework open source Weka*. El algoritmo C4.5 construye árboles de decisión a partir de un conjunto de datos de entrenamiento de la misma forma que lo construye el algoritmo ID3, es decir, usando el concepto de entropía (ver detalles en la sección 4.3). Cabe destacar que el algoritmo C4.5 originalmente se diseñó para trabajar con atributos discretos, mutuamente exclusivos y exhaustivos. En nuestra investigación el conjunto de entrenamiento está formado por las distribuciones de probabilidad normal o Gaussiana de las variables DS_{SV} y DS_{SNV} (calculadas a partir del recuento de rasgos semánticos), sin embargo, la clase a predecir es discreta. Existen distintas extensiones al algoritmo C4.5 para tratar las variables continuas y consiste en dividir en múltiples intervalos el atributo continuo basándose en un umbral. Los detalles del algoritmo C4.5 se puede consultar en [41,49].

Una vez construido el árbol de probabilidad se transforma en un conjunto de expresiones que serán evaluadas de forma dinámica durante el proceso de inferencia. Por tanto, podemos distinguir dos algoritmos diferentes, uno para la fase del aprendizaje automático y otro para la fase de inferencia. El Algoritmo 3 es el utilizado durante la fase de aprendizaje.

Algoritmo 3.- Aprendizaje Automático. Transformación de nodos del árbol en expresiones conjuntivas.

Detalle del Algoritmo aprendizaje automático.

```
buildExpressions() {
    data = Recuperar instancias (SV, SNV)
    Optimizar parámetro numfolds del algoritmo.
    J48.buildClassifier(data);
```

```
// El árbol nos permite encontrar umbrales de probabilidad para
// cada una de las clases EA_presente y EA_ausente
Expresions =Transformar árbol en expresiones conjuntivas.
// Se calcula la probabilidad la probabilidad de acierto
// en la clasificación de dicha expresión.
Calcular TPR/TNR por cada expresión.}
```

El Algoritmo 4 describe cómo se utilizan las expresiones condicionales generadas con el algoritmo anterior, durante el proceso de inferencia.

Algoritmo 4.- Inferencia. Evalúa una expresión conjuntiva y asigna la probabilidad TPR o TNR.

Detalle del Algoritmo inferencia árbol de probabilidad.

```
Inferir_probabilidad_arbol(Atributos_corpus cat) {
  Para todas las expresiones {
    Evaluar_expresion(cat);
    Si (cat cumple expresión conjuntiva) {
          Return TPR/TNR;
    }
  }
```

6.3.2 Inferencia CLG en BN.

Como se ha indicado anteriormente, las BN híbridas están compuestas por un subconjunto de variables discretas y un subconjunto de variables continuas, dando lugar a distintos tipos de relaciones causales y métodos de inferencia. Cada uno de estos tipos de variables lleva asociado un método de inferencia:

- Variables discretas cuyos padres son variables discretas. En este caso se propagan las probabilidades a posteriori como si de una BN discreta se tratará.
- Variables discretas que son hijas de variables continuas. Para este tipo de relaciones causales, en la fase de aprendizaje de la BN se construyen arboles de probabilidad con el algoritmo C4.5, tal y como se ha explicado en la sección anterior.
- Variables continuas que no tienen padres. En este caso se aplica la fórmula descrita anteriormente en el Algoritmo 1:

$$P(EA_{presente}) = 1 - \int_{Y}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]} dx \qquad (6.9)$$

- Variables continuas que son hijas de otras variables continuas, que es el objeto de esta sección y se detallará a continuación.

Como se ha indicado anteriormente, la EA produce un deterioro semántico focalizado en estadios tempranos. Con este algoritmo buscamos patrones en la producción oral de

rasgos semánticos, a partir del deterioro semántico diferencial, ponderando el grado de predicción de cada variable del corpus.

Partiendo de estos descubrimientos científicos [1,13,7] sobre la EA, se considera que las categorías semánticas se pueden expresar como una expresión lineal de sus variables predictoras –los rasgos semánticos. Un ejemplo del modelo causal es el representado en la Figura 21, donde la probabilidad de sufrir déficit léxico-semántico-conceptual en la categoría semántica ***manzana*** se expresa como una expresión lineal de los bloques semánticos que la componen. En este caso, la variable ***manzana*** es una variable continua y es hija de los bloques conceptuales (*taxonómico, tipos, parte-todo, funcional, evaluativo, lugar/hábitat, conducta, causa/genera, procedimental, ciclo vital* y *otros)* de dicha categoría semántica.

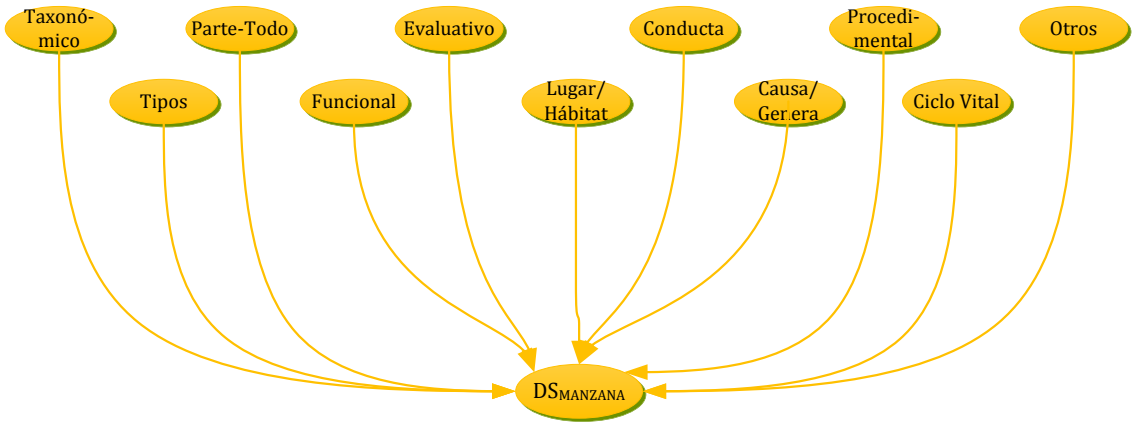


Figura 21.- Fragmento de la CLG BN para la variable intermedia $DS_{MANZANA}$.

La variable ***manzana*** (déficit léxico-semántico-conceptual de la categoría semántica manzana) se expresa como una expresión lineal de todos sus rasgos semánticos:

$$cs = \sum_{i \in Rasgos_{cs}} \alpha_i * rasgo_i \qquad (6.10)$$

donde

cs: categoría semántica, ∈ {*manzana, perro, pino, coche, silla, pantalón*}

rasgo: rasgo semántico, ∈ {*taxonómico, tipos, parte-todo, funcional, evaluativo, hábitat, conducta, causa, procedimental, ciclo vital* y *otros*}.

α: es el coeficiente de correlación o asociación de la EA con cada rasgo semánticos.

Las variables intermedias y de interés tienen más de una variable continua como padre, por tanto, la media de estas variables depende linealmente sus padres, es decir, depende linealmente de los rasgos semánticos (unidades menores y significativas). La fórmula utilizada para calcular la distribución de probabilidad de las variables intermedias es:

$$P(cs|rasgos_{cs}) = 1 - \int_{cs} \mathbb{N}(cs, \sum_{z \in rasgos}^{x\ cs} (\mu_{xz} + \alpha_{xz}),\ \sigma_{cs}^2) \qquad (6.11)$$

donde

$\mu_{xz} + \alpha_{xz}$*:* Se calcula sumando la media de la producción oral de rasgos semánticos, más el coeficiente de asociación.

σ_{cs}^2: Varianza de la categoría semántica, calculada a partir de la expresión lineal.

El algoritmo de aprendizaje automático del modelo cuantitativo requiere dos fases. En la primera fase, se calculan los coeficientes: μ, σ, coeficiente de correlación de Pearson, distancia euclídea modificada, coeficientes de regresión, ganancia de información y pesos de atributos por producción oral de rasgos semánticos. En la segunda fase, se calculan las variables intermedias y los coeficientes de regresión que forman la expresión lineal. Una vez obtenido el resultado de las expresiones lineales se calculan la μ y σ para cada una de las variables intermedias (ver ecuación 6.12).

En el apartado de experimentos se muestran los resultados conseguidos con las CLG BN.

6.3.3 Método de inferencia aproximado.

El método de inferencia aproximado comparte los métodos de inferencia con el método anterior, excepto para las variables continuas cuyos padres son también variables continuas. Por tanto, en esta sección nos centraremos en este caso.

Se considera la siguiente premisa *la producción oral de rasgos semánticos de una categoría semántica, es la suma de la producción oral de rasgos semánticos de cada uno de sus rasgos*, tal y como se expresa en la fórmula (6.12).

$$POA_{categoria\ semántica} = \sum POA_{rasgos} \qquad (6.12)$$

donde

$POA_{categoria\ semántica}$: Recuento de la producción oral de rasgos semánticos para las categorías semánticas ∈ {*manzana, perro, pino, coche, silla, pantalón*}

POA_{rasgos}: Recuento de la producción oral de rasgos semánticos para las categorías semánticas ∈ {*taxonómicos, tipos, parte-todo, funcional, evaluativo, lugar y hábitat, comportamiento, causa/genera, procedimental, ciclo vital y otros*}

Partiendo de esta premisa, podemos definir que la probabilidad de padecer un déficit léxico-semántico-conceptual en una categoría semántica determinada, es la probabilidad de padecer un déficit léxico-semántico-conceptual de cada uno de sus rasgos que la componen, pero además, ponderado por un factor predictivo de la *EA* y dividido por el número total de rasgos, tal y como se muestra en las siguientes ecuaciones:

$$P(cs) = \sum_{z \in Rasgos} \frac{\alpha_z P(z)}{N}$$

$$P(\neg cs) = \sum_{z \in Rasgos} \frac{\alpha_z P(\neg z)}{N}$$

$$1 = \delta(P(cs) + P(\neg cs)) \qquad (6.13)$$

donde

P(cs): Probabilidad de padecer un déficit léxico-semántico-conceptual en la categoría semántica *cs*. La probabilidad se calcula con la fórmula:

$$P(EA_{presente}) = 1 - \int_{x}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]} dx \qquad (6.14)$$

P(¬cs): Probabilidad de no padecer un déficit léxico-semántico-conceptual en la categoría semántica (cs), que se calcula *1 – P(*$EA_{presente}$*)*

N: Número de rasgos de cada categoría semántica.

α_z: Coeficiente de asociación o correlación.

δ: Factor correctivo.

Hemos desarrollado una variante del método de inferencia aproximado para el cual se utilizan dos coeficientes de regresión (ver detalles en la sección 4.3), uno para la muestra de sujetos sanos y otro para la muestra de sujetos enfermos de EA. Por tanto, esta variante respecto a la ecuación 1=$\delta(P(cs) + P(\neg cs))$) es:

$$P(cs) = \sum_{z \in Rasgos} \frac{r_{1z} P(z)}{N}$$

$$P(\neg cs) = \sum_{z \in Rasgos} \frac{r_{2z} P(\neg z)}{N}$$

$$1 = \delta\big(P(cs) + P(\neg cs)\big) \qquad (6.15)$$

El método de inferencia aproximado pretende construir una expresión lineal que haga que la $P(EA_{presente})$ tienda a 1, para las personas que realmente padecen la EA, y haga que la $P(EA_{ausente})$ tienda a 0, para las personas que realmente están cognitivamente sanas.

Como indica el título de esta sección, este método de inferencia es aproximado, es decir, la probabilidad a posteriori de las variables intermedias no se basa en frecuencias, ni en recuentos. El fundamento principal de este método es el grado de déficit léxico-semántico-conceptual de una categoría semántica determinada, es igual a la suma ponderada del grado de déficit léxico-semántico-conceptual de cada uno de los rasgos que en los que se ha segmentado la definición oral. Como se ha indicado anteriormente, el grado de déficits es un problema estadístico, donde tratamos de encontrar una medida

cuantitativa de estos déficits. Algunos de estos déficits se identifican mediante construcciones hipotéticas de determinadas teorías de investigación. Se podrá comprobar en los experimentos los resultados obtenidos con la BN híbrida con inferencia aproximada. Los déficits léxico-semánticos-conceptuales apuntan la presencia de un deterioro de la memoria semántica y estos a su vez indican la presencia del DC, el cual es el precursor de la EA.

Capítulo

Estrategias evolutivas para la optimización de BNs

7

Como se ha venido indicando a lo largo de esta tesis doctoral, el método de aprendizaje del modelo cuantitativo de las BNs se ha aplicado utilizando por un lado, el corpus lingüístico de definiciones orales de Peraita y Grasso [1], y por otro lado, estudios epidemiológicos obtenidos de la literatura científica. Más concretamente, la TPC de la variable ***EA*** se construye a partir del estudio epidemiológico [18], junto con un algoritmo de aprendizaje automático basado en *Naive Bayes*. Por otro lado, las variables ***Edad*** y ***Nivel educativo*** en las BNs discretas ejercen una influencia implícita en el método de aprendizaje automático del modelo cuantitativo durante el proceso de discretización. El objetivo del capítulo es describir cómo se ha mejorado el rendimiento de las BNs, optimizando la TPC asociada a la variable de interés ***EA*** con un algoritmo de estrategias evolutivas, aunque podría aplicarse a cualquier variable de la BN. Cada individuo de la población está constituido por los parámetros de la TPC de la variable ***EA*** y cada TPC da lugar a una solución suboptimal, ya que el algoritmo de estrategias evolutivas utiliza un algoritmo memético para inicializar la población.

La computación evolutiva se basa en procesos biológicos que han servido de inspiración a los investigadores [38], proporcionando ideas y metáforas. Como su nombre indica las estrategias evolutivas se inspiran en el proceso natural de la evolución de las especies y la principal de estas metáforas es el mecanismo evolutivo de resolución de problemas de tipo *prueba y error*. En esta tesis se aporta un *framework* de estrategias evolutivas, adaptado a nuestro método de diagnóstico, que se describe a continuación.

Hemos optado por las estrategias evolutivas porque se han aplicado con éxito a un amplio número de dominios distintos y su uso se está incrementando continuamente; se integra muy bien con nuestro framework y es muy flexible, en el sentido de que podemos proporcionar en la entrada, conjuntamente o individualmente, todas las TPCs (parámetros) que necesitemos, con un número indeterminado de parámetros, y obtenemos en la salida esos parámetros optimizados. Cuando buscamos un solucionador de problemas naturales podemos encontrar dos candidatos, el primero en el campo de la neurocomputación y el segundo en la computación evolutiva; en esta tesis se ha optado por la computación evolutiva.

7.1 Introducción.

El objetivo del capítulo es crear un framework para optimizar las TPCs de una BN. En esta tesis hemos limitado el problema a una sola TPC (variable ***EA***) por varias razones:

- La TPC de la variable ***EA*** se construye a partir de investigaciones obtenidas de la literatura científica. El cálculo de esta TPC se realiza de forma automática utilizando un simplificador.
- La TPC de la variable ***EA*** es la que tiene mayor número de componentes o parámetros, lo cual, hace más tedioso su construcción o revisión manual.
- El coste computacional de esta técnica aplicada a todas las variables es muy elevado y se deja para posteriores investigaciones con más medios.
- No se ha considerado la optimización conjunta de todas las TPCS de las BNs porque supone un gran número de parámetros y sería necesario un corpus de datos más extenso y estratificado, principalmente por edad y nivel educativo.

Como se ha indicado anteriormente, se ha implementado un *framework* de estrategias evolutivas específico para este problema. Las estrategias evolutivas ofrecen un gran número de métodos y operadores, los cuales se han implementado en el framework con el objetivo de seleccionar aquellos que mejor se adecuan al problema. Se puede consultar más detalle en [38]. Los métodos y operadores implementados son:

- Se implementa cuatro mecanismos para detener el algoritmo de estrategias evolutivas.
- Se implementa tres operadores de mutación: dinámico, con retroalimentación y dos operadores autoadaptativas.
- Se implementa tres métodos de penalización del *fitness*: uno estático, otro dinámico y un tercero autoadaptativo.
- Se implementa cuatro métodos para el control de tamaño de la población: (u, λ) y (μ+λ), que son estáticos; GAVaPS y PROFIGA, que son dinámicos.
- Se implementa un mecanismo para inicializar la población basado en un algoritmo memético, el cual construye cada TPC que constituye cada individuo, utilizando *Naive Bayes* y variando las prevalencias de la EA dentro de su intervalo de confianza. De esta forma se garantiza que el algoritmo comience con individuos que representan soluciones suboptimal. El algoritmo memético permite converger en pocas generaciones a una solución óptima, reduciendo así el enorme coste computacional.

El Algoritmo 5 se ha implementado utilizando una representación de números reales de *n* componentes o dimensiones, que varía en función del tamaño de la TPC. El framework permite una entrada indeterminada de parámetros, lo cual nos permite optimizar de forma conjunta un número indeterminado de TPCs. El pseudocódigo del algoritmo es el siguiente:

Algoritmo 5.- Pseudocódigo del algoritmo de optimización basado en estrategias evolutivas.

```
Procedure StrategiesEvolutive
Initializate
UpdateFitness
Si el método de control de crecimiento de GAVaPS entonces
Inicializar Lifetime de todos los individuos
Fin si
While (condition terminations is not reach)
Actualizar_parametros(mutación con retroalimentación,
penalización con retroalimentación)
reproduction();
survivorSelection();
Actualizar_parametros(finalización, PROFIGA, terminación)
End while
End Procedure
```

7.2 Representación del problema.

El *framework* implementa un algoritmo para transformar la representación de las TPCs a un vector de componentes que representan cada individuo de la población. En la Figura 22 se representa un fragmento del resultado de la transformación de una TPC a un individuo, y viceversa.

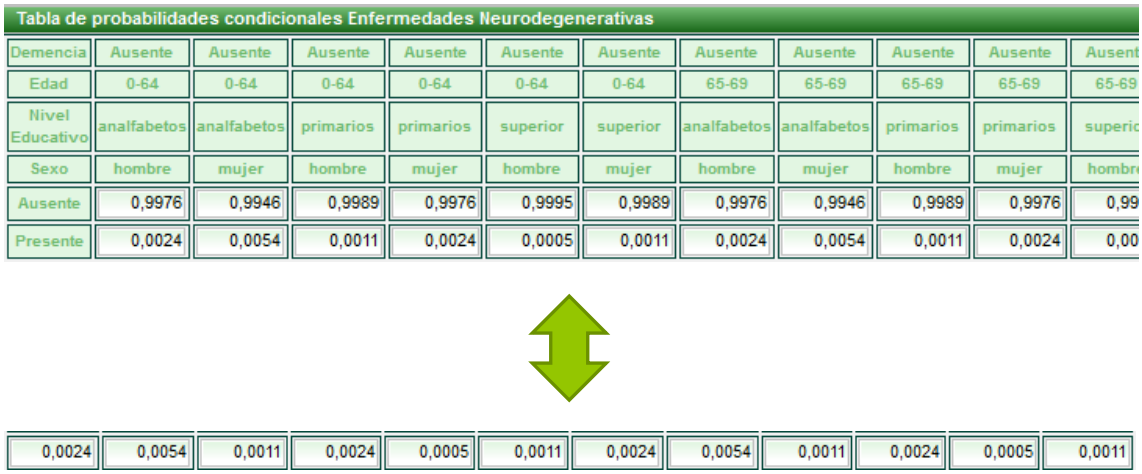


Figura 22.- Representación del proceso de transformación TPC – Componentes de un individuo.

7.2.1 Configuración del algoritmo.

Se han implementado varias estrategias de mutación, de recombinación, de control de la población y de penalización del *fitness*. El objetivo es buscar la estrategia más eficiente,

tanto desde la perspectiva de la eficacia de las BNs como del coste computacional. Las distintas estrategias vienen definidas por los siguientes factores:

- **Maximización o minimización:** indica si la función de optimización consiste en buscar un máximo o un mínimo. En los experimentos se utiliza como *fitness* el AUC que puede tomar un valor comprendido entre 0 a 1, siendo 1 el valor más óptimo. Por tanto, se trata de un problema de maximización. Es posible, en el campo de la IA y de las curvas ROC, utilizar otras métricas de rendimiento que haría necesario la utilización de la minimización, por esta razón el *framework* implementa las dos estrategias.
- **Población (IndividualsNumber):** indica el número de individuos de la población (μ).
- **Descendencia (offspring):** indica el número de descendientes por cada generación (λ). Normalmente $\lambda > \mu$.
- **Condición de terminado:** determina el umbral de precisión utilizado para la finalización del algoritmo. Es decir, el algoritmo finaliza si llega a un valor $\left(bestFitness \pm \varepsilon > optimalValue\right)$.
- **Condición de terminación sin mejora del fitness (stopK):** es un parámetro utilizado para la finalización del algoritmo. Su valor contiene el número de generaciones que se permiten sin obtener ninguna mejora del mejor fitness.
- **Número de generaciones (generationsNumber):** indica el número de ciclos máximos para finalizar el algoritmo. En el caso de que no se encuentre la solución óptima indicada en el parámetro anterior, se finalizará con un número de ciclos máximo del algoritmo.
- **Clase fitness:** indica la clase java que contiene el método de evaluación de la adaptabilidad de los individuos. Esta clase es arbitraria pero debe cumplir la interfaz java *FitnessFunction* que se debe encontrar en el classpath.
- **Operador de mutación:** indica el método de mutación que son: *Uncorrelated Mutation with One Step Size*, *Uncorrelated Mutation with n Step Size*, *Mutación dinámica o determinista* y *Rechenberg*.
- **Umbral de desviación típica (thresholdDesvTip):** indica el valor de desviación típica mínimo.
- **Límite inferior (loBound):** indica el valor del límite inferior a aplicar a todos los componentes del cromosoma.
- **Límite superior (upBound):** indica el valor del límite superior a aplicar a todos los componentes del cromosoma.
- **Precisión:** indica el número de decimales con los que va a trabajar el algoritmo.
- **Método de recombinación:** indica el método de recombinación para las variables. Se ha implementado *Intermediary recombination* y *Discrete recombination*.
- **Alcance de recombinación:** indica si se va a utilizar tanto para las variables como para los parámetros, la recombinación global –multiparental—o recombinación local –biparental.
- **Variable de recombinación:** indica el método de recombinación para los parámetros y, al igual que recombinationModel, se ha implementado *Intermediary recombination* y *Discrete recombination*, y se pueden seleccionar independientemente del método de recombinación utilizado para las variables.
- **Método de selección de supervivientes:** indica el método de selección de supervivientes, he implementado tanto para *(λ, μ)*, *(λ+μ)*, *AVGaPS* y *Profiga*.

- **Ratio de aprendizaje.** se utiliza en el método de mutación *Uncorrelated Mutation with One Step Size*. Si no se indica, de forma automática se establece el valor $\frac{1}{\sqrt{n}}$.
- **TauP:** es el parámetro *learning rate* utilizado para *Uncorrelated Mutation with n Step Size*. Su valor es aproximadamente $\frac{1}{\sqrt{2n}}$.
- **Tau:** este parámetro se utiliza para *Uncorrelated Mutation with n Step Size*. Su valor es aproximadamente $\frac{1}{\sqrt{2\sqrt{n}}}$
- **C:** constante utilizada para el método de mutación de *Rechenberg* y su valor puede contener $0{,}817 \leq c \leq 1$.
- **K:** utilizada para el método de *Rechenberg* y controla el número de generaciones antes de la actualización del parámetro sigma.
- **Método de penalización (penalty):** es el método de penalización del fitness para cuando alguno de los componentes del individuo se sale de rango. Los tipos de penalización son: estática, dinámica y auto adaptativa.
- **W:** Es el peso utilizado para la penalización estática.
- **B1:** Representa el parámetro β_1 del método de penalización dinámica.
- **B2:** Representa el parámetro β_2 del método de penalización dinámica.
- **MINLT:** Representa el *Lifetime* mínimo para el método de control de la población *GAVaPS*.
- **MAXLT:** Representa el *Lifetime* máximo para el método de control de la población *GAVaPS*.
- **GAVaPS:** Método de cálculo del *Lifetime* para *GAVaPS*. Los métodos permitidos son proporcional, lineal y bilineal.
- **Factor de incremento:** Factor de incremento para el método de control de la población *PROFIGA*.
- **V:** Parámetro utilizado por *PROFIGA* para determinar el número de generaciones sin mejoras.
- **Reducción:** Valor comprendido entre 0 y 1 para determinar la razón de reducción.

En los siguientes apartados se detallara el uso de estos parámetros.

7.2.2 Inicialización de la población.

La inicialización de la población utiliza un algoritmo memético, es decir, la TPC que se pretende optimizar se inicializa, utilizando el algoritmo de aprendizaje automático *Naive Bayes* descrito en el capítulo 5. De forma aleatoria, pero dentro del intervalo de confianza de las prevalencias, el algoritmo varía los parámetros de la TPC por mutación y recombinación. Los valores del vector varían en función de unas desviaciones típicas cuyo rango está comprendido entre el valor mínimo *thresholdDesvTip* y con valor máximo $\frac{n}{1.96}$,

donde

n: es el número de dimensiones del vector –el número de estados *Presente* de la variable ***EA***.

7.3 Condición de Terminación.

Se establecen cuatro métodos para finalizar el algoritmo:

- Al encontrar un valor de *fitness* óptimo en función del parámetro condición de terminación. El valor de *fitness* tiene que estar comprendido entre el valor óptimo ± ε para que la función de evaluación del *fitness* finalice el algoritmo.
- Cuando el algoritmo no mejora su *fitness* en el número de generaciones determinado por el parámetro *stopK*.
- Cuando se llega a un número máximo de generaciones –*generationsNumbers*— sin alcanzar un *fitness* óptimo.
- Cuando utilizamos *PROFIGA* o *GAVaPS* como algoritmo de control de población y la población queda reducida a menos de 2 individuos. Esta condición se establecen para hacer más robusto el algoritmo ante determinadas situaciones excepcionales en las que la población puede verse reducida a menos de 2 individuos.

7.4 Métodos de mutación.

Los métodos de mutación que se han implementado y que propone [38] son: a) *Uncorrelated Mutation with One Step Size*, b) *Uncorrelated Mutation with n Step Size* c) *Deterministic or dynamic mutation* d) *Adaptive mutation with feedback*. El Algoritmo 6 se utiliza para seleccionar el método de mutación.

Algoritmo 6.- Algoritmos de mutación.

```
public void mutate() {

if (mutationOperator == 1) {

uncorrelatedOneStep();

} else {

if (mutationOperator == 2) {

uncorrelatedNStep();

} else {

if (mutationOperator == 3) {

deterministicMutation();

} else {
```

```
if (mutationOperator == 4) {
adaptativeMutation();
        }
  }
 }
}
```

El discriminador *mutationOperator* se inicializa en el constructor de la clase *Individual*.

7.4.1 Uncorrelated Mutation with One Step Size.

Uncorrelated Mutation with One Step Size utiliza la misma función de distribución para mutar cada componente x_i. Así mismo, la desviación típica evoluciona junto con los valores, aunque sólo existe una desviación típica para todos los componentes. Para la evolución de las desviaciones típicas se utiliza el parámetro *learning rate*, que es un parámetro definido por el usuario y suele tener un valor aproximado $\frac{1}{\sqrt{n}}$. El ajuste de la desviación típica está limitado entre un valor mínimo, configurado en el valor *thresholdDesvTip*, y el valor máximo, que se calcula con la siguiente fórmula $\frac{n}{1.96}$. La fórmula que se ha utilizado para mutar cada uno de los componentes del individuo es:

$$\sigma' = \sigma \cdot e^{\tau \mathbb{N}(0,1)}$$
$$x_i' = x_i + \sigma' \cdot \mathbb{N}(0,1) \quad (7.1)$$

donde

τ: es un parámetro de usuario *ratio de aprendizaje* y su valor es $\propto \frac{1}{\sqrt{n}}$

n: es la dimensión del vector.

7.4.2 Uncorrelated Mutation with n Step Size.

Este operador es similar al anterior, salvo que se aplica un paso de mutación para cada variable de forma independiente. Al igual que en el caso anterior, el ajuste de la desviación típica está limitada entre un valor mínimo, configurado en el parámetro *thresholdDesvTip,* y el valor máximo, que se calcula con la siguiente fórmula $\frac{n}{1.96}$. Las fórmulas empleadas para mutar cada uno de los componentes del individuo son:

$$\sigma' = \sigma \cdot e^{\tau' \cdot \mathbb{N}(0,1) + \tau \mathbb{N}_i(0,1)}$$
$$x_i' = x_i + \sigma_i' \cdot \mathbb{N}_i(0,1) \quad (7.2)$$

donde

$$\tau' \propto \frac{1}{\sqrt{2n}}$$

$$\tau \propto \frac{1}{\sqrt{2\sqrt{n}}}$$

n: es la dimensión del problema.

7.4.3 Método de mutación dinámica.

Se define el tamaño de mutación como $\sigma(t) = 1 - 0.9 \times \frac{t}{T}$,

donde

t: es el número actual de generación

T: es el número máximo de generaciones.

Cada uno de los componentes del vector $\bar{x}$ es remplazado por la mutación gaussiana $x_i' = x_i + \sigma \cdot N\ (0,1)$.

7.4.4 Método de mutación adaptativa con realimentación.

Para este método utilizamos la regla de éxito de *Rechenberg*:

$$\sigma' = \begin{cases} \sigma / c & \text{si } p_s > 1/5 \\ \sigma \cdot c & \text{si } p_s < 1/5 \\ \sigma & \text{si } p_s = 1/5 \end{cases}$$

donde

c: es $0.817 \leq c \leq 1$. Los valores de los componentes del vector $\bar{x}$ se remplaza por: $x_i' = x_i + \sigma' \cdot N\ (0,1)$

7.5 Selección de padres.

Es una selección aleatoria con distribución uniforme sobre toda la población. Por cada recombinación se seleccionan dos o más padres –en función de si se han configurado el algoritmo con recombinación global o local. La única restricción es que en una misma recombinación no se utilice dos veces el mismo padre.

7.6 Recombinación.

Se han implementado dos estrategias de recombinación: *intermediary recombination* y *discrete recombination*. Se pueden seleccionar de forma independiente cualquiera de las estrategias, tanto para las variables como para los parámetros. Es decir, se puede seleccionar el método *intermediary recombination,* para la recombinación de las variables, y *discrete recombination,* para la recombinación de los parámetros. Además, se han extendido estas dos estrategias con la variante multiparental, es decir, recombinación global. Para generar un descendiente se seleccionan dos padres distintos con una distribución uniforme y se recombinan según el método que se haya configurado en el algoritmo. Para generar λ descendientes el proceso se repite tantas

veces como λ descendientes existan. Si la recombinación es global se utilizan dos padres diferentes por cada variable. Este proceso recombina los valores de los parámetros de penalización *W,* mencionados anteriormente.

Algoritmo 7.- Algoritmo para la recombinación

```
public ArrayList<Individual> recombination(ArrayList<Individual> parents, int offspringsNumber, int recombinationScope, int methodRecombination, int methodVariableRecombination) {

if (parents.size() <= 1) {

return new ArrayList<Individual>();

}

ArrayList<Individual> offsprings = new ArrayList<Individual>();

int pos1 = 0;

int pos2 = 0;

for (int i = 0; i < offspringsNumber; i++) {

// Seleccionamos dos padres diferentes por cada recombinación

pos1 = 0;

pos2 = 0;

while (pos1 == pos2) {

pos1 = (int) (RandUtility.nextDouble() * parents.size());

pos2 = (int) (RandUtility.nextDouble() * parents.size());

// Tiene que haber un mínimo de 4 padres

if (parents.size() <= 4) {

break;

    }

  }

Individual parent1 = parents.get(pos1);

Individual parent2 = parents.get(pos2);

Individual offspring = parent1.clone();

ArrayList<Double> xs = new ArrayList<Double>();

ArrayList<Double> desvTip = new ArrayList<Double>();

int n = parent1.getXs().size();
```

```
int nDesvTip = parent1.getDesvTip().size();

double w = 0;

// Se recombinan todos los componentes de los valores

for (int j = 0; j < n; j++) {

// Se combinan los valores

if (methodRecombination == 2) {

if (RandUtility.nextBoolean()) {

w = parent1.getW();

xs.add(parent1.getXs().get(j));

} else {

w = parent2.getW();

xs.add(parent2.getXs().get(j));

      }

} else {

w = (parent1.getW() + parent2.getW()) / 2;

double vxs          =(parent1.getXs().get(j).doubleValue() + parent2.getXs().get(j).doubleValue()) /
2;

xs.add(new Double(vxs));

   }

// Se combinan las desviaciones típicas

if (j < nDesvTip) {

if (methodVariableRecombination == 2) {

if (RandUtility.nextBoolean()) {

desvTip.add(parent1.getDesvTip().get(j));

} else {

desvTip.add(parent2.getDesvTip().get(j));

      }

} else {

double vdestip   =(parent1.getDesvTip().get(j).doubleValue()                                    +
parent2.getDesvTip().get(j).doubleValue()) / 2;
```

```
desvTip.add(new Double(vdestip));
    }
  }
// Si el ámbito es global vuelvo a seleccionar dos padres
if (recombinationScope == 2) {
pos1 = 0;
pos2 = 0;
while (pos1 == pos2) {
pos1 = (int) (RandUtility.nextDouble() * parents.size());
pos2 = (int) (RandUtility.nextDouble() * parents.size());
// Tiene que haber un mínimo de 4 padres
if (parents.size() <= 4) {
break;
      }
   }
  parent1          = parents.get(pos1);
  parent2          = parents.get(pos2);
 }
}
offspring.setW(w);
offspring.setXs(xs);
offspring.setDesvTip(desvTip);
offsprings.add(offspring);
}
return offsprings; }
```

7.7 Selección de sobrevivientes.

Para la selección de supervivientes se han implementado dos métodos: (μ,λ) y $(\mu+\lambda)$. El primero, sólo tiene en cuenta aquellos descendientes con mejor valor *fitness,* y el segundo, tiene en cuenta tanto los padres como los descendientes con mejor *fitness*.

7.7.1 Método (μ, λ).

Este método ordena la población de descendientes de menor a mayor y selecciona los últimos elementos, que son los que disponen mejor *fitness*.

Algoritmo 8.- Selección de sobrevivientes. Método (μ, λ)

```
public ArrayList<Individual> selectionOffspring(ArrayList<Individual> offsprings, int individualsNumber) {

ArrayList<Individual> population = new ArrayList<Individual>();

Collections.sort(offsprings);

for (int i = 0; i < individualsNumber; i++) {

population.add(offsprings.get(offsprings.size() - i - 1));

  }

return population;

}
```

7.7.2 Método (μ + λ).

Este método es similar al anterior, pero tiene en cuenta, para la selección de los mejores individuos, los padres y la descendencia.

Algoritmo 9.- Selección de sobrevivientes. Método (μ + λ)

```
public ArrayList<Individual> selectionParentsOffspring(ArrayList<Individual> parents, ArrayList<Individual> offsprings, int individualsNumber) {

ArrayList<Individual> population = new ArrayList<Individual>();

parents.addAll(offsprings);

Collections.sort(parents);

for (int i = 0; i < individualsNumber; i++) {

population.add(parents.get(parents.size() - i - 1));

  }

return population;

}
```

7.8 Métodos de control de población.

Se han desarrollado dos métodos para controlar la población: *GAVaPS* y *PROFIGA*, que son mutuamente exclusivos. Estos métodos para el control de la población se describen a continuación.

7.8.1 GAVaPS.

Cada individuo cuando nace, ya sea en la inicialización o en la recombinación, se le asigna una duración de vida proporcional a su valor de *fitness*, pero limitado por un tiempo de vida máximo y mínimo. En cada generación se va disminuyendo el tiempo de vida de cada individuo hasta que llega a 0. En ese momento se elimina el individuo de la población. En cada generación, una vez que los nuevos individuos se han creado por recombinación y se han mutado, se calcula su *fitness* y se fija un tiempo de vida. En cada generación se crean tantos descendientes como se hayan establecido en el parámetro *Offspring* y se añaden directamente a la población tantos individuos como se especifiquen en el parámetro *IndividualsNumber,* en función de su *fitness*.

No se establece ningún mecanismo de selección, es decir, las nuevas generaciones se añaden a la población sin selección previa y será cuando su *timelife* llegue a 0, cuando se eliminen de la población. En el Algoritmo 10 se define el método GAVaPS.

Algoritmo 10.- Control de la población GAVaPS

```
public ArrayList<Individual> selectionGAVSaPS(ArrayList<Individual> parents,
ArrayList<Individual> offsprings, int methodGAVaPS, int minLT, int maxLT, int
individualsNumber) {

ArrayList<Individual> population = parents;

ArrayList<Individual> offspring  = new ArrayList<Individual>();

if (offsprings.size() > 0) {

Collections.sort(offsprings);

for (int i = 0; i < individualsNumber; i++) {

offspring.add(offsprings.get(offsprings.size() - i - 1));

        }

// Los padres se mantienen

        double[] fitness   = computeFitness(parents, offspring);

      computeLifeTime(offspring, fitness, methodGAVaPS, minLT    , maxLT);

// Se resta una generación a los padres

substractOneGeneration(parents);

// Añado los descendientes al resto de la población

population.addAll(offspring);  }

return removeOutGeneration(population);

}
```

El Algoritmo 10 genera por cada generación, tantos individuos como se haya indicado en el parámetro *individualsNumber*. Posteriormente, se resta 1 al tiempo de vida restante para toda la población y se eliminan todos los individuos cuyo *lifetime* sea igual a 0.

En primer lugar, para calcular el *lifetime* de cada individuo se calcula el *fitness* máximo, mínimo y la vida media de toda la población. Se utilizan tres posibles métodos para calcular el *lifetime*.

- Proporcional:

$$Lifetime = min\left(MinLT + \eta \frac{fitness[i]}{AvgFit}, MaxLT\right) \quad (7.3)$$

donde

$$\eta = \frac{1}{2} \cdot (MaxLT - MinLT)$$

AvgFit: El *fitness* de media de la población.

MaxLT: Es el número de generaciones máximas.

MinLT: Es el número de generaciones mínimas.

Fitness[i]: Es el valor del *fitness* del individuo *i*.

- Lineal:

$$LifeTime = MinLT + 2\eta \frac{fitness[i] - AbsFitMin}{AbsFitMax - AbsFitMin} \quad (7.4)$$

donde

AbsFitMax: Es el valor absoluto del *fitness* máximo.

AbsFitMin: Es el valor absoluto del *fitness* mínimo.

- Bilineal:

$$TimeLife = \begin{cases} MinLT + \eta \cdot \frac{fitness[i] - MinFit}{AvgFit - MinFit} \; si \; AvgFit \geq fitness[i] \\ \frac{1}{2} \cdot (MinLT + MaxLT) + \eta \cdot \frac{fitness[i] - AvgFit}{MaxFit - AvgFit} \; si \; AvgFit < fitness[i] \end{cases} \quad (7.5)$$

7.8.2 PRoFIGA.

Otro método desarrollado en esta investigación para el control de la población es *PRoFIGA*. Este método ajusta el tamaño de la población en función de las mejoras del fitness que se van obteniendo en un número determinado de generaciones (parámetro *V* de la configuración). El Algoritmo 11 define la implementación de *PRoFIGA*.

Algoritmo 11.- Método para control de la población PRoFIGA.

```
public ArrayList<Individual> selectionProfiga(...) {
double bestfitness = getBestFitness(offsprings);
if (bestfitness > evst.getLastFitness()) {
GROW_POPULATION_1
} else {
if (evst.getVcount() == v && bestfitness > evst.getVFitness() ) {
GROW_POPULATION_2
} else {
REDUCE_POPULATION
    }
}
```

GROW_POPULATION_1 y GROW_POPULATION_2 son mecanismos de crecimiento de la población en función del valor *X:*

$$X = increaseFactor \cdot evst.getGenerationsNumber() - evst.getGeneration()) \cdot \frac{maxFitness_{new} - maxFitness_{old}}{initMaxFitness} \quad (7.6)$$

donde

increaseFactor: es un parámetro definido en la configuración comprendido entre 0 y 1.

Evst.getGenerationNumber(): es el número máximo de generaciones.

Evst.getGeneration(): es el número de generación actual.

$maxFitness_{new}$*:* es el mejor *fitness* actual.

$maxFitness_{old}$*:* es el mejor *fitness* anterior.

initMaxFitness: es el mejor *fitness* inicial.

El procedimiento *REDUCE_POPULATION* realiza una reducción de la población según el coeficiente de reducción indicado por el usuario en el parámetro *reducción*. La población se aumenta o reduce en función de la variable *X,* siguiendo la fórmula (7.7):

$$X = tamaño\ de\ la\ población \cdot \%reducción \quad (7.7)$$

7.9 Función de evaluación.

La función de evaluación del *fitness* requiere de una gran capacidad computacional. Cada individuo de la población representa la TPC de la variable ***EA***. Para calcular la

idoneidad del individuo a la solución, se propagan las evidencias de un subconjunto de casos del corpus, por la BN y se modifica de forma dinámica la TPC de la variable ***EA*** por cada individuo; una vez se ha creado el individuo por recombinación y se ha mutado. Las probabilidades a posteriori para cada caso del subconjunto se guardan y posteriormente se utilizan para calcular el AUC.

El algoritmo realiza dos funciones, la inicialización y el cálculo del *fitness*. Tal y como se indicó anteriormente, para la inicialización de la población se utiliza un algoritmo memético, el cual parte de una solución suboptimal. Los parámetros calculados con el algoritmo de aprendizaje *Naive Bayes* se mutan de forma aleatoria utilizando la fórmula (7.8):

$$probabilidad_{edad,sexo,nivel\ educativo,deterioro\ cognitivo} = probabilidad_{edad,sexo,nivel\ educativo,deterioro\ cognitivo} \pm \sigma_x \quad (7.8)$$

donde

σ_x: es la desviación típica seleccionada de forma aleatoria, cuyo valor está comprendido entre un valor mínimo *thresholdDesvTip* y un valor máximo $\frac{n}{1.96}$,

n: es el número de dimensiones del vector.

