Máster en Inteligencia Artificial Avanzada, Métodos y Aplicaciones

Técnicas de Aprendizaje Automático para el control de Depuradoras. Resolución del problema del Nitrógeno

REALIZADO POR:

María del Carmen Prieto Estravid

DIRIGIDO POR:

Félix Hernández del Olmo

Madrid, 2018

Hasta el infinito y más allá

Buzz Lightyear

AGRADECIMIENTOS

A todos los que me animaron a continuar.

A mi marido Julio Enrique, el mayor de mis apoyos, sin el que nunca habría llegado hasta aquí.

A Félix Hernández del Olmo por indicarme el camino correcto.

A Antonio Juano, al que conocí en un grupo de prácticas de la UNED y se ha convertido en uno de mis mejores amigos.

A Ángeles Manjarrés que siempre está dispuesta a ayudar aunque ya no seas su alumna.

A Javier Blasco por escuchar siempre las partes negativas.

A mis padres Miguel y Maruja que me enseñaron a perseverar, aunque no comprendan porque sigo estudiando a mi edad.

A mis hijos Jaime, Julio y Ernesto, que se han hecho mayores en mi largo periplo por la UNED.

Resumen

En 2015, la ONU aprobó la Agenda 2030 sobre el Desarrollo Sostenible, un conjunto de objetivos globales para erradicar la pobreza, proteger el planeta y asegurar la prosperidad para todos.

El Objetivo 6 consiste en garantizar la disponibilidad de agua y su gestión sostenible y el saneamiento para todos. El día Mundial del Agua 2017 se dedicó a la importancia del tratamiento de aguas residuales y a fomentar su reautilización.

La depuración eficaz de las aguas residuales se ve influenciada por la automatización en la captura de datos, ya que supone un aumento considerable de la información suministrada por los sensores para realizar los controles de las plantas. Disponer de toda la información, ordenarla y relacionarla de la manera más eficiente, mediante el análisis de datos, y obtener el conocimiento necesario para una correcta toma de decisiones, sirve para optimizar el proceso, predecir momentos críticos en la depuración y ajustar los costes de explotación y operación.

El 70 % de los costes totales de una EDAR se deben a la aireación en el proceso biológico por lo que resulta de vital importancia ajustar este proceso al máximo para conseguir que las aguas depuradas cumplan la legislación con el menor gasto posible.

Este trabajo analiza los datos proporcionados por el benchmark BSM1_LT durante un periodo de un año y medio para realizar una clasificación. En la primera parte se realiza una clasificación de los datos en función del tiempo atmosférico debido a que la cantidad de contaminantes se ve influenciada por volumen de agua a tratar y las características ambientales y esto afecta al volumen de oxígeno que ha de ser aportado por las soplantes para conseguir que el amonio final se encuentre dentro de los límites exigidos por la normativas.

Por otro lado se clasifican los valores de salida del nitrógeno orgánico o amoniacal, con ventanas temporales, para comprobar si el oxígeno suministrado es el adecuado o se puede modificar para optimizar los costes.

Para mejorar la operatividad de la planta se realiza una predicción del valor del amonio del efluente, tambien utilizando ventanas tempororales, lo que permitirá variar la cantidad de oxígeno proporcionado por las soplantes del sistema de aireación y ajustarlo a las necesidades del momento de manera eficiente.

Palabras Clave

Aprendizaje Automático, Clasificación Supervisada, Clasificación no Supervisada, ARIMA, Depuración de Aguas Residuales, Nitratos, BSM1_LT, Optimización Energética.

Abstract

In 2015, UN adopted the 2030 Agenda for Sustainable Development and its Sustainable Development Goals will mobilize efforts to finish all forms of poverty, tackle climate change, while ensuring that no one is left behind.

Goal 6 ensure access to water and sanitation for everyone. World Water Day 2017 theme was wastewater and the reduce and reuse of wastewater.

The effective purification of wastewater is influenced by the automation in data capturing, because it supposes a considerable increase of the information provided by the sensors to carry out the controls of the plants. Once you have all the information, order and relate it in the most efficient way, through data analysis, and obtain the necessary knowledge for a correct decision making, serves to optimize the process, predict critical moments in the debugging and adjust the costs of operation.

70 % of the total costs of an WWTP are due to the aeration in the biological process, so it is vital to adjust this process the most you can, in order to achieve that the treated water complies with the legislation with the lowest cost.

This paper analyzes the data provided by benchmark BSM1_LT during a period of one and a half years to make a classification. On one hand, a classification of the data according to the weather is carried out because of the quantity of pollutants is influenced by the volume of water to be treated and the environmental characteristics. This affects to the volume of oxygen that has to be contributed by the blowers to ensure that the final ammonium is within the limits required by regulations.

On the other hand, the output values of organic or ammoniacal nitrogen are classified using temporary windows, to check whether the oxygen supplied is adequate or can be modified to optimize costs.

To improve the operability of the plant, a prediction of the ammonium value of the effluent is made, with temporary windows too, which will allow to vary the amount of oxygen provided by the aeration system blowers and adjust it efficiently.

Key Words

Machine learning, Supervised learning, Clustering, ARIMA, WasteWater Treatment, Nitrate, BSM1_LT, Energy Efficiency.

Nomenclatura

AA Aprendizaje Automático

ACF Funciones de Autocorrelación

ADP Programación Adaptativa Dinámica

ANFIS Sistemas de Inferencias Borrosas Basados en Redes Adaptables

ARIMA Autoregressive Integrated Moving Average

ASM Modelos de Lodos Activados

ASM1 Modelo de Lodos Activados nº 1

BP algoritmo de aprendizaje de propagación hacia atrás

BSM1 Benchmark Simulation Model nº 1

BSM1-LT Long-Term Benchmark Simulation Model nº 1

CC red neuronal artificial Cascada Correlación

COST European Cooperation in the field of Scientific and Technical Research

CV Validación Cruzadar

DMC Matriz Dinámica

DO Oxígeno Disuelto

DQO Demanda Química de Oxígeno

DTFT Transformada de Fourier de Señales Discretas

ECMN Error Cuadrático Medio Normalizado

EDAR Estación Depuradora de Aguas Residuales

GA Algoritmos Genéticos

GA-NFS Sistema Difuso Neuronal basada en Algoritmo Genético

IWA International Water Association

k-NN K Nearest Neighbour

LDA Análisis Discriminante Lineal

MGM Método Multi-Gradiente

MINLP Programación No Lineal Entera Mixta

MLP Perceptrón Multicapa

MLPANN Redes Neuronales Artificiales con Perceptrón Multicapa

MLVSS Sólidos Suspendidos Volátiles en Licor de Mezcla

MPC Control Predictivo

N Nitrógeno

NH4-N Nitrógeno Amoniacal

NMMPC Modelo Predictivo con Control Multiobjetivo no Lineal

NNOMC Método de Control y Modelado Online con Redes Neuronales

ORP Potencial de Oxidación Reducción

OSMC Open Source Modelica Consortium

P Fósforo

PACF Funciones de Autocorrelación Parcial

PAO microoganismo acumulador de fosfatos

PCA Análisis de las Componentes Principales

RBF Función de Base Radial

RBFANN Redes Neuronales Artificiales en Función de Base Radial

RF Random Fores

RNA Red Neuronal Artificial

RSE Error Cuadrático Medio

RVM Máquina de Vectores de Relevancia rápida

SCLD Sistema de Control basado en Lógica Difusa

SE Sistema Experto

SORBF Red Neuronal con función de auto-organización de base radial

SS Sólidos en Suspensión

SVI Índice Volumétrico de Lodos

SVM Máquinas de Vectores Soporte

T temperatura

TKN nitrógeno total Kjeldahl

TN Nitrógeno Total

TP Fósforo Total

UE Unión Europea

Índice general

Índice de figuras

Índice de tablas

ÍNDICE DE TABLAS

Capítulo 1

Introducción

1.1. Contexto

En el año 2000, patrocinado por las Naciones Unidas, los lideres mundiales se comprometieron a buscar formas de combatir la pobreza en todas sus formas. Esto se tradujo en un marco de actuación denominado Objetivos de Desarrollo del Milenio, compuesto por ocho grandes campos de actuación que abarcaban temas tan fundamentales como erradicar la pobreza extrema, conseguir la educación universal, promover la igualdad de género, disminuir la mortalidad infantil o garantizar la sostenibilidad del medio ambiente. En el año 2015 tras finalizar su ciclo de aplicación se pudo comprobar que el planteamiento de objetivos ambiciosos había conseguido involucrar a buena parte de la sociedad consiguiendo que hasta los países más desfavorecidos alcanzaran un progreso sin precedentes.

El Objetivo 7, garantizar la sostenibilidad del medio ambiente, incluía el acceso al agua potable y al saneamiento. Según el informe publicado en 2015 por Naciones Unidas [47] :

- En 2015, 91 % de la población mundial utiliza una fuente de agua mejorada, en comparación al 76 % en 1990.
- Desde 1990, de los 2.600 millones de personas que obtuvieron acceso a fuentes de agua potable mejorada, 1.900 millones lo hicieron a través de agua potable suministrada por cañería hasta su propio hogar. Más de la mitad de la población mundial (58 %) ahora disfruta de este nivel más alto de servicio.
- En todo el mundo, 147 países han cumplido con la meta del acceso a una fuente de agua potable, 95 países han alcanzado la meta de saneamiento y 77 países han cumplido ambas.

- A nivel mundial, 2.100 millones de personas han obtenido acceso a saneamiento mejorado. El porcentaje de personas que defecan al aire libre se ha reducido casi a la mitad desde 1990.

A pesar de estos grandes avances la escasez de agua afecta a más del 40 % de la población mundial, y se estima que esto aumentará.

Tras concluir el periodo de los Objetivos del Milenio las Naciones Unidas adoptaron la Agenda 2030 para el Desarrollo Sostenible que incluye 17 ambiciosos objetivos [48]. El Objetivo 6 consiste en garantizar la disponibilidad de agua, su gestión sostenible y el saneamiento para todos, para ello hay que conseguir reducir a la mitad la proporción de aguas residuales no tratadas y aumentar el reciclaje de agua y su reutilización segura. En la actualidad, más de 2 mil millones de personas sufren problemas por la escasez de agua, problema que aumentará con el crecimiento de la población y los efectos del cambio climático.

Para intentar concienciar a la sociedad cada año, el 22 de Marzo, se celebra el Día Mundial del Agua para centrar la atención en el manejo sostenible de los recursos de agua dulce. El Día Mundial del Agua 2017 se dedicó a la importancia del tratamiento de aguas residuales y a fomentar su reutilización. En palabras de la Directora General de la UNESCO, Irina Bokova.

> Limitar la liberación en la naturaleza de aguas residuales sin tratar no solo salva vidas y mejora la salud de los ecosistemas, sino que, además, puede contribuir a fomentar el crecimiento sostenible.....Ante la demanda creciente, las aguas residuales pueden constituir una alternativa fiable como fuente de abastecimiento de agua; para ello es preciso cambiar el paradigma de la gestión de las aguas residuales, pasando de “tratar y desechar” a “reducir, reutilizar, reciclar y recuperar”. Las aguas residuales ya no deberían verse como un problema, sino como parte de la solución para problemas a los que se están enfrentando todas las sociedades. Debemos situar la mejora de la gestión de las aguas residuales en el centro de una economía circular, logrando un equilibrio entre el desarrollo y la protección y el uso sostenible de los recursos naturales. Ello aportará amplios beneficios, con repercusiones en la seguridad alimentaria y energética y en la atenuación de los efectos del cambio climático.

1.2. Aguas Residuales

Aunque las bases modernas de la depuración de aguas no se empezaron a crear hasta principios el siglo XX, la idea del saneamiento y la conducción de aguas residuales ya era conocida en al antigüedad. Como ejemplo destacable podemos señalar la red de saneamiento con la que contaba la ciudad de Roma, en el año el 600 A.C. y denominada «Cloaca Máxima», que vertía los residuos de la ciudad al río Tíber.

La construcción de las primeras redes de alcantarillado puso de manifiesto que aunque contribuían a la reducción del número de puntos de vertido, mejorando las condiciones locales respecto a la situación anterior, se producía una mayor concentración de la contaminación en el punto de vertido, lo que produjo un agravamiento del estado de los ríos, creando condiciones higiénicas y ambientales inaceptables. Debido a esto sugirió la idea de que el vertido de aguas residuales no debería realizarse directamente en los ríos sino que debería utilizarse para fertilizar el suelo, de esta manera se completaba el anterior concepto de saneamiento y se proponía el primer sistema de tratamiento basado en la recogida y transporte del agua residual.

A partir de este punto, se desarrollan los primeros sistemas de depuración, inicialmente dirigidos a la eliminación de materias sólidas y posteriormente complementados con la de la materia orgánica soluble mediante los tratamientos biológicos, utilizando primero los filtros percoladores en 1897 y, posteriormente los fangos activados, sistema desarrollado por Arden y Locket en 1914.

Hacia finales del siglo XIX se estableció en Massachusetts la estación experimental Lawrence para el estudio del tratamiento del agua de abastecimiento y de las aguas residuales, muchos de los métodos posteriores de tratamiento intensivo se basaron en los preparados en Lawrence [43].

En el año 1927 en Apeldoorn (Holanda) se trataron las aguas residuales procedentes de un matadero usando un proceso de lodos activados y antes de la 2ª Guerra Mundial ya existían plantas de este tipo en Europa, América, Japón, la India o Sudáfrica.

La historia de modelado de sistemas de lodos activados se puede dividir en tres períodos [41]:

1º Periodo. Criterio Empírico. El período comprendido entre el descubrimiento proceso y principios de la década de 1950 se caracterizó por el diseño empírico y las conjeturas. Los métodos de diseño iniciales de los tanques de lodos activados eran simples y fácilmente modificables buscando adaptarse a los requerimientos o condiciones del momento.

2º Periodo. Estado estacionario. Con utilización de crecimiento microbiano y sustrato orgánico. El periodo entre los años 1950 y 1980 se caracteriza por el uso de la cinética química para relacionar (en estado estacionario) el crecimiento microbiano y la utilización de sustratos orgánicos en condiciones aeróbicas. La mayoría de los estudios cinéticos iniciales se originaron a partir de la cinética química y se centraron en la solución de problemas particulares, bien definidos, tales como el crecimiento microbiano en condiciones de estado estacionario con las condiciones ambientales totalmente controladas. Esto dio lugar a modelos que tenían poca aplicabilidad y eran incapaces de predecir las diversas reacciones de adaptación bajo condiciones ambientales cambiantes.

3º Periodo. Uso de modelos dinámicos complejos. Desde 1980 hasta la actualidad los modelos alcanzan una gran complejidad gracias a la aplicación de principios de ingeniería del reactor en combinación con grandes matrices de expresiones cinéticas y constantes estequiométricas. Se ha conseguido definir el comportamiento de los componentes del influente de aguas residuales (fracciones orgánica, de nitrógeno (N) y de fósforo (P)) y componentes de la biomasa (heterótrofos, nitrificantes y PAOs (microorganismos acumuladores de fosfatos) con otras subdivisiones) y la estequiometría de la reacción y la cinética.

1.2.1. Depuración de aguas residuales urbanas

El vertido de aguas residuales urbanas sin depurar ejerce sobre los cauces receptores toda una serie de efectos negativos, de entre los que cabe destacar [56]:

- Aparición de fangos y flotantes. Los sólidos en suspensión sedimentables originan acumulación de residuos en el fondo de los cauces. Además, los no sedimentables pueden formar capas flotantes en la superficie y las orillas de los ríos. Los depósitos de fangos y flotantes pueden llegar a terminar con el oxigeno disuelto en el agua y originar malos olores.

- Agotamiento del contenido de oxígeno presente en las aguas. Los componentes de las aguas residuales fácilmente oxidables comenzarán a ser degradados vía aerobia por la flora bacteriana de las aguas del cauce, si el contenido en oxígeno disuelto disminuye por debajo de los valores mínimos necesarios para el desarrollo de la vida acuática puede generar problemas en los ecosistemas fluviales. Los procesos de degradación vía anaerobia generarán olores desagradables al liberarse gases.

- Aportes excesivos de nutrientes. Las aguas residuales contienen nutrientes (N y P principalmente) causantes del crecimiento descontrolado de algas y otras plantas en los cauces (eutrofización) que puede llegar a impedir el empleo de estas aguas para usos domésticos e industriales.

- Daños a la salud pública. Los vertidos de aguas residuales sin tratar a cauces públicos pueden fomentar la propagación de organismos perjudiciales para el ser humano (virus, bacterias, protozoos y helmintos). Entre las enfermedades que pueden propagarse a través de las aguas contaminadas por los vertidos de aguas residuales urbanas, destacan: el tifus, el cólera, la disentería o la hepatitis A.

1.2.2. Tratamientos de las aguas residuales

Las estaciones depuradoras sirven para eliminar una elevada proporción de los contaminantes presentes en las aguas residuales, vertiendo efluentes depurados, que puedan ser asimilados de forma natural por los cauces receptores.

La legislación actual (Real Decreto-Ley 11/95) contempla en la depuración de aguas tres tipos de tratamientos [35]:

1.2.2.1. Tratamientos Primarios

El Real Decreto-Ley 11/95 define al tratamiento primario como:

> El tratamiento de aguas residuales urbanas mediante un proceso físico o fisicoquímico que incluya la sedimentación de sólidos en suspensión, u otros procesos en los que la DBO_5(Demanda Bioquímica de Oxígeno a los 5 días) de las aguas residuales que entren, se reduzca, por lo menos, en un 20 % antes del vertido, y el total de sólidos en suspensión en las aguas residuales de entrada se reduzca, por lo menos, en un 50 %.

El principal objetivo de los tratamientos primarios se centra en la eliminación de sólidos en suspensión, consiguiéndose además una cierta reducción de la contaminación biodegradable, dado que una parte de los sólidos que se eliminan está constituida por materia orgánica. Los tratamientos primarios más habituales son la decantación primaria y los tratamientos fisicoquímicos.

- Decantación primaria: su objetivo es la eliminación de la mayor parte posible los sólidos sedimentables, bajo la acción exclusiva de la gravedad. La retirada

de estos sólidos es muy importante ya que, en caso contrario, originarían fuertes demandas de oxígeno en el resto de las etapas de tratamiento de la estación.

- Tratamientos fisicoquímicos: en este tipo de tratamiento, mediante la adición de reactivos químicos, se consigue incrementar la reducción de los sólidos en suspensión, al eliminase además sólidos coloidales, al aumentar el tamaño y densidad de los mismos mediante procesos de coagulación-floculación.

1.2.2.2. Tratamientos Secundarios

El Real Decreto-Ley 11/95 define tratamiento secundario como:

> El tratamiento de aguas residuales urbanas mediante un proceso que incluya un tratamiento biológico con sedimentación secundaria u otro proceso en el que se consiga la eliminación de materia orgánica.

El tratamiento biológico se realiza con la ayuda de microorganismos, fundamentalmente bacterias, que en condiciones aerobias actúan sobre la materia orgánica presente en las aguas residuales gracias a la acción metabólica y enzimática de los microorganismos, consiguiendo así la formación de flóculos con peso suficiente para poder separarse de la masa de agua.

El proceso biológico es un mundo ecológico en si mismo donde se obtiene un rendimiento óptimo en unas situaciones concretas de caudal y de carga. El equipo encargado del sistema no tiene que preocuparse del mecanismo funcional, es el sistema biológico el que se adaptará a las modificaciones.

El aporte de oxígeno para el mantenimiento de las reacciones de oxidación, síntesis y respiración endógena, se efectúa introduciendo, generalmente aire en los recipientes en que se llevan a cabo estas reacciones, recipientes que se conocen con el nombre de reactores biológicos o cubas de aireación. Las nuevas bacterias que van apareciendo en los reactores, como consecuencia de las reacciones de síntesis, tienden a unirse (floculación), formando agregados de mayor densidad que el líquido circundante, y en cuya superficie se va adsorbiendo la materia en forma coloidal.

Para la separación de estos agregados, conocidos como lodos o fangos, el contendido de los reactores biológicos (licor de mezcla) se conduce a una etapa posterior de sedimentación (decantación o clarificación secundaria), donde se consigue la separación de los lodos de los efluentes depurados por la acción de la gravedad.

De los lodos decantados una fracción se purga como lodos en exceso, mientras que otra porción se recircula al reactor biológico para mantener en él una concentración determinada de microorganismos. Este proceso se conoce como lodos activados.

1.2.2.3. Tratamientos Terciarios

Este tipos de tratamiento se contemplan en el Real Decreto-Ley 11/95 como tratamiento de aguas para zonas sensibles.

Los tratamientos terciarios permiten obtener efluentes finales de mejor calidad para que puedan ser vertidos en zonas donde los requisitos son más exigentes o puedan ser reutilizados.

La finalidad de los tratamientos terciarios es eliminar la carga orgánica residual y aquellas otras sustancias contaminantes no eliminadas en los tratamientos secundarios, como por ejemplo, los nutrientes, fósforo y nitrógeno. Estos procesos son de naturaleza biológica o físicoquímica, siendo el proceso unitario más empleado el tratamiento fisicoquímico que consta de una coagulación – floculación y una decantación. Para la eliminación de nutrientes (nitrógeno y fósforo), se recurre cada vez más al empleo de procesos biológicos, sin embargo en el caso del fósforo, los procesos de precipitación química empleado sales de hierro y de aluminio, continúan siendo los de mayor aplicación.

1.2.3. Eliminación del nitrógeno

En la eliminación biológica de nitrógeno se opera de forma secuencial, bajo condiciones óxicas y anóxicas, que dan como resultado final su liberación a la atmósfera en forma de nitrógeno gaseoso.

Según la Directiva 91/271/CCE, para el caso de vertidos de instalaciones de tratamiento de aguas residuales urbanas realizados en zonas sensibles cuyas aguas sean eutróficas o tengan tendencia a serlo en un futuro próximo, se deberán cumplir las condiciones descritas en la Tabla 1.1.

Los requisitos para instalaciones individuales pueden no aplicarse, si la reducción de la carga total de todas las instalaciones que vierten a la zona sensible es del 75 % para el N total.

No obstante, las autorizaciones de vertido de las instalaciones de tratamiento de aguas residuales urbanas, podrán imponer requisitos más rigurosos, cuando ello sea necesario para garantizar que las aguas receptoras cumplan con los objetivos de calidad fijados en la normativa vigente.

Según el séptimo informe sobre la aplicación de la Directiva, publicado en el año 2013 [18], la contaminación era importante en el 22 % de las masas de agua de la Unión Europea (UE) y la eutrofización afectaba a casi el 30 % de las aguas en 17 de los Estados miembros.

Tabla 1.1: Requisitos para los vertidos procedentes de instalaciones de depuración de aguas residuales urbanas mediante tratamiento más riguroso

Parámetros	Concentración		Porcentaje mínimo de reducción
	10.000 a 100.000 h-e	> 100.000 h-e	
Nitrógeno total	15 mg/L N	10 mg/L N	70-80 %

Gracias a la aplicación de esta normativa la exportación total de nitrógeno se ha reducido un 9 %, principalmente debido a una reducción de las emisiones de fuentes puntuales, pero a pesar de esto, basándose en estas estimaciones, las cargas totales anuales de contaminación procedentes de las aguas residuales urbanas que incumplían la Directiva fueron de aproximadamente 603 KT/año de nitrógeno, 78 KT/año de fósforo y 3.900 KT/año de contaminación orgánica total. El nitrógeno generado por la fracción de las aguas residuales que incumplían la Directiva representa aproximadamente el 15 % del nitrógeno total vertido a los mares.

1.3. Automatización en el tratamiento de aguas residuales

La primera conferencia de Ingeniería en Automatización y Control Industrial, bajo el patrocinio de la predecesora de International Water Association (IWA), la International Association on Water Pollution Research (IWPR), se celebró en Londres en 1973.

Las primeras aplicaciones en la década de 1970, se realizaron en procesos de lodos activados para la eliminación de materia orgánica especialmente, en la Universidad de Ciudad del Cabo bajo la dirección del Prof. Marais. Los requerimientos de salida del efluente fueron principalmente para controlar la demanda bioquímica de oxígeno (DBO) y lo sólidos en suspensión, mientras que no se consideraba la eliminación de nitrógeno ni de fósforo. Esto se canalizó en 1982 a través del Grupo de Trabajo sobre Modelización de Lodos Activados de la IWA, consiguiendo que de los fenómenos biológicos y las reacciones fisicoquímicas responsables de la eliminación de compuestos orgánicos de carbono, nitrógeno y fósforo se hayan traducido en los Modelos de Lodos Activados (ASM) [32]. Dentro de los diferentes modelos propuestos para describir los fenómenos biológicos que tienen lugar en el proceso de fangos activados, los modelos propuestos por la IWA se han convertido en un estándar de facto. Los modelos de la familia ASM (ASM1, ASM2, ASM2d, ASM3) son utilizados en muchos de los estudios de investigación basados en simulación. Estos modelos no solo han mejorado

la comprensión de los procesos clave, sino que también han proporcionado un lenguaje y nomenclatura común e implementaciones verificadas.

Durante las tres últimas décadas la implementación de sistemas de control industrial ha evolucionado de la tecnología analógica a la tecnología digital. Sin embargo, la teoría de control que estaba disponible en la década de 1970 todavía pueda resolver la mayoría de los problemas de control de procesos en el tratamiento de aguas residuales [49].

El objetivo a principios de la década de 1990 era utilizar el sistema para la calibración de modelos en línea automatizada, validación de datos, diagnóstico y control de procesos [52]. Para realizar esto con rigor es necesario implementar unas estrategias de control.

La simulación proporciona un medio apropiado, de bajo coste y efectivo, para la evaluación de estrategias de control y operación. El problema planteado en un proceso tan complejo como una Estación Depuradora de Aguas Residuales (EDAR) sugirieron la necesidad de una cierta estandarización respecto al protocolo a seguir para poder comparar estrategias de control de una manera efectiva. Para poder realizar una comparación efectiva, cada estrategia de control debe ser simulada bajo el mismo escenario y las mismas condiciones.

La idea de producir un "punto de referencia de simulación" estandarizado fue sugerida por primera vez por Bengt Carlsson (Universidad de Uppsala, Suecia) en la Conferencia ICA de 1993 en Hamilton, Canadá. Esta idea fue desarrollada por el primer Grupo IAWQ sobre Control Basado en Respirometría del Proceso de Lodos Activados [62] y posteriormente modificada por los Grupos de Trabajo del European Cooperation in the field of Scientific and Technical Research (COST) Action 624 y 628 [16] .

Utilizando como índice de referencia las acciones COST se realizan simulaciones con el banco de pruebas Benchmark Simulation Model nº 1 (BSM1) , que consiste en un sistema con un método de control asociado, un procedimiento de evaluación comparativa y un criterio de evaluación. BSM1 usa el modelo de lodos activados ASM1 para el modelado de las reacciones biológicas. En el modelo BSM1 el periodo de evaluación es de 14 días.

