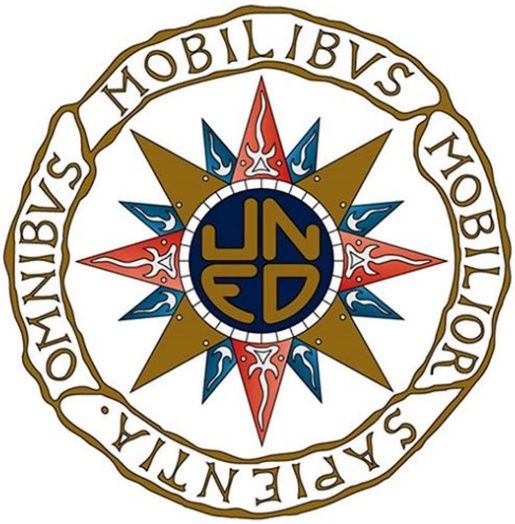
Análisis del efecto sobre el aprendizaje de la evaluación formativa en vídeos en el contexto de un MOOC real sobre Inteligencia Artificial

Autor: Jesús Ramos Guillou
Director: Prof. Emilio Letón Molina

Septiembre 2019



Universidad Nacional de Educación a Distancia
Máster Universitario en I.A. Avanzada

Je me sens mieux si je sais que j’ai une petite étoile, minuscule,
mais ferme

Índice general

Agradecimientos

Me gustaría agradecer a todas las personas que me han ayudado de una u otra manera a realizar este trabajo, empezando por Emilio, mi director. Gracias a mis hermanos por ofrecerse a hacer de *beta testers*; gracias a Raquel Viejo y a todo el equipo del CEMAV por la ayuda con el vídeo de difusión del curso, y al profesor Félix de la Paz por su participación en el mismo con Nao; gracias a UNED Abierta por la difusión y por mostrarse siempre solícitos con todas las peticiones que les realizábamos en relación con el curso; y, por supuesto, gracias a todas esas personas que están cerca, ellas saben quiénes son, por aguantarme durante esta última época, especialmente difícil.

Resumen

El presente trabajo analiza el efecto que tiene sobre el aprendizaje, en el contexto real de un MOOC (*Massive Open Online Course*) sobre aprendizaje automático para todos los públicos, la posición en la cual, respecto a un vídeo pódcast, se realiza la evaluación formativa (y se aporta el *feedback* relacionado). En concreto, a partir de la investigación existente en el área del aprendizaje multimedia y sobre el efecto del *feedback* en el aprendizaje y en la motivación, se compara la realización de preguntas de refuerzo en puntos intermedios de los vídeo pódcast (lo que en inglés se denomina *in-video quizzes*) con la realización de esas mismas preguntas al finalizar el vídeo (*post-video quizzes*). Derivado de la información empírica disponible, se esperaba obtener un resultado que arrojara diferencias significativas a favor del grupo experimental tanto en los resultados de retención, como de transferencia, además de en la tasa de alumnos que finalizan el curso y en la valoración subjetiva que los estudiantes hacen del mismo.

De las 712 personas que participaron en el curso, 317 fueron asignadas aleatoriamente a formar el grupo experimental y 395 el grupo de control. Tan sólo 35 personas de cada grupo ($8{,}9\,\%$ del grupo de control y $11\,\%$ del grupo experimental) completaron las seis lecciones y cinco test evaluados del curso. Los resultados indican que existe una ventaja significativa del grupo experimental sobre el de control en la puntuación de las preguntas orientadas a retención (Mann-Whitney-Wilcoxon: $p = 0{,}0017$), y que se obtienen resultados ligeramente superiores en transferencia, pero, en este caso, a pesar de lo esperado, sin significancia estadística. De la misma manera, no se detecta una diferencia significativa en la tasa de finalización del curso, a pesar de que el porcentaje de personas del grupo experimental que finalizan el curso es mayor que el del grupo de control. La valoración del curso es en general bastante buena, tanto en el grupo de control como en el experimental, sin diferencia entre ellos.

A pesar de tratarse, hasta donde sabemos, del primer intento de comparar, mediante un experimento aleatorio, el uso de preguntas de refuerzo intercaladas en los vídeos con preguntas de refuerzo realizadas al finalizar el vídeo, el presente trabajo tiene la limitación de no haber podido usar preguntas puramente embebidas en los vídeos. Futuros trabajos deberían tratar de replicar los resultados en otros MOOCs usando preguntas realmente embebidas, además de trabajar con una muestra mayor y afinando más los instrumentos de evaluación para tratar de delimitar si el uso de preguntas embebidas tiene impacto o no sobre los resultados en transferencia.

1. Introducción

En los últimos años, sobre todo con el auge y expansión del acceso a Internet, el uso de los entornos virtuales de aprendizaje se ha convertido en algo bastante común tanto en contextos formales como informales, y posiblemente lo sea aún más en los próximos años. Los contenidos multimedia, y especialmente el vídeo, son los motores actuales de estos entornos. Los MOOCs, La Khan Academy, el propio YouTube o la existencia de plataformas de gestión de cursos como Moodle son buena muestra de este fenómeno. Puesto que el vídeo es el medio principal por el que se intenta facilitar el aprendizaje, es de vital importancia investigar cuáles son las características de estos vídeos que permiten llegar más fácilmente al estudiante, posibilitando así un mayor compromiso, motivación y aprendizaje final.

No se debe perder de vista que, en estos entornos, el aprendizaje se convierte en un proceso en el que, el estudiante, para bien o para mal, es eminentemente autónomo y que la guía que le puede prestar un profesor es indudablemente menor que en un entorno presencial. Por ello, además de que los vídeos estén diseñados de forma que potencien el aprendizaje (qué presentan y cómo lo presentan), se vuelve también muy importante el posibilitar tanto que el alumno pueda saber en cada momento si está cumpliendo con los objetivos de aprendizaje pretendidos, mediante el uso de técnicas de evaluación formativa y de presentación de *feedback*, como el conseguir mantener su motivación intrínseca.

En este contexto, el presente trabajo investiga la influencia de una técnica bastante común en los entornos virtuales de aprendizaje como es la de intercalar preguntas de evaluación formativa en los vídeos (denominadas comúnmente en inglés *in-video quizzes*) y la compara con el uso de las mismas preguntas realizadas al finalizar cada vídeo (*post-video quizzes*) en el entorno de un MOOC real.

1.1. Objetivos

El principal objetivo del trabajo de investigación ha sido el comprobar, mediante la realización de un experimento aleatorio usando para ello un MOOC en un entorno real de aprendizaje, el efecto de intercalar preguntas orientadas a la evaluación y consolidación del propio aprendizaje en los vídeos respecto a la realización de esas mismas preguntas al finalizar los vídeos. Las preguntas concretas que se buscan

contestar con el experimento son: *¿Afecta de alguna manera el uso de preguntas embebidas en los vídeos a la tasa de alumnos que finalizan el curso con éxito? ¿Cómo afecta al rendimiento del aprendizaje el uso de preguntas intercaladas en los vídeos respecto al uso de preguntas al final de los vídeos? ¿Existe diferencia en la influencia por nivel de aprendizaje (recuerdo vs. transferencia)? ¿La experiencia subjetiva de los alumnos varía en función de la situación de las preguntas de refuerzo?*

Tanto en la formulación de las hipótesis como en el análisis de resultados, se tendrán en cuenta distintos experimentos previos y la teoría existente respecto al uso de la evaluación formativa y el *feedback* (Kluger y DeNisi (1996); Black y Wiliam (1998); Hattie y Timperley (2007); Shute (2008); van der Kleij et al. (2015)); distintas teorías del aprendizaje, como la teoría de la carga cognitiva (*Cognitive Load Theory*, Sweller et al. (1998); Sweller (2010)) o la teoría cognitiva del aprendizaje multimedia (*Cognitive Theory of Multimedia Learning*, Mayer y Moreno (2003)); o de la motivación (*Self Determination Theory*, Deci et al. (1999)).

Como objetivo secundario se plantea evaluar el entorno MOOC, y en concreto la plataforma Open edX (implantada en UNED Abierta), como plataforma de experimentación, comprobando si los mecanismos de configuración y de extracción de datos son suficientes para la realización de futuros experimentos o si cabe plantear alguna mejora.

1.2. Metodología

Para la experimentación se ha utilizado la plataforma MOOC Open edX, implantada en UNED Abierta, mediante la concepción, diseño, desarrollo, difusión e impartición de un curso, abierto al público en general, de introducción al aprendizaje automático orientado a personas que no contaran con una especial experiencia previa en informática, estadística o matemáticas. El curso diseñado consta de 6 lecciones, cada una de las cuales tiene contenido en vídeo (1 o 2 vídeos por lección de entre 10 y 18 minutos de duración) y secciones de texto de apoyo y ampliación asociadas. Se ha estimado una dedicación total por parte de los alumnos de unas 12 horas para finalizar el curso.

En la plataforma del curso se han configurado dos grupos o cohortes, una denominada *de control* y otra *experimental*. En la cohorte de control al finalizar ciertos vídeos aparecen una o dos preguntas que tienen el objetivo de que el alumno pueda comprobar si ha comprendido aspectos fundamentales de la materia y que en la presente memoria se van a denominar *preguntas de refuerzo*. En el grupo experimental, aunque el vídeo en sí es idéntico, estas mismas preguntas de refuerzo aparecen nada más finalizar la explicación del concepto asociado, deteniendo el vídeo sin esperar a la finalización del vídeo completo, provocando, por tanto, la división del vídeo en

distintas secciones. Es decir, la única diferencia entre los dos grupos es la cercanía de las preguntas de refuerzo a la explicación de su concepto asociado. La asignación de cada alumno a su correspondiente cohorte la realiza, de forma aleatoria, la plataforma en el primer acceso del alumno al material del curso.

El curso se abrió al público el día 27 de mayo de 2019 y contó con atención tutorial para abordar dudas durante 3 semanas, hasta el día 16 de junio de 2019. Aunque el curso tiene como fecha de finalización, ya sin atención tutorial, el día 15 de octubre de 2019, a fecha 20 de julio de 2019, fecha en la que se realizó la recogida de datos, se habían apuntado al curso 712 personas, habiendo accedido todas ellas en alguna ocasión al curso y, por tanto, con cohorte asignada. 317 de ellas pasaron a formar el grupo experimental y las otras 395 el grupo de control. Tan sólo 35 personas del grupo de control (el 8,9 % del total del grupo) finalizaron completamente el curso, de las cuales aprobaron (obtuvieron más de la mitad de los puntos en el total de los test) 33 personas, mientras que en el caso del grupo experimental fueron también 35 personas las que finalizaron el curso, el 11 % del total, y aprobaron 34 personas.

1.3. Estructura de la memoria

La memoria se ha estructurado como sigue:

- En el capítulo 2 se realiza una revisión bibliográfica de los distintos aspectos relacionados con el experimento. Se revisa la experimentación relacionada con el vídeo como material de enseñanza y su comparación con otros materiales, además de la evidencia empírica acerca de los beneficios y perjuicios que los distintos tipos de vídeo tienen sobre el aprendizaje. Se repasan estudios relativos a los MOOCs, al mejor uso del vídeo dentro de este tipo de cursos y estudios previos relacionados con el uso de las preguntas embebidas en vídeos, en especial en el contexto de los MOOCs. También se analiza la bibliografía relacionada con el *feedback* y con el uso del *feedback* en entornos virtuales y se examinan las principales teorías del aprendizaje multimedia, con énfasis en aquellos aspectos que puedan tener incidencia en el experimento.
- A continuación, en el capítulo 3, se describe el diseño del experimento y se formulan las hipótesis de trabajo que se ponen en relación con la bibliografía analizada en el capítulo anterior. Una vez descrito el experimento, se continúa con la descripción de los aspectos relacionados con la concepción, diseño y construcción del MOOC que da soporte al experimento. Se tratará desde la motivación para desarrollar específicamente ese MOOC, hasta decisiones de diseño que pueden tener importancia en el experimento, entre otras: duración del MOOC, grado de dificultad, necesidad de experiencia previa por parte de los alumnos o duración de los vídeos.

- En el capítulo 4 se detalla la ejecución del experimento, se describe con detalle la fase de extracción de datos, se analizan los datos extraídos y los resultados se ponen en contexto con las hipótesis de partida y la bibliografía asociada.
- Para concluir, en el capítulo 5, se resumen los resultados, se analizan posibles aspectos mejorables y limitaciones del experimento, además de plantear posibles mejoras y líneas de trabajo futuras.
- Se incluyen, como anexo A, los contenidos de una de las lecciones del MOOC, la lección 5, titulada «Selección de modelo y análisis del error».

2. Estado del arte

2.1. El fenómeno MOOC: luces y sombras

Desde su surgimiento en el año 2012 (Jordan (2015), Chuang y Ho (2016)) los Cursos Online Masivos y Abiertos (COMA, más conocidos por su abreviatura en inglés: MOOCs) han permitido a un gran número de personas complementar su formación en múltiples áreas de forma gratuita o a un muy bajo coste. Desde las iniciativas iniciales como Coursera (www.coursera.org), edX (www.edx.org) o Udacity (www.udacity.com), se ha pasado a un ecosistema de plataformas muy diverso en la que múltiples organizaciones cuentan con su propia oferta de cursos abiertos (por ejemplo, en el caso de España, plataformas como MiriadaX o UNED Abierta, por nombrar solamente dos, o FutureLearn a nivel europeo). Estas plataformas de enseñanza online suelen contar todas ellas con los mismos componentes principales que caracterizan a los MOOCs (Brinton et al. (2016)): (1) vídeos de las lecciones como vehículo principal de enseñanza; (2) componentes de evaluación en forma de ejercicios y de test que buscan tanto servir de refuerzo al aprendizaje como de forma de evaluación de los alumnos; y (3) un sistema de foros como principal mecanismo de interacción social y de comunicación entre los alumnos y el equipo docente.

Sin duda, uno de los elementos más atractivos de los MOOCs es la capacidad que tienen de llegar, a través de Internet, a un gran número de alumnos potenciales de todo origen y condición, posibilitando la eliminación de múltiples barreras –sociales, económicas o culturales– en el acceso a una educación –en principio– de calidad. Es decir, este paradigma parece, a priori, acercarnos a la promesa de un ideal como es el de la democratización de la educación. Como ejemplo, las iniciativas de Harvard y del MIT (denominadas HarvardX y MITx respectivamente) han contado entre 2012 y 2016 con 2,4 millones de usuarios únicos en 290 cursos (Chuang y Ho (2016)). Sin embargo, a pesar de lo que pudiera parecer inicialmente, Jordan (2015) nos alerta de que ciertos estudios han detectado que son los alumnos que ya cuentan con una mayor formación los que más habitualmente finalizan este tipo de cursos, lo cual contribuiría a aumentar la brecha en el acceso a la educación en lugar de contribuir a cerrarla. En la misma línea, Kizilcec et al. (2013) analiza la distribución de alumnos «activos» (aquellos que además de apuntarse al curso han accedido al mismo en algún momento) en función del Índice de Desarrollo Humano de sus países de origen para tres cursos de informática de distintos niveles de especialización –«*Computer*

Science 101», «*Algorithms: Design and Analysis*» y «*Probabilistic Graphical Models*»– y encuentra que, por ejemplo, para el curso con un nivel de especialización inferior, «*Computer Science 101*», tan sólo el 18 % de los alumnos provenían de países con un Índice de Desarrollo Humano medio o bajo, mientras que un 69 % de los alumnos procedían de los países en el cuartil más desarrollado. Los datos son similares para los cursos con un nivel más elevado («*Algorithms: Design and Analysis*» y «*Probabilistic Graphical Models*»). Además, hasta un 80 % de los alumnos de países con un Índice de Desarrollo Humano medio o bajo se encuadraban en aquellos alumnos que solamente veían los vídeos durante uno o dos períodos de evaluación, abandonando el curso posteriormente, siendo este porcentaje de un 69 % en el caso de alumnos provenientes de países en el primer cuartil del IDH.

Otros problemas que se han achacado a los MOOCs son el gran número de alumnos que puede llegar a tener cada profesor o el no disponer del acceso presencial a un docente (Brinton et al. (2016)), ambas relacionadas dado que influyen en el grado de interactividad alumno-profesor, que en un MOOC normalmente queda relegado al intercambio de mensajes en los foros del curso. Pero si hay una gran crítica que se le ha realizado a este tipo de cursos es la del bajísimo nivel de alumnos que los finalizan con éxito si se compara con el número de personas que inicialmente se inscribe en ellos, una mediana de tan sólo el 12,6 % en datos de 129 cursos (Jordan (2015)). Esta tasa de finalización de los cursos tiene una gran variabilidad y en el mismo análisis se identifican como factores muy importantes tanto la longitud del curso como el tipo de evaluación que se realiza, siendo la evaluación por pares, en la que los alumnos evalúan los trabajos de sus propios compañeros, una de las características que mayor influencia tiene en una menor tasa de finalización. Sin embargo, debe tenerse cierta precaución al analizar la tasa de finalización con éxito de los cursos como único indicador posible respecto a su utilidad puesto que un curso puede ser útil para distintos alumnos de múltiples maneras y la baja barrera de entrada a los mismos propicia la existencia de múltiples posibles estrategias de utilización por parte de los alumnos, que no necesariamente tienen que coincidir con la que interesa a los promotores del MOOC.

Si uno de los factores principales que influyen en la tasa de finalización de los cursos es su duración, parece evidente que contar con cursos más cortos y con una mayor flexibilización de los mismos, de manera que el alumno pueda adaptar los mismos a sus intereses y necesidades, puede ser una buena idea para conseguir aumentar la tasa de seguimiento y finalización de los cursos. Con este objetivo se han realizado distintas propuestas tendentes a conseguir una mayor modularidad de este tipo de cursos (Jordan (2015); Challen y Seltzer (2014)) de manera que cada MOOC trate un contenido muy concreto y que se permita de manera más simple el diseño de distintos itinerarios formativos enlazando distintos cursos, ya sea de la misma o de distintas instituciones, pudiendo adaptar los mismos a las necesidades del alumno. Pero lograr la modularidad es posible a distintos niveles, y, dando un paso más,

Letón y Molanes-López (2014) proponen, en ese sentido, una estrategia de diseño de vídeos formativos, denominados Minivídeos Docentes Modulares (MDM) que, entre otras características, promueva la reutilización de los mismos en distintos contextos. Para ello, se deben diseñar específicamente con esta idea en mente, centrando el vídeo en la explicación de un sólo concepto, sin realizar alusiones a ningún otro vídeo o contenido. Este «encapsulamiento» de los distintos conceptos en los vídeos, de la misma manera que ocurre en el diseño de software, facilita su reutilización en distintos cursos o componentes y el mantenimiento a la larga de los mismos, e incluso podría facilitar la construcción de cursos adaptativos en los que no todos los alumnos vean a priori todo el contenido posible, sino que éste se le vaya ofreciendo o recomendando por un motor de personalización en función de sus interacciones con el sistema.