Debido al gran coste computacional de la evaluación del *fitness* se selecciona de forma aleatoria un subconjunto de casos para la evaluación del *fitness*. Por otro lado, el algoritmo de estrategia evolutiva trabaja con vectores de números reales, tanto para la mutación como para la recombinación. Además, el algoritmo permite optimizar de forma simultánea las TPCs de cuantas variables se decidan. Un mismo individuo puede contener varias TPCs de distintas variables y esta representación de las TPCs no es compatible con las de la BNs, lo cual ha dado lugar a una función transformación de las TPCs a un vector de números reales y viceversa. El Algoritmo 12 define los pasos seguidos durante el proceso de inicialización de la población.

Algoritmo 12.- Inicialización de la población.

```
public void inicializacion {
calcular_TPC_DSD();  // Naive Bayes. Ver capítulo 5;
inicializacion_poblacion();
seleccion_aleatoria_de_casos();
configurar_algoritmo();
transformar_TPC_Individuo();
bNet = loadNet();

}
```

El algoritmo de estrategia evolutiva podría trabajar con cualquier función de *fitness*, es decir, se podría utilizar cualquier parámetro de rendimiento para evaluar la función *fitness*. Al igual que antes, es necesario transformar la representación del individuo, en TPCs compatibles con las BNs. Para cada uno de los casos seleccionados se propagan las evidencias por la BN y se almacena la probabilidad a posteriori de la variable de interés en una estructura de datos intermedia. Posteriormente, con los datos almacenados se realiza un análisis ROC y se calcula el AUC. Este proceso se repite por

cada individuo de la población y para todas las generaciones, lo que requerirá millones de operaciones matemáticas para optimizar una sola TPC. El Algoritmo 13 detalla el procedimiento para el cálculo del *fitness*.

Algoritmo 13.- Cálculo del fitness

```
public double computeFitness(Individuo individuo) {
  for (String idnode: nodes) {
    tpc = transformar_Individuo_TPC();
    aplicar_TPC(bNet, tpc, "DSD");
  }
  double auc = 0;
  listCatC   = buscarListaDeCasos();
  listCatC   = seleccionar_Casos_Aleatoriamente();
  for (Classificationcontinuous cat: listCatC) {
   inference = evidencesContinuousBean.propagate(net, ages, els, sexs, total.get(0),rel_DemP_AlzP,
rel_DemA_AlzA, cat);
   results = guardarResultados(inference);
  }
  auc = calcularAUC(results, "DSD");
  // Factores de penalización.
  if (initialization.getPenaltyMethod() != 0) {
   double w = 0;
   // Si penalización estática
   if (initialization.getPenaltyMethod() == 1 ) {
     w = initialization.getW();
   }
  // Si penalización feedback
   if (initialization.getPenaltyMethod() == 2 ) {
     w = EvolutionState.getInstance().getW();
   }
  // Si penalización autoadaptativa
  if (initialization.getPenaltyMethod() == 3 ) {
     w = individual.getW();
  }
  double penalty = 0;
  for (Double v: individual.getXs()) {
     if (v.doubleValue() < individual.getLoBound()
```

```
        || v.doubleValue() > individual.getUpBound()) {
        penalty += (double) 1 / individual.getXs().size();
    }
  }
  // Si se trata de un problema de maximización se tiene que multiplicar por -1.
  if (maximization == 1) {
    penalty *= -1;
  }
  auc = auc + (w * penalty);
}
if (maximization == 2) {
  auc = auc * -1;

}
auc = round(auc, initialization.getPrecision());
return auc;}
```

En la parte final del Algoritmo 13 se incluyen varios mecanismos de penalización del *fitness*, que se pueden clasificar en: a) estático, b) dinámico y c) auto adaptativo. Estos métodos de penalización se explicarán en la siguiente sección.

7.10 Funciones de penalización del fitness

Se han empleado tres métodos de penalización del *fitness* y todos estos métodos tienen en común que se penaliza su *fitness* cuando algunos de los componentes del vector $\bar{x}$ están fuera del rango (parámetros *loBound* y *upBound*). Los parámetros del algoritmo son:

- Sin penalización
- Penalización estática que utilizan la siguiente formula:

$$eval(\bar{x}) = f(\bar{x}) + W \times penalty(\bar{x}) \qquad (7.9)$$

donde

W: es un parámetro definido por el usuario y $penalty(\bar{x})$ toma el valor 0 o 1, en función de si tiene alguna penalización o no.

- Penalización dinámica. Emplea la fórmula siguiente:

$$W(t+1) = \begin{cases} \left(1/\beta_1\right) \times W(t) & si\ todos\ los\ elementos\ del\ vector\ \bar{x}\ son\ válidos \\ \beta_2 \times W(t) & si\ ningún\ elemento\ \bar{x}\ es\ válido \\ W(t) & en\ caso\ contrario \end{cases} \quad (7.10)$$

donde

β_1 y β_2: son parámetros definidos por el usuario

W(T): es el peso de la penalización va evolucionando de forma similar a la mutación y recombinación

- Penalización auto-adaptativa. El valor de *W* el peso de la penalización va evolucionando de forma similar a la mutación y recombinación.

PARTE

Experimentos

III

Capítulo

Evaluación de los modelos de BNs discretas y de las estrategias de discretización

8

En el capítulo 5 se ha descrito en detalle, el diseño del modelo cualitativo y cuantitativo de tres BNs discretas. En este capítulo se realizan dos experimentos cuyos objetivos son:

- El primer experimento analiza el rendimiento del método de diagnóstico, segmentado los conglomerados de atributos por edad y nivel educativo durante el proceso de discretización. El objetivo de este experimento es seleccionar la mejor estrategia de discretización.
- El segundo experimento contrasta los resultados de las distintas propuestas de BNs, respecto al diagnóstico dado por neurólogos, es decir, se decide qué modelo de BN es mejor (se ajusta mejor a las conclusiones de los neurólogos). En este experimento se utiliza la mejor estrategia de discretización obtenida en el primer experimento.

En todos los experimentos se propagan las evidencias, extraídas de los casos del corpus [1], por la BN y se guardan las probabilidades a posteriori de las variables de interés para su posterior análisis e interpretación.

La estructura del capítulo es la siguiente. En la sección 8.1 se realiza una introducción al capítulo, donde se detallan las métricas de rendimiento que se utilizan. En la sección 8.2 se mide la influencia de la edad y nivel educativo en la producción oral de rasgos semánticos, y su impacto en el proceso de inferencia. En la sección 8.3 se mide la eficacia del método de diagnóstico. En la sección 8.4 se realizan las discusiones del capítulo.

8.1 Introducción.

Los experimentos se han llevado a cabo con los casos del corpus lingüístico [1], el cual está constituido por 42 sujetos cognitivamente sanos y 39 sujetos enfermos de EA. En todos los experimentos se ha utilizado *Leave-One-Out Cross-Validation*, que es una variación de *n-fold Cross-Validation*. *Leave-One-Out Cross-Validation* divide el conjunto de casos en *n* particiones, de tal forma que se usan *(n-1)* particiones para el entrenamiento y la restante para la validación. El proceso se repite *n* veces, de tal forma que por cada iteración se utilizan distintas particiones para el entrenamiento y una partición para la validación. La TPC de la variable ***EA*** se ha optimizado con los algoritmos de estrategias evolutivas para los modelos de BN 2 y 3.

El Algoritmo 14 describe el procedimiento seguido para la realización de los experimentos.

Algoritmo 14.- Procedimiento seguido para realizar los experimentos.

```
Discretizar todos los atributos numéricos de todos los casos del corpus.
Repetir N veces {
        Seleccionar de forma aleatoria {
                (N-1) Casos del corpus lingüístico para el aprendizaje de la BN.
                1 Caso del corpus lingüístico para la validación de la BN.
        }
        Propagar las evidencias del caso seleccionado para validación por la BN.
        Almacenar las probabilidades a posteriori de las variables de interés
        y variables intermedias.
}
Analizar e interpretar los resultados.
Generar métricas de rendimiento.
```

Para comparar el rendimiento de los clasificadores utilizamos las curvas *Receiver Operating Characteristics* (ROC). Las curvas ROC es una técnica muy útil para visualizar, organizar y seleccionar clasificadores basado en su rendimiento. Las curvas ROC se utilizan comúnmente en la toma de decisiones médicas y en los últimos años, cada vez se utilizan más en proyectos de minería de datos. Además de tratarse de un método gráfico para medir el rendimiento, tiene propiedades que la hacen especialmente útil en los dominios donde existen una distribución de clases sesgadas y clasificaciones de costes de error desiguales. Es decir, las curvas ROC miden el rendimiento de los clasificadores sin tener en cuenta la distribución de clases o los costes de los errores.

Para la evaluación de los clasificadores se consideran cuatro posibles resultados dada una instancia. Si una instancia es positiva y es clasificada como positiva, se cuenta como *True Positive* (TP); si es clasificada como negativa, se cuenta como *False*

Negative (FN); si una instancia es negativa y es clasificada como negativa, se cuenta como *True Negative* (TN), y si es clasificado como positivo se cuenta como *False Positive* (FP).

Las ecuaciones (8.1) representan distintas métricas calculadas a partir de la matriz de confusión.

$$fp\ rate\ (FPR) = \frac{FP}{N} \qquad tp\ rate\ (TPR) = \frac{TP}{P} \quad (8.1)$$

donde

N: representa el número de casos negativos

P: representa el número de casos positivos.

$$precisión = \frac{TP}{TP + FP} \qquad exactitud = \frac{TP + TN}{P + N} \quad (8.2)$$

Otras métricas rendimiento utilizadas en los experimentos son *Mean Squared Error* y *Root Mean Squared Error*. Estás métricas miden la distancia entre la probabilidad predicha por la BN y su valor esperado. Siendo el valor esperado 1, para indicar que el sujeto padece la EA, y 0, para indicar que el sujeto esta cognitivamente sano. En ambas métricas cuanto más cercano sea su valor a 0 mejor es el resultado. Las fórmulas aplicadas para calcular dichas métricas son:

$$mean\ squared\ error = \frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n} \quad (8.3)$$

$$root\ mean\ squared\ error = \sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}} \quad (8.4)$$

donde

p_x: Es la probabilidad predicha por la BN para el caso *x*

a_x: Es el valor esperado para el caso *x*.

Las curvas ROC son gráficos bidimensionales en los cuales el TPR se representa en el eje *Y* y el FPR se representa en el eje *X*. Las curvas ROC representan el equilibrio relativo entre el beneficio de los TP y el coste de los FP. El *area under ROC* (AUC) es una métrica muy importante para nuestra investigación, ya que es la métrica que se ha utilizado en las estrategias evolutivas para medir el *fitness* de cada individuo (ver capítulo 7 para más detalle). El AUC mide la proporción del área por unidad cuadrada, su valor está comprendido entre 0 y 1. Por ejemplo, un clasificador aleatorio produce una línea diagonal que va desde los puntos (0,0) al (1,1) dando como resultado un AUC de 0,5. Un clasificador cuyo AUC esté por debajo a 0,5 es un clasificador no realista. Se puede decir que un punto en el espacio ROC es mejor que otro si está al noroeste (TPR

es mayor, FPR es menor, o ambos) del primero. Los clasificadores que aparecen en el lado izquierdo de la curva ROC, cerca del eje X, pueden ser considerados como conservativos, ya que suelen realizar clasificaciones positivas sólo si hay fuertes evidencias; estos clasificadores suelen cometer pocos errores FP, pero también suelen tener un TPR bajo. Los clasificadores situados en la parte derecha de la curva ROC pueden ser considerados como más liberales, infieren clasificaciones positivas con evidencias muy débiles. La línea diagonal y = x representa una estrategia aleatoria para adivinar la clase, es decir, traza una línea diagonal de (0,0) a (1,1). Los puntos de la curva ROC más representativos son el punto (0,0), que representa la estrategia por la cual el clasificador nunca clasifica los casos como positivos, es decir, no hay FP pero tampoco hay TP. La estrategia opuesta es representada en el punto (1,1). El punto (0,1) representa la clasificación perfecta (ver detalles en [50]). El AUC está estrechamente relacionado con el coeficiente *Gini* [51], el cual es dos veces el área entre la diagonal y la curva *ROC*. Hand y Till [52] demostraron la relación entre el AUC y el coeficiente *Gini* con la siguiente fórmula ($Gini + 1 = 2 * AUC$).

Como se ha indicado anteriormente, utilizamos *Leave-One-Out Cross-Validation* para evitar el sobreajuste. El método que se ha utilizado en esta investigación para generar la curva ROC es el más sencillo, se recogen las probabilidades predichas de todas las validaciones junto con la etiqueta de la clase a la que pertenece la instancia y se genera una lista única de probabilidades predichas. Las probabilidades generadas por la BN son comparables de una iteración a otra, es decir, se encuentra en la misma escala y además, el conjunto de entrenamiento es representativo de la muestra.

8.2 Influencia de la edad y nivel educativo en la producción oral de rasgos semánticos.

Como se ha indicado anteriormente, se han implementado varias estrategias para la discretización de los atributos numéricos. El objetivo del experimento es analizar la influencia de la edad y el nivel educativo en la producción oral de rasgos semánticos. Tal y como se indicó en el capítulo 5, establecer enlaces causales entre estos factores y todas las variables del corpus, obliga a crear más de 130 enlaces causales adicionales, aumentando la complejidad del aprendizaje automático de las TPCs. Para modelar esta situación se ha creado una jerarquía de clúster multinivel, la cual permite crear dos clúster por cada segmento de edad o nivel educativo, categoría semántica y rasgo semántico.

En este experimento se ha realizado una comparativa entre las tres estrategias de discretización:

- La primera estrategia crea dos clúster por cada categoría y rasgo semántico.
- La segunda estrategia crea dos clúster por cada segmento de edad, categoría semántica y rasgo semántico.

- La tercera estrategia crea dos clúster por cada nivel educativo, categoría semántica y rasgo semántico.

En la Tabla 20 se muestran las distribuciones de casos por nivel de la enfermedad. Se ha utilizado *leave-one-out cross validation* para evitar el sobreajuste. Las probabilidades predichas durante la validación del modelo de diagnóstico se han almacenado y se han utilizado posteriormente para generar la curva ROC. En el apéndice B se puede consultar este mismo experimento realizado con la BN con razonamiento deductivo.

Nivel Enfermedad	Nº Casos
Leve	33
Moderado	6
Sanos	42

Tabla 20.- Estratificación de casos por estado cognitivo

En este experimento se analizan las probabilidades a posteriori de la variable de interés ***DSD***, en lugar de la variable ***EA.*** La razón de utilizar la variable ***DSD*** y no la variable ***EA***, es que en la variable de interés ***EA*** existen los enlaces causales: ***edad→EA***, ***sexo→EA*** y ***nivel educativo→EA,*** y por tanto, el proceso de discretización no tiene efecto alguno sobre esta variable. Sin embargo, en este experimento analizamos la influencia de estos factores en el proceso de discretización y no como una influencia informativa.

Clusterización por categoría semántica y rasgo.

Para este test se crea un clúster por cada categoría semántica y rasgo, es decir, se crean 132 clústeres (6 categorías semánticas * 11 rasgos * 2 clúster para cada uno de los estados de la variable).

Clusterización por edad, categorías semánticas y rasgos.

Para este test se crea un clúster por cada segmento de edad, categorías semánticas y rasgos, es decir, se crean 792 clústeres (6 segmentos de edad * 6 categorías semánticas * 11 rasgos * 2 clúster para cada uno de los estados de la variable rasgo).

Clusterización por nivel educativo, categorías semánticas y rasgos.

Para este test se crea un clúster por cada nivel educativo, categorías semánticas y rasgos, es decir, se crean 396 clústeres (3 niveles educativo * 6 categorías semánticas * 11 rasgos * 2 clúster para cada uno de los estados de la variable rasgo).

En la Figura 23 se representan tres curvas ROC correspondientes a las tres estrategias de discretización por análisis de clúster llevada a cabo en este experimento. Al igual que en el experimento anterior, en el eje de abscisa se representa el *TP rate* y en el eje de ordenadas se representa el *FP rate*. Este experimento se ha realizado con el modelo 3 de BN.

Cabe destacar que aunque se representan las tres curvas en la misma gráfica, cada curva se genera con un clasificador diferente, ya que la estrategia de discretización utilizada causa que los modelos cuantitativos de cada BN varíen significativamente. Estas variaciones del modelo cuantitativo se deben a que los centroides obtenidos por análisis

de clúster, varían al segmentar los conglomerados de atributos por edad o nivel educativo y consecuentemente da lugar a que los resultados de la discretización varíen. Esto a su vez, da lugar a que los resultados de las probabilidades condicionales, calculadas por el algoritmo de aprendizaje automático, también varíen.

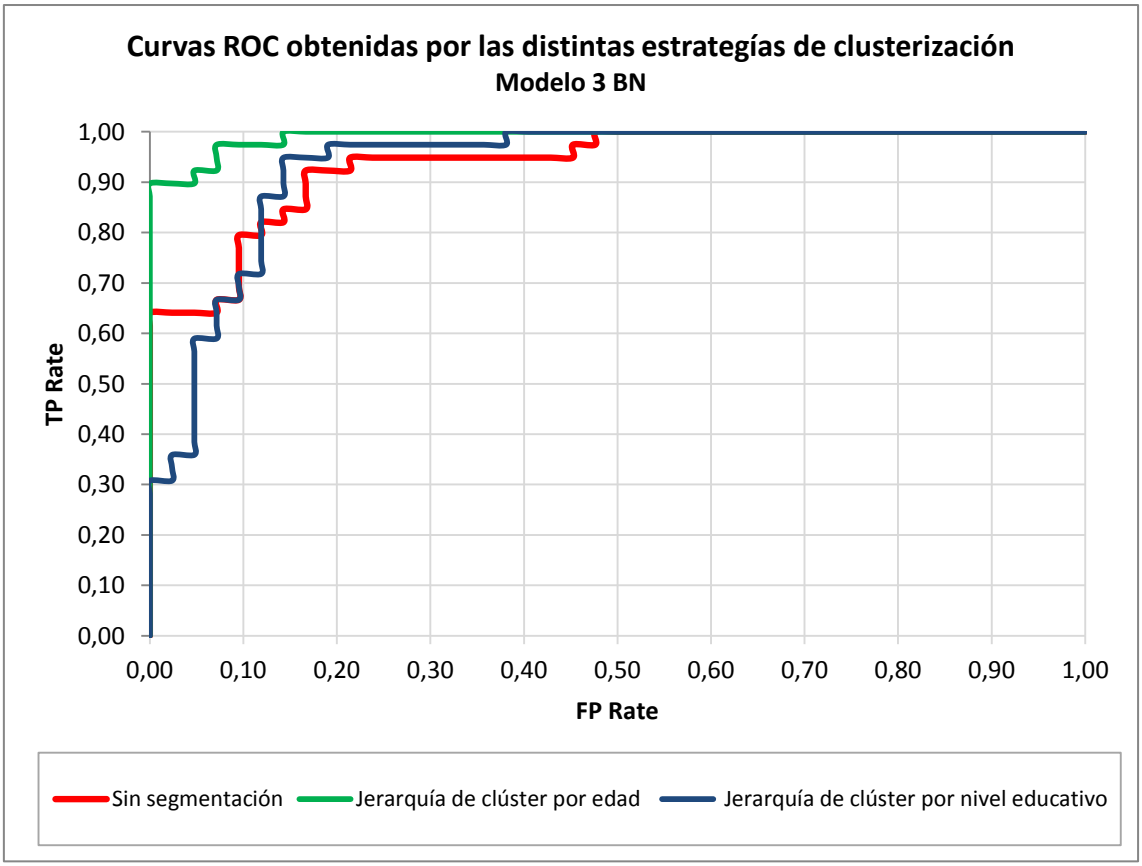


Figura 23.- Curvas ROC obtenidas por el modelo 3 y las distintas estrategias de discretización.

En la Figura 23 se puede observar que la BN que mejores resultados proporciona, es aquella que segmenta los conglomerados por edad. Sin embargo, al segmentar la producción lingüística por nivel educativo, la mejora del rendimiento es más sutil.

En la Tabla 21 se detallan las métricas de rendimiento obtenidas a partir de las distintas estrategias de discretización por análisis de clúster. En esta tabla se pone de manifiesto que al crear una jerarquía de clúster por edad, se consiguen mejores resultados.

Tabla 21.- Métricas de rendimiento obtenidas para las distintas estrategias de discretización por análisis de clúster.

	Sin segmentación	Por edad	Por nivel educativo
True Positive (TP)	36	35	38
True Negative (TN)	35	41	33
False Positive (FP)	7	1	9
False Negative (FN)	3	4	1
TP rate	0,9231	0,8974	0,9744
FP rate	0,1667	0,0238	0,2143

	Sin segmentación	Por edad	Por nivel educativo
Precisión	0,8372	0,9722	0,8085
Exactitud	0,8765	0,9383	0,8765
Mean Squared Error	0,1067	0,0538	0,1063
Root Mean Squared Error	0,3267	0,232	0,326
AUC	0,9371	0,9915	0,9261

A continuación se analizan los resultados obtenidos midiendo la eficacia en el diagnóstico, en relación al diagnóstico proporcionado por los neurólogos siguiendo los criterios NINCDS-ADRDA, es decir, mide la eficacia en relación a su valor esperado. Para ello se segmenta las probabilidades a posteriori de la variable de interés EA en los tramos: (0-25%, 25-50%, 50-75% y 75-100%). Posteriormente, se hace un recuento de casos por cada tramo de probabilidad inferida y se normaliza. El recuento se hace por un lado, para la muestra de sujetos sanos y la variable $\boldsymbol{EA_{ausente}}$, y por otro lado, para la muestra sujetos enfermos de EA y la variable $\boldsymbol{EA_{presente}}$.

En la Figura 24 se comparan las distintas estrategias de discretización utilizando el modelo 3, con la muestra de sujetos sanos. Se puede deducir como segmentado los conglomerados de atributos por edad o nivel educativo, la BN infiere una probabilidad comprendida entre el 80%-100% de que el sujeto esté cognitivamente sano para un mayor número de casos. Sin embargo, sin la segmentación de los conglomerados de atributos por edad o nivel educativo, la mayor parte de los casos se infiere una probabilidad comprendida entre el 60%-80% de que el sujeto este sano, es decir, la distancia a su valor esperado es mayor.

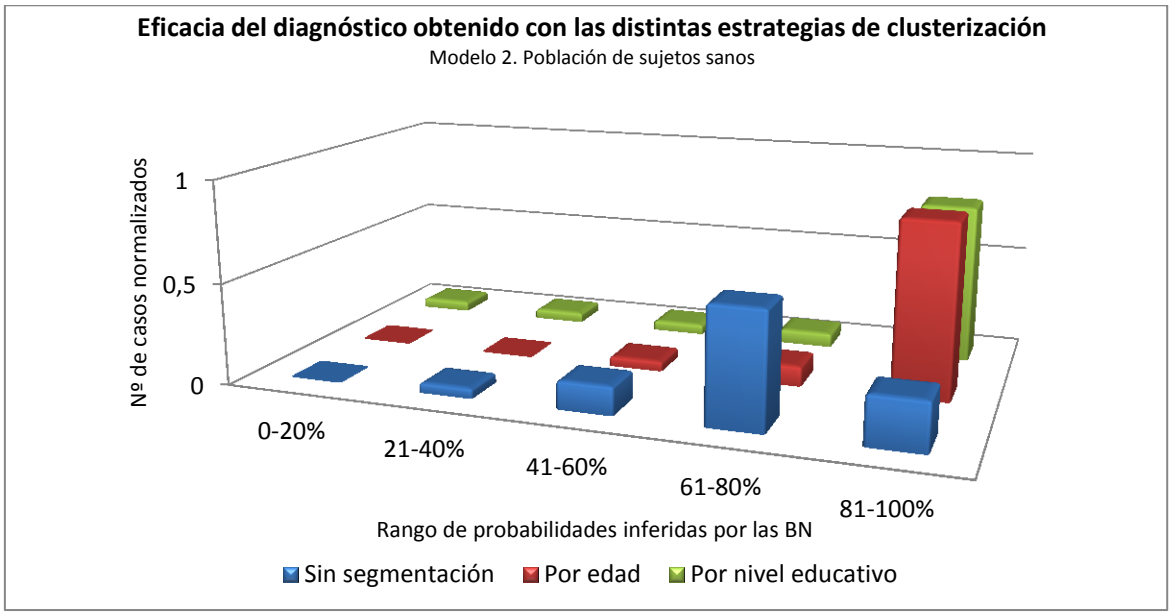


Figura 24.- Eficacia del diagnóstico obtenido con las distintas estrategias de discretización. Modelo 3 BN para la muestra de sujetos sanos.

La Figura 25 es similar a la anterior pero para la muestra de sujetos enfermos de EA. Al igual en la gráfica anterior, se demuestra que la segmentación de atributos por edad y nivel educativo, ocasiona que la BN infiera una probabilidad comprendida entre el 80% y 100% de padecer la EA para un mayor número de casos. Es decir, las probabilidades a posteriori están más cercanas a su valor esperado.

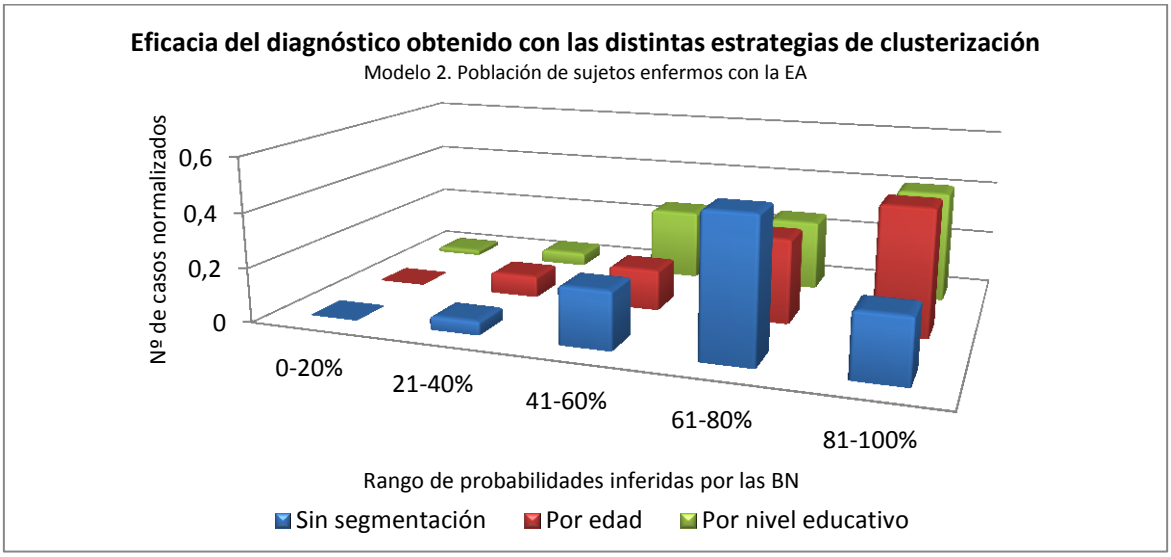


Figura 25.- Eficacia del diagnóstico obtenido con las distintas estrategias de discretización. Modelo 3 BN para la muestra de sujetos enfermos de EA.

8.3 Eficacia del método de diagnóstico.

El propósito de este experimento es comprobar la eficacia del método de diagnóstico. Como el número de casos es muy reducido, se usa la misma muestra del corpus tanto para el entrenamiento como para la validación, para ello se usa *Leave-One-Out Cross-Validation* escogiendo 79 casos para el entrenamiento y 2 casos para la validación (sano y enfermo de EA) en cada iteración. El proceso se repite 42 veces, de tal forma que por cada iteración se selección de forma aleatoria distintos casos para el entrenamiento y para la validación. Dada la dificultad que hemos tenido para adquirir nuevos casos, se ha optado por Leave-One-Out Cross-Validation porque nos ha permitido evaluar el método de diagnóstico, con un error muy bajo y con un número reducido de casos. Las probabilidades predichas para el caso de validación, se almacenan en una base de datos y posteriormente se utilizan para construir las curvas ROC, siguiendo el procedimiento descrito en la sección anterior. Para realizar este experimento, la TPC de la variable ***EA*** para las BNs con razonamiento abductivo, se han optimizado con algoritmos de estrategias evolutivas. En el apéndice B, se puede consultar el resultado de este mismo experimento, pero sin optimizar la TPC de esta variable.

En este experimento se generan tres curvas ROC. En la Figura 26 se representa la curva ROC generada a partir del modelo 1 de BN (ver sección 5.3) que utiliza inferencia por razonamiento deductivo. Para este experimento se segmenta la producción oral de rasgos semánticos por tramos de edad.

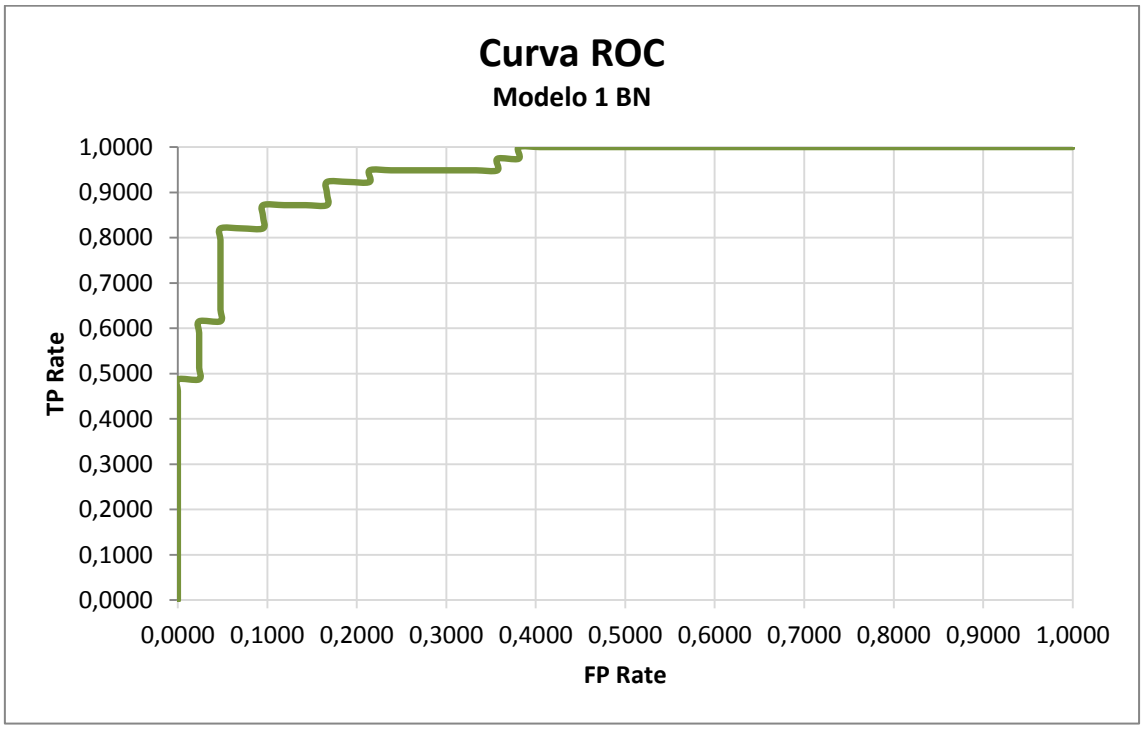


Figura 26.- Curva ROC obtenida a partir del modelo 1 de BN con segmentación de atributos por edad.

Se puede observar en la curva ROC de la Figura 26 que el clasificador tiene un rendimiento muy bueno. A partir de las probabilidades a posteriori se busca un *threshold* que permita clasificar correctamente el mayor número posible de casos. A partir de este *threshold* se construye la matriz de confusión y se calculan otras métricas de rendimiento, además del AUC.

La segunda curva ROC se genera con el modelo 2 de BN (ver sección 5.4). Esta BN utiliza razonamiento abductivo y no tiene los enlaces causales ***EA→SV*** y ***EA→SNV***. Para este experimento se utiliza una TPC de la variable ***EA*** optimizada, tal y como se indica en el capítulo 10.

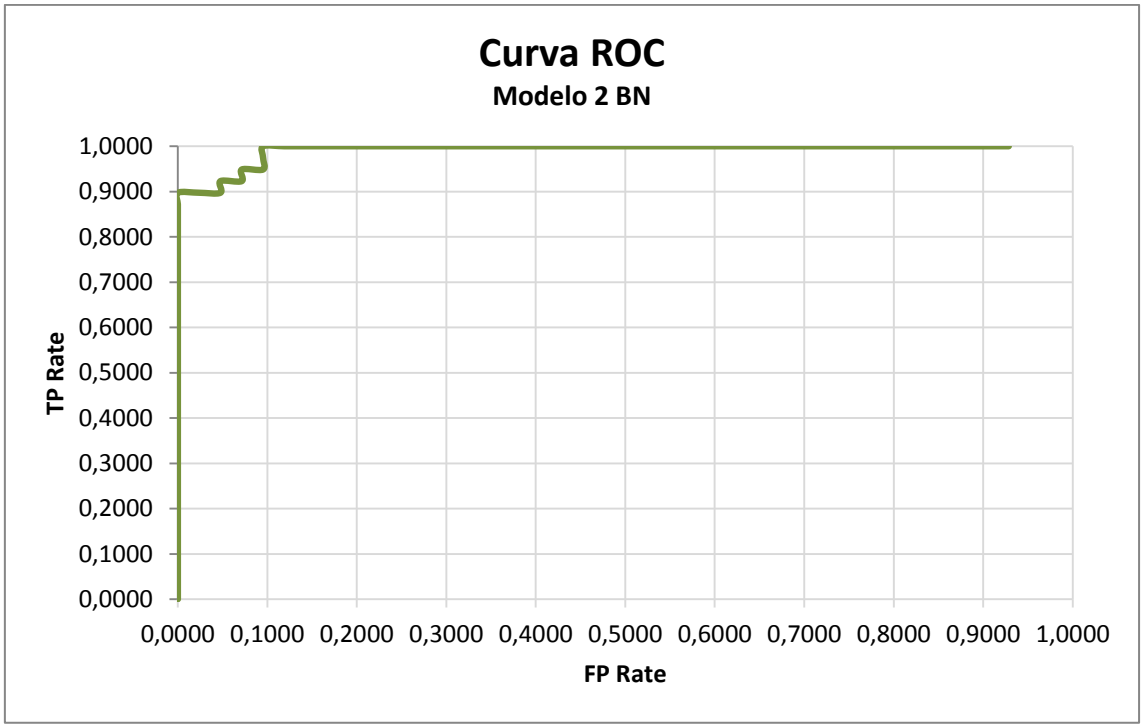


Figura 27.- Curva ROC obtenida a partir del modelo 2 de BN con segmentación de atributos por edad.

La tercera curva ROC se genera con el modelo 3 de BN (ver sección 5.5) que utiliza inferencia se realiza por razonamiento abductivo. En esta BN se crean los enlaces causales ***EA→SV*** y ***EA→SNV***. En la Figura 28 se muestra la curva ROC esta BN. Para este experimento se utiliza una TPC de la variable ***EA*** optimizada, tal y como se indica en el capítulo 10.

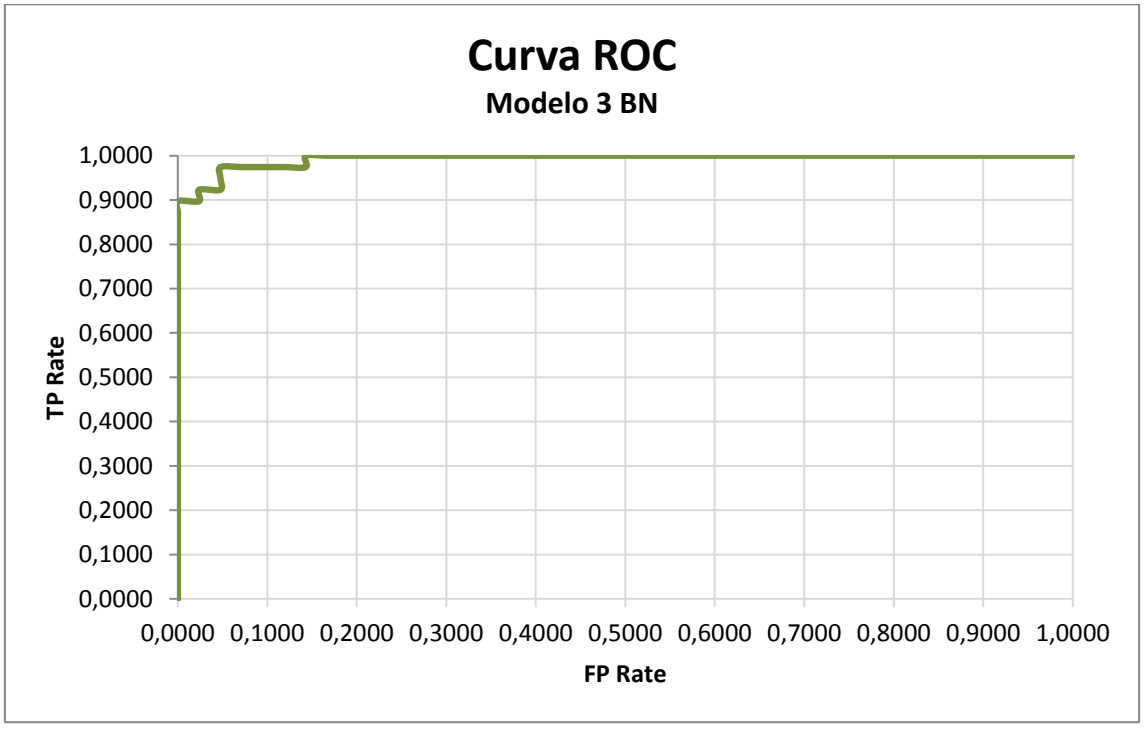


Figura 28.- Curva ROC obtenida a partir del modelo 3 de BN con segmentación de atributos por edad.

De las curvas ROC de las Figura 26, 24 y 25, se pueden determinar que todos los modelos de BN tienen un rendimiento excelente, siendo la BN del modelo 3 la que produce mejor resultado. Esta BN tiene un rendimiento superior a la BN del modelo 2; a continuación se analizará con más detalle está diferencia en el rendimiento.

En la Tabla 22 se analizan otras métricas de rendimiento de este experimento. Se puede comprobar en esta tabla que el rendimiento de la BN del modelo 3 supera en la mayoría de las métricas, los modelos 1 y 2. Cabe destacar que el *threshold* no se trata de una métrica, sino de un umbral de probabilidad a partir del cual se clasifican los casos como positivos o negativos.

Tabla 22.- Métricas de rendimiento del experimento 1 para los modelos 1, 2 y 3 de BN.

	Modelo 1	Modelo 2	Modelo 3
True Positive (TP)	34	39	38
True Negative (TN)	38	37	39
False Positive (FP)	4	5	3
False Negative (FN)	5	0	1
TP rate	0,8718	1	0,9744
FP rate	0,0952	0,119	0,0714
Precisión	0,8947	0,8864	0,9268
Exactitud	0,8889	0,9383	0,9506
Mean Squared Error	0,1223	0,0867	0,0598
Root Mean Squared Error	0,3497	0,2944	0,2446
AUC	0,9493	0,9921	0,9933
Threshold	0,506785	0,259378	0,357941

En la Tabla 23 se analiza las métricas de rendimiento de las variables intermedias obtenidas en el experimento. En la Tabla 23 se muestra como con algunas variables intermedias se consiguen mejores métricas de rendimiento que con otras.

Tabla 23.- Comparativa de las métricas de rendimiento obtenidas con las variables intermedias en la BN discreta con razonamiento abductivo.

	BN DISCRETA RAZONAMIENTO ABDUCTIVO		
Variables	**AUC**	**FP RATE**	**TP RATE**
DS_{SV}	0,9823	0,0714	0,8974
DS_{SNV}	0,9792	0,0952	0,9744
$DS_{Manzana}$	0,9621	0,0714	0,9231
DS_{Perro}	0,8974	0,1667	0,7949
DS_{Pino}	0,9438	0,1429	0,8974
DS_{Coche}	0,9536	0,0476	0,8462
DS_{Silla}	0,9640	0,0476	0,8718
$DS_{Pantalón}$	0,9475	0,0952	0,8718

En las Figura 29 y 27 se analizan las probabilidades a posteriori de la variable de interés ***EA,*** para todos los casos de la muestra, inferidas por cada una de las BNs discretas. En esta gráfica cada punto representa la probabilidad a posteriori de la variable de interés ***EA*** para los estados *ausente* y *presente*, las cuales se representan en la misma figura pero en curvas distintas. La muestra de sujetos sanos se representa en la Figura 29 y la muestra de sujetos enfermos de EA se representa en la Figura 30. La probabilidad de estar sano se representa por un rombo de color verde, mientras la probabilidad de padecer un deterioro semántico se representa por un círculo rojo. En las curvas, por claridad, se dibuja una línea de trazado entre las probabilidades de $\boldsymbol{EA_{presente}}$ y $\boldsymbol{EA_{ausente}}$ respectivamente. El número de caso se representa en el eje de abscisa y la probabilidad de estar cognitivamente sano, inferido por la BN, se representa en el eje de ordenadas.