BSM1 presenta algunas limitaciones como trabajar con un periodo de evaluación demasiado corto o no tener en cuenta el efecto estacional. Para intentar solventar estos problemas se planteó el Long-Term Benchmark Simulation Model nº 1 (BSM1_LT) [25]. El diseño BSM1_LT es una extensión del BSM1 en el que el periodo de evaluación

se incrementa de manera significativa, a un año, y además se considera la temperatura variable, lo que afecta a varios parámetros.

Como el efecto de estacionalidad influye notablemente en las operaciones de tratamiento de aguas residuales es importante realizar una monitorización de al menos un año y comenzar el periodo de evaluación en un mes de verano para minimizar los riesgos del estudio. Al aumentar de manera tan grande el periodo de evaluación en BSM1_LT respecto a BSM1, 364 frente a 14 días, es necesario tener en cuenta también una variación en la temperatura que en BSM1 se considera constante e igual a 15ºC. BSM1_LT considera una variación sinusoidal de la temperatura con valores máximos en torno a 20.5º a principios de agosto y mínimos, sobre 9.5º, a principios de febrero. El coeficiente de transferencia de oxígeno Kla depende de la temperatura por lo que será uno de los factores a tener en cuenta a la hora de diseñar la planta.

En el modelo BSM1, la inicialización consiste en simular la planta a evaluar durante 100 días con entrada constante, seguido de dos semanas de datos de entrada dinámica de tiempo seco, lluvioso o tormentoso. Esto se hace para obtener un estado casi constante y poder realizar un buen estudio posterior. El período estacionario debe ser más prolongado en BSM1_LT debido a que las estrategias de control que actúan en escalas de tiempo grandes, necesitan también una fase inicial extensa, por lo que se inicializará durante de seis meses para alcanzar un estado cuasi estacionario. Para que el proceso sea más realista se incluyen eventos lluviosos y tormentosos durante el periodo de inicialización que va desde el inicio de enero hasta finales de junio, y debe de ser diferente de los últimos 6 meses del período de evaluación.

Existen múltiples lenguajes enfocados a la realización de simulaciones entre las que cabe destacar Modelica [22], un lenguaje de modelado libre, orientado a objetos y aplicable a variados campos de la ingeniería. Modelica contiene herramientas para definir las relaciones entre los distintos componentes y las particularidades de cada componente en sí. Entre sus bibliotecas hay una especifica para las aguas residuales, WasteWater, adaptada a los modelos ASM. Su especificación es abierta, lo que permite que existan sistemas comerciales como Dymola de Dassault Sistemes, MapleSim de Maplesoft o MathModelica de MathCore o de código abierto, tales como JModelica, desarrollada por ModelonAB y OpenModelica del Open Source Modelica Consortium (OSMC), con aplicaciones tanto académicas como de uso industrial.

1.4. Aprendizaje Automático

El Aprendizaje Automático (AA) , o Machine Learning por su nombre en inglés, es la rama de la IA que tiene como objetivo desarrollar técnicas que permitan aprender a las computadoras. De forma más concreta, se trata de crear algoritmos capaces de generalizar comportamientos y reconocer patrones a partir de una información suministrada en forma de ejemplos. Es un proceso de inducción del conocimiento, es decir, un método que permite obtener un enunciado general a partir de enunciados que describen casos particulares. Podríamos decir que una de las tareas del AA es intentar extraer conocimiento sobre algunas propiedades no observadas de un objeto basándose en las propiedades que sí han sido observadas de ese mismo objeto (o incluso de propiedades observadas en otros objetos similares), es decir, predecir comportamiento futuro a partir de lo que ha ocurrido en el pasado. En el AA se considera aprendizaje a aquello que la máquina pueda aprender a partir de la experiencia, no a partir del reconocimiento de patrones programados a priori.

En el año 1997 Mitchel [44] definió el AA como:

> Se dice que un programa aprende de la experiencia con respecto a cierta clase de tareas T y una medida del rendimiento P si la medida del rendimiento de tareas en T medidas por P aumenta con la experiencia E.

Aplicando esta definición a la depuración de aguas y más concretamente al control de la cantidad de amonio presente en el flujo efluente de vertido podríamos definir los conceptos:

- Tarea T: obtener vertidos de aguas residuales que se adapten a la legislación vigente.
- Medida del rendimiento P: porcentaje de vertidos que se encuentren por debajo del límite.
- Experiencia de entrenamiento E: datos de salida de la EDAR.

El AA se divide en tres tipos [40]:

1. Aprendizaje supervisado, donde se va dirigiendo al sistema en el proceso de entrenamiento, se genera una función que establece una correspondencia entre las entradas y las salidas deseadas del sistema, donde la base de conocimientos del sistema está formada por ejemplos etiquetados a priori.

2. Aprendizaje no supervisado, donde no se corrige al sistema en su proceso de entrenamiento, el proceso de modelado se lleva a cabo sobre un conjunto de ejemplos formados únicamente por entradas al sistema, sin conocer su clasificación correcta, por lo que se busca que el sistema sea capaz de reconocer patrones para poder etiquetar las nuevas entradas.

3. Aprendizaje por refuerzo, en el que no se es importante el valor de salida, sólo si ha clasificado bien o no. En este caso el algoritmo aprende observando el mundo que le rodea y con un continuo flujo de información en las dos direcciones (del mundo a la máquina, y de la máquina al mundo) realizando un proceso de ensayo-error, y reforzando aquellas acciones que reciben una respuesta positiva en el mundo.

En este trabajo nos vamos a centrar en los dos primeros, aunque trabajaremos principalmente con el aprendizaje de tipo supervisado.

1.5. Objetivos y alcance del TFM

El principal objetivo del presente trabajo consiste en el desarrollo de una técnica de AA para el control de depuradoras que lleve a cabo la supervisión del Nitrógeno para adaptar el vertido a las condiciones ambientales necesarias.

Para cumplir este objetivo se utilizará software libre. Las simulaciones se realizarán con Open Modelica y la mencionada biblioteca WasteWater y para realizar las clasificaciones y predicciones se utilizará el lenguaje de programación Python con Scikit-learn, que es una biblioteca de AA, de software libre, que cuenta con varios algoritmos de clasificación, regresión y agrupación. Para realizar las estimaciones se usará Statsmodels, un módulo de Python que proporciona clases y funciones para la estimación de muchos modelos estadísticos diferentes, así como para realizar pruebas estadísticas y explorar los datos.

En la primera parte de este trabajo se va a realizar una clasificación de los datos en función del tiempo atmosférico ya que la cantidad de contaminantes varía en cada caso y esto influye en la cantidad de oxígeno que ha de ser aportado por las soplantes para conseguir que el amonio final se encuentre dentro de los límites exigidos por la normativas ($S_{NH} < 4\ gNm^{-3}$). Durante el tiempo lluvioso se produce un aumento de caudal sin añadir más contaminantes por lo que la cantidad de contaminantes totales en este tipo de eventos disminuye al verse diluidos por el aumento del caudal, en el

tiempo tormentoso se produce tanto un aumento del caudal como de los contaminantes presentes.

Tras hacer un preprocesamiento de los datos, se realizarán clasificaciones supervisada y no supervisada con una validación posterior utilizando varios métodos.

Después de esta clasificación se efectuará otra en función del amonio de salida para ver si se adapta o no a la normativa de vertidos. Para ello será necesario realizar previamente una simulación con BSM1_LT, utilizando OpenModelica y su biblioteca WasteWater, que se ejecuta durante 609 días (63 estabilización + 182 entrenamiento+ 364 evaluación).

Con los datos de salida se realizan clasificaciones utilizando un conjunto de entrenamiento que va variando día a día según se van añadiendo los datos del día en curso, para que el modelo se adapte a la forma en que se trabajaría con una EDAR real.

Hasta el momento se han buscado clasificaciones del valor del nitrógeno efluente en función de si se encontraba o no dentro de los límites admitidos en la legislación de vertidos. Un paso adicional consiste en predecir el valor que tendrá el amonio de salida y no solo si es admisible, para poder ajustar de manera más eficiente el oxigeno suministrado por las soplantes. Para ello se intentará elaborar una predicción para que el control de planta se pueda realizar de una manera más eficiente. Para ejecutar la predicción se procede a eliminar la tendencia y la estacionalidad y se aplica el modelo AutoRegressive Integrated Moving Average (ARIMA) que nos sirve en una serie temporal para proyectar su valor en el futuro inmediato.

1.6. Contenido de la Memoria

El contenido de este trabajo está distribuido en seis capítulos, con los siguientes temas:

En el capitulo primero se desarrolla esta introducción. En el segundo se realiza una revisión bibliográfica. En el tercero se exponen las características de los benchmark BSM1 y BSM1_LT. El cuarto capítulo muestra las clasificaciones realizadas con el tiempo atmosférico. En el quinto se realiza una predicción del nivel de amonio de salida utilizando distintos métodos. El sexto capítulo finaliza con las conclusiones y posibles trabajos futuros.

Capítulo 2

Revisión de la Bibliografía

2.1. Introducción

La depuración de las aguas residuales es un tema prioritario en la sociedad actual. El control de estos procesos es fundamental, tanto por motivos ecológicos como por cuestiones de ahorro energético.

Una EDAR es una instalación destinada a transformar las aguas residuales procedentes de vertidos sanitarios o industriales en aguas «limpias» para el medio ambiente que se ajusten a la legislación tanto española como europea.

Es un hecho que el vertido de aguas residuales sin depurar ocasiona daños, en ocasiones irreversibles, al medio ambiente, por otro lado, el vertido de aguas residuales no tratadas supone riesgos para la salud pública. Es por esto por lo que es preciso el tratamiento de estas aguas antes de su vertido.

El principal problema actual para afrontar estos tratamientos es de tipo económico, no solo por los costes de construcción de las plantas sino fundamentalmente por los costes de mantenimiento y explotación. En el caso español la depuración de los 4.000 Hm^3/año de aguas residuales urbanas supone el 1 % del consumo energético nacional.

En el tratamiento de las aguas residuales éstas se someten a una serie de procesos físicos, químicos y biológicos que tienen por objeto reducir la concentración de los contaminantes y permitir el vertido de los efluentes depurados, minimizando los riesgos tanto para el medio ambiente, como para las poblaciones. En las grandes y medianas aglomeraciones urbanas el procedimiento más habitual para el tratamiento de los vertidos líquidos se conoce como «lodos activados», que desde sus primeras aplicaciones a principios del siglo XX se ha convertido en el tratamiento más extendido a nivel mundial

.

2.2. Sistemas de control

Durante las tres últimas décadas la implementación de sistemas de control industrial ha evolucionado de la tecnología analógica a la tecnología digital.

A partir de 1950 empezó la introducción de controladores automáticos en el dominio industrial, iniciándose la nueva era del control automático. Estos controladores se denominaron PID, porque reaccionaban al error ya producido de forma Proporcional (P), Integral (I) y Derivativa (D). El control de un proceso con las tres acciones de un PID tiene una dinámica compleja, el resultado puede conllevar problemas de estabilidad ya que las variables críticas del proceso son difíciles de controlar. En muchos casos se prefiere dejar el proceso bajo control manual como cuando existen retardos puros. Aunque este controlador es útil en muchos casos, en otros, su rendimiento es pobre o inadecuado, y en cualquier caso tienen que ser ajustados por operadores expertos.

El campo del Aprendizaje Automático es uno de los que más aportes ha realizado al control optimizado de una EDAR minimizando la intervención de personal técnico, con aplicaciones basadas en Sistemas Expertos, Redes Neuronales o sistemas con Agentes Inteligentes, obteniendo resultados cada vez más satisfactorios.

Varios autores han aplicado técnicas de Inteligencia Artificial en estudios recientes para conseguir alcanzar la optimización de una EDAR.

2.2.1. Sistemas de Control mediante Redes Neuronales

Las Redes Neuronales Artificiales (RNAs) son un campo muy importante dentro de la Inteligencia Artificial. Inspirándose en el comportamiento conocido del cerebro humano, principalmente el referido a las neuronas y sus conexiones, trata de crear modelos artificiales que solucionen problemas de difícil solución mediante técnicas algorítmicas convencionales.

Una de las ventajas de modelado mediante redes neuronales es su capacidad de aprender relaciones a través de los datos por si mismo en lugar de asumir la forma funcional de las relaciones como se hace en los métodos cognitivistas. A través de un proceso de aprendizaje iterativo, la red neuronal extrae la información del conjunto de entrenamiento y la almacena. La red neuronal puede modelar cualquier relación con cualquier grado de precisión siempre que haya datos suficientes para el modelado, pero también puede realizar representaciones con datos imprecisos o incompletos.

A diferencia de los modelos cognitivos convencionales, una red neuronal puede adaptarse dinámicamente a las nuevas condiciones del proceso a través del aprendizaje

continuo. Después de realizar el entrenamiento de la red puede ser utilizada para proporcionar predicciones para nuevas entradas.

Qiao et al [57] proponen una estrategia de control para conseguir la optimización del problema de la depuración de aguas residuales mediante Control Adaptativo Óptimo con Data-Driven (DDAOC) basado en Programación Adaptativa Dinámica (ADP) para el control de Oxígeno Disuelto (DO) y nitratos. El DDAOC consiste en dos redes neuronales, una RNA de evaluación y otra de optimización que utilizan un método de aprendizaje iterativo para aproximar la solución óptima del problema. En este trabajo se optimizan únicamente los nitratos y nitritos (S_{NO}) y el oxígeno disuelto (S_O). El DDAOC basa sus raíces teóricas en el principio de optimización de Bellman y en la teoría basada en datos. En el enfoque basado en datos la información sobre la dinámica de sistemas se consigue de un conjunto de datos suficientemente grandes para poder aprender la política óptima a través de los datos de entrada y de salida. Debido a que no es necesario el uso de un modelo mecanicista DDAOC es fácil de implementar. Los resultados obtenidos con esta simulación DDAOC comparados con el control tradicional mediante PID representan una disminución en el consumo de energía total entorno al 8.5 % y un aumento de S_{NO} en un 8.7 % junto con una disminución del S_O en un 21.0 %. Los resultados muestran que esta estrategia de optimización basada en datos puede ser utilizado eficazmente para el control del oxígeno disuelto y la optimización del nivel de nitrato.

Un Método de Control y Modelado Online con Redes Neuronales (NNOMC) es propuesto por Quiao et al [58] para el control multivariable de procesos de tratamiento de aguas residuales. Utiliza una red neuronal feedforward o prealimentada, para describir un tipo de sistema que reacciona a los cambios en su entorno, normalmente para mantener algún estado concreto del sistema. Un sistema que exhibe este tipo de comportamiento responde a las alteraciones de manera predefinida. Para conseguir la estabilidad del método NNOMC es necesario limitar la tasa de aprendizaje de la red neuronal. Los resultados obtenidos mediante la simulación del benchmark muestran que el método NNOMC consigue tanto una buena aproximación como el control del rendimiento.

Además de estudios sobre optimización de una EDAR se han presentado algunas patentes sobre el tema, como la propuesta por Jun-Fei Quiao y Fan-Jun Li [1], con un método para el tratamiento de aguas residuales que realiza la predicción de fósforo total del efluente basado en una Red Neuronal en Cascada Auto-organizada. La red neuronal artificial conocida como Cascada Correlación (CC) propuesta por Fahlman

y Lebiere [1990], está diseñada siguiendo el esquema de crecimiento de red. Se comienza con una red mínima sin capas ocultas, es decir, con sólo algunas entradas y uno o más nodos de salida. Las neuronas ocultas son agregadas una a una en la red, obteniendo de esta manera una estructura multicapa, que permite aplicar las técnicas de regularización utilizadas para perceptrones multicapa. Los modelos en cascada no contemplan interconexiones entre neuronas de la misma capa, ni tampoco interconexiones de retroalimentación. La red auto-organizada presenta un aprendizaje no supervisado, es decir que no precisa de conjunto de entrada previo. Con el fin de resolver el problema de la medición en línea en tiempo real del fósforo total del efluente, el método diseña la estructura de la red en cascada con crecimiento auto-organizado, con la tasa de contribución de las fórmulas variables, así como métodos diferenciados de entrenamiento de pesos, a través de ajustes en tiempo real relacionados con la configuración de la red, mejorando el proceso de tratamiento de las aguas residuales urbanas y la precisión de la predicción de fósforo total del efluente antes del proceso de vertido al alcantarillado de la ciudad, mediante el ajuste de la estructura y el peso de las conexiones de la red CC. El resultado del experimento muestra que el método de predicción inteligente puede medir de forma adecuada y precisa la concentración total de fósforo en el efluente en el proceso de tratamiento de las aguas residuales urbanas. La patente propone un ajuste de enlaces correlacionados de control en el proceso de tratamiento de aguas residuales y en las reacciones bioquímicas, mejorando así la calidad del efluente de la depuradora de aguas residuales y la prestación de apoyo teórico y técnico para la mejora del funcionamiento seguro y estable del proceso de tratamiento.

2.2.2. Modelos de Aprendizaje Automático utilizando Máquinas de Vectores Soporte

El objetivo del estudio realizado por Guo et al [28] es establecer dos modelos de aprendizaje automático, RNA y Máquinas de Vectores Soporte (SVM), con el fin de predecir la concentración del Nitrógeno Total (TN) durante el intervalo de un día, del efluente de una planta de tratamiento de aguas residuales en Ulsan, Corea. En general, la calidad del agua de una depuradora es sensible a parámetros tales como el pH, la temperatura (T), las concentraciones de sustratos y los contaminantes. Esto se debe a que las aguas residuales son tratadas mediante procesos metabólicos de los microorganismos. Sin embargo, el tratamiento biológico todavía exhibe características variables en el tiempo y altamente no lineales, afectado por diversos parámetros, tanto conocidos como desconocidos. Debido a estas características el modelo de aprendizaje

automático ha demostrado ser una herramienta útil porque tiene una precisión relativamente alta para hacer frente a sistemas complejos. Por otra parte, una ventaja clave de estos modelos para la evaluación del desempeño de una EDAR es que se pueden predecir directamente los valores de salida partiendo de los valores de entrada tras realizar entrenamiento y validación. Las RNAs se diseñan mediante un perceptrón multicapa. El perceptrón multicapa es una red de alimentación hacia adelante (feedforward) compuesta por una capa de unidades de entrada (sensores), otra capa de unidades de salida y un número determinado de capas intermedias de unidades de proceso, también llamadas capas ocultas porque no se ven las salidas de dichas neuronas y no tienen conexiones con el exterior. Cada sensor de entrada está conectado con las unidades de la segunda capa, y cada unidad de proceso de la segunda capa está conectada con las unidades de la primera capa y con las unidades de la tercera capa, así sucesivamente. Las unidades de salida están conectadas solamente con las unidades de la última capa oculta. Con esta red se pretender establecer una correspondencia entre un conjunto de entrada y un conjunto de salidas deseadas. Utiliza como algoritmo de retroprogramción el aprendizaje con momentos y como función de transferencia una función sigmoidea tangente. Los conjuntos de parámetros óptimos de los nodos ocultos, tasa de aprendizaje, y el momento para el modelo se determinaron mediante algoritmos de búsqueda de patrones. Las SVM son estructuras de aprendizaje basadas en la teoría estadística del aprendizaje. Se basan en transformar el espacio de entrada en otro de dimensión superior en el que el problema puede ser resuelto mediante un hiperplano óptimo (de máximo margen). La formulación de las máquinas de vectores se basa en el principio de minimización estructural del riesgo que ha demostrado ser superior al principio de minimización del riesgo empírico. Las SVM presentan un buen rendimiento al generalizar en problemas de clasificación pese a no incorporar conocimiento específico sobre el dominio. La solución no depende de la estructura del planteamiento del problema. Como Kernel se utiliza la función de base radial que da lugar a un modelo óptimo para la predicción de la calidad del agua efluente y los parámetros clave del modelo se determinaron por el algoritmo de optimización.

Estudiando las concentraciones de TN medidas y predichas por los modelos de RNA y SVM con la aplicación de los parámetros óptimos, se observa que la exactitud de la predicción de la SVM fue ligeramente mayor que la precisión de la RNA durante las etapas de formación, mientras que durante las etapas de validación ambos modelos presentan una precisión muy similar. En consecuencia, se observó un mayor rendimiento de predicción de SVM frente a RNA. Estos dos modelos también se

podría aplicar para estimar la concentración TN del efluente cuando la vigilancia en tiempo real o el muestreo no sean posibles, y para estimar el rango de los parámetros de salida que cumpla con las normas de calidad del agua de vertido.

Xu Yuge et al [2] presentan una patente con un método de medición online para el tratamiento de aguas residuales basado en una Máquina de Vectores de Relevancia rápida (RVM). Las RVM utilizan el tratamiento bayesiano de una función de decisión similar a la de una SVM, donde previamente se introducen los pesos mediante un conjunto de hiperparámetros. En comparación con una SVM, los pesos cero en la RVM representan ejemplos prototípicos de clases, que se denominan vectores de relevancia. La RVM utiliza muchas menos funciones de base que la SVM correspondiente y su rendimiento es por lo general superior. El método propuesto consta de las siguientes etapas: estimación de los hiperparámetros mediante un algoritmo de probabilidad marginal rápido, para obtener un valor de peso y un valor de desviación de la muestra de un modelo, a continuación, se establece un modelo de predicación en línea de la RVM, se optimizan los parámetros del modelo y se consigue una medición exacta y rápida de la DBO de las aguas residuales. Se establece un modelo de predicción óptima, por lo que se mejora la precisión de la predicción, el efecto es evidente y las actuaciones se mejoran. Este método consigue un alto grado de precisión de la predicción, tiene gran capacidad de generalización y minimiza el tiempo de actualización, por lo que se consigue una importante reducción de los gastos de operación de la EDAR, lo que redunda en la calidad de los vertidos gracias a la realización de un sistema de control automático para el tratamiento de las aguas residuales.

2.2.3. Control Online con Métodos de Regresión y de Clasificación

La medición en línea de la DQO y del Nitrógeno Amoniacal (NH4-N) en el flujo de entrada de una EDAR es crucial para el seguimiento de los procesos de tratamiento biológico de las aguas residuales y para el desarrollo de estrategias de control avanzadas para la optimización del control de la planta. Como conseguir una medición directa de DQO y NH4-N es difícil y caro Kern et al [38] proponen un trabajo en el que se realiza una estimación de la DQO y NH4-N basada en medición espectroscópica en línea del sistema de flujo mediante técnicas de aprendizaje automático.

El estudio se realiza utilizando tanto Métodos de Regresión como:

- Método de Mínimos Cuadrados Parciales (PLS)
- Perceptrón Multicapa (MLP)

- Vectores Soporte para Regresión (SVR)
- Mínimos Cuadrados Parciales (PLS)

Y métodos de Clasificación tales como:

- Análisis de Discriminantes Lineales (LDA)
- Random Forest (RF)
- Máquinas de Vectores Soporte (SVM)

Tras comparar los resultados obtenidos con las técnicas de Regresión y de Clasificación, los resultados muestran que la estimación de DQO se puede lograr usando unicamente mediciones estándar en línea, utilizando RF con Regresión con un Error Cuadrático Medio Normalizado (ECMN) de 0,3, que es suficientemente preciso para las aplicaciones prácticas. En el caso de NH4-N, se consigue una buena estimación utilizando Regresión con PLS con un ECMN de 0,16 solo si se utiliza una combinación de medidas estándar y espectroscópicas en línea. Además, la comparación de métodos de regresión y clasificación muestra que ambos métodos funcionan igualmente bien en la mayoría de los casos. Este análisis evidencia que es posible conseguir suficientes resultados de la estimación mediante la clasificación en lugar de utilizando los métodos de regresión. Esto no sólo facilita la presentación transparente de las concentraciones de flujo de entrada para el operador, sino también permite el entrenamiento adaptado específicamente para centrarse en la correcta estimación de parámetros concretos. Para su aplicación práctica se recomienda estimar DQO utilizando un equipo estándar o con la instalación adicional en línea de una sonda de turbidez, mientras que la estimación de NH4-N es una opción válida para las plantas que ya tienen una sonda de DQO en línea en la entrada de la EDAR.

2.2.4. Sistemas de Control mediante Lógica Difusa

La lógica difusa es una alternativa de control valiosa para procesos con un desarrollo complejo. Un Sistema de Control basado en Lógica Difusa (SCLD) utiliza una descripción del proceso mediante reglas desarrolladas a partir del conocimiento del proceso y proporciona un algoritmo que pueda convertir un conjunto de reglas basadas en la experiencia de un operador humano en una estrategia de control automático. Las principales ventajas del SCLD son: su facilidad de implementación, la velocidad con la que se obtiene una salida fiable, lo que permite solucionar problemas de control automático de manera sencilla y sin necesidad de conocer el modelo matemático para

poderlo controlar y el ser una forma rápida y económica de solventar un problema. Por otro lado presenta algunos inconvenientes como la necesidad de un tiempo de aprendizaje para obtener buenos resultado o que si el problema tiene solución matemática esta suele ser mejor que la obtenida mediante lógica difusa.

Los SCLDs combinan las reglas de un sistema experto con una especificación flexible de sus parámetros de salida. La estructura de un SCLD está generalmente basada en la monitorización de valores de referencia del sistema. Las medidas pueden ser realizadas mediante sensores o basadas en observaciones manuales. Un SCLD está compuesto por dos niveles: el numérico y el lingüístico. Mientras que el proceso tiene que trabajar con valores discretos, el SCLD trabaja con variables lingüísticas. Para realizar su transformación es necesario que los valores discretos de las medidas del proceso sean caracterizados en términos cualitativos (por ejemplo, bajo, normal, alto, etc.) y asociados a ellos con un determinado grado de pertenencia, de manera que se obtendrán los valores difusos o fuzzy. A continuación, con la interacción de los expertos que conocen el proceso a controlar, se definen un conjunto de reglas heurísticas que definirán el tipo de actuación en función de las condiciones de entrada. Las reglas son sentencias del tipo 'si-entonces' que determinan la acción de control del sistema difuso en función de las variables de entrada. El conjunto de reglas conforman la matriz de decisión o base de reglas. Finalmente, la salida difusa tiene que ser traducida de nuevo a un valor discreto (defuzzificación) para que la acción de control pueda ejecutarse en el proceso [23].

El sistema de control basado en la lógica difusa aplicado a una EDAR utiliza complejos algoritmos de control para optimizar el consumo energético y mejorar la estabilidad del proceso biológico mediante el control de la aireación de los reactores biológicos. Este sistema de control establece lazos de control independientes para el oxígeno disuelto en cada tanque y para la presión de descarga de las soplantes. La concentración de oxígeno disuelto en cada tanque se mantiene en el valor de consiga modificando el grado de apertura de la válvula de control mientras que la presión de descarga en la conducción de aire se controlará modificando en número de soplantes en funcionamiento y las velocidades de giro de la soplante que disponga de variador. Mediante esta propuesta se airearán todos los tanques mediante el conjunto de soplantes.