2.2. El vídeo como vehículo de enseñanza y los marcos conceptuales del aprendizaje multimedia.

Si hay un aspecto que caracteriza a los MOOCs es el uso nuclear que hacen del vídeo formativo transmitido en línea por Internet –también llamados vídeo pódcast, *vodcasts* o *webcasts* (Kay (2012))– como herramienta formativa (Guo et al. (2014)). Es por ello que, aunque siempre ha sido un formato que ha atraído la atención de los investigadores, la explosión del fenómeno MOOC, junto con la universalización de servicios como YouTube o plataformas como la Khan Academy, han provocado que la investigación acerca de cómo diseñar los vídeos para conseguir maximizar el compromiso y la adquisición de conocimientos por parte de los alumnos haya adquirido una importancia especial en los últimos años.

Existen, entre otros, dos marcos conceptuales muy importantes que es conveniente revisar antes de analizar los aspectos más relevantes de la investigación en el uso de los vídeo pódcast como herramienta formativa. Estos marcos teóricos sirven como guía en la investigación de cuáles pueden ser los aspectos de mayor importancia en el diseño de los vídeos, y en general de los elementos multimedia. Son la «Teoría de la Carga Cognitiva» (*Cognitive Load Theory*, Sweller et al. (1998)) y la «Teoría Cognitiva del Aprendizaje Multimedia» (*Cognitive Theory of Multimedia Learning*, Mayer y Moreno (2003)) y en adelante se hará referencia a las mismas como CLT y CTML respectivamente.

2.2.1. La Teoría de la Carga Cognitiva (CLT)

La idea fundamental de la CLT (Sweller et al. (1998); Sweller (2003); Hasler et al. (2007); Paas et al. (2010); Sweller (2010)) es tratar de mantener la carga impuesta a la memoria de trabajo bajo control, dado que es conocida la baja capacidad de

esta memoria, tanto en espacio como en tiempo de retención (Baddeley (2003); Barrouillet y Camos (2007)), en contraposición a la memoria a largo plazo, que no sufre este problema (Sweller (2003)). La memoria de trabajo debe tener en todo momento la suficiente capacidad para que los conocimientos nuevos se integren con los *esquemas* (conocimiento) preexistentes en la memoria a largo plazo. Si en algún momento la memoria de trabajo no es capaz de tratar con todos los elementos que se le presentan, bien por la propia complejidad de la materia tratada, bien por las características de los materiales usados en el aprendizaje, éste no se podrá producir convenientemente y se deberán buscar estrategias para disminuir esa *carga cognitiva*.

La CLT distingue entre tres tipos de carga cognitiva, todas ellas definidas en función de las interacciones entre los distintos elementos que forman parte de la tarea concreta y del material usado para el aprendizaje (Sweller (2010)), de forma que, cuantas más interacciones entre elementos existen en una determinada tarea, ya sea por la dificultad de la tarea en sí o por los materiales empleados para abordarla, mayor es la carga cognitiva impuesta:

- **Carga intrínseca** (***intrinsic load***): es la carga cognitiva inherente de una determinada materia. Es fija para una tarea específica y para una persona concreta con unos determinados conocimientos previos y no es posible disminuirla mediante el diseño de los materiales de enseñanza, sino solamente modificando la tarea en sí. Un detalle fundamental es que el concepto «elemento» será diferente para personas con distinta experiencia previa y, por tanto, también las interacciones entre esos elementos y, por ende, la carga cognitiva. En este sentido, en Sweller (2010) se pone el ejemplo de un alumno que está aprendiendo a leer y para el cual cada letra, o parte de ésta, se corresponde con un símbolo, mientras que para un adulto, que aprendió a leer siendo niño, los símbolos los forman las palabras o incluso frases completas.
- **Carga extraña** (***extraneous load***)**:** es la carga innecesaria que impone un determinado material de aprendizaje cuando hace que las interacciones entre elementos sean mayores que los que serían imprescindibles para abordar la tarea deseada. Es importante darse cuenta de que la diferencia entre la carga intrínseca y la carga extraña solamente tiene sentido desde el punto de vista de alguien previamente entrenado en esa determinada tarea puesto que el alumno no es capaz de discriminar qué información de la presentada es imprescindible y cuál prescindible.

La carga cognitiva total impuesta por la combinación de una determinada tarea y las características del material utilizado es la suma de la carga intrínseca y de la carga extraña Sweller (2010). Si la suma (o la propia carga intrínseca) superan la capacidad de la memoria de trabajo, el aprendizaje no se produce.

- **Carga pertinente** (***germane load***): la carga pertinente es un concepto un tan-

to diferente que hace referencia a la cantidad de recursos que la persona que intenta aprender dedica a lidiar con la **carga intrínseca** y es, por tanto, la que posibilita el aprendizaje. Si el material utilizado en el aprendizaje está diseñado para que el alumno pueda dedicar los recursos de su memoria de trabajo a tratar principalmente con las interacciones entre elementos impuestas de forma intrínseca por la tarea, la carga pertinente y, en consecuencia, el aprendizaje, alcanzarán su punto máximo. Puesto que se ha puesto de relevancia que los símbolos y sus interacciones dependen de la tarea, del material y de la experiencia previa del alumno, podemos inferir que para lograr maximizar el aprendizaje los materiales deben estar expresamente diseñados para un determinado nivel de experiencia.

De la CLT se derivan una serie de principios que guían el diseño de los materiales de enseñanza (Sweller (2003, 2010)). Se revisan aquellos que se han considerado más pertinentes para el presente trabajo:

- **Efecto del ejemplo resuelto** (***worked example effect***): este efecto pone de manifiesto que cuando se cuenta con alumnos con poca experiencia previa, éstos pueden aprender más fácilmente mediante ejemplos resueltos que resolviendo el mismo número de problemas por ellos mismos. La razón, en principio, es que el ejemplo resuelto actúa como andamiaje que sustituye la falta de experiencia previa (de un esquema mental preexistente en la memoria a largo plazo que pueda guiar el proceso) y, por ello, al poder seguir paso a paso el proceso necesario para resolver el problema, se evita que el alumno deba enfrentarse a todos los elementos a la vez, reduciendo así la carga cognitiva.
  Es bastante habitual que esta técnica del ejemplo resuelto se complemente con la asignación al alumno de un problema muy similar al resuelto para que lo intente él mismo. Este arreglo posibilita que alumno y profesor puedan observar si el aprendizaje está siguiendo la senda correcta.
- **Efecto de modalidad** (***modality effect***): el efecto de modalidad deriva del efecto, comprobado empíricamente (Sweller (2003)), de que la capacidad de la memoria de trabajo aumenta al usar el canal auditivo y el canal visual de forma conjunta. Por ejemplo, en Griffin et al. (2009) se describe un experimento que obtuvo mejores resultados con el uso de transparencias de PowerPoint sincronizadas con audio que entregando el PowerPoint y el audio por separado, sin sincronizar. Este uso del doble canal es también una de las asunciones fundamentales de la CTML (Mayer y Moreno (2003)) que se analizará posteriormente. Hay que especificar que este efecto se produce cuando la información solamente es comprensible mediante la combinación de ambas fuentes de información, la obtenida por el canal auditivo y la obtenida por el canal visual. Si la información es comprensible con la información transmitida por uno sólo de los canales, nos encontramos con el efecto definido en el punto siguiente.

- **Efecto de la redundancia** (***redundacy effect***): el efecto de la redundancia se da cuando el aportar información adicional, en lugar de favorecer el aprendizaje, lo perjudica. En general se produce cuando la información adicional es innecesaria para la comprensión de la materia por parte **de un determinado alumno** y, por ello, está muy relacionado con la experiencia previa que éste tenga en la materia. Al presentar información innecesaria (carga extraña) se está obligando a la memoria de trabajo a analizar esa información para determinar si es necesaria y/o importante y, por tanto, obliga a manejar en la memoria de trabajo, en un mismo momento, un mayor número de elementos y sus relaciones de forma innecesaria.
- **Efecto de reversión por experiencia** (***reversal expertise effect***): directamente derivado de efecto de la redundancia se produce el efecto de que, conforme un alumno avanza en el aprendizaje de una materia, los apoyos al aprendizaje que dejan de ser necesarios no solamente dejan de ser positivos para pasar a ser neutrales, sino que, en ocasiones, pueden llegar a tener un efecto negativo en el aprendizaje. Es decir, muchos de los efectos postulados a partir de la CLT que sirven como guía de diseño de materiales de enseñanza (como por ejemplo el efecto del ejemplo resuelto), pueden tener, para alumnos con la suficiente experiencia previa, el efecto contrario al esperado.

2.2.2. La Teoría Cognitiva del Aprendizaje Multimedia (CTML)

La CTML (Mayer y Moreno (2003); Mayer (2005)) es una teoría que busca entender cómo procesa la mente humana los contenidos multimedia, entendidos éstos como la combinación de palabras (ya sea en formato escrito u oral) e imágenes (ya sean estáticas o dinámicas), con el objetivo de enunciar recomendaciones sobre su diseño que mejoren el aprendizaje. La CTML, de igual manera que la CLT, busca mantener bajo control la carga de la memoria de trabajo. Se basa en tres asunciones principales: (1) la *asunción de los canales duales*, que, igual que se propone en la CLT, dice que la mente dispone de dos canales diferentes de procesamiento, el auditivo y el visual; (2) la *asunción de la capacidad limitada*, que dice que ambos canales, visual y auditivo, tienen una capacidad muy limitada (nuevamente, una asunción muy similar, si no idéntica, a la de la CLT); y (3) la *asunción del aprendizaje activo*, que dice que para que un aprendizaje real sea posible debe producirse una cantidad significativa de procesamiento que Mayer divide en tres fases: seleccionar el material significativo, organizarlo de manera que se construya un modelo mental coherente y, por último, integrar este modelo mental con el conocimiento previo preexistente. Las fases iniciales de selección y organización de la información estarían dirigidas por el conocimiento previo con el que cuenta el estudiante.

Los vídeos, por su propia naturaleza, presentan la información de forma transitoria –información que en un momento dado estaba en la pantalla, en el momento

siguiente puede haber desaparecido–, lo que provoca que el procesamiento de la información se deba realizar a un ritmo que puede no ser el que conviene al estudiante y que obliga a mantener en la memoria de trabajo las relaciones espaciales y espacio-temporales entre elementos (por ejemplo, se puede dar el caso de un determinado elemento que ya no se presenta en la pantalla pero que es fundamental para entender un concepto mediante su relación con un elemento que sí está presentándose) aumentando, de esta manera, la carga cognitiva que sufre el alumno (Hasler et al. (2007)). Por ello, en este tipo de materiales es muy importante tener en cuenta los principios de diseño que, de la misma manera que en la CLT, se derivan a partir de la CTML y que buscan una reducción de la sobrecarga cognitiva. Algunas de ellas son fundamentalmente las mismas que en la CLT, como el efecto de modalidad o el efecto de redundancia, pero hay algunos otros efectos específicos de la CTML, que en este marco de trabajo se suelen denominar principios de diseño multimedia, como el principio de señalización o el de segmentación, que resultan especialmente relevantes.

El principio de **señalización** (Mayer y Moreno (2003); Mautone y Mayer (2001)) hace referencia a la técnica de aportar pistas o ayudas al estudiante acerca de cómo seleccionar u organizar mentalmente la información que se le presenta. La señalización puede abarcar cualquier efecto que ayude a guiar la atención del alumno hacia aquella información importante o que le permita deducir con mayor facilidad la estructura y relaciones entre los elementos que conforman la materia que se está trabajando. Como demostración empírica, en Mautone y Mayer (2001) se describen tres experimentos aleatorios en los que se presentó a los alumnos los principios físicos del vuelo de un aeroplano. En cada uno de los tres experimentos se utilizó para un grupo de alumnos un contenido con señalizaciones y para el otro grupo el mismo contenido sin ellas. Los experimentos se diferenciaban entre sí en que en el primer experimento el material se componía de texto escrito, en el segundo de texto narrado, y en el tercero de una animación con texto narrado. Los resultados de los tres experimentos, para los tres tipos de materiales, marcaron una mejoría significativa sobre los test de transferencia (aplicación de los conocimientos a un contexto distinto) de los grupos que disfrutaron del material con señalización, pero no así en los de retención, para los cuales no se apreció un efecto consistente. El efecto de la señalización será más evidente para aquellas materias que tengan una alta carga intrínseca, como pueden ser las matemáticas o las ciencias, o para materiales que impongan una alta carga externa, como puede ser el caso de los vídeos, dependiendo de la velocidad con la que se presenten los conceptos. Cuando no es el caso, la señalización puede actuar como distracción para el alumno (Schrader y Rapp (2016)). En un metaanálisis realizado a partir de los datos de 32 estudios, Xie et al. (2017) informan de un efecto significativo del uso de la señalización en materiales multimedia tanto sobre la percepción subjetiva de carga cognitiva como sobre el rendimiento en aprendizaje (tanto retención como transferencia).

El principio de **segmentación** (Mayer y Moreno (2003); Mayer y Chandler (2001)), como su nombre indica, se basa en crear segmentos de la información, con pausas intermedias para permitir la reflexión por parte del alumno y el procesado completo de la información presentada con anterioridad. Se trata de un efecto que tiene especial importancia en el caso de los vídeos, debido a la carga que imponen por la forma dinámica que tienen de presentar la información (Spanjers et al. (2010)). En Mayer y Chandler (2001), se reprodujo este efecto mediante dos experimentos en los que usaron como material una animación explicando la formación de los rayos. En el primer experimento un grupo veía en primer término la animación dividida en 16 partes y posteriormente de forma completa (sin la capacidad de detenerla), mientras que otro grupo lo hacía en el orden contrario. En el segundo experimento el primer grupo visualizaba la animación en dos ocasiones, en ambas de forma segmentada, y el segundo grupo realizaba dos visualizaciones completas de la animación, sin la capacidad de manejar la animación. En ambos experimentos, el primer grupo, que en el primer experimento veía la animación segmentada en primer lugar, y que en el segundo la veía en las dos ocasiones segmentada en partes, obtuvo unos resultados en transferencia significativamente superiores al otro grupo. Este efecto, especialmente en los niveles altos de aprendizaje (transferencia), se ha visto refrendado en distintos experimentos, como, por ejemplo, en Moreno (2007) o en Schittek Janda et al. (2005).

Además de la mera función de pausa, Spanjers et al. (2010) proponen también que la segmentación puede actuar como una forma de efecto de señalización temporal, permitiendo que el alumno comprenda más fácilmente qué eventos tienen relación directa entre sí y cuáles deben separarse, capturando así la estructura interna de la materia que se está tratando. En cuanto a los contextos en que la segmentación puede resultar más útil, Ibrahim (2012) y Spanjers et al. (2011) indican que principalmente en los casos en que la materia tratada es compleja, debido a su carga intrínseca; cuando el ritmo del vídeo es rápido, lo cual provoca que la información fluya más rápido, causando una mayor carga cognitiva; y cuando los alumnos son inexpertos en la materia. A este respecto, Boucheix y Guignard (2005) y Spanjers et al. (2011), en sendos experimentos, encuentran que la segmentación es mucho más eficiente para alumnos sin una experiencia previa que para alumnos expertos, en un caso claro de efecto de reversión por experiencia.

2.2.3. Investigación sobre el vídeo como soporte formativo

Aunque la investigación sobre el uso del vídeo como forma de transmitir conocimientos se remonta a bastante tiempo atrás (Kozma (1986)), no ha sido hasta que la tecnología no ha avanzado lo necesario como para permitir la suficiente interactividad en los vídeos y una entrega del control del ritmo del vídeo al alumno, cuando las posibilidades formativas de este soporte se han empezado a descubrir en todo

su esplendor. Así lo demuestran distintos experimentos (Schwan y Riempp (2004); Zhang et al. (2006); Hasler et al. (2007); Merkt et al. (2011)) que dejan patente que el verdadero potencial del vídeo formativo solamente aparece cuando va acompañado de componentes interactivos. Esta interactividad puede ser de distinto tipo, desde el simple control del ritmo del vídeo que permita al alumno regularlo para ajustarlo a sus preferencias y capacidades, hasta la formulación de preguntas o la entrega de *feedback*. En realidad, no se trata de un fenómeno que se haya descubierto con los avances tecnológicos de los últimos veinte años, puesto que en Merkt et al. (2011) nos cuentan que ya en 1994, Wetzel, Radtke y Stern (Wetzel et al. (1994)) concluyeron que altos niveles de interactividad en entornos de vídeo por ordenador estaban relacionados con mejores resultados. Si ponemos estos resultados en el contexto de la CLT, vemos que tienen todo el sentido, debido a la alta carga cognitiva que impone este medio cuando no hay mecanismos de regulación de la misma (Hasler et al. (2007)).

Se ha encontrado que cuando se le añade la interactividad al vídeo, éste tiene la capacidad de igualar o incluso, en ocasiones, de superar al mismo contenido representado en papel, a pesar de que históricamente diversos estudios habían encontrado que, salvo en niños pequeños, el texto impreso conseguía ventaja sobre el vídeo (sin control de flujo) (Merkt et al. (2011)). Lógicamente, cuando uno se enfrenta a un texto impreso, cuenta por defecto con mecanismos que le permiten adecuar la lectura a sus propias capacidades, gustos o entendimiento y no es hasta que no se aporta la misma capacidad a los vídeos cuando éstos obtienen cierta ventaja, ya sea en la transmisión de conocimiento procedimental (Schwan y Riempp (2004); Hasler et al. (2007)) o declarativo (Merkt et al. (2011)). La importancia que tiene que el alumno pueda autorregularse, o al menos disfrutar de ciertas pausas en el procesamiento de la información, ya se había intuido cuando se ha tratado el efecto de la segmentación, por lo que podemos decir que estos resultados están también en línea con los propuestos por la CTML.