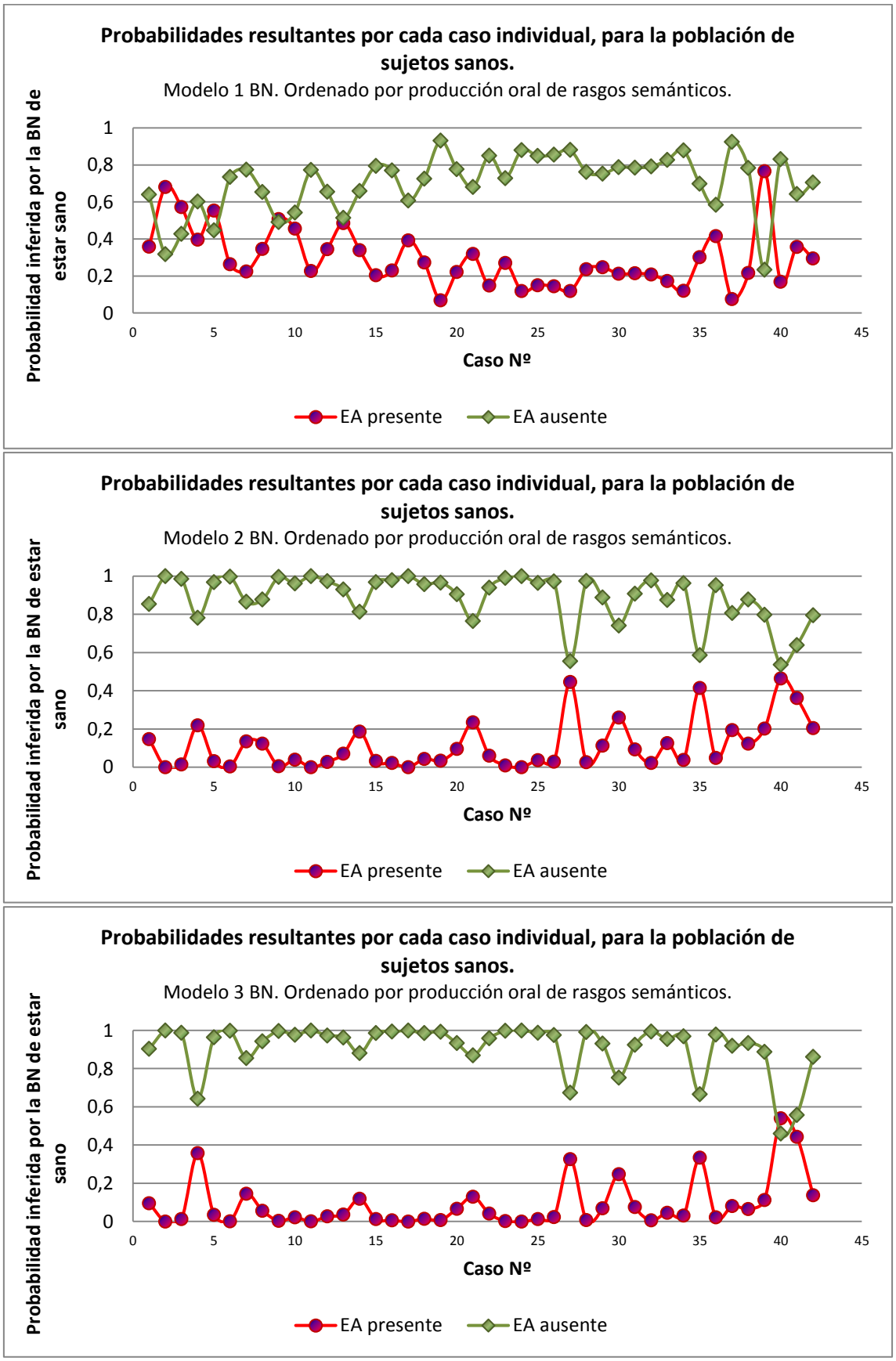


Figura 29.- Probabilidades resultantes por cada caso individual, para la muestra de sujetos cognitivamente sanos a) modelo 1 b) modelo 2 c) modelo 3

En la Figura 30 se muestra como las probabilidades inferidas por el modelo 3 de BN están más cercanas a su valor esperado. Cabe destacar que para este experimento los casos se han ordenado por producción oral de rasgos semánticos en forma descendente, sin embargo no se aprecia una correspondencia lineal entre las probabilidades a posteriori y la producción oral de rasgos semánticos, por consiguiente podemos afirmar que las BNs son capaces de encontrar, en este experimento, relaciones complejas de interacción entre la EA y la producción oral de rasgos semánticos.

De la misma manera, en la Figura 30 se representan las probabilidades inferidas por las BNs para la muestra de sujetos enfermos de EA.

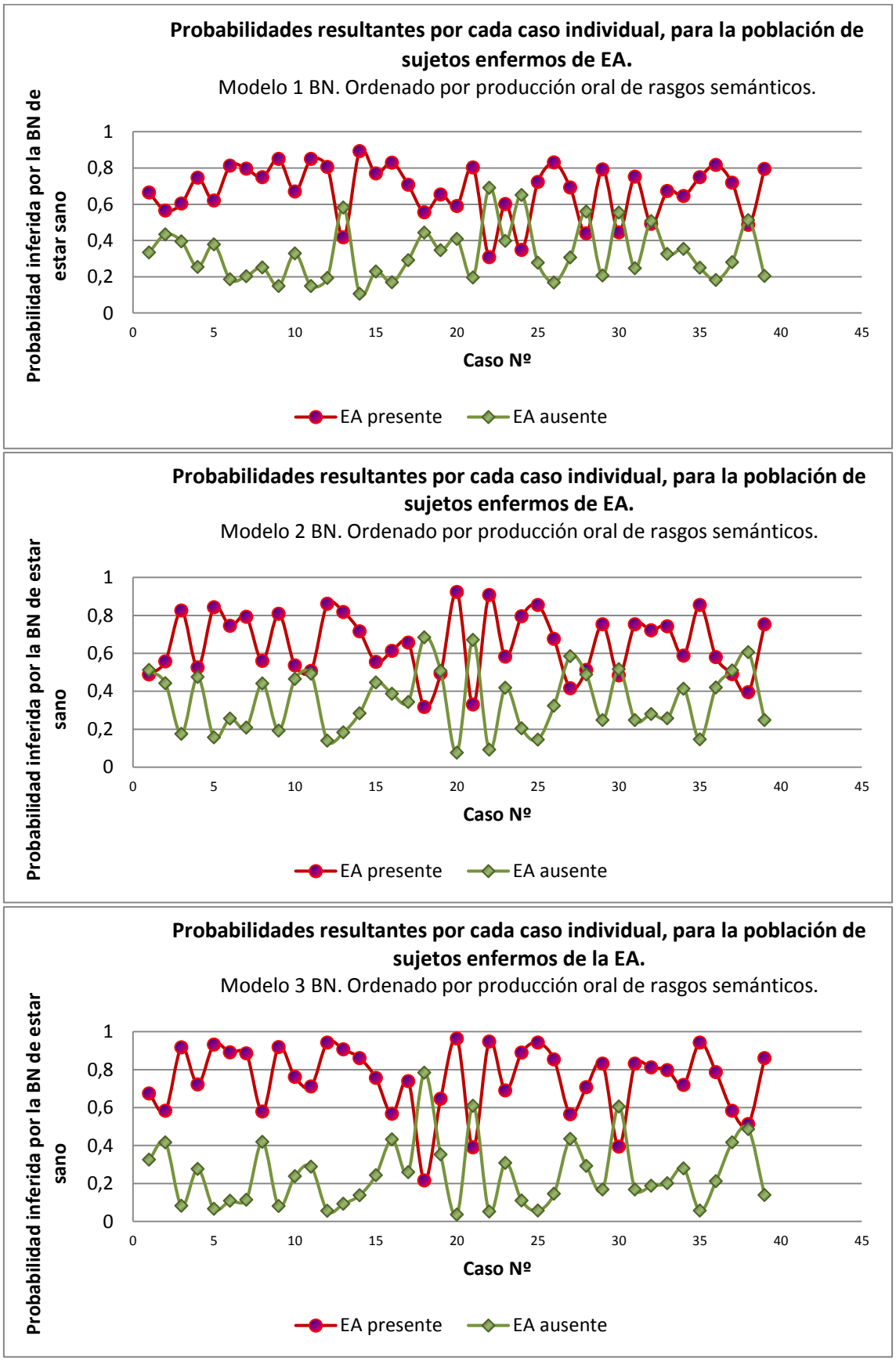


Figura 30.- Probabilidades resultantes por cada caso individual para la muestra de sujetos enfermos de EA a) modelo 1 b) modelo 2 c) modelo 3

En la Tabla 24 se muestran las probabilidades a posteriori de la variable de interés EA para todos los casos de la muestra inferidas por el modelo 3 de la BN. La columna *Attributes count,* representa la producción oral de rasgos semánticos de todo el test oral; la columna *Class True,* representa el diagnóstico proporcionado por neurólogos; la columna *Class Hyp,* es diagnóstico predicho, y *Score,* representa la probabilidad a posteriori de la variable $\boldsymbol{EA_{presente}}$. Cabe destacar que no se encuentra una relación lineal entre la producción oral de rasgos semánticos y la probabilidad a posteriori de la variable $\boldsymbol{EA_{presente}}$.

Tabla 24.- Probabilidades inferidas, clasificación de todas las instancias y score de la curva ROC.

Attributes Count	Class True	Class Hyp	Score	Attributes Count	Class True	Class Hyp	Score	Attributes Count	Class True	Class Hyp	Score
102	A	A	0,00	**95**	A	A	0,07	**27**	P	P	0,71
111	A	A	0,00	**87**	A	A	0,08	**18**	P	P	0,71
62	A	A	0,00	**102**	A	A	0,08	**16**	P	P	0,72
58	A	A	0,00	**43**	A	A	0,10	**35**	P	P	0,72
78	A	A	0,00	**46**	A	A	0,11	**43**	P	P	0,74
81	A	A	0,00	**96**	A	A	0,12	**28**	P	P	0,76
87	A	A	0,00	**103**	A	A	0,13	**10**	P	P	0,76
87	A	A	0,01	**60**	A	A	0,14	**37**	P	P	0,79
120	A	A	0,01	**69**	A	A	0,14	**37**	P	P	0,80
127	A	A	0,01	**76**	P	A	0,22	**22**	P	P	0,81
120	A	A	0,01	**72**	A	A	0,25	**34**	P	P	0,83
70	A	A	0,01	**89**	A	A	0,33	**34**	P	P	0,83
104	A	A	0,01	**71**	A	A	0,33	**0**	P	P	0,85
108	A	A	0,01	**44**	A	A	0,36	**42**	P	P	0,86
104	A	A	0,01	**70**	P	P	0,39	**30**	P	P	0,86
101	A	A	0,02	**34**	P	P	0,39	**32**	P	P	0,89
85	A	A	0,02	**44**	A	P	0,44	**11**	P	P	0,89
84	A	A	0,02	**53**	P	P	0,51	**2**	P	P	0,89
76	A	A	0,03	**69**	A	P	0,54	**15**	P	P	0,91
81	A	A	0,03	**50**	P	P	0,57	**32**	P	P	0,92
85	A	A	0,04	**60**	P	P	0,57	**20**	P	P	0,92
87	A	A	0,04	**81**	P	P	0,58	**23**	P	P	0,93
85	A	A	0,04	**52**	P	P	0,58	**12**	P	P	0,94
85	A	A	0,05	**66**	P	P	0,58	**23**	P	P	0,94
45	A	A	0,06	**34**	P	P	0,65	**16**	P	P	0,94
46	A	A	0,07	**23**	P	P	0,67	**45**	P	P	0,95
75	A	A	0,07	**56**	P	P	0,69	**56**	P	P	0,96

8.4 Discusiones.

Los resultados que se consiguen en estos experimentos son excelentes, destacando los que se obtiene en el experimento 1 con el modelo 3 de BN. Esta BN utiliza razonamiento abductivo y modela el deterioro semántico diferencial que pueden presentar algunos enfermos de EA cuando la enfermedad es incipiente. En el experimento de la sección 8.3, en las Figura 27 y 28, o en la Tabla 22; se puede comprobar una mejora sutil en el rendimiento en la BN que modela el deterioro semántico diferencial. Analicemos un caso concreto para ver esta mejora, por ejemplo, el caso nº 27 del modelo 3 se clasifica correctamente, pero sin embargo, en el modelo 2 se podría considerar como un falso negativo si se establece un *threshold* menos conservador. El caso nº 27 se corresponde con una persona de 73 años de edad, con estudios universitarios y que fue diagnosticado por neurólogos como EA leve. Esta diferencia en el rendimiento como consecuencia de los enlaces causales ***EA→SV*** y ***EA→SNV***, es especialmente importante para esta tesis doctoral, ya que demuestra que este síntoma de la EA puede permitir diagnosticar la enfermedad en fase leve.

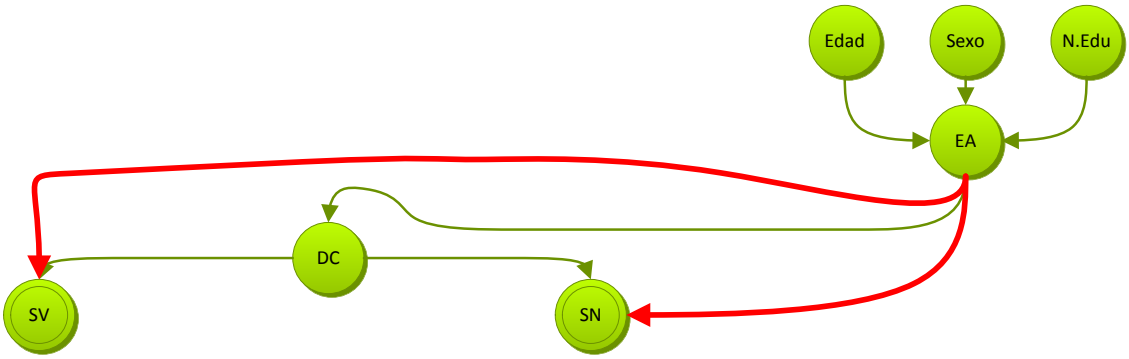


Figura 31.- Fragmento BN 3 que representa el deterioro selectivo entre los dominios SV y SNV.

En la Figura 32 se analiza un fragmento de las probabilidades a posteriori inferidas con Elvira [33]. El grosor de los enlaces causales representa el peso en el diagnóstico, durante el proceso de inferencia, de estos enlaces causales. Se pueden apreciar como los enlaces ***EA→SV*** y ***EA→SNV*** influyen de forma significativa en la probabilidad a posteriori de la variable ***EA***. Además, se puede resaltar que este paciente en concreto parece mostrar un deterioro semántico diferencial entre los dominios semánticos SV y SNV.

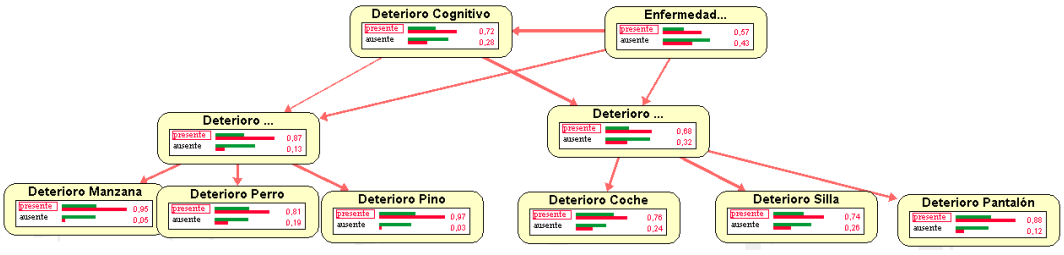


Figura 32.- Probabilidades a posteriori y transmisión de influencia por los enlaces para el modelo 3.

La Figura 33 es similar a la anterior, pero se ha obtenido con el modelo 2 de BN. Se observa como las probabilidades a posteriori de todas las variables intermedias son similares a la inferida por el modelo 3, sin embargo la probabilidad a posteriori de la variable $\boldsymbol{EA}_{\text{presente}}$ es sólo de un 41%, frente a un 57% que se obtiene con el modelo 3.

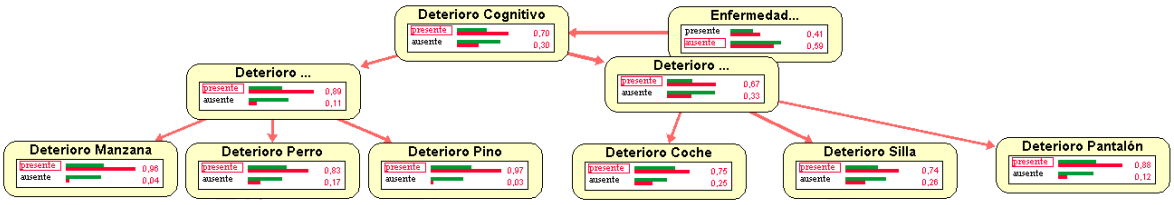


Figura 33.- Probabilidades a posteriori y transmisión de influencia por los enlaces para el modelo 2.

De este experimento se puede determinar que el rendimiento de esta BN es excelente, con una precisión y exactitud por encima del 90%. Obviamente faltaría determinar la especificidad del método de diagnóstico con enfermos de otras ENs no-EA.

En el experimento 2 son destacables las mejoras en el rendimiento de las BNs que se consiguen con las distintas estrategias de discretización, fundamentalmente la estrategia en la que se segmentan los recuentos de la producción oral de rasgos semánticos por tramos de edad. Con esta mejora innovadora se consigue mejorar el rendimiento de las BNs en aproximadamente un 10%. Sin embargo, la mejora conseguida al segmentar los recuentos de la producción oral de rasgos semánticos por nivel educativo, es más sutil. A pesar de esta mejora sutil, creemos que la estrategia es buena y nos hace plantear la pregunta ¿podrían existir otros factores que influyan en la producción lingüística de atributos? Probablemente sí, pero no se han podido estudiar otros factores por la dificultad en la obtención de nuevos casos. Es posible que el nivel de ocupación intelectual sea un factor influyente en la producción oral de rasgos semánticos. Una posible explicación de que sea el nivel de ocupación y no el nivel educativo el más influyente en la producción oral de rasgos semánticos, es porque las personas con más edad normalmente suelen tener un menor nivel de ocupación intelectual debido a que podrían llevar más tiempo jubilado. Esta hipótesis podría ser una explicación a la dispersión existente en los datos que disponemos, ya que evidentemente no todos los jubilados mantienen un nivel de ocupación intelectual bajo. Lo mismo puede ocurrir con el nivel educativo, las personas con un nivel educativo medio/bajo, pero con un nivel de ocupación alto ¿podrían producir más rasgos semánticos que una persona con un nivel educativo alto pero con un nivel de ocupación muy bajo? Dentro de las posibilidades de nuestra investigación se han analizado estos factores y podemos afirmar que se mejora el rendimiento de las BNs, si se tienen en cuenta esos factores durante el proceso de discretización.

Capítulo

Evaluación de los modelos de BNs híbridas

9

En el capítulo 6 se detallaron las técnicas de modelado e inferencia utilizadas con la BN híbrida. En este capítulo se analiza el rendimiento de esta BN y para ello se realizan tres experimentos. Se omite la introducción porque la metodología seguida en este capítulo es similar a la del capítulo anterior, es decir, se utilizan los mismos casos, las mismas métricas de rendimiento y *leave-one-out cross-validation* para evitar el sobreajuste.

Los experimentos de este capítulo persiguen los siguientes objetivos:

- El primer experimento pone de relieve la importancia de la segmentación de las definiciones orales en los once bloques conceptuales que propone el corpus de Peraita y Grasso [1]. Para ello se crea una BN Híbrida simplificada, cuya estructura se representa en la Figura 34.
- El segundo experimento se realiza con una CLG BN utilizando todas las variables del corpus, todas las variables intermedias o latentes y todas las variables de interés. Este experimento mide el rendimiento de la CLG BN utilizando distintas ecuaciones de regresión lineal durante la inferencia de las variables latentes o intermedias.
- El tercer experimento es similar al segundo experimento en cuanto a la utilización de los coeficientes de predicción, pero en lugar de utilizar una CLG BN se utiliza una BN híbrida con un algoritmo de inferencia aproximada.

Para el aprendizaje de los parámetros de la BN híbrida sólo se ha utilizado la muestra de sujetos cognitivamente sanos.

9.1 Importancia de la segmentación de atributos en once bloques conceptuales.

Cuando la EA se encuentra en fase avanzada, el daño cerebral es tan omnipresente que el DS se hace evidente. Sin embargo, cuando la enfermedad es incipiente, el DS puede presentarse de forma selectiva, siendo esta es la base para el método de diagnóstico que presentamos en esta investigación.

En este experimento se utiliza la BN de la Figura 34, la cual no tiene en cuenta la segmentación de las definiciones orales en los once bloques conceptuales que propone el corpus de Peraita y Grasso [1].

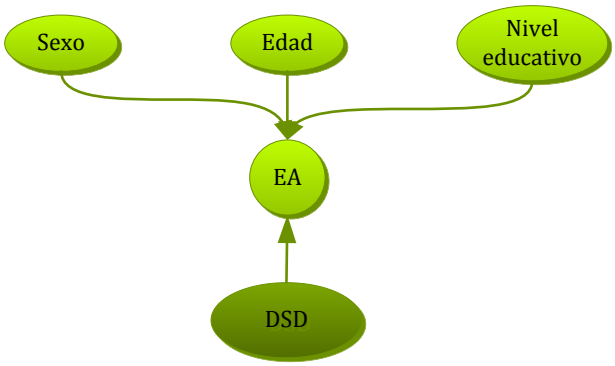


Figura 34.- BN híbrida reducida

En la BN de la Figura 34 se hace un recuento de la producción oral de rasgos semánticos de todas las categorías semánticas. Este modelo funciona bien para los enfermos de EA en estado avanzado, pero cuando la enfermedad se encuentra en sus primeras fases el clasificador comete más errores en el diagnóstico.

En la Figura 35 se representan tres curvas ROC: la curva de color rojo, es la curva obtenida a partir de la BN híbrida reducida; la curva de color verde, se ha obtenido con una CLG BN que utiliza todas las variables del corpus (variables predictoras) y variables intermedias (variables latentes); y la curva de color azul, se ha generado con una BN en la que se emplea un método de inferencia aproximada para las variables intermedias o latentes. En la CLG BN, se ha utilizado el coeficiente de correlación de Pearson, y en la BN híbrida con inferencia aproximada, se ha utilizado la ganancia de información, para la inferencia de las variables latentes.

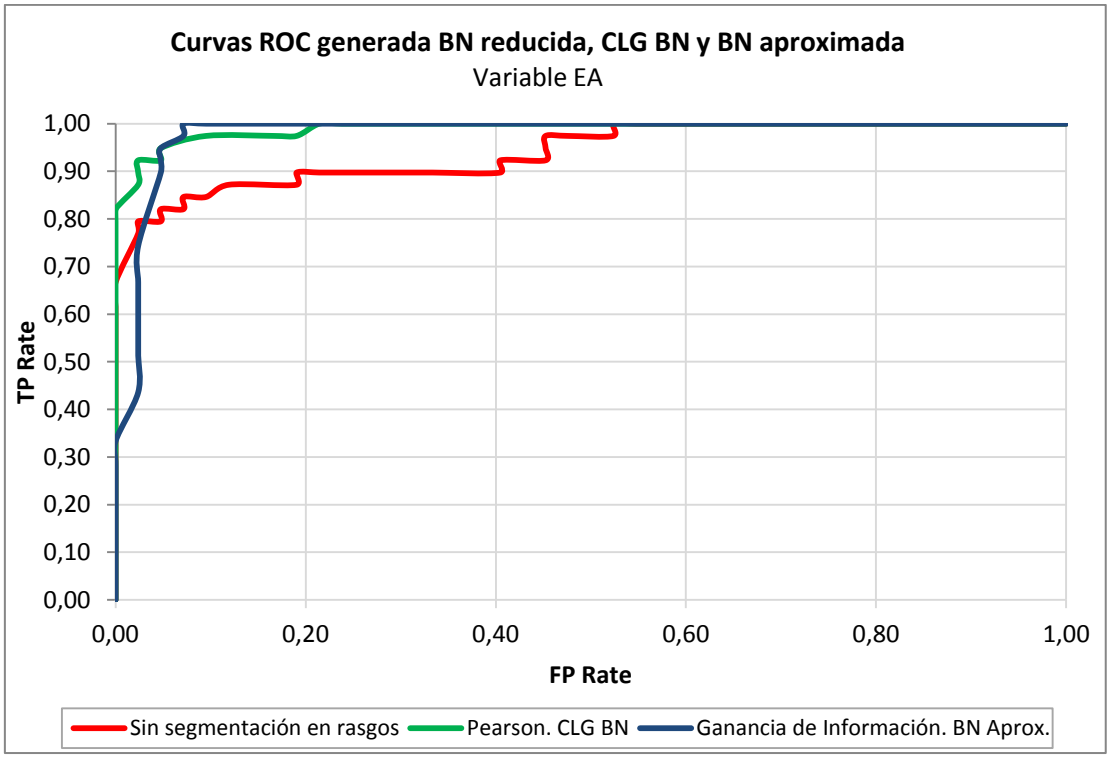


Figura 35.- Curvas ROC para comprobar la importancia de la segmentación de la producción oral de atributos lingüísticos en unidades menores y significativas (variable *EA*).

En la Tabla 25 se detallan las métricas de rendimiento obtenidas con las tres BNs del experimento.

Tabla 25.-Métricas de rendimiento para comprobar la importancia de la segmentación de la producción oral de atributos lingüísticos en unidades menores y significativas (variable *EA*).

	Sin segmentación en rasgos	Pearson. CLG BN	Ganancia de Información. BN Aproximada
True Positive (TP)	31	36	39
True Negative (TN)	40	41	39
False Positive (FP)	2	1	3
False Negative (FN)	8	3	0
TP rate	0,7949	0,9231	1
FP rate	0,0476	0,0238	0,0714
Precisión	0,9394	0,973	0,9286
Exactitud	0,8765	0,9506	0,963
Mean Squared Error	0,0995	0,0615	0,0699
Root Mean Squared Error	0,3154	0,2479	0,2645
AUC	0,9402	0,989	0,9799
Threshold	0,960581	0,511505	0,613875

Merece la pena seguir profundizando en las diferencias del rendimiento de la BN reducida, respecto a la CLG BN y a la BN con inferencia aproximada. Para ello, en la Figura 36 se analiza las probabilidades a posteriori de la variable de interés ***DSD*** para todos los casos de la muestra. En este análisis se selecciona la variable ***DSD*** en lugar de la variable ***EA***, porque el objetivo es estudiar la importancia de la segmentación de las

definiciones orales en los once bloques conceptuales que propone el corpus [1], sin tener en cuenta factores de riesgo ni de protección. El formato de las gráficas es similar a las del capítulo anterior, es decir, cada punto representa la probabilidad a posteriori de la variable de interés ***DSD*** en los estados *ausente* y *presente*. La Figura 36 se ha generado con las probabilidades a posteriori de la variable ***DSD***, obtenidas para la muestra de sujetos sanos.

En la Figura 37 se representan las probabilidades a posteriori, de todos los casos del corpus, inferidas para la muestra de sujetos enfermos de EA.

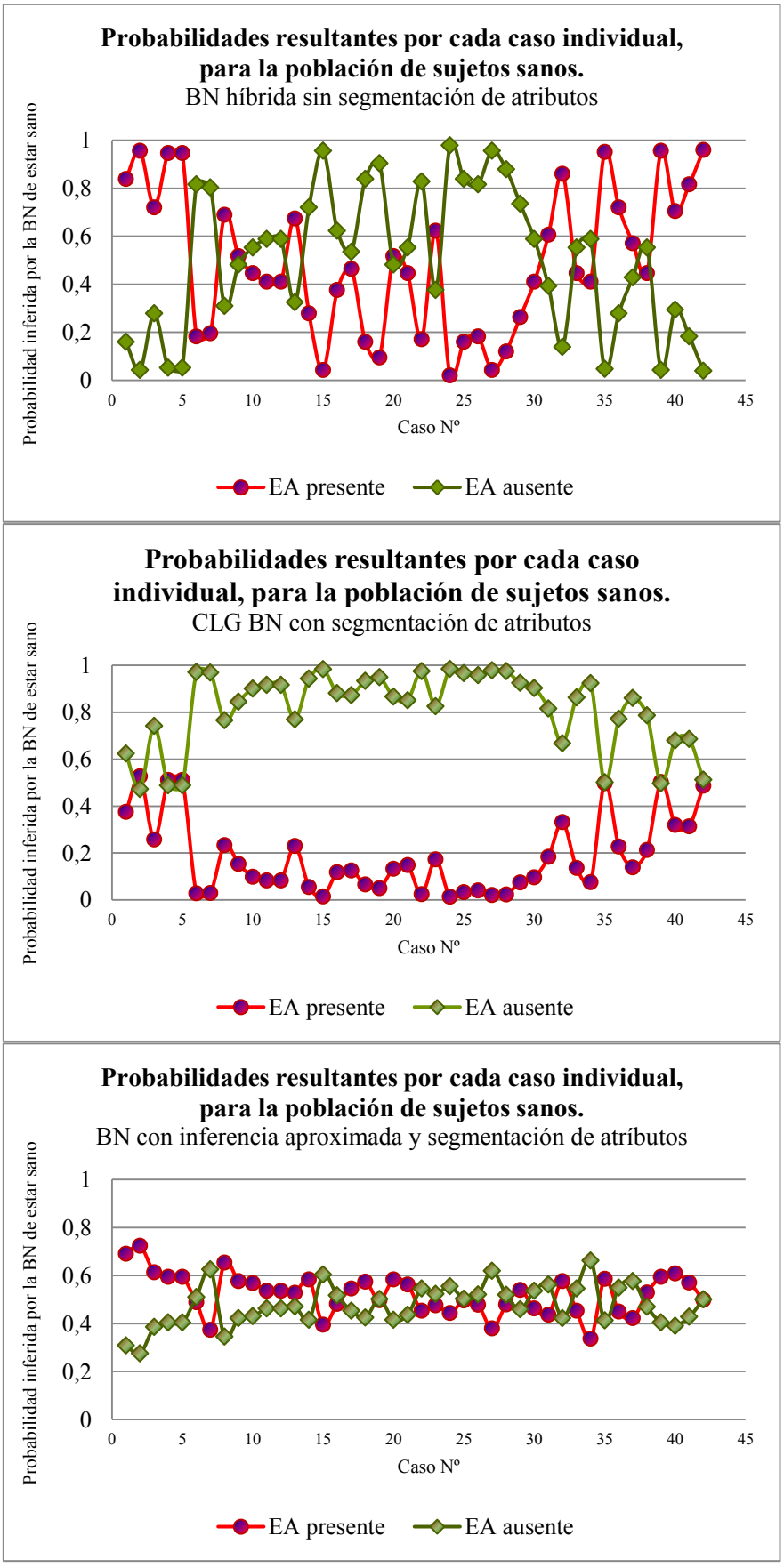


Figura 36.- Probabilidades a posteriori de la variable DSD del experimento 1 de las BN híbridas a) BN reducida, b) CLG BN, c) BN con inferencia aproximada. Muestra de sujetos sanos.

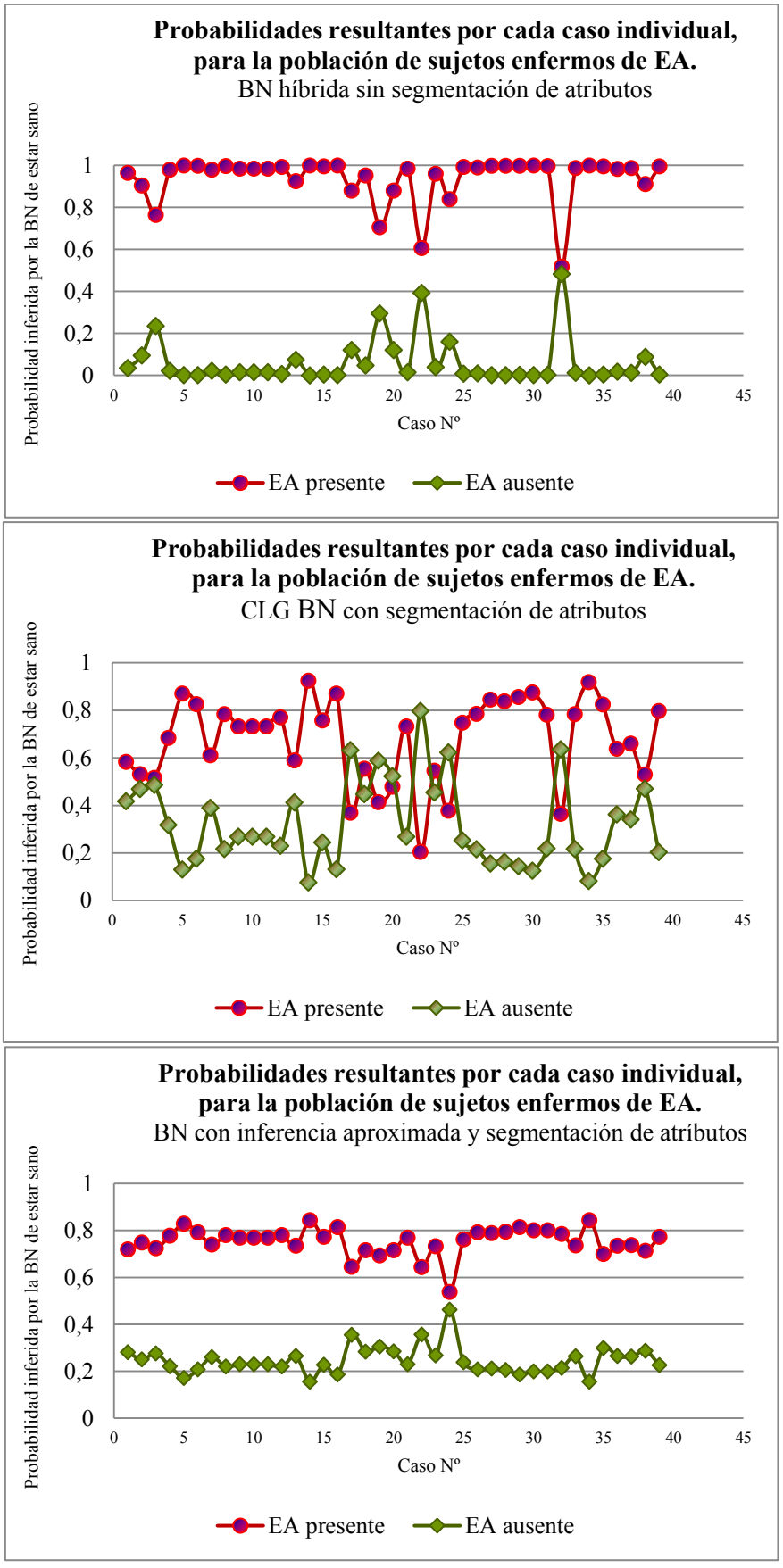


Figura 37.- Probabilidades a posteriori del experimento 1 con BN híbridas (muestra de sujetos enfermos de EA).

En la Tabla 26 se detallan las distintas métricas obtenidas para la variable ***DSD***. En estas métricas la BN que mejor resultado proporciona es la BN aproximada.

Tabla 26.- Métricas de rendimiento para comprobar la importancia de la segmentación de la producción oral de atributos lingüísticos en unidades menores y significativas (variable *DS*).

	Sin segmentación en rasgos	Pearson. CLG BN	Ganancia de Información. BN Aproximada
True Positive (TP)	28	33	38
True Negative (TN)	41	39	38
False Positive (FP)	1	3	4
False Negative (FN)	11	6	1
TP rate	0,7179	0,8462	0,9744
FP rate	0,0238	0,0714	0,0952
Precisión	0,9655	0,9167	0,9048
Exactitud	0,8519	0,8889	0,9383
Mean Squared Error	0,1785	0,0959	0,1766
Root Mean Squared Error	0,4225	0,3097	0,4202
AUC	0,9396	0,9707	0,9805

En la Tabla 27 se puede comprobar que las BNs que modelan la segmentación de las definiciones orales en estos once bloques conceptuales propuestos en [1], mejoran las métricas de rendimiento respecto a las BNs que no modelan esta segmentación, tanto en las BN híbridas como en las BN discretas. Las diferencias entre ambos tipos de BNs son sutiles pero las BNs que modelan esta segmentación mejoran su eficacia. Cabe destacar que en las dos CLG BN, se han generado las curvas ROC con las probabilidades a posteriori de la variable ***DS*** en lugar de la variable ***EA***, en una BN en la que no se tienen en cuenta los factores de contexto. Estás métricas se han obtenido utilizando en el experimento *leave-one-out cross validation*.

Tabla 27.- Comparativa de las métricas de rendimiento obtenidas con BNs que segmentan las definiciones orales en rasgos semánticos VS BNs que no realizan esta segmentación.

BNs	AUC	FP RATE	TP RATE	PRECISIÓN	EXACTITUD
Discreta	0,9628	0,0952	0,9231	0,900	0,9136
Discreta reducida	**0,9493**	**0,0714**	**0,8462**	**0,9167**	**0,8889**
CLG BN	0.9713	0.0476	0.8462	0.9429	0.9012
CLG BN reducida	**0,9396**	**0,0238**	**0,7179**	**0,9655**	**0,8519**

9.2 Eficacia de los distintos coeficientes de regresión en las CLG BN.

En esta sección se realiza una comparativa de rendimiento de distintos métodos de inferencia de la CLG BN. Cada método de inferencia construye ecuaciones de regresión (ver detalles en el capítulo 4) para especificar el efecto hipotetizado de ciertas variables denominadas predictoras sobre otras variables denominadas criterio (ver detalles en el capítulo 6).

En la Figura 38 se representan distintas curvas ROC obtenidas con distintos métodos de inferencia de la CLG BN. Cada método de inferencia utiliza un coeficiente de correlación o asociación para construir ecuaciones de regresión. Estos coeficientes de correlación o asociación se describieron en detalle en el capítulo 4 y son: correlación de Pearson, distancia euclídea modificada, regresión simple, ganancia de información y pesos de atributos. En este método de diagnóstico existen tantos coeficientes como variables tiene el corpus lingüístico.

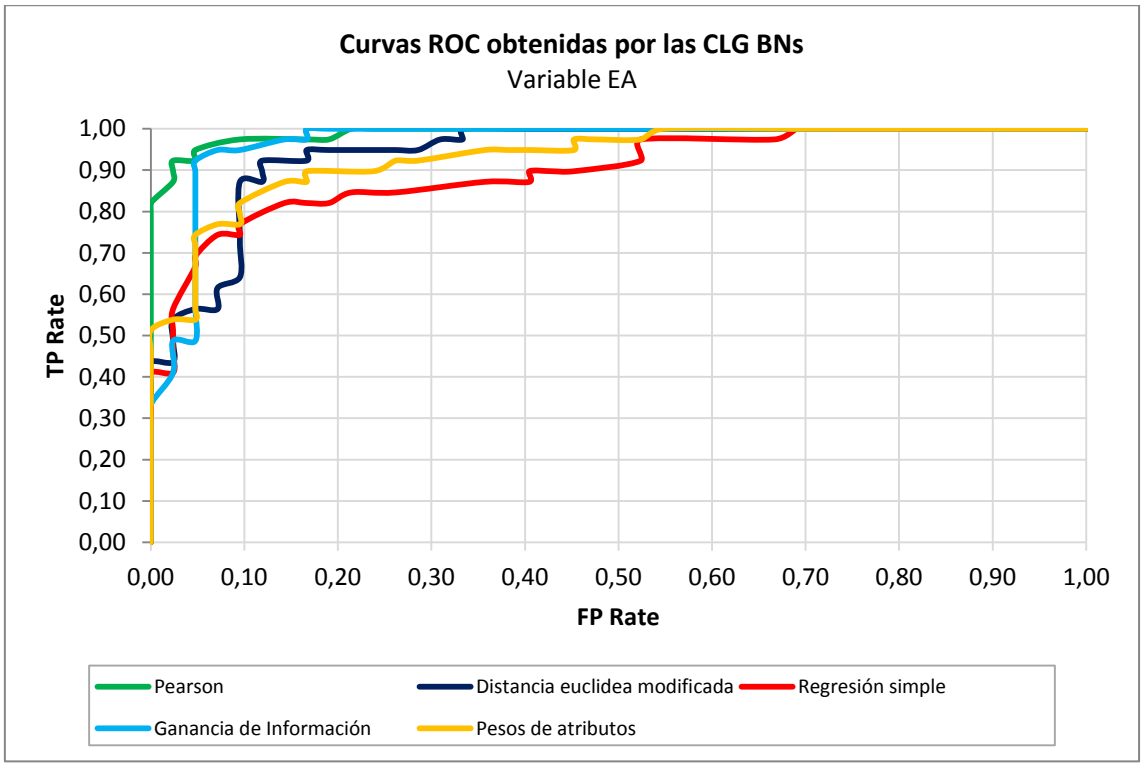


Figura 38.- Curvas ROC obtenidas con las CLG BN utilizando distintos coeficientes de asociación o correlación.

En la Tabla 28 se muestran las métricas de rendimiento obtenidas con la CLG BN, empleando distintos coeficientes en las ecuaciones de regresión para el cálculo de las variables latentes.

Tabla 28.- Métricas de rendimiento para la CLG BN empleando distintos coeficientes de asociación o correlación.