Se han realizado estudios del control del rendimiento de una EDAR utilizando lógica difusa [15] mediante un módulo de razonamiento experto que usa reglas obtenidas de conocimientos teóricos para construir el módulo de razonamiento. La incertidumbre y la subjetividad asociada a la descripción de conocimientos heurísticos se enfocan

mediante el uso de lógica difusa para clasificarla como baja, normal, media o alta para obtener el grado de pertenencia o posibilidad de que un hecho suceda. El uso de este módulo experto consigue en las salidas de la simulación unos buenos valores en el efluente.

En la EDAR de Molina de Segura en la Región de Murcia se ha utilizado un sistema de control por lógica difusa para controlar la apertura de la válvula que regula el caudal de aire que pasa por los difusores en función de la concentración de oxígeno disuelto. Este sistema permite establecer un control de oxígeno independiente para cada zona de aireación lo que proporciona una mayor estabilidad al proceso biológico y de esta forma consigue unos resultados muy buenos de eliminación de nutrientes, especialmente nitrógeno [6].

2.2.5. Sistemas basados en Algoritmos Genéticos

Con el fin de reducir el consumo energético producido por la aireación de los tanques en los procesos de depuración de aguas (más del 50 % del consumo energético total), Ozturk et al [50] presentan un trabajo en el que estudian el problema utilizando Programación No Lineal Entera Mixta (MINLP) que resuelven mediante Algoritmos Genéticos (GA). En este trabajo, nos muestran que es posible reducir la tasa total de aireación de un tanque de lodos activados a través de la utilización de una distribución no uniforme de los caudales de aire locales, utilizando un modelo de tanque de lodos activados referenciado a la EDAR de Stickney en Chicago (EEUU). Para realizar la optimización se plantea un problema MINLP de seis variables que resuelven mediante GA. Gracias a este pequeño número de variables de optimización y la capacidad de GA para escapar de los mínimos locales, la solución se consigue acercar bastante al mínimo global del problema. El algoritmo genético crea un cierto número de soluciones candidatas, llamada población inicial, y las evalúa para ver su aptitud frente a las condiciones definidas. Las soluciones con mejor condición física (élites) se pasan a la siguiente iteración (generación), además de nuevas soluciones (hijos) que se crean a través de soluciones de ajuste (padres). Entonces, uno de estos padres pueden ser modificados (mutación) o ambos padres se puede utilizar para generar una solución (cruzado), que se transmite a la siguiente generación. A través de la iteración de las generaciones, el GA finalmente converge en un mínimo local del problema. El perfil de aireación obtenido en la solución del problema de optimización MINLP insta a los aireadores en el primer y último tanque aeróbico a tasas relativamente más alta aireación, al tiempo que permite a los internos utilizar valores mucho más bajos. Dado que los perfiles de aireación óptimos trabajan con la mínima cantidad de aireación

necesaria para llevar a cabo el proceso de tratamiento y satisfacer las limitaciones, un aumento del 30 % en estos perfiles óptimos (como un factor de seguridad) puede mejorar en gran medida la respuesta dinámica del proceso para adaptarse a los eventos transitorios.

En la investigación realizada por Bagheri et al [7], se han desarrollado los modelos híbridos de Redes Neuronales Artificiales con Perceptrón Multicapa (MLPANN) y Redes Neuronales Artificiales en Función de Base Radial (RBFANN) aplicando Algoritmos Genéticos, para predecir con precisión el Índice Volumétrico de Lodos (SVI) para la EDAR de Ekbatan en Teherán (Irán), utilizando datos reales de operación entre los años 2011 y 2013 con diversas condiciones de funcionamiento. Los parámetros de funcionamiento, incluyendo los Sólidos Suspendidos Volátiles en Licor de Mezcla (MLVSS), el pH, el DO, la T y los Sólidos en Suspensión (SS), DQO y TN del influente, se utiliza en los procesos de modelado de la Red Neuronal. El GA se utilizó para optimizar las neuronas, los pesos, las funciones, y los umbrales de las MLPANN y RFBANN.

La estructura de una RBFANN básica consiste en una capa de entrada, una capa de salida, y una capa oculta con N nodos. La función de base radial simétrica se utiliza como función de activación de los nodos ocultos. La transformación de los nodos de entrada a los nodos ocultos se realiza de forma no lineal, y el entrenamiento de esta parte de la red se realiza sin supervisión. El entrenamiento de los parámetros de la red (peso) entre las capas ocultas y de salida se produce de una manera supervisada basándose en los objetivos de salida. La RBFANN tiene como ventaja el aprendizaje rápido porque tiene una etapa de aprendizaje sin ninguna iteración de actualización de pesos, una capacidad robusta y buena capacidad de adaptación. Sin embargo, estas redes requiere grandes cantidades de datos para realizar el entrenamiento, y necesitan realizar un buen ajuste de los parámetros de la función de activación de los nodos ocultos. Los parámetros se determinan por la experiencia o mediante el uso de métodos óptimos, como el GA, para sintonizar los parámetros de red y los pesos. La RBFANN aplica diferentes funciones de red, como la función newrbe, a los datos de entrada. La función newrbe crea una red de dos capas con sesgos para ambas capas. La función newrbe crea una red de base radial añadiendo una neurona en cada iteración. La red añade una nueva neurona hasta que el error total de la red alcanza un error objetivo o bien se alcanza un número máximo de neuronas. Los pesos de la capa de base radial se inicializan con los valores de los patrones, los pesos de la capa lineal se obtienen simulando las salidas de la primera capa y resolviendo una expresión lineal. Una de las estructuras RNA más utilizadas en los problemas de

clasificación es MLPANN con un algoritmo de aprendizaje de propagación hacia atrás (BP). Esto es útil cuando el sistema de diagnóstico debe ejecutarse en tiempo real y manejar una gran cantidad de señales. Por otra parte, esta red neural es robusta, especialmente con respecto al ruido. La MLPANN está formada por neuronas simples llamadas perceptrones. La estructura de la MLPANN básica consiste en una capa de entrada, una capa de salida, y una capa oculta con N nodos. El perceptrón calcula una sola salida con las entradas múltiples haciendo una combinación lineal de acuerdo con sus pesos de entrada y a continuación determina de la salida a través de una función de transferencia no lineal. La MLPANN aplica diferentes funciones de red, tales como la función newff a los datos de entrada. La función newff crea una red de retropropagación de alimentación hacia adelante. La primera capa de la red tiene pesos procedentes de las entradas y cada capa posterior tiene un peso procedente de la capa anterior. La función de transferencia de esta red puede ser cualquier función diferenciable de transferencia. La MLPANN es entrenada con diferentes algoritmos de aprendizaje (como la retropropagación de errores, la retropropagación incrementales, el algoritmo de Levenberg-Marquardt, retropropagación basada en el descenso de gradiente y tasas de aprendizaje adaptativas con retropropagación). Las funciones de transferencia de las capas oculta y de salida se determinan de forma iterativa mediante el desarrollo de varias redes.

Para obtener la mejor solución entre todas las posibles se realiza una optimización mediante GA utilizando búsqueda aleatoria (algoritmos evolutivos) para la optimización de una función de aptitud por medio de la codificación de los parámetros. El GA obtiene buenos resultados utilizando operadores como la creación de la población, la selección, cruce y mutación. La precisión de la predicción de las redes en los modelos de RNA depende del número de neuronas ocultas, las funciones y la tasa de aprendizaje, por lo que estas variables fueron las elegidas para optimizar la estructura RNA por el programa de GA. Basándose en el resultado de este estudio, los modelos óptimos se obtuvieron con una capa oculta formada por 9 neuronas. Las RNAs seleccionadas fueron utilizados para predecir el SVI para diferentes entradas en los dominios de entrenamiento y las pruebas. Basándose en los resultados de este estudio el modelo MLPANN-GA obtiene mejores valores que el modelo RBFANN-GA para simular el rendimiento del SVI en la EDAR de Ekbata, comparando los resultados de verificación y validación, es decir, mejores capacidades de predicción y generalización. La predicción del SVI fue un éxito tanto para el modelo MLPANN-GA como para el RBFANN-GA ya que los modelos indican una coincidencia casi perfecta entre el grupo experimental y los valores predichos de SVI. También se observa que la precisión

y la exactitud de todos los modelos aumentaron cuando GA se aplicó a los modelos de RNA. El promedio de error medio en la predicción del SVI para los modelos de entrenamiento y pruebas no supera el 3 % de los valores de entrada de SVI.

En el trabajo realizado por Huang et al [36], se propone un sensor de software utilizando un Sistema Difuso Neuronal basada en Algoritmo Genético (GA-NFS) para la estimación en tiempo real de las concentraciones de nutrientes en un proceso de tratamiento biológico de aguas residuales. Para conseguir mejorar el rendimiento de la red se empleó el algoritmo de c-medias con agrupamiento difuso autoadaptado y se utilizó un algoritmo genético para extraer y optimizar la estructura de la red. El GA-NFS se aplicó a un proceso de tratamiento biológico de aguas residuales para la eliminación de nutrientes. La arquitectura del sistema GA-NFS está formada por cinco componentes principales: un sistema de inferencia borrosa de entrada y de salida del preprocesador, un generador de sistema difuso, una red neuronal adaptativa que representa el sistema difuso y un optimizador del algoritmo genético. El sistema de inferencia borrosa es de tipo Sugeno (basado en reglas) y presenta grandes ventajas como ser computacionalmente eficiente, trabajar bien con técnicas lineales (por ejemplo como lo disponible para controladores PID), trabajar bien con técnicas de optimización y control adaptable, tener garantizada una superficie de control continua y estar bien adaptado al análisis matemático. Su sistema adaptativo es de tipo Sistemas de Inferencias Borrosas Basados en Redes Adaptables (ANFIS). Los parámetros de entrada y de salida se seleccionan o se generan a partir de los parámetros que se utilizan comúnmente para la descripción del sistema. El modelo GA-NFS tiene cinco capas. Los nodos en la capa 1 son nodos de entrada que representan variables lingüísticas de entrada. La capa 5 es la capa de salida. Los nodos en la capa 2 son nodos término que actúan como funciones de pertenencia para representar los términos de cada variable lingüística. Cada nodo en la capa 3 es un nodo regla que representa una regla lógica difusa, por lo tanto todos los nodos en la capa 3 forman una base de reglas difusas. En los nodos en la capa 4, se utiliza la parte consecuente del modelo Takagi-Sugeno-Kang (TSK). Los vínculos a nivel de capa 3 definen las condiciones previas de los nodos de reglas. Los vínculos a nivel de capa 2 representan una conexión completa entre los nodos lingüísticos y sus correspondientes nodos término. El algoritmo BP se utiliza para entrenar el sistema GA-NFS en la actualización de los parámetros de las funciones de pertenencia y los pesos de los enlaces entre las capas 4 y 5 sobre la base de la primera etapa de formación. Durante la etapa de formación de BP, la estructura de la GA-NFS es estática. Cuando el sistema difuso neuronal se somete al entrenamiento, la información de error de realimentación entre la salida real de los nodos de

la capa 5 y los correspondientes valores objetivo se propaga hacia atrás a través del sistema. La propagación hacia atrás de las señales de error se utiliza para actualizar los pesos de las conexiones entre las capas 4 y 5 . En la actualización de parámetros, el objetivo es reducir al mínimo la función de error. Después de entrenar el modelo la inferencia se realizó de acuerdo con nueve reglas lingüísticas difusas, también se incluyeron otras reglas heurísticamente para realizar la comparación entre los valores de entrada y salida. Los parámetros de entrada-salida seleccionados incluyen como entradas del sistema: pH, Potencial de Oxidación Reducción (ORP) y DO y como salidas del sistema: DQO, NO_3^- y PO_4^{3-}. De acuerdo con los resultados obtenidos en la simulación, el sistema GA-NFS podría predecir la dinámica de nutrientes de la operación anóxica/óxica. La estimación en tiempo real de las concentraciones de DQO, NO_3^- y PO_4^{3-} basada en GA-NFS funcionó de manera efectiva utilizando la información online proporcionada por el sistema anóxico/óxico. Para comparar los resultados entre GA-NFS y NFS se utilizan como criterios de evaluación RMSE, MAPE y R. Cuanto más pequeño sean RMSE y MAPE y más grande R mejor será el rendimiento.

Teniendo en cuenta el alto nivel de complejidad en el proceso anóxico/óxico, la gran cantidad de información existente en el conjunto de datos y los amplios intervalos de concentración, la predicción de los modelos GA-NFS para los parámetros fueron bastante acertados. El sensor de software, basado en el modelo GA-NFS se puede aplicar de manera efectiva a los procesos anóxicos/óxicos con el fin de hacer frente a las variaciones del afluente, que son típicos de las aguas residuales municipales además la duración de la fase de la operación se puede optimizar y ajustar en tiempo real. El sensor de software puede lograr eficazmente los objetivos ambientales y económicos del sistema en un tiempo real, cumpliendo en todo momento con la normativa ambiental.

2.2.6. Control usando Sistemas Multiagente

Bongards et al [8] proponen el control de una EDAR y de una red de alcantarillado mediante Sistemas Multiagente, utilizan agentes para representar cada uno de los tanques y la planta depuradora, para realizar con posterioridad el control del sistema de alcantarillado al que vierten.

El sistema sigue un enfoque económico, donde los tanques de agua de tormenta actúan como competidores en un mercado de agua virtual. La ventaja de este enfoque es que presenta un menor esfuerzo de configuración y una mayor capacidad de expansión del sistema en comparación con las estrategias de control convencionales, además de obtener mejores resultados. Esto se logra mediante la autoorganización

y la cooperación de los agentes del sistema. Las estrategias de control para la minimización de la contaminación del agua de los ríos o la optimización de la planta de tratamiento de aguas residuales se pueden ajustar fácilmente mediante la regulación de uno o dos parámetros de los agentes.

Este estudio realiza una analogía entre las funciones de los componentes principales de un sistema de depuración con las del mercado. En ambos sistemas, hay un producto, un proveedor de almacenamiento, un comprador del producto, y un sistema de transporte común. Estas similitudes conducen a la hipótesis de que también serán similares las reglas por las que se rige el sistema. Esto dio como resultado el enfoque para describir el comportamiento de un mercado con un sistema multi-agente. Los participantes individuales en el mercado están representados por agentes de software con diferentes funciones. El negociador es la instancia entre el comprador y el vendedor. Cada tanque ofrece su producto a los intermediarios. El intermediario lleva la oferta del comprador (la EDAR) y trata de obtener la cantidad necesaria de los vendedores lo más barato posible. Para ello, recoge todas las ofertas en cada ronda y determina la mejor oferta. El vendedor con el precio más bajo puede vender su cantidad. Las rondas comerciales se terminan tan pronto como la cantidad total ha sido comprado o no hay más productos ofertados.

2.2.7. Predicciones basadas en Árboles de Decisión

Para realizar la estimación de los parámetros de salida, DBO, DQO y SS, de una planta depuradora, Celik et al [13] presentan en la International Conference on Environmental Science and Technology – 2013 (ICOEST'2013 - Cappadocia) un estudio con modelos predictivos mediante árboles de decisión basado en el índice de Gini, para la estimación de los parámetros del efluente y el pH.

Se realiza el Análisis de las Componentes Principales utilizando el método de correlación lineal para su posterior aplicación en los árboles de decisión. El índice de Gini es un método basado en la división binaria del conjunto de datos. Los valores de atributo se dividen en dos grupos que van generando ramas. Los valores de atributo de cada elemento binario del grupo se coloca en ramas separadas, estas ramas se utilizan para crear los grupos con los valores de atributos. De acuerdo con los resultados de las pruebas el rendimiento del modelo desarrollado se encuentra a un nivel aceptable aunque este sistema no es perfecto para el tratamiento de aguas de una EDAR. Este método se puede mejorar dividiendo los valores de salida en nuevos grupos aunque seguirán existiendo algunos problemas de precisión. El problema puede ser resuelto

mediante la adición de otros algoritmos híbridos para que la precisión de los resultados alcance un punto óptimo.

2.2.8. Control Predictivo

El Control Predictivo (MPC) es un conjunto de métodos de control que hacen uso explícito de un modelo del proceso para obtener la señal de control minimizando una función objetivo. Estos métodos utilizan controladores que tienen básicamente la misma estructura y los siguientes elementos principales:

- Uso explícito de un modelo para predecir la evolución del proceso en los instantes futuros.
- Minimización de una función objetivo.
- Utilización de un horizonte de control finito y deslizante que implica el cálculo de la secuencia de control para todo el horizonte pero con la aplicación de la primera señal de la secuencia y la repetición de todo el proceso en el siguiente instante.

Con el fin de obtener una estrategia de control multivariable adecuado para el tratamiento de las aguas residuales Han et al [29] proponen un trabajo utilizando un Modelo Predictivo con Control Multiobjetivo no Lineal (NMMPC). El NMMPC propuesto utiliza una Red Neuronal con función de autoorganización de base radial (SORBF) y un controlador de múltiples objetivos con el método de optimización multi gradiente. Este NMMPC basa su simplicidad en el diseño y la eficiencia del Modelo de Control Predictivo para hacer frente a la complejidad computacional. Uno de los factores clave de la estrategia NMMPC es encontrar un modelo de control apropiado para la planta utilizando un método basado en datos. Debido a su fácil diseño, la capacidad de generalización, la fuerte tolerancia al ruido de entrada y la capacidad de aprendizaje en línea, se utilizaron las redes neuronales en Función de Base Radial (RBF) en MPC para el modelado no lineal del sistema . Esta red neuronal RBF propuesta no fija el número de nodos ocultos pudiéndose modificar tanto la estructura de la red (el número de nodos ocultos) como los parámetros (los pesos). La red SORBF con estructura concurrente y aprendizaje de parámetros se utiliza para conseguir un modelo en línea bastante exacto y compacto basado en las características de la EDAR, por lo que la red neural SORBF puede ser utilizada para reemplazar algunos instrumentos existentes cuando se obtengan mediciones no fiables. En el diseño NMMPC, el índice de rendimiento se utiliza para ajustar las propiedades del sistema de lazo cerrado,

mediante el Método Multi Gradiente (MGM) para manejar múltiples objetivos. La idea clave es reducir al mínimo las funciones de costos mediante el seguimiento de la dirección multi gradiente. Los resultados obtenidos en esta simulación, según los criterios BSM1, muestran que la estrategia NMMPC propuesta consigue buenos resultados en el control de los valores del DO y los nitratos, incluso en una planta con gran número de perturbaciones, aunque no se han realizado aplicaciones del NMMPC a una EDAR real para poder comprobar si estos resultados teóricos se ajustan a la realidad.

La aplicación del Control Predictivo Multivariable a gran escala para el proceso de lodos activados en una planta de tratamiento de aguas residuales municipales de Vikinmäki en Helsinki (Finlandia) se discute en los trabajos realizados por Mulas et al [45].

El modelo MPC se ha convertido en un enfoque atractivo para un número considerable de aplicaciones en la depuración de aguas residuales, principalmente debido a la capacidad de los algoritmos MPC para hacer frente a los problemas de control multivariable de una manera óptima, a través de la utilización sistemática de modelos simples y generalmente lineales. La estrategia de Control mediante Matriz Dinámica (DMC) utiliza un modelo de proceso de respuesta finita lineal y un modelo de perturbación aditiva de salida constante. La elección de una estrategia de DMC es preferible sobre una configuración relativamente simple de controladores de realimentación desacoplados porque se considera menos sensible a la elección óptima de los puntos de ajuste del controlador que además es mejor para una gama más amplia de condiciones de funcionamiento. La idea básica del algoritmo MPC es calcular en cada control de paso una secuencia que minimice una cierta función objetivo. La secuencia de control se calcula en base a un modelo simplificado del proceso y a las mediciones de salida. Los modelos se obtienen mediante el análisis de la respuesta del amonio y los nitratos cuando se aplica una función de paso de diferentes amplitudes a las variables manipuladas (el DO y el caudal de recirculación interno). En este trabajo se hace hincapié en la selección de una configuración de control que contribuya a reducir al mínimo los costes económicos al tiempo que mejora el rendimiento de eliminación de los compuestos de nitrógeno. Para ello se utiliza un algoritmo de control por matriz dinámica que se ve favorecido por el control de las concentraciones de nitrógeno en el extremo del proceso biológico.

Para demostrar la efectividad de la propuesta se consideran diferentes configuraciones de control y se comparan con las estrategias de control de aireación reales

utilizadas en la planta de Vikinmäki. Para la comparación, se implementan una serie de bucles básicos de control de retroalimentación, siete en total. La configuración seleccionada implica un controlador DMC con tres variables manipuladas (el DO, los puntos de ajuste en las primeras zonas del reactor biológico y el caudal de recirculación interno) y dos variables controladas (el nitrógeno amoniacal y el nitrato, medidos en el extremo de la zona de desgasificación), junto con un control de lodos de bajo nivel, que manipula el exceso de caudal de lodos para el control de los SS en la última zona del biorreactor. Basándose en los resultados de la simulación, este trabajo muestra la potencialidad del control mediante DMC, que es capaz de disminuir los costes de consumo de energía y, al mismo tiempo, reducir los picos de amonio y la concentración de nitrato en el efluente.

2.2.9. Control mediante Sistemas Expertos

En el control de las plantas de tratamiento de aguas residuales existe un importante problema debido a las fluctuaciones tanto del caudal de entrada como de la composición de este caudal, para solventar este problema es necesario el control constante por parte de un operador para realizar el ajuste de la planta. El uso de Sistemas Expertos (SE) es habitual para ayudar al operario en la toma de decisiones [51].

Se han realizado estudios de control de EDARs utilizando el enfoque emergente de Aprendizaje Automático mediante técnicas de Aprendizaje por Refuerzo [34][33]. Utilizando como índice de referencia las acciones COST [21] se realizan simulaciones con el banco de pruebas BSM1, que consiste en un sistema con un método de control asociado, un procedimiento de evaluación comparativa y un criterio de evaluación. BSM1 usa el modelo de lodos activados ASM1 para el modelado de las reacciones biológicas [3].

El agente propuesto para realizar el control inteligente tiene dos entradas: las medidas de NH_4 y O_2 obtenidas a partir de dos sensores colocados en el último tanque aeróbico y una única salida, el valor de DO . El agente actúa sobre la planta cambiando este punto de ajuste DO. El objetivo del agente es reducir los costos de energía de la planta, manteniendo las condiciones de salida del efluente de manera óptima. Se realizaron pruebas con dos escenarios distintos demostrándose que la estrategia de control realizado por el agente fue distinta para cada planta sin la necesidad de la intervención de un operador de planta o ingeniero durante el proceso. Este agente también puede supervisar y modificar los valores de referencia de la planta en caso de cambios ambientales. De este modo, se obtiene un agente autónomo libre para

optimizar los procesos de la planta de forma permanente sin necesidad de intervención humana.

Qin J. y Guo H. [59] proponen la gestión de la depuración de aguas residuales mediante la utilización de un SE que gestione la toma de decisiones de lo datos obtenidos online.

Primeramente se realiza un estudio para optimizar la toma de datos debido a la gran complejidad que presentan este tipo de sistemas. Una vez determinado el número de instrumentos de seguimiento automático para controlar la calidad del agua y donde deben estar ubicadas las sondas con los sensores, en función de las necesidades de cambios y gestión de procesos, la información adquirida se puede visualizar y registrar en el tiempo real. Los datos adquiridos a través de la medición en línea y por los sistemas de adquisición han de satisfacer las necesidades de gestión necesarias para el proceso de tratamiento de aguas residuales, así como para diseñar el SE con una mayor fiabilidad. Los datos sobre el tratamiento de aguas residuales se obtienen de dos maneras, una es los datos adquiridos de forma automática a partir de los instrumentos de control en línea, que es la manera principal de obtención datos, la otra es mediante los datos introducidos puntualmente por el personal de gestión, por ejemplo, algunos datos introducidos cuando se produce una desviación o algún dato que es fácil de adquirir artificialmente pero no es fácil de adquirir mediante los instrumentos.

Posteriormente un sistema de gestión del conocimiento establecido con la base de datos relacional consigue que la base de datos del conocimiento tenga la capacidad de almacenar, mantener y extender el conocimiento. Las reglas se dividen en los elementos que la forman, los elementos de regla se pueden combinar juntos de nuevo a partir de una regla de producción mediante el uso de la relación entre la clave principal y la clave externa. Entre los parámetros que se monitorizan se encuentran NO_3^- , PO_4^{3-}, pH, DQO, Fósforo Total (TP), TN, ORP y NH_4^+

Se utiliza un proceso denominado BCFS (Biologisch-Chemische-Fosfaat-Stikstof Verwij Dering) para la eliminación del fósforo, este proceso está especialmente diseñada para optimizar la actividad de las bacterias denitrificantes en la eliminación de P. En este proceso se utilizan DO y ORP como los principales parámetros de control porque es difícil conseguir datos de los demás parámetros en tiempo real. La base de conocimiento basada en reglas se construye a través de la discusión con expertos en la materia. A medida que el sistema realiza el aprendizaje en función de los datos adquiridos, el razonamiento hacia adelante impulsado por datos va mejorando la base del conocimiento. Para diseñar el sistema experto se utiliza el lenguaje CLIPS.

2.3. Conclusión

Debido al importante impacto que producen las aguas residuales en el medio ambiente, su depuración se ha convertido en un tema prioritario a nivel internacional. Las legislaciones sobre el tema son cada vez mas restrictivas haciendo hincapié en áreas de especial interés ecológico, que en el caso español representan más de la cuarta parte del territorio. Además el control de las EDARs es fundamental por motivos de ahorro energético.

Para solventar estos problemas se han realizado en los últimos años grandes avances en los sistemas de control de las EDARs, utilizando para ello la modelización del proceso de fangos activos según la línea de trabajo del grupo de fangos activos ASM de la IWA.

El campo de la Inteligencia Artificial es uno de los que más aportes ha realizado, con aplicaciones basadas en Sistemas Expertos, Lógica Difusa, Redes Neuronales, Algoritmos Genéticos, Máquinas de Vectores Soporte o Sistemas Multiagente, obteniendo resultados cada vez más satisfactorios y observando que los modelos de Aprendizaje Automático pueden ser un método fiable para la predicción de la calidad del agua al realizar la alerta temprana del control de la calidad del tratamiento de aguas residuales.

Capítulo 3

Modelización de una Estación Depuradora de Aguas Residuales

3.1. Introducción

Las plantas de tratamiento de aguas residuales son grandes sistemas no lineales sometidos a importantes perturbaciones de flujo y de carga, además existen incertidumbres relativas a la composición de las aguas residuales entrantes. Sin embargo, estas plantas tienen que operarse de forma continua, cumpliendo normas estrictas y rigurosas.

Se han propuesto muchas estrategias de control pero su evaluación, ya sea en modo práctico o basado en la simulación, es difícil debido a la gran complejidad del proceso, pero también a la falta de criterios de evaluación estándar causados por las variaciones geográficas, ambientales y, por supuesto, políticas.

Una vez conocidas las características de las aguas a depurar y de las condiciones necesarias para realizar los vertidos, se deben escoger las configuraciones posibles de la planta para poder realizar posteriormente las simulaciones que permitan predecir su comportamiento sin los problemas técnicos y económicos que suponen la realización de una planta real.