Es probablemente la obtención de estos resultados prometedores de los *video podcasts* lo que ha provocado que aproximadamente desde 2005 la investigación sobre el uso de este medio en la educación se haya multiplicado. Kay (2012), en su revisión de la literatura sobre el uso de los pódcast de vídeo en educación, indica que con anterioridad a 2006, tan sólo 8 artículos con revisión por pares se habían escrito sobre esta materia, pasando a 52 en el período 2006-2012. Es importante que se produzca esta investigación ya que, aunque, como se ha visto, existe cierta evidencia de que el uso del vídeo formativo puede tener un efecto positivo, no está totalmente resuelto en qué condiciones se da ese efecto, para qué materias, con qué características deben contar los vídeos o si la edad, experiencia o gustos de los alumnos influyen en cómo les afectan los vídeos. En los párrafos siguientes se realiza una revisión de algunos de los aspectos en este sentido que ya han sido estudiados.

Uno de los aspectos importantes de los pódcast de vídeo es entender en qué con-

texto se utilizan, puesto que la utilidad de ciertos tipos de vídeo puede depender de ese contexto concreto. Éste puede ser, por ejemplo, como parte de un MOOC o entorno virtual de enseñanza (Guo et al. (2014)), como complemento a las clases tradicionales (conocido en inglés como *blended learning*) (Griffin et al. (2009); Wieling y Hofman (2010)), o como forma de que los alumnos puedan tener una primera aproximación a la materia **antes** de asistir a clase (*flipped classroom*) (Kay (2012); Muñoz-Merino et al. (2017)) de forma que en las clases el profesor se pueda centrar en aquellos aspectos de la materia que cuentan con mayor dificultad, a resolver dudas, o a realizar ejercicios.

En cuanto a los distintos tipos de vídeo respecto a la forma en que son grabados, aunque existen propuestas de clasificación exhaustiva para enfocar su estudio de forma homogénea (Letón et al. (2012)), hay unos pocos tipos de vídeo bastante comunes, como son (Guo et al. (2014); Chen y Wu (2015)):

- **Grabación de la clase:** se trata de uno de los tipos de vídeo más usados por su facilidad de producción y consiste en grabar al instructor impartiendo una clase real.
- **Grabación en estudio:** es cuando se realiza el vídeo grabando en un estudio al instructor a cuerpo completo y la imagen de éste se alterna o se superpone a los contenidos en postproducción.
- **Grabación en oficina**: similar a la anterior, pero la grabación del instructor se realiza en una mesa de oficina que hace que sólo sea visible el torso y la cabeza del instructor.
- **Narración:** el instructor narra la clase sobre una serie de diapositivas que se presentan en pantalla. Se suele ver complementada con una pequeña imagen del instructor.
- **Estilo Khan:** es el estilo puesto de moda por la Khan Academy y que consiste en que el instructor va narrando mientras dibuja en una tableta digital a mano.

Lógicamente, es común encontrarse con mezclas de estos estilos, como podría ser, por ejemplo, un vídeo estilo «narración» en la que el instructor aparece en la mesa de su oficina y, por tanto, solamente se le ve la parte superior del cuerpo y que, además, usa diapositivas sin mucho texto en las que va escribiendo o dibujando, mediante una tableta, lo que va explicando.

En cuanto al contenido pedagógico de los vídeos, se suelen encontrar dos tipos principales (Guo et al. (2014); Kay (2012)): los vídeos que explican conceptos y los que explican la resolución de problemas, o, más en general, «tutoriales» acerca de algún procedimiento concreto.

Derivado de los tipos descritos de pódcast de vídeo, una de las características más analizadas en la literatura es cuál es la influencia sobre los alumnos de la aparición

en la imagen del profesor, y de la forma concreta de esta aparición (por ejemplo, Guo et al. (2014); van Gog et al. (2014); Chen y Wu (2015); Korving et al. (2016); van Wermeskerken y van Gog (2017); Wang y Antonenko (2017); Hong et al. (2018)). Los resultados obtenidos por los distintos experimentos no son concluyentes, no habiéndose encontrado en distintos casos ninguna diferencia entre la presencia o ausencia del profesor en la imagen (Kizilcec et al. (2014); Homer et al. (2008)) o habiendo obtenido solamente una mejora en retención en vídeos sin excesiva dificultad (Wang y Antonenko (2017)). Sin embargo, sí parece estar más claro que los alumnos perciben la aparición del profesor como algo positivo para su propio aprendizaje (Guo et al. (2014); Kizilcec et al. (2014); Wang y Antonenko (2017)) y ello a pesar de que parece contrastado que el profesor, en ocasiones, tiene el efecto de desviar la atención de la información que se muestra en la pantalla (Wang y Antonenko (2017)) por lo que es probable que sea más apropiada una aparición del docente en ciertos momentos en que la materia sea más simple o que se alterne la imagen del profesor con la de la materia.

Sin duda, el efecto que puede tener la aparición del profesor en los vídeos depende de múltiples factores por lo que es necesaria una investigación mucho más extensa sobre el particular para poder dilucidar de qué manera es apropiada la aparición del instructor, si es que en algún caso lo es. Por ejemplo, Hong et al. (2018) midieron cómo afectaba la aparición o no del profesor en función de que el contenido del vídeo fuera declarativo o procedimental y, en su contexto concreto, llegaron a la conclusión de que la aparición del profesor aumentaba la carga cognitiva en los vídeos con contenido procedimental, pero no así en los de contenido más declarativo, para los cuales la presencia del profesor mejoraba el aprendizaje. En van Gog et al. (2014), sin embargo, en un experimento que comparó la influencia de ver la cara del instructor o no verla, en un vídeo que demostraba la resolución de un problema, obtuvieron un mejor resultado cuando la cara del profesor era visible. Curiosamente, este resultado no pudo ser replicado posteriormente en un experimento prácticamente idéntico (van Wermeskerken y van Gog (2017)).

En cuanto a las distintas formas en que el profesor puede aparecer en la imagen, en Korving et al. (2016) se compararon seis vídeos de diferentes profesores, con diferentes características, en tres versiones distintas: sin la aparición del profesor, con el profesor en tamaño pequeño y con el profesor en un tamaño grande. Los resultados no fueron concluyentes. Por su parte, Chen y Wu (2015) analizaron las diferencias entre un vídeo tipo «grabación de clase», un segundo vídeo estilo «narración», y un tercer vídeo con el profesor grabado en estudio y superpuesto en la imagen, concluyendo que, en el contexto del experimento, el vídeo en formato narración fue inferior a los otros dos tipos de vídeo. Por último, es importante destacar la limitada existencia de investigación del efecto de la aparición del instructor cuando los consumidores de los vídeos no son adultos.

Más allá de la presencia o no del profesor en el vídeo, ya se ha visto en la sección

dedicada a los MOOC (sección 2.1) la gran influencia que tiene la duración de los vídeos. Efectivamente, se trata también del mayor indicador del tiempo de interacción (*engagement*) con los vídeos según las conclusiones expuestas por Guo et al. (2014) a partir de los datos de cuatro cursos de edX (6,9 millones de vídeos visualizados). Otros datos interesantes aportados son que la mediana de tiempo de visualización de los distintos vídeos no supera los 6 minutos, sea cual sea la duración del vídeo, y que la duración de los vídeos afecta a la tasa de contestación de las preguntas realizadas posteriormente. Estos efectos son coherentes con los obtenidos de las distintas experiencias con la segmentación y nos indican que es mejor diseñar vídeos muy breves, que abarquen un sólo concepto y que apuesten por la interacción, en línea con las propuestas realizadas en Letón y Molanes-López (2014).

Un tipo especial de animación, que cada vez cada vez se ve con mayor frecuencia en los vídeos, es la de la escritura o dibujo a mano, por ejemplo la utilizada en los vídeos estilo Khan. No existen demasiados experimentos que se hayan centrado en analizar el efecto de la escritura a mano respecto al uso de la tipografía de imprenta. En Cross et al. (2013) compararon los dos estilos y después de analizar la opinión al respecto de 150 personas, concluyeron que la escritura a mano resultaba más cercana y atractiva mientras que el uso de la letra de imprenta resultaba más clara y legible. Luzón y Letón (2015) realizaron un experimento aleatorio con alumnos de secundaria en el que compararon el efecto de la animación, mediante la escritura a mano, de fórmulas matemáticas con respecto a presentarlas escritas en PowerPoint, obteniendo una mejora significativa en el resultado sobre el aprendizaje, tanto en retención como en transferencia, para el grupo que disfrutó de la versión del vídeo con animación de las fórmulas.

Pero no sólo el tipo de escritura puede influir en los resultados. En una interesante propuesta, Rodriguez-Ascaso et al. (2018), realizan, primero, una necesaria llamada sobre la atención que se debe prestar a las pautas de accesibilidad a la hora de diseñar el contenido formativo, en este caso los vídeos, de forma que la mayor parte de la población pueda hacer uso del mismo, evitando así, en la medida de lo posible, discriminaciones involuntarias; y, segundo, comprueban experimentalmente que el realizar un diseño que tenga en cuenta la accesibilidad, no perjudica en absoluto a los alumnos sin discapacidad, sino más bien al contrario ya que los 228 niños de 12 años participantes en el experimento encontraron el vídeo accesible significativamente más atractivo que su contraparte no accesible.

2.3. La importancia de la evaluación formativa y del *feedback*

2.3.1. Definiciones y caracterización del *feedback*.

Evaluación formativa y *feedback* son conceptos muy íntimamente relacionados. Black y Wiliam (1998) definen la evaluación formativa (*formative assessment*) como «todas aquellas actividades llevadas a cabo por maestros o alumnos que proporcionan información que pueda ser utilizada como *feedback* para modificar las actividades de enseñanza y aprendizaje en las que participan», en contraposición a una evaluación clásica, más pensada para calificar al alumno, conocida en inglés como *summative assesment* y que se suele traducir en la literatura como evaluación sumativa. Por su parte, tanto en Hattie y Timperley (2007) como en van der Kleij et al. (2015) se recoge una excelente definición del *feedback* (formativo) de Winne y Butler (1994), que lo definen como «aquella información a partir de la cual un alumno puede confirmar, completar, sobrescribir, ajustar o reestructurar información en la memoria, ya sea esa información conocimiento del dominio, conocimiento metacognitivo, creencias sobre sí mismo o sobre las tareas, o tácticas y estrategias cognitivas». El *feedback*, por tanto, es una parte fundamental del ciclo que aporta valor a la evaluación formativa, ya que sin esta información la evaluación carece de efecto, al ser el mecanismo que posibilita que el alumno (o el profesor) disponga de la información necesaria sobre su propio progreso con respecto a un objetivo concreto y pueda, o bien tener la seguridad de que ha comprendido correctamente un determinado concepto, o bien descartar esquemas mentales incorrectos antes de construir sobre ellos. Se trata, por tanto, de una herramienta que guía la autorregulación del alumno con respecto a su propio aprendizaje (ver, entre otros, Black y Wiliam (1998); Shute (2008)).

Aunque parece no haber consenso respecto al mecanismo que permite que el uso del *feedback* potencie el aprendizaje (van der Kleij et al. (2015)), a decir de Hattie y Timperley (2007), a partir de la información de 12 metaanálisis, el efecto promedio del uso del *feedback* en clase estaría entre las técnicas que mayor influencia pueden tener en el aprendizaje, aunque es necesario puntualizar que los distintos metaanálisis están basados en estudios muy heterogéneos y a menudo aportan resultados contradictorios (van der Kleij et al. (2015)). Pero no todos los tipos de feedback son igual de efectivos y, de hecho, ciertos tipos de feedback pueden ser directamente contraproducentes (fundamentalmente cuando se usan recompensas externas que poco tienen que ver con la tarea o cuando la retroalimentación se usa como una forma de control sobre el alumno –eliminando su capacidad de autorregulación–).

Es habitual ver los distintos tipos de *feedback* clasificados de la siguiente manera (Shute (2008); Narciss (2008); van der Kleij et al. (2015)):

- **Conocimiento del resultado (*KR, Knowledge of results*)**: indica únicamente

si la respuesta es correcta o no, sin aportar más información y, en su forma más básica, sin permitir varios intentos.

- **Contestación hasta la respuesta correcta** (***AUC*, *Answer until correct***): variante de la anterior en la que se permite contestar la pregunta hasta que la cuestión se contesta correctamente.
- **Conocimiento de la respuesta correcta** (***KCR*, *Knowledge of the correct response***): indica cuál es la respuesta correcta.
- ***Feedback* elaborado** (***EF*, *Elaborated feedback***): cuando el *feedback* aporta información adicional que permite al alumno darse cuenta del error, ayudarle a tomar una acción correctiva o reforzar el entendimiento de un concepto, si es correcto. Puede tomar muchas formas distintas como, por ejemplo: realizar un análisis del error, dar una explicación de la respuesta correcta, aportar pistas al fallar la respuesta o sugerir material adicional (Shute (2008)). A menudo la variación del KR en la que se permite contestar la pregunta hasta dar con la respuesta correcta (AUC) se considera *feedback* elaborado (van der Kleij et al. (2015)).

En general, la información empírica disponible indica que es esperable que el EF tenga un impacto bastante mayor que los tipos de *feedback* más básicos.

Para analizar en mayor profundidad el efecto de las distintas formas posibles de *feedback*, Hattie y Timperley (2007) diferencian entre el funcionamiento de esta información a cuatro niveles a los cuales la información entregada/obtenida puede apuntar:

- **Nivel de tarea**: es información que indica cómo se ha realizado o comprendido una tarea o concepto. Probablemente es el tipo de *feedback* más común y es efectivo siempre y cuando desencadene un proceso de reflexión o una acción de autorregulación. A este nivel hacen referencia los *feedback* de tipo KR o KCR, si no se combinan de ninguna manera con alguna característica de *feedback* elaborado.
- **Nivel de proceso**: hace referencia al proceso que es necesario seguir para la realización de una tarea o la comprensión de un concepto. Es beneficioso para alcanzar niveles más profundos de aprendizaje que el *feedback* a nivel de tarea.
- **Nivel de autorregulación**: es información que trata de influir o guiar la forma concreta (acciones) en la que el alumno se autorregula en el proceso de aprendizaje o de realización de una tarea y por tanto tiene que ver con conceptos como la autoevaluación, autonomía o autocontrol del alumno.
- **Nivel personal** (**del *yo***): es información que, aunque lleva implícita información sobre si la tarea se ha realizado bien o mal, está directamente dirigida a características personales del alumno (Hattie y Timperley (2007) ponen como

ejemplo: «eres un gran estudiante» o «bien hecho, esa es una respuesta inteligente»). Es el menos efectivo de los tipos de retroalimentación y, a menudo, sus efectos llegan a ser negativos. Si una información a otro nivel se combina con *feedback* a nivel personal, hace que éste tenga menos efecto.

En general, el *feedback* a nivel de proceso o autorregulación es más efectivo que el *feedback* a nivel de tarea, y, como norma, es mejor evitar cualquier referencia al nivel personal.

Para que el *feedback* sea beneficioso, el alumno debe prestar atención a su importancia y tener la suficiente capacidad estratégica para saber extraerle valor a esa información (Hattie y Timperley (2007); Winne y Butler (1994)). Es decir, para que el alumno sepa interpretar y realizar acciones correctoras a partir del *feedback* se requiere: (1) que posea cierta capacidad de autorregulación, y (2) que el *feedback* sea consumido prestándole la atención necesaria (Black y Wiliam (1998)). Por ello, es importante que el *feedback* tenga en cuenta el nivel actual del alumno y esté diseñado de forma que estimule el procesamiento activo (Hattie y Timperley (2007)).

En relación con el nivel del alumno con respecto a las actividades generales de aprendizaje (autorregulación) y con respecto a la materia concreta, son precisamente estas dos las principales interacciones del *feedback* con otras variables (Hattie y Timperley (2007); Shute (2008); Johnson-Glenberg (2010)). Las tareas simples se benefician de un *feedback* dirigido a nivel de tarea en mayor medida que las complejas (Hattie y Timperley (2007)) y un alumno de un nivel bajo se beneficiará más de un *feedback* elaborado que un alumno avanzado, para el que es mejor simplemente comunicarle si la tarea es correcta o no (Shute (2008)), probablemente por la misma razón que la reversión de los efectos debido a la experiencia en las teorías del aprendizaje multimedia: un alumno va necesitando menos apoyos conforme se va volviendo más experto en una materia, pudiendo, incluso, llegar a ser contraproducentes.

Hay un aspecto del *feedback* para el que, a pesar de haberse estudiado bastante, no se han obtenido resultados claros y unívocos: es el de la temporalidad del *feedback* (Hattie y Timperley (2007); Shute (2008)). En general se suelen distinguir dos tipos de *feedback* en función del momento en que se da la retroalimentación con respecto al momento en que se responde o finaliza una tarea. El *feedback* se considera *inmediato* si se da inmediatamente después de la contestación y se considera *diferido* en caso contrario (van der Kleij et al. (2015)). Para esta variable, Hattie y Timperley (2007) comunican que un *feedback* inmediato es mejor cuando el *feedback* está dirigido al nivel de tarea pero que, en cambio, si el *feedback* está dirigido al nivel de proceso, es mejor que su entrega sea diferida. Además, comunican también una interacción entre la temporalidad del *feedback* y la dificultad de la tarea, obteniendo el *feedback* diferido mejores resultados conforme la tarea es más compleja (Clariana et al. (2000)). Por su parte, en Shute (2008) informan de resultados conflictivos

que en general parecen indicar que es mejor un *feedback* diferido para conseguir resultados en transferencia y que el *feedback* inmediato es mejor para tareas como las matemáticas o la programación (tareas procedimentales), especialmente en el corto plazo. En cualquier caso, sí parece claro que los alumnos prefieren un *feedback* inmediato a uno diferido y que dedican más tiempo a leer un *feedback* inmediato que uno diferido (Miller (2009); van der Kleij et al. (2012)).