	Pearson	Distancia Euclídea Modificada	Regresión Simple	Ganancia de Información	Pesos de Atributos
True Positive (TP)	36	36	29	37	35
True Negative (TN)	41	37	38	39	32
False Positive (FP)	1	5	4	3	10
False Negative (FN)	3	3	10	2	4
TP rate	0,9231	0,9231	0,7436	0,9487	0,8974
FP rate	0,0238	0,119	0,0952	0,0714	0,2381
Precisión	0,973	0,878	0,8788	0,925	0,7778
Exactitud	0,9506	0,9012	0,8272	0,9383	0,8272
Mean Squared Error	0,0615	0,1194	0,123	0,0746	0,1209

	Pearson	Distancia Euclídea Modificada	Regresión Simple	Ganancia de Información	Pesos de Atributos
Root Mean Squared Error	0,2479	0,3456	0,3507	0,2731	0,3477
AUC	0,989	0,9414	0,8968	0,9664	0,931

En la Tabla 29 se analiza el rendimiento que se consigue en la clasificación de sujetos enfermos de EA y sujetos sanos, estudiando las probabilidades a posteriori de las variables intermedias de las CLG BN. Para realizar la comparativa de la Tabla 29 se genera una curva ROC por cada variable intermedia y se sigue el mismo procedimiento que el utilizado con las variables de interés.

Tabla 29.- Comparativa de las métricas de rendimiento obtenidas con las variables intermedias en la CLG BN.

	CLG BN RAZONAMIENTO DEDUCTIVO		
Variables	AUC	FP RATE	TP RATE
DS_{SV}	0,9695	0,1905	0,9744
DS_{SNV}	0,9328	0,2619	0,9744
$DS_{Manzana}$	0,9261	0,0952	0,8205
DS_{Perro}	0,8871	0,119	0,7692
DS_{Pino}	0,9023	0,2619	0,9231
DS_{Coche}	0,9322	0,2619	0,9744
DS_{Silla}	0,8730	0,0714	0,6667
$DS_{Pantalón}$	0,862	0,2857	0,8462

9.3 Eficacia de los distintos coeficientes de regresión en una BN híbrida con inferencia aproximada.

Esta sección es similar a la anterior pero utiliza una BN híbrida con algoritmos de inferencia aproximados (ver detalles en el capítulo 6).

En la Figura 39 se representan varias curvas ROC correspondientes a cada método de inferencia aproximada. La diferencia entre un clasificador y otro, es el coeficiente utilizado para ponderar la importancia predictora de las variables.

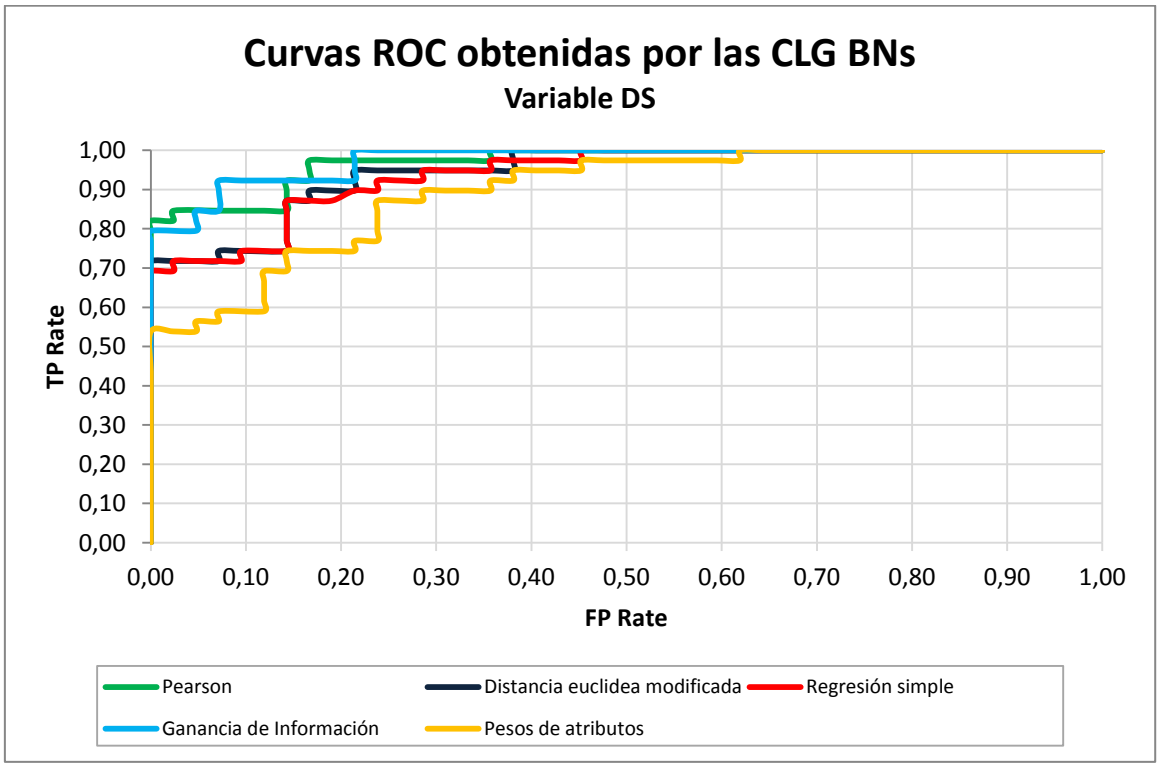


Figura 39.- Curvas ROC obtenidas con las BN con inferencia aproximada utilizando distintos coeficientes para las ecuaciones de regresión en la inferencia de las variables latentes.

En la Tabla 30 se detallan otras métricas de rendimiento obtenidas con la BN híbrida con inferencia aproximada.

Tabla 30.- Métricas de rendimiento para la BN con inferencia aproximada utilizando distintos coeficientes para las ecuaciones de regresión en la inferencia de las variables latentes.

	Pearson	Distancia Euclídea Modificada	Regresión Simple	Ganancia de Información	Pesos de Atributos
True Positive (TP)	33	28	28	36	21
True Negative (TN)	39	42	40	38	42
False Positive (FP)	3	0	2	4	0
False Negative (FN)	6	11	11	3	18
TP rate	0,8462	0,7179	0,7179	0,9231	0,5385
FP rate	0,0714	0	0,0476	0,0952	0
Precisión	0,9167	1	0,9333	0,9	1
Exactitud	0,8889	0,8642	0,8395	0,9136	0,7778
Mean Squared Error	0,0959	0,1064	0,1296	0,0891	0,1734
Root Mean Squared Error	0,3097	0,3261	0,3601	0,2985	0,4165
AUC	0,9707	0,9451	0,9389	0,9756	0,8938

9.4 Discusiones.

Las BNs híbridas han permitido realizar experimentos que proporcionan un valor añadido a la investigación de Peraita y Grasso [1], ya que se pone de relieve la importancia de la segmentación de la producción oral de rasgos semánticos en unidades menores y significativas. En las curvas de la Figura 35 se muestra como el rendimiento de la BN reducida, es inferior al rendimiento de la CLG BN y de la BN híbrida con inferencia aproximada. Se puede apreciar en la Tabla 25 como el AUC de la BN híbrida reducida es inferior al AUC de la CLG BN y al AUC de la BN aproximada.

En la Figura 36 se puede comprobar que son numerosos los casos de la BN reducida cuya probabilidad a posteriori se aleja mucho de su valor esperado, además se infiere, para muchos sujetos cognitivamente sanos, una probabilidad por encima de 0,8 de padecer la EA. Cabe recordar que para clasificar los casos como positivos o negativos se utiliza un umbral (threshold) cuyos valores son: 0,960581 para la BN reducida, 0,511505 para la CLG BN y 0,613875 para la BN híbrida.

En la Figura 37 a) se muestra como la mayor parte de las probabilidades $EA_{presente}$, se sitúan en torno a su valor esperado. En este experimento el *threshold* se sitúa en 0.96, dando lugar a once falsos negativos y un solo falso positivo.

Para la CLG BN el *threshold* (umbral) se establece en 0,51, permitiendo discriminar más fácilmente las personas sanas de las personas enfermas.

Hay que recurrir a Figura 36 para poder valorar la capacidad de discriminación de la BN con inferencia aproximada. Para la muestra de sujetos enfermos de EA, la probabilidad inferida de padecer la EA se sitúa por encima de 0,7, sin embargo para la población de sujetos sanos se sitúa por debajo de 0,6, permitiendo discriminar más fácilmente a las personas cognitivamente sanas de las personas con DS.

Desafortunadamente para esta investigación no hay posibilidad realizar exploraciones complementarias a estas personas concretas, ya que el corpus se ha publicado en el año 2010 y estos experimentos se han realizado en octubre del 2012.

En el segundo experimento se analizan distintos métodos de inferencia para la CLG BN. Cada método de inferencia incide directamente en la inferencia de las variables latentes, es decir, en cada método de inferencia se construyen distintas ecuaciones lineales estructurales para calcular las probabilidades a posteriori de dichas variables. De los resultados de la CLG BN de la Figura 38 se deduce que el método de inferencia con el que se obtienen mejores resultados, para esta BN, es aquel que utiliza el coeficiente de Pearson (color verde). Existen otros coeficientes que funcionan bastante bien en la CLG BN como el ratio de la ganancia de información o la distancia euclídea modificada.

En el tercer experimento es similar al segundo, salvo que utiliza para el cálculo de las probabilidades a posteriori de las variables latentes o intermedias, un método de inferencia aproximado. Con el método de inferencia aproximada, los mejores resultados se consiguen con el ratio de ganancia de información y con el coeficiente de correlación de Pearson. Por el contrario, el peor resultado se obtiene utilizando los pesos de las

variables según la producción oral de atributos lingüísticos. Las mejoras métricas se obtienen utilizando los coeficientes de Pearson y ganancia de información en la inferencia de las variables latentes.

Las BNs híbridas propuestas en esta tesis doctoral podrían abrir nuevas vías de investigación desde la perspectiva de la neuropsicología cognitiva para responder a preguntas como: ¿Por qué existen unas categorías semánticas que ayudan a predecir mejor la EA que otras? ¿Se trata de una característica de esta muestra en concreta o realmente los enfermos de EA presentan algún patrón donde por ejemplo, la categoría natural ***Manzana*** sirve para predecir mejor la EA? ¿Por qué si se segmenta la producción oral de atributos en rasgos semánticos se consigue un clasificador más eficiente?, etc. En el capítulo 12 se analizarán estas preguntas desde la perspectiva de la estadística y la IA, y se podrá comprobar que los resultados convergen con los resultados de la neuropsicología cognitiva.

Capítulo

Evaluación de las estrategias evolutivas 10

En el capítulo 7 se describió en profundidad las técnicas de optimización, basadas en estrategias evolutivas, para las TPCs. En este capítulo se optimiza la TPC de la variable ***EA*** de las distintas BNs propuestas en tesis doctoral, tanto discretas como híbridas. El algoritmo es aplicable a todas las TPCs de la BN, sin embargo, debido al elevado coste computacional que requiere, se ha limitado el problema a una sola TPC, concretamente se ha limitado a la optimización de la TPC de la variable ***EA***. Se ha seleccionado esta variable por dos razones: es la principal variable de interés y su TPC se calcula con el simplificador *Naive Bayes*. En este capítulo se evalúan los métodos y operadores (penalización del *fitness*, de mutación, de selección de la población y de control de la población) que mejor se adecuan a nuestro método de diagnóstico. Los objetivos fundamentales de este capítulo son:

- Encontrar los métodos (penalización del *fitness*, mutación y control de población) más apropiados para el problema.
- Aplicar el algoritmo de estrategia evolutiva a toda la población, para encontrar una TPC que optimice el rendimiento de las BNs.

La estructura del capítulo es la siguiente. En la sección 10.1 se evalúan los distintos métodos de penalización del *fitness*, de mutación y de control de la población. En la sección 10.2 se realizan una batería de experimentos, cuyo fin es analizar la mejora del rendimiento obtenida al optimizar la BN. En la sección 10.3 se plantean una serie de discusiones y se analiza el problema del sobreajuste del modelo 1 de BN (inferencia por razonamiento deductivo). En esta sección se demuestra que la técnica es apropiada para este método de diagnóstico, aunque en esta tesis no se haya podido contrastar con precisión los resultados.

10.1 Evaluación del algoritmo.

Comparar la calidad de los resultados producidos por el algoritmo de estrategias evolutivas y afinar los parámetros del algoritmo para conseguir mejorar el rendimiento de los clasificadores, es muy importante para el método de diagnóstico. Por ello, en esta sección se evalúan las distintos operadores y métodos implementados por el *framework*.

El proceso de evaluación se ha dividido en dos fases:

- La primera fase analiza los distintos métodos de penalización, mutación y control de la población aplicadas sobre un conjunto reducido de casos del corpus.
- La segunda fase, una vez encontrada los métodos que mejor se adecuan a nuestro problema, se optimizará la BN utilizando en el cálculo del *fitness* todos los casos del corpus lingüístico.

Debido a la naturaleza estocástica de las estrategias evolutivas, el proceso de optimización se repite 10 veces con objeto de obtener unas medidas de rendimiento de naturaleza estadística. Estas medidas de rendimiento son:

- Tasa de éxito o *SR* (success rate). Mide el porcentaje de ejecuciones en los que algoritmo finaliza con éxito.
- Eficacia o *MBF* (mean best fitness). *MBF* es la media de los mejores *fitness* obtenidos en todas las ejecuciones.
- Eficiencia o *AES* (average number or evaluations to a solution). Mide el número de generaciones que necesita el algoritmo para llegar a una solución óptima.

Esta sección comienza analizando el funcionamiento de las distintas estrategias de penalización del *fitness*.

10.1.1 Métodos de penalización del fitness.

Como se ha indicado anteriormente, se han desarrollado tres métodos de penalización del *fitness*: a) penalización estática, b) penalización dinámica, c) penalización auto-adaptativa. En la Tabla 31 se detallan los parámetros utilizados en el *framework* de estrategias evolutivas para analizar estos tres métodos de penalización del *fitness*.

Tabla 31.- Configuración utilizada para determinar la estrategia de penalización del fitness.

penaltyMethod	➢ Sin penalización del fitness ➢ Estática ➢ Dinámica ➢ Autoadaptativa.
μ	45
λ	150
Número de generaciones	1000
Threshold Desv. Tip.	0.01
Recombination Method	Intermediary
Recombination Scope	Local
Finalización sin mejora	8
W	3

	Este parámetro se utiliza para la penalización estática.
B1	1.8. Este parámetro se utiliza para la penalización dinámica.
B2	2.3 Este parámetro se utiliza para la penalización dinámica.
survivorSelection	(μ+λ)
Mutation	Uncorrelated Mutation with One Step Size
Casos del corpus lingüístico	20 sujetos sanos. 20 sujetos enfermos de EA.

En la Figura 40 se representa el MBF para 10 ejecuciones independientes del proceso de optimización. Cabe recordar que la inicialización de la población utiliza un algoritmo memético y por tanto, el problema parte de una solución suboptimal. Se aprecia en la Figura 40 que todos los métodos de penalización del *fitness* consiguen un *fitness* óptimo (AUC ≥ 0.99).

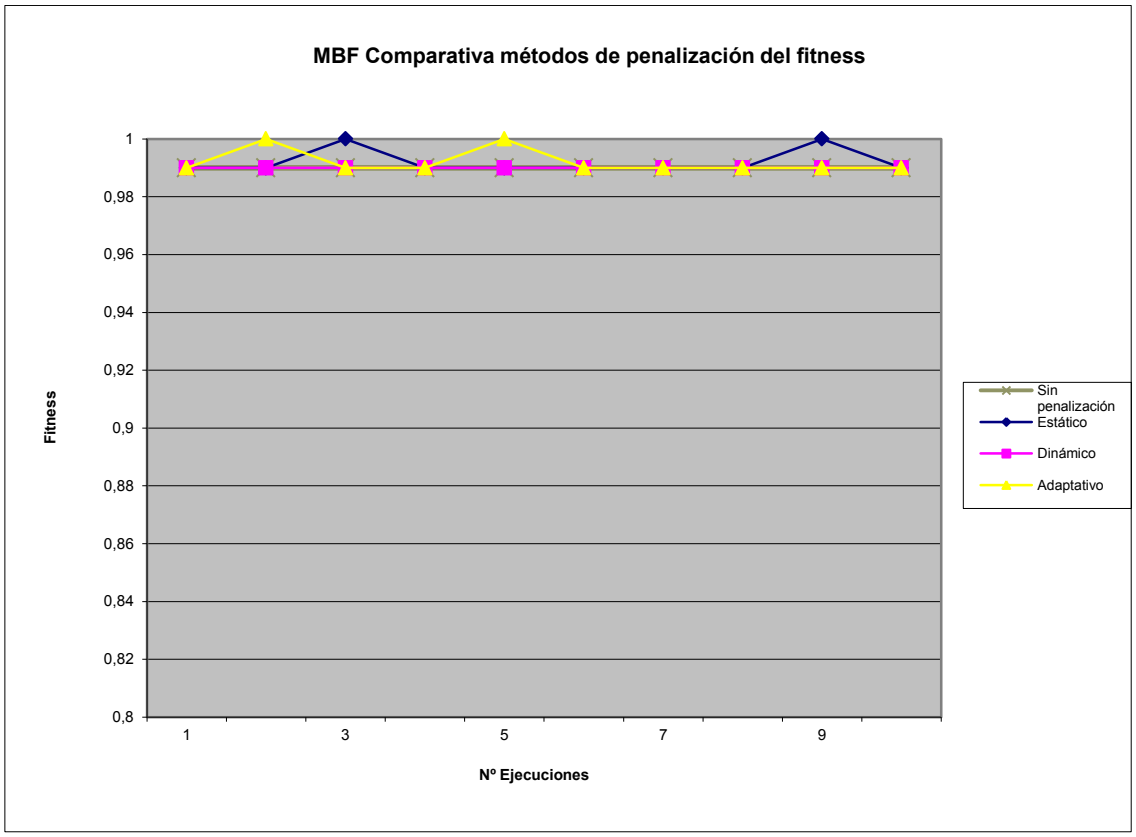


Figura 40.- MBF. Comparativa métodos de penalización del fitness.

En la Figura 44 se analiza, por cada método de penalización, el número de generaciones necesarias para converger a una solución óptima. Cabe destacar que existe una restricción importante en los métodos de mutación y recombinación, ya que todos los

componentes de la TPC de las BNs deben estar comprendidos entre 0 y 1. El algoritmo de penalización utiliza la formulación siguiente para la penalización del *fitness*:

$$p_{edad,sexo,ne,dc} = \begin{cases} 0 - \varepsilon \ si \ p_{edad,sexo,ne,dc} < 0 \\ p_{edad,sexo,ne,dc} \ si \ está \ comprendida \ entre \ 0 \ y \ 1 \\ 1 - \varepsilon \ si \ p_{edad,sexo,ne,dc} > 1 \end{cases} \quad (10.1)$$

donde

ε: es un valor de residuo cuya función es hacer posible todos los estados de la variable.

En la Figura 41 se muestra que el método *sin penalización de fitness* consigue converger a la solución óptima antes que las estrategias de penalización estática, dinámica y adaptativa.

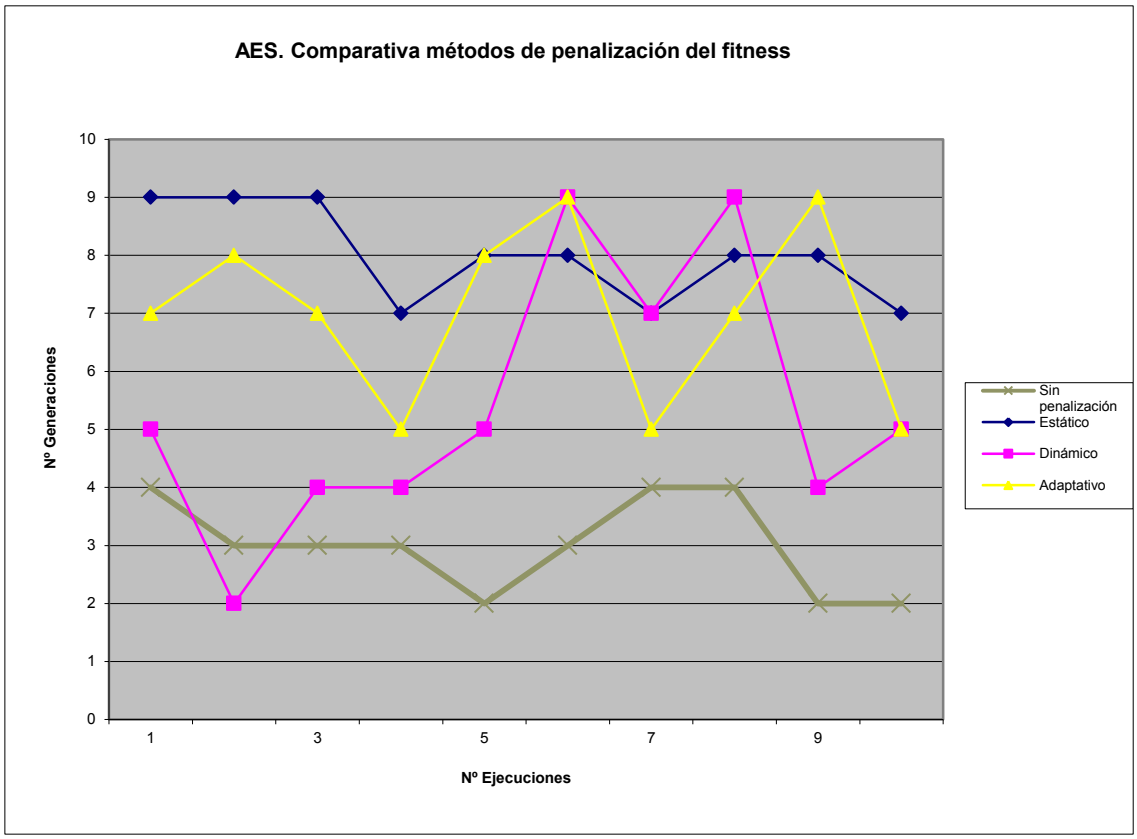


Figura 41.- AES.- Comparativa métodos de penalización del fitness.

La Figura 42 analiza la tasa de éxito de cada método de penalización. Todos los métodos de penalización han convergido hacia una solución óptima.

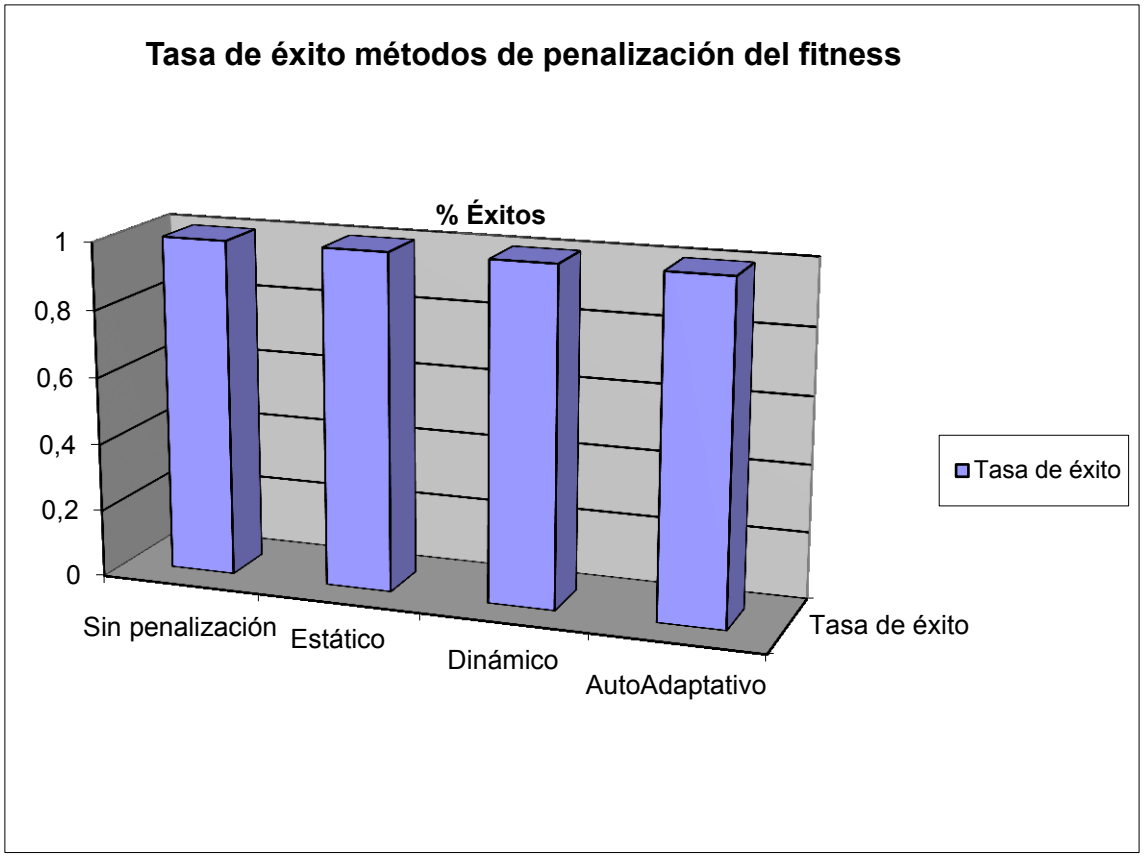


Figura 42.- SR. Tasa de éxito de los distintos métodos de penalización.

Se puede deducir de esta sección que los métodos de penalización más eficientes son: *sin penalización de fitness* y *penalización de fitness dinámica*.

10.1.2 Métodos de Mutación.

Se han desarrollado cuatro métodos de mutación distintos: a) *Uncorrelated Mutation with One Step Size*, b) *Uncorrelated Mutation with n Step Size*, c) *Deterministic Mutation* d) *Adaptative Mutation*.

En esta sección se analiza el rendimiento de cada método de mutación teniendo en cuenta la tasa de éxito, eficacia y eficiencia.

Tabla 32.- Configuración utilizada para determinar la estrategia de mutación más apropiada.

Método de mutación	➢ Uncorrelated Mutation with One Step Size ➢ Uncorrelated Mutation with n Step Size ➢ Determinista ➢ Autoadaptativo
Recombinación variables	Intermediary
Recombinación parámetros	Intermediary
Survivor Method	$(\mu+\lambda)$.

μ	45
λ	150
Nº Generaciones	1000
Rango	$0 \leq x_i \leq 1$
thresholdDesvTip	0.01
terminationK	20
terminationCondition	0.01
Penalty	Penalización dinámica
C	0.817. Constante de Rechenberg
K	3. Numero de iteraciones de Rechenberg
Casos del corpus lingüístico	20 sujetos sanos. 20 sujetos enfermos de EA.

En la Figura 43 se compara el mejor *fitness* de todas las generaciones y para ello, se ha ejecutado el algoritmo 10 veces. En la Figura 43 se hace evidente que los métodos de mutación más eficaces son: *Uncorrelated Mutation with One Step Size* y *Uncorrelated Mutation with n Step Size*. Para este problema se pueden descartar los métodos: *Deterministic Mutation* y *Adaptative Mutation*, ya que ninguno de los métodos alcanzan la solución óptima.

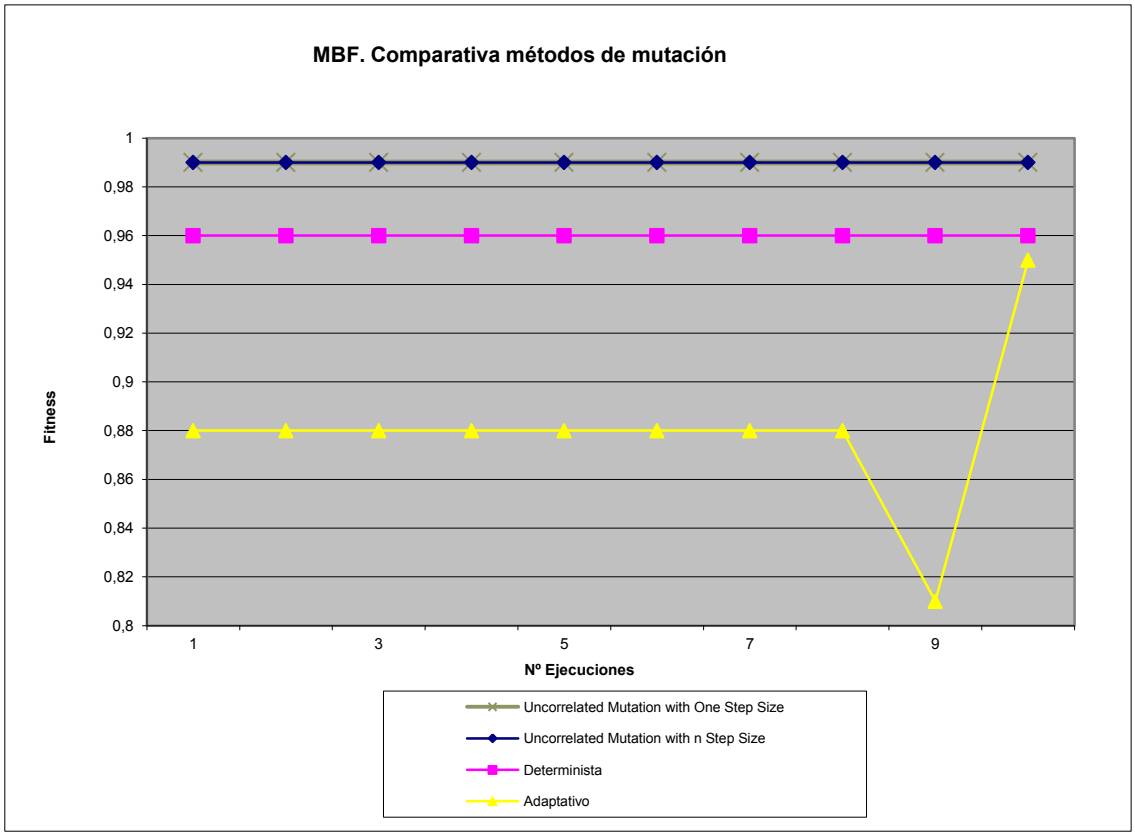


Figura 43.- MBF. Comparativa métodos de mutación.

En la Figura 44 se analiza el número de generaciones que necesita cada método de mutación para converger a una solución óptima. Se puede determinar en la Figura 44, que el método más eficiente es el *Uncorrelated Mutation with One Step Size.*

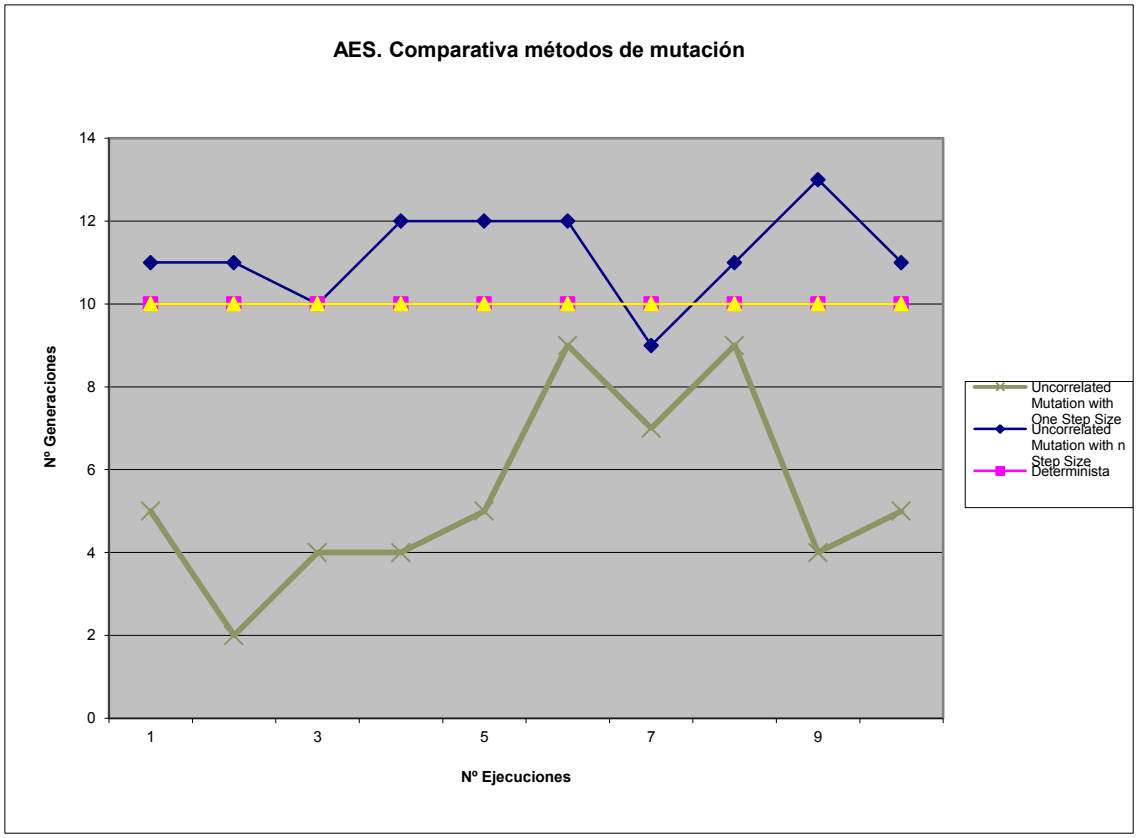


Figura 44.- AES comparativa métodos de mutación.

Por último, se analiza la tasa de éxito de cada método de mutación, siendo evidente que los métodos más eficaces con las BNs propuestas en esta tesis son: *Uncorrelated Mutation with One Step Size* y *Uncorrelated Mutation with n Step Size.*

Podemos concluir de esta sección que el método de mutación más eficiente y eficaz para la optimización de esta BN es el *Uncorrelated Mutation with One Step Size.*

10.1.3 Métodos de control de la población.

Como se ha indicado anteriormente, se ha implementado los métodos *AVGaPS* y *PRoFIGA* para el control de la población. En los métodos de control de población (además de analizar *MBF*, *AES* y *SR*) se analiza la evolución del crecimiento de la población. Un método óptimo es aquel que converja a una solución óptima en pocas generaciones y con la menor población posible. Cuanto mayor es la población, más requisitos computacionales exige el algoritmo. En la Tabla 33 se detallan los parámetros del algoritmo.

Tabla 33.- Configuración utilizada para determinar la estrategia de control de la población más apropiada.

Método de control de la población	➢ AVGaPS Proporcional ➢ AVGaPS Lineal ➢ AVGaPS Bi-lineal ➢ PRoFIGA
penaltyMethod	Método de penalización dinámica.

B1	1.8. Este parámetro se utiliza para la penalización dinámica.
B2	2.3 Este parámetro se utiliza para la penalización dinámica.
Mutation	Uncorrelated Mutation with n Step Size
MIN_LT	1 Parámetro utilizado para AVGaPS
MAX_LT	6 Parámetro utilizado para AVGaPS
IncreaseFactor	0.8 Parámetro utilizado por PRoFIGA
V	3 Número de incrementos sin mejoras, utilizado por PRoFIGA
Reducción	0.01 porcentaje de reducción.

En la Figura 45 se muestran los mejores *fitness* conseguidos por *AVGaPS* y *PRoFIGA*. Para *AVGaPS*, como se indicó en el capítulo 7, se implementan tres métodos de control de la población: a) proporcional, b) lineal, c) bilineal. Todos los métodos de control de la población encuentran un *fitness* óptimo. Además, algunos métodos convergen hacia un AUC = 1, que es la clasificación perfecta. Es muy importante tener en cuenta que se está seleccionando un subconjunto de casos del corpus y puede existir un sobreajuste de las probabilidades condicionales (en la sección 10.3 se analiza el sobreajuste). Por otro lado, esta selección se realiza de forma aleatoria y es posible que existan variaciones entre ejecuciones distintas del algoritmo.

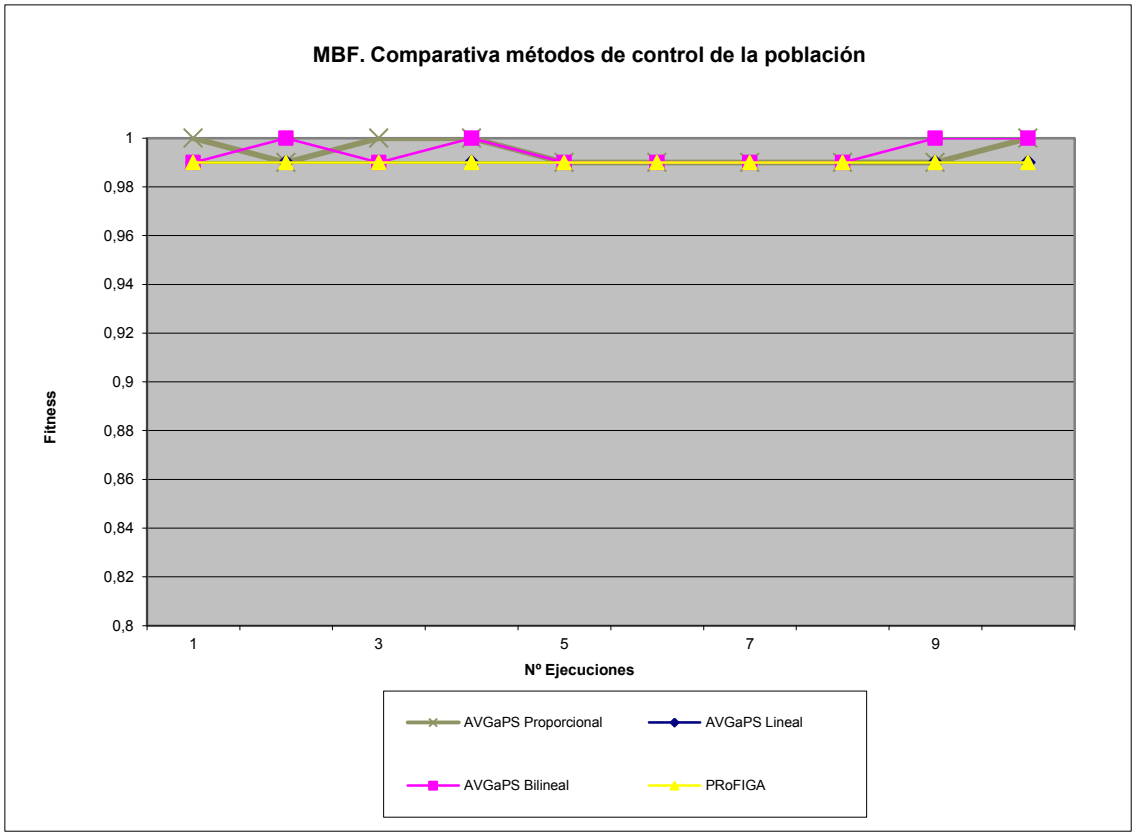


Figura 45.- MBF. Métodos de control de la población.

En la Figura 46 se analiza el número de generaciones que necesita cada método de control de población para alcanzar una solución óptima. Se puede determinar que *PRoFIGA* y, *AVGaPS* Proporcional y Bilineal, son los algoritmos que menos generaciones necesitan para converger a una solución óptima.

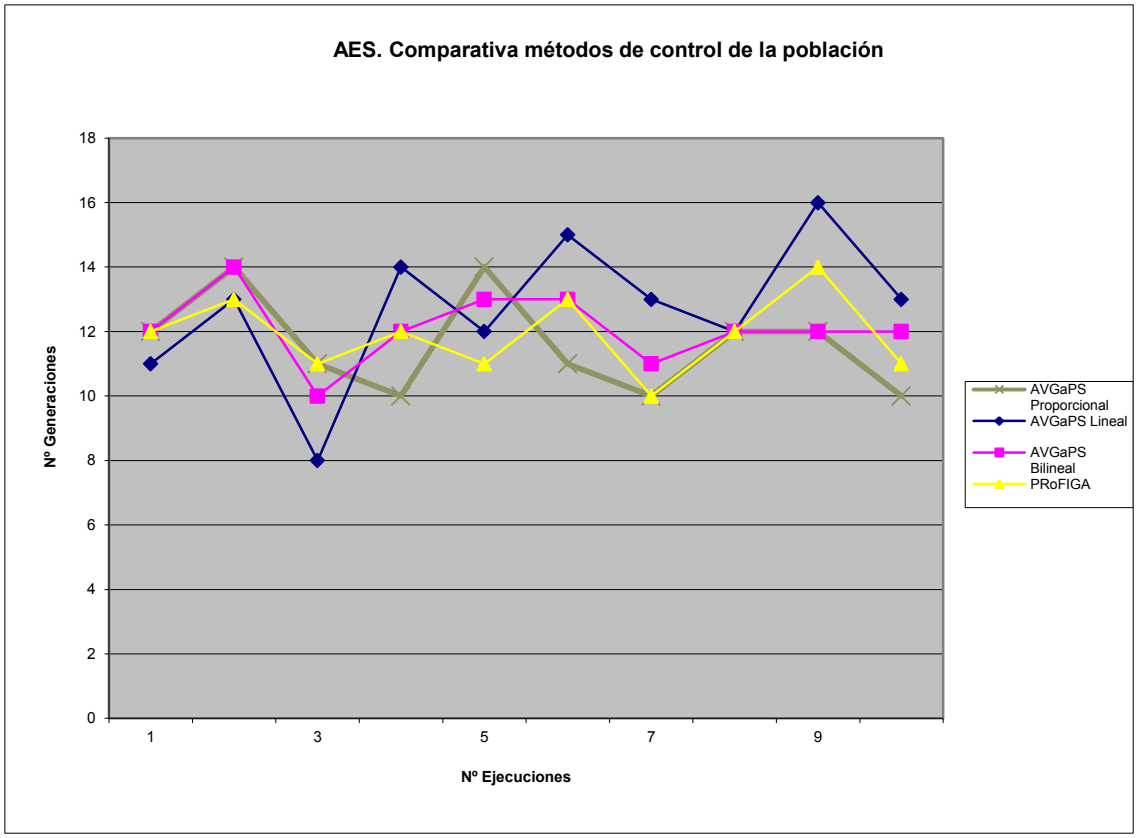


Figura 46.- AES. Comparativa métodos de control de la población.