En el campo de la modelización del proceso de fangos activos, se sigue la línea de trabajo del grupo de modelización de fangos activos de la IWA

Existen hoy cuatro generaciones de modelos de la IWA [32]: el ASM1 original y el mejorado ASM3, capaces de predecir la degradación de la materia orgánica mediante técnicas de nitrificación y denitrificación, y el ASM2 y su versión modificada ASM2d que incluyen además la eliminación biológica del fósforo.

El desarrollo de una referencia de evaluación se inició en Europa por los Grupos de Trabajo COST Action 682 y 624 [16], este trabajo fue posteriormente continuado

por un Task Group de la IWA.

El índice de referencia es un entorno de simulación que define una distribución de la planta, un modelo de simulación, cargas del afluente, procedimientos de ensayo y criterios de evaluación. Una vez que se ha validado el código de simulación se puede aplicar cualquier estrategia de control y el rendimiento puede ser evaluado de acuerdo a distintos criterios en función de las necesidades concretas de cada planta.

3.2. Modelización de una EDAR

Las herramientas de simulación permiten prever los cambios de comportamiento del sistema, realizar una optimización de los costes de explotación de la instalación o diseñar propuestas que mejoren la efectividad de la EDAR.

3.2.1. Modelo de fangos activados para la eliminación de materia orgánica y nitrógeno ASM1

El modelo ASM1 surgió ante la necesidad de un modelo matemático dinámico para controlar la eliminación de nitrógeno y elementos orgánicos en los procesos de lodos activados. Este modelo incorpora todas las transformaciones y componentes necesarias para describir la degradación de la materia orgánica en condiciones anaeróbicas y anóxicas (denitrificación) y el proceso de nitrificación posterior en el que el amonio es oxidado a nitratos.

El modelo fue presentado utilizando la notación matricial y comprende:

- 13 componentes en total: 7 disueltos y 6 de partículas en suspensión.
- 8 procesos: 3 de crecimiento, 2 de decaimiento y 3 de hidrólisis.

La materia orgánica carbonosa del modelo está dividida en Demanda Química de Oxígeno (DQO) biodegradable, DQO no biodegradable (materia orgánica inerte) y Biomasa. La fracción biodegradable se divide a su vez en una fracción rápidamente biodegradable (S_s, soluble) y en una fracción lentamente biodegradable (X_s, en suspensión). Se toma como hipótesis que la fracción rápidamente biodegradable está compuesta de materia orgánica soluble que se absorbe y metaboliza rápidamente por los microorganismos, mientras que la fracción X_s compuesta de partículas, coloides y materia orgánica compleja, sufre una hidrólisis enzimática para poder ser absorbida, por lo que es de asimilación más lenta.

La fracción no biodegradable de la DQO está dividida en una fracción soluble inerte (S_I) y en una fracción particulada (X_I) que no se ven afectadas por el proceso. S_I abandona la planta con el efluente del decantador secundario mientras que X_I se atrapa en el lodo purgado. La biomasa activa se divide en dos tipos de microorganismos, heterótrofos (X_{BH}) y autótrofos (X_{BA}). Una variable adicional, X_P, se introduce para modelar la fracción inerte de productos procedentes del decaimiento de la biomasa, pero en la realidad, en el lodo no se puede diferenciar X_P de X_I.

El nitrógeno total presente en el sistema incluye por un lado los nitratos y nitritos (S_{NO}), combinados en un solo componente para simplificar el modelo, y por otro el nitrógeno total Kjeldahl (TKN). El TKN se fragmenta en nitrógeno amoniacal (S_{NH}, que incluye el $N-NH4^+$ y $N-NH3$), nitrógeno orgánico y nitrógeno contenido en la biomasa. El nitrógeno orgánico también se divide en una fracción soluble y otra particulada, cada una teniendo su fracción biodegradable y no biodegradable. Solo las fracciones biodegradables, soluble (S_{ND}) y en suspensión (X_{ND}), aparecen en el modelo ASM1.

3.2.2. Modelos de fangos activados para la eliminación de materia orgánica, nitrógeno y fósforo ASM2 y ASM2d

El modelo ASM2 surgió ante la necesidad de eliminar el fósforo en las EDAR. Este modelo además de la eliminación del nitrógeno y elementos orgánicos incluye el proceso de acumulación del fósforo y la eliminación posterior que se produce gracias a las bacterias acumuladoras de P (X_{PAO}). En este modelo se observó una carencia respecto a la capacidad de las bacterias X_{PAO} de permanecer activas en condiciones anóxicas por lo que se amplió al modelo ASM2d que si tiene en cuenta estas características.

El modelo ASM2d incluye:

- 19 componentes en total: 9 componentes solubles y 10 suspendidos.
- 17 procesos biológicos, llevados a cabo por grupos de microorganismos heterótrofos, autótrofos y acumuladores de fósforo.

En los modelos ASM2/ASM2d el componente S_F representa el sustrato fermentable soluble y S_A la materia orgánica fermentable (ácido acético), el fósforo viene representado, en el efluente, por el fósforo soluble S_{PO4}. Los Poli-Hidroxi-Alcanatos (X_{PHA}) sirvan para generar el crecimiento de las bacterias X_{PAO} en condiciones anóxicas para almacenar el fósforo en forma de polifosfatos (X_{PP}).

En estos modelos el nitrógeno viene representado, en el efluente, por el amonio (S_{NH4}) y el nitrato (S_{NO3}). La actividad de las bacteria X_{BH} y la nitrificación de las bacterias X_{AUT} tienen el mismo funcionamiento que en el modelo ASM1, al igual que las fracciones particuladas X_I y X_S .

En los modelos ASM2/ASM2d se consideran los procesos de precipitación y redisolución de las sales ortofosfóricas pero no existe distinción entre X_I y X_P lo que supone una simplificación del modelo.

3.2.3. Modelo de fangos activados para la eliminación de materia orgánica y nitrógeno ASM3

El modelo ASM3 presenta una versión mejorada del ASM1 al detectarse algunas limitaciones en su funcionamiento e incluye:

- 10 componentes en total.
- 7 procesos bioquímicos.

Uno de los principales cambios fue la inclusión de compuestos almacenados en células heterótrofas, de manera que el crecimiento deja de depender en exclusiva del sustrato externo. Otra diferencia importante es que en ASM1 el decaimiento de los microorganismos sucede a una velocidad determinada mientras que en ASM3 la respiración endógena integra todas las formas de pérdida de biomasa y necesidades energéticas y no está relacionado con el crecimiento por lo que reduce la importancia del proceso de hidrólisis en las predicciones del modelo.

En el modelo ASM3 las bacterias autótrofas (X_A) llevan a cabo un proceso de nitrificación oxidando el S_{NH} a S_{NO} para posteriormente junto con X_{BH} sufrir un proceso de respiración endógena en condiciones aeróbicas y anóxicas convirtiéndose finalmente en producto inerte X_I.

3.3. Entornos de simulación

La simulacion proporciona un medio apropiado, de bajo coste y efectivo, para la evaluacion de estrategias de control y operacion, pero en un proceso tan complejo como una EDAR, la elevada cantidad de posibles escenarios sugirieron la necesidad de una cierta estandarizacion en el protocolo a seguir para poder comparar estrategias de control de una manera efectiva. Para poder realizar esta comparación, cada estrategia de control debe ser simulada bajo el mismo escenario y las mismas condiciones. Con

este proposito, y puesto que para el proceso de fangos activados existen modelos estandar ampliamente aceptados, se realizaron, bajo el paraguas de acciones COST de la union europea, unos protocolos de simulacion que han permitido estandarizar los estudios.

El IWA/COST Benchmark Simulation Model No. 1 (BSM1) se basa en el modelado del tratamiento secundario, es decir, donde tiene lugar el tratamiento biologico del agua residual. El BSM1 define el diseno de la planta, la composición del afluente (que consta de 15 variables), los procedimientos para llevar a cabo las simulaciones y los criterios para evaluar los resultados.

Para realizar las simulaciones se tienen en cuenta las condiciones del afluente definidas en la descripción 'benchmark simulatión' [4] [63]. Hay tres estados afluentes, cada uno representando un evento atmosférico diferente. Estos tres estados a estudiar son: el periodo seco, eventos de tormenta y eventos de lluvia.

Durante el tiempo lluvioso se produce un aumento de caudal sin añadir más contaminantes por lo que la cantidad de contaminantes totales en este tipo de eventos disminuye al verse diluidos por el aumento del caudal. La hora del día para el inicio de la lluvia es una variable aleatoria, con una distribución uniforme, es decir, cualquier momento del día es igual de probable para el comienzo del evento de lluvia. Lo mismo es aplicable para la finalización del evento La intensidad de la lluvia también puede ser variable en el tiempo.

En el tiempo tormentoso se produce tanto un aumento del caudal como de los contaminantes presentes. La intensidad de la tormenta y el tiempo transcurrido desde la última determinan el aumento del flujo de contaminantes. Un evento de tormenta puede combinarse con un evento de lluvia para modelar fenómenos de aumento de polución junto con la lluvia. También , para el evento de tormenta, el tiempo de inicio y de parada debe ser aleatorio.

En el modelo BSM1 el periodo de evaluación es de 14 días.

El diseño BSM1_LT [60] es una extensión del BSM1 en el que el periodo de evaluación se incrementa de manera significativa a un año y además se considera la temperatura variable lo que afecta a varios parámetros.

Como el efecto de estacionalidad influye de manera significativa en las operaciones de tratamiento de aguas residuales es importante tanto realizar una monitorización de al menos un año como comenzar el periodo de evaluación en un mes de verano para minimizar los riesgos del estudio.

Al aumentar de manera tan grande el periodo de evaluación en BSM1_LT respecto a BSM1, 364 frente a 14 días, es necesario tener en cuenta también una variación en la

temperatura que en BSM1 se considera constante e igual a 15ºC. BSM1_LT considera una variación sinusoidal de la temperatura con valores máximos en torno a 20.5º a principios de agosto y mínimos, sobre 9.5º, a principios de febrero. El coeficiente de transferencia de oxígeno Kla depende de la temperatura por lo que será uno de los factores a tener en cuenta a la hora de diseñar la planta.

En el BSM1, la inicialización consiste en simular la planta a evaluar durante 100 días con entrada constante, seguido de dos semanas de datos de entrada dinámica de tiempo seco, lluvioso o tormentoso. Esto se hace para obtener un estado casi constante y poder realizar un buen estudio posterior.

El período inicial debe ser más prolongado en BSM1_LT debido a que las estrategias de control que actúan en escalas de tiempo grandes, necesitan también una fase de estabilización extensa, por lo que se inicializará durante de seis meses para alcanzar un estado cuasi estacionario. Para que el proceso sea más realista se incluyen eventos lluviosos y tormentosos durante este periodo de inicialización que va desde el inicio de enero hasta finales de junio, y debe de ser diferente de los últimos 6 meses del período de evaluación.

3.4. Diseño de la planta

El primer diseño (BSM1) [26] combina la nitrificación con una denitrification previa, esta es una configuración que se utiliza comúnmente para conseguir la eliminación biológica de nitrógeno en plantas a gran escala. La planta de referencia se compone de un reactor de lodos activados compuesto por cinco compartimentos, dos tanques anaeróbicos seguido de tres tanques aeróbicos, a este reactor de lodos activados le sigue un decantador secundario. Se propone una estrategia básica de control para poner a prueba el punto de referencia: su objetivo es controlar el nivel de oxígeno disuelto en el compartimiento final del reactor mediante la manipulación del coeficiente de transferencia de oxígeno, y controlar el nivel de nitrato en el último compartimento anóxico mediante la manipulación de la caudal de reciclado interno.

Estos modelos utilizan el Modelo de Lodos Activados nº 1 (ASM1) [32] para modelar las reacciones biologicas que se producen en la planta depuradora y el decantador de tipo Takács de 10 capas y ha de tener las siguientes características:

- 5 tanques biológicos en serie con un decantador secundario.
- Volumen total de 5999 m^3 (tanques 1 y 2, cada uno con $1000m^3$ y tanques 3, 4 y 5 con $1333m^3$ de volumen).

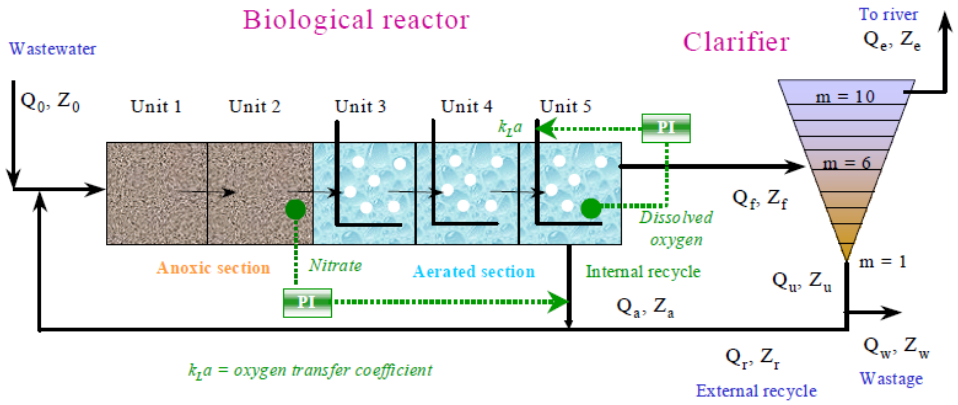

Figura 3.1: Diseño esquemático de la planta BSM1/BSM1_LT [26]

- Tanques 1 y 2 anaeróbicos pero recibiendo flujo mezclado.
- Aireación de los tanques 3, 4 y 5 activada, con un valor máximo de Kla = 10 h^{-1} .
- Valores de Kla por defecto en los tanques 3 y 4 y Kla = 3.5 h^{-1} en el tanque 5.
- Saturación del óxigeno disuelto con valor constante en los tanques 3, 4 y 5, So_sat = 8 $gO_2\, m^{-3}$.
- Decantador secundario formado por 10 capas de 1500 m^2 de superficie y un espesor de 4 m. El volumen total del decantador será de 6000 m^3 .
- Un punto de alimentación del decantador situado a 2.2 m del fondo, situado en el centro de la sexta capa.
- Dos recirculaciones internas:
  1. Recirculación interna desde el tanque 5 al 1 con flujo de Qa = 55338 $m^3 d^{-1}$.
  2. Flujo de recirculación (RAS) desde el fondo del decantador secundario al comienzo de la planta con un flujo predeterminado de Qr = 18446 $m^3 d^{-1}$.
- Flujo de lodos (WAS) bombeados desde el fondo del decantador secundario a un ritmo predeterminado de Qw = 385 $m^3 d^{-1}$.

	Configuración	Unidades
Volumen - Tanque 1	1000	m^3
Volumen - Tanque 2	1000	m^3
Volumen - Tanque 3	1333	m^3
Volumen - Tanque 4	1333	m^3
Volumen - Tanque 5	1333	m^3
Espesor - Decantador	4	m
Superficie - Decantador	1500	m^2
Volumen - Decantador	6000	m^3

Tabla 3.1: Características físicas del biorreactor y el decantador

	Caudales	Unidades
Caudal del Afluente	18446	$m^3 dia^{-1}$
Caudal de Recirculación	18446	$m^3 dia^{-1}$
Caudal de Recirculación Interna	55338	$m^3 dia^{-1}$
Caudal de Lodos	385	$m^3 dia^{-1}$
KLa – Tanque 1	–	–
KLa – Tanque 2	–	–
KLa – Tanque 3	10	hr^{-1}
KLa – Tanque 4	10	hr^{-1}
KLa – Tanque 5	3.5	hr^{-1}

Tabla 3.2: Configuración de variables del sistema

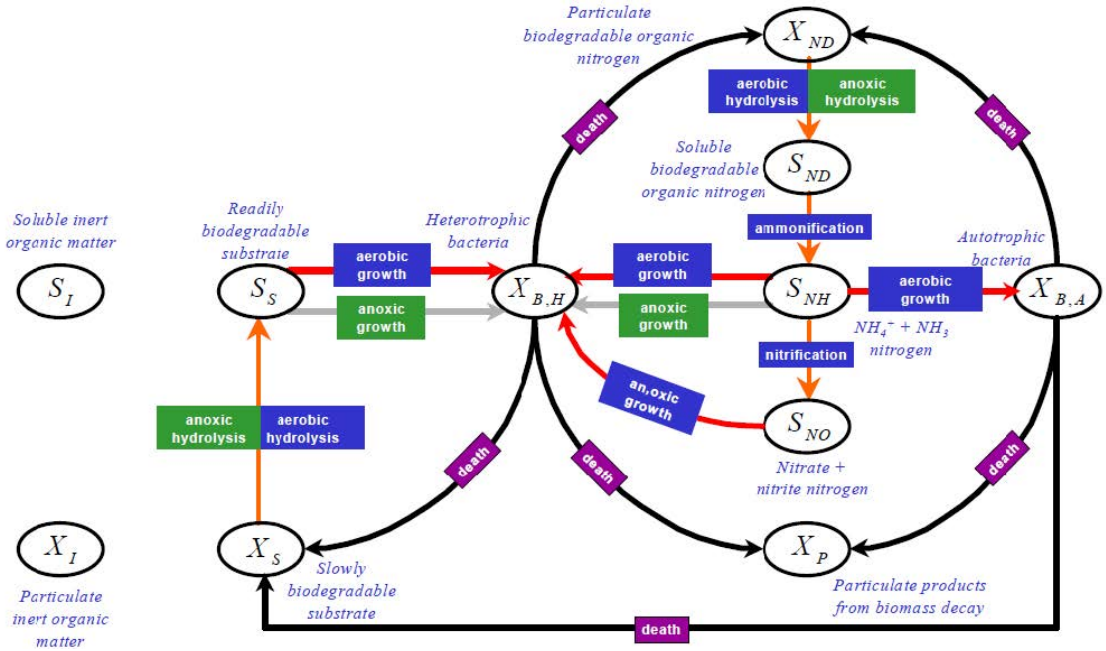

Figura 3.2: Funcionamiento del modelo ASM1

Las características físicas del biorreactor y el decantador se enumeran en la Tabla 3.1 y las variables necesarias para configurar el sistema en la Tabla 3.2.

BSM1 también determina los valores de los parámetros estequiométricos y cinéticos del proceos ASM1 a una temperatura de 15^{o}C y como modelo de sedimentación la función doble exponencial de velocidad de sedimentación de Takács [64], basada en el concepto de flujo de sólidos, y que es aplicable tanto en condiciones de sedimentación impedida o retardada como en la floculación.

3.5. Procesos biológicos

Los fenómenos biológicos que tienen lugar en el reactor se modelizan según los criterio especificados en ASM1, descritos de manera esquemática en la Figura 3.2.

ASM1 utiliza 13 componentes (variables de estado) y 8 procesos. Las variables de estado se describen en la Tabla 3.3.

3.6. Conclusión

La creación de modelos de simulación de referencia para el control de EDARs es un campo que lleva desarrollándose por parte de la IWA desde hace más de de quince

Definición	Notación	Unidades
Materia orgánica inerte soluble	Si	g COD m^{-3}
Sustrato fácilmente biodegradable	Ss	g COD m^{-3}
Partículas de materia orgánica inerte	Xi	g COD m^{-3}
Sustrato lentamente biodegradable	Xs	g COD m^{-3}
Biomasa activada heterótrofa	Xbh	g COD m^{-3}
Biomasa activada autótrofa	Xba	g COD m^{-3}
Partículas de productos derivados de la descomposición de la biomasa	Xp	g COD m^{-3}
Oxígeno	So	g COD m^{-3}
Nitratos y Nitritos	Sno	g N m^{-3}
Amonios $NH_4^+ + NH_3$	Snh	g N m^{-3}
Nitrógeno orgánico soluble biodegradable	Snd	g N m^{-3}
Partículas de nitrógeno orgánico biodegradable	Xnd	g N m^{-3}
Alcalinidad	Salk	mol L^{-1}

Tabla 3.3: Variables de estado ASM1

años.

La evaluación y comparación práctica o basada en la simulación de una EDAR es difícil. Esto se debe en parte a la variabilidad del inuente, a la complejidad de los fenómenos biológicos y bioquímicos y a la gran variedad del tiempo de operación (de unos pocos minutos a varios días). La falta de criterios de evaluación estándar es también una tremenda desventaja. Para mejorar realmente la aceptación de estrategias de control innovadoras, dicha evaluación debe basarse en una metodología rigurosa que incluya un modelo de simulación, diseño de la planta, controladores, sensores, criterios de rendimiento y procedimientos de prueba, es decir, un protocolo completo.

El BSM1 se ha convertido en la herramienta de simulación estándar para la evaluación del rendimiento de las técnicas de control aplicadas a las plantas de tratamiento de aguas residuales.

El modelo BSM1_LT además de contar con las características de evaluación del protocolo BSM1 incluye un tiempo de operación más realista, la variabilidad estacional del caudal influente y la dependencia de la temperatura.

Capítulo 4

Estudio del tiempo atmosférico

4.1. Introducción

Para que el tratamiento de las aguas residuales sea más eficaz y eficiente es importante conocer, al comienzo del proceso, las características del agua afluente. Los índices químicos y biológicos como el pH, la conductividad, el DO, parámetros de nutrientes, carga orgánica, sólidos, lodos, caudal, etc, son proporcionados por sensores [35] pero resulta de interés conocer el tiempo atmosférico correspondiente a cada punto de entrada porque es una ayuda fundamental para controlar la cantidad de oxígeno que deben suministrar los aireadores y de esta forma ser energéticamente más eficientes .

Hoy en día hay gran cantidad de datos que se recogen y almacenan en bases de datos de todo el mundo y se han creado múltiples algoritmos para extraer la información necesaria de estos grandes conjuntos de datos. Existen varias metodologías diferentes para abordar este problema: clasificación, regla de asociación, agrupación, etc. En este trabajo se utilizarán técnicas de clasificación.

La clasificación supervisada consiste en predecir un determinado resultado basado en una entrada dada. Con el fin de predecir el resultado, el algoritmo procesa un conjunto de entrenamiento que contiene un grupo de atributos y el propio resultado, usualmente llamado objetivo o atributo de predicción. El algoritmo intenta descubrir relaciones entre los diversos atributos que harán posible predecir el resultado. A continuación se le da al algoritmo un conjunto de datos no visto antes, llamado conjunto de predicción, que contiene el mismo conjunto de atributos, excepto el atributo de predicción, aún no conocido. El algoritmo analiza la entrada y produce una predicción. La precisión de la predicción define cómo es de «bueno» el algoritmo.

En la clasificación no supervisada el objetivo es descubrir grupos de instancias o relaciones similares dentro de los datos. En este enfoque no existe información

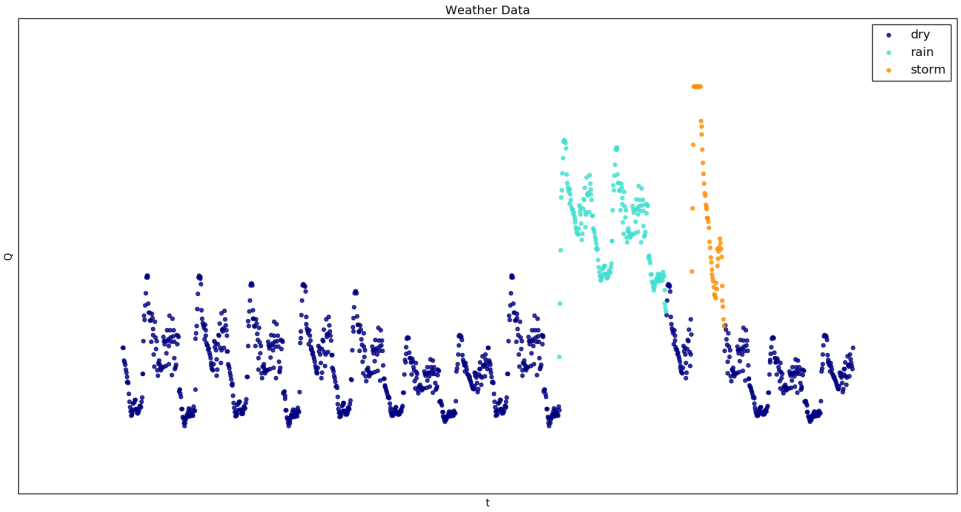

Figura 4.1: Caudales del Influente BSM1

sobre la etiqueta de clase de datos o cuántas clases hay. La técnica de aprendizaje no supervisado más importante es el clustering, que inicialmente creará diferentes grupos de datos de entrada y será capaz de poner cualquier nueva entrada en el clúster apropiado.

Para realizar las clasificaciones se utiliza el lenguaje Python junto con el paquete scikit-learn.

Scikit-learn [53] es una biblioteca Open Source de Aprendizaje Automático escrita en Python que contiene numerosos algoritmos y utilidades para realizar tareas con conjuntos de datos.

4.2. Clasificación supervisada en BSM1

Los datos de influente utilizados para trabajar en BSM1 incluyen un atributo para indicar el tiempo atmosférico (seco, lluvia o tormenta) que hay en cada momento como se representa en la Figura 4.1.

4.2.1. Transformación de los datos

La estandarización del conjunto de datos es un requisito común para muchos estimadores de aprendizaje implementados en scikit-learn, ya que los datos podrían comportarse de manera incorrecta si las características individuales no se parecen más o menos a los datos estándar normalmente distribuidos, con media cero y varianza uno.

En la práctica, a menudo ignoramos la forma de la distribución y simplemente transformamos los datos para centrarlos, eliminando el valor medio de cada característica para luego escalarlos dividiendo por su desviación estándar.

Por ejemplo, muchos elementos utilizados en la función objetivo de un algoritmo de aprendizaje asumen que todas las características están centradas alrededor de cero y tienen varianza en el mismo orden. Si una característica tiene una varianza mayor que las demás podría dominar la función objetivo y hacer que el estimador no pueda aprender de otras característica que le puedan proporcionar una información adecuada.

4.2.2. Selección de atributos

Al tener un gran volumen de datos, muchas tuplas con muchos atributos, la aplicación de técnicas para reducir la dimensionalidad antes de tratar el problema es indispensable.

Se puede valorar si los atributos de un determinado subconjunto producen una mejor clasificación que otro subconjunto. Mediante un sistema de votos, y tras un gran número de repeticiones, se obtendrá una ordenación de los atributos que son necesarios para producir buenas clasificaciones.

4.2.2.1. Contraste de hipótesis χ^2

La prueba χ^2 mide la dependencia entre las variables estocásticas, por lo que usar esta función elimina las características que son más probables de ser independientes de la clase y por lo tanto irrelevantes para la clasificación. Esta prueba solo se puede aplicar a conjuntos de datos que contengan únicamente características no negativa, como es el caso del influente BSM1.

Se calculan las estadísticas χ^2 entre cada característica y su clase. Esta puntuación se puede utilizar para seleccionar las características con los valores más altos, que serán las que nos darán una mejor clasificación [54].