Pero el uso del *feedback* no contribuye solamente a potenciar –directa o indirectamente– los resultados del aprendizaje. Existen sobradas muestras de que también influye en un aspecto muy importante: la motivación del alumno (ver, entre otros, Ryan (1982); Hidi (1990); Deci et al. (1999); Narciss y Huth (2004, 2006); Lazowski y Hulleman (2015); Faber et al. (2017)). La motivación es un factor fundamental para que se produzca el aprendizaje (Hidi (1990)) y en sí misma influye sobre los resultados, pero no sólo en ellos, sino que también influye notablemente sobre la predisposición, comportamiento y el esfuerzo de los alumnos (Lazowski y Hulleman (2015)). De entre las múltiples teorías de la motivación existentes en el ámbito de la psicología, la que más habitualmente aparece mencionada en los estudios relacionados con la evaluación formativa y con el *feedback* es, probablemente, la *Self-Determination Theory* (Ryan y Deci (2000)). Esta teoría distingue un tipo especial de motivación: la motivación intrínseca. Es aquella que procede de nosotros mismos, de nuestros intereses y curiosidad, y Ryan y Deci la vinculan con el aprendizaje y la creatividad. Aquellas personas que se mueven motivados exclusivamente de esta manera, en contraposición con «motivaciones» impuestas, tienen una mayor confianza, rendimiento y creatividad. Una de las formas de potenciar la motivación intrínseca es a través de un *feedback* apropiado (Deci et al. (1999)), sobre todo aquel que proporciona la sensación de competencia y autonomía. Este último punto es importante: el *feedback* no debe pretender controlar las futuras acciones del alumno si quiere mantener su efecto positivo, ya que la autonomía es un aspecto fundamental. Es lo que Deci y Ryan denominan *feedback* de control vs. *feedback* de información. Es importante señalar que Deci et al. (1999) encuentran mejora en la motivación mediante el uso de *feedback* en alumnos adultos, pero no así en niños. Curiosamente, el uso de un *feedback* apropiado, aquel en sintonía con el nivel actual del alumno, que promociona su autonomía y que no hace referencias personales, mejora tanto la motivación como el rendimiento, pero, a su vez, cierta evidencia empírica indica que los alumnos más motivados prestan mayor atención al *feedback* (ver referencias en Faber et al. (2017)), consiguiendo, de esta manera, una suerte de círculo virtuoso.

Respecto a la evaluación formativa en sí misma, un mecanismo más que es posible que influya en los resultados sobre el aprendizaje, en este caso especialmente sobre la capacidad de recuerdo, aunque también tiene influencia sobre la transferencia, es el denominado *testting effect* (Roediger III y Karpicke (2006); Roediger III y Butler (2011)). Esto es, que la práctica de la recuperación repetida de conocimientos del cerebro propicia mejoras en la retención de los conceptos, incluso por encima del

estudio repetido de un tema en la retención a medio plazo. Este efecto es aún mayor si la evaluación formativa se combina con su correspondiente *feedback.* Una posible explicación de este efecto es la visión neurocognitiva de que tratando de recordar con un cierto esfuerzo se provoca la activación de los mismos caminos neuronales que codifican la información y que esa activación contribuye a reforzar el recuerdo (Johnson-Glenberg (2010)).

2.3.2. El *feedback* en entornos virtuales.

El potencial beneficio que se puede obtener del uso de técnicas de evaluación formativa y de *feedback* no depende solamente de la forma que tome la información de retroalimentación, de la experiencia previa del alumno, o de la dificultad de las tareas, sino que va a depender también del contexto interpersonal en el cual se presenta (Deci et al. (1999)) o de los riesgos –de cualquier tipo– que pueda percibir el alumno a la hora de contestar a la evaluación formativa (Black y Wiliam (1998)). En este sentido, los entornos virtuales de aprendizaje podrían llegar a representar una ventaja importante ya que podemos hipotetizar que tienen la capacidad de conseguir simplificar el contexto interpersonal y reducir los riesgos percibidos por el alumno al tratarse de interacciones (en principio privadas) alumno-ordenador. Ya sea por este motivo o por otros, Shute (2008) nos informa de que los efectos del *feedback* son en general mayores en entornos virtuales que en un entorno de clase normal.

En 2015, Van der Kleij, Feskens y Eggen (van der Kleij et al. (2015)) publicaron un metaanálisis sobre los efectos del *feedback* en los resultados del aprendizaje cuando se aplica mediante entornos de aprendizaje por ordenador. Se trata de un estudio bastante exhaustivo que construye en buena medida sobre los trabajos, referencia en la materia, de Hattie y Timperley (2007) y Shute (2008), a la vez que centra la atención en un contexto más concreto. Se basa en 40 estudios con resultados que son, normalmente, pero no siempre, favorables al *feedback*, y que son analizados en función de diversos factores. A partir de ese análisis, se especifican como principales conclusiones: (1) que el *feedback* elaborado –categoría en la cual en este estudio se incluyen la respuesta repetida hasta lograr contestar correctamente (AUC) y la explicación de la respuesta correcta– es el que más efecto tiene, seguido del KCR, y que el *feedback* que menor efecto tiene es el KR; (2) que los efectos del *feedback* son mayores en tareas de alto nivel (transferencia) que en tareas de bajo nivel, basadas en la retención; (3) que los efectos son muy importantes en áreas relacionadas con las matemáticas, de un nivel intermedio para áreas de ciencias o de ciencias sociales, y pequeño para áreas relacionadas con los idiomas; (4) que el *feedback* en entornos virtuales tiene una mayor influencia en alumnos adultos que en niños/adolescentes; y (5) que en general se obtuvo mayor efecto con el *feedback* inmediato que con el diferido.

Se ha visto anteriormente (sec. 2.2.1) cómo la carga cognitiva puede afectar al aprendizaje de forma importante y cómo esta carga es más fácil que afecte a alumnos sin experiencia en la materia que están tratando, debido a que no cuentan con un esquema previo que pueda guiar la adquisición e integración de nueva información. De la misma manera, en la sección 2.2.3, al revisar la investigación referente al uso de pódcast de vídeo en el aprendizaje, se ha resaltado que el hecho de que la información se presente en los vídeos de forma transitoria hace del vídeo un medio propicio para el surgimiento de situaciones de sobrecarga cognitiva. Pues bien, la aportación de *feedback* al alumno puede ayudar a prevenir posibles situaciones de sobrecarga reduciendo la carga cognitiva, si se ha diseñado con ese objetivo (Shute (2008)). En Moreno (2004), mediante dos experimentos, se obtiene que, para este propósito de reducir la carga cognitiva, resulta más beneficioso un *feedback* elaborado que un *feedback* que cuente exclusivamente con características correctivas, en alumnos sin experiencia cuando el material de aprendizaje es un juego interactivo multimedia. En general, los distintos estudios refuerzan la idea de que el *feedback* elaborado es superior al resto de tipos de *feedback*, aunque con ciertos matices. Por ejemplo, en Letón et al. (2017) se compara un *feedback* KCR con un EF en forma de vídeo en un curso de estadística, resultando mejor el *feedback* elaborado. En Petrović et al. (2017), en el contexto de un curso de procesamiento de señales digitales, se analiza la interacción entre los distintos tipos de *feedback* y la dificultad, encontrando que, independientemente de la dificultad, cualquier tipo de *feedback* es mejor a no usar ninguno (con un efecto notable para este dominio, mayor cuanta más dificultad tengan las preguntas) y que el EF, en forma de explicación de la respuesta correcta, es mejor que el KCR cuando la dificultad es alta. Por último, en Attali y van der Kleij (2017) se estudian los distintos tipos de *feedback* en función de los resultados obtenidos en cuestiones similares –mismo concepto y estructura– a las planteadas en la evaluación formativa efectuadas con posterioridad. La conclusión es que los alumnos que se ven expuestos al EF –una vez más en forma de explicación de la respuesta correcta– obtienen mejores resultados solamente cuando fallaron inicialmente la pregunta de la evaluación formativa, lo cual sería indicativo de su falta inicial de conocimiento o entendimiento de la materia.

Dadas las interacciones del *feedback* y su forma concreta con la dificultad de la materia y el nivel previo del alumno, un avance lógico es plantearse la posible adaptación automática de la evaluación formativa y del *feedback* a la condición y gustos del alumno mediante la aplicación de mecanismos de modelado del estudiante. No se trata de una idea nueva, ya que los denominados «sistemas de tutorización inteligente» (*intelligence tutoring systems, ITS*) tienen la adaptación de las actividades educacionales al nivel del alumno –entre ellas el feedback– como uno de sus objetivos fundamentales (Brusilovsky y Millán (2007)). Estudios recientes que abordan este aspecto, con el *feedback* adaptativo como objetivo central son, por ejemplo, Basu et al. (2017) donde se propone un *framework* para conseguir un soporte o andamiaje

(*scaffolding*) adaptativo o Tacoma et al. (2017) en el que se propone y experimenta un modelado del estudiante para la entrega individualizada de *feedback* en un curso introductorio de estadística.

2.3.3. El uso de preguntas embebidas en los vídeos

Una forma muy común de practicar la evaluación formativa en el contexto de los MOOCs es el uso de preguntas de carácter formativo, con su *feedback* correspondiente, al final de los vídeos (*post-video quizzes*) o de forma embebida en el propio vídeo (*in-video quizzes*). Sin embargo, a pesar de su extenso uso, se trata de una práctica poco estudiada y no hay demasiada información sobre la forma más beneficiosa de realizar esta evaluación formativa relacionada con los vídeos.

En Cummins et al. (2016) se detallan las características que deben tener las *in-video quizzes*, en contraposición con las *post-video quizzes*:

- Deben aparecer durante la reproducción del vídeo, pausándose éste automáticamente.
- Deben mostrarse en el momento apropiado que generalmente será después de la explicación de un concepto.
- Deben aportar *feedback* al alumno para que éste pueda tomar una acción correctiva lo antes posible, si es necesario.

Los (pocos) artículos que hacen referencia a preguntas embebidas se centran en estudiar, a partir de los *logs* de interacción de los cursos, el comportamiento de los alumnos con respecto a las preguntas o los vídeos, o en si su uso influye en el compromiso (*engagement*) de los alumnos con respecto a la tasa de abandono de los vídeos o al número efectivo de preguntas contestadas (Cummins et al. (2016); Kovacs (2016)). Sin embargo, no se compara el aprendizaje entre vídeos con preguntas embebidas, situaciones en que se realizan las preguntas después del vídeo o la ausencia de cuestiones formativas y, cuando se hace referencia a este hecho, es comparando los resultados obtenidos con vídeos diferentes. De hecho, en Cummins et al. (2016) se plantea esta línea como propuesta de investigación futura. Solamente se ha podido localizar un artículo, Vural (2013), donde se realiza una medición del impacto en el aprendizaje de intercalar preguntas entre clips de vídeo respecto a no realizar preguntas, mediante una configuración cuasiexperimental. Un aspecto importante en este estudio es que el grupo que disfrutaba del material con preguntas no podía ver el siguiente clip de vídeo sin haber contestado primero de forma correcta a las preguntas, restricción que no tenía el otro grupo que podía navegar libremente por los vídeos. Los alumnos que usaron el entorno formativo con preguntas obtuvieron un mejor resultado en el aprendizaje.

En Kovacs (2016) se analiza cómo cambia la forma de ver el vídeo por parte de los alumnos cuando existen preguntas embebidas, pasando de una visualización lineal del vídeo a una visualización no lineal, donde se hace un mayor uso de los controles de manejo del vídeo. Por ejemplo, los alumnos realizan 55 veces más búsquedas en el contenido anterior del vídeo desde una pregunta embebida que desde otros puntos del vídeo y en muy raras ocasiones avanzan el vídeo saltándose una cuestión. Además, los alumnos que comienzan a ver un vídeo con preguntas embebidas tienden a ver un mayor porcentaje del vídeo por lo que el uso de preguntas embebidas podría ayudar a paliar el abandono de los vídeos.

Hay otra publicación, Wachtler et al. (2016), que estudia el efecto de la situación de las preguntas dentro de los vídeos y del intervalo entre ellas. En él también se comparan los resultados de una clase con metodología "*flipped classroom*", usando los vídeos con las preguntas embebidas, y otra clase con metodología normal (sin vídeos). En la comparación, usando un examen estándar que no comprende solamente los temas tratados en los vídeos, se obtiene que la clase que usa los vídeos consigue unos resultados mejores a los de la otra clase, siendo la diferencia significativa.

3. Diseño experimental y formulación de hipótesis

El presente capítulo está dedicado a exponer el diseño concreto del experimento realizado –comparación, mediante un experimento aleatorio, entre el uso de *in-video quizzes* vs. *post-video quizzes* en el contexto de un MOOC real– y los trabajos relativos al diseño y construcción del MOOC que da soporte al mismo. Se comienza exponiendo las características básicas del experimento realizado y las cuestiones relacionadas con el curso que se consideraron que podrían afectar de manera más potente al resultado del experimento. Se continúa con la formulación de hipótesis en relación con el diseño experimental y con la literatura revisada en el capítulo 2. El siguiente apartado trata diversos aspectos relacionados con el diseño e implementación de los contenidos del curso propiamente dichos, entre otros se tratan: la motivación para la elección de la materia del curso –una introducción básica al aprendizaje automático para todos los públicos–, el enfoque del curso, las lecciones y conceptos tratados en cada una, el diseño y grabación de los vídeos, la «virtualización» del curso o la configuración de la plataforma para la realización del experimento. A continuación, se describen los trámites realizados para su puesta en marcha en UNED Abierta y las actividades dedicadas a la difusión del vídeo, y se describe la forma de acceder al MOOC y se aportan enlaces a los vídeos desarrollados para el mismo.

3.1. Diseño experimental

En el diseño del experimento se contemplan dos grupos de alumnos, cada uno de los cuales visualiza ciertos contenidos del curso, los relacionadas con los vídeos, de forma diferente. Uno de los grupos –grupo de control– visualiza los vídeos completos y después de algunos vídeos cuenta con preguntas de refuerzo relacionadas con el contenido de cada vídeo. El otro grupo –grupo experimental– visualiza exactamente los mismos vídeos, pero para presentar las preguntas de refuerzo no se espera a finalizar el vídeo, sino que se presentan al finalizar la explicación del concepto tratado en el vídeo que tenga relación con la pregunta. Ambos grupos ven las mismas preguntas de refuerzo, pero en diferente momento y, salvo por esta diferencia, todo el resto del curso es idéntico para los dos grupos.

En este punto es necesario explicar que la plataforma Open edX no soporta de forma nativa, a día de hoy, la configuración de cuestiones embebidas o incrustadas en el vídeo como tal, de forma que el vídeo se pare automáticamente en una marca de tiempo y en su lugar se presente una pregunta o ejercicio, y que, una vez contestada ésta, se continúe con la reproducción del vídeo. Para conseguir este efecto es necesario configurar una extensión al producto –lo que en la plataforma Open edX se denominan *Xblocks*– desarrollada por la universidad de Stanford para este fin[1]. Desgraciadamente, por motivos administrativos ajenos a la propia UNED Abierta, fue imposible conseguir la instalación de esta extensión en la plataforma por lo que no se pudo llevar a cabo el plan original, que consistía en presentar al grupo experimental preguntas *embebidas* en el vídeo, y fue necesario optar por presentar las cuestiones *intercaladas* entre segmentos del vídeo en lugar de *embebidas* como tal. Este aspecto se trata posteriormente en el apartado 3.3.4 que trata la virtualización del curso, donde se puede ver un ejemplo exacto de cómo se presentan las preguntas al grupo experimental, y en el capítulo 5, al tratar las limitaciones del estudio.

Las «preguntas de refuerzo» consisten en preguntas orientadas a asegurar la comprensión de los conceptos tratados en el vídeo y a su aplicación a contextos diferentes, es decir, son preguntas de transferencia. El alumno tiene la posibilidad de contestarlas cuantas veces desee y con cada contestación se le presenta el correspondiente *feedback* que le indica si la respuesta ha sido correcta o no y, en caso de serlo, una explicación de la respuesta correcta. Esto es, el *feedback* utilizado en el experimento es del tipo KCR (AUC) + EF en forma de explicación de la respuesta correcta. Como se ha visto en el capítulo 2, este *feedback* es sobre todo apropiado cuando la materia es compleja y el nivel del alumno bajo.

La evaluación de los resultados sobre el aprendizaje se realiza mediante un test al finalizar cada una de las lecciones del MOOC. Cada uno de esos test se puede contestar solamente en una ocasión y consiste en diez preguntas, de las cuales seis están diseñadas para evaluar un aprendizaje memorístico –«preguntas de recuerdo»– y las otras cuatro a la aplicación de lo aprendido en contextos diferentes –«preguntas de transferencia»–. Al finalizar el curso se presenta al alumno un test de valoración del MOOC que permite extraer conclusiones sobre la percepción subjetiva del conjunto de estudiantes en cada grupo. Además, al comienzo del curso se realizan dos test que sirven para analizar la equivalencia de ambos grupos. El primero se ha denominado «calentando motores» y, debido a que el MOOC tiene cierta carga de matemáticas, tiene además el objetivo que el alumno tenga la seguridad de que cuenta con las habilidades matemáticas mínimas necesarias para seguir el curso. El alumno puede contestar este test tantas veces como considere necesario de forma que si al realizar el test se da cuenta de que no recuerda algún concepto, puede repasarlo y volver a realizar el test. El segundo test que se realiza al inicio del curso es

[1] xblock-in-video-quiz: https://github.com/Stanford-Online/xblock-in-video-quiz o su reciente *fork*: https://github.com/raccoongang/edx-xblock-in-video-quiz

un test de conocimientos previos que, igual que los test de cada lección, solamente se puede contestar en una ocasión.

Usar un MOOC real como instrumento para la realización de un experimento aleatorio es una empresa un tanto arriesgada, dado que no es posible saber a priori cuánta gente se apuntará al mismo. Para poder llevar a buen término el experimento, era necesario el uso de un MOOC que interesara a un número suficiente de personas y que preferiblemente pudiera ser realizado por la mayor parte de la población, además de finalizado en un corto espacio de tiempo –esto es, un «minicurso»–, con el fin de minimizar en lo posible la tasa de abandono y así conseguir un tamaño muestral apropiado que aportara el suficiente poder estadístico. Por otro lado, para tratar de maximizar el efecto de la posición de las preguntas de refuerzo, y de acuerdo con la evidencia empírica (Shute (2008); van der Kleij et al. (2015) y otros), era deseable que los alumnos fueran adultos en su mayor parte, la materia sobre la que versara el MOOC tuviera cierta complejidad, preferiblemente estuviera relacionada con las matemáticas o las ciencias, y que el conocimiento previo de los alumnos sobre la misma fuera limitado.