Respecto a la tasa de éxito, tal y como se representa en la Figura 47, todos los métodos alcanzan un *fitness* óptimo.

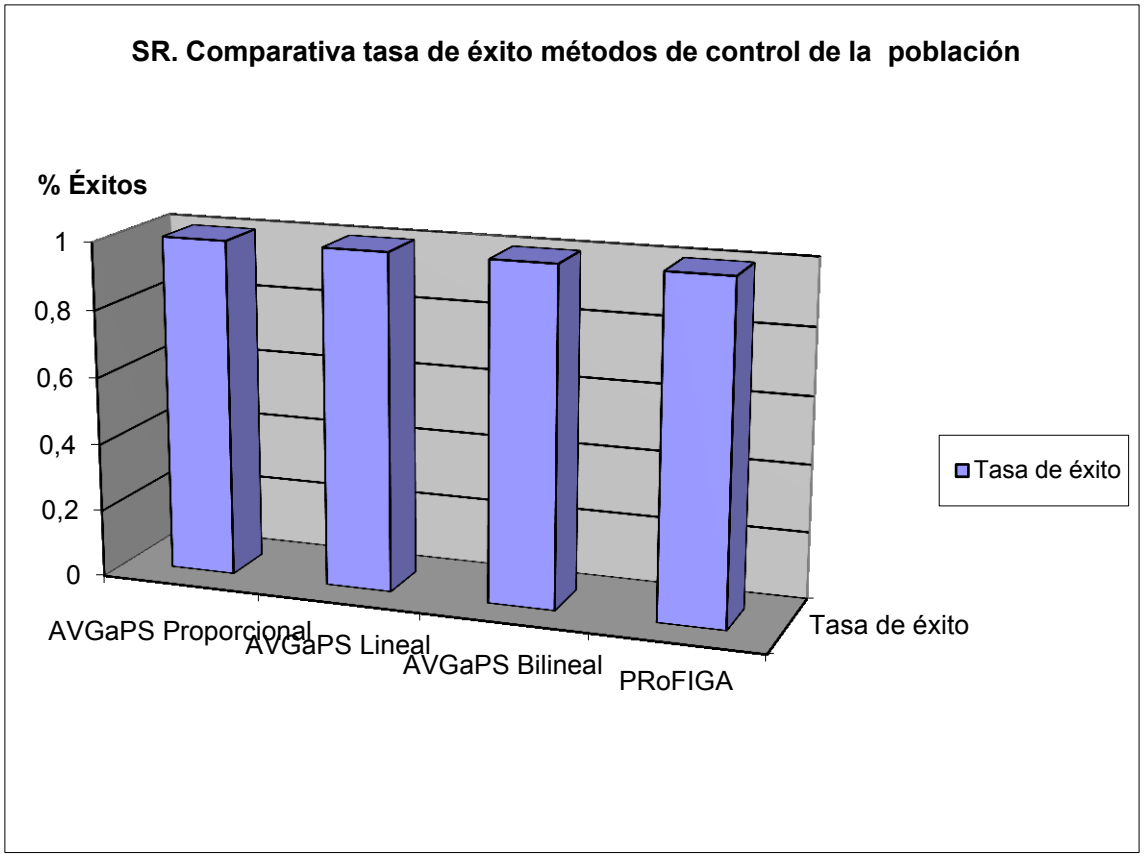


Figura 47.- SR. Comparativa tasa de éxito métodos de control de la población.

En los métodos de control de la población es interesante analizar el crecimiento de la población por cada generación. Los resultados que se recogen en la Figura 48, parte del resultado obtenido de la mejor iteración –mejores resultados en cuanto al tamaño de la población—del algoritmo por cada método de control de población, es decir, se ha seleccionado una ejecución donde se haya conseguido un *fitness* óptimo en el menor número de generaciones. De la Figura 48 se puede concluir que un buen método para el control de la población es *PRoFIGA*, ya que tiene un menor crecimiento de la población por cada generación. Por otro lado, la población inicial se inicializa con pocos individuos, debido a las exigencias de recursos computacionales que necesita el algoritmo con poblaciones muy grandes. En otros problemas que requieran menos recursos computacionales, es posible que algoritmo comience con muchos individuos y se vaya reduciendo a medida que evolucionan las generaciones.

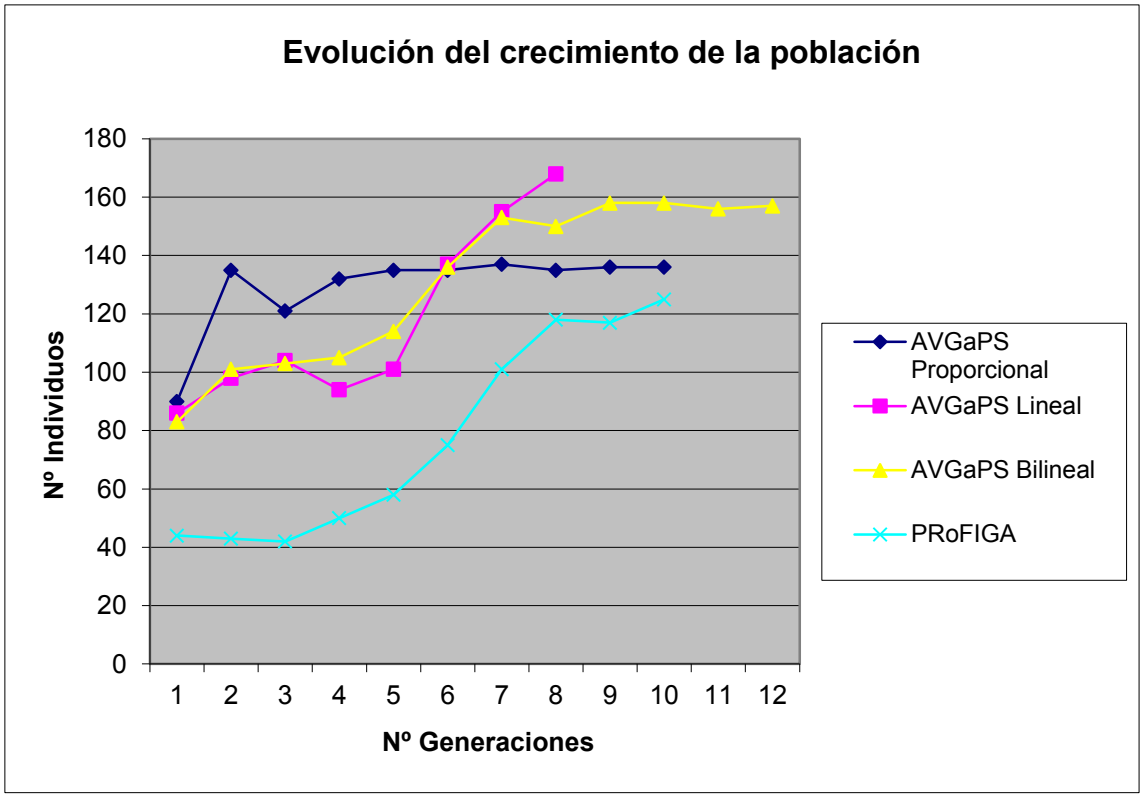


Figura 48.- Evolución del crecimiento de la población. Comparativa métodos de control de la población.

10.2 Validación del método de optimización.

En las secciones anteriores, se han analizado distintas estrategias para la optimización de la TPC de la variable ***EA***. En esta sección se aplicará el algoritmo de estrategia evolutiva a toda la población del corpus y se podrá comprobar cómo esta técnica de optimización mejora el rendimiento de las BNs.

Para validar el método de optimización se analiza la curva ROC y el AUC, obtenido antes y después de aplicar la optimización. Las BNs a las que se les van aplicar las estrategias evolutivas son: el modelo 1 "*Inferencia por razonamiento deductivo*" de la BN discreta y a la CLG BN.

En esta sección para optimizar estas BNs se utilizan todos los casos del corpus con los siguientes parámetros:

Tabla 34.- Parámetros utilizado del algoritmo de optimización para la BN Discreta.

Tipo de BN	Discreta
Segmentación de atributos	Por edad.
Población	Española.
Método de control de la población	PRoFIGA

Método de penalización	Sin penalización del fitness.
Método de Inicialización	Algoritmo memético de inicialización de la población.
B1	1.8. Este parámetro se utiliza para la penalización dinámica.
B2	2.3 Este parámetro se utiliza para la penalización dinámica.
Mutation	Uncorrelated Mutation with One Step Size
IncreaseFactor	0.8 Parámetro utilizado por PRoFIGA
V	3 Número de incrementos sin mejoras, utilizado por PRoFIGA
Reducción	0.01 porcentaje de reducción.
Nº de casos del corpus	42 personas sanas. 39 personas enfermas de EA.

10.2.1 Validación de la BN Discreta.

En primer lugar, se va a validar el rendimiento obtenido para el modelo 1 de la BN segmentando los conglomerados de atributos por edad, combinando el estudio epidemiológico [18] con el corpus lingüístico y aplicando un algoritmo memético para la inicialización de la población**.** La muestra utilizada está constituida por 42 personas cognitivamente sanas y 39 enfermos de EA.

En la Figura 49 se presentan dos curvas ROC, la primera, se ha obtenido con la BN antes de optimizar la TPC de la variable ***EA***, y la segunda, después de optimizar la TPC de dicha variable. La utilización de las curvas ROC se analiza en detalle en el capítulo 8.

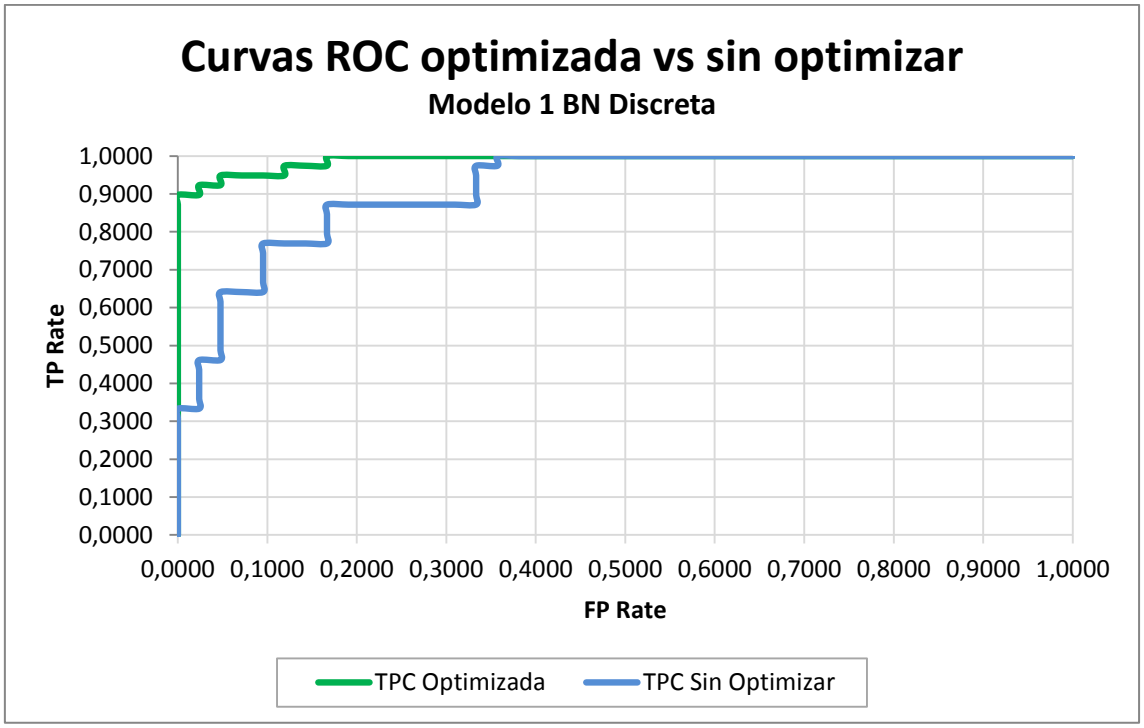


Figura 49-. Curvas ROC modelo 1 BN discreta optimizada vs sin optimizar

En la Tabla 35 se muestran las métricas de rendimiento para el modelo 1 de BN discreta optimizada y sin optimizar.

Tabla 35.- Métricas de rendimiento para el modelo 1 BN discreta optimizada vs sin optimizar

	Métricas de Rendimiento BN Optimizada	Métricas de Rendimiento BN Sin optimizar
True Positive (TP)	35	39
True Negative (TN)	42	26
False Positive (FP)	0	16
False Negative (FN)	4	0
TP rate	0,8974	1
FP rate	0	0,381
Precisión	1	0,7091
Exactitud	0,9506	0,8025
Mean Squared Error	0,131	0,1378
Root Mean Squared Error	0,362	0,3712
AUC	0,9908	0,9158

10.2.2 Validación de la BN Continua.

En esta sección se repite el mismo proceso que en la sección anterior pero con la CLG BN. Los parámetros de la CLG BN se describen en la Tabla 36.

Tabla 36.- Parámetros utilizado del algoritmo de optimización para la CLG BN.

Tipo de BN	BN híbrida
Método de cálculo variables continuas padres de variables continuas	Por aproximación / Ganancia de Información
Población	Española
Población medidas estadísticas	Sanos
Población validación	42 Sanos 39 Enfermos de EA

En la Figura 50 se presentan dos curvas ROC, la primera, se ha obtenido con la BN híbrida antes de optimizar la TPC de la variable ***EA***, y la segunda, después de optimizar la TPC de dicha variable.

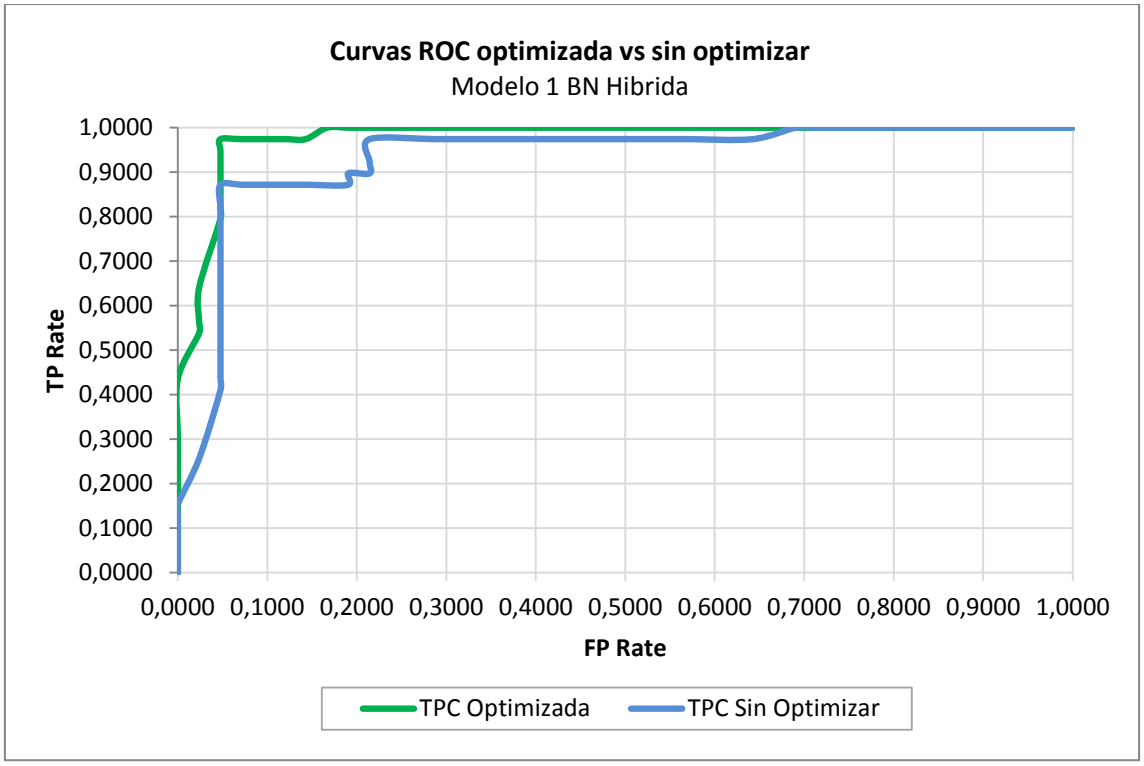


Figura 50.- Curvas ROC modelo 1 BN híbrida optimizada vs sin optimizar

En la Tabla 37 se muestran las métricas de rendimiento para el modelo 1 de la BN hibrida optimizada y sin optimizar.

Tabla 37.- Métricas de rendimiento para el modelo 1 BN híbrida optimizada vs sin optimizar.

	Métricas de Rendimiento BN Optimizada	Métricas de Rendimiento BN Sin optimizar
True Positive (TP)	38	34
True Negative (TN)	40	39
False Positive (FP)	2	3
False Negative (FN)	1	5
TP rate	0,9744	0,8718
FP rate	0,0476	0,0714
Precisión	0,95	0,9189
Exactitud	0,963	0,9012
Mean Squared Error	0,1196	0,1138
Root Mean Squared Error	0,3458	0,3373
AUC	0,978	0,9322

10.3 Discusiones.

Se puede concluir de estos experimentos que las estrategias evolutivas mejoran el rendimiento de las BNs discretas e hibridas. En la Figura 49 se puede observar como la curva ROC generada con el modelo 1 de la BN discreta, produce mejor resultado que la misma BN antes de aplicarle el algoritmo de optimización. Tal y como se indicó en la parte III de esta tesis, un punto de la curva ROC es mejor que otro si este se encuentra a noroeste del primero (*tp rate* es mayor y *fp rate* es mejor, o ambos). En la Figura 49 se puede apreciar como todos los puntos, generados con la BN optimizada, están más cercanos al punto (0, 1).

En la Tabla 35 se muestran las métricas de rendimiento y se deduce que la BN optimizada tiene un mejor rendimiento que la BN sin optimizar. Para calcular estas métricas se busca, con el algoritmo J48, un umbral (*threshold*) que maximice el *TP* y el *TN*, de tal forma que si este umbral varía las métricas varían.

La misma conclusión se puede extraer de la Figura 50 y la Tabla 37, donde se optimiza la TPC de la variable ***EA*** para una CLG BN, es decir, la CLG BN en la que se optimiza la TPC de la variable ***EA*** muestra un mejor comportamiento que la misma BN antes de optimizarse.

No obstante, dado el número de casos disponibles en el corpus [1] y el número de componentes de la TPC de la variable ***EA***, en los experimentos anteriores es posible que se esté produciendo un sobreajuste de los parámetros durante el proceso de optimización. A modo recordatorio, en la Figura 51 se muestra un fragmento de los enlaces causales del modelo 1 de la CLG BN.

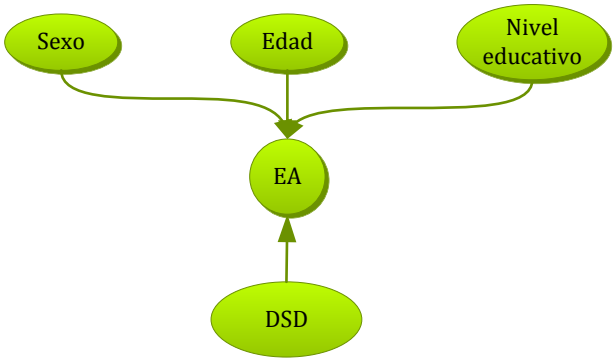


Figura 51.- Fragmento de la CLG BN Híbrida.

Dada las relaciones causales de este fragmento de la estructura de la CLG BN, la TPC de la variable ***EA*** tiene 144 componentes (6 segmentos de edad * 2 sexo * 3 nivel educativo * 2 estados variable ***DSD*** * 2 estados variable ***EA***). De los 144 componentes, la estrategia evolutiva optimiza 72, en concreto, sólo se optimizan el estado de la variable ***EA*** *presente*.

En la sección 2 se han empleado 40 casos entre personas sanas y enfermas de EA para buscar los métodos y parámetros más adecuado para la optimización de estas BNs. Por tanto, es evidente de que existen componentes de la TPC cuyos valores no tienen ninguna influencia sobre el AUC, ya que al no existir una muestra representativa de estos segmentos edad, sexo, nivel educativo y salud cognitiva, el algoritmo puede asignar valores arbitrarios a estos parámetros. El problema es que ante nuevos casos, no se puede garantizar la fiabilidad del resultado.

En la sección 2 de este capítulo se han empleado todos los casos del corpus, es decir, 42 casos correspondientes a personas sanas y 39 correspondientes a personas cognitivamente sanas. Este enfoque presenta dos problemas, en algunos segmentos (edad, sexo, nivel educativo y salud cognitiva) no hay una muestra significativa de la población que propicie que estos valores se optimicen. Otro problema es que para optimizar la TPC no existe un conjunto de casos para la validación del método lo suficientemente amplio como para comprobar la no existencia del sobreajuste.

En esta sección se demuestra que la técnica es aplicable al método de diagnóstico que proponemos y para ello se utiliza el modelo 3 de BN Discreta *"Inferencia por razonamiento abductivo y estudio del deterioro semántico diferencial entre los dominios SV y SNV"*. En la Figura 52 se representa un fragmento del modelo causal de esta BN.

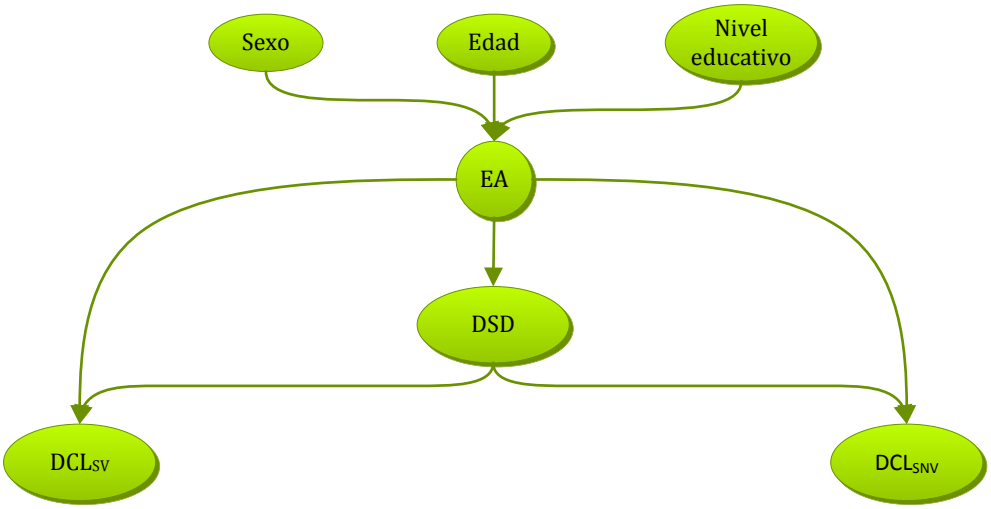


Figura 52.- Fragmento de la BN Discreta.

Debido a los enlaces causales de este modelo de BN, la TPC de la variable ***EA*** tiene tan sólo 36 componentes. Esta reducción de los componentes de la TPC respecto se debe al modelado de los enlaces causales. El procedimiento que se ha seguido para optimizar esta TPC es:

- Se selecciona de forma aleatoria 30 casos sanos y 30 casos de personas con la EA.
- Se optimiza la TPC según estos 60 casos seleccionados de forma aleatoria.
- Se valida el rendimiento de la BN optimizada, con 42 casos sanos y 39 casos enfermos, es decir, se valida el modelo con 21 casos más de los utilizados en el proceso de optimización.
- Se repite el proceso 5 veces para contrastar resultados.

En la Figura 53 se muestran las curvas ROC de las 5 repeticiones del proceso de optimización y la curva ROC del mismo modelo de BN sin optimizar. Se puede observar en la Figura 53 como se mejora el AUC en todas las repeticiones del experimento.

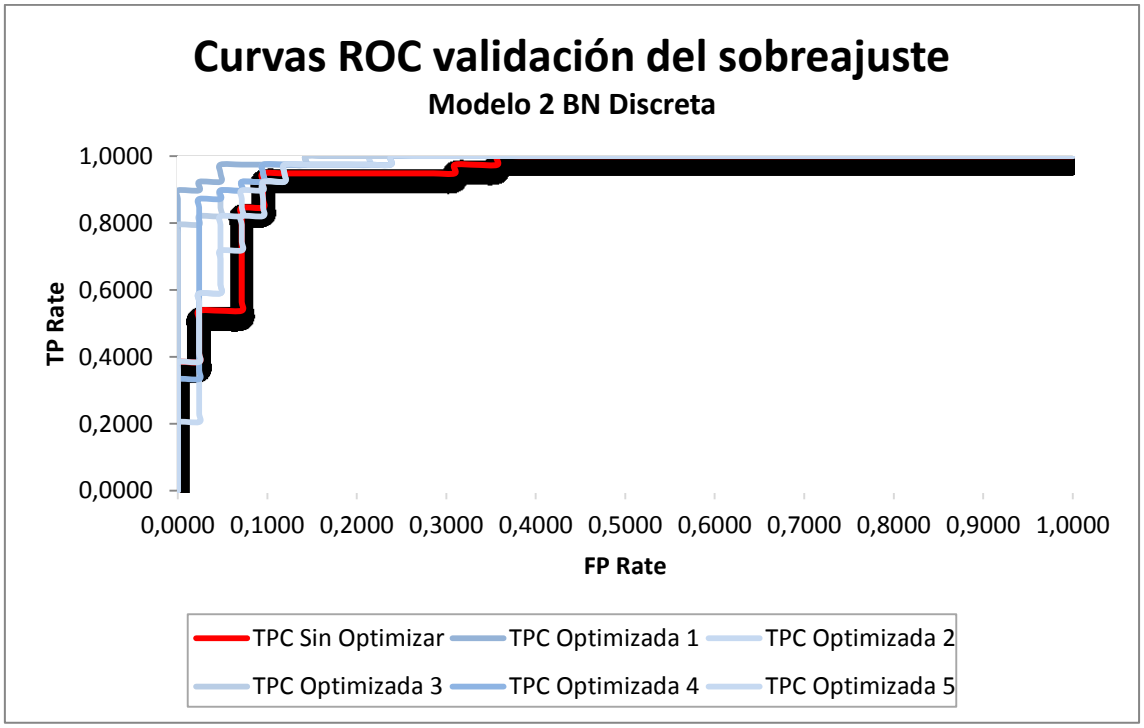


Figura 53.- Curvas ROC validación del sobreajuste

La curva roja de la Figura 53 presenta más puntos por debajo del sureste del resto de las curvas. Esto nos indica que cuando optimizamos las TPC, la BN tiene un rendimiento mejor. En la Tabla 38 se muestran las métricas de rendimiento de cada una de las repeticiones del proceso de optimización. En la cabecera de la tabla se enumeran las distintas ejecuciones del algoritmo (BN1, BN2,.., BN5).

Tabla 38.- Métricas de rendimiento para la BN discreta para analizar el sobreajuste

	BN sin optimizar	BN 1	BN 2	BN 3	BN 4	BN 5
True Positive (TP)	37	38	38	38	38	38
True Negative (TN)	38	39	37	38	38	36
False Positive (FP)	4	3	5	4	4	6
False Negative (FN)	2	1	1	1	1	1
TP rate	0,9487	0,9744	0,9744	0,9744	0,9744	0,9744
FP rate	0,0952	0,0714	0,119	0,0952	0,0952	0,1429
Precisión	0,9024	0,9268	0,8837	0,9048	0,9048	0,8636
Exactitud	0,9259	0,9506	0,9259	0,9383	0,9383	0,9136
Mean Squared Error	0,2312	0,0598	0,0763	0,0647	0,0736	0,0731
Root Mean Squared Error	0,4808	0,2446	0,2762	0,2544	0,2712	0,2704
AUC	0,9475	0,9933	0,956	0,9853	0,9731	0,964

En la Tabla 38 se pone de manifiesto que en todas las iteraciones del proceso de optimización se mejora el AUC. Otros datos interesantes de la tabla anterior son: *Mean Squared Error* y *Root Mean Squared Error*, que indican el error cuadrático medio. Se puede observar como las BNs optimizadas reducen este error.

De estos experimentos se puede destacar que los métodos de penalización del *fitness* que mejor comportamiento han tenido con estas BNs son: *sin penalización* y *penalización dinámica.* Respecto al método de mutación, el métodos que mejor se ha comportado en este problema es *Uncorrelated Mutation with One Step Size*. Y por último respecto al método de control de población, con el método *PRoFIGA* es el más eficiente.

De esta sección se puede deducir que las estrategias evolutivas mejoran el rendimiento de nuestras BNs. Se ha comprobado, aunque no de forma exhaustiva, que no existe sobreajuste de las probabilidades condicionales al conjunto de validación.

Las estrategias evolutivas podrían abaratar costes en la adquisición de nuevos casos, ya que este método es aplicable al muestreo incidental en la adquisición de nuevos casos. No obstante, es muy importante contar con una muestra estratificada en función de los segmentos de edad y nivel educativo.

Capítulo

Otras técnicas de minería de datos

11

El método de diagnóstico que proponemos utiliza un clasificador probabilístico, entre otras técnicas de IA. Es posible utilizar otros algoritmos de minería de datos como alternativa a nuestro método de diagnóstico. En este capítulo se analizan varios algoritmos de minería de datos para el diagnóstico de la EA y otras patologías asociadas.

La estructura del capítulo es la siguiente. En la sección 11.1 se describen las técnicas de minería de datos seleccionadas en el descubrimiento de patrones de interacción entre la EA y la producción oral de rasgos semánticos. En la sección 11.2 se detalla la fase de transformación de los datos, la cual consta de dos métodos (selección de atributos y discretización) para hacer más favorable el proceso de aprendizaje automático. En esta sección se descartan otras técnicas que suelen aplicarse en la fase de transformación de datos de los proyectos de minería de datos, estas técnicas son: el muestreo, la proyección de datos y la limpieza de datos [40]. En las secciones 11.3, 11.4 y 11.5 se realizan experimentos utilizando los algoritmos de aprendizaje automático: *C4.5* (arboles de decisión), *Naive Bayes* y *k-Means* (análisis de clúster). Se han seleccionado algoritmos de aprendizaje automático que generan reglas legibles por las personas para permitir analizar e interpretar los resultados.

11.1 Introducción.

La minería de datos es una tecnología relativamente nueva pero con una gran proyección de futuro. Son muchísimos los casos de éxito de estas técnicas en problemas complejos y con gran cantidad de datos. Estas metodologías y herramientas ayudan a analizar, comprender y extraer conocimiento a partir de grandes cantidades de información.

En este capítulo se utiliza *Weka workbench* que es una colección de algoritmos de aprendizaje automático del estado del arte y herramientas de procesamiento de información. *Weka* ofrece un amplio soporte para todo el proceso de minería de datos experimental, incluyendo la preparación de los datos de entrada, la evaluación estadísticamente de los esquemas de aprendizaje y la visualización de los resultados de salida [40]. También incluye una amplia variedad de herramientas de preprocesamiento.

Weka incluye métodos para los principales problemas de minería de datos como regresión, clasificación, clustering, reglas de asociación y selección de atributos.

Weka fue desarrollado por la Universidad de Waikato en Nueva Zelanda, está escrito en Java y distribuido bajo el término de *GNU General Public License*. Uno de los principales patrocinadores de *Weka* es la empresa Pentaho.

En este capítulo se seleccionan un conjunto de algoritmos de minería de datos con el propósito de extraer patrones interpretables de la producción oral de rasgos semánticos, para el diagnóstico de las posibles alteraciones cognitivas que afectan a la memoria semántica en sus aspectos declarativos. Además, los algoritmos seleccionados en este capítulo fueron identificados como top 10 por los organizadores del *International Data Mining Conference* [53], en la Tabla 39 se muestra el resultado de esta conferencia.

Tabla 39.- Top 10 algoritmos en minería de datos.

	Algoritmo	Categoría
1	C4.5	Classification
2	k-means	Clustering
3	SVM	Statistical learning
4	Apriori	Association analysis
5	EM	Statistical learning
6	PageRank	Link mining
7	Adaboost	Ensemble learning
8	kNN	Classification
9	Naïve Bayes	Classification
10	CART	Classification

A lo largo de esta tesis doctoral se ha comprobado como se mejora la eficacia del método de diagnóstico al segmentar las definiciones orales en los once bloques conceptuales que propone el corpus de Peraita y Grasso [1]. En este capítulo, todos los experimentos se realizan a partir de las variables del corpus, es decir, segmentando la las definiciones orales en estos bloques conceptuales. El objetivo es realizar una comparativa de los algoritmos seleccionados en este capítulo, con el método de diagnóstico propuesto en esta tesis doctoral, para poder así poner de relieve las mejoras introducidas a nuestro método de diagnóstico.

Aunque existen metodologías como *Cross Industry Standard Process for Data Mining* (CRISP-DM), en este capítulo no se considera necesario seguir ninguna metodología para realizar esta comparativa porque el objetivo y los requisitos están muy bien definidos. Además, se puede prescindir de algunas de las fases y tareas que propone la metodología que son innecesarias para alcanzar nuestros objetivos.

CRISP-DM fue concebido a finales de 1996 por Daimler Chrysler, los cuales ya disponían de bastante experiencia en la aplicación de minería de datos en organizaciones industriales y comerciales. SPSS que ha estado proporcionando

servicios basados en minería de datos desde 1990 y había lanzado su primer *workbench* comercial de minería de datos en 1994, denominado *Clementine*. *NCR* es un producto que nace como parte estratégica para proporcionar un valor añadido a sus clientes de su producto *Teradata data warehouse*. La metodología de minería de datos *CRISP-DM* esta descrita en términos de un modelo de proceso jerárquico, consistente en un conjunto de tareas descritas en cuatro niveles de abstracción (de general a específico):

- **Fases:** el proceso está organizado en un número determinado de fases, cada fase está compuesta por varias tareas genéricas de segundo nivel.
- **Tareas genéricas:** tiene la intención de generalizar bastante para cubrir todas las posibles situaciones de la minería de datos.
- **Tareas especializadas:** describen como las acciones en las tareas genéricas deberían ser llevadas a cabo en ciertas situaciones específicas.
- **Instancia de proceso:** es un registro de las acciones, decisiones y de los resultados del proceso de minería de datos actual.

Por otro lado, en este capítulo se utilizan las mismas métricas de rendimiento que las utilizadas en capítulos anteriores. Para evitar el sobreajuste se utiliza cross-validation, repitiendo el proceso 10 veces para obtener estimadores más fiables.

Los experimentos de este capítulo se realizan con la herramienta *Experimenter* de *Weka*, que permite llevar a cabo experimentos automatizados, los cuales ejecutan distintos algoritmos de aprendizaje automático, distintos ajustes de sus parámetros, colección de estadísticas de rendimiento y examen de significación sobre los resultados.

11.2 Transformación de datos.

La fase de transformación de datos en el proceso de minería de datos, puede examinar distintas formas en los que la entrada de datos puede ser manipulada para conseguir métodos de aprendizaje más eficaces. A menudo la transformación de datos consta de las siguientes tareas: selección de atributos, discretización de atributos, proyección de datos, muestreo, limpieza de datos y conversión de problemas multiclases a problemas de dos clases. Para desarrollar este capítulo sólo son necesarios la selección de atributos y la discretización. A continuación se justifica el uso o no, de estas técnicas:

- Selección de atributos. Consiste en descartar del corpus aquellos atributos que son claramente irrelevantes o redundantes. Tal y como se ha indicado en el capítulo 3, existen bloques conceptuales que son comunes a todas las categorías semánticas y otros que no son comunes. Sin embargo, en nuestro método de diagnóstico no se ha descartado ninguna variable. La relevancia o irrelevancia de los atributos se han representado a través de las probabilidades condicionales.
- Discretización de atributos numéricos. Es absolutamente esencial si en el esquema de aprendizaje sólo se puede trabajar con datos categóricos. Todos los clasificadores que utilizamos en este capítulo pueden trabajar con atributos numéricos.

- Proyección de datos. Permite añadir nuevos atributos sintéticos o variables intermedias al modelo para facilitar el aprendizaje automático. La transformación de datos se ha venido utilizando a lo largo de la tesis doctoral, aunque no se ha eliminado o sustituido ninguna variable del corpus, sino al contrario, se han añadido variables latentes o intermedias al modelo causal.
- Muestreo. Cuando el proyecto de minería de datos utiliza grandes cantidades de información, puede dar lugar a una reducción del corpus del conjunto de instancias mediante selección aleatoria. Esta tarea es innecesaria en estos experimentos, ya que el corpus lingüístico está formado por sólo 81 casos.
- Conversión de problemas multiclases a problemas de dos clases. Aunque el corpus lingüístico distingue entre personas cognitivamente sanas, enfermos de EA en fase leve y en enfermos de EA en fase moderada, se han simplificado las clases a $\boldsymbol{EA_{presente}}$ y $\boldsymbol{EA_{ausente}}$.

En esta sección se van a evaluar varios algoritmos evaluadores de atributos, algunos de estos algoritmos no hacen uso de algoritmos de inducción.

11.2.1 Algoritmo CfsSubsetEval.

CfsSubsetEval evalúa la capacidad predictiva de cada atributo individual y el grado de redundancia entre ellos, prefiriendo los atributos que estén altamente correlacionados con la clase, siempre que en el conjunto de atributos no contenga otro atributo cuya correlación con el atributo en cuestión sea mayor. Este algoritmo se utiliza en combinación con el método de búsqueda *Best-First*, ya que es menos probable que este algoritmo se atasque con mínimos locales en comparación con otros algoritmos. La búsqueda *Best-First* es un método que no termina cuando el rendimiento empieza a disminuir y además, mantiene una lista de todos los subconjuntos de atributos evaluados hasta el momento ordenados según su rendimiento, de modo que puede volver a consultar configuraciones anteriores. *Best-First* puede comenzar con un conjunto vacío de atributos o por el contrario, comenzar con el conjunto de atributos de la muestra y buscar hacía atrás.

En la Tabla 40 se pueden observar los resultados del algoritmo utilizando todo el conjunto de entrenamiento. Los atributos seleccionados según este algoritmo son:

- Nivel educativo
- De la categoría natural manzana: tipos, funcional, evaluativo y otros.
- De la categoría natural perro: otros.
- De la categoría natural ***pino***: *tipos*, *evaluativo* y *lugar*.
- Del objeto básico ***coche***: *taxonómico* y *tipos*

Tabla 40.- Resultado de la selección de atributos del algoritmo CfsSubsetEval

```
=== Run information ===

Evaluator:    weka.attributeSelection.CfsSubsetEval

Search:weka.attributeSelection.BestFirst -D 1 -N 5

Relation:     CORPUS
```

Evaluation mode:evaluate on all training data

=== Attribute Selection on all input data ===

Search Method:

Best first.

Start set: no attributes

Search direction: forward

Stale search after 5 node expansions

Total number of subsets evaluated: 1194

Merit of best subset found: 0.538

Attribute Subset Evaluator (supervised, Class (nominal): 70 leveldisease):

CFS Subset Evaluator

Including locally predictive attributes

Selected attributes: 2,5,7,8,14,25,27,30,31,37,38,47,48,58,60 : 15

Educationallevel, mantip, manfun, maneva, manotr, perotr, pintip, pineva, pinlug, coctax, coctip, cocotr, siltax, silotr, pantip

Weka da la posibilidad de utilizar cross-validation en la selección de atributos, en la Tabla 41 se muestran los resultados obtenidos con este algoritmo. El experimento se repite tres veces y en cada ejecución se indica el número de veces que el atributo se selecciona por cada fold, así como el porcentaje de veces que el atributo ha sido seleccionado. Se puede comprobar en la Tabla 41 que en las distintas ejecuciones del algoritmo se seleccionan un conjunto de atributos comunes. Los atributos no comunes a todas las ejecuciones se han marcado de color rojo. Por simplicidad, los atributos se han representado con un alias de tal forma que los tres primeros caracteres corresponden a la categoría semántica y los tres últimos al bloque conceptual. Para las categorías semánticas los alias utilizados son: ***man*** (manzana), ***per*** (perro), ***pin*** (pino), ***coc*** (coche), ***sil*** (silla) y ***pan*** (pantalón). Para los bloques semánticos los alias utilizados son: *tax* (taxonómico), *tip* (tipo), *par* (parte-todo), *fun* (funcional), *eva* (evaluativo), *lug* (lugar y hábitat), *con* (conducta), *cau* (causa/genera), *pro* (procedimental) y *cic* (ciclo vital y otros).

Tabla 41.- Resultados de CfsSubsetEval para la selección de atributos.