Se utiliza un contraste de hipótesis usando la distribución χ^2. El estadístico de contraste será

$$\chi 2 = \sum \frac{(y_i - z_i)^2}{y_i} \tag{4.1}$$

El numerador de cada término es la diferencia entre la frecuencia observada y la frecuencia esperada. Por tanto, cuanto más cerca estén entre sí ambos valores más pequeño será el numerador, y viceversa. El denominador permite relativizar el tamaño del numerador. Cuanto menor sean el valor del estadístico χ^2, más coherentes serán las

observaciones obtenidas con los valores esperados. Por el contrario, valores grandes de este estadístico indicarán falta de concordancia entre las observaciones y lo esperado. En este tipo de contraste se suele rechazar la hipótesis nula (los valores observados son coherentes con los esperados) cuando el estadístico es mayor que un determinado valor crítico.

Se calculan las estadísticas χ^2 entre cada característica no negativa y su clase. Esta puntuación se puede utilizar para seleccionar las características que darán una mejor clasificación.

Aplicando esta prueba al influente BSM1 se obtiene:

Fit Scores:
[4,340e+06 1,846e+03 3,599e+02 1,686e+03 3,599e+03 1,309e+03 1,690e+03 4,921e+02 1,016e+04]

Las mejores variables son, por este orden: Q, Si, Ss y Xs.

4.2.2.2. Regresión Logística

El modelo de regresión logística se utiliza para pronosticar la probabilidad de que ocurra o no un suceso determinado. Sirve para estudiar el impacto que tiene cada una de las variables explicativas en la probabilidad de que ocurra el suceso en estudio, en el caso del influente BSM1 se estudiará que variables influyen en la clasificación y cuales no.

Es una técnica multivariante de dependencia ya que trata de estimar la probabilidad de que ocurra un suceso en función de la dependencia de otras variables.

Se parte de un conjunto de variables independientes X_1, X_2,...,X_p que caracterizan a los n datos, la regresión logística clasifica cada dato en una de las dos categorías de la variable Y (True o False). La probabilidad de que un sujeto «i» pertenezca a una de ellas será la combinación lineal $Z = b_1X_1 + b_2X_2 + ... + b_pX_p + b_0$ será igual a:

$$P_i = \frac{1}{1 + e^{-Z}} \tag{4.2}$$

Si la probabilidad P_i de que el sujeto este encuadrado en esa categoría es mayor que 0,5 se le asigna True, si es menor se le asignará False.

Se trata de predecir esta probabilidad ya no sólo para los sujetos observados (los de la muestra que están en los datos) sino para la población en general. Para ello se calcula la función Z considerando una muestra aleatoria del conjunto de datos, después se realiza sobre la totalidad y se mide si el número de aciertos es suficientemente elevado. Esto constituye una prueba de su capacidad de predicción.

Al aplicar Regresión Logística a BSM1 para seleccionar 4 variables,

```
Num Features: 4
Selected Features: [ True True False False False False True False True]
Feature Ranking: [1 1 2 3 4 5 1 6 1]
```

Las variables seleccionadas mediante este método son : Q, Si, Xi y Xs.

4.2.2.3. Extra Trees Classifier

El objetivo de los métodos de ensamblado consiste en construir un modelo predictivo por integración de múltiples modelos mejorando el rendimiento de las predicciones. Se consigue combinando las predicciones de varios modelos realizados con uno o varios algoritmos de aprendizaje y generalizándolos para dar como resultado un modelo único.

Dentro de los métodos de ensamblado el Extra Trees Classifier constituye un metaestimador que ajusta un número de árboles de decisión aleatorio en varias submuestras del conjunto de datos original usando el promedio para mejorar la precisión de la predicción y controlar el sobreajuste. Se trata de árboles de decisión sin poda que responden al clásico procedimiento top-down y forman parte de la técnica perturbación-y-combinación con dos diferencias importantes frente a los árboles de decisión: por un lado, los nodos se crean a partir de puntos de corte acordados completamente al azar; y por otro, para el crecimiento de los árboles se utiliza el total de la muestra [27]. Su aplicación se suele abordar cuando se pretende aumentar la precisión de la predicción dada por un árbol de decisión.

En el influente BSM1 se consigue:

```
Features Importance: [ 0.16 0.507 0.076 0.104 0.073 0.014 0.014 0.02 0.031]
```

Por lo que las variables con mejor puntuación son por orden: Si, Q, Snd y Ss.

4.2.3. Reducción de la dimensionalidad

Los métodos de reducción de dimensionalidad son técnicas estadísticas que distribuyen el conjunto de los datos en subespacios derivados del espacio original, de menor dimensión, que permiten hacer una descripción de los datos de manera más eficiente.

Es decir, la reducción consiste en transformar el conjunto de variables iniciales en un conjunto de menor dimensión que sea capaz de retener la mayor parte de la información.

La ventaja de estos métodos es que no utilizan información de la variable respuesta, y por tanto se pueden utilizar datos no etiquetados (análisis semisupervisado). Haciendo esto conseguimos una mejor representación de la información, ya que los

modelos de extracción aprenden relaciones entre predictores que luego serán de utilidad para fases posteriores. A cambio, el inconveniente es que tras la transformación de los datos estos dejan de tener cualquier interpretación sencilla, ya que de hecho ni siquiera van a tener expresiones de las variables originales, salvo las combinaciones lineales que se obtienen en PCA [61].

4.2.3.1. Análisis de las Componentes Principales (PCA)

Dado un conjunto de puntos en un espacio multidimensional, PCA realiza un cambio del sistema de coordenadas de manera que las primeras dimensiones en dicho sistema recojan la mayor variabilidad posible de los datos. PCA asume que los datos siguen distribuciones normales y que tienen una representación lineal en cierta base.

En general la técnica de PCA se usa para reducir la dimensión de los datos, en nuestro caso el número de variables, teniendo en cuenta que PCA es efectivo cuando la correlación entre variables es alta, es decir; la linealidad entre las variables no es cero.

Este método realiza una transformación lineal de las variables iniciales, para proyectar vectores propios ortonormales denominados componentes principales [54].

El objetivo deseado es reducir las dimensiones de un conjunto de datos d-dimensional al proyectarlo en un subespacio k-dimensional, donde k $<$ d, con el fin de aumentar la eficiencia computacional, manteniendo la mayor parte de la información. Para ello se calculan los vectores propios (los componentes principales) de un conjunto de datos y se recogen en una matriz de proyección. Cada uno de esos vectores propios está asociado con un valor propio que puede interpretarse como la longitud o magnitud del vector propio correspondiente. Es razonable reducir del conjunto de datos a través de PCA a un subespacio dimensional más pequeño al eliminar los pares que proporcionan menos información. Los vectores propios (componentes principales) determinan las direcciones del nuevo espacio de características, y los valores propios determinan su magnitud. Es decir, los valores propios explican la varianza de los datos a lo largo de los nuevos ejes de características.

Al aplicar PCA de 2 componentes a BSM1 se obtiene,

PCA Meaning of the 2 components: 0.12 Q - 0.29 Si - 0.36 Snd - 0.35 Snh - 0.36 Ss - 0.37 Xbh - 0.34 Xi - 0.37 Xnd - 0.36 Xs 0.74 Q - 0.50 Si - 0.00 Snd - 0.11 Snh + 0.00 Ss + 0.17 Xbh + 0.36 Xi + 0.17 Xnd + 0.09 Xs

Que queda representado en la Figura 4.2 junto con el de 3 componentes principales.

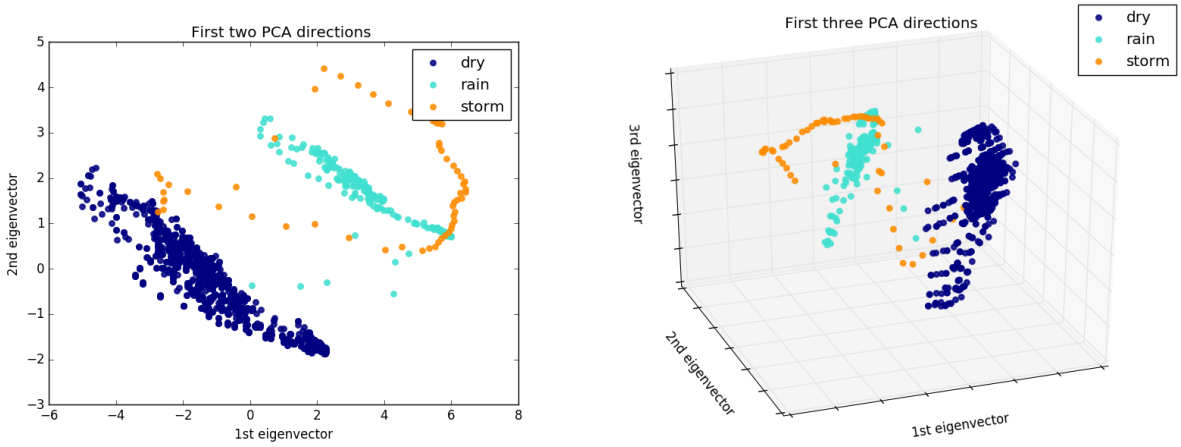

Figura 4.2: Análisis de las Componentes Principales PCA 2 y 3 componentes

4.2.3.2. Análisis Discriminante Lineal (LDA).

LDA también tiene como objetivo encontrar las direcciones que maximizan la separación (o discriminación) entre las diferentes clases, que pueden ser útiles en el problema de clasificación de patrones.

Se recomienda para tareas de aprendizaje supervisado y clasificación de clases, ya que proyecta los datos a una baja dimensión en comparación de los datos originales, pero garantiza la máxima dispersión entre clases para poder reducir al mínimo posible la dispersión interna de cada clase.

LDA, en contraste con PCA, es un método supervisado y utiliza las etiquetas de clase.

Los resultados del análisis LDA en BSM1 son:

```
LDA homogeneity score: 0.804236
LDA overall mean: [ -6.245e-16 4.871e-15 -2.012e-16 -1.388e-16 3.816e-16 -7.355e-16
-5.343e-16 6.870e-16 5.343e-16]
```

Y se muestran representados en la Figura 4.3.

4.2.4. Clasificación y validación

Se realizan clasificaciones supervisadas con los datos BSM1 estandarizados aplicando los algoritmos Naive Bayes, K Nearest Neighbor (k-NN), Classification Trees, Random Forest y SVM, posteriormente se comprueba su idoneidad mediante Validación Cruzada. Los algoritmos han sido elegidos teniendo en cuenta su adecuación a las características de los datos [12].

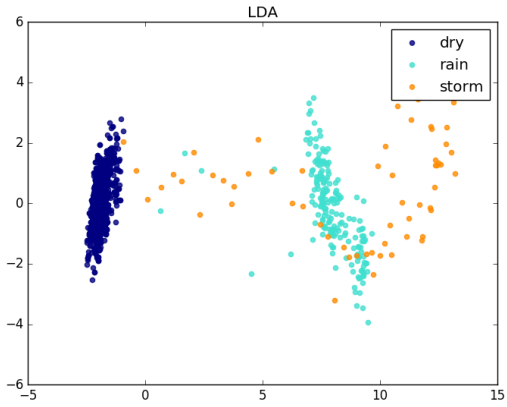

Figura 4.3: Análisis Discriminante Lineal (LDA)

Naive Bayes es uno de los clasificadores más utilizados por su simplicidad y rapidez. Está basada en el Teorema de Bayes, también conocido como teorema de la probabilidad condicionada, se basa en encontrar la hipótesis más probable que describa a ese caso concreto [44].

El algoritmo k-NN se fundamenta en la idea básica de que cada un nuevo caso se va a clasificar en la clase más frecuente a la que pertenecen sus K vecinos más cercanos. El paradigma se fundamenta en una idea muy simple e intuitiva, lo que unido a su fácil implementación hace que su uso sea muy extendido [17].

Entre los posibles mecanismos para obtener predicciones de manera fiable, una de las que más destaca es la creación de Árboles de Decisión (Classification Trees), que proporcionan un conjunto de reglas que se van aplicando sobre los ejemplos nuevos para decidir qué clasificación es la más adecuada a sus atributos. Un árbol de decisión está formado por un conjunto de nodos de decisión y de nodos-respuesta (hojas), cada nodo del árbol es un atributo (campo) de los ejemplos, y cada rama representa un posible valor de ese atributo [44].

Random Forest es uno de los algoritmos más potentes usados hoy en día. Consiste en una combinación de múltiples árboles de decisión que juntos conforman un "bosque", a cada uno de los árboles de decisión se le asigna una porción de los datos de ejemplo o entrenamiento. El resultado final será la combinación del criterio de todos los árboles que formen el bosque [10].

Las Máquinas de Vectores de Soporte o Support Vector Machines (SVM) son un conjunto de algoritmos de aprendizaje supervisado desarrollados por Vladimir Vapnik. Consiste en construir un hiperplano en un espacio de dimensionalidad muy alta, o incluso infinita, que separe las clases que tenemos. Una buena separación entre las

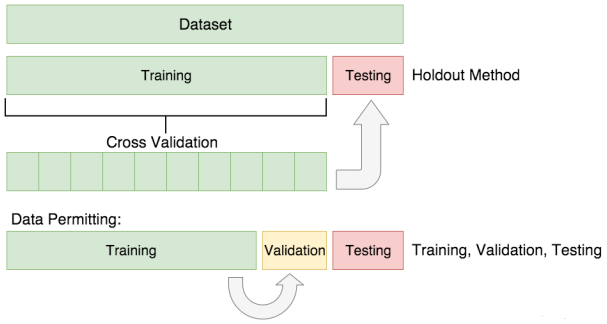

Figura 4.4: Conjuntos de Entrenamiento, Prueba y Validación.

clases permitirá un clasificación correcta de la nueva muestra, es decir, necesitamos encontrar la máxima separación a los puntos más cercanos a este hiperplano [19].

Para validar los modelos generados se recurrirá a la Validación Cruzada (CV), que es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. Consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones. Se utiliza en entornos donde el objetivo principal es la predicción y se quiere estimar cómo de preciso es un modelo que se llevará a cabo a la práctica [20]. En la validación cruzada de K iteraciones o K-fold Cross-Validation los datos de muestra se dividen en K subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento. El proceso de validación cruzada es repetido durante K iteraciones, con cada uno de los posibles subconjuntos de datos de prueba. Finalmente se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado.

El valor de K suele ser 5 o 10, pero no hay una norma establecida. A medida que K aumenta, la diferencia de tamaño entre el conjunto de entrenamiento y los subconjuntos de remuestreo se reduce. A medida que esta diferencia disminuye, el sesgo se vuelve más pequeño. Con estos valores de K se ha demostrado empíricamente que se pueden producir estimaciones con tasas de error que no sufren de sesgo excesivamente alto (como una sobreestimación de la habilidad del modelo) ni de varianza muy alta (que puede cambiar mucho en función de los datos utilizados para ajustarse al modelo [39].

En valor de K = 10 elegido es muy común en el campo del aprendizaje automático aplicado.

Para analizar los datos BSM1 se usará un conjunto de entrenamiento formado por el 80 % de los datos, utilizando el 20 % restante para la prueba, como se ilustra en la Figura 4.4.

Algoritmo	Matriz de confusión	Validación cruzada
Gaussian Naive Bayes	226 0 0 0 28 3 0 4 8	0.9656133829
k-Nearest Neighbor	226 0 0 1 30 0 2 4 6	0.984200743494
Random Forest	226 0 0 1 30 0 0 3 9	0.986918215613
Decision Tree	226 0 0 0 31 0 0 3 9	**0.986988847584**
Support Vector Machine	226 0 0 1 30 0 1 4 7	0.978624535316

Tabla 4.1: Clasificaciones con Validación Cruzada.

Los resultados de las clasificaciones y su correspondiente validación se muestran en la Tabla 4.1. El mejor valor de la validación cruzada se obtiene con los Árboles de Decisión.

El árbol de decisión, basado en el índice de Gini, de profundidad 13 con la clasificación para BSM1 se presenta en la Figura 4.5.

4.2.5. Índice de intensidad de precipitación

Para intentar mejorar la clasificación con un parámetro que tenga en cuenta la duración de los eventos atmosféricos, se introduce el índice *«n»* de la intensidad de precipitación, método propuesto por Monjo en 2009 [37]. El índice *«n»* describe el comportamiento de la precipitación en función del tiempo a lo largo del evento y permite su clasificación según un tipo de curva que caracteriza la intensidad de los eventos. Esta metodología clasifica el evento según la regularidad de la precipitación respecto al tiempo por lo que ayudará a diferenciar los eventos lluviosos de los tormentosos.

El índice de precipitación *«n»* viene definido por:

$$I = I_0(\frac{t_0}{t})^n \tag{4.3}$$

$$n = \frac{\log(\frac{I}{I_0})}{\log(\frac{t_0}{t})} \tag{4.4}$$

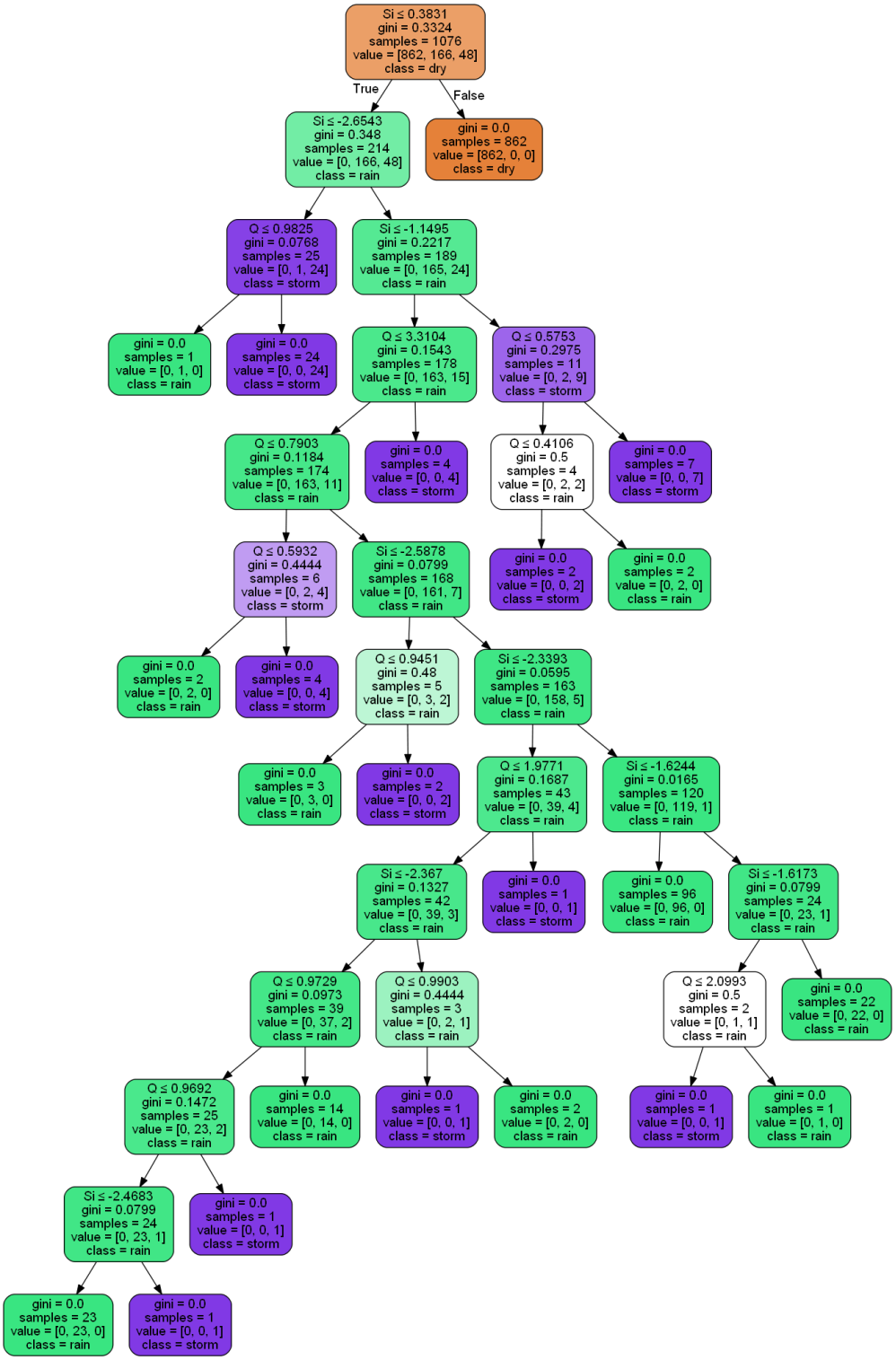

Figura 4.5: Árbol de Decisión de profundidad 13 para BSM1

Algoritmo	Matriz de confusión	Validación cruzada
Gaussian Naive Bayes	226 0 0 0 28 3 0 3 9	0.9656133829
k-Nearest Neighbor	226 0 0 1 29 1 1 3 8	0.983271375465
Random Forest	226 0 0 0 29 2 0 3 9	0.986847583643
Decision Tree	226 0 0 0 30 1 0 3 9	**0.986988847584**
Support Vector Machine	226 0 0 1 30 0 2 3 7	0.979553903346

Tabla 4.2: Clasificaciones con Validación Cruzada con el índice «*n*».

Al realizar de nuevo las clasificaciones y su validación posterior, añadiendo esta nueva variable a las utilizadas con anterioridad se obtienen los resultados de la Tabla 4.2, muy similares a los obtenidos sin introducir el índice de intensidad de precipitación, sin embargo el árbol de decisión resultante tiene una profundidad de 14.

4.2.6. Transformada de Fourier

La transformada de Fourier es una función matemática empleada para transformar señales entre el dominio del tiempo y el de la frecuencia [9].

Aplicando esta transformación a los datos BSM1 se intenta clasificar los eventos atmosféricos en función de la frecuencia con que estos ocurren, de manera que queden claramente diferenciados los tiempos seco, lluvioso y tormentoso.

Dadas las características del conjunto BSM1 se aplica una Transformada de Fourier de Señales Discretas (DTFT). La DTFT se aplica en este caso a una señal discreta en el tiempo$x[n] = Q$, con periodo de muestreo $ts = \frac{1}{f_s} = 0{,}25$ y aperiódica y se obtiene una función $X(f)$, que es continua como función de la frecuencia y periódica con periodo $F = \frac{1}{T_S}$. El periodo de muestreo coincide con el intervalo de los datos, que se toman cada 15 minutos.

Partiendo de una señal aperiódica discreta $x[n]$ la transformamos en una señal periódica continua $X(f)$ mediante la DTFT, centrando la frecuencia a 0. Los resultados gráficos se ven en la Figura 4.6.

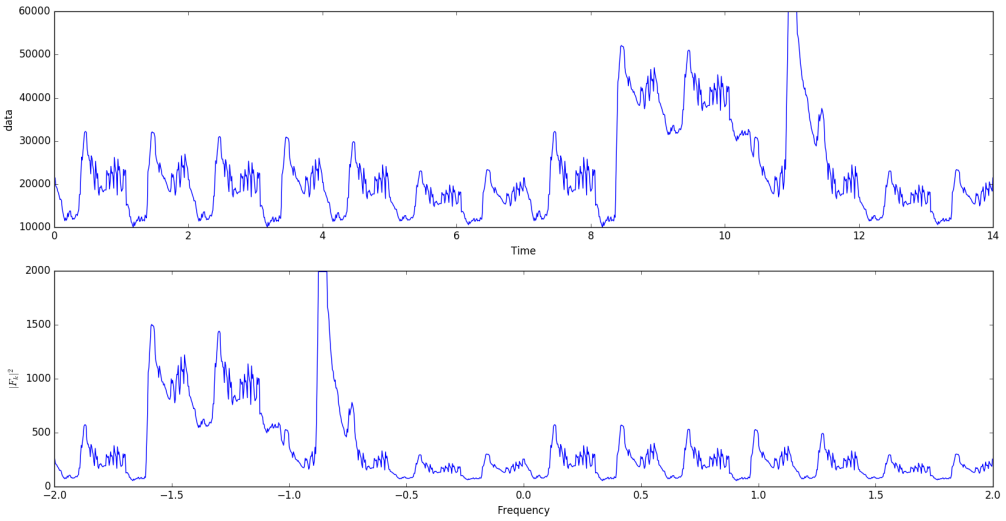

Figura 4.6: Transformada Discreta de Fourier

Los resultados de aplicar los algoritmos de clasificación validados con CV a los datos transformados con Fourier se muestran en la Tabla 4.3.

4.2.7. Clasificación utilizando validación con 5X2 CV Paired test de Dietterich

Para evaluar con otro enfoque las diferencias entre las técnicas se procedió a realizar una estimación y validación cruzada siguiendo el algoritmo de Dietterich 5x2 Cross-Validation Paired t-Test [5].

En este método de validación cruzada el valor cinco se refiere al número de replicaciones durante el proceso de entrenamiento y el valor 2 se refiere al número de grupos en los que se dividen los datos originales.

El conjunto de datos global S se divide en dos grupos S1 y S2 que deben cumplir: $S = S_1 \cup S_2$ y $S_1 \cap S_2 = \varnothing$

Para llevar a cabo el proceso de entrenamiento y estimación. Este proceso se repite 5 veces e incluye los siguientes pasos:

Primero el clasificador se entrena usando S_2 y se usa para clasificarS_2 y S_1, en el segundo paso se entrena usandoS_1 y se usa para clasificar S_1 y S_2 . La Figura 4.7 representa na interpretación gráfica de este test.

Para realizar este test con SKLearn se utiliza una combinación de GridSearch y CrossValidation. GridSearch es una técnica de optimización de los hiperparámetros de un modelo. En scikit-learn esta técnica se proporciona en la clase GridSearchCV.

Algoritmo	Matriz de confusión	Validación cruzada
Gaussian Naive Bayes	226 0 0 0 26 5 0 6 6	0.956319702602
k-Nearest Neighbor	226 0 0 1 29 1 2 5 5	0.978624535316
Random Forest	226 0 0 0 29 2 0 6 6	0.9820342007435
Decision Tree	226 0 0 0 30 1 0 4 8	**0.980483271375**
Support Vector Machine	226 0 0 1 30 0 3 6 3	0.959107806691

Tabla 4.3: Clasificaciones con Validación Cruzada aplicando transformada de Fourier.

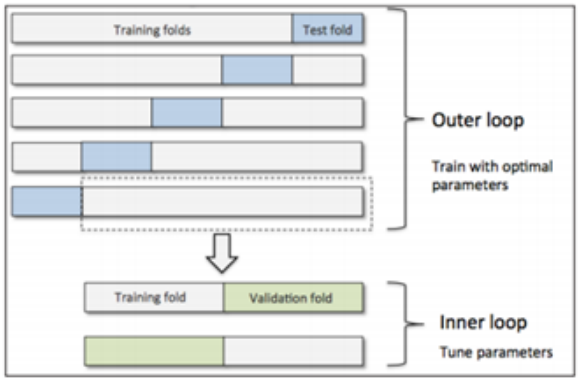

Figura 4.7: 5X2 CV Paired test de Dietterich

	BSM1	BSM1+n	BSM1 Fourier
Gaussian Naive Bayes	0.967	0.967	0.951
k-Nearest Neighbor	0.956 (n_neighbor=5)	0.956 (n_neighbor=5)	0.953 (n_neighbor=3)
Random Forest	0.990 (n_estimators=20)	0.989 (n_estimators=10)	0.983 (n_estimators=20)
Decision Tree	0.988 (max_depth=13)	0.980 (max_depth=13)	**0.988 (max_depth=7)**
Support Vector Machine	0.986 (kernel=rbf)	0.986 (kernel=rbf)	0.983 (kernel=rbf)

Tabla 4.4: Clasificación con validación 5X2 CV

Se debe proporcionar el modelo a estimar y una conjunto de valores para probar. GridSearchCV construye y evalúa el modelo para cada combinación de parámetros.