Aunque, de acuerdo con la documentación de Open edX, parecía perfectamente posible contar con dos condiciones diferentes para dos grupos de alumnos distintos, e, incluso, parecía que el sistema se encargaba de realizar la asignación de cada persona a cada uno de los grupos experimentales, para tener mayor seguridad, al tratarse de un aspecto crítico, se realizó una prueba de concepto en una instalación local de Open edX (versión *ficus*) que resultó satisfactoria. Aunque se verá con más detenimiento posteriormente, reseñar que con la configuración elegida la asignación aleatoria de cada alumno a uno de los grupos se realiza la primera vez que éste accede a los contenidos del curso.

3.2. Formulación de hipótesis

Se considera que la utilización de preguntas de refuerzo embebidas en el vídeo, con respecto a la realización de estas mismas preguntas nada más finalizar el vídeo, puede influir de distintas maneras tanto directamente sobre la capacidad del alumno de procesar la información y de construir un modelo mental correcto de la materia, como en la motivación y el compromiso, y, por tanto, a su vez nuevamente sobre el aprendizaje.

En primer lugar, el uso de preguntas embebidas es una evolución de los mecanismos de interactividad en los vídeos que, recordemos, permiten a los alumnos ejercer una autorregulación en el procesamiento de la información y un aprendizaje más activo a través de vídeos (Mayer y Chandler (2001); Zhang et al. (2006)). Al realizar una pregunta de refuerzo, con un *feedback* apropiado, en el mismo instante en que se termina de explicar un concepto, se permite al alumno comprobar si ha comprendido

correctamente y corregir concepciones erróneas cuanto antes. Como sucede en Kovacs (2016), se espera que las preguntas embebidas provoquen un mayor uso de los sistemas de control de los vídeos, es decir, un aumento del uso del autocontrol por parte del alumno. Aunque ambos grupos reciben exactamente el mismo *feedback*, y la diferencia entre recibir la evaluación formativa mediante una pregunta embebida o mediante una pregunta posterior al vídeo es solamente de minutos, se espera que el *feedback* tenga mayor influencia cuando permite al alumno tener la seguridad de que su comprensión de un concepto es correcta antes de pasar al siguiente concepto, el cual, además, en muchas ocasiones depende del anterior. Dado que estamos hablando sobre todo de «comprensión», se espera que el efecto sea más acusado en los niveles altos del aprendizaje.

En segundo lugar, tal y como se ha observado en Cummins et al. (2016) y en Kovacs (2016), la tasa de abandono de los vídeos se retrasa cuando el vídeo dispone de preguntas embebidas, lo cual puede indicar que el uso de este tipo de preguntas tiene un efecto positivo en la motivación general del alumno. De la misma manera, Ryan y Deci (2000) nos dicen que potenciar la autonomía y la sensación de competencia del alumno mejora su motivación intrínseca, algo que hipotetizamos se consigue mejor con las *in-video quizzes* que con las *post-video quizzes* al permitirles autorregularse de manera más eficiente.

Por último, la interpolación de preguntas en los vídeos, dado que paran su flujo hasta que el alumno contesta a la pregunta, provocan *de facto* una segmentación del mismo (Mayer y Moreno (2003)), lo cual, como ya se ha visto en el capítulo 2, permite al alumno reflexionar sobre la información y procesarla adecuadamente, reduciendo la carga cognitiva provocada por el dinamismo del vídeo a la hora de presentar la información. Otro efecto que puede ayudar a reducir la carga cognitiva del alumno es el de la señalización que provoca la aparición de la pregunta embebida. Esta señalización es tanto temporal, permitiendo al alumno diferenciar más fácilmente la explicación de un concepto de la explicación del siguiente, como sirviendo de aviso sobre la importancia de un determinado concepto sobre el que se pregunta. La investigación relacionada con ambos principios –segmentación y señalización– nos dice que son sobre todo relevantes cuando la materia es compleja y el nivel de los alumnos en esa materia es bajo, como es el caso.

Para modular todos estos efectos potenciales, es necesario tener en cuenta el entorno elegido: un MOOC. Los MOOCs tienen unas características especiales que pueden dificultar la detección de ciertos efectos: (1) los alumnos interaccionan con el curso como más les pueda interesar para la consecución de sus propios objetivos; (2) la experiencia de aprendizaje de un MOOC no se circunscribe a los vídeos sino que habitualmente se cuenta con material complementario y con la participación e interacción en los foros, ambos, aspectos que se mantienen para este experimento; y (3) en los MOOCs no se suele limitar el material que los alumnos tienen disponible a la hora de realizar los test calificados, es decir, es muy común que los estudiantes

revisen el material audiovisual o sus propias notas mientras están contestando a las preguntas de los test calificados. Tampoco se establece limitación alguna en el caso que nos ocupa.

Derivadas de los anteriores razonamientos, se establecen las siguientes hipótesis para el estudio:

H_1: Intercalar preguntas de refuerzo tendrá un efecto positivo en la disminución de la tasa de abandono global del curso.

Debido a los efectos que pueden tener las preguntas intercaladas sobre una mejor capacidad de autorregularse, una mayor interactividad con el vídeo y una mayor sensación de autonomía, se espera que la motivación de los alumnos pertenecientes al grupo con preguntas embebidas sea mayor, y por ello, que estos alumnos mantengan un mayor compromiso con los vídeos de las distintas lecciones y, por extensión, que su tasa de abandono del curso sea menor.

H_2: Intercalar preguntas de refuerzo en los vídeos tendrá un efecto positivo sobre los resultados de aprendizaje, tanto en recuerdo como en transferencia, siendo el efecto mayor sobre los resultados de transferencia.

Al resultar un mecanismo que se estima que permite reducir la carga cognitiva, una mejor autorregulación en el procesamiento de la información por parte del alumno y la mayor facilidad para la construcción de un modelo mental correcto y de su integración con el conocimiento previo, el intercalar preguntas de refuerzo en los vídeos debería tener un efecto positivo en el aprendizaje. Este efecto se entiende que debería ser mayor para niveles altos de aprendizaje –transferencia– ya que las preguntas de refuerzo van dirigidas a ese nivel, que las preguntas intercaladas facilitan la construcción de un modelo mental coherente y que la literatura afirma que la segmentación de los vídeos es especialmente útil para niveles altos de aprendizaje.

H_3: Intercalar preguntas de refuerzo en los vídeos tendrá un efecto positivo en la percepción de los alumnos sobre el curso.

Derivado de la potenciación de la motivación y de la capacidad de autonomía y de la sensación de competencia, junto con la esperable mejora en el aprendizaje, se espera que la percepción sobre el curso del grupo que trate con las preguntas intercaladas sea mejor que la del grupo que cuenta con las preguntas después del vídeo.

3.3. Diseño y desarrollo de los contenidos del MOOC

3.3.1. Motivación para la realización de un MOOC sobre aprendizaje automático para todos los públicos

Al tratar de decidir el tema concreto que se desarrollaría para el MOOC se tuvieron en cuenta, en primer lugar, los requisitos de que la materia contara con cierta com-

plejidad, estuviera relacionada con las matemáticas o las ciencias y que estuviera dirigido a personas que no contaran con experiencia previa. Es decir, debido a que existía la posibilidad de que el potencial efecto de las preguntas intercaladas en el vídeo se viera diluido por el uso de un MOOC real, en el que el alumno cuenta con más información que la del propio vídeo y la de las preguntas de refuerzo y en el que además cuenta con la libertad propia de este tipo de cursos, se buscaba realizar el experimento en un contexto propicio tanto para el uso del tipo de *feedback* elegido como para el posible impacto en el aprendizaje de la segmentación producida en los vídeos por la inclusión de preguntas intercaladas.

Además, el curso debía resultar atractivo para un gran número de personas para conseguir obtener una muestra lo suficientemente grande. Estos condicionantes llevaron a decidir que una introducción al aprendizaje automático, eliminando las partes matemáticas complejas de forma que lo pudiera realizar cualquiera con habilidades matemáticas de secundaria, podría ser la elección oportuna. Pero no es que se considerara solamente que se trataba de la opción óptima para el estudio experimental, se consideraba, y aún se considera, que, debido a todo el ruido existente alrededor de la «IA» en la sociedad en general, **se trata de un curso necesario** para el que en aquel momento no parecía existir oferta alternativa (ni siquiera en inglés). Por otro lado, cada vez en el ámbito de más dominios se están tratando de realizar aproximaciones al uso del aprendizaje automático, pero, en muchos casos, las empresas u organismos no cuentan con personal especializado que posea tanto el conocimiento del negocio como ciertos conocimientos en el área de la Inteligencia Artificial que les permita dirigir el proceso o que, al menos, les permita comunicarse con expertos externos sin sentirse totalmente perdidos. Por todo esto, se estimó que el curso debía tratar de cumplir con cuatro objetivos:

1. Servir de introducción al aprendizaje automático para todo tipo de alumnos, incluso aquellos que posteriormente tuvieran pensado realizar un curso más avanzado en la materia.
2. Desmitificar la «IA» tal y como se puede estar entendiendo por la sociedad. Tratar de que se comprenda el estado actual de la técnica y que el alumno adquiera la capacidad de diferenciar lo que ahora mismo es realidad de la (todavía) ciencia ficción.
3. Abordar cuáles son los retos sociales reales y actuales que tiene la aplicación en general de las técnicas de aprendizaje automático.
4. Que los alumnos comprendan el proceso de construcción de los modelos de aprendizaje, por qué es así y que adquieran la capacidad de comunicarse con los expertos en el tema.

Estos objetivos han configurado un curso, titulado finalmente «*Entendiendo la Inteligencia Artificial: los fundamentos básicos al alcance de todos*», que se encuentra a caballo

Figura 3.1.: Presentación del curso en UNED Abierta

entre la mera divulgación y un curso introductorio al aprendizaje automático para personas más especializadas, en el cual se tratan temas divulgativos, como por ejemplo qué es un algoritmo o qué son las redes neuronales, hasta temas más complejos como el compromiso sesgo-varianza, pasando por el impacto del sesgo en los datos o el problema de la falta de equidad de los modelos.

Durante la fase de diseño y construcción de este curso –que comenzó en enero de 2018– empezaron a aparecer otras propuestas con un objetivo muy similar como son «*Elements of AI*», de la Universidad de Helsinki y la empresa Reaktor, cuya primera parte comenzó el 14 de mayo de 2018, o «*AI for Everyone*» que se puso en marcha en Coursera el 28 de febrero de 2019. La aparición de estos cursos de introducción a la IA orientados también a un público no técnico, fue un espaldarazo a la idea de que, efectivamente, el presente curso era una buena idea y de que se trataba de una oferta necesaria.

3.3.2. Estructura y contenidos del curso

El curso diseñado está compuesto de: (1) seis lecciones, cada una de las cuales cuenta con uno o dos vídeos, material escrito complementario y, con la excepción de la primera lección, preguntas de refuerzo y una evaluación calificativa formada por diez preguntas –seis de recuerdo y cuatro de transferencia–; (2) de una sección preliminar, denominada «antes de comenzar», en la cual se da la bienvenida a los alumnos, se explica el funcionamiento del curso y se realizan los test «calentando motores» y «conocimientos previos»; y (3) de una sección final en la cual se posibilita al alumno realizar la valoración del curso y se le facilitan ciertos recursos interesantes para continuar aprendiendo, como por ejemplo los másteres relacionados con la materia que ofrece la UNED y otros MOOCs más avanzados que el presente. En total, para el curso se han desarrollado nueve vídeos con una duración total de más de dos horas (detalles en tabla 3.2), setenta y cinco cuestiones entre las preguntas de los distintos test –excluyendo el test de valoración– y las preguntas de refuerzo y un texto ilustrado de acompañamiento equivalente a setenta páginas de Word.

A continuación se detallan los contenidos y características concretas de cada lección.

Lección 1: ¿Inteligencia Artificial? Aclarando conceptos

Se trata de una lección de introducción, que tiene como objetivo aclarar qué es eso de la inteligencia artificial y poner en contexto el aprendizaje automático como una de las ramas de la IA relacionándola con otros términos muy utilizados (entre otros: *big data*, ciencia de datos o analítica predictiva). Se trata de aclarar también el concepto de algoritmo, para que el alumno sepa que no se trata de un concepto nuevo y que no sólo son «algoritmos» aquellos que son generados con «IA», y de explicar la gran ventaja que nos aportan las técnicas de aprendizaje automático cuando no es posible, o es muy difícil, escribir el algoritmo a mano. Es una lección con un contenido bastante sencillo, sin carga matemática, que cuenta con un vídeo que no tiene asociado ni test de evaluación, ni preguntas de refuerzo y que, por tanto, es idéntica para ambas cohortes.

Lección 2: Aprendizaje automático

Una vez que se ha puesto en contexto qué es el aprendizaje automático y cuáles son sus ventajas, la lección 2 se encarga de explicar qué se entiende por aprendizaje en este contexto y cuáles son las principales clases de aprendizaje existentes (supervisado, no supervisado y por refuerzo), haciendo un especial hincapié en el aprendizaje supervisado, para el que se ven en detalle las diferencias entre regresión y clasificación mediante el uso de la regresión lineal y de la regresión logística,

Figura 3.2.: Extracto de las preguntas 1 y 2 del test «Conocimientos previos»

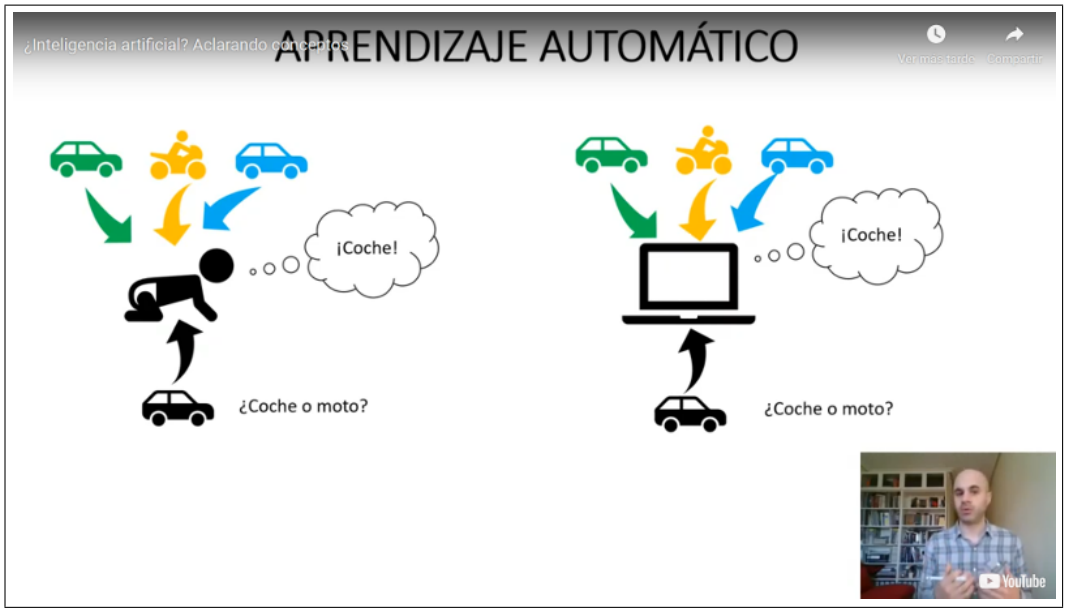


Figura 3.3.: Imagen del vídeo «¿Inteligencia Artificial? Aclarando conceptos» de la lección 1

con un ejemplo de cada una. Dado que el curso pretende que el alumno comprenda conceptualmente las bases de las técnicas pero ocultando la carga matemática, el aprendizaje del valor de los parámetros se trata como una caja negra y los ejemplos (conjunto de datos de velocidad y distancia de frenado para regresión y conjunto de datos de lirios para clasificación) se centran en que se comprenda que es posible tratar de hacer que un ordenador «aprenda» a realizar predicciones a partir de la existencia de un conjunto de datos etiquetados.

Es la lección con una mayor carga matemática del curso y probablemente una de las que tienen una mayor complejidad habida cuenta del público potencial del curso. Contiene dos vídeos, uno centrado en la regresión y otro en la clasificación, cada uno con una pregunta de refuerzo.

Lección 3: Construcción de un modelo de aprendizaje

En la lección 2 no se entra en cuál es el proceso que permite aprender a partir de los datos etiquetados, ni se menciona que el resultado debe ser validado con ejemplos «limpios». Por ello, esta tercera lección se centra en explicar que el propósito del aprendizaje supervisado es poder anticipar la etiqueta de datos nuevos, que todavía no se han visto. Se introduce el concepto de conjunto de datos de prueba y cuál es el proceso de construcción de un modelo, explicando, de manera informal, en qué consiste el entrenamiento de un modelo. Se finaliza con una introducción

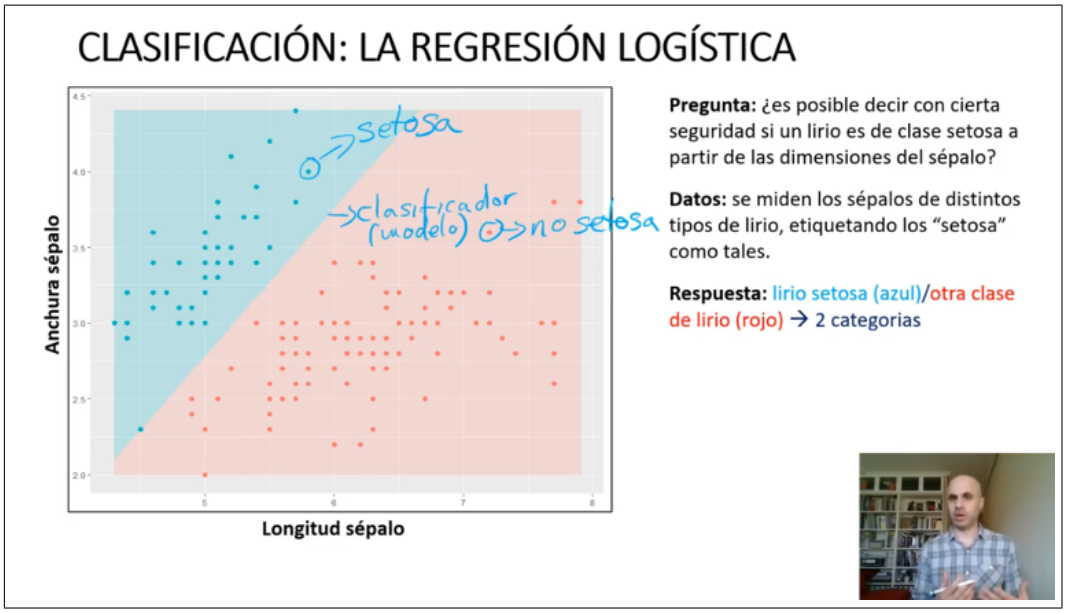


Figura 3.4.: Imagen del vídeo «Aprendizaje supervisado: introducción a la clasificación» de la lección 2

básica a la evaluación de clasificadores, tratando conceptos como los tipos de error –falsos positivos y falsos negativos–, la matriz de confusión y con una revisión de las métricas más utilizadas.