=== Run information ===					
Evaluator: weka.attributeSelection.CfsSubsetEval					
Search:weka.attributeSelection.BestFirst -D 1 -N 5					
Evaluation mode:10-fold cross-validation					
Ejecución 1		Ejecución 2		Ejecución 3	
number of folds (%)	attribute	number of folds (%)	attribute	number of folds (%)	attribute
10(100 %)	2 el	10(100 %)	2 el	10(100 %)	2 el
1(10 %)	4 mantax	1(10 %)	4 mantax	1(10 %)	4 mantax
10(100 %)	5 mantip	10(100 %)	5 mantip	10(100 %)	5 mantip
10(100 %)	7 manfun	10(100 %)	7 manfun	8(80 %)	7 manfun
5(50 %)	8 maneva	6(60 %)	8 maneva	6(60 %)	8 maneva
1(10 %)	12 manpro	1(10 %)	12 manpro	1(10 %)	9 manlug
3(30 %)	13 mancic	3(30 %)	13 mancic	1(10 %)	12 manpro
5(50 %)	14 manotr	5(50 %)	14 manotr	4(40 %)	13 mancic
1(10 %)	15 pertax	2(20 %)	15 pertax	7(70 %)	14 manotr
1(10 %)	18 perfun	1(10 %)	18 perfun	2(20 %)	15 pertax
6(60 %)	19 pereva	4(40 %)	19 pereva	1(10 %)	16 pertip
1(10 %)	20 perlug	1(10 %)	21 percon	5(50 %)	19 pereva
10(100 %)	25 perotr	10(100 %)	25 perotr	1(10 %)	21 percon
5(50 %)	26 pintax	4(40 %)	26 pintax	10(100 %)	25 perotr
9(90 %)	27 pintip	8(80 %)	27 pintip	4(40 %)	26 pintax
3(30 %)	29 pinfun	2(20 %)	29 pinfun	9(90 %)	27 pintip
10(100 %)	30 pineva	10(100 %)	30 pineva	2(20 %)	29 pinfun
10(100 %)	31 pinlug	10(100 %)	31 pinlug	10(100 %)	30 pineva
2(20 %)	36 pinotr	2(20 %)	36 pinotr	10(100 %)	31 pinlug
10(100 %)	37 coctax	10(100 %)	37 coctax	2(20 %)	36 pinotr
10(100 %)	38 coctip	9(90 %)	38 coctip	10(100 %)	37 coctax
1(10 %)	39 cocpar	1(10 %)	41 coceva	10(100 %)	38 coctip
10(100 %)	47 cocotr	9(90 %)	47 cocotr	1(10 %)	41 coceva
10(100 %)	48 siltax	10(100 %)	48 siltax	10(100 %)	47 cocotr
10(100 %)	58 silotr	1(10 %)	49 siltip	10(100 %)	48 siltax
10(100 %)	60 pantip	10(100 %)	58 silotr	2(20 %)	49 siltip
1(10 %)	64 panlug	10(100 %)	60 pantip	10(100 %)	58 silotr
		1(10 %)	62 panfun	1(10 %)	59 pantax
		1(10 %)	64 panlug	10(100 %)	60 pantip
				1(10 %)	64 panlug

11.2.2 Algoritmo Relief (Recursive Elimination of Features).

El algoritmo *Relief* asigna un peso a cada atributo que indica el nivel de relevancia de la característica con respecto al concepto a deducir. *Relief* es un algoritmo aleatorio, es decir, las instancias del corpus se seleccionan de forma aleatoria del conjunto de entrenamiento. Por cada atributo del corpus de datos se actualizan unos indicadores de relevancia basándose en la diferencia entre las instancias seleccionadas y, las *n* instancias más cercanas de la misma clase y la clase opuesta. *Relief* puede operar sobre tipos de datos discretos y continuos. En dominios reales, *Relief* da una alta correlación con la clase pero sin eliminar muchos de los atributos relevantes en sentido débil.

El método de búsqueda utilizado es *Ranker*, que devuelve una lista ordenada de los atributos según su calidad.

Se utilizan los parámetros por defecto del algoritmo:

- **numNeighbours**: Número de vecinos, se deja por defecto 10.
- **sampleSize**: Número de instancias del ejemplo. Por defecto -1 indica que se usarán todas las instancias.
- **Seed**: Semilla para la selección aleatoria de los ejemplos
- **Sigma**: Determina la influencia de los vecinos más cercanos.
- **weightByDistance**: Pesos de los vecinos más cercanos derivados de su distancia.

Para el método de búsqueda *Ranker* se establece el *threshold* a 0.

En la Tabla 42 se representan los resultados obtenidos con el algoritmo *Relief*, utilizando el método de búsqueda *Ranker*. En esta tabla se han eliminado los atributos clasificados por el algoritmo como irrelevantes en sentido fuerte. Los atributos están ordenados en función del grado relevancia. La selección de atributos se ha realizado con *cross-validation nfold* 10. Se pueden observar tres columnas, la primera columna *average merit,* es el promedio –por cada fold— del grado de relevancia junto con la desviación típica; la segunda columna *average rank,* es el promedio del ranking obtenido por fold, y *attribute*, que es el nombre del atributo. Es decir, el algoritmo da como resultado un ranking de atributos. Al utilizar cross-validation se realiza un promedio de este ranking y se calcula la desviación típica. Por ejemplo, para el atributo ***perpar*** –categoría semántica ***perro*** y bloque semántico *parte*— el promedio en el ranking ha sido 29 y su desviación típica es de 5.9.

Tabla 42.- Resultados de Relief para la selección de atributos.

average merit	average rank	attribute	average merit	average rank	attribute
0.162 +- 0.032	1.2 +- 0.4	ed.level	0.021 +- 0.008	27.4 +- 6.79	pinotr
0.141 +- 0.016	1.8 +- 0.4	mantip	0.02 +- 0.006	27.4 +- 5.31	pertip
0.097 +- 0.012	3.8 +- 0.87	pantip	0.021 +- 0.01	27.7 +- 8.67	paneva
0.085 +- 0.014	5.3 +- 2.05	perotr	0.018 +- 0.006	29.1 +- 5.96	perpar
0.086 +- 0.013	5.5 +- 1.63	coctax	0.019 +- 0.007	29.3 +- 7.99	maneva
0.08 +- 0.012	6.2 +- 2.09	cocotr	0.019 +- 0.011	30.6 +-10.85	panotr

average merit	average rank	attribute	average merit	average rank	attribute
0.077 +- 0.018	6.9 +- 2.55	coctip	0.017 +- 0.006	30.8 +- 5.65	pantax
0.064 +- 0.018	9.5 +- 3.41	mantax	0.017 +- 0.004	30.8 +- 4.69	pineva
0.061 +- 0.009	9.8 +- 1.89	manfun	0.018 +- 0.005	31 +- 4.86	perfun
0.057 +- 0.014	10.6 +- 3.14	pintax	0.017 +- 0.009	31.7 +- 7.56	panfun
0.058 +- 0.013	11 +- 3.52	pintip	0.016 +- 0.005	32.3 +- 5.08	cocpar
0.057 +- 0.013	11.6 +- 2.29	siltax	0.015 +- 0.005	33 +- 5.98	pincau
0.047 +- 0.011	13.7 +- 3.1	pinfun	0.013 +- 0.006	33.8 +- 7.05	pinlug
0.046 +- 0.007	13.7 +- 1.95	siltip	0.013 +- 0.013	36.9 +-10.65	silotr
0.04 +- 0.014	16.6 +- 6.15	manotr	0.011 +- 0.005	37.6 +- 7.59	percon
0.037 +- 0.012	17.6 +- 4.05	sileva	0.01 +- 0.009	38.1 +-10.01	perlug
0.032 +- 0.003	18.4 +- 1.8	silfun	0.01 +- 0.006	39.2 +- 8.21	mancau
0.033 +- 0.009	19.7 +- 6.81	cocfun	0.01 +- 0.006	39.2 +- 7.92	panlug
0.032 +- 0.006	19.7 +- 3.41	panpar	0.009 +- 0.004	39.4 +- 3.88	silpro
0.028 +- 0.007	21 +- 4.02	pinpar	0.005 +- 0.007	44.7 +-10.12	percic
0.032 +- 0.018	21.1 +- 7.63	pertax	0.004 +- 0.005	45.7 +- 8.33	sillug
0.024 +- 0.009	24.9 +- 8.49	Coccau	0.003 +- 0.005	46.5 +- 6.52	manlug
0.021 +- 0.006	27.2 +- 5.79	Pereva	0.002 +- 0.003	49 +- 6.08	pancic

11.2.3 Algoritmo ConsistencySubsetEval.

El algoritmo *ConsistencySubsetEval* evalúa un conjunto de atributos por el grado de consistencia respecto a los valores de la clase, cuando las instancias de entrenamiento son proyectadas sobre el conjunto.

La consistencia de cualquier subconjunto de atributos nunca puede ser inferior a la de todo el conjunto completo, por tanto, es normal usar este subconjunto evaluador en combinación con un algoritmo de búsqueda aleatoria o exhaustiva con el propósito de buscar el subconjunto más pequeño posible, cuya consistencia sea la misma para todo el conjunto completo de atributos.

El algoritmo de búsqueda utilizado es *GreedyStepwise*, el cual, puede ser forward o backward a través del espacio de atributos. El algoritmo puede comenzar por un conjunto vacío de atributos o por conjunto constituido por todos los atributos del corpus de datos y se detiene cuando la adición o eliminación de algunos atributos, tiene como resultado una disminución en la evaluación. Este algoritmo sólo funciona con clases nominales.

En la Tabla 43 se detallan los resultados obtenidos con el algoritmo *ConsistencySubsetEval* para la selección de atributos. Al igual que en la Tabla 41, en la primera columna se indica el número de veces que el atributo se selecciona por cada fold, así como el porcentaje de veces que el atributo ha sido seleccionado.

Tabla 43.- Resultados de ConsistencySubsetEval para la selección de atributos.

number of folds (%)	attribute
7(70 %)	educationallevel
5(50 %)	mantax
8(80 %)	mantip
6(60 %)	manfun
5(50 %)	maneva
4(40 %)	mancic
2(20 %)	manotr
1(10 %)	pertax
4(40 %)	pereva
1(10 %)	perlug
8(80 %)	perotr
2(20 %)	pintax
3(30 %)	pintip
1(10 %)	pinfun
2(20 %)	pineva
1(10 %)	pinlug
1(10 %)	coctax
6(60 %)	coctip
1(10 %)	cocotr
2(20 %)	silotr
1(10 %)	pantip

11.3 Arboles de decisión.

Los árboles de decisión es uno de los métodos prácticos, dentro del aprendizaje automático inductivo, más ampliamente utilizados en minería de datos. Este algoritmo deriva del algoritmo simple divide y conquistarás. Para este experimento utilizamos el algoritmo C4.5 que puede trabajar con atributos numéricos. Para consultar más detalles del algoritmo véase los capítulos 3 y 6, o en [49].

Los árboles de decisión completamente expandidos con frecuencia contienen estructuras innecesarias. Durante la construcción del árbol se pueden adoptar dos estrategias en relación a la poda que son: *postpoda* y *prepoda*. La *prepoda* intenta decidir durante la construcción del árbol cuando parar en la construcción de subárboles –esta estrategia evita el trabajo de construir subárboles para luego desecharlos. La *postpoda* parece que tiene algunas ventajas respecto a la *prepoda*, como por ejemplo en las situaciones en las que dos atributos individuales no tengan valor predictivo de forma individual, pero si lo tienen cuando se combinan. Se pueden considerar diferentes operaciones en la post poda como *subtree replacement* y *subtree raising*. Por cada nodo, el esquema de aprendizaje podría decidir si debería realizar un remplazo del sub-árbol, elevar el subárbol o dejar el

subárbol tal y como está, sin podar. La idea es seleccionar algunos sub-árboles y remplazarlos con hojas simples.

Los parámetros que se han considerado para este algoritmo son:

- **Confidence Factor o exactitud**: Es un factor que establece el threshold para la medida del ratio de ganancia de información. Para este experimento se utiliza el valor por defecto 0,25.
- **Binary Splits**: Indica si las instancias utilizan atributos nominales cuando se construye el árbol.
- **minNumObj**: Número mínimo de instancias por hoja. Para este experimento se utilizará el valor por defecto de *Weka* que es 2.
- **numFolds:** Determina el tamaño del conjunto de poda. Los datos se dividen de forma equitativa en un número de folds y el último folds es usado para la poda. Este parámetro se optimizará con *Weka Experimenter*.
- **reducedErrorPruning:** Indica si se usa un error reducido para la poda en lugar de la poda de C4.5. Este parámetro se optimizará con *Weka Experimenter*.
- **subtreeRaising**: Indica si se considera la operación *subtree raising* cuando se poda. Se utiliza el valor por defecto *true*.
- **Unpruned**: Indica si se realiza la poda. Para este experimento se utiliza el valor por defecto *false.*
- **useLaplace**: Indica si el algoritmo debe contar las hojas basado en suavizado de Laplace. Para este experimento se establece el valor a *true.*

Para este experimento se utilizará *Weka Experimenter* para comprobar la eficacia del clasificador siguiendo el esquema de la Figura 54:

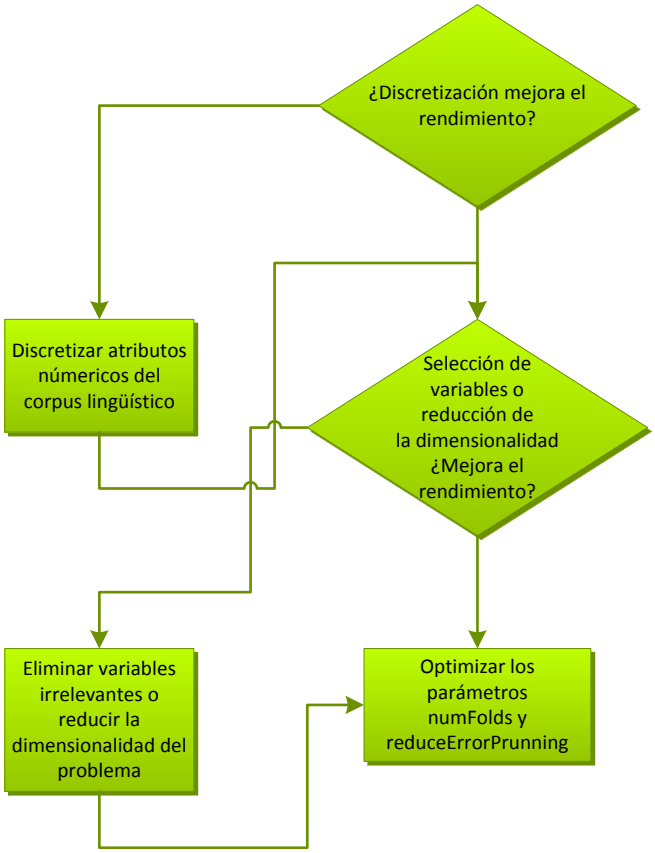


Figura 54.- Esquema de aprendizaje para el algoritmo J48.

En la Tabla 44 se muestran métricas de rendimiento obtenidas con el algoritmo *J48*, utilizando distintos algoritmos para la selección de atributos. En la primera columna se indica el algoritmo utilizado para la selección de atributos y en la segunda columna los porcentajes de instancias bien clasificadas. El resto de métricas son iguales a las que se han venido utilizando en esta tesis doctoral.

Tabla 44.- Resultados obtenidos con los árboles de decisión J48.

Key: trees.J48 '-C 0.25 -M 2 -A' -217733168393644444						
Dataset	**% correcto**	**TPR**	**FPR**	**Precisión**	**Exactitud**	**AUC**
ConsistencySubsetEval	75.60	0.80	0.29	0.78	0.80	0.82
CsfSubsetEval	80.07	0.79	0.19	0.84	0.79	0.83
Original	76.43	0.77	0.23	0.81	0.77	0.79
Relief	76.82	0.77	0.22	0.81	0.77	0.80

Se puede observar que J48 funciona mejor con el algoritmo para selección de atributos *CsfSubsetEval*. Los parámetros utilizados para J48 son: *(confidence 0,25)*, *(minNumObj 2)* y *(Laplace true)*.

Los parámetros de rendimiento de la Tabla 44 se obtuvieron con *datasets* que contienen atributos numéricos. Sin embargo, los árboles de decisión también permiten trabajar con valores cualitativos. En la Tabla 45 se puede comprobar algunas de las métricas

obtenidas con el algoritmo J48, discretizando los atributos numéricos y aplicando la selección de atributos. Los resultados mejoran sensiblemente al discretizar los atributos.

Tabla 45.- Resultado obtenido con los árboles de decisión J48 y datasets discretos.

Key: trees.J48 '-C 0.25 -M 2 -A' -217733168393644444						
Dataset	% correctos	TPR	FPR	Precisión	Exactitud	AUC
ConsistencySubsetEval	82.97	0.76	0.10	0.90	0.76	0.91
CsfSubsetEval	83.97	0.83	0.15	0.87	0.83	0.89
Original	82.36	0.83	0.18	0.84	0.83	0.88
Relief	81.29	0.81	0.18	0.83	0.81	0.87

La Tabla 44 y Tabla 45 se han generado con *Weka Experimenter* con varios *datasets* obtenidos con distintos algoritmos para la selección de atributos y distintas configuraciones de J48.

Tabla 46.- Mejor configuración encontrada con el corpus lingüístico para el algoritmo J48.

Criterio / Parámetro	**Valores**
Tipos de atributos	Discretos
Atributos seleccionados con ConsistencySubsetEval	Nivel educativo. Manzana (taxonómico, tipos, funcional, ciclo vital). Perro (otros) Pino (taxonómico, funcional)
numFolds	5
reduceErrorPruning	True
subtreeRaising	True
Laplace	False

En la Tabla 47 se muestran las métricas de rendimiento obtenidas con el algoritmo J48.

Tabla 47.- Métricas de rendimiento obtenidas con el algoritmo J48.

J48	
True Positive (TP)	35
True Negative (TN)	33
False Positive (FP)	9
False Negative (FN)	4
TP rate	0,84
FP rate	0,156
Precisión	0,845
Exactitud	0,84
Mean Squared Error	0,2151
Root Mean Squared Error	0,3402
AUC	0,91

Si contrastamos las métricas de la Tabla 47 con las obtenidas en los experimentos de los capítulos 7 y 8, se puede determinar que nuestro método de diagnóstico es más eficaz que el algoritmo J48 en el diagnóstico de la EA.

En la Figura 55 se muestra el árbol de decisión inferido con el algoritmo de inducción J48. Los rombos representan los nodos de decisión –antecedentes—y los rectángulos los consecuentes. En los consecuentes se muestran (valor entre paréntesis) el número de casos afectados por la regla y el número de excepciones de la regla. Cabe destacar que el algoritmo J48 implícitamente desecha algunos atributos por la propia naturaleza del algoritmo.

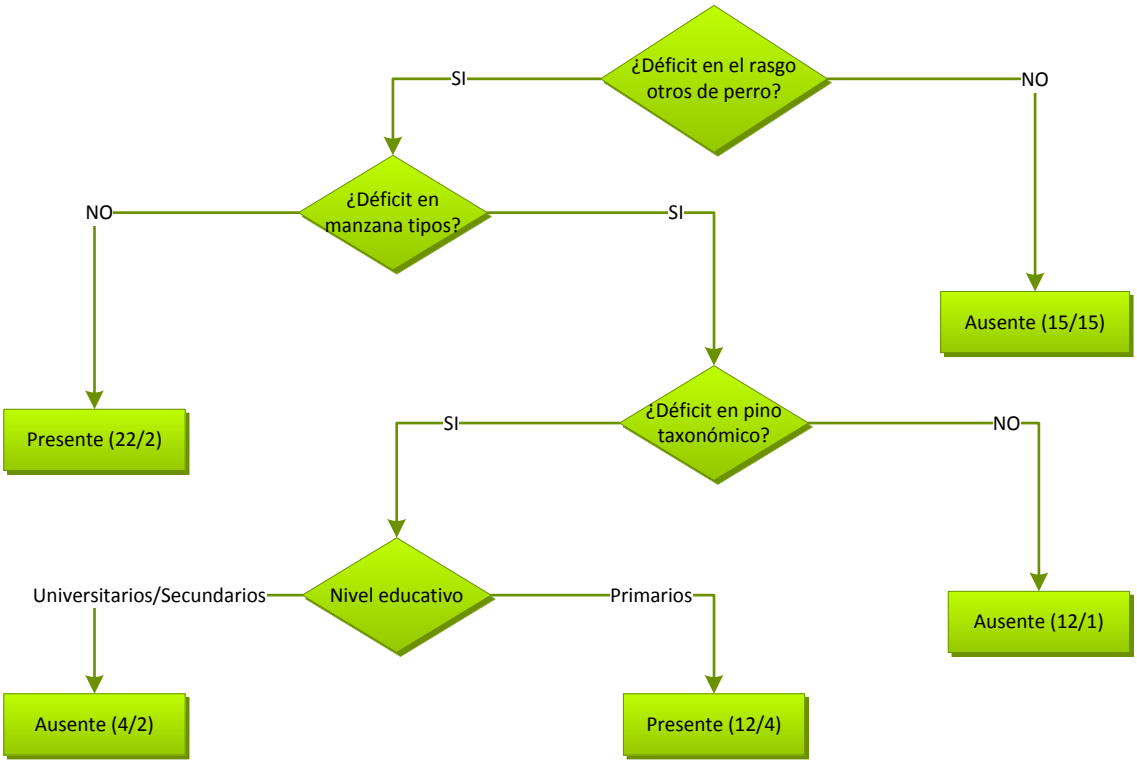


Figura 55.- Árbol de decisión inferido con J48

Según este árbol de decisión, una persona con la EA en sus primeros estadios, que conserve intacto la categoría semántica ***perro*** será clasificado como EA ausente, sin tener otras categorías semánticas que pudieran estar deterioradas como consecuencia de un deterioro semántico focalizado.

11.4 Naive Bayes.

Naive Bayes es uno de los modelos gráficos probabilísticos más simples de clasificación [54]. *Naive Bayes* está en la lista de Top 10 de los algoritmos de minería de datos gracias a su alta simplicidad representacional y computacional, sin renunciar a un alto rendimiento en tareas de clasificación. *Naive Bayes* parte de la suposición de la independencia condicional de los atributos dada la clase. En la Figura 56 se representa

un fragmento del modelo probabilístico construido a partir del corpus lingüístico con Naive Bayes.

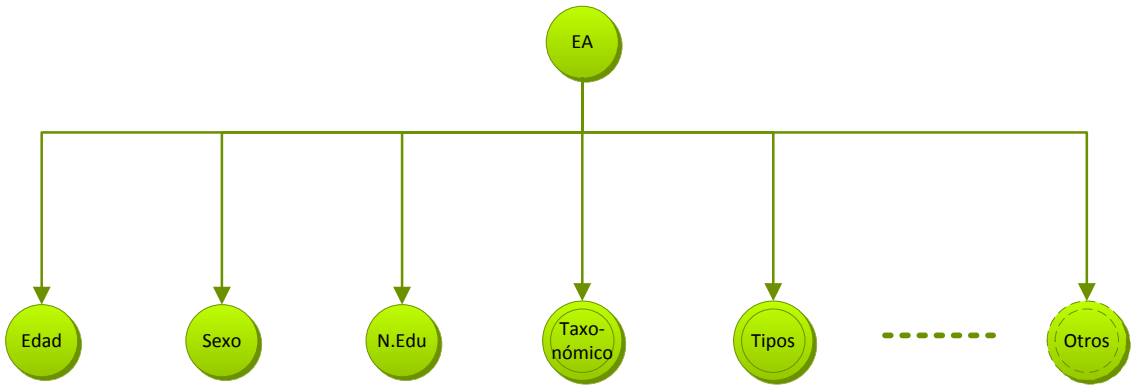


Figura 56.- Fragmento del modelo probabilístico para la detección de EA con Naive Bayes.

El aprendizaje del modelo cuantitativo de esta BN es muy sencillo. La distribución de probabilidad condicional para cada P(edad | EA) se calcula por la frecuencia relativa de cada estado de cada variable según la clase.

$$P(RS_{CAT}|EA) = \frac{N(RS_{CAT}, EA)}{N(EA)} \quad (11.1)$$

donde

RS: rasgo semántico, ∈ {*taxonómico, tipos, parte-todo, funcional, evaluativo, lugar/hábitat, comportamiento, causa/genera, procedimental, ciclo vital, otros*}

CAT: categorías semánticas, ∈ {*manzana, perro, pino, coche, silla pantalón*}

d*om(RS), dom(EA)*: dominio de *RS* y *EA*, ∈ {ausente, presente}

N(X): Recuento de casos que satisface las condiciones *X*.

Este experimento se realiza con *Weka Explorer* y no con *Weka Experimenter,* debido a que el algoritmo no requiere, dada su simplicidad, de una configuración muy precisa. El algoritmo requiere de atributos nominales, por lo que los atributos numéricos deben ser discretizado. El algoritmo *Naive Bayes* de *Weka* permite realizar una discretización supervisada o no supervisada, en este experimento vamos a establecer el parámetro *UseSupervisedDiscretize=true*.

En la Tabla 48 se muestran las métricas de rendimiento obtenidas con el algoritmo *Naive Bayes*. El algoritmo Naive Bayes es una BN cuya clase a predecir es un nodo padre de todos los síntomas y factores de riesgo. El resultado obtenido con este clasificador es bastante bueno, sin embargo, no supera a nuestro método de diagnóstico.

Tabla 48.- Métricas de rendimiento obtenidas con el algoritmo Naive Bayes.

	Naive Bayes
True Positive (TP)	31
True Negative (TN)	39
False Positive (FP)	8
False Negative (FN)	3
TP rate	0,864
FP rate	0,141
Precisión	0,869
Exactitud	0,864
Mean Squared Error	0,1335
Root Mean Squared Error	0,3564
AUC	0,921

11.5 k-Means.

Este algoritmo de aprendizaje no supervisado ha sido utilizado en esta tesis doctoral para la discretización de los atributos numéricos del corpus lingüístico. *k-Means* es de la familia de las técnicas de clustering, las cuales dividen las instancias en grupos naturales sin que exista una clase a predecir. Puede consultar más detalles sobre el algoritmo en el capítulo 3 o en [49,40].

Con este algoritmo no es necesario discretizar los atributos numéricos, pero si es interesante realizar la selección de atributos. En la Tabla 49 se muestran las métricas obtenidas con este algoritmo. Las columnas de la tabla indican el método para la selección de atributos utilizado y las filas indican las métricas de rendimiento. En la Tabla 49 se puede deducir que el mejor resultado de *k-Means++* se obtiene con el algoritmo de selección de atributos *CsfSubsetEval*. El rendimiento de este algoritmo es inferior al de los algoritmos J48 o Naive Bayes.

Tabla 49.- Métricas de rendimiento obtenidas con el algoritmo *k-Means*.

	Corpus Original	Selección de atributos ConsistencySubsetEval	Selección de atributos CsfSubsetEval	Relief
True Positive (TP)	38	38	38	38
True Negative (TN)	27	29	31	27
False Positive (FP)	15	13	11	15
False Negative (FN)	1	1	1	1
TP rate	0,97	0,97	0,97	0,97
FP rate	0,64	0,69	0,74	0,64
Precisión	0,72	0,75	0,78	0,72
Exactitud	0,80	0,83	0,85	0,80

11.6 Discusión.

Aunque los resultados de estos algoritmos son muy buenos, el rendimiento obtenido en todos los casos es inferior al obtenido con el método diagnóstico propuesto en esta tesis doctoral. En la Tabla 50 se realiza una comparativa de las métricas de rendimiento obtenidas con las BNs diseñadas en esta tesis doctoral, aplicando las estrategias evolutivas, versus las obtenidas con otros algoritmos de minería de datos. Se puede comprobar en la Tabla 50 como nuestra propuesta (color rojo) mejora todas las métricas de rendimiento respecto a los algoritmos de minería de datos: J48, Naive Bayes y k-Means. Esta comparativa pone de relieve la eficacia de las mejoras innovadoras que aporta esta tesis doctoral al método de diagnóstico.

Tabla 50.- Comparativa de métricas de las BN utilizadas en esta tesis VS otras técnicas de IA.

	BN Discreta Modelo 3	Pearson. CLG BN	Ganancia de Información. BN Aproximada	J48	Naive Bayes	KMeans
True Positive (TP)	38	36	39	35	31	38
True Negative (TN)	39	41	39	33	39	31
False Positive (FP)	3	1	3	9	8	11
False Negative (FN)	1	3	0	4	3	1
TP rate	0,97	0,92	1,00	0,84	0,86	0,97
FP rate	0,07	0,02	0,07	0,16	0,14	0,74
Precisión	0,93	0,97	0,93	0,85	0,87	0,78
Exactitud	0,95	0,95	0,96	0,84	0,86	0,85
Mean Squared Error	0,06	0,06	0,07	0,22	0,13	
Root Mean Squared Error	0,24	0,25	0,26	0,34	0,36	
AUC	0,99	0,99	0,98	0,91	0,92	

PARTE

CONCLUSIONES IV

Capítulo

Conclusiones y trabajos futuros

12

12.1 Conclusiones

El objetivo de la tesis ha sido encontrar y desarrollar las soluciones que la Inteligencia Artificial ofrece para el diagnóstico temprano de la EA, basándose en el conocimiento adquirido desde una serie de test de definiciones orales. Desde el principio se ha planteado como primera hipótesis que las redes bayesianas son la herramienta más adecuada para ello, utilizando para el aprendizaje de sus parámetros probabilísticos, el análisis, interpretación y segmentación de definiciones orales con restricción temporal de determinadas categorías semánticas, en particular desde el corpus lingüístico utilizado.

Vamos primero a mostrar las conclusiones que hemos podido extraer de las diferentes etapas de la investigación y experimentos realizados con: BNs discretas, BNs híbridas, estrategias evolutivas y otros algoritmos de aprendizaje automático.

Conclusiones a partir del análisis estadístico.

La primera conclusión que se extrae del análisis estadístico es que algunas variables tiene un mayor grado de asociación/correlación, que otras, con la EA. Otro dato que se extrae de este análisis, es que los participantes, tanto enfermos de EA como sanos, produjeron de media más rasgos semánticos en la categoría semántica ***Perro*** que en la categoría semántica ***Manzana***; aunque la desviación típica también es mayor. Respecto a los dominios semánticos SV y SNV, la diferencia en el grado de asociación/correlación es más sutil, aunque se ha conseguido un mejor clasificador al tener en cuenta sólo el dominio semántico SV. Por otro lado, se seleccionaron, con distintos algoritmos de minería de datos, las variables más relevantes para predecir el deterioro de la memoria semántica, y en algunos casos, como por ejemplo con el algoritmo *ConsistencySubsetEval,* sólo se seleccionaron variables del dominio semántico ***SV***, desechando todas las variables del dominio semántico ***SNV***.

Otra conclusión que se puede extraer del análisis estadístico, es que no se encontraron medidas estadísticas que permitieran clasificar con cierta precisión las personas cognitivamente sanas y las personas enfermas de EA, principalmente porque en la

muestra hay mucha dispersión (desviación típica significativa) en la producción oral de rasgos semánticos. Por ello, no ha sido posible diseñar un método de diagnóstico de la EA basado en la estadística, sin embargo, ha sido interesante utilizar la estadística descriptiva para calcular determinados parámetros, que unido a las técnicas de IA, permitieron construir unas BNs híbridas con las que se consiguieron unos resultados muy buenos.

Conclusiones de los experimentos con las BNs discretas.

La primera conclusión que se ha podido extraer de las BNs discretas es la importancia de ejercer influencias informativas sobre las variables del corpus de determinados factores de contexto, como la edad y el nivel educativo. Se propuso una estrategia para ejercer esta influencia informativa durante el proceso de discretización; con un número reducido de casos, sin aumentar la complejidad de los modelos causales y se consiguieron unos resultados excelentes. Esta estrategia innovadora aplicada en el proceso de discretización podría extenderse a otros dominios de aplicación.

Otra conclusión que se ha podido extraer de las BNs discretas es la eficacia del método de diagnóstico que se ha propuesto en esta tesis. El método de diagnóstico se ha validado con varios experimentos respecto al diagnóstico dado por los neurólogos, en los que se utilizaron distintos modelos de BNs, distintas técnicas de modelado y distintos algoritmos de aprendizaje automático; se consiguieron unos resultados muy buenos.

Otra conclusión, y además coincide con las medidas de asociación/correlación del análisis estadístico, es que la categoría semántica ***Manzana*** sirvió para identificar mejor el DS, mientras que con la categoría semántica ***Perro***, se obtuvieron peores resultados. Para llegar a esta conclusión se generaron curvas ROC a partir de las probabilidades a posteriori de las variables intermedias $\boldsymbol{DS_{SV}}$, $\boldsymbol{DS_{SNV}}$, $\boldsymbol{DS_{Manzana}}$, $\boldsymbol{DS_{Perro}}$, $\boldsymbol{DS_{Pino}}$, $\boldsymbol{DS_{Coche}}$, $\boldsymbol{DS_{Silla}}$ y $\boldsymbol{DS_{Pantalón}}$. En algunos enfermos de EA, el déficit léxico-semántico-conceptual de la categoría semántica ***Perro*** es menos significativo que el déficit léxico-semántico-conceptual de la categoría semántica ***Manzana.***

Con la BN discreta del modelo 3 "*Inferencia por razonamiento abductivo y estudio del deterioro semántico diferencial entre SV y SNV"* se consiguió mejorar el diagnóstico de la EA en fase leve de un número reducido de casos, al modelar explícitamente el deterioro semántico diferencial entre los dominios ***SV*** y ***SNV***.

Otra conclusión que se extrae de las BNs discretas, es que no se ha encontrado una relación lineal entre la producción oral de rasgos semánticos y las probabilidades a posteriori inferidas por las BNs. Esto demuestra que tanto la segmentación de las definiciones orales en los once bloques conceptuales que propone el corpus [1], como las mejoras a las técnicas de IA diseñadas en esta tesis doctoral, han mejorado la eficacia del método de diagnóstico, permitiendo encontrar patrones complejos de interacción entre la EA y la producción oral de rasgos semánticos. En esta tesis no se ha podido comprobar la especificidad del DS causados por otras ENs de tipo no-EA.

Conclusiones de los experimentos con las BNs híbridas.

La primera conclusión que se puede extraer de las BNs híbridas es demostrar la mejora que se consigue en el diagnóstico de la EA en fase leve, al segmentar las definiciones orales en bloques conceptuales. Con la segmentación de la producción oral en los once bloques conceptuales que propone el corpus [1] se mejora la eficacia del diagnóstico. Todos los modelos de BN que se proponen en esta tesis doctoral, se han diseñado teniendo en cuenta este posible deterioro de la memoria semántica focalizado para mejorar el diagnóstico de los enfermos de EA en fase leve. Las BNs híbridas han permitido medir la influencia de la segmentación de las definiciones orales en los once bloques conceptuales que propone el corpus [1].

Otra conclusión, que coincide con el análisis estadístico y con las BN discretas, es que la categoría semántica ***Manzana*** sirvió para identificar mejor el DS, mientras que se consiguió un resultado peor con la categoría semántica ***Perro***.

Con las BN Híbridas se construyeron distintas redes de ecuaciones lineales estructurales, se utilizó en cada red un determinado coeficiente de correlación, un determinado grado de asociación o un determinado ratio. Con estos coeficientes se les dio en los experimentos más importancia a unas variables que ha otras durante el proceso de inferencia, en función del grado predictivo de cada variable. Los coeficientes que se han utilizado en las BNs híbridas y con los que se consiguieron mejores resultados fueron: el coeficiente de correlación de Pearson y el ratio de la Ganancia de información.

Conclusiones de la utilización de las estrategias evolutivas.

La primera conclusión que se puede extraer de las estrategias evolutivas es que se ha conseguido reducir el coste computacional inicializando la población con un algoritmo memético. Sería posible mejorar el algoritmo de estrategia evolutiva extendiendo el uso de los algoritmos meméticos en los métodos de mutación y recombinación. Existen otras conclusiones más controvertidas, desde la perspectiva del coste computacional, como por ejemplo el método de penalización del *fitness*, de mutación y de control de la población. Se puede concluir que para optimizar la TPC de la variable EA, el mejor método de penalización del *fitness* es, penalización de *fitness* dinámica; el mejor método de mutación es, *uncorrelated Mutation with One Step Size*; y el mejor método de control de la población es, PRoFIGA.

Por otro lado, a lo largo de la tesis se ha podido comprobar la aportación de los algoritmos de estrategias evolutivas en la tarea de optimización de la TPC de la variable ***EA***. Aunque no se ha podido contrastar de forma exhaustiva los resultados obtenidos con las estrategias evolutivas, se pone de manifiesto como está técnica puede mejorar el rendimiento de los clasificadores, si se dispusiera de una muestra lo suficientemente representativa y estratificada. Sin embargo, se puede concluir que con las estrategias evolutivas se podrían abaratar costes en la elaboración del corpus lingüístico, ya que

pueden corregir los efectos del sesgo en el proceso de inferencia debido a la falta de aleatoriedad de la muestra.

Conclusiones de otras técnicas de minería de datos.

Una conclusión que se extrae de la aplicación de otras técnicas de minería de datos es la validez del método de diagnóstico aplicando otros algoritmos de aprendizaje automático, ya que se consiguieron con estos algoritmos unos resultados aceptables, aunque no superan a los resultados del método de diagnóstico propuesto en esta tesis doctoral. No se puede concluir que el diagnóstico del deterioro de la memoria semántica se resuelve única y exclusivamente con BNs, ya que existen algoritmos de aprendizaje automático muy eficaces. Sin embargo si podemos concluir, que es muy importante modelar este problema desde el conocimiento del dominio, tanto en la estructura, como en los algoritmos de aprendizaje automático de los parámetros. Por otro lado, se pone de relieve la necesidad de la existencia, en el algoritmo de aprendizaje automático, de algún mecanismo para establecer una influencia informativa sobre las variables del corpus lingüístico [1] de determinados factores de contexto.

Otra conclusión importante que se extrae de la aplicación de otros algoritmos de minería de datos, en comparación con el método de diagnóstico propuesto, es la segmentación de las definiciones orales en rasgos semánticos y la necesidad de dar un peso específico a cada variable del corpus, en función de su ratio predictivo durante el proceso de inferencia.

Conclusiones generales.

El método de diagnóstico que se ha propuesto en esta tesis doctoral, es un método que podría estar presente en la evaluación clínica diaria con un bajo coste económico y muy accesible a grandes poblaciones de personas con riesgo de padecer alguna demencia. Actualmente el diagnóstico basado en biomarcadores [19,20,11] (neuroimagen, líquido cefalorraquídeo y pruebas genéticas) han avanzado enormemente; pero desgraciadamente, estos sistemas no llegan a todo el mundo, pues son costosos y no están disponibles en la clínica diaria. Pasará aún un tiempo para que estas pruebas sean rutinarias, por tanto, actualmente se siguen aplicando criterios clínicos y neuropsicológicos, pudiéndose complementar estos criterios con el método de diagnóstico propuesto en esta tesis.

Tal y como se ha venido indicando a lo largo de esta memoria, tanto el corpus lingüístico, como el método de diagnóstico propuesto en esta tesis doctoral, pueden extenderse a otras ENs.

Otra conclusión que puede extraer de esta tesis es que el mejor modelo de BN discreta es el modelo 3 “Inferencia por razonamiento abductivo y estudio del deterioro semántico diferencial entre los dominios SV y SNV”, junto con la estrategia de clusterización por edad. Sin embargo, queremos resaltar algunos matices respecto a las CLG BN y las BN híbridas con inferencia aproximada; por ejemplo, si existe la posibilidad de ampliar el corpus lingüístico, las BN híbridas dan la posibilidad de

utilizar un corpus de datos para el aprendizaje automático constituido únicamente por personas sanas, lo cual facilita y abarata mucho este trabajo de campo. Si se dispusiera de una muestra lo suficientemente amplia y estratificada, un buen clasificador podría ser la CLG BN junto con la segmentación de la muestra por edad y nivel educativo.

El desarrollo del método de diagnóstico usando técnicas de IA es original de esta tesis. El corpus lingüístico de definiciones orales y la investigación sobre la memoria semántica, pertenece a la investigación de Peraita y Grasso [1], estando el capítulo 3 sección 1 dedicado a una descripción metodológica de esta investigación.

12.2 Trabajos futuros.

Estamos convencidos que el método de diagnóstico que se propone en esta tesis doctoral podría dar lugar a nuevas vías de investigación, como por ejemplo, diseñar un método automático para la adquisición de evidencias para la BN o a investigar con otros modelos de BNs alternativos que puedan considerar otros síntomas, factores de riesgo o incluso distintos estadios de la enfermedad.

En esta sección se proponen nuevas vías investigación. La primera propuesta y la más importante para la incorporación del método de diagnóstico propuesto en esta tesis en la clínica diaria, consiste en extender el software para analizar e interpretar la producción de rasgos semánticos de forma automática o semiautomática. En la segunda propuesta se sugiere una BN con nuevas variables para poder realizar un diagnóstico más preciso. En la tercera propuesta se modela una BN dinámica, la cual tiene en cuenta los diferentes trastornos psicológicos y del comportamiento en la EA, en función del estadio en la que suelen aparecer estos trastornos. Por último, se proponen varios diagramas de influencia (DI) para medir el grado de utilidad de determinadas exploraciones complementarias, tratamientos farmacológicos y terapias cognitivas.

12.2.1 Automatización/Semiautomatización de la captura de evidencias para las BNs.

Diseñar una metodología para la captura de las evidencias es fundamental para que el método de diagnóstico, propuesto en esta tesis, pueda convertirse en un sistema comercial para el diagnóstico de la EA y otras ENs. Se proponen dos enfoques distintos para esta tarea, un método de captura de las evidencias semiautomático y otro método de captura de las evidencias totalmente automatizado. De forma resumida estos enfoques son:

- El primer método propuesto para la captura de evidencias de forma semiautomática, consiste en realizar unos test orales dentro de un marco teórico, dónde se les preguntan a los pacientes sobre distintos rasgos de objetos básicos. El sujeto enumera verbalmente los rasgos que se le ocurra. De esta forma la automatización del reconocimiento de voz sólo tiene que reconocer las pausas que produce el paciente entre los distintos atributos. La enumeración de rasgos debe ser supervisado por personal dedicado a esta tarea para eliminar el ruido

que puedan producir los sujetos. La ventaja respecto al método actual, es que reduce la complejidad del reconocimiento de voz, el análisis e interpretación de las definiciones orales.

- El segundo método para la captura de evidencias de forma automática, consiste en elaborar unos test asistidos por ordenador, dónde se les preguntan a las sujetos sobre determinadas categorías semánticas, con distintos medios audiovisuales.