Se utiliza la validación cruzada para evaluar cada modelo individual especificando el argumento cv al constructor GridSearchCV. Una vez evaluado el modelo (utilizando CrossValidation con cv=2) mediante GridSearcCV se realiza CrossValidation con

Al construir esta clase se debe proporcionar un diccionario de hiperparámetros para evaluar su precisión.

Se aplica esta técnica de validación a los cinco algoritmos utilizados anteriormente con el conjunto de datos del benchmark BSM1 y añadiendo el índice *«n»* de precipitación y también a los datos transformados con Fourier. Los valores de clasificación obtenidos así como los parámetros que optimizan cada uno de los algoritmos se muestran en la Tabla 4.4.

Los valores obtenidos al clasificar utilizando la técnica 5X2 CV Paired test son mejores que los que se consiguieron al aplicar validación cruzada, rozando el 99 %. Además el árbol de decisión basado en el índice de Gini obtenido con los datos transformados con Fourier, representado en la Figura 4.8, tiene una profundidad de 7, mucho menor que los árboles de profundidad 13 obtenidos anteriormente.

4.3. Clasificación Supervisada en BSM1_LT

El modelo BSM1_LT, a diferencia de BSM1, no incluye una variable que indique el tiempo atmosférico existente en cada instante estudiado. Como este parámetro es importante a la hora de controlar la cantidad de oxígeno que hay que suministrar en la EDAR para conseguir un efluente que satisfaga las condiciones de vertido, se va a intentar realizar una clasificación supervisada basándose en los resultados previamente obtenidos al clasificar el modelo BSM1. Para ello se crea un fichero añadiendo a los

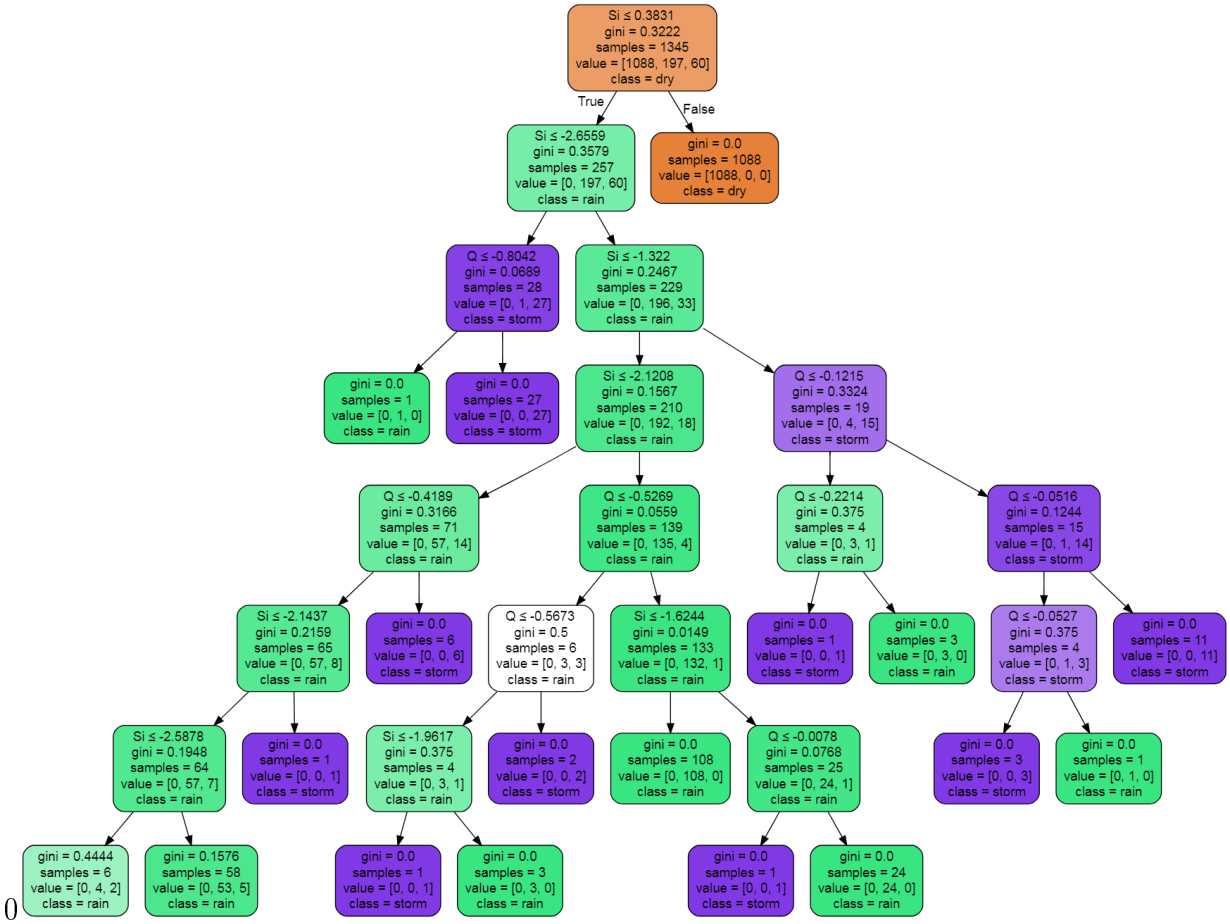

Figura 4.8: Árbol de Decisión de BSM1 transformado con Fourier.

	BSM1	BSM1_LT
tiempo seco	70 %	85 %
tiempo lluvioso	20 %	14 %
tiempo tormentoso	10 %	1 %

Tabla 4.5: Porcentaje de datos por tipo de tiempo atmosférico

datos de BSM1 ya clasificados, los de BSM1_LT (58465 datos), como un cuarto grupo independiente para ver como se realiza la clasificación supervisada.

Se clasifica utilizando los mismos algoritmos, Naive Bayes, k-NN, Classification Trees, Random Forest y SVM, y validando posteriormente con Validación Cruzada y 5x2 Cross-Validation Paired t-Test.

Analizando las matrices de confusión se observa que el grupo con los datos procedentes de BSM1_LT se clasifica aparte y no queda integrado en los grupos 0 (tiempo seco), 1 (tiempo lluvioso) y 2 (tiempo tormentoso), por lo que parece que la clasificación supervisada integrando los dos archivos BSM1 y BSM1_LT no es factible.

Los dos archivos son incompatibles entre si, fundamentalmente por las diferencias entre por los porcentajes de datos de cada tiempo atmosférico [24], como se puede observar en la Tabla 4.5.

4.4. Clasificación No Supervisada en BSM1_LT

En este tipo de clasificación se trabaja con datos que tiene un conjunto de características, de las que se desconoce a que clase o categoría pertenecen, por lo que la finalidad es el descubrimiento de grupos de datos cuyas características afines permitan separar las diferentes clases.

Para realizar la clasificación no supervisada se aplican los algoritmos Cluster K-Means, Mean-Shift, Ward Hierachical Clustering y BIRCH por ser una buena representación de los algoritmos de clustering [65].

Se utilizará como archivo de datos BSM1_LT al que se ha añadido el indice *«n»* de intensidad de precipitación para poder diferenciar de manera más precisa los eventos lluviosos de los tormentosos.

Una manera muy útil de segmentar y entender el tipo de datos que se está estudiando es ver cómo se organizan estos datos de forma natural. Para eso se utiliza el algoritmo K-Means, uno de los más usados de la ciencia de datos en general y para hacer clustering en particular [30]. El algoritmo K-Means es un sencillo método para

dividir, en base a una serie de variables, una población en un número k de segmentos (o clusters) determinado por el usuario.

El primer paso del algoritmo es la inicialización. Para ello se seleccionan k individuos de la población al azar. Esos individuos son los representantes de cada clúster y se llaman también centroides. Una vez hecho esto, el algoritmo repite un número de veces dos pasos: primero se realiza la asignación de individuos, donde cada individuo se asigna al clúster a cuyo centroide está más cerca. Para ello se mide la distancia entre él y cada centroide, normalmente usando la distancia euclídea. Esto genera una partición de los datos, posteriormente se procede a la actualización de los centroides en la que el nuevo representante del grupo se traslada al centro (media) de todos los individuos asignados a él. Estos pasos se repiten hasta que los grupos se mantienen iguales o, lo que es lo mismo, no hay nuevas reasignaciones de individuos, por lo que el algoritmo habrá convergido.

El algoritmo Mean-Shift [14] es un método iterativo no paramétrico que funciona encontrando las modas de unas distribuciones, pero sin necesitar saber cuántas modas se tienen. Considera que el espacio de datos es una función de densidad de probabilidad muestreada y para cada punto del conjunto de datos, encuentra la moda más cercana, para ello, define una región alrededor de ese punto y encuentra su media, cambiando la situación de la media actual a la nueva (shift). Repite el proceso hasta que los datos converjan. Una vez finalizado el proceso se observará el número de clusters en que se ha dividido la población.

Los llamados métodos jerárquicos, como el Ward Hierachical Clustering, tienen por objetivo agrupar clusters para formar uno nuevo, de tal forma que sucesivamente se va efectuando este proceso de aglomeración hasta que se minimice alguna distancia o bien se maximice alguna medida de similitud. El método de Ward [46] es un procedimiento jerárquico en el que, en cada etapa, se unen los dos clusters para los cuales se tenga el menor incremento en el valor total de la suma de los cuadrados de las diferencias, dentro de cada clúster, de cada individuo al centroide del clúster.

Otro algoritmo de clustering jerárquico es BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies). Este método [66] almacena para cada clúster un triplete de datos que contiene el número de objetos que pertenecen a ese grupo, el valor de la suma de todos los valores de los atributos de los objetos pertenecientes al grupo, y la suma de los cuadrados de los atributos de los objetos que pertenecen al clúster. Con esta información construye un árbol de grupos llamado CF-tree (Cluster Features tree).

	Cluster K-Means	Mean-Shift	Ward	Birch
Tiempo Seco	42.22 %	63.14 %	39.80 %	24.77 %
Lluvia	50.99 %	33.12 %	51.19 %	44.54 %
Tormenta	6.79 %	3.74 %	9.01 %	30.69 %

Tabla 4.6: Clasificación no supervisada BSM1_LT con 9 variables

	Cluster K-Means	Mean-Shift	Ward	BIRCH
Tiempo Seco	50.13 %	47.39 %	58.58 %	9.14 %
Lluvia	7.12 %	51.01 %	34.60 %	54.19 %
Tormenta	42.75 %	1.60 %	6.82 %	36.67 %

Tabla 4.7: Clasificación no supervisada BSM1_LT con 9 variables más Temp y «n»

El procedimiento del algoritmo BIRCH es el siguiente, primero se genera un CF-tree inicial, leyendo los datos y asignándolos a una rama o a otra. Si la distancia entre un objeto nuevo y los anteriores se hace mayor que cierto parámetro T, se crea una rama nueva. posteriormente se revisar el árbol creado para ver si es demasiado grande, y moldearlo modificando el valor del parámetro T, si el valor de este parámetro aumenta, las ramas del árbol se juntan al no haber distinción de grupos. Tras esto se aplicar algún procedimiento de clustering, como el K-Means, sobre los nodos de cada nivel y finalmente se redistribuyen los datos según los centroides encontrados en el paso anterior, logrando un mayor refinamiento en el agrupamiento.

Para realizar la clasificación no supervisada en BSM1_LT se utilizarán distintas combinaciones de variables teniendo en cuenta las selecciones realizadas con anterioridad y añadiendo además del índice de intensidad de precipitación, la temperatura, parámetro de gran importancia en BSM1_LT por abarcar un periodo de tiempo muy extenso.

Primero se realiza una clasificación aplicando todos los algoritmos con las variables Q, Si, Snd, Snh, Ss, Xbh, Xi, Xnd y Xs. Los porcentajes de clasificación de cada tiempo atmosférico se muestran en la Tabla 4.7, y sus representaciones gráficas en la Figura 4.9.

En una segunda prueba se amplían las variables añadiendo la Temperatura y la intensidad de precipitación. Los resultados se muestran en la Tabla 4.7

Se realizan de nuevo los cálculos con las variables caudal (Q), temperatura e índice de intensidad de precipitación (n), variando los parámetros de los algoritmos pero ninguno de los resultados obtenidos se considera satisfactorio.

Posteriormente se clasifican de nuevo los datos utilizando Q, Si, Ss, Xi, Temp y n, es decir las variables seleccionadas como las que mejor clasificaban en BSM1 a las

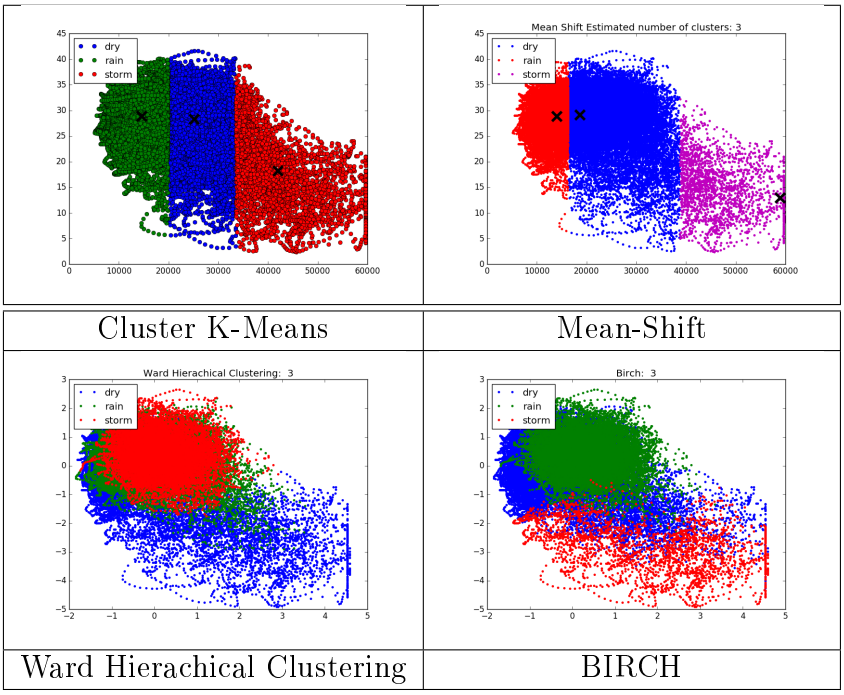

Figura 4.9: Clasificación no supervisada BSM1_LT

	Cluster K-Means (cluster = 3)	Mean-Shift (bandwidth = 2140)	Ward (n_neighbor =10)	BIRCH (clusters=3, T=600)
Tiempo Seco	52.73 %	57.10 %	54.89 %	83.73 %
Lluvia	9.19 %	39.49 %	17.94 %	14.89 %
Tormenta	38.08 %	3.41 %	27.17 %	1.38 %

Tabla 4.8: Clasificación no supervisada BSM1_LT con Q, Si, Ss, Xi, Temp y «n»

que se añaden la temperatura y el índice de intensidad de precipitación y se varían los parámetros de los algoritmos obteniendo los resultados de la Tabla 4.8.

El algoritmo BIRCH con 3 clusters y un valor de threshold (T) = 600 obtiene los porcentajes de clasificación de cada tiempo atmosférico que se muestran en la Tabla 4.9 y en la Figura 4.10. Se observa que la clasificación obtenida presenta errores inferiores al 1 %.

El índice de intensidad de precipitación ‹‹n›› que no mejoraba los resultados de clasificación en BSM1 si parece influenciar al trabajar con BSM1_LT, ya que este conjunto de datos además de ser muchísimo mayor se ve afectado por el efecto de la estacionalidad.

Para asegurase de que la clasificación obtenida es óptima se comprueba que los datos no se han clasificado de forma aleatoria, como se puede apreciar en la Figura 4.11, donde se muestran los valores en los distintos periodos de BSM1_LT, 63 días

	BIRCH	Esperado
Tiempo Seco	83.73 %	85 %
Lluvia	14.89 %	14 %
Tormenta	1.38 %	1 %

Tabla 4.9: Clasificación BIRCH (clusters=3, T=600)

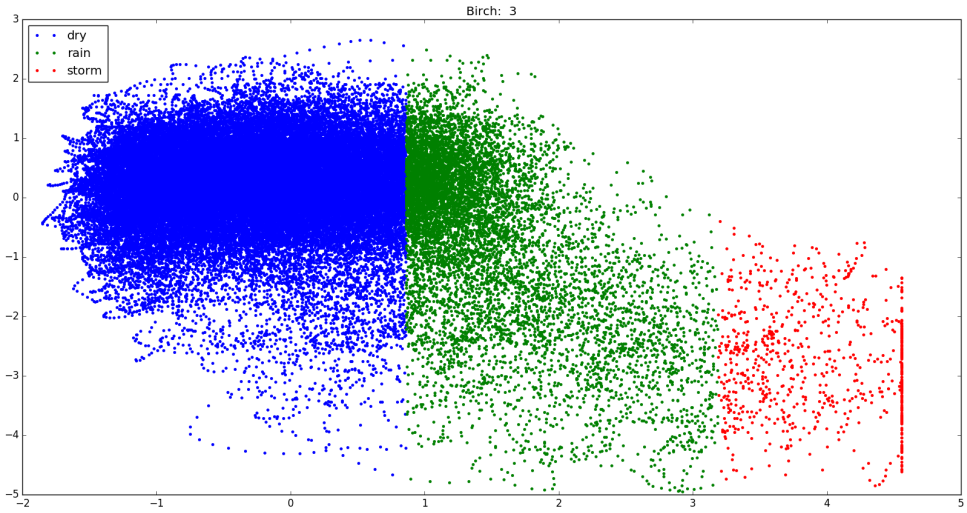

Figura 4.10: Clasificación BIRCH (clusters=3, T=600)

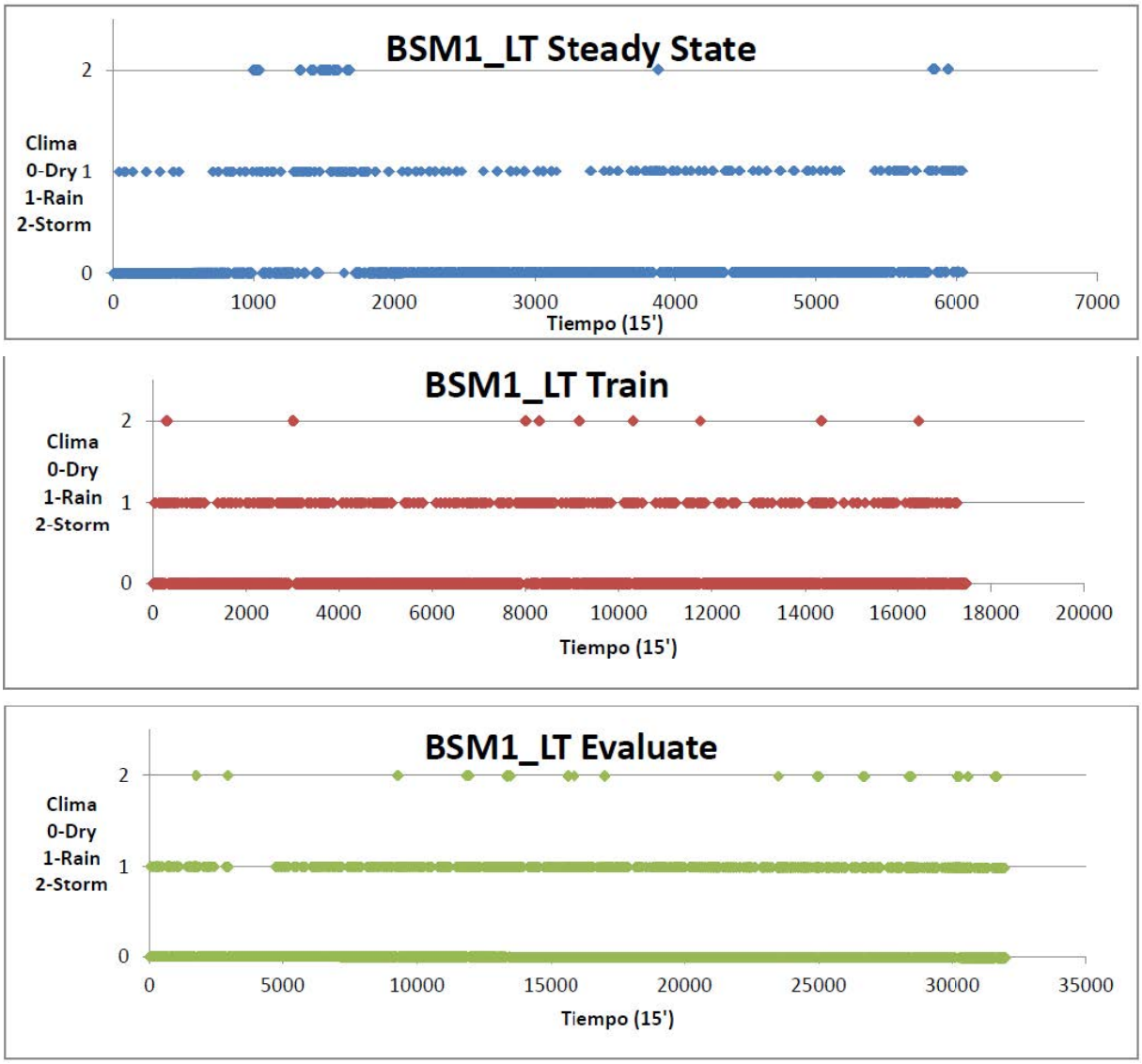

Figura 4.11: Clasificación BIRCH (clusters=3, T=600) por periodos

de estabilización, 182 días de entrenamiento y 364 días del periodo de evaluación.

4.5. Clasificación supervisada BSM1_LT

Para comprobar si la clasificación no supervisada realizada con el algoritmo BIRCH a los datos proporcionados por BSM1_LT se asemeja a la clasificación de los datos BSM1 se realiza una clasificación supervisada utilizando como variable de tiempo atmosférico la clasificación proporcionada por el algoritmo BIRCH, obteniendo los resultados que se muestran en la Tabla 4.10, consiguiendo validaciones por encima del 85 %, lo que parece verificar los resultados obtenidos previamente mediante el

	Validación Cruzada	5x2 CV
Gaussian Naive Bayes	0.80328	0.803
k-Nearest Neighbor	0.86249	0.880
Random Forest	0.85193	0.858
Decision Tree	0.85019	0.877

Tabla 4.10: BSM1_LT. Clasificación supervisada con validación

algoritmo BIRCH no supervisado.

4.6. Conclusión

Dada la importancia del tiempo atmosférico a la hora de gestionar el funcionamiento de los soplantes que aportan el oxígeno necesario para la eliminación del amonio en el control de las aguas residuales de una EDAR, resulta muy interesante conocer su valor como paso previo a la depuración.

Se ha comprobado que se puede realizar una clasificación no supervisada de los datos BSM1_LT con el algoritmo BIRCH con errores de clasificación inferiores al 1 %.

Capítulo 5

Estudio del Nitrógeno

5.1. Introducción

Una vez realizada la clasificación del tiempo atmosférico interesa conocer los valores de nitrógeno presentes en el efluente vertido para ajustar de manera más precisa la cantidad de oxigeno que es necesario aportar para su eliminación.

El nitrógeno es un contaminante presente en las aguas residuales que debe ser eliminado con anterioridad al vertido de éstas en los cursos de aguas, porque el exceso de nitrógeno reduce el oxígeno disuelto de las aguas superficiales, es tóxico para el ecosistema acuático, entraña un riesgo para la salud pública y junto al fósforo son responsables del crecimiento desmesurado de organismos fotosintéticos (eutrofización). Todos estos factores hacen que la legislación sea cada vez más restrictiva en cuanto a los límites máximos permitidos para este parámetro.

La forma más común empleada para la eliminación del nitrógeno se basa en un doble proceso biológico de nitrificación y denitrificación. En la primera etapa, la de nitrificación, el amonio es convertido primero en nitrito y éste, a su vez, en nitrato, mediante el uso de bacterias nitrificadoras que utilizan carbono inorgánico como fuente de carbono y obtienen la energía necesaria para su crecimiento de las reacciones químicas de la nitrificación. La segunda etapa, la de denitrificación, consiste en la conversión del nitrato en nitrógeno gas, el cual se libera a la atmósfera. Esta conversión la llevan a cabo unas bacterias en condiciones anaerobias, las cuales utilizan el nitrato como aceptor final de electrones y la materia orgánica presente en el agua como fuente de carbono.

Para poder estudiar si los valores de nitrógeno vertidos se adaptan a la legislación mediante una clasificación es necesario realizar previamente una simulación con los datos de entrada proporcionados por BSM1_LT. La simulación se realizará en Modelica, un lenguaje de modelado libre, orientado a objetos y aplicable a múltiples

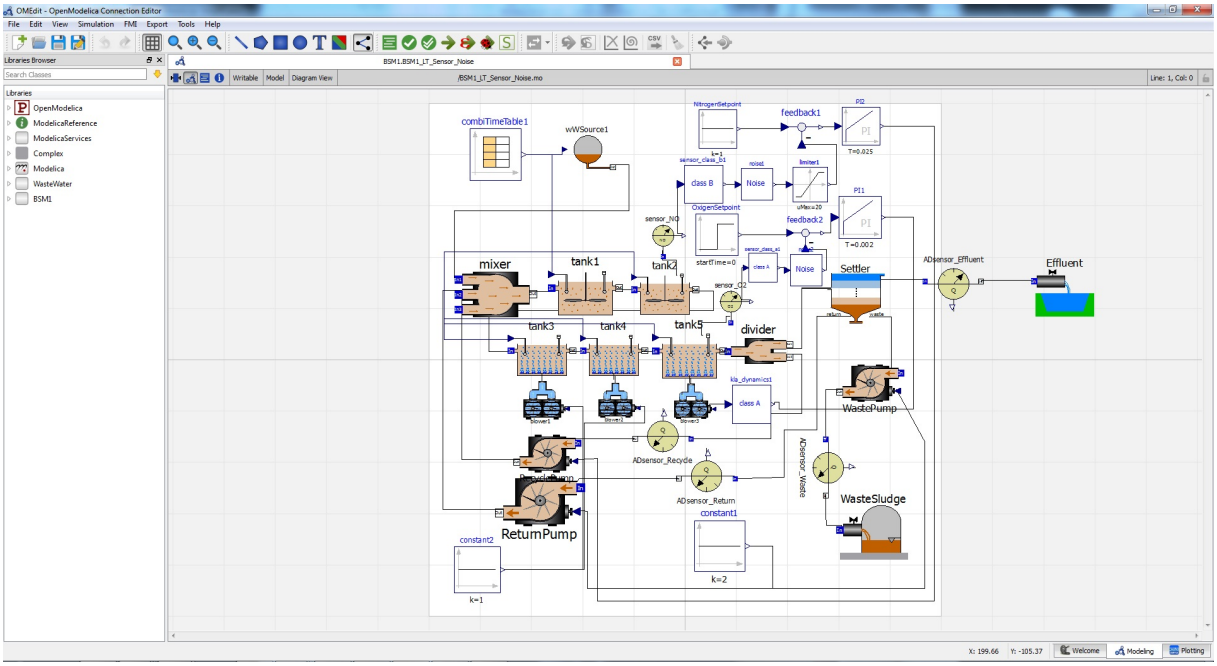

Figura 5.1: Simulación de planta con bucle cerrado, sensores y ruido

campos de la ingeniería. Modelica contiene herramientas para definir las relaciones entre los distintos componentes y las particularidades de cada componente en sí, y está dotado de una biblioteca de aguas residuales, WasteWater, que proporciona los módulos necesarios para simular una planta con las características descritas en el benchmark BSM1.