Es una lección que se considera de dificultad media, con bastante seguridad más sencilla que la anterior, y que cuenta con dos vídeos, el primero centrado en el proceso iterativo de generación de modelos de aprendizaje en el cuál, además, se hace una demostración de entrenamiento y predicción utilizando la herramienta «H2o Flow»[2] y el segundo dedicado a explicar la matriz de confusión y las métricas de exactitud, precisión y sensibilidad, y el compromiso entre estas dos últimas. El primer vídeo no cuenta con preguntas de refuerzo y el segundo cuenta con dos preguntas.

Lección 4: Redes neuronales y aprendizaje profundo

En la lección 4 llega el momento de abordar los problemas no linealmente separables, que son necesarios clasificadores mucho más complejos para resolverlos y se introducen las redes neuronales como la técnica más en boga actualmente para conseguir este tipo de clasificadores. Se comienza explicando el concepto de perceptrón y los cálculos que realiza, posteriormente se introduce la neurona sigmoidea

[2]https://www.h2o.ai/, descarga desde http://h2o-release.s3.amazonaws.com/h2o/latest_stable.html

Si pensamos en la superficie de esta gráfica con dos parámetros (w, b) y el valor del error como si de un paisaje con montañas y valles se tratara y nos imaginamos a nosotros mismos en uno de los picos (zonas rojas) con el objetivo de llegar lo más rápidamente posible a la zona más baja de los alrededores (zona azul, a ser posible), ¿cómo actuaríamos? Lo más fácil para empezar sería comenzar a bajar en aquella dirección en que la pendiente es más pronunciada y cada cierto tiempo ir cambiando la dirección que hemos tomado para continuar en aquella dirección en que el descenso es más rápido hasta que llegue un momento en que el terreno vuelva a ascender de forma importante. Ese es exactamente el proceso que sigue el descenso del gradiente salvo que en lugar de montañas trabaja con una función (que normalmente tendrá muchos más de dos parámetros) y para calcular la pendiente utiliza matemáticas.

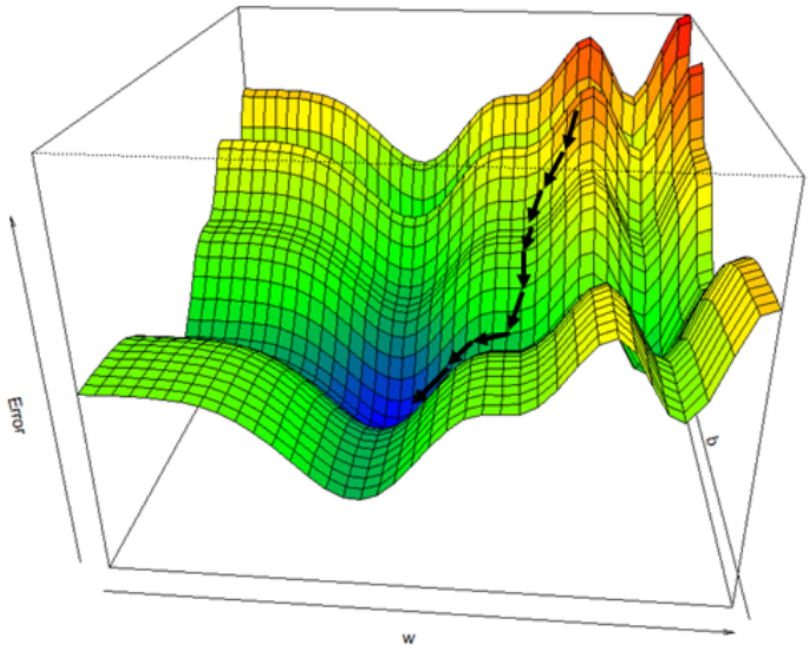


Figura 3.5.: Extracto de la explicación escrita del proceso de entrenamiento de la lección 3

y, por último, la conexión de múltiples neuronas en redes. Solamente se nombra la retropropagación, sin ahondar en qué consiste, y se pone el foco en el número de parámetros que tiene una red como forma de valorar la complejidad de las hipótesis que puede generar la red –anticipando lo que en la siguiente lección se ve en mayor profundidad, el sobreajuste y el compromiso sesgo-varianza–. Por último, se trata el concepto de aprendizaje profundo como una red con un gran número de capas ocultas.

Una vez vistas las redes neuronales se propone una pequeña práctica utilizando el *Playground* de *TensorFlow*[3] que es una web que permite entrenar distintas redes neuronales en el navegador para resolver algunos problemas simples. Es muy visual porque permite seguir de forma dinámica el proceso de entrenamiento y la actualización de los pesos y probar distintas arquitecturas, con distinto número de capas y de neuronas, además de modificar la mayoría de los principales parámetros de la red, desde la función de activación hasta el tipo de regularización.

Esta lección cuenta con un sólo vídeo centrado en explicar el perceptrón y en cómo calcular su salida a partir de las entradas, y en las redes neuronales. Tiene dos preguntas de refuerzo. Se considera que este tema tiene una dificultad similar a la anterior (media) ya que solamente se usan matemáticas para ver ejemplos muy simples con el perceptrón y en ningún momento se tratan los aspectos de entrenamiento de una red.

Lección 5: Selección de modelo y análisis del error

Se trata, probablemente, junto con la lección 2, de la lección más compleja del curso. El motivo de esta lección es exponer, aunque sólo sea de manera intuitiva, algunos de los conceptos principales que guían el proceso de desarrollo de los modelos de aprendizaje, como son: la selección de modelo –introduciendo los conceptos de conjunto de datos de validación y la validación cruzada–, el problema del sobreajuste y del compromiso existente entre la capacidad de aproximación y la capacidad de generalización de un modelo, la descomposición del error en error debido al sesgo y en error debido a la varianza y cómo influyen en el comportamiento del error tanto la complejidad del modelo como el número de datos disponibles para realizar el entrenamiento. Se menciona el concepto de regularización y se ve cómo es posible utilizar todo lo anterior para obtener un conjunto de recomendaciones que permitan, a partir del análisis del error que comete un modelo, saber qué opciones son las más optimas a probar en la siguiente iteración.

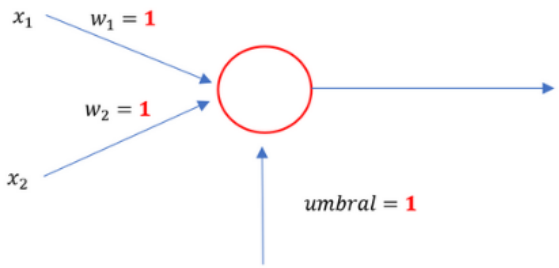
Esta lección cuenta con un único vídeo que tiene dos preguntas de refuerzo y se centra, sobre todo, dado lo abstracto de esos conceptos, en tratar de explicar el compromiso entre aproximación y generalización y la descomposición del error.

[3] http://playground.tensorflow.org/ creada por Daniel Smilkov y Shan Carter

Pregunta 6
¿Cuándo se empiezan a explorar las posibilidades de las neuronas artificiales?

- ◯ Antes de 1970
- ◯ En los inicios del siglo XXI
- ◯ En los años 90
- ◯ A partir de 1980

Pregunta 7
A partir del siguiente perceptrón



¿Cuáles deben ser los valores de x_1 y x_2 para que la neurona se active?

- ◯ $x_1 = 0; x_2 = 0$
- ◯ $x_1 = 1; x_2 = 0$
- ◯ $x_1 = 0; x_2 = 1$
- ◯ $x_1 = 1; x_2 = 1$

Figura 3.6.: Extracto del test calificado de la lección 4. La pregunta 6 orientada a evaluar la retención y la 7 la capacidad de transferencia.

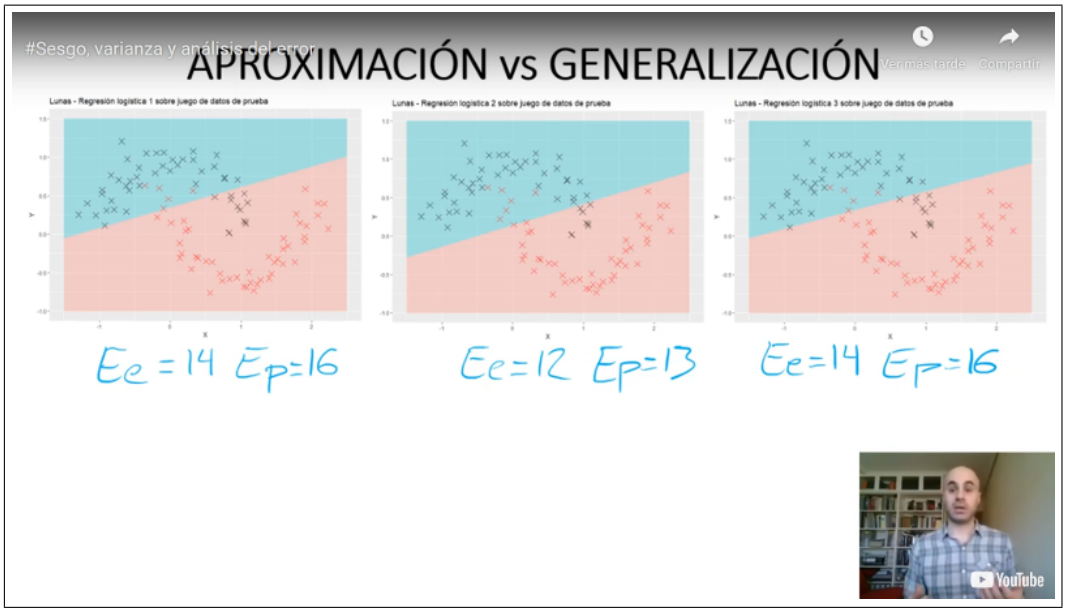


Figura 3.7.: Imagen del vídeo «Sesgo, varianza y análisis del error» de la lección 5

Lección 6: Consideraciones sociales de la IA

Esta última lección del MOOC es una lección que se considera fundamental debido a que, en muchas ocasiones, da la impresión de que en la sociedad en general se debate en exceso sobre potenciales efectos perversos de la inteligencia artificial que, si bien puede que en un futuro se conviertan en realidad, a día de hoy suenan todavía a ciencia ficción, mientras que se relegan a un segundo plano, o directamente no se tratan, otros efectos muy importantes que ya hoy son una realidad o, al menos, lo pueden ser en un futuro muy próximo. Estamos hablando fundamentalmente del problema de los sesgos en los datos con los que se entrenan y validan los modelos; de la potencial falta de equidad o justicia (*fairness*) que pueden tener los modelos con ciertos subgrupos de la población; de la falta de transparencia y capacidad de interpretación por parte de los humanos de los resultados de los algoritmos de aprendizaje más potentes; de cómo afectan estas nuevas capacidades a la privacidad de las personas; o del potencial impacto de la IA en el mercado de trabajo.

Esta lección se ha abordado de manera un tanto diferente, tratando, fundamentalmente, de plantear y definir la existencia de estos retos de forma que el alumno pueda construir su propia opinión informada, ya que en estas cuestiones aún existen más preguntas que respuestas. Además, se consideró interesante incluir un aspecto que no se suele tratar y que se considera que toda persona que tenga relación con sistemas que apliquen técnicas de aprendizaje automático –e, idealmente, todo

	Nombre	*Dificultad estimada*	*# vídeos*
Lec. 1	¿Inteligencia Artificial? Aclarando conceptos	Baja	1
Lec. 2	Aprendizaje automático	Alta	2
Lec. 3	Construcción de un modelo de aprendizaje	Media	2
Lec. 4	Redes neuronales y aprendizaje profundo	Media	1
Lec. 5	Selección de modelo y análisis del error	Alta	1
Lec. 6	Consideraciones sociales de la IA	Media	2

Tabla 3.1.: Distribución del contenido del MOOC en lecciones y dificultad estimada.

ciudadano– debería conocer: el marco normativo existente en nuestro entorno (el Reglamento General de Protección de Datos) y en cómo afecta a la toma de decisiones automatizada y a la elaboración de perfiles que usen datos personales o que permitan inferir características personales. Se pone especial atención en el artículo 22 del Reglamento y en la prohibición general que expresa (según la interpretación que a día de hoy tiene como oficial el Comité Europeo de Protección de Datos) a realizar una toma de decisiones automática (sin participación humana), incluida la elaboración de perfiles, cuando la decisión resultante pueda tener efectos jurídicos o afectar de forma significativa la vida de las personas[4].

En esta lección hay dos vídeos, el primero dedicado a introducir los problemas del sesgo en los datos, la equidad de los modelos y la falta de capacidad de interpretación, que cuenta con una pregunta de refuerzo. El segundo trata los aspectos relacionados con la normativa y cuenta con otra pregunta de refuerzo. Esta lección se considera de dificultad media, pero es necesario tener en cuenta que la segunda parte de la misma, al tratar los aspectos normativos, tiene unas características notablemente diferentes al resto del curso.

3.3.3. Proceso de creación de los vídeos

En este apartado se describe el proceso seguido para la creación de los vídeos del curso, que es quizás el tipo de contenido que más complejidad tiene producir. Se tratan desde los principios de diseño tenidos en cuenta, hasta las herramientas utilizadas para su grabación y posterior edición.

[4] «Reglamento General de Protección de Datos» (https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:32016R0679&from=ES) y las «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles» (https://edpb.europa.eu/node/71)

Cuestión sobre los datos usados en el entrenamiento vs los datos con los que se aplicará el modelo

1/%(possible) de puntos (no calificados)

Un equipo tiene el encargo de desarrollar un sistema de detección y clasificación de señales de tráfico de limitación de velocidad, destinado a ser montado en un vehículo, de forma que el vehículo adquiera la capacidad de detectar las señales y avisar al conductor. Para generar el modelo que clasifique las señales, el equipo dispone de una base de datos enorme de imágenes de España con y sin señales de limitación de velocidad, de diferentes carreteras y en diferentes condiciones de luz y atmosféricas.

Si el vehículo va a ser distribuido en el mercado de USA, ¿la forma de proceder está siendo correcta?

Señal de límite de velocidad con el formato común en Europa.
Autor: administración del Reino Unido, Open Government Licence v1.0

Señal de límite de velocidad con el formato común en USA.
Fuente: Manual on Uniform Traffic Control Devices

- ◯ No, no es posible generar un modelo tan avanzado como ese
- ◯ SI, si tenemos muchos datos para entrenar el modelo, debería ser algo factible
- ◉ No, el detector de señales probablemente no detecte las señales de USA ✔
- ◯ SI, siempre que la distribución de los datos entre los conjuntos de entrenamiento, validación y pruebas se realice de forma correcta

Respuesta
Correcto:
Si se entrena al modelo para detectar y clasificar una serie de señales, el modelo aprenderá a clasificar esas señales y no otras. Si se aplica el modelo con señales que tengan otras características, como son las que se utilizan en USA, el modelo no será capaz de clasificarlas.

Figura 3.8.: Pregunta de refuerzo (y *feedback* correspondiente) asociado del vídeo «Sesgo, equidad, interpretabilidad y aprendizaje automático» de la lección 6

#	*Título vídeo**	*Lección*	*Duración (# pregs.)*	*momento preguntas***
1	¿Inteligencia Artificial? Aclarando conceptos	1	15:40 (0)	
2	Aprendizaje supervisado: introducción a la regresión	2	12:22 (1)	9:41
3	Aprendizaje supervisado: introducción a la clasificación	2	11:05 (1)	9:30
4	Proceso de construcción de un modelo de aprendizaje	3	17:51 (0)	
5	Evaluación de un modelo de clasificación	3	10:57 (2)	3:07/9:46
6	Introducción a las redes neuronales	4	13:31 (2)	5:39/12:19
7	Sesgo, varianza y análisis del error	5	17:00 (2)	10:08/15:45
8	Sesgo, equidad, interpretabilidad y aprendizaje automático	6	11:42 (1)	5:21
9	La elaboración de perfiles de personas y el Reglamento General de Protección de Datos	6	13:09 (1)	10:44

*El título es un enlace al vídeo en YouTube. Para ver las preguntas de refuerzo es necesario acceder a la plataforma del curso (https://iedra.uned.es)

**Solamente para el grupo experimental, el grupo de control tiene las preguntas de refuerzo al finalizar el vídeo.

Tabla 3.2.: Vídeos del curso (y enlace a YouTube) con su duración, el número de preguntas de recuerdo asociadas y el momento en que aparecen.