12.2.2 BN con variables de información y contexto adicionales.

En esta tesis doctoral se utiliza exclusivamente como instrumento metodológico un corpus de definiciones orales, pero sería posible tener en cuenta nuevas variables que representen distintos síntomas, factores de riesgo e información de contexto [21,1]; para permitir hacer un diagnóstico más preciso y determinista.

La BN que se propone en la Figura 57 es similar a las BNs de los capítulos 5 y 6, pero se añaden nuevos déficits asociadas al deterioro cognitivo leve en la tercera edad y a veces prodrómico a la EA. El inconveniente de este modelo es que no tiene en cuenta los distintos estadios de la enfermedad, ni los trastornos asociados a cada fase.

Las variables que son causas del DC leve:

- Déficit léxico-semántico-conceptual en el dominio semántico SV, definido anteriormente.
- Déficit léxico-semántico-conceptual en el dominio semántico SNV, definido anteriormente.
- Problemas de la función ejecutiva.
- Déficits práxicos.
- Problemas de memoria.
- Problemas de denominación de categorías conceptuales.

Las variables asociadas de tipo sociodemográficas y clínicas:

- Nivel de ocupación.
- Edad.
- Sexo.
- Nivel educativo.
- Relaciones sociales (relación social en el momento actual, vivir sólo o acompañado).
- Comorbilidad.
- Tabaquismo, alcohol

Otros tipos de demencias en las que está presente un deterioro cognitivo.

- Alzheimer (EA)
- Demencia Vascular (DV)
- Demencia Parkinson (DP)

- Demencia por cuerpos de Lewy (DCL).
- Demencia frontotemporal (DFT)
- Hidrocefalia a presión normal (HPN).

En la Figura 57 se representa la estructura de una BN propuesta, extendiendo las BNs de esta tesis doctoral a otros campos para poder hacer un diagnóstico más certero.

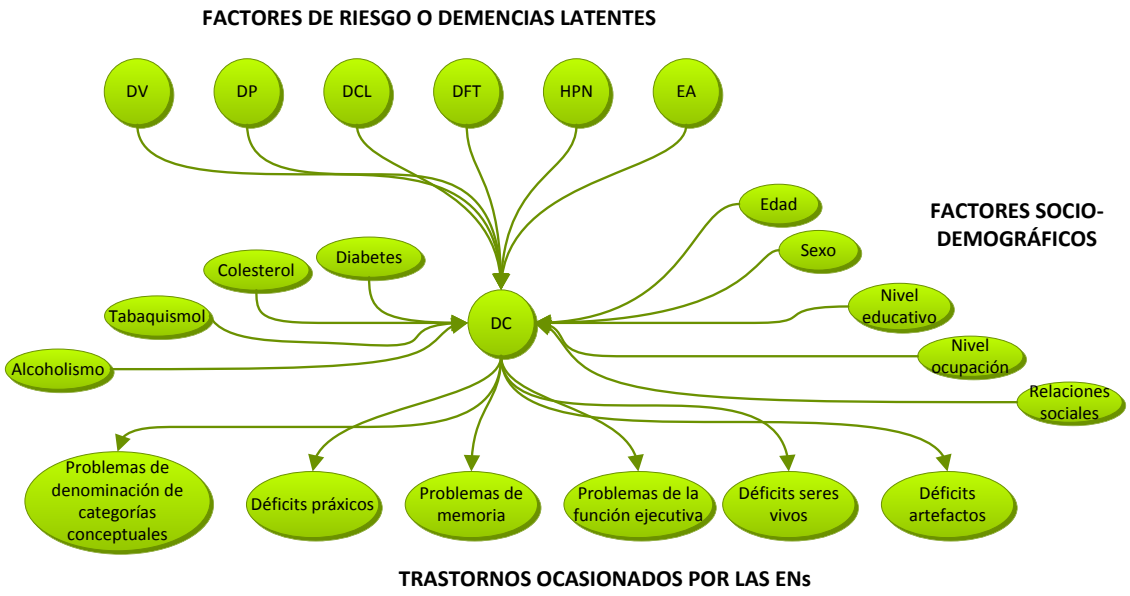


Figura 57.- BN con nuevas variables.

12.2.3 BN Dinámica.

La BN dinámica proporciona un fenómeno de modelado, el cual, evoluciona con el tiempo, es decir, las relaciones causales entre las variables representan la EA en un punto del tiempo. El modelo de BN que se presenta en esta propuesta no es exhaustivo, sólo tiene por objeto proponer una vía de investigación y por tanto un nuevo método de diagnóstico. Construir el modelo cuantitativo y cualitativo de esta BN es extremadamente complejo y requiere de estudios epidemiológicos específicos.

En la Figura 58 se modela una posible BN dinámica. En esta BN se consideran cuatro estadios de la enfermedad, en cada estadio aparecen una serie de trastornos que podrían ayudar a determinar el grado de avance de la EA. Por ejemplo, cuando la enfermedad es incipiente, existe una disminución en la capacidad organizativa, y cuando la enfermedad es moderada, hay una afectación del lenguaje y del sistema motor.

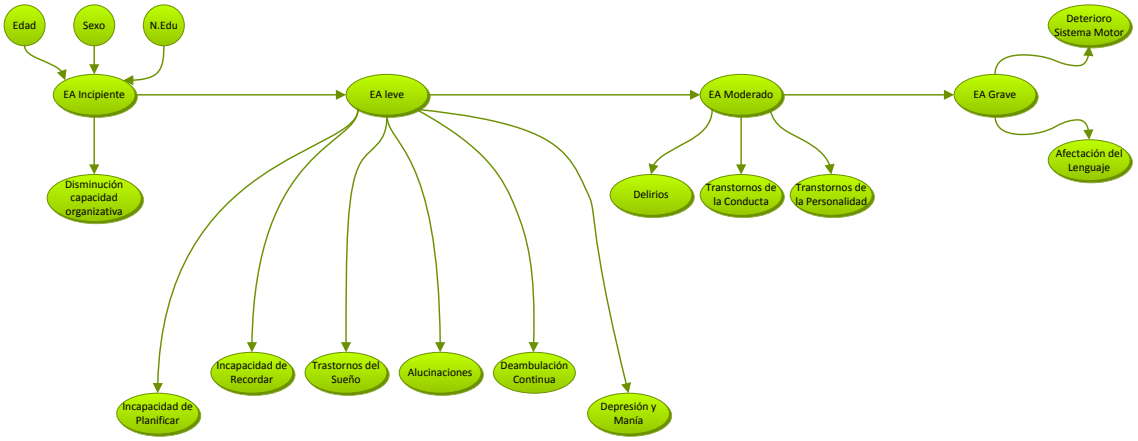


Figura 58.- BN Dinámica.

12.2.4 Diagramas de influencia (DI).

Un DI es un grafo dirigido acíclico con nodos causales, decisiones y nodos de utilidad. Los nodos de utilidad no pueden tener hijos y no tienen estados. Los nodos de decisión y nodos causales, tienen un número finitos de estados.

Los DI permiten tomar decisiones de forma normativa, por ejemplo, un DI podría usarse como una guía práctica clínica para el diagnóstico de la EA. Estos DIs permiten determinar cuál es la política de actuación más adecuada, incluso en los casos en que no es tan evidente y el juicio clínico del médico es incapaz de encontrar la mejor solución, es decir, podría ayudar al profesional de la medicina a determinar cuál es la mejor decisión en cuanto a la realización de exploraciones complementarias o la recomendación de algún medicamento pese a la presencia de la incertidumbre en el diagnóstico. Otra ventaja que podría proporcionar este método en el análisis de decisiones, es que pueden combinar de forma explícita y sistemática las opiniones de diferentes expertos con datos experimentales, tales como los datos de estudios publicados en la literatura médica. Los DIs son unas técnicas muy flexibles, es posible modificar los parámetros para adaptarlos a los de un país diferente, añadir nuevas pruebas diagnósticas o nuevos tratamientos, de una forma fácil.

La idea de este DI es que el paciente realice el test para determinar si padece alguna alteración cognitiva que afecte a la memoria semántica en sus aspectos declarativos. A partir de las probabilidades a posteriori obtenidas con nuestro método de diagnóstico, se calcula la utilidad esperada de determinadas exploraciones complementarias, tratamientos farmacológicos o terapias cognitivas. Al igual que en la BN de la sección anterior, estos DIs sólo introducen una posible vía de investigación. La propuesta que se hace en esta tesis es incorporar nuestro método de diagnóstico a la guía práctica clínica.

En esta propuesta se incorporan tres DIs: el primer DI, ayuda a determinan si se deben realizar exploraciones complementarias; el segundo DI, ayuda a decidir sobre los tratamientos farmacológico más adecuados, y el tercer DI, ayuda a decidir sobre las terapias cognitivas más apropiadas en función del estadio en el que se encuentra la

enfermedad. Un ejemplo de los distintos tipos de decisiones se tiene en cuenta en el DI se obtiene de [21] y son:

- Exploraciones complementarias: CBC, VSG, bioquímica, serología, electrocardiograma, neuroimagen, punción lumbar, electroencefalograma, spect, pet, etc.
- Tratamientos farmacológicos: ansiolíticos, colinérgico, hipnóticos, antidepresivos, neurolépticos, etc.
- Terapias no farmacológicas: memoria, apraxia, afasia, función ejecutiva, etc.

12.2.5 Exploraciones Complementarias.

Este DI calcula la utilidad esperada para realizar determinadas exploraciones complementarias, en función de la probabilidad de padecer la EA. Las elipses representan los nodos causales, los cuadrados representan las decisiones que debe tomar el sistema y los rombos representan las utilidades esperadas de cada decisión. Las utilidades esperadas pueden tratarse como valores subjetivos, pueden representar el coste económico de las exploraciones o los efectos secundarios del tratamiento.

Se ha simplificado el DI de la Figura 59, es decir, se han eliminado la mayor parte de los nodos causales con objeto de que el DI sea más legible. Las variables aleatorias se toman de las BNs propuestas en esta tesis doctoral, los nodos de decisión se han tomado de [21] y los nodos de utilidad se proponen para futuras investigaciones.

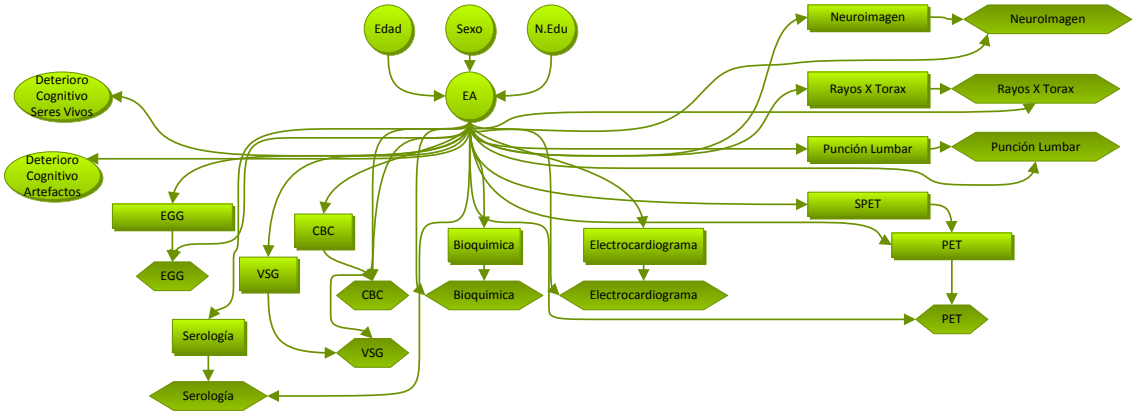


Figura 59.- Diagrama de influencia. Exploraciones complementarias.

12.2.6 Tratamientos Farmacológicos.

En esta sección se propone un DI para calcular la utilidad esperada de la aplicación de determinados tratamientos farmacológicos, una vez conocida la probabilidad de padecer la EA. Al igual que en la sección anterior, las elipses representan los nodos causales, los cuadrados representan las decisiones que debe tomar el sistema y los rombos representan las utilidades esperadas de cada decisión. Los nodos de decisión se ha tomado de [21] y las utilidades pueden usar diversos criterios como: efectos secundarios que afecten a la calidad de vida del paciente, frente a los beneficios que se obtiene con dicho tratamiento; también se pueden utilizar criterios económicos, etc. En la Figura 60 se representa el DI para los tratamientos farmacológicos.

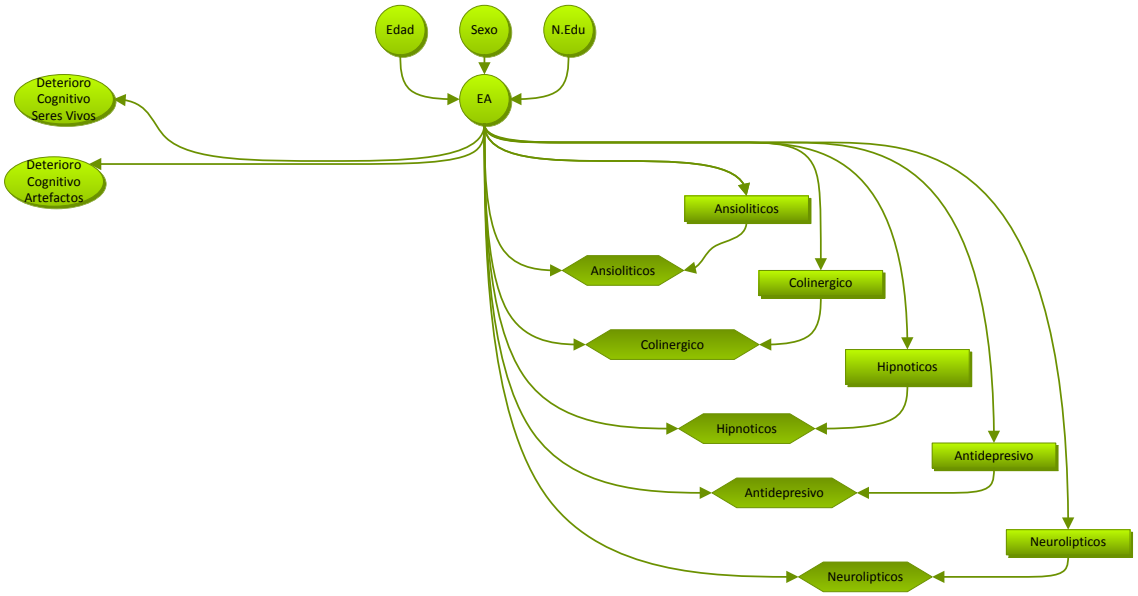


Figura 60.- Diagrama de Influencia. Tratamientos Farmacológicos.

12.2.7 Terapias no farmacológicas.

Existen investigaciones [55] que señalan la eficacia de las terapias no farmacológicas en la mejora de la calidad de vida de las personas que padecen la EA. Según los resultados de [55] las terapias no farmacológicas pueden contribuir de forma realista y asequible a la mejora y administración de cuidados de los enfermos de EA. Las terapias cognitivas influyen en el retardo de los efectos de la enfermedad en relación al DC.

Al igual que en la sección anterior, las elipses representan los nodos causales, los cuadrados representan las decisiones que debe tomar el sistema y los rombos representan las utilidades esperadas de cada decisión. La utilidad esperada puede atender a criterios como: costes económicos, beneficios para el paciente en relación a su calidad de vida, etc. La Figura 61 representa el DI para recomendar determinadas terapias cognitivas.

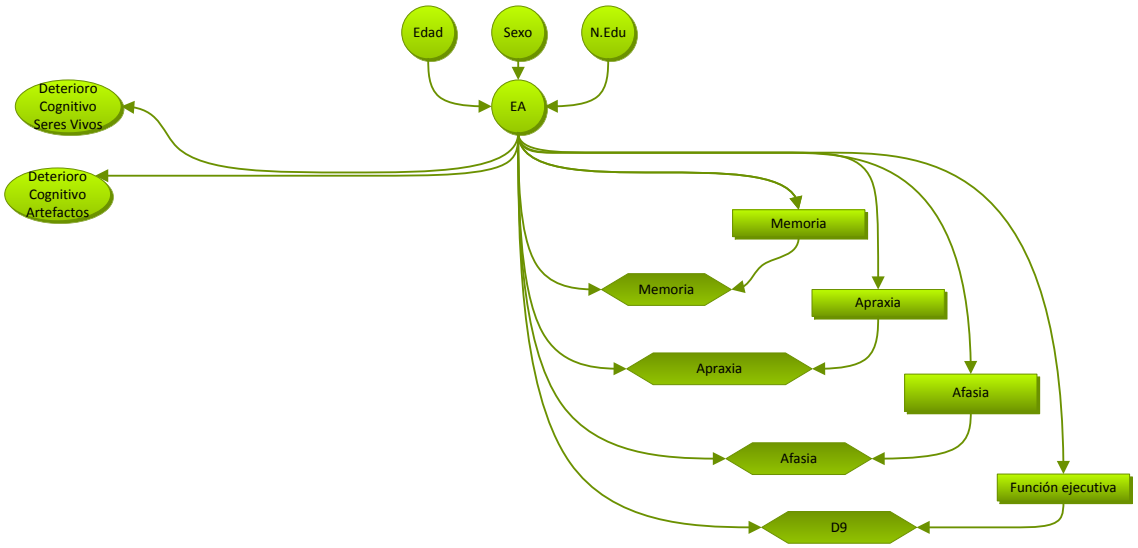


Figura 61.- Diagrama de Influencia. Terapias cognitivas.

Bibliografía.

1. Peraita Adrados H, Grasso L. Corpus lingüístico de definiciones de categorías semánticas de personas mayores sanas y con la enfermedad de Alzheimer. Una investigación transcultural hispano-argentina. Fundación BBVA [in Spanish] [Internet]. 2010. Available from: http://www.fbbva.es/TLFU/dat/DT3_2010_corpus_linguistico_peraita_web.pdf ; www.uned.es/investigacion-corpuslinguistico.

2. Commission E. Public Health [Internet]. 2013 [cited 2013 03 23. Available from: http://ec.europa.eu/health/major_chronic_diseases/diseases/alzheimer/index_en.htm.

3. Lambon Ralph MA, Lowe C, Rogers TT. Neural basis of category-specific semantic deficits for living things: evidencefrom semantic dementia, HSVE and a neural network model. Brain. 2007 April; 130(Pt 4).

4. Cohen G, Johnstone RA, Plunkett K. Exploring cognition: Damaged brains and neural networks. 1st ed.: Psychology Press; 2002.

5. T. Rogers T, C. Plaut D. Connectionist perspectives on category-specific deficits. In Forde E, Humphreys G, editors. Category Specificity in Brain and Mind. East Sussex, England: Psychology Press; 2002. p. 251-290.

6. Capitani E, Laiacona M, Mahon B, Caramazza A. What are the facts of semantic category-specific deficits? A critical review of the clinical evidence. Cogn Neuropsychol. 2003 May; 20(3).

7. Bentler PM. Multivariate Analysis with Latent Variables. Annu Rev Psychol. 1980 February; 31(419-456): 419-456.

8. Ken McRae GSCMSSaCM. Semantic feature production norms for a large set of living and nonliving things. Behav Res Methods. 2005 November; 37(4): 547-559.

9. Anderson JR. Rules of the Mind. 1st ed. Hillsdale, New Jersey: Lawrence Erlbaum Associates; 1993.

10. Eastman Q. EMORY [Internet]. 2012 [cited 2012 Diciembre 05. Available from: http://news.emory.edu/stories/2012/08/alzheimers_blood_test.

11. Nordberg A. Amyloid imaging in Alzheimer's disease. Neuropsychologia. 2008;(46): 1636–1641.

12. Klunk WE, Engler H, Nordberg A, Wang Y, Blomqvist G, Holt DP, et al. Imaging Brain Amyloid in Alzheimer's Disease with Pittsburgh Compound-B. Ann Neurol. 2004 March; 55(3).

13. Peraita Adrados H, Galeote Moreno MÁ, González Labra MJ. Deterioro de la memoria semántica [in Spanish]. Psicothema. 1999 Marzo; 11(4): 917-937.

14. Stern Y, Gurland B, Tatemichi TK, Xin Tang M, Wilder D, Richard M. Influence of Education and Occupation on the Incidence of Alzheimer's Disease. 1994; 271(13).

15. Guerrero Triviño JM, Martínez-Tomás R, Peraita Adrados H. Bayesian Network-based Model for the Diagnosis of Deterioration of Semantic Content Compatible with Alzheimer's Disease. In Proceding IWINAC 2011. Foundations on Natural and Artificial Computation.; 2011; Isla de La Palma. Canarias. p. 419-430.

16. Guerrero Triviño JM, Martínez Tomás R, Díaz Mardomingo MC, Peraita Adrados H. Utilidad/uso especifica/o de un Corpus de Definiciones de Categorías Semánticas. Modelo basado en Redes Bayesianas para el Diagnóstico del Deterioro del Contenido Semántico Compatible con la Enfermedad de Alzheimer [in Spanish]. In Procede del III congreso internacional del lingüística del corpus.; 2011; Valencia: Universitat politècnica de València. p. 827,836.

17. Singh-Manoux A, Kivimaki M, Glymour MM, Elbaz A, Berr C, Ebmeier KP, et al. Timing of onset of cognitive decline: results from Whitehall II prospective cohort study. BMJ. 2012 January; 344.

18. Fernández Martínez M, Castro-Flores J, Pérez-de las Heras S, Mandaluniz-Lekumberri A, Gordejuela M, Zarranz J. Prevalencia de la demencia en mayores de 65 años en una comarca del País Vasco [in Spanish]. Rev Neurol. 2008 Enero; 46(2): 89-96.

19. Maddalena A, Papassotiropoulos A, Jung HH, Müller-Tillmanns B, Hegi T, Nitsch RM, et al. Biochemical Diagnosis of Alzheimer Disease by Measuring the Cerebrospinal Fluid Ratio of Phosphorylated tau Protein to β-Amyloid Peptide42. JAMA. 2003 Sep; 60(1202).

20. Hu WT, Holtzman DM, Fagan AM, Shaw LM, Perrin R, Arnold SE, et al. Plasma multianalyte profiling in mild cognitive impairment and Alzheimer disease. Neurology. 2012 August; 79(9): 897-905.

21. Peña-Casanova J. Enfermedad de Alzheimer. Del diagnóstico a la terapia: conceptos y hechos [in Spanish]. Fundación la Caixa [Internet]. 1999. Available from: http://www.fundacio1.lacaixa.es/webflc/wpr0pres.nsf/wurl/alndream1pcos_esp%

5EOpenDocument/index.html.

22. Ruben Armañanzas. PLCB. Ensemble transcript interaction networks: A case study. Comput Methods Programs Biomed. 2012;: 442-450.

23. Evanthia E. Tripoliti DIF. A supervised method to assist the diagnosis of Alzheimer's Disease. In Proceedings of the 29th Annual International Conference of the IEEE EMBS; 2007; Lyon. p. 3426-3429.

24. Sun Y, Lv S, Tang Y. Construction and Application of Bayesian Network in Early Diagnosis of Alzheimer Disease's System. In Proceeding of IEEE/ICME International Conference on Complex Medical Engineering; 2007; Dalian Univ. of Technol., Dalian. p. 924- 929.

25. Matic A, Mehta P, Rehg JM, Osmani V, Mayora O. Monitoring Dressing Activity Failures through RFID and Video. Methods Inf Med. 2012 April; 51(1): 45–54.

26. Kearns WD, Nams VO, Fozard JL. Tortuosity in Movement Paths Is Related to Cognitive Impairment. Methods Inf Med. 2010 March; 49(6): 592–598.

27. Sun Y, Tang Y, Ding S, Lv S, Cui Y. Diagnose the mild cognitive impairment by constructing Bayesian network. Elsevier. 2011 January; 38(1).

28. Díez FJ, Iturralde E, Mira J, Zubillaga S. DIAVAL, a Bayesian expert system for echocardiography. Artif Intell Med. 1997 May; 10(1): 59-73.

29. Triviño JMG. Diagnóstico del deterioro cognitivo compatible con enfermedades neurodegenerativas con Redes Bayesianas [in Spanish]. Trabajo Fin de Máster. Madrid: UNED, Departamento de IA Avanzada; 2011.

30. Díez FJ. Aplicaciones de los modelos gráficos probabilistas en medicina [in Spanish]. In Martín JAG, Callejón JMP, editors. Procede del curso de verano de la U.C.L.M.; 1998; Albacete. p. 239-263.

31. Arthur D, Vassilvitskii S. k-means++. The Advantages of Careful Seeding. In SODA '07 Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms; 2007; New Orleans: Society for Industrial and Applied Mathematics Philadelphia. p. 1027 - 1035.

32. E. Neapolitan R. Learning Bayesian Networks. 1st ed. Upper Saddle River, New Jersey: Prentice Hall Series in Artificial Intelligence; 2004.

33. UNED. Elvira [Internet].. Available from: http://www.ia.uned.es/~elvira/index-en.html.

34. Valls-Pedret C, Molinuevo JL, Rami L. Diagnóstico precoz de la enfermedad de Alzheimer: fase prodrómica y preclínica [in Spanish]. Rev. Neurología. 2010 Septiembre; 51: 51(8):471-480.

35. Grasso L, Mardomingo MdC, Peraita Adrados H. Análisis preliminar de rasgos de definiciones de categorías semánticas del Corpus lingüístico de sujetos sanos y con Enfermedad de Alzheimer [in Spanish]. Facultad de Psicología. UNED, Departamento de Psicología Básica 1; 2009.

36. Kjaerulff U, Madsen A. Bayesian Networks and Influence Diagrams. 1st ed. Jordan M, Kleinberg J, Schölkopf B, editors. New York: Springer; 2007.

37. Díez-Vegas FJ. Teoría probabilista de la decisión en medicina [in Spanish] [Internet]. 2007. Available from: http://www.cisiad.uned.es/techreports/decision-medicina.php.

38. A.E.Eiben , J.E.Smith. Introduction to Evolutionary Computing. 1st ed. G.Rozenberg , editor. New York: Springer; 2003.

39. Poli R, Langdon WB. Schema Theory for Genetic Programming with One-Point Crossover and Point Mutation. School of Computer Science. The University of Birmingham.

40. Witten IH, Frank E, Hall MA. Data Mining. Practical Machine Learning Tools and Techniques. 3rd ed. Kaufmann M, editor. New York: Morgan Kaufmann; 2011.

41. Quilan JR. Induction of decision trees. Machine Learning. 1986;: 81-106.

42. Chan TF, Golub GH, LeVeque RJ. Algorithms for Computing the Sample Variance: Analysis and. American Statistician. 1983; 37.

43. The Apache Commons Mathematics Library [Internet].. Available from: http://commons.apache.org/math/.

44. Blashfield RK, Aldenderfer MSN. Handbook of multivariate experimental psychology. Perspectives on individual differences. The methods and problems of cluster analysis. 2nd ed. R. J, Cattell RB, editors. New York, NY, US: Plenum Press; 1988.

45. Henry DB, Tolan PH, Gorman-Smith D. Cluster Analysis in Family Psychology Research. Journal of Family Psychology. 2005 Mar; 19(1).

46. Estarelles R, Fuente Idl, Olmedo P. Aplicación y valoración de diferentes algoritmos no-jerarquicos en el analisis cluster y su representación grafica.

Anuario de Psicologia. 1992;(55).

47. Wikipedia [Internet].. Available from: http://es.wikipedia.org/wiki/Distribucion_normal.

48. Quilan JR. C4.5: Programs for Machine Learning. Morgan Kaufmann. 1993;: 81-106.

49. Mitchell TM. Machine Learning: McGraw-Hill Science/Engineering/Math; 1997.

50. Fawcett T. ROC Graphs: Notes and Practical Considerations for Researchers. Notes and Practical Considerations for researchers. 1501 Page Mill Road, Palo Alto: HP Laboratories; 2004.

51. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. 1st ed. Champman , Hall , editors. Belmont, CA: CRC Press; 1984.

52. Hand DJ, Till RJ. A simple generalization of the area under the ROC curve to multiple class. Machine Learning. 2001 November; 45(2): 171-186.

53. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. Springer-Verlag. 2008 January; 14(1): 1-37.

54. Friedman N, Geiger D, Goldszmidt M. Bayesian Network classifiers. In Provan G, Langley P, Smyth P, editors. Machine Learning. Netherlands: Kluwer Academic Publishers.; 1997. p. 131-163.

55. Olazarán J, Reisberg B, Clare L, Cruz I, Peña-Casanova J, Ser Td, et al. Nonpharmacological Therapies in Alzheimer's Disease: A Systematic Review of Efficacy. Dement Geriatr Cogn Disord. 2010 September; 30(2).

56. Oracle. Java Enterprise Edition [Internet].. Available from: http://www.oracle.com/technetwork/java/javaee/overview/index.html.

PARTE

Apéndices

V

Apéndice

Descripción de la IU

A

En este capítulo se describe brevemente la interfaz de usuario del software desarrollado en esta tesis doctoral. Este software se ha desarrollado en entorno Web, 100% Java y cumple la especificación Java EE 6.

La elección de la arquitectura software utilizada en esta tesis doctoral no es arbitraria, sino que se ha seleccionado para enmarcar el proyecto dentro de los proyectos TIC e-Salud y poder así facilitar la financiación privada, aunque en el momento de escribir esta tesis doctoral no se ha conseguido financiación.

El objetivo de este capítulo es dar una visión en amplitud, más que en profundidad, de la interfaz de usuario (IU) del software. El capítulo comienza una con una introducción y continúa con una descripción de todos los subsistemas que componen el software.

A.1. Introducción.

Para elaborar esta tesis doctoral se ha desarrollado un software para el aprendizaje automático, métodos de inferencia innovados en esta tesis, cálculo de métricas de rendimiento, optimización de las TPCs, experimentos, etc. Este software utiliza *framework opensource* como, *Elvira*[8], *Apache Math* [43], *Richfaces* o *Java EE 6* [56].

El software se ha desarrollado teniendo en cuenta la internacionalización. Los navegadores soportado son:

- Entornos Linux.
  - Firefox 3.0 y superior.
  - Opera 9.5 y superior.
- Entorno Windows.
  - Firefox 3.0 y superior.
  - Google Chrome.
  - Internet Explorer 6.0 y superior.
  - Opera 9.5 y superior.
  - Safari 3.0 y superior.

[8] http://www.ia.uned.es/~elvira/index-en.html

- Mac OS environments.
  - Safari 3.0 y superior.
  - Firefox 3.5 y superior.

Este software se divide en cinco bloques:

- **Aprendizaje**: Contiene toda la funcionalidad necesaria para el aprendizaje automático del modelo cuantitativo. Entre estas funcionalidades se encuentran: gestión de prevalencias procedentes de estudio epidemiológicos, gestión de prevalencias calculadas a partir de la base de casos, gestión de la base de casos y cálculo de parámetros para las BN discretas e Híbridas.
- **Inferencia**: Esta herramienta permite inferir un diagnóstico a partir del análisis de las definiciones orales, una vez segmentada las definiciones orales en unidades menores y significativas, es decir, en rasgos semánticos o atributos.
- **Análisis de los modelos**. Herramienta que permite medir el rendimiento de los clasificadores, y en general, del método de diagnóstico.
- **Optimización**: Esta herramienta permite optimizar la TPC de la variable de interés EA.
- **Configuración**: Contiene la funcionalidad necesaria para establecer todos los parámetros del sistema para su configuración.

A.2. Aprendizaje.

El subsistema de aprendizaje automático de los modelos cuantitativos, consta de una serie de herramientas cuyo objetivo fundamental es el aprendizaje de los modelos cuantitativos de las distintas BNs diseñadas en esta tesis doctoral. Del mismo modo, existe la posibilidad de establecer todos los parámetros de forma manual.

A.2.1. Prevalencias.

Actualmente existen numerosas investigaciones sobre la EA que dan lugar a distintos estudios epidemiológicos, y por tanto, la información relativa a las prevalencias de la EA, por edad, por sexo y por nivel educativo, pueden variar de forma continuada. Por esta razón es necesario que el software disponga de una pantalla de usuario para configurar esta información.

En la Figura 62 se pueden observar varias tablas que responden a la prevalencia de la EA estratificada por edad, sexo y nivel educativo. Estas prevalencias son fundamentales para el aprendizaje del modelo cuantitativo. Los campos que contiene cada una de las tablas son:

- Identificador de la prevalencia. Es un identificador único de la prevalencia.
- Descripción. Es una descripción de la prevalencia que estamos tratando.
- Prevalencia. Es un número decimal comprendido entre 0 y 1 con el valor de la prevalencia.
- IC. Intervalo de confianza para la prevalencia.

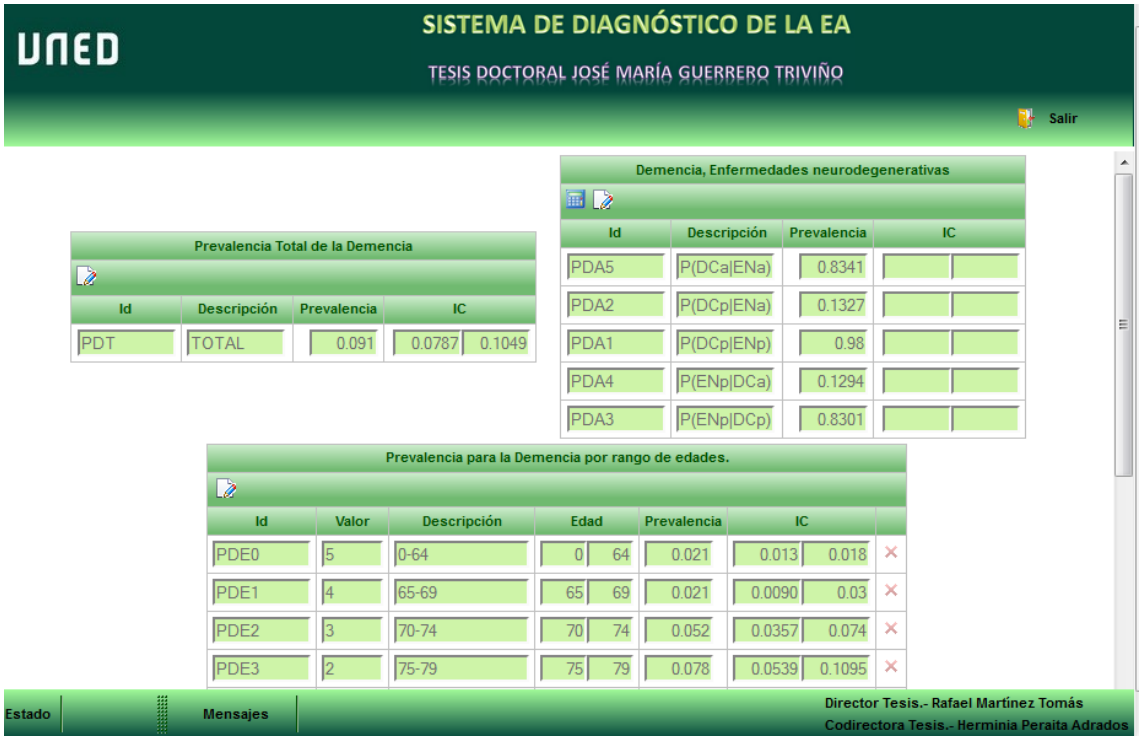


Figura 62.- Gestión de prevalencias procedentes de estudios epidemiológicos.

El sistema también da la posibilidad de calcular las prevalencias a partir de los datos del corpus lingüístico [1].

A.2.2. Base de casos del corpus lingüístico.

El software incluye una gestión para las instancias que constituyen el corpus lingüístico. La Figura 63 es la IU que permite filtrar la información por distintos criterios de consulta.

Figura 63.- IU para la gestión de casos del corpus lingüístico de definiciones orales.

La Figura 64 es la IU para el alta, baja y modificación de casos individuales del corpus.

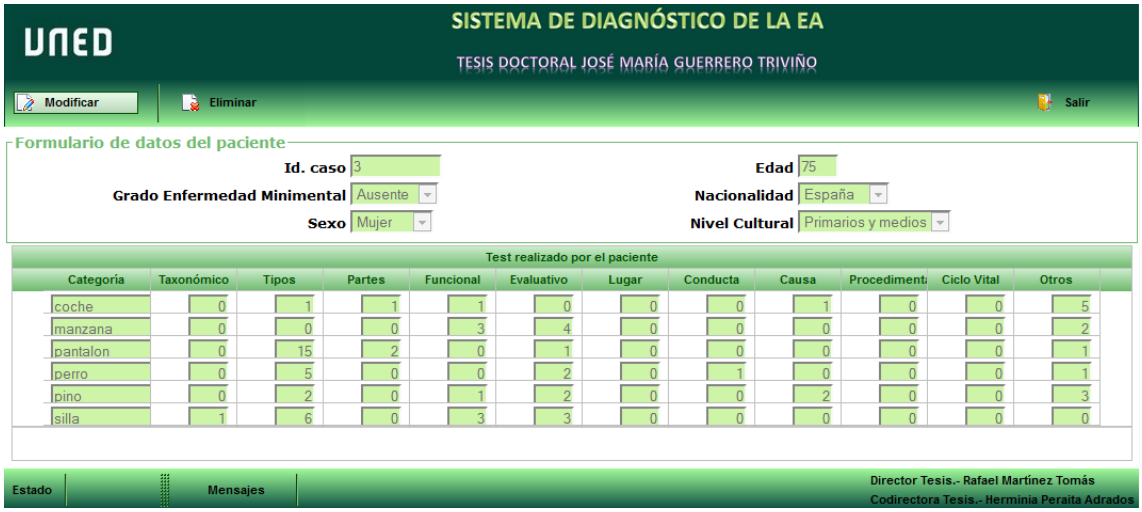


Figura 64.- IU de detalle de los casos del corpus lingüístico.

A.2.3. Aprendizaje del modelo cuantitativo de las BN discretas.

El software implementa algoritmos para el aprendizaje automático del modelo cuantitativo de las BN discretas. En el capítulo 6 se describen en detalle estos algoritmos.

El software no sólo calcula parámetros, sino que también permite realizar una pequeña gestión de las TPC de todas las variables de la BN. Los objetivos de esta IU son:

- Aprendizaje automático del modelo cuantitativo de las BN discretas.
- Gestión de las TPCs.
- Sincronización de las TPCs con Elvira.
- Optimización de las TPCs con algoritmos de estrategias evolutiva.
- Carga de las TPCs optimizadas con los algoritmos de estrategias evolutivas.

En la Figura 65 se muestra un fragmento de la IU para el aprendizaje automático del modelo cuantitativo. Cabe destacar que son muchos los parámetros de esta BN, por lo que se ha dividido la IU en pestañas por cada grupo de variables: variables de interés, variables intermedias, categoría semántica ***manzana***, categoría semántica ***perro***, categoría semántica ***pino***, categoría semántica ***coche***, categoría semántica ***silla*** y categoría semántica ***pantalón***.

Figura 65.- IU para el aprendizaje automático del modelo cuantitativo de las BN discretas.

En esta IU se permite establecer probabilidades condicionales de forma manual, es decir, se puede modificar cualquier TPC de la BN y sincronizar con Elvira.

A.2.4. Aprendizaje del modelo cuantitativo de las BN híbridas.

Al igual que las BNs discretas, las BNs híbridas requieren de un aprendizaje del modelo cuantitativo. Las BNs híbridas, a diferencia de las BNs discretas, sólo necesitan una muestra de sujetos sanos para el aprendizaje del modelo cuantitativo. A diferencia de la IU de la sección anterior, estos parámetros no se pueden modificar.

Los objetivos de esta IU son:

- Aprendizaje automático del modelo cuantitativo de las BN híbridas.
- Sincronización de las TPCs con Elvira.
- Optimización de las TPCs con algoritmos de estrategias evolutiva.
- Carga de las TPCs optimizadas con los algoritmos de estrategias evolutivas.

La Figura 66 es un fragmento de la IU para el aprendizaje automático de los parámetros de las BN híbridas. Al igual que en IU anterior la pantalla se divide en distintas pestañas por cada grupo de variables.

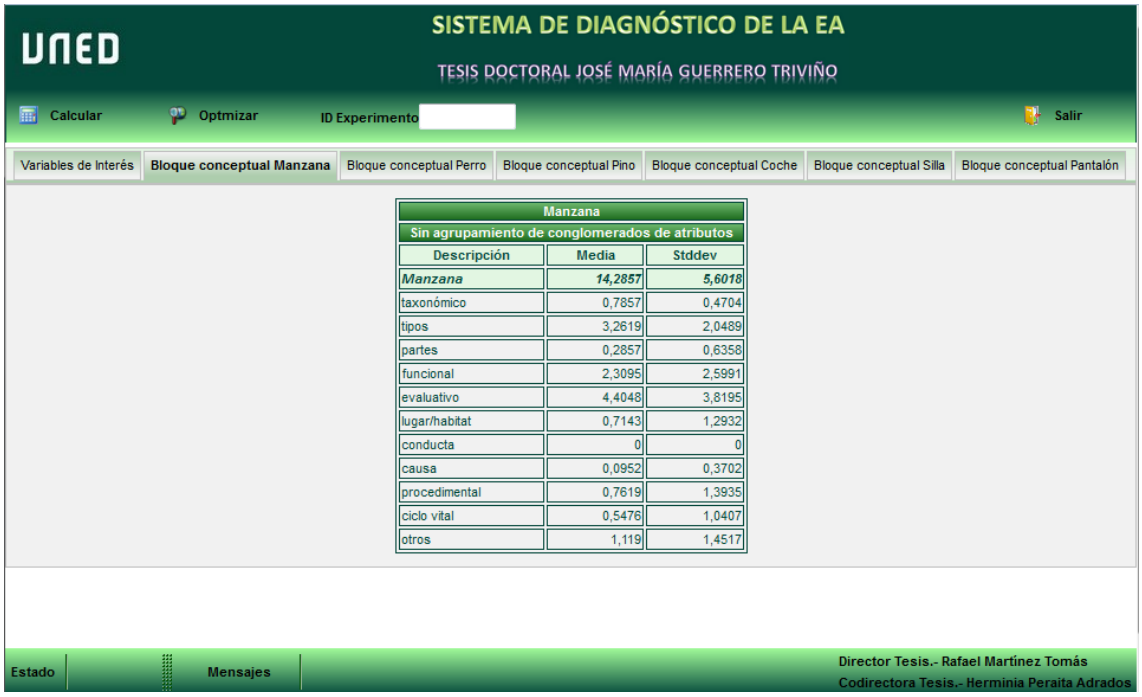


Figura 66.- IU para el aprendizaje automático del modelo cuantitativo de las BN híbridas (μ y σ)

La Figura 67 es la IU para el cálculo de los coeficientes de correlación, regresión, ponderación de atributos y ganancia de información. Estos coeficientes se utilizan en la CLG BN y la BN con inferencia aproximada (ver detalles en el capítulo 7).