El comportamiento del proceso de lodos activados se reproduce en un simulador que actúa como una plataforma de pruebas en tiempo real y que también se utiliza para la identificación de los modelos de entrada-salida de varias variables para el controlador predictivo.

La simulación se realiza utilizando OpenModelica, un entorno de simulación y modelado, de código abierto, multiplataforma y destinado tanto a uso industrial como académico.

5.2. Simulación

La planta creada en OpenModelica para la simulación sigue las especificaciones descritas en el benchmark BSM1, basadas en ASM1 para definir las reacciones biológicas. Se implementan estrategias de control basadas en bucles cerrados, tal como se muestra en la Figura 5.1.

Los objetivos de control primarios son para mantener la concentración de Nitrato en el segundo tanque con un valor de ajuste predeterminado de 1 g m-3 y la concentración de oxígeno disuelto en el quinto tanque a un valor de ajuste predeterminado de 2 $g.(COD)m^{-3}$. Para ello necesitaremos modelar varios sensores. Para realizar el control se utilizarán controladores de tipo PI (controlador de acción Proporcional e Integral) , cuya función de transferencia responde a la ecuación:

$$Gs(s) = \frac{Y(s)}{E(s)} = K_P(\frac{1}{T_S + 1}) \tag{5.1}$$

con los valores de PI Nitrogeno, PI2(k = 0.10842, T = 0.025) y PI Oxigeno, PI1(k = 0.1042, T = 0.002).

Para el control del oxígeno en el tanque 5 a un valor de ajuste predeterminado de 2 $g.(-COD).m^{-3}$, necesitamos un sensor de clase A, con un rango de medida de 0 a $10g.(COD).m^{-3}$ y un ruido de medición de 0,25 $g.(COD).m^{-3}$. La variable que se manipula mediante este sensor es el coeficiente de transferencia de oxígeno, Kla (5). El sensor clase A tendrá un tiempo de respuesta (T90) de 1 minuto y una orden del sistema de n = 2. Esto se traduce en una relación de T10 (tiempo de retardo) para T90 (tiempo de respuesta) $R_{T10/T90} = 0{,}133$. Así, el retardo de transporte es sólo una pequeña fracción del tiempo de respuesta. La función de transferencia será:

$$Gs(s) = \frac{1}{1 + T_S} * \frac{1}{1 + T_S} \tag{5.2}$$

Con un valor de $T = T90/3{,}89$

El control del sistema de aireación (KLa1-KLa5) se realiza con un sensor de clase A donde se considera que un tiempo de respuesta de T90 = 4 minutos. La constante de tiempo de cada retardo es $T = T90/3,89 = 1,03min.$

La medición de Nitrato en el tanque 2 se realiza mediante un sensor de clase B con un rango de medida de 0 a 20 $g.N.m^{-3}$. El ruido de la medición es igual a 0,5 $g.N.m^{-3}$. La variable manipulada es el caudal de reciclado interno que vuelve desde el tanque 5 al tanque 1. Para los sensores de clases B se supone un orden del sistema de n = 8. Esto conducirá a una relación del tiempo de retardo para el tiempo de respuesta de $R_{T10/T90} = 0,392$. En este caso el tiempo de retardo ya es de aproximadamente 40 % del tiempo de respuesta, esto supone tener en cuenta el efecto significativo del transporte de la muestra. Para la el sensor de clase B se utilizará un tiempo de respuesta de 10 minutos. La función de transferencia será:

$$Gs(s) = \frac{1}{1 + T_S} * \frac{1}{1 + T_S} * \frac{1}{1 + T_S} * \frac{1}{1 + T_S} * \frac{1}{1 + T_S} * \frac{1}{1 + T_S} * \frac{1}{1 + T_S} * \frac{1}{1 + T_S} \tag{5.3}$$

Con un valor de $T = T90/11{,}7724$

Las señales de medición reales siempre incluyen la medición de ruido que puede llevar a acciones de control no deseados o reducir la velocidad de la reacción. Por lo tanto el ruido ha de estar incluido en el modelo de sensor. La elección de una señal de ruido aleatoria habría requerido ejecutar cada simulación de referencia un gran número de veces con el fin de eliminar la influencia de la señal aleatoria, por lo que la señal de ruido se elige con una desviación estándar de 1, que se multiplica con el nivel de ruido definido (2,5 % del valor de medición máximo). Para las simulaciones se utilizarán los valores:

- Sensor de oxígeno Clase A, el rango de medición: 0-10 $g.(COD).m^{-3}$ y el ruido de medición = 0,25$g.(COD).m^{-3}$.
- Sensor de nitrato: Clase B, con un rango de medición de 0-20 $g.N.m^{-3}$ y medición del ruido = 0,5 $g.N.m^{-3}$.
- Noise: noise_sampling=15/60/24.

5.3. Clasificación supervisada

Al realizar cualquier clasificación normalmente no se tiene en cuenta el aspecto temporal de los datos sino el conjunto de ellos, esto implica que se pueden utilizar unos datos temporalmente posteriores para predecir los valores anteriores.

Esto no es aplicable al funcionamiento real de una EDAR ya que los datos sobre los que se ha de predecir son siempre datos pasados.

Para que el estudio del comportamiento del nitrógeno vertido sea lo más parecido posible a un funcionamiento real se realiza una clasificación teniendo en cuenta solo los valores pasados y no los futuros. Para ello se parte de una ventana temporal con los datos de una semana del valor del Shh_out, que es el amonio de salida tras simular la planta con OpenModelica y la variable Snh_Ok, variable binaria que toma el valor 0 si Snh_out $<$ 4 (válido) y 1 si Snh_out $>=$ 4.

Con estos datos se crea un conjunto de entrenamiento con el que se entrena el algoritmo de clasificación y se utiliza como test el día siguiente. Es decir, se comienza trabajando con ocho días, siete de entrenamiento y uno para la comprobación.

A continuación el valor predicho se añade al conjunto de entrenamiento y se utiliza como test el día posterior. De esta forma se va ampliando la ventana hasta terminar con un conjunto de entrenamiento formado por 611 días que predecirá el último día del conjunto de datos BSM1_LT (Figura 5.2).

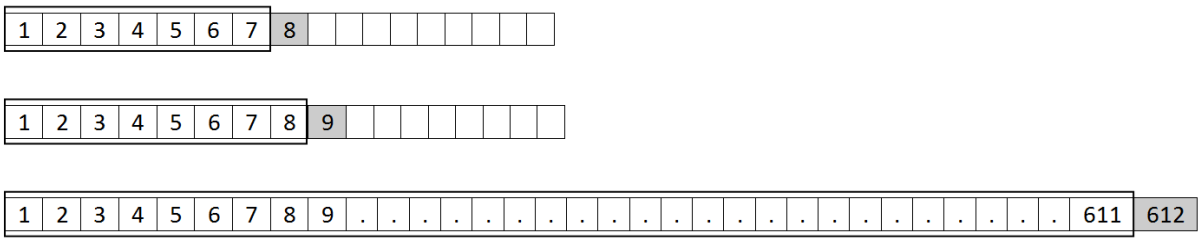

Figura 5.2: Ventanas de predicción

5.3.1. Clasificación y validación.

Se clasifica con los mismos algoritmos utilizados para realizar las clasificación del tiempo atmosférico, Naive Bayes, k-NN, Classification Trees y Random Forest, y para contrastar su validez se utiliza los valores del Error Cuadrático Medio (RSE), la Desviación Estándar (RMSE) y la Precisión.

El Error Cuadrático Medio de un estimador mide el promedio de los errores al cuadrado, es decir, la diferencia entre el estimador y lo que se estima. El RSE es una función de riesgo, correspondiente al valor esperado de la pérdida del error al cuadrado o pérdida cuadrática. La diferencia se produce debido a la aleatoriedad o porque el estimador no tiene en cuenta la información que podría producir una estimación más precisa.

El valor RMSE estima la magnitud global del error de cada modelo. Interesa conseguir valores de RMSE cercanos a cero ya que eso significa que las predicciones del modelo son buenas.

Precisión se refiere a la dispersión del conjunto de valores obtenidos de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor la precisión. La precisión es la cercanía de agrupación entre un grupo de resultados y se define como la proporción de verdaderos positivos contra todos los resultados positivos (tanto verdaderos positivos, como falsos positivos), e interesa un valor lo más cercano a 1 posible [55].

Los resultados se ven en la Tabla 5.1 y 5.2 y el árbol de decisión, basado en el índice de Gini, de profundidad 6 con la clasificación para el nitrógeno vertido se presenta en la Figura 5.3.

La Desviación Estándar de la clasificación obtenida con el Árbol de Decisión de profundidad 6 alcanza un valor RMSE del límite Snh admisible de 0.239732 y una Precisión de 0.9425.

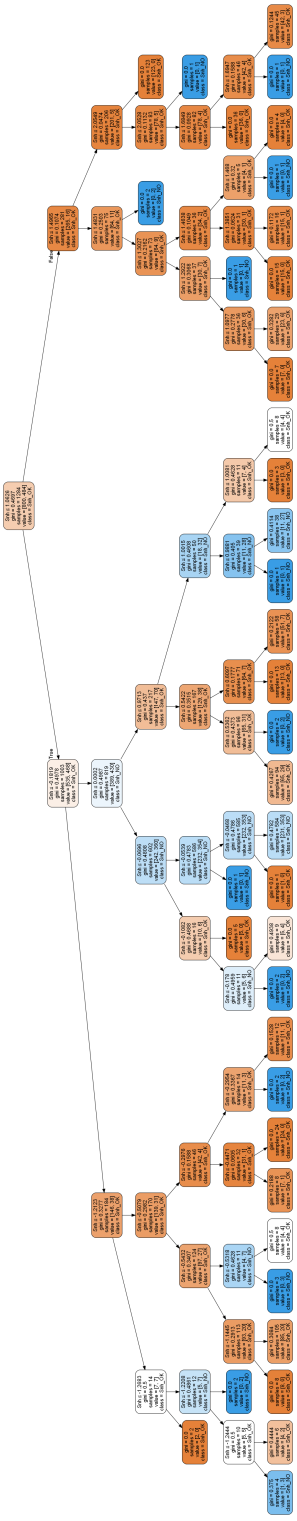

Figura 5.3: Árbol de Decisión de profundidad 6 para Snh

	MSE	RMSE	Precisión
Gaussian Naive Bayes	0.436782	0.66089	0.5632
k-Nearest Neighbor (n_neighbor = 5)	0.3390	0.114934	0.885057
Random Forest (n_estimators = 20)	0.35548	0.126437	0.89655
Decision Tree (max_depth = none)	0.17241	0.4152	0.8276

Tabla 5.1: Clasificación Snh

	MSE	RMSE	Precisión
Decision Tree (max_depth = none)	0.17241	0.4152	0.8276
Decision Tree (max_depth = 6)	**0.05747**	**0.239732**	**0.9425**
Decision Tree (max_depth = 4)	0.08046	0.283654	0.91954
Decision Tree (max_depth = 10)	0.08046	0.283654	0.91954
Decision Tree (max_depth = 13)	0.149425	0.38656	0.8506

Tabla 5.2: Clasificación Snh con Árboles de Decisión

5.3.2. Clasificación con Redes Neuronales.

Como estudio adicional se realiza una nueva clasificación utilizando Redes Neuronales para ver si se pueden conseguir mejores clasificaciones.

Las Redes Neuronales Artificiales imitan a las redes neuronales reales en el desarrollo de tareas de aprendizaje [31].

Una RNA está implementada mediante un Perceptrón Multicapa (MLP) . Un MLP es una red de neuronas simples llamados perceptrones. El perceptrón calcula una sola salida de múltiples entradas de valor real mediante la formación de una combinación lineal de acuerdo con sus pesos de entrada y, a continuación, produce una salida a partir de la aplicación de una función umbral a la media ponderada. Si se conectan las salidas de unas neuronas como entradas de otras se obtiene una red neuronal.

Uno de los ejemplos más típicos de red neuronal es el la Back Propagation Neural Network. Consta de una capa de entrada con tantos neuronas como variables de entrada se vayan a introducir en el modelo, una capa oculta que realiza la mayor parte del cálculo y una capa de salida con tantas neuronas como posibles clases existan.

Se realiza la clasificación con ventanas temporales de manera similar a la utilizada en la clasificación supervisada. La Red tendrá una neurona de entrada, el nitrógeno contenido en el efluente y una neurona de salida que indica si la cantidad de nitrógeno se adecua a los límites exigidos o no.

Se realizan pruebas con distintas RNAs y se contrasta su validez utilizando los mismos estimadores anteriores, como se muestra en la Tabla 5.3.

RNA	MSE	RMSE	Precisión
[1, (2), 1]	0.02443	0.1563	0.93571
[1, (3), 1]	**0.011492**	**0.1072**	**0.98850**
[1, (4), 1]	0.08105	0.2847	0.91628
[1, (5), 1]	0.14010	0.3743	0.88742

Tabla 5.3: Clasificación Snh mediante RNA

La clasificación obtenida con la RNA [1, (3), 1] formada por una capa de entrada con 1 neuronas, una capa ocultas con 3 neuronas y una capa de salida con una neurona, consigue una precisión suprior al 98 % lo que representa un excelente resultado. Esta RNA se representa gráficamente en la Figura 5.4.

5.4. Predicción

Hasta el momento se han realizado clasificaciones del valor del nitrógeno efluente en función de si se encontraba o no dentro de los límites admitidos en la legislación de vertidos. Un análisis adicional consiste en predecir el valor que tendrá el amonio de salida y no solo si es admisible, para poder ajustar de manera más eficiente el oxigeno suministrado por las soplantes.

En 1970, Box y Jenkins desarrollaron un modelo estadístico de series temporales en el que se tiene en cuenta la interacción entre los datos porque cada uno de ellos se presenta como una función de los valores anteriores [42]. Estos modelos ARIMA permiten describir cada valor como una función lineal de los datos anteriores, errores debidos al azar e incluso un componente estacional.

La técnica ARIMA busca modelar la tendencia de los datos a lo largo del tiempo para extrapolarla posteriormente en el futuro y poder tomar decisiones sobre tendencias reales.

La metodología ARIMA se compone de cuatro fases [11]:

- Primero se identifica un posible modelo ARIMA, para ello hay que transformar la serie eliminando la estacionalidad y la tendencia.

- En segundo lugar se estiman los coeficientes del modelo “p” , “d” y “q” utilizando las Funciones de Autocorrelación (ACF) y Autocorrelación Parcial (PACF) . Para aplicar un modelo ARIMA a la serie temporal, es necesario encontrar los valores óptimos para los tres parámetros del modelo (p, d, q):

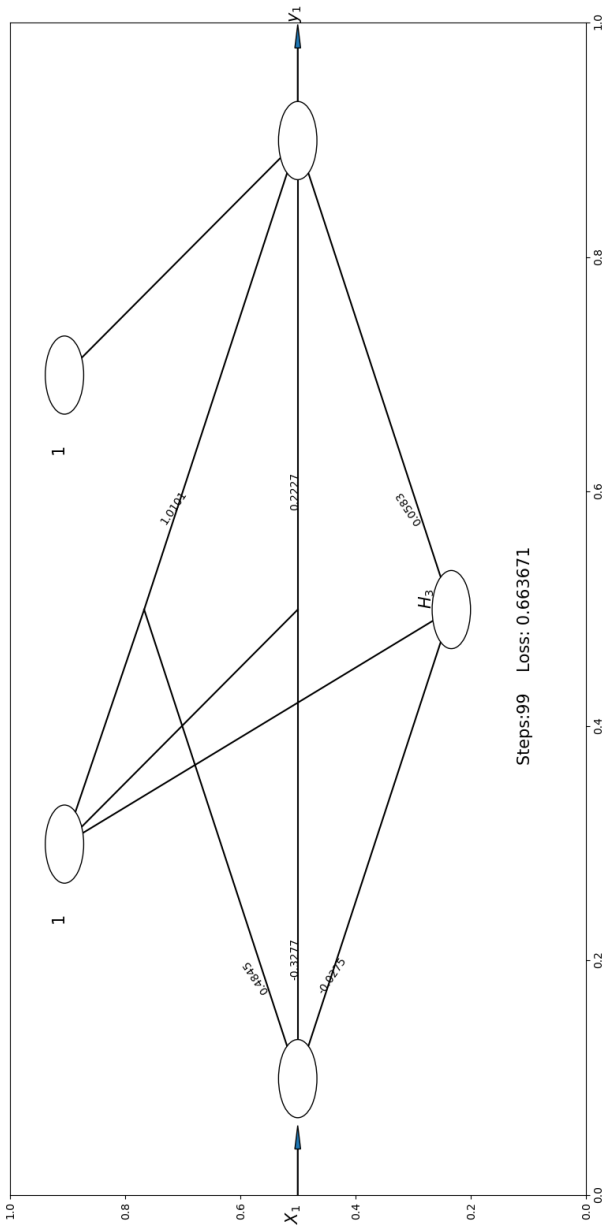

Figura 5.4: RNA [1, (3), 1]

- El número de términos autorregresivos (AR) (p): el término "p" representa el número de retrasos de la serie temporal.
- El número de términos de media móvil (MA) (q): el término "q" es el número de errores de pronóstico en la ecuación de predicción.
- El número de diferencias (d): "d" es la diferencia utilizada para convertir la serie en estacionaria.

Suele ser conveniente seleccionar varios modelos para ser estimados y contrastados posteriormente.

- En la tercera fase se utilizan distintos procedimientos para contrastar la validez de los modelos seleccionados.
- Finalmente, una vez seleccionado el modelo ARIMA adecuado se realiza la predicción.

Para poder realizar las predicciones es necesario transformar los datos BSM1_LT para que incluyan una variable que muestre el tiempo de la forma d-mm-aaaa hh:mm:ss.

La representación temporal de los datos del amonio de salida, como se muestra en la Figura 5.5, revela como la serie presenta una tendencia, lo que significa que no es estacionaria por lo que habrá que eliminar los factores de tendencia y estacionalidad del modelo. En la Figura 5.6 se representan estos factores que han de ser corregidos para poder aplicar las funciones ACF y PACF que ayuden a estimar los parámetros para definir el modelo ARIMA.

La forma más común de eliminar tanto la tendencia como la estacionalidad es calculando las diferencias entre observaciones consecutivas de la serie temporal, esto normalmente confirma la estacionalidad de las series temporales.

Después de aplicar ACF y PACF se procede a buscar los parámetros que que minimicen la Desviación Estándar del modelo ya que esto significa que las predicciones del modelo serán mejores al estar más cerca de los valores reales. Los resultados se muestran en la Tabla 5.5.

ARIMA realizará un pronóstico basándose en el conjunto de datos con los que se entrena el modelo, luego comparará los resultados de sus predicciones con los valores reales de los datos. ARIMA realiza un pronóstico progresivo, creando un nuevo modelo ARIMA después de recibir cada nueva predicción, y mantendrá un seguimiento de todas las observaciones en una lista creada con los datos de entrenamiento y a la que se vana añadiendo nuevas predicciones en cada iteración.

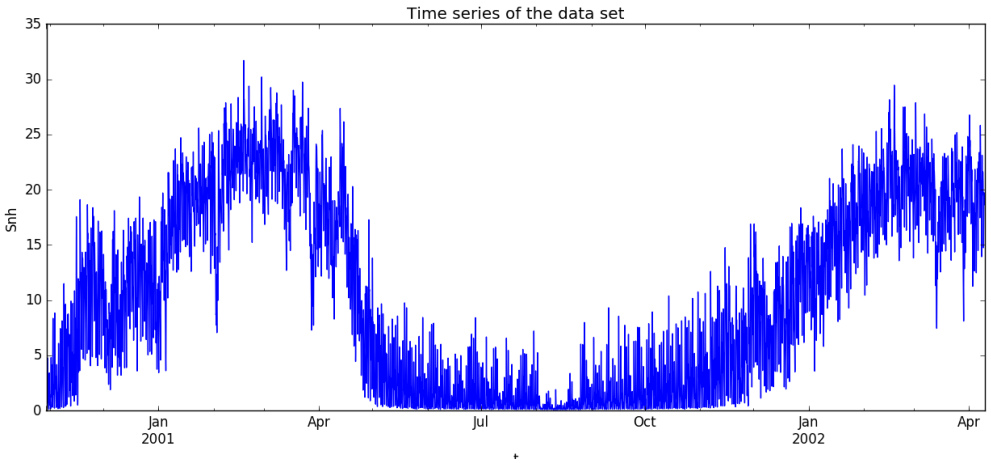

Figura 5.5: Representación temporal del Amonio de salida.

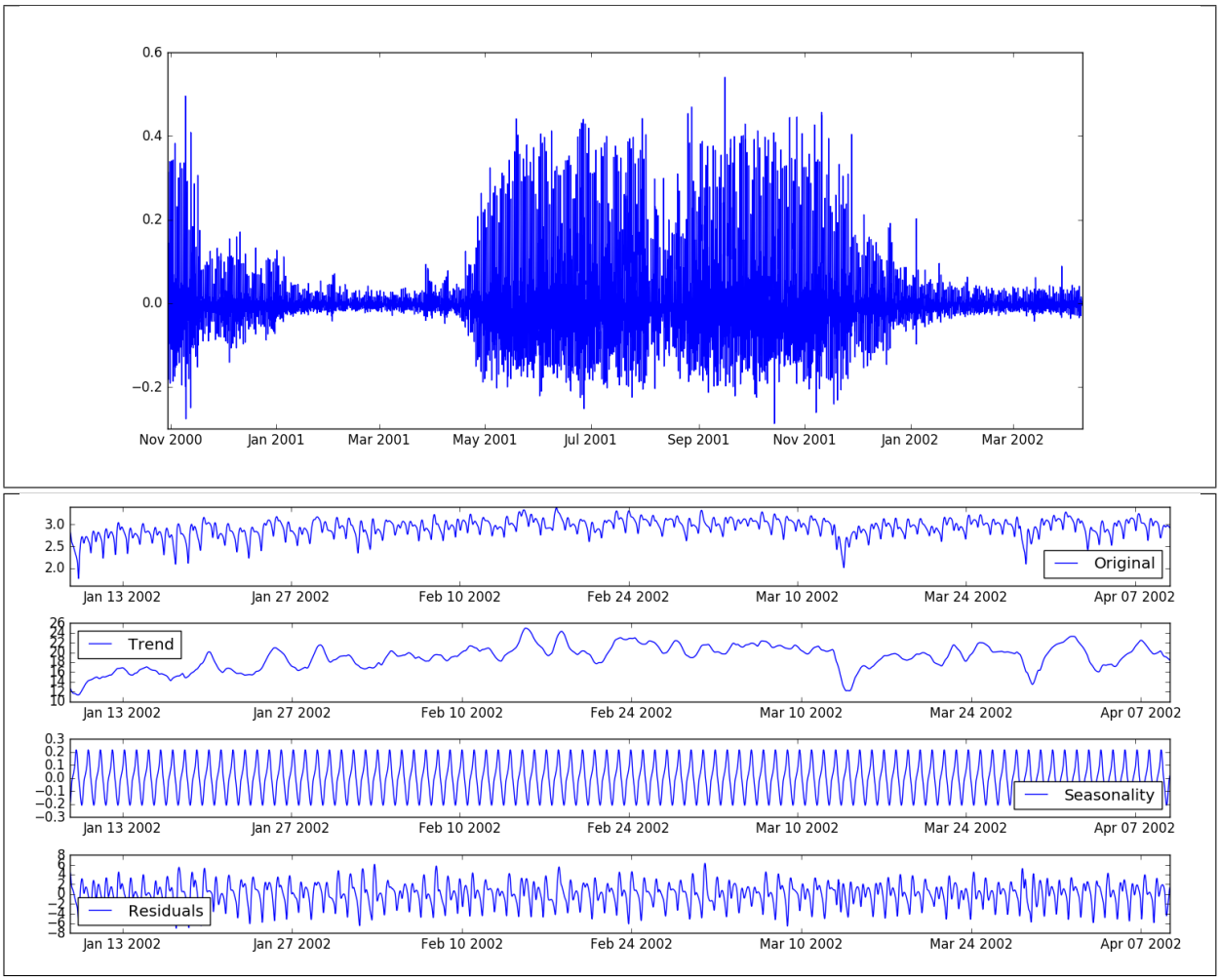

Figura 5.6: Descomposición de la tendencia, estacionalidad y residuos

Modelo ARIMA	RMSE
ARIMA(0,1,1)	9.6715
ARIMA(1,1,1)	**1.4838**
ARIMA(1,1,2)	1.5864
ARIMA(1,1,3)	1.6795
ARIMA(2,1,1)	2.7920
ARIMA(2,1,2)	2.5456
ARIMA(2,1,3)	2.3859

Tabla 5.4: ARIMA búsqueda de parámetros óptimos.

Errores	Número	Porcentaje
Falso negativo	56	0.166058773 %
Falso positivo	19	0.056341369 %
Total errores	75	0.222400142 %

Tabla 5.5: Errores de predicción

Para validar el modelo se comparan las predicciones con los valores reales calculando la Desviación Estándar..

El modelo ARIMA(1,1,1) tiene una desviación estándar de 1.4838. En la Figura 5.7 se puede apreciar que los valores de RMSE están muy cercanos a cero lo que implica que las predicciones del modelo son precisas, por lo que se realizan las predicciones que obtienen un Error Cuadrático Medio MSE de 0.000554 y se representan en la Figura 5.8, donde se observa que las predicciones están en la escala adecuada y mantienen la tendencia de los valores originales.

Se realiza una comprobación adicional estudiando si los valores de salida del amonio predichos se encuentran dentro de los límites admisibles (Snh <4) (Tabla 5.5) y se calcula el error para compararlos con los errores de clasificación del nitrógeno obtenidos anteriormente con los Árboles de Decisión y las RNAs. El valor RMSE del límite Snh admisible es de 0.3074, inferior a los obtenidos previamente (Tabla 5.6) y su Precisión 0.9966, más alta que con los métodos anteriores.