3.3.3.1. Principios de diseño y metodología

Los principios de diseño que se han tenido en cuenta para el desarrollo de los vídeos son los establecidos en el concepto de Minivideos Docentes Modulares (MDMs) explicados en Letón y Molanes-López (2014). Fundamentalmente se han tenido en cuenta los siguientes aspectos:

- Que los vídeos fueran lo más cortos posible. Idealmente deberían estar entre 5 y 10 minutos, aunque esto es algo que, como queda patente en la tabla 3.2, a pesar del esfuerzo de síntesis realizado, no se ha podido conseguir. Es más, en un principio, estaba previsto que cada lección del curso contara con un sólo vídeo, idea que fue necesario modificar posteriormente debido a la desmesurada extensión de los mismos.
- Que los vídeos fueran autocontenidos. Es decir, deben explicar el concepto que traten de forma completa, sin depender de otros vídeos o material, ni hacer referencia a ellos. Esto permite reutilizar los vídeos en otros contextos diferentes. Dado que en los vídeos no se hace referencia a los conceptos tratados con anterioridad, la asociación de los distintos conceptos que se ven a lo largo del curso se ha realizado mediante el texto de apoyo.
- Cada diapositiva debería tener espacio suficiente para escribir, para albergar los subtítulos y para superponer una ventana con la imagen del instructor.
- El vídeo debe contar un resumen del contenido al finalizar el mismo.
- Por último, debido a su mejor legibilidad (Rodriguez-Ascaso et al. (2018)), se decide usar un color azul para escribir en las diapositivas.

Respecto a la metodología, se han seguido las recomendaciones de de la Fuente Sánchez et al. (2013) que identifica las siguientes fases para el diseño de los vídeos:

- Selección de los contenidos y temas a tratar. Con especial atención a aquellos conceptos con mayor dificultad para el alumno o que se consideran básicos.
- Identificación de las características de la población a la que va destinado el vídeo.
- Concretar los medios técnicos que se usarán.
- Para cada vídeo: definir los conceptos que tratará, preparar el guion, preparar las transparencias necesarias, ensayar la grabación y realizar la grabación.

Este último punto es un proceso iterativo, con un refinado sucesivo, y, sin duda, así lo fue para el diseño y la grabación de los vídeos de este curso, cuyos contenidos, la forma de explicarlos y presentarlos y la propia grabación dieron lugar a múltiples versiones diferentes de los vídeos.

Además de estas recomendaciones, se han tratado de seguir también los consejos contenidos en el MOOC de edX «*Creating Video for the edX Platform*», en especial las recomendaciones relativas a iluminación y sonido.

3.3.3.2. Medios técnicos utilizados

Para la grabación de los vídeos se han utilizado los siguientes medios:

- Ordenador portátil con Windows 10 y MS PowerPoint (versión 1902 en Office 365), que ha sido el software utilizado para realizar la grabación como tal mediante la *webcam*, con un ratón para realizar el paso de las animaciones y de las diapositivas.
- Apple Pencil y iPad 6ª generación con la *app* EasyCanvas, la cual, mediante la conexión del iPad al PC por USB, hace posible escribir sobre las diapositivas.
- Ordenador portátil secundario para hacer las veces de *teleprompter* mediante el software Imaginary TelePrompter sobre Linux Ubuntu.
- Teclado inalámbrico para pausar, volver poner en marcha y manejar el ritmo de avance del *teleprompter*.
- Aplicación Shotcut 18.05 para la edición final de los vídeos.
- Aplicación Captura 7.0.1 para la captura de pantalla. Utilizada para la parte de demostración del vídeo «Construcción de un modelo de aprendizaje».
- Micrófono Bird UM1 con soporte trípode.

3.3.3.3. Proceso de diseño y grabación

Como ya se ha dicho, el proceso de diseño y grabación fue un proceso de refinado continuo. Fruto de la inexperiencia en esta tarea, se diseñó gran cantidad de material que posteriormente debió ser descartado dada la gran extensión que adquirían los vídeos –resulta difícil darse cuenta *a priori* de las pocas cosas que pueden contarse en diez minutos–. Para la elaboración de los vídeos se dieron los siguientes pasos:

1. **Determinación de los conceptos a tratar en el vídeo:** tratando de que los vídeos resultaran lo más modulares posible y de incidir especialmente en los aspectos más complejos o que se consideraron fundamentales de cada lección.

2. **Diseño de las diapositivas:** Para diseñar las diapositivas que formarían parte de los vídeos se utilizó, como primer paso, un diseño en papel, dividiendo una hoja de papel tamaño A4 en cuatro cuadrantes y diseñando en cada cuadrante

Figura 3.9.: Imagen del «estudio» de grabación.

una diapositiva. Este espacio reducido evita que se trate de llenar las diapositivas con excesiva información. Además, se puso especial atención en dejar suficiente espacio para poder escribir ciertos contenidos sobre las diapositivas y de usar imágenes como forma de mantener la atención de los alumnos. Al trasladar el diseño a PowerPoint, se buscó hacer un uso apropiado de las animaciones, utilizándolas cuando se les viera un recurso didáctico interesante y siempre como apoyo y sincronizadas con la explicación verbal.

3. **Elaboración del guion:** Una vez elaboradas las diapositivas, se creó un guion específico para cada vídeo para mostrarlo en el *teleprompter.* Este guion no contiene sólo toda la explicación hablada, sino que además se añadieron etiquetas con información sobre cuándo avanzar las animaciones, pasar de diapositiva y cuándo y qué escribir sobre las diapositivas. Aunque el guion ayuda mucho a la hora de tener claro qué explicar, de no olvidarse nada importante y de no llenar el vídeo de titubeos sobre lo que se está explicando, se ha percibido que el resultado es más natural si se usa el guion como referencia, pero permitiéndose ciertas libertades sobre el mismo.
   Un ejemplo de guion elaborado es el siguiente:

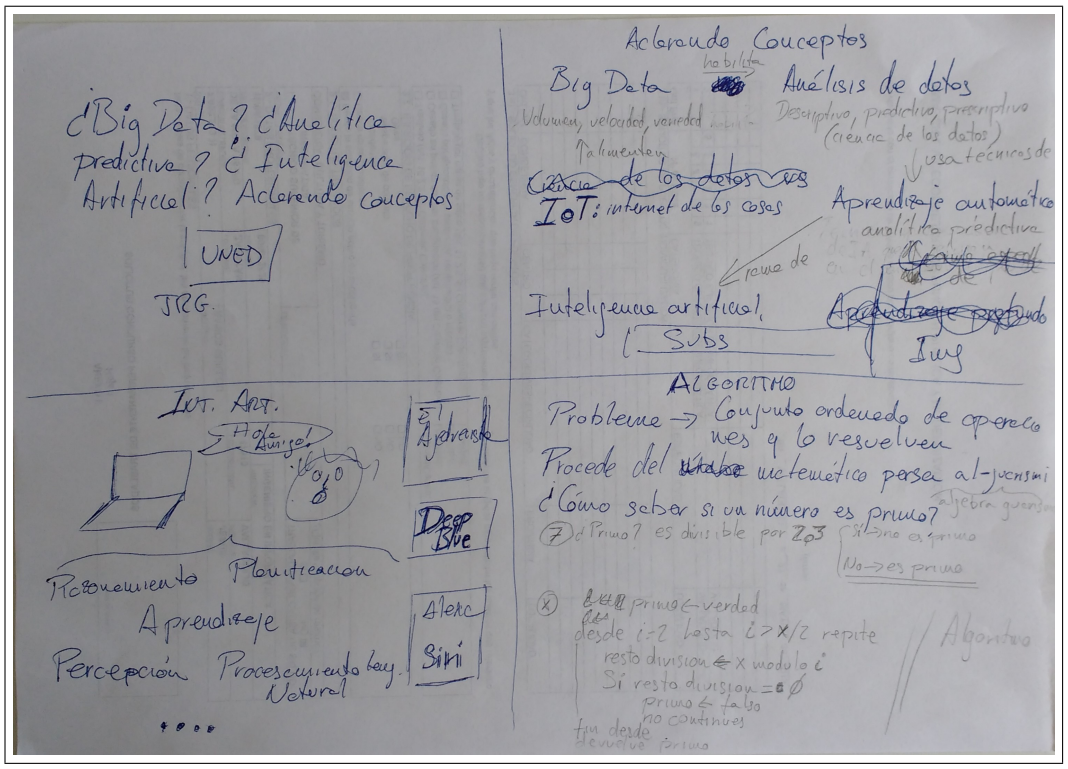


Figura 3.10.: Primera versión del boceto de las diapositivas iniciales del vídeo «¿Inteligencia Artificial? Aclarando conceptos».

```
[...] Vamos a empezar usando los tres juegos de datos de entrenamiento
para generar una regresión logística para cada uno. <click>
Vemos que cada uno de los tres modelos generados a partir de los
tres conjuntos de datos distintos es bastante parecido. Es decir,
el resultado no depende demasiado del conjunto de datos concreto
que tenemos disponible. Sin embargo, vemos a simple vista que no
son capaces de ajustar a los datos disponibles, algo normal en este
caso, dado que con la regresión logística solamente podemos obtener
separaciones lineales. Vamos a analizar qué error estamos cometiendo
en cada uno de los casos. En el juego de datos de más a la izquierda,
el clasificador comete 14 errores <pausa><escribir "Ee=14"><continuar>,
el segundo clasificador, en el centro, comete 12 fallos
<pausa><escribir "Ee=12"><continuar> y el de la derecha comete,
como el primero, 14 fallos <pausa><escribir "Ee=14"><continuar>.
Más allá de la posibilidad que tenemos de verlo gráficamente en este
ejemplo al estar trabajando solamente con dos dimensiones, esta es
la medida que nos indica que los modelo no están siendo capaces de
ajustar a los datos de entrenamiento todo lo que nos gustaría.
Vamos a ver qué tal lo hace cada modelo sobre los datos que hemos
reservado para probar <click> [...]
```

4. **Grabación y edición del vídeo:** Para realizar las grabaciones se ha utilizado la opción de PowerPoint «grabar presentación con diapositivas» que permite grabar una presentación superponiendo, si así se desea, la imagen de la persona que está realizando la presentación en la imagen de la diapositiva. Además, permite también escribir sobre las diapositivas. Es una opción muy cómoda para realizar las grabaciones puesto que permite tanto hacer la grabación de la presentación completa, como realizarla a partir de una diapositiva determinada, y al finalizar la grabación cada diapositiva conserva la porción de vídeo que le corresponde lo que permite tener flexibilidad a la hora de cambiar el orden de las diapositivas, eliminar diapositivas, pegar diapositivas de otras versiones de grabación o volver a grabar sólo ciertas diapositivas. Solamente es necesario tener precaución y no hablar o moverse entre las transiciones de las diapositivas puesto que entre dos diapositivas siempre se produce un pequeño corte en el vídeo. Así mismo, si se está usando la opción de mostrar la imagen de la cámara, se recomienda fijar una postura a adoptar en los cambios de diapositiva para minimizar la sensación de salto en el vídeo. Además, si se desea contar con toda la flexibilidad que permite el mover, quitar y poner diapositivas ya grabadas, y las grabaciones se van a realizar durante varios días, se aconseja utilizar la misma vestimenta en las distintas grabaciones y tratar de realizar éstas con las mismas condiciones de iluminación.
   Una vez realizadas las grabaciones, éstas se pueden exportar a un fichero de vídeo. En el caso que nos ocupa, una vez exportado, se realizó una edición del vídeo con la aplicación gratuita Shotcut con el fin de suavizar ciertos saltos entre diapositivas, recortar ciertos trozos de vídeo y, en definitiva, tratar de mejorar la calidad de los mismos.

Figura 3.11.: Pantallazo de la opción de PowerPoint «grabar presentación con diapositivas».

5. **Carga en YouTube y subtitulado:** la plataforma Open edX consume los vídeos desde YouTube por lo que el siguiente paso fue cargarlos en esta plataforma de vídeos. Aprovechando que era necesario utilizar YouTube, se hizo uso de las herramientas disponibles en YouTube Studio para la generación de los subtítulos. En concreto, YouTube permite cargar un texto y sincronizarlo con el sonido del vídeo, convirtiendo el texto a formato de subtítulos con las marcas de tiempo incorporadas. Para la sincronización se reutilizaron los guiones de cada uno de los vídeos, desprovistos previamente de las etiquetas con la información sobre cómo actuar durante la grabación. Una vez que se disponía de la versión preliminar de los subtítulos, se realizó una revisión y corrección manual de todos ellos para enmendar los errores de sincronización –francamente pocos–, y para adecuar los subtítulos a lo que realmente se decía en el vídeo ya que en múltiples ocasiones no se siguió el guion al pie de la letra. Obtenidos la versión final de los subtítulos, se descargaron en formato «srt» y se cargó cada uno de estos ficheros en la plataforma Open edX puesto que la plataforma cuenta con su propia forma de presentar las transcripciones de los vídeos que, además, permite navegar en el vídeo haciendo clic en el texto transcrito.

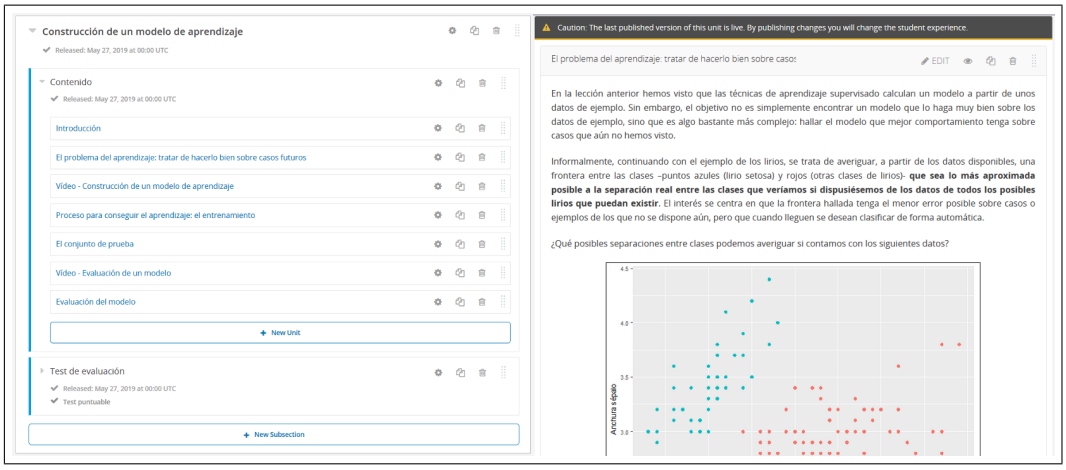


Figura 3.12.: Muestra de la interfaz de la herramienta Studio. A la izquierda estructura de la lección «construcción de un modelo de aprendizaje», a la derecha contenido (texto ilustrado) de una de sus unidades.

3.3.4. Virtualización de los contenidos y configuración para el experimento

Una vez que se dispuso de todos los contenidos, el siguiente paso fue construir el curso como tal en la plataforma Open edX de UNED Abierta. Para ello, previamente se solicitó el permiso formal de realización del curso. Los contenidos en la plataforma se cargan, de forma bastante fácil e intuitiva, a través de la herramienta web Studio. Los cursos en Open edX se estructuran en secciones, subsecciones y unidades, por lo que se hizo corresponder cada una de las lecciones planteadas para el MOOC con una sección que contuviera dos subsecciones: «contenido» y «test de evaluación» (salvo para el caso de la lección 1, que tiene una sola sección al no disponer de un test de evaluación). En la subsección «contenido» de cada lección se virtualizaron, organizados en unidades, los diferentes contenidos del curso: texto ilustrado, vídeos y preguntas de refuerzo.

Probablemente la parte con más interés en cuanto a la configuración del curso sea cómo se ha realizado la configuración para que cada una de las cohortes vea los contenidos de manera diferente. Existen un par de opciones en Open edX para conseguir este efecto, una mediante lo que se llama *«Content Experiments»* y, la otra, habilitando lo que en edX denominan *cohortes*. Ambas opciones tenían la posibilidad de satisfacer las necesidades que tenía este experimento por lo que se optó por la que se antojó más sencilla: el uso de cohortes. Las cohortes son grupos de usuarios que tendrán alguna característica diferente en el curso, habitualmente cierto contenido. La asignación de los estudiantes a cada cohorte se puede hacer de manera manual, automática o híbrida, habiendo elegido para el presente MOOC la opción automáti-

ca, que asigna de forma aleatoria a cada estudiante a una de las cohortes existentes la primera vez que accede al curso. Se crearon dos cohortes, una para cada grupo experimental y, para configurar las dos maneras diferentes de ver las unidades que contuvieran los vídeos y las preguntas de refuerzo, se usó lo que se denomina «grupos de contenido», que es el concepto que en Open edX permite definir que distintos alumnos vean distinto contenido. Se crearon dos grupos de contenido, uno para el grupo de control y otro para el grupo experimental, y cada uno de los grupos se asignó a la cohorte correspondiente de manera que los estudiantes asignados a una cohorte quedaran automáticamente incluidos en el grupo de contenidos. Como último paso, se configuró cada componente de vídeo de manera apropiada asignando su visibilidad únicamente a uno u otro grupo de contenido. La figura 3.13 muestra cómo ve cada grupo una unidad con vídeo y pregunta de refuerzo. En el caso del grupo de control se tiene un único componente para el vídeo completo y, debajo de éste, otro u otros, dependiendo del número de preguntas, para mostrar las preguntas de evaluación formativa. En el caso del grupo experimental la unidad contiene más de un componente de vídeo –configurados para mostrar tramos de éste (se configura en la plataforma el punto del vídeo de comienzo y de fin de reproducción)–, tantos como sean necesarios para albergar entre los tramos del vídeo los componentes de las preguntas de refuerzo.

3.3.5. Pruebas, difusión y puesta en marcha del MOOC

Con el curso ya construido se planificó su puesta en marcha. Se estimó que 3 semanas era un tiempo más que suficiente para realizar el curso con una dedicación de 4 horas semanales, por lo que se planificó la apertura del curso para el día 27 de mayo de 2019 y se optó por mantener la atención tutorial, con contestación de dudas en los foros del curso, durante las tres primeras semanas (hasta el día 16 de junio). A su vez, a partir de un guion que se preparó para ese objetivo, el CEMAV compuso un fantástico vídeo promocional en el que también se contó con la participación del profesor Félix de la Paz a través del robot Nao[5]. El vídeo puede verse en la página inicial del curso en UNED abierta[6] o bien, directamente, mediante el siguiente enlace de YouTube: https://youtu.be/gPFeCbURrqM.

Llegado este punto, se inició de forma paralela una fase de pruebas *beta* del curso y la fase de difusión del MOOC a través, fundamentalmente, de redes sociales y de los foros de la UNED, con el objetivo de captar alumnos.