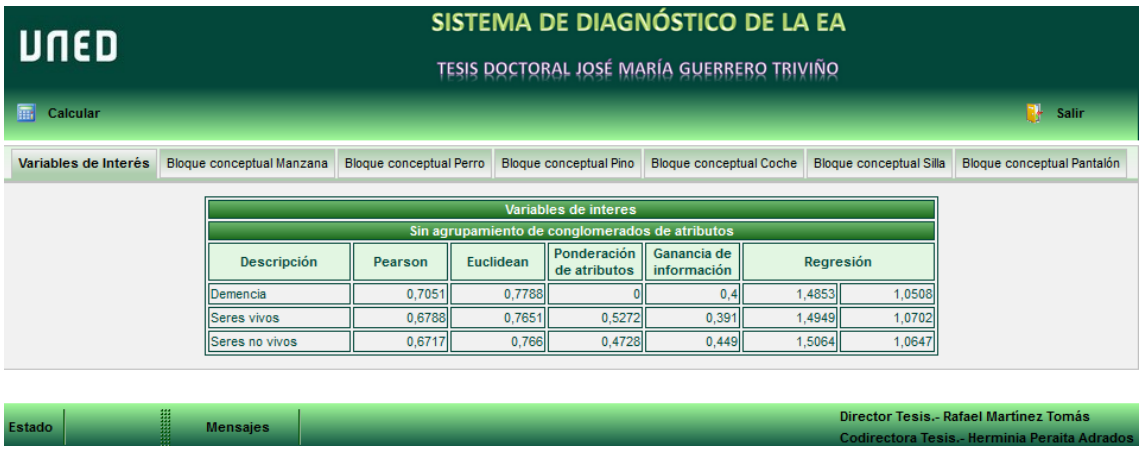


Figura 67.- IU para el aprendizaje automático de distintos coeficientes de correlación.

A.2.5. Consulta de clúster.

Las BNs discretas requieren de un proceso de discretización de los atributos numéricos. Este proceso de discretización se realiza con *k-Means++,* donde se buscan dos centroides por cada variable. Cada centroide representa el estado que puede tomar cada variable.

La Figura 68 es la IU para consultar los centroides calculados con el algoritmo *k-Means++*.

Figura 68.- IU para la consulta de los centroides obtenidos con k-Means++

A.3. Inferencia.

El subsistema de inferencia está constituido por una única IU. El objetivo de esta IU es inferir un diagnóstico a partir de las definiciones orales de las categorías naturales y objetos básicos, una vez segmentada estas definiciones orales en rasgos semánticos.

La Figura 69 es la IU para la introducción de evidencias y su propagación por la BN. En esta IU se muestran las probabilidades a posteriori de la variable de interés ***EA*** y el ***DSD***. En los capítulos 5 y 6 se detallaron las técnicas utilizadas durante el proceso de inferencia.

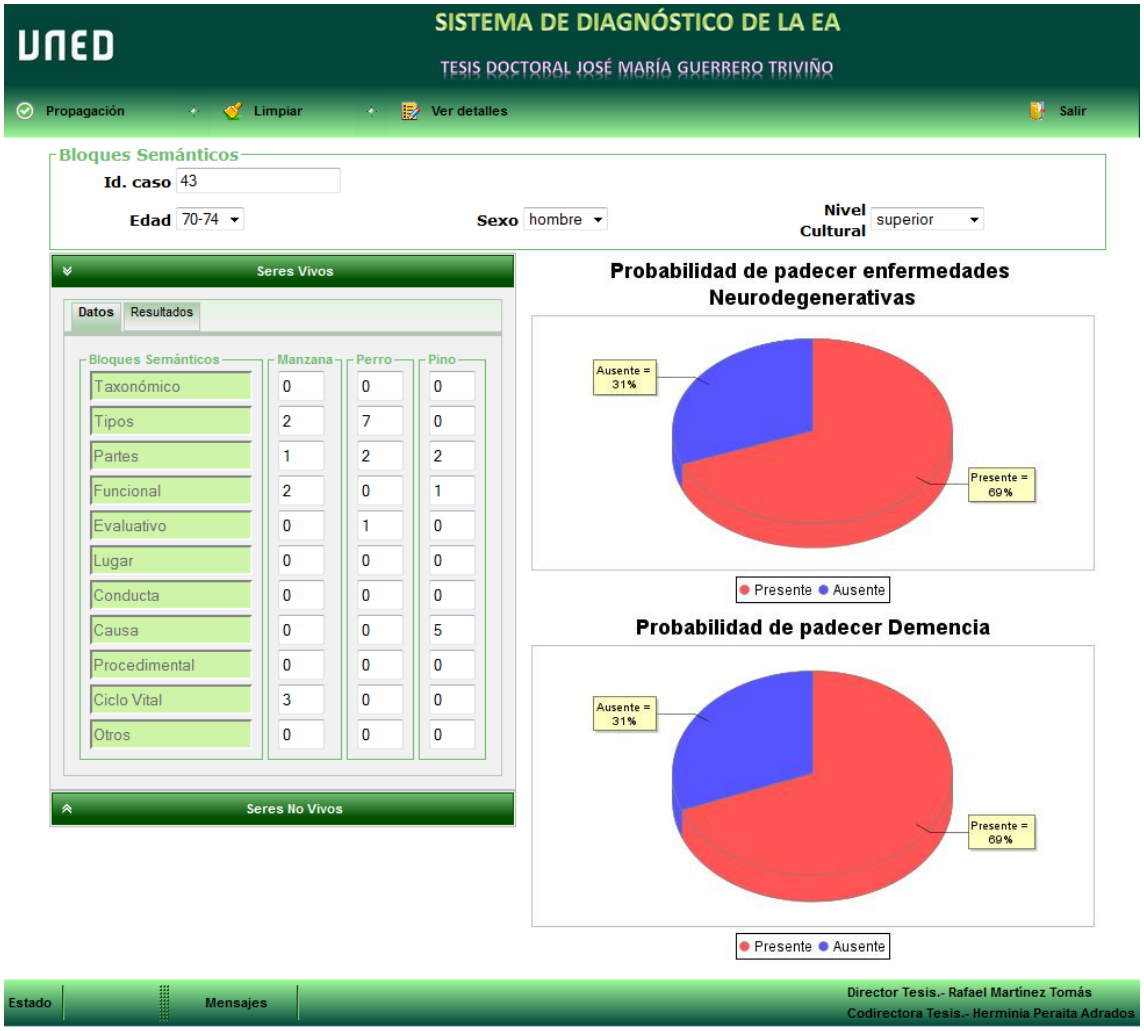


Figura 69.- IU para la inferencia de casos individuales.

También es posible consultar las probabilidades a posteriori de las variables intermedias, tal y como se muestra en la Figura 70.

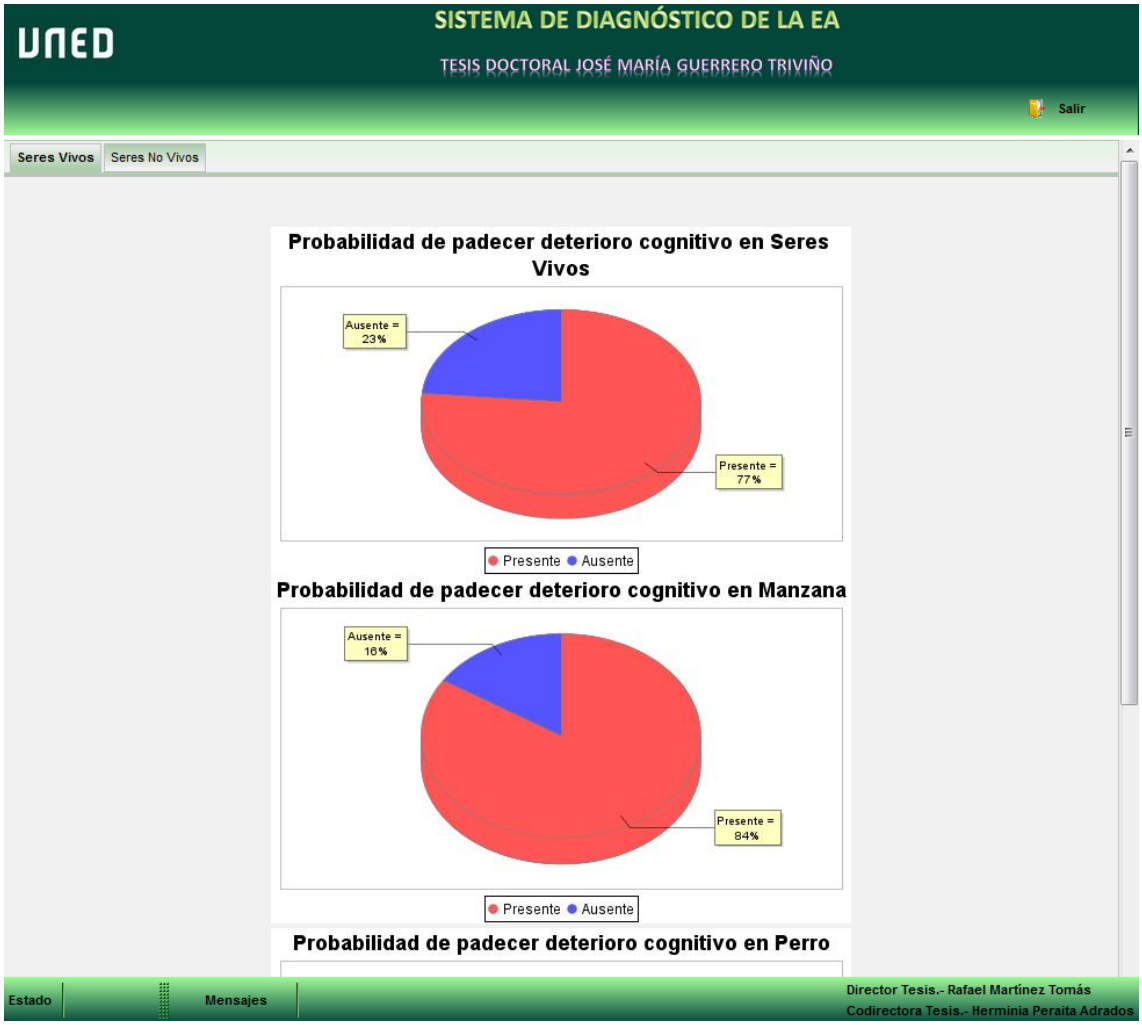


Figura 70.- IU para representar las probabilidades a posteriori de las variables intermedias SV y SNV.

A.4. Medidas de rendimiento.

El software desarrollado en esta tesis genera diversas medidas de rendimiento, con el fin de comparar la eficacia de los distintos clasificadores. Para generar estas métricas se propagan todos los casos por la BN y se guardan las probabilidades a posteriori de todas las variables de interés y de las variables intermedias. Existen diversas estrategias para obtener estas probabilidades a posteriori y evitar el sobreajuste:

- Utilizar el mismo conjunto de datos para el test y validación.
- Utilizar *One Leave Out Cross Validation.*
- Utilizar *nFold Cross Validation.*

En la Figura 71 se representa tantas gráficas como variables de interés e intermedias tiene la BN. En el capítulo 7 se explica en detalle las gráficas que aparecen en la Figura 71. Esta IU divide la pantalla en diferentes pestañas y en cada pestaña se representan las gráficas por cada variable de interés e intermedia.

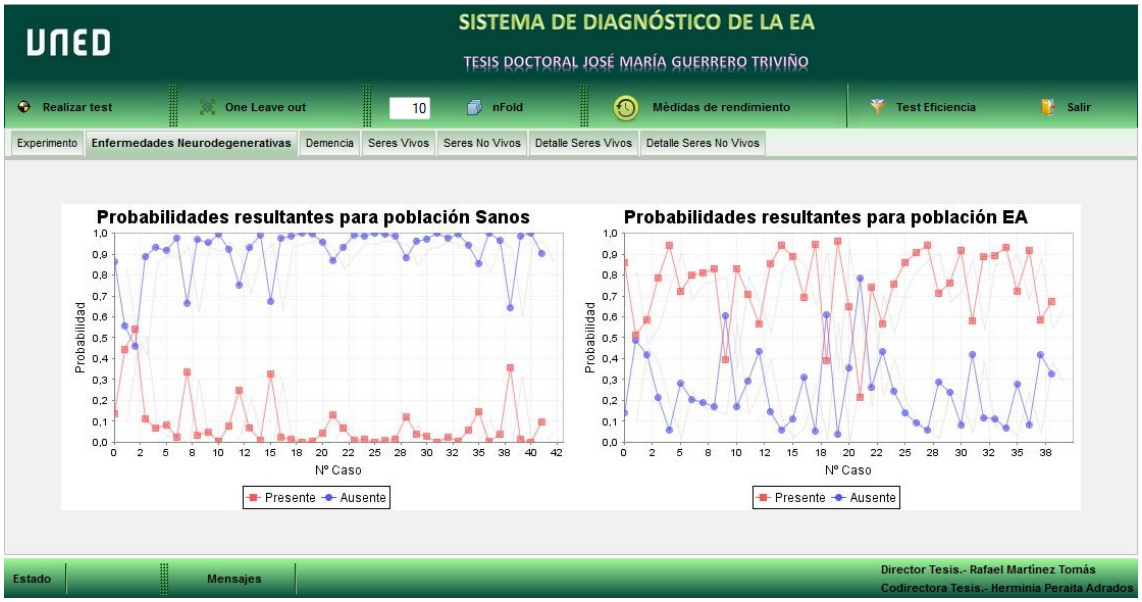


Figura 71. IU para obtener métricas de rendimiento. Propagación de todos los casos.

Desde la IU de la Figura 71, es posible obtener otras métricas de rendimiento, tal y como se puede comprobar en la Figura 72. Estas métricas son relativas a las curvas ROC (ver detalles en el capítulo 7).

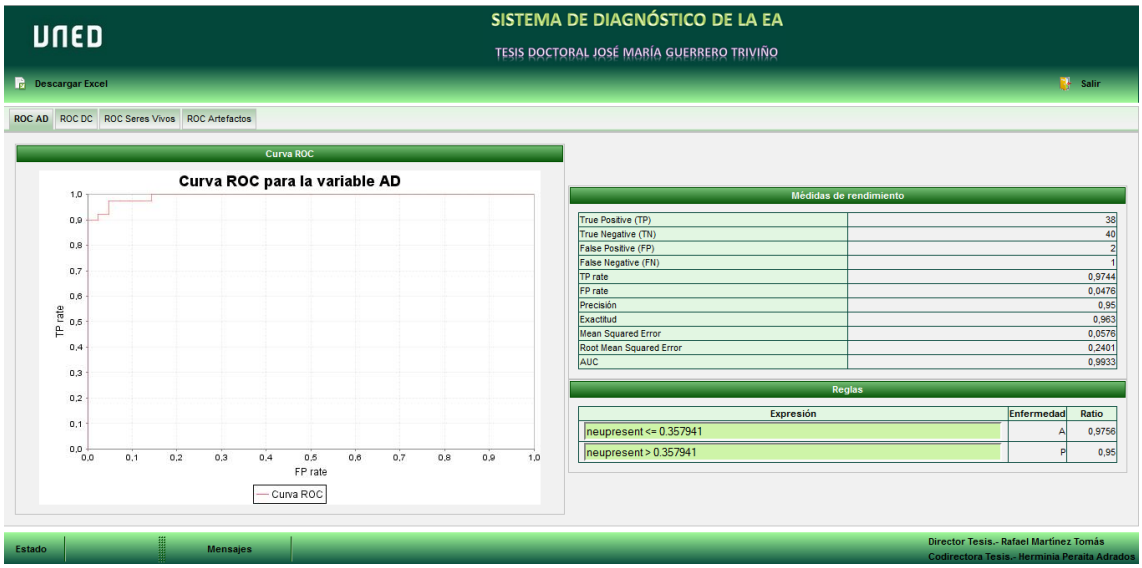


Figura 72.- IU con métricas relativas a las curvas ROC.

También se ha implementado otra métrica de rendimiento, basada en la distancia respecto a su valor esperado. La Figura 73 se muestra la gráfica que representa la métrica que relaciona las probabilidades a posteriori de las variables de interés, con su valor esperado.

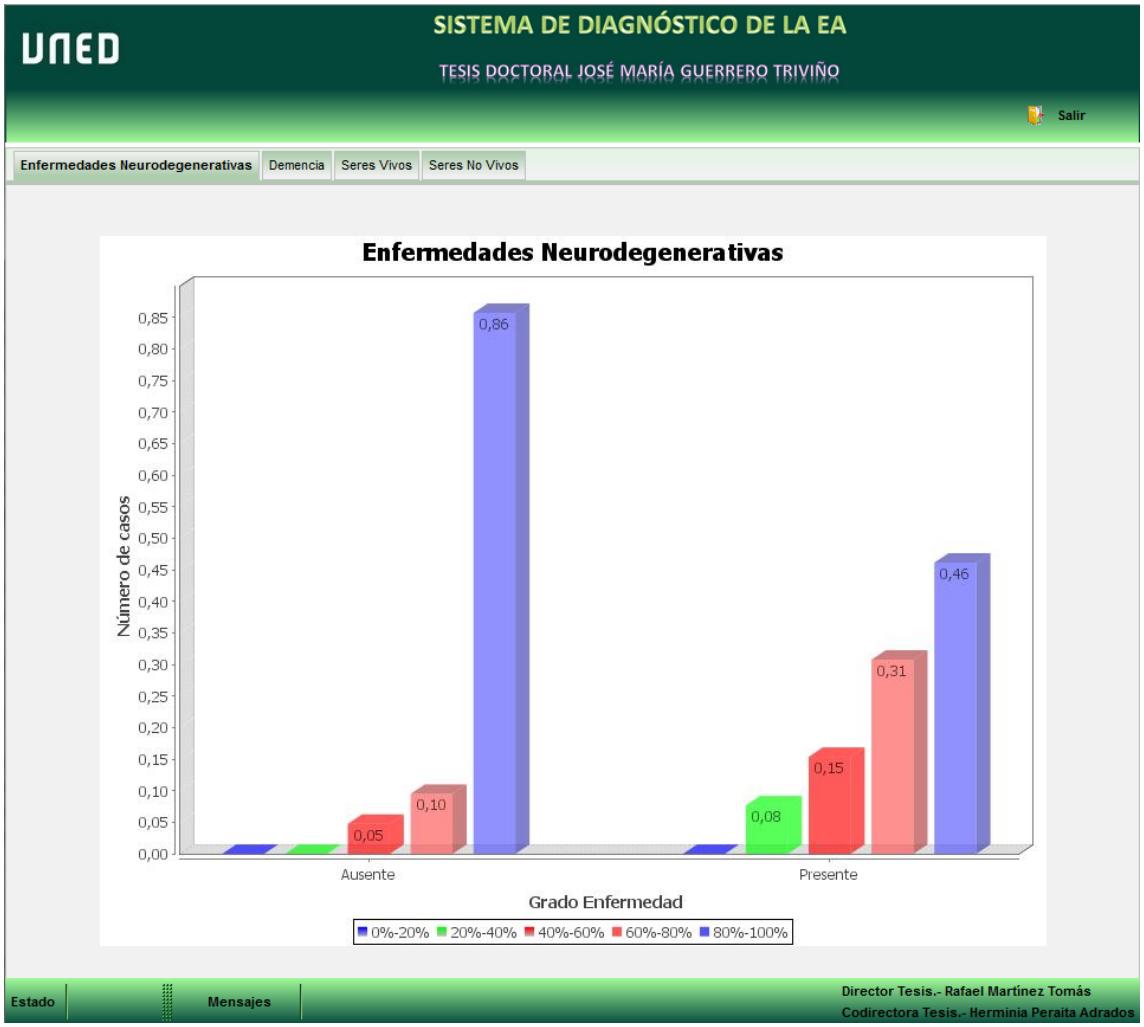


Figura 73.- IU de métricas respecto a su valor esperado.

Puede consultar más detalles sobre está gráfica en los capítulos 8 y 9.

A.5. Configuración.

Este software requiere de una configuración donde se indica entre otros parámetros, el modelo de BN a utilizar, método de discretización a emplear, método de propagación de evidencias por la BN, etc.

La Figura 74 es la IU encarga de establecer los valores a los parámetros del sistema.

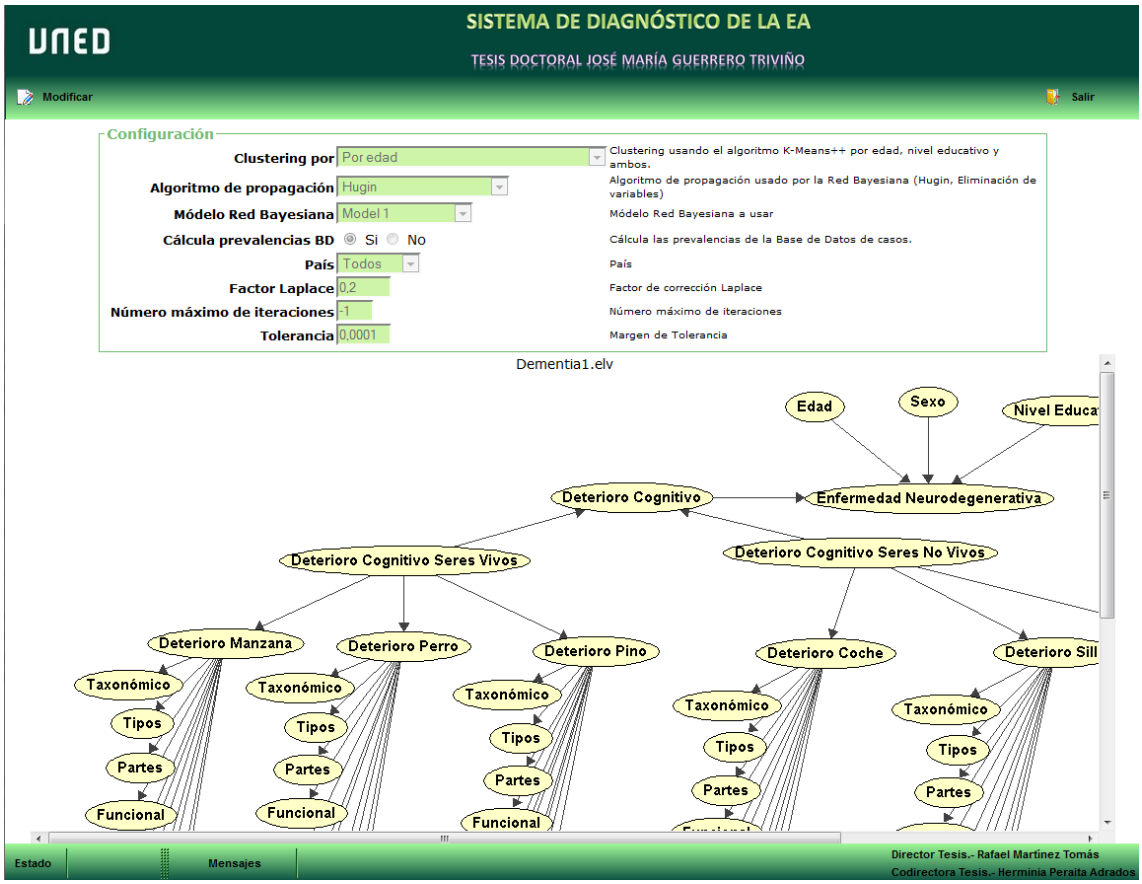


Figura 74.- IU para configuración del sistema.

El software lo se ha desarrollado con capacidad de internacionalización y actualmente todas las pantallas soportan el español y el inglés.

Apéndice

Experimentos complementarios

B

Han sido numerosos los experimentos realizados en esta tesis para validar el método de diagnóstico; se han utilizado distintas estrategias discretización, se ha optimizado la TPC de la variable EA con estrategias evolutivas, se han creado distintas BNs con distintas técnicas de modelado y se ha creado un algoritmo de aprendizaje automático que combina el corpus lingüístico con estudios epidemiológicos obtenidos de la literatura científica [18].

Se ha considerado muy importante en esta investigación tener en cuenta las relaciones informativas entre la edad, el sexo y nivel educativo, y la variable de interés ***EA***; dando lugar a una TPC con un gran número de parámetros para los que ha sido necesario utilizar el simplificador *Naive Bayes* para su aprendizaje**.** Por otro lado, como se ha venido indicando a lo largo de la tesis doctoral, la elaboración del corpus lingüístico [1] ha sido costosa, siendo necesaria la colaboración de varios departamentos de neurología de varios hospitales de la CAM. Por esta razón, no ha sido posible contar con un corpus de datos, para la optimización de la TPC de la variable ***EA***, y otro corpus de datos distinto, para realizar los experimentos. Del mismo modo se ha recurrido, en un modelo de BN, a la literatura científica [18] para aprender estos parámetros.

La estructura del apéndice es la siguiente. El objetivo de la sección B.1 es analizar la eficacia del método de diagnóstico, pero a diferencia de la sección 8.3, no utiliza técnicas de optimización. En objetivo de la sección B.2 es analizar la influencia de la edad y el nivel educativo en la producción oral de rasgos semánticos, pero a diferencia de la sección 8.2, utiliza el modelo 1 de BN en lugar del modelo 3. El objetivo de la sección B.3 es analizar la importancia de la segmentación de atributos en once bloques conceptuales, pero a diferencia de la sección 9.1, utiliza la ganancia de información en la CLG BN y añade un nuevo modelo de BN híbrido reducido con inferencia aproximada.

B.1. Eficacia del método de diagnóstico.

Este experimento es similar al de la sección 8.3 pero no se optimiza la TPC de la variable ***EA***. Para el modelo 1 de BN discreta "*Inferencia por razonamiento deductivo*" la TPC de la variable ***EA*** se calcula a partir del estudio epidemiológico [18] y para los

modelos 2 y 3 de BN "*Inferencia por razonamiento abductivo"*, esta TPC se calcula a partir del corpus lingüístico. Este experimento se realiza con *One-leave-out cross validation*.

En la Figura 75 se representa las curvas ROC generadas a partir de los modelos 1, 2 y 3 de BN (ver sección 5.3). En este experimento se segmenta la producción oral de rasgos semánticos por tramos de edad.

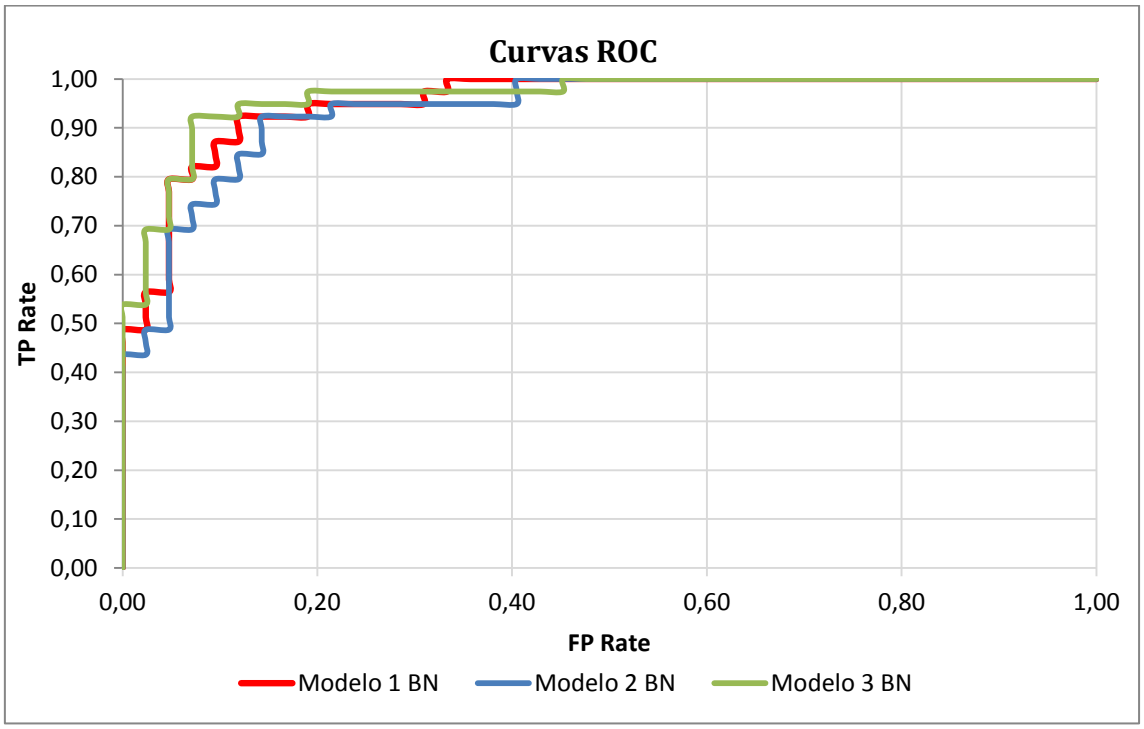


Figura 75.- Curva ROC obtenida a partir del modelo 1 de BN con segmentación de atributos por edad.

Al igual que en el capítulo 8, la BN que mejor rendimiento tiene es el modelo 3 de BN "*Inferencia por razonamiento abductivo y estudio del deterioro semántico diferencial entre los dominios SV y SNV*". Cabe destacar el rendimiento del modelo 1.

En la Tabla 51 se analizan otras métricas de rendimiento de este experimento.

Tabla 51.- Métricas de rendimiento del experimento 1 para los modelos 1, 2 y 3 de BN.

	Modelo 1	Modelo 2	Modelo 3
True Positive (TP)	36	36	36
True Negative (TN)	37	36	39
False Positive (FP)	5	6	3
False Negative (FN)	3	3	3
TP rate	0,9231	0,9231	0,9231
FP rate	0,119	0,1429	0,0714
Precisión	0,878	0,8571	0,9231
Exactitud	0,9012	0,8889	0,9259
Mean Squared Error	0,1201	0,1152	0,0957
Root Mean Squared Error	0,3465	0,3394	0,3094
AUC	0,953	0,9371	0,9628
Threshold	0,433433	0,560985	0,712579

B.2. Influencia de la edad y nivel educativo en la producción oral de rasgos semánticos.

El propósito de este experimento es similar al de la sección 8.2, pero se utiliza el modelo 1 de BN discreta en lugar del modelo 3. En este experimento se utiliza leave-one-out cross validation. Las probabilidades a posteriori analizadas son de la variable ***DSD*** y por tanto, dada la estructura de esta BN, los factores de contexto no tienen influencia sobre esta variable. Con este experimento se pretende confirmar, junto con el experimento de la sección 8.2, la mejora en el rendimiento que se obtiene al discretizar, segmentando los conglomerados de atributo por edad o nivel educativo.

En la Figura 76 se representan tres curvas ROC correspondientes a las tres estrategias de discretización por análisis de clúster llevada a cabo en este experimento. Cabe destacar que aunque se representan las tres curvas en la misma gráfica, cada curva se genera con un clasificador diferente.

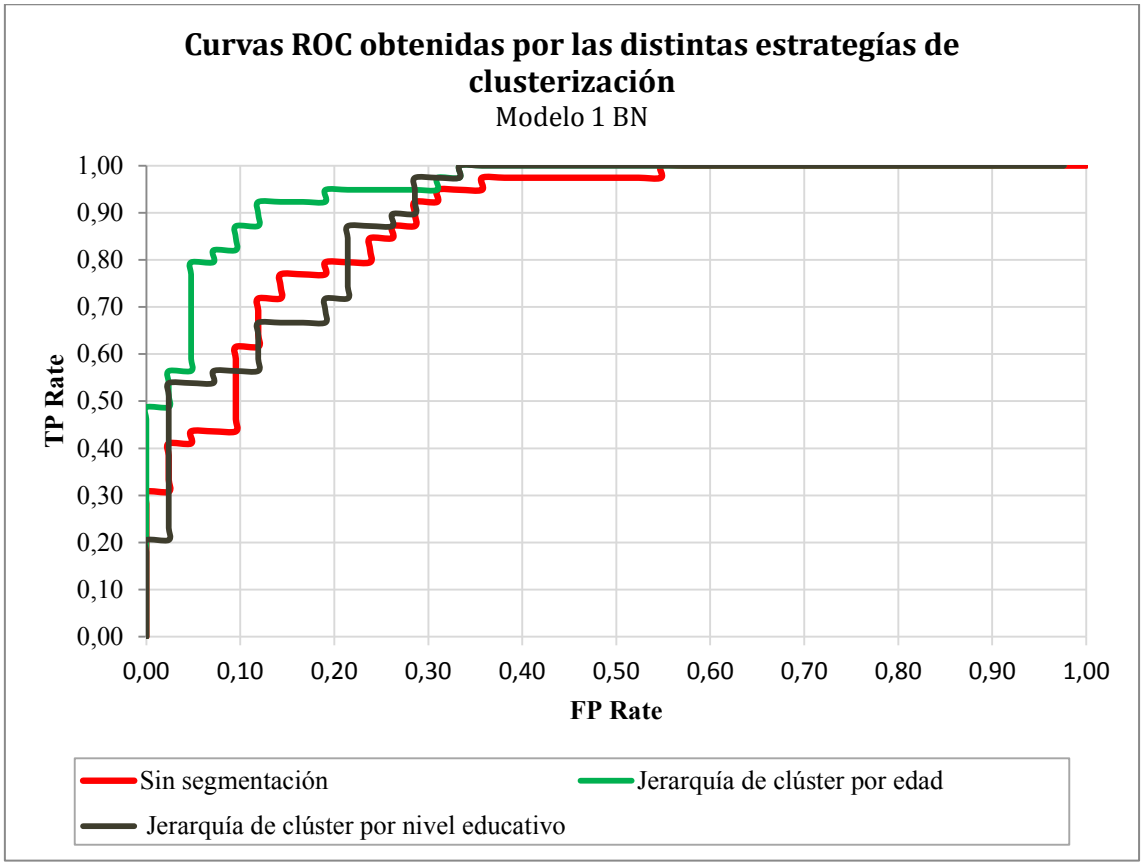


Figura 76.- Curvas ROC obtenidas por el modelo 3 y las distintas estrategias de discretización.

En la Tabla 52 se detallan las métricas de rendimiento obtenidas a partir de las distintas estrategias de discretización por análisis de clúster. Al igual que en la sección 8.2, en la Tabla 52 se pone de manifiesto que al crear una jerarquía de clúster por edad, se consiguen mejores resultados mientras que por nivel educativo el resultado es más controvertido.

Tabla 52.- Métricas de rendimiento obtenidas para las distintas estrategias de discretización por análisis de clúster.

	Sin segmentación	Por edad	Por nivel educativo
True Positive (TP)	37	36	38
True Negative (TN)	28	37	30
False Positive (FP)	14	5	12
False Negative (FN)	2	3	1
TP rate	0,9487	0,9231	0,9744
FP rate	0,3333	0,119	0,2857
Precisión	0,7255	0,878	0,76
Exactitud	0,8025	0,9012	0,8395
Mean Squared Error	0,1455	0,1201	0,1394
Root Mean Squared Error	0,3815	0,3465	0,3734
AUC	0,8895	0,953	0,898

B.3. Importancia de la segmentación de atributos en once bloques conceptuales.

Este experimento complementa al de la sección 9.1 y al igual que en la sección 9.1 se utiliza One-leave-cross validation en los experimentos. En la Figura 77 se representan cuatro curvas ROC: la curva de color rojo, es la curva obtenida a partir de la BN híbrida reducida sin segmentación de las definiciones orales en rasgos semánticos; la curva de color verde, se ha obtenido con una CLG BN que utiliza todas las variables del corpus (variables predictoras) y variables intermedias (variables latentes); la curva de color azul fuerte, se ha generado con una BN en la que se emplea un método de inferencia aproximada para las variables intermedias o latentes; y la curva de color celeste, se ha generado con una BN reducida la cual no segmenta las definiciones orales en rasgos semánticos y para las variables latentes utiliza un algoritmo de inferencia aproximado. En la CLG BN y en la BN Híbrida con inferencia aproximada se ha utilizado el ratio de la ganancia de información, para la inferencia de las variables latentes.

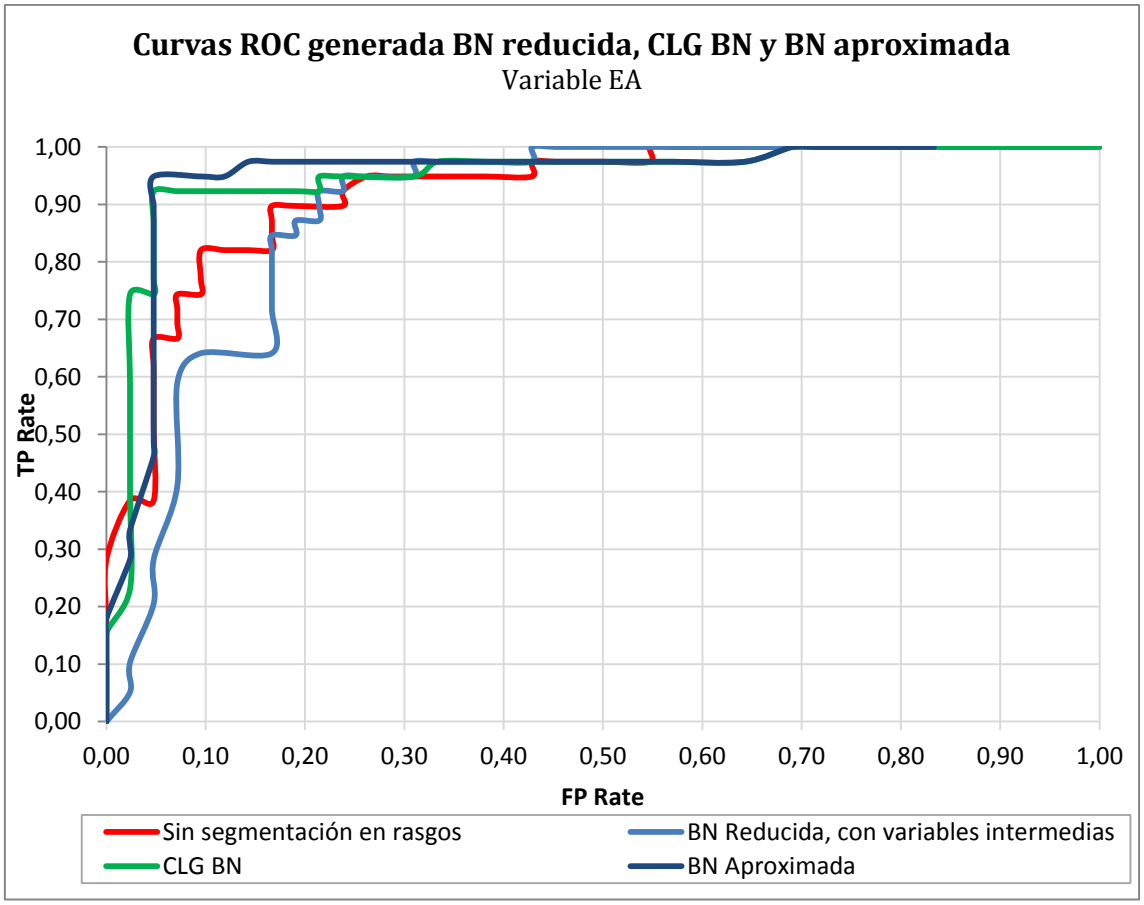


Figura 77.- Curvas ROC para comprobar la importancia de la segmentación de la producción oral de atributos lingüísticos en unidades menores y significativas (variable EA).

En la Tabla 53 se detallan las métricas de rendimiento obtenidas con las tres BNs del experimento.

Tabla 53.- Métricas de rendimiento para comprobar la importancia de la segmentación de la producción oral de atributos lingüísticos en unidades menores y significativas (variable EA).

	Sin segmentación en rasgos	Ganancia Información. CLG BN	Ganancia de Información. BN Aproximada	BN Reducida, con inferencia aproximada
True Positive (TP)	31	36	37	37
True Negative (TN)	40	39	38	31
False Positive (FP)	2	3	4	11
False Negative (FN)	8	3	2	2
TP rate	0,7949	0,9231	0,9487	0,9487
FP rate	0,0476	0,0714	0,0952	0,2619
Precisión	0,9394	0,9231	0,9024	0,7708
Exactitud	0,8765	0,9259	0,9259	0,8395
Mean Squared Error	0,0995	0,098	0,0928	0,1654
Root Mean Squared Error	0,3154	0,313	0,3046	0,4067
AUC	0,9402	0,9463	0,9481	0,8907

Al igual que en la sección 9.1, la segmentación de las definiciones oral en rasgos semánticos mejora la eficacia del clasificador.