Método	MSE	RMSE	Precisión
Árboles de Decisión (depth = 6)	0.0575	0.2397	0.9425
RNA [1,(3),1]	0.0115	0.1072	0.98850
ARIMA (1,1,1)	0.0945	0.3074	0.9966

Tabla 5.6: Comparación de Errores

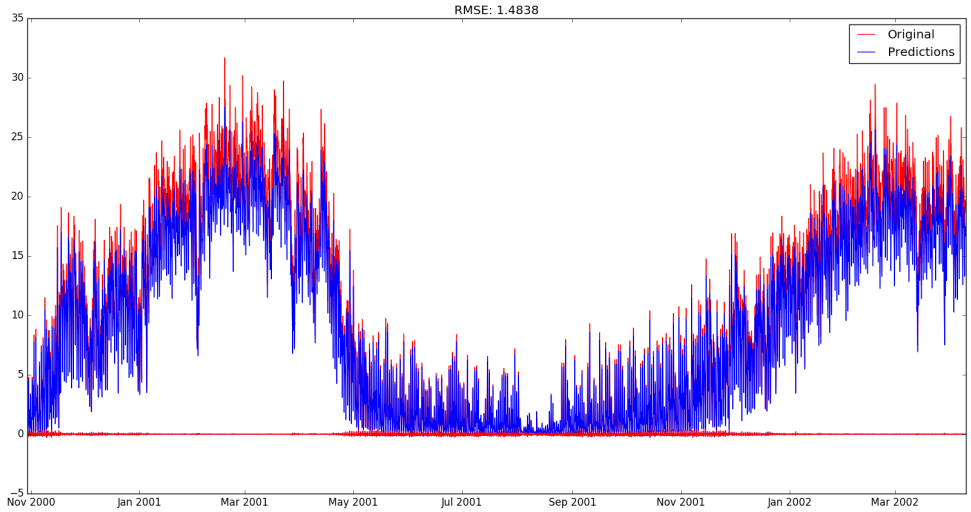

Figura 5.7: RMSE ARIMA(1,1,1)

5.5. Conclusión

Con el objetivo de conseguir determinar los valores del amonio presentes en el vertido de la EDAR se han realizado varias clasificaciones supervisadas entre las que destacan los Árboles de Decisión con Precisión del 94 % y Redes Neuronales Artificiales, consiguiendo unos valores de Precisión superiores al 98 %, lo que representan unos muy buenos resultados.

Adicionalmente se formulan predicciones utilizando el modelo ARIMA y se observa que los valores predichos no solo se ajustan de forma muy precisa a los valores reales sin no que además el número de valores que pueden provocar error en el límite admisible de los vertidos es inferior al 0.25 %.

Comparando los Errores Cuadráticos y las Precisiones de los tres métodos se aprecia que las predicciones realizadas con ARIMA presentan un error menor, por lo que parece el método más adecuado.

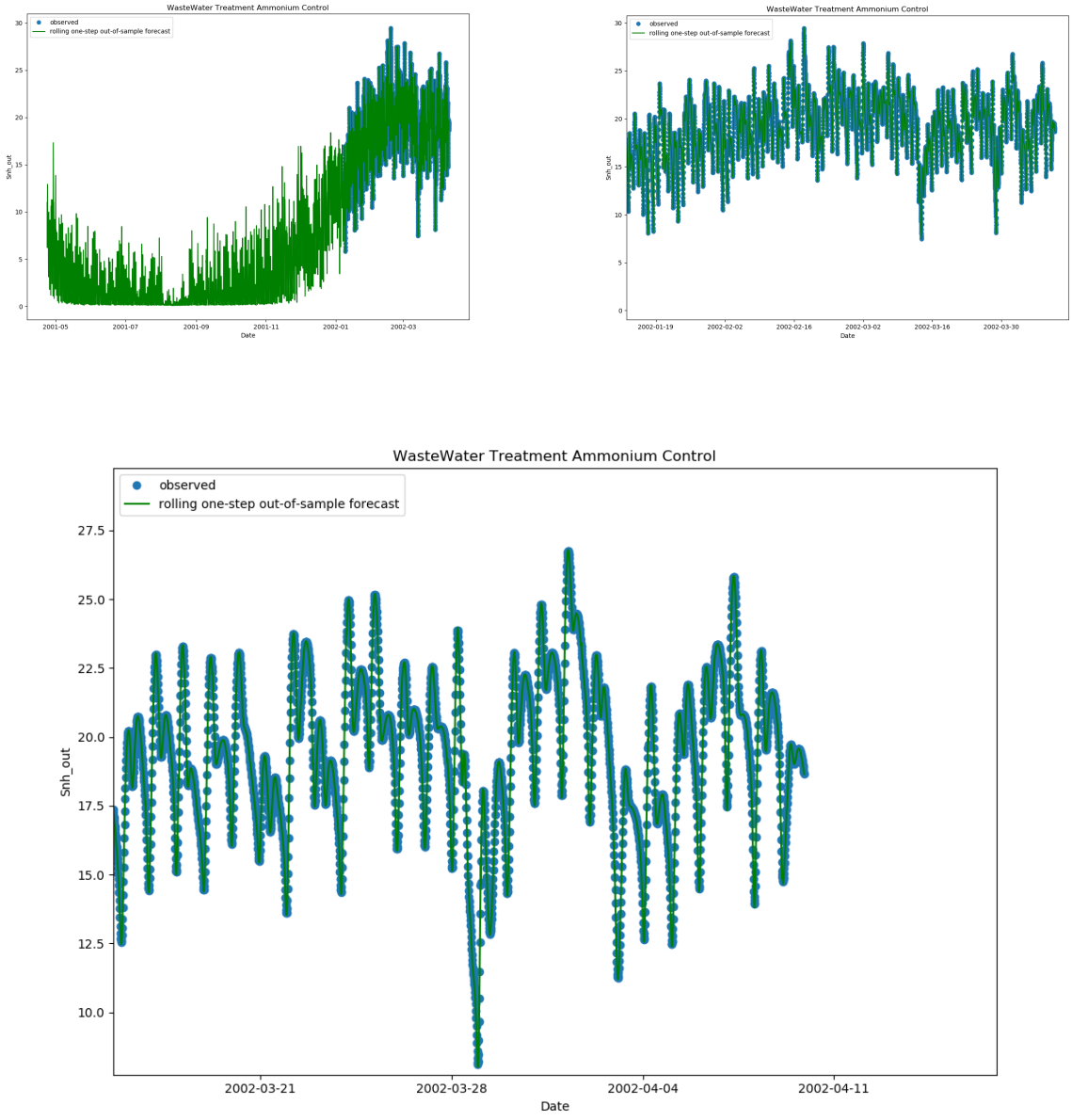

Figura 5.8: Predicciones ARIMA(1,1,1)

Capítulo 6

Conclusiones y trabajos futuros

6.1. Conclusión

La depuración de aguas residuales es un tema prioritario en la gestión medioambiental y conseguir que el amonio de salida se encuentre dentro de los límites admisibles es uno de los factores a controlar de mayor relevancia.

Para intentar solucionar el conflicto entre las aguas limpias y el coste que esto conlleva se han realizado múltiples investigaciones con el objetivo de optimizar este proceso en términos de eficiencia energética, ya que el coste de la energía debido a los equipos de aireación que realiza el control del oxígeno en los tanques aeróbicos, supone más del 70 % del consumo energético de una EDAR y alrededor del 25 % de los costes totales.

El avance a pasos agigantados de la Inteligencia Artificial y la integración en la vida diaria con aplicaciones en labores cotidianas lleva a pensar que utilizar el Aprendizaje Automático en el control de una Estación Depuradora de Aguas Residuales conseguirá un beneficio importante para el conjunto de la población.

Dentro del Aprendizaje Automático, encontramos técnicas de clasificación que nos permiten agrupar los datos de acuerdo a criterios o métodos, estas técnicas son la clasificación supervisada y la no supervisada.

La clasificación supervisada cuenta con un conocimiento a priori, es decir para la tarea de clasificar un objeto dentro de una categoría o clase se cuenta con modelos ya clasificados (objetos agrupados que tienen características comunes). El modelo BSM1 cuenta con una clase que muestra el tiempo atmosférico. Se ha comprobado que este dato es importante a la hora de determinar la cantidad de oxígeno que es necesario aportar a la planta depuradora para llevar a cabo los procesos de nitrificación en los tanques aeróbicos, ya que mantener bajo control la cantidad de oxígeno suministrada

es fundamental porque representa el coste más importante de todo el proceso de depuración de aguas.

La selección de atributos nos permite escoger las características que mejor clasifican el tiempo climático, las variables seleccionadas son Q, Si, Ss y Xi, proporcionadas por BSM1, a las que se añade el índice de intensidad de precipitación *«n»*.

Se ha realizado la clasificación supervisada del tiempo atmosférico en BSM1 utilizando los algoritmos Naive Bayes, k-NN, Classification Trees, Random Forest y SVM y se han validado los resultados con dos técnicas diferentes, Validación Cruzada y aplicando el el algoritmo de Dietterich 5x2 Cross-Validation Paired t-Test .

Los resultados obtenidos muestran valores superiores al 90 % en todos los algoritmos alcanzando valores en torno al 98 % al validar utilizando la técnica 5X2 CV Paired test.

Aplicando la transformada de Fourier a los datos para convertir los valores dependientes del tiempo en valores dependientes de la frecuencia, se generan árboles de decisión con una profundidad de 7, mucho menor que los árboles de profundidad 13 obtenidos sin realizar la transformación, y una validación del 98.4 %.

Se intentó extender la clasificación supervisada obtenida con BSM1 a BSM1_LT, ya que este último modelo no cuenta con una variable que clasifique el clima, pero la diferencia en los rangos de las variables imposibilitó esta opción por lo que se realiza una clasificación no supervisada en el modelo BSM1_LT.

En la clasificación no supervisada, a diferencia de la supervisada, no contamos con conocimiento a priori, por lo que tendremos un área de entrenamiento disponible para la tarea de clasificación. En este tipo de clasificación contamos con muestras que tiene un conjunto de características, de las que no sabemos a que clase o categoría pertenece, entonces la finalidad es el descubrimiento de grupos de "objetos" cuyas características afines nos permitan separar las diferentes clases.

En el conjunto de datos proporcionado por BSM1_LT se seleccionan las variables Q, Si, Ss , Xi y la Temperatura a los que se añade también el índice de intensidad de precipitación.

Se utilizan los algoritmos Cluster K-Means, Mean-Shift, Ward Hierachical Clustering y BIRCH para realizar la clasificación.

El algoritmo BIRCH con 3 clusters y un valor de threshold (T) = 600 obtiene unos porcentajes de clasificación de cada tiempo atmosférico con errores inferiores al 1 %.

Como forma de comprobar que la clasificación no supervisada realizada con el algoritmo BIRCH a los datos proporcionados por BSM1_LT se asemeja a la clasi-

ficación de los datos BSM1 se realiza una clasificación supervisada utilizando como variable de tiempo atmosférico la clasificación proporcionada por el algoritmo BIRCH obteniendo, mediante la técnica 5X2 CV Paired test un resultado de validación del 87.7 % en los Árboles de Decisión.

Con el fin de estudiar si los valores de nitrógeno vertidos se adaptan a la legislación, mediante una clasificación, es necesario realizar previamente una simulación con los datos de entrada proporcionados por BSM1_LT. La simulación se implementará en lenguaje Modelica utilizando la biblioteca de aguas residuales, WasteWater, que proporciona los módulos necesarios para simular una planta con las características descritas en el benchmark BSM1.

Se clasifican los valores de nitrógeno vertidos con los mismos algoritmos utilizados con el tiempo atmosférico, Naive Bayes, k-NN, Classification Trees y Random Forest, utilizando un método de ventanas temporales para crear el conjunto de entrenamiento, mediante los datos previamente conocidos, es decir no se usaran valores futuros para predecir valores pasados. Los resultados obtienen una precisión del 94 %, se realizan pruebas adicionales con Redes Neuronales Artificiales para intentar mejorar estos valores.

La clasificación obtenida con la RNA [1, (3), 1] formada por una capa de entrada con 1 neuronas, una capa ocultas con 3 neuronas respectivamente y una capa de salida con una neurona, obtiene una Precisión superior al 98 % lo que supone una mejora en la clasificación.

Una vez clasificado el nitrógeno contenido en las aguas de vertido resulta interesante realizar una predicción de los valores para poder actuar con antelación a la hora de modificar la cantidad de oxígeno que deben de suministrar las soplantes a los tanques aeróbicos. Para ello se utiliza la técnica ARIMA, que busca modelar la tendencia de los datos a lo largo del tiempo para extrapolarla posteriormente en el futuro y poder tomar decisiones sobre tendencias reales.

ARIMA realiza un pronóstico basándose en el conjunto de datos con los que se entrena el modelo y luego compara los resultados de sus predicciones con los valores reales de los datos.

El modelo ARIMA(1,1,1) tiene una desviación estándar de 1.4838 y a las predicciones obtenidas corresponde un error cuadrático medio de 0.000554, valores que están muy cercanos a cero lo que implica que las predicciones del modelo son precisas.

También se observa que las predicciones están en la escala adecuada y mantienen la tendencia de los valores originales.

Finalmente se comprueba que los valores predichos no solo se ajustan de forma muy precisa a los valores reales sin no que además el número de valores que pueden provocar error a la hora de determinar si el nitrógeno se encuentra dentro del límite admisible de los vertidos es inferior al 0.25 %, y se comparan las Precisiones de los distintos métodos.

Se ha visto que las técnicas utilizadas para clasificar y predecir los valores de nitrógeno amoniacal presentes en los vertidos de las aguas residuales depuradas alcanzan unos valores de precisión muy elevados, llegando en el caso del método ARIMA a ser superior al 99 %, lo que significa una importante ayuda a la hora de determinar la cantidad de oxígeno a suministrar a la planta depuradora. Esto puede suponer un importante ahorro energético, que facilitará alcanzar el equilibrio buscado entre coste de operación y la adecuación ambiental.

6.2. Trabajos futuros

Una linea interesante de trabajos futuros sería la aplicación de estas técnicas de AA al control del fósforo presente en los vertidos de las aguas depuradas.

Según el séptimo informe sobre la aplicación de la Directiva 91/271/EEC, publicado en el año 2013 el porcentaje de fósforo generado por la fracción de las aguas residuales que incumplían la Directiva representa aproximadamente el 35 % del fósforo total. El fósforo es uno de los nutrientes que contribuyen en mayor grado a la eutrofización de lagos y aguas naturales. Su presencia causa muchos problemas en la calidad del agua incluyendo aumentos en los costes de purificación, dificultad en la conservación del lagunaje o pérdida de las poblaciones naturales

La eliminación del fósforo se consigue normalmente mediante precipitación química, que resulta ser cara y causa el aumento del volumen de lodo hasta un 40 %. Una alternativa es la eliminación biológica del fósforo mediante una tecnología que se basa en una modificación del proceso de fangos activos que requiere una configuración de planta más específica y una operación más compleja. Presenta la misma configuración que un sistema normal de fangos activos pero con un volumen anaerobio previo a la zona de aireación.

El control del oxígeno para realizar conjuntamente la eliminación de nitrógeno y fósforo ha de realizarse de manera rigurosa ya que la presencia de nitratos permite a las bacterias heterótrofas desnitrificantes competir por el sustrato e inhiben la eliminación del fósforo.

Referencias

[1] Based on the prediction method of sewage treatment effluent total phosphorus cascade self-organizing neural networks, August 5 2015. URL http://www.google.com/patents/CN105160422A?cl=en. CN Patent 105,160,422.

[2] Online soft measurement method for sewage treatment based on quick relevance vector machine, June 3 2015. URL http://www.google.com/patents/CN104680015A?cl=en. CN Patent App. CN 201,510,093,369.

[3] J Alex, L Benedetti, J Copp, KV Gernaey, U Jeppsson, I Nopens, M-N Pons, L Rieger, C Rosen, JP Steyer, et al. Benchmark simulation model no. 1 (bsm1).

[4] J Alex, JF Beteau, JB Copp, C Hellinga, Ulf Jeppsson, S Marsili-Libelli, MN Pons, H Spanjers, and H Vanhooren. Benchmark for evaluating control strategies in wastewater treatment plants. In *Control Conference (ECC), 1999 European*, pages 3746–3751. IEEE, 1999.

[5] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2014.

[6] Pedro Simón Andreu, Carlos Lardín Mifsut, and Manuel Abellán Soler. Optimización energética en edar de la región de murcia. *Ingeniería civil*, (168):93–112, 2012.

[7] Majid Bagheri, Sayed Ahmad Mirbagheri, Zahra Bagheri, and Ali Morad Kamarkhani. Modeling and optimization of activated sludge bulking for a real wastewater treatment plant using hybrid artificial neural networks-genetic algorithm approach. *Process Safety and Environmental Protection*, 95:12–25, 2015.

[8] Michael Bongards, Daniel Gaida, Oliver Trauer, and Christian Wolf. Intelligent automation and it for the optimization of renewable energy and wastewater treatment processes. *Energy, Sustainability and Society*, 4(1):1–12, 2014.

[9] Ronald Newbold Bracewell and Ronald N Bracewell. *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York, 1986.

[10] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[11] J. Brownlee. *Introduction to Time Series Forecasting with Python: How to Prepare Data and Develop Models to Predict the Future*. Jason Brownlee, 2017.

[12] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168. ACM, 2006.

[13] Ufuk Celik, N Yuntay, and C Sertkaya. Wastewater effluent prediction based on decision tree. *Journal of Selcuk University Natural and Applied Science*, pages 138–148, 2013.

[14] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619, 2002.

[15] J Comas, I Rodríguez-Roda, M Poch, KV Gernaey, Christian Rosén, and Ulf Jeppsson. Extension of the iwa/cost simulation benchmark to include expert reasoning for system performance evaluation. *Water science and technology*, 53 (4-5):331–339, 2006.

[16] John B Copp. *The COST Simulation Benchmark: Description and Simulator Manual: a Product of COST Action 624 and COST Action 682*. EUR-OP, 2002.

[17] Scott Cost and Steven Salzberg. A weighted nearest neighbor algorithm for learning with symbolic features. *Machine learning*, 10(1):57–78, 1993.

[18] Council of European Union. Council regulation (EU) no 91/271/2013, 2013. http://ec.europa.eu/environment/water/water-urbanwaste/implementation/pdf/Technical%20assessment%20UWWTD.pdf.

[19] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[20] Pierre A Devijver and Josef Kittler. *Pattern recognition: A statistical approach*. Prentice hall, 1982.

[21] Union européenne. Direction générale de la recherche. *The COST simulation benchmark: description and simulator manual*. Directorate-General for Research, 2002.

[22] Peter Fritzson. *Principles of object-oriented modeling and simulation with Modelica 3.3: a cyber-physical approach*. John Wiley & Sons, 2014.

[23] Jesús Colprim Galcerán, Ignasi Rodríguez-Roda Layret, Angel Freixó Rey, and Mireia Fiter. Control basado en lógica difusa para los sólidos en suspensión: Desarrollo e implementación en la edar granollers. *Ingeniería química*, (437): 104–112, 2006.

[24] Krist V Gernaey, Christian Rosén, and Ulf Jeppsson. Bsm2: A model for dynamic influent data generation. *Department of Industrial Electrical Engineering and Automation, Lund University, Lund, Sweden*, 2005.

[25] KV Gernaey, Christian Rosén, and Ulf Jeppsson. Wwtp dynamic disturbance modelling an essential module for long-term benchmarking development. *Water science and technology*, 53(4-5):225–234, 2006.

[26] KV Gernaey, Lyngby DTU, Denmark U Jeppsson, and S Winkler. Benchmark simulation model no. 1 (bsm1), 2008.

[27] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[28] Hong Guo, Kwanho Jeong, Jiyeon Lim, Jeongwon Jo, Young Mo Kim, Jong pyo Park, Joon Ha Kim, and Kyung Hwa Cho. Prediction of effluent concentration in a wastewater treatment plant using machine learning models. *Journal of Environmental Sciences*, 32:90 – 101, 2015. ISSN 1001-0742. doi: http://dx.doi.org/10.1016/j.jes.2015.01.007. URL http://www.sciencedirect.com/science/article/pii/S1001074215001278.

[29] Hong-Gui Han, Hu-Hai Qian, and Jun-Fei Qiao. Nonlinear multiobjective model-predictive control scheme for wastewater treatment process. *Journal of Process Control*, 24(3):47 – 59, 2014. ISSN 0959-1524. doi: http://dx.doi.org/10.1016/j.jprocont.2013.12.010. URL http://www.sciencedirect.com/science/article/pii/S0959152413002655.

[30] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[31] Simon S Haykin, Simon S Haykin, Simon S Haykin, and Simon S Haykin. *Neural networks and learning machines*, volume 3. Pearson Upper Saddle River, NJ, USA:, 2009.

[32] Mogens Henze. *Activated sludge models ASM1, ASM2, ASM2d and ASM3*, volume 9. IWA publishing, 2002.

[33] Felix Hernandez-del Olmo, Elena Gaudioso, and Antonio Nevado. Autonomous adaptive and active tuning up of the dissolved oxygen setpoint in a wastewater treatment plant using reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(5):768–774, 2012.

[34] Félix Hernández-del Olmo, Félix H Llanes, and Elena Gaudioso. An emergent approach for the control of wastewater treatment plants by means of reinforcement learning techniques. *Expert Systems with Applications*, 39(3):2355–2360, 2012.

[35] Aurelio Hernández Muñoz. *Depuración y desinfección de aguas residuales*. Colegio de Ingenieros de Caminos, Canales y Puertos, Quinta edición. España, 2001.

[36] Mingzhi Huang, Yongwen Ma, Jinquan Wan, and Xiaohong Chen. A sensor-software based on a genetic algorithm-based neural fuzzy system for modeling and simulating a wastewater treatment process. *Applied Soft Computing*, 27: 1–10, 2015.

[37] Robert Monjo i Agut. El índice n de la precipitación intensa. 2009.

[38] Patrick Kern, Christian Wolf, Daniel Gaida, Michael Bongards, and Seán McLoone. Cod and nh 4-n estimation in the inflow of wastewater treatment plants using machine learning techniques. In *Automation Science and Engineering (CASE), 2014 IEEE International Conference on*, pages 812–817. IEEE, 2014.

[39] Max Kuhn and Kjell Johnson. *Applied predictive modeling*, volume 26. Springer, 2013.

[40] Pierre Lison. An introduction to machine learning, 2015.

[41] Jacek Makinia. *Mathematical modelling and computer simulation of activated sludge systems*. Iwa Publishing, 2010.

[42] Richard McCleary, Richard A Hay, Erroll E Meidinger, and David McDowall. *Applied time series analysis for the social sciences*. Sage Publications Beverly Hills, CA, 1980.

[43] LeonardEddy Metcalf, Harrison P Leonard Metcalf, and Harrison P Eddy. *Tratamiento y depuración de las aguas residuales*. Number 543.2. Labor, 1981.

[44] Tom M Mitchell. Machine learning (mcgraw-hill international editions computer science series). 1997.

[45] Michela Mulas, Stefania Tronci, Francesco Corona, Henri Haimi, Paula Lindell, Mari Heinonen, Riku Vahala, and Roberto Baratti. Predictive control of an activated sludge process: An application to the viikinmäki wastewater treatment plant. *Journal of Process Control*, 35:89–100, 2015.

[46] Fionn Murtagh and Pierre Legendre. Wards hierarchical agglomerative clustering method: which algorithms implement wards criterion? *Journal of classification*, 31(3):274–295, 2014.

[47] United Nations. Department of Economic. *The Millennium Development Goals Report 2015*. United Nations Publications, 2015.

[48] United Nations. Department of Economic. *The Sustainable Development Goals Report 2017*. United Nations Publications, 2017.

[49] Gustaf Olsson and Bob Newell. *Wastewater treatment systems*. IWA publishing, 1999.

[50] Mustafa Cagdas Ozturk, Fernando Martin Serrat, and Fouad Teymour. Optimization of aeration profiles in the activated sludge process. *Chemical Engineering Science*, 139:1–14, 2016.

[51] PA Paraskevas, IS Pantelakis, and Themistocles D Lekkas. An advanced integrated expert system for wastewater treatment plants control. *Knowledge-Based Systems*, 12(7):355–361, 1999.

[52] Gilles G Patry and Imre Takacs. Simulation: a key component in the development of an integrated computer-based approach to wastewater treatment plant control. In *The 1994 IEEE Conference on Control Applications. Part 2(of 3), Glasgow, UK, 08/24-26/94*, pages 1023–1028, 1994.

[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[54] Daniel Peña. *Análisis de datos multivariantes*, volume 24. McGraw-Hill Madrid, 2002.

[55] Daniel Peña. *Fundamentos de estadística*. Alianza Editorial, 2014.

[56] Alianza por el Agua. Manual de depuración de aguas residuales urbanas. *Monográficos Agua en Centroamérica, Ideasamares*, 2008.

[57] Jun-Fei Qiao, Ying-Chun Bo, Wei Chai, and Hong-Gui Han. Adaptive optimal control for a wastewater treatment plant based on a data-driven method. *Water Science and Technology*, 67(10):2314–2320, 2013. ISSN 0273-1223. doi: 10.2166/wst.2013.087. URL http://wst.iwaponline.com/content/67/10/2314.

[58] Jun-fei Qiao, Guang Han, and Hong-gui Han. Neural network on-line modeling and controlling method for multi-variable control of wastewater treatment processes. *Asian Journal of Control*, 16(4):1213–1223, 2014. ISSN 1934-6093. doi: 10.1002/asjc.758. URL http://dx.doi.org/10.1002/asjc.758.

[59] Jianjun Qin and Huajun Guo. Expert system development on on-line measurement of sewage treatment based process. *Sensors & Transducers*, 164(2):227, 2014.

[60] Christian Rosén, Ulf Jeppsson, and Peter A Vanrolleghem. Towards a common benchmark for long-term process control and monitoring performance evaluation. *Water Science and Technology*, 50(11):41–49, 2004.

[61] Lindsay I Smith et al. A tutorial on principal components analysis. *Cornell University, USA*, 51(52):65, 2002.

[62] Henri Spanjers, Peter Vanrolleghem, Gustaf Olsson, and Peter Dold. Respirometry in control of the activated sludge process. *Water Science and Technology*, 34(3-4):117–126, 1996.

[63] Henri Spanjers, Peter Vanrolleghem, Khanh Nguyen, Henk Vanhooren, and Gilles G Patry. Towards a simulation-benchmark for evaluating respirometry-based control strategies. *Water Science and Technology*, 37(12):219–226, 1998.

[64] Imre Takács, Gilles G Patry, and Daniel Nolasco. A dynamic model of the clarification-thickening process. *Water research*, 25(10):1263–1271, 1991.

[65] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.

[66] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: an efficient data clustering method for very large databases. In *ACM Sigmod Record*, volume 25, pages 103–114. ACM, 1996.