Las pruebas las ejecutaron dos personas que no sabían nada del MOOC y que no contaban con conocimientos sobre la materia con el fin de pulir posibles errores de

[5]Como aclaración, mi participación en este vídeo promocional se circunscribe exclusivamente al guion del mismo.

[6]https://iedra.uned.es/courses/course-v1:UNED+InteligArtific_001+2018/about

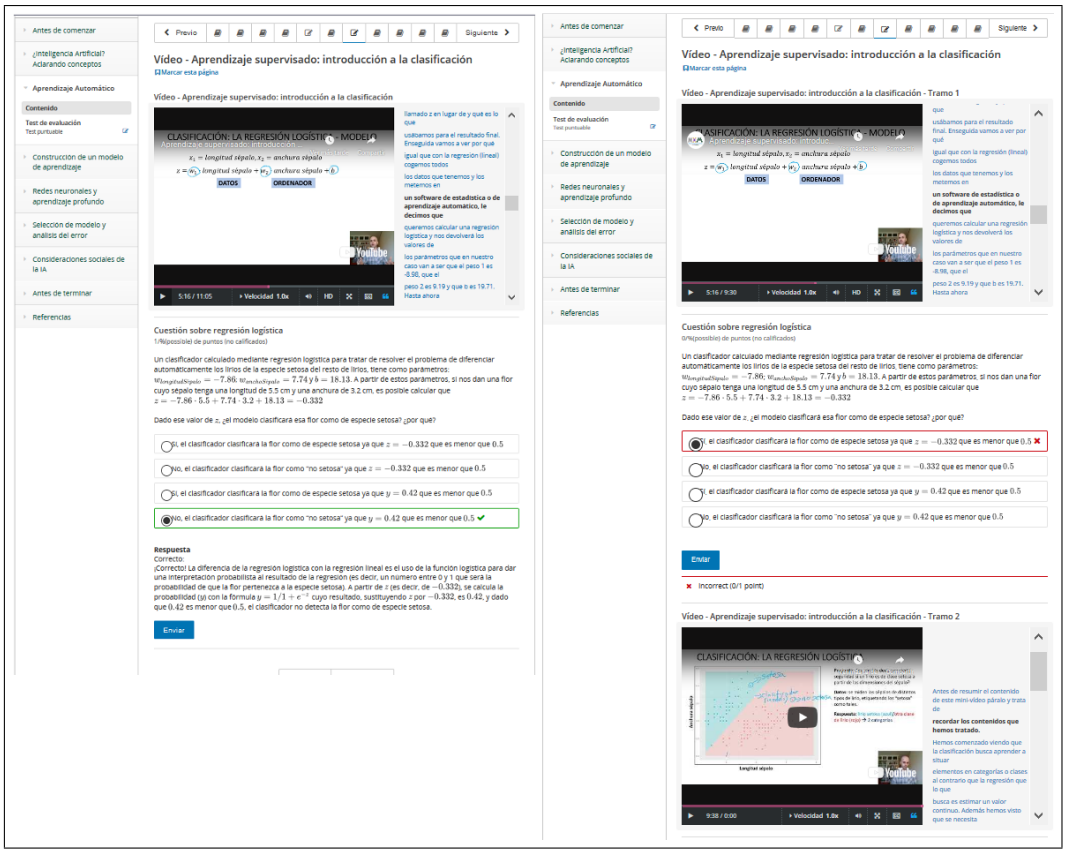


Figura 3.13.: Comparación de la vista de la unidad correspondiente al vídeo «Aprendizaje supervisado: introducción a la clasificación» en el grupo de control (izquierda) y en el grupo experimental (derecha).

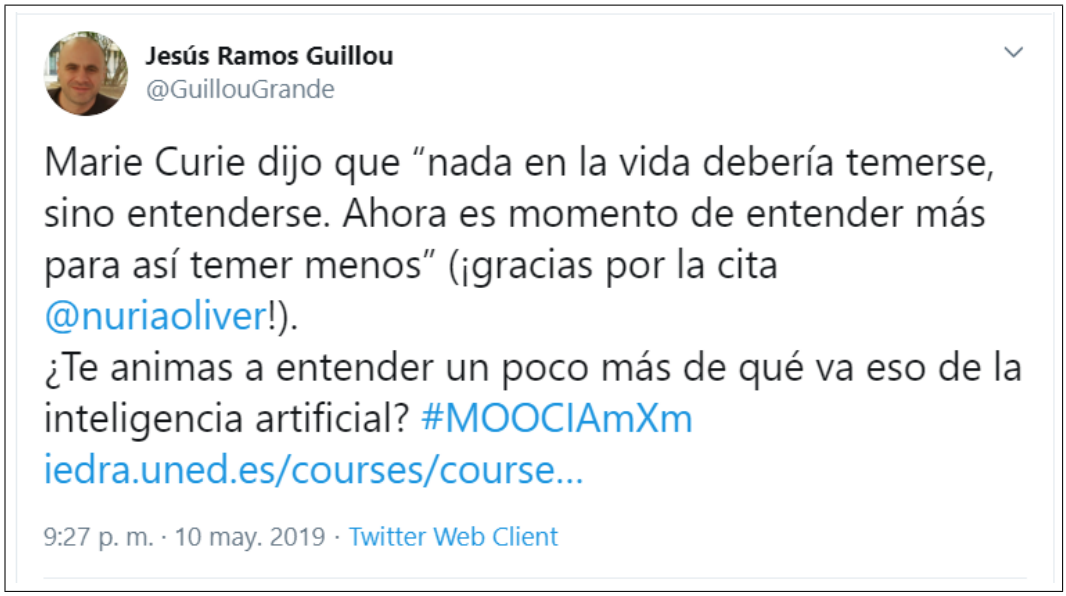


Figura 3.14.: Uno de los tuits usados para tratar de difundir el curso.

contenido, de las preguntas o de la configuración del curso. Una de ellas realizó el curso completo como si fuera un alumno más mientras que la otra se centró en comprobar aspectos más típicos de pruebas funcionales a través de la interfaz de usuario. Esta fase sirvió para corregir algunos errores, la mayoría de escasa importancia, y para tener cierta seguridad de que un alumno sin experiencia previa podía seguir el curso completo (la persona que hizo el curso completo obtuvo 48 puntos sobre 50 posibles, revisando que los errores no fueron provocados por errores en las preguntas o en el material del curso y comprendiendo el origen de los mismos).

Respecto a la difusión, la mayor parte del esfuerzo se centró en la red social Twitter, aunque también se difundió el curso en círculos próximos y a través de los cursos de la UNED. Se escogió utilizar la etiqueta #MOOCIAmXm como forma de agrupar los distintos tuits y en la difusión, además de usar la cuenta del autor como se puede ver en la figura 3.14, se contó con la inestimable ayuda de Emilio Letón (@emilioleton), de UNED Abierta, a través de la cuenta @UNEDAbierta, así como de @IA_UNED e, incluso, de la cuenta oficial de la UNED (@UNED).

3.3.6. Acceso al curso

Para poder acceder al MOOC, que estará activo hasta el día 15 de octubre, se puede entrar a la plataforma Open edX de UNED Abierta (https://iedra.uned.es) y buscar el curso –se recuerda el título: «Entendiendo la inteligencia artificial: los fundamentos básicos al alcance de todos»–, o bien usar el siguiente enlace directo:

Fase	*Fecha inicio*	*Fecha finalización*	*Esfuerzo (horas)*
Diseño del curso y virtualización contenidos	01/01/2018	30/11/2018	276
Preparación, grabación vídeos, subtitulado y configuración en la plataforma	01/12/2018	10/03/2019	168
Pruebas, correcciones y difusión	11/03/2019	26/05/2019	27
Ejecución curso	27/05/2019	16/06/2019	23
Total			**494**

Tabla 3.3.: Distribución por fases del esfuerzo en la realización del curso.

https://iedra.uned.es/courses/course-v1:UNED+InteligArtific_001+2018/about.

Para tener acceso es necesario disponer de una cuenta en la plataforma o bien crear una nueva. El acceso al curso de esta manera será con el rol de un alumno normal, y, por ello, en el primer acceso se producirá la asignación de esa cuenta a una de las dos cohortes del curso, debido a lo cual se verá el curso bien como un alumno del grupo de control (viendo todos los vídeos al completo y apareciendo en algunos casos preguntas posteriormente), o bien como un alumno del grupo experimental (ciertos vídeos, en los que hay preguntas, se verán separados en tramos con preguntas de refuerzo entre los segmentos del vídeo).

Es posible acceder a YouTube para ver los vídeos fuera de la plataforma pinchando en el título de cada uno de los vídeos en la tabla 3.2. Debe tenerse en cuenta que accediendo de esta manera solamente se tiene acceso al vídeo y que no se verá ningún tipo de pregunta.

4. Ejecución del experimento, resultados y puesta en contexto

En este capítulo se detallan los aspectos relacionados con la ejecución del experimento, con la extracción de datos de la plataforma Open edX y preprocesado de datos, con el análisis de los datos y presentación de resultados, y con discusión de los resultados en relación con las hipótesis planteadas.

4.1. Ejecución del experimento, extracción y preprocesado de datos

4.1.1. Ejecución del experimento

El experimento tuvo lugar entre los días 27 de mayo de 2019, fecha de apertura del MOOC, y el día 20 de julio de 2019, fecha de recogida de los datos. El curso se planteó de manera que la mayor parte de las personas interesadas lo realizaran en tres semanas, desde el 27 de mayo hasta el 16 de junio. Para conseguirlo, se publicitó ese calendario en la información del curso, se realizó una planificación específica para esas tres semanas, se comunicó que la atención tutorial a través de los foros del curso (común para ambas cohortes) se daría únicamente durante ese tiempo y se enviaron correos semanales animando a los alumnos y recordándoles cuál era el objetivo para esa semana.

Una vez finalizada la atención tutorial, se decidió retrasar durante algún tiempo la recogida de datos para permitir que un mayor número de personas finalizara el curso y así aumentar el tamaño muestral. Finalmente, la extracción de datos se realizó el día 20 de julio de 2019, fecha en la que habían completado el curso 70 personas, 35 del grupo experimental y 35 del de control.

4.1.2. Extracción y preprocesado de datos

Para preparar el análisis de datos, se descargaron desde la plataforma Open edX algunos de los distintos ficheros en formato «csv» que la plataforma ofrece a los instructores. En concreto, se utilizaron dos tipos de ficheros:

Figura 4.1.: Ejemplo de correo enviado a los alumnos durante el curso.

1. **Informe denominado «informe de calificación»:** contiene la información de todos los estudiantes inscritos en el curso, de la cohorte asignada y de la puntuación obtenida en los test puntuables del curso. Se utilizó para extraer la información del grupo al que pertenecía cada alumno.
2. **Informe de las respuestas a un problema dado (*«student state»*):** en la interfaz de Open edX, introduciendo el código identificador de cada test o cuestión, es posible descargarse un informe detallado de las respuestas de cada alumno para una prueba concreta. Se descargó este fichero para todos los test del curso: «calentando motores», «conocimientos previos», los test de cada una de las lecciones y la valoración del curso. El informe consiste en un fichero en formato csv que contiene en uno de sus campos una estructura en formato json con distinta información sobre las respuestas del alumno, como, por ejemplo, la corrección o incorrección de cada una de las preguntas o cuál ha sido la respuesta del alumno a cada una de ellas.
   En la figura 4.2 se puede ver un ejemplo del json de un alumno correspondiente al test «calentando motores» (que consta de cinco preguntas), resumido para mostrar las partes utilizadas para extraer los datos usados en la fase de análisis del presente trabajo. Este json se ha procesado para extraer la puntuación de cada alumno para cada uno de los test (fueran puntuables para el curso o no).

```
{
    "correct_map":{
        "029c2dd77ef84a34a79a2a74d6269fe2_2_1":{
            "correctness":"correct",
        },
        "029c2dd77ef84a34a79a2a74d6269fe2_3_1":{
            "correctness":"correct",
        },
        "029c2dd77ef84a34a79a2a74d6269fe2_4_1":{
            "correctness":"incorrect",
        },
        "029c2dd77ef84a34a79a2a74d6269fe2_5_1":{
            "correctness":"correct",
        },
        "029c2dd77ef84a34a79a2a74d6269fe2_6_1":{
            "correctness":"correct",
        }
    },
    "student_answers":{
        "029c2dd77ef84a34a79a2a74d6269fe2_2_1":"choice_0",
        "029c2dd77ef84a34a79a2a74d6269fe2_3_1":"choice_1",
        "029c2dd77ef84a34a79a2a74d6269fe2_4_1":"choice_3",
        "029c2dd77ef84a34a79a2a74d6269fe2_5_1":"choice_0",
        "029c2dd77ef84a34a79a2a74d6269fe2_6_1":"choice_1"
    }
}
```


Figura 4.2.: Ejemplo (retocado para eliminar las partes no utilizadas) del json que incluye el *«student state»* de un alumno, correspondiente al test «calentando motores».

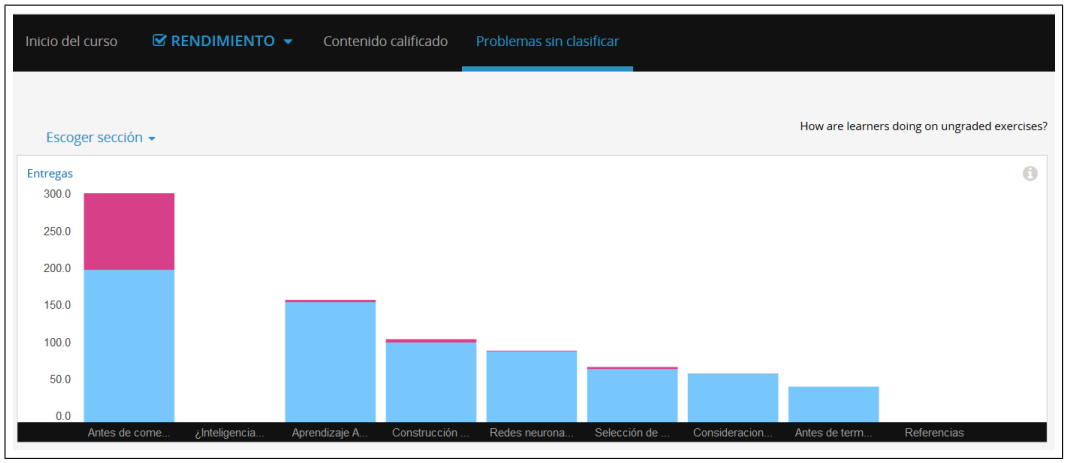


Figura 4.3.: Detalle de gráfica que muestra edX Insights.

Todo el preprocesado de datos se ha realizado mediante R, utilizando el paquete «jsonlite» para procesar el json, con el objetivo de obtener un sólo conjunto de datos en el que constaran todas las puntuaciones de los distintos alumnos para los distintos test y el grupo experimental al que pertenecían, de forma que se facilitara el posterior análisis.

Para finalizar este apartado, unos comentarios acerca de otras posibles fuentes de datos que se pueden llegar a utilizar en otros experimentos:

- UNED Abierta cuenta con una plataforma donde se pueden ver distintas gráficas interesantes a partir de información agregada. Es el Open edX Insights. Esta web muestra información de, entre otros aspectos: las características de los alumnos, procedencia de éstos, vídeos visualizados y durante cuánto tiempo, o resultados obtenidos en las distintas pruebas. La información agregada se puede descargar, pero **no contiene la información de la cohorte**. Esto, unido con que al tratarse de información agregada no es posible eliminar los datos de los usuarios de prueba, hace que estos datos sean difícilmente utilizables para un análisis comparativo.

- Open edX registra muchos de los eventos de interacción de los alumnos con la plataforma. Especialmente, aquellos relacionados con los vídeos –por ejemplo, los eventos *play, pause, seek, stop*–. Esta información no está, en principio, disponible de una manera sencilla para los investigadores en el entorno disponible en la UNED, y se considera que contar con esta información es fundamental para abordar cualquier estudio avanzado acerca de cómo interaccionan los alumnos con la plataforma. Por ejemplo, como se resalta en el capítulo de conclusiones (capítulo 5), sería muy interesante utilizar esta información para comprobar si hay diferencias entre los grupos en la tasa de abandono de

cada vídeo o en si el tiempo total de visualización de los vídeos es distinto. Ejemplos de estudios que hacen uso de estos datos los podemos encontrar, por ejemplo, en Guo et al. (2014); Kovacs (2016) o en Brinton et al. (2016).

4.2. Análisis de datos

Este apartado detalla el análisis de los datos extraídos del experimento. Por mayor claridad, se recuerdan aquí las hipótesis de partida:

H_1: Intercalar preguntas de refuerzo tendrá un efecto positivo en la disminución de la tasa de abandono global del curso.

H_2: Intercalar preguntas de refuerzo en los vídeos tendrá un efecto positivo sobre los resultados de aprendizaje, tanto en recuerdo como en transferencia, siendo el efecto mayor sobre los resultados de transferencia.

H_3: Intercalar preguntas de refuerzo en los vídeos tendrá un efecto positivo en la percepción de los alumnos sobre el curso.

4.2.1. Metodología

Para analizar los datos resultantes del experimento se plantea realizar las siguientes comparaciones entre los resultados del grupo de control y del grupo experimental:

- Para cerciorarnos de que ambos grupos son similares, se realiza una comparación de los resultados de ambos grupos en los test "calentando motores" y "conocimientos previos". Para realizar esta comparativa se usará un test-t independiente, si se cumplen las asunciones de normalidad y homocedasticidad –que se evaluarán mediante test de Shapiro-Wilk y Levene, respectivamente– o un test de Mann-Whitney-Wilcoxon en caso de que las asunciones no se cumplan.
- Para comparar los datos de la tasa de abandono se realiza una comparativa del número de personas de cada cohorte que realiza el test de conocimientos previos con respecto al número de personas que realiza el último test evaluado del curso, el test de la lección 6. Para ello se realizará un test Chi-cuadrado. La medida del efecto se calculará mediante la razón de productos cruzados (*odds ratio*).
- Si la tasa de abandono fuera significativamente diferente entre los dos grupos, estaríamos ante un supuesto de posible *sesgo de selección* (Bareinboim et al. (2014); Hernán y Robins (2019)) ya que la propia condición experimental interferiría con la probabilidad de finalizar el experimento con posterioridad a la asignación aleatoria de cada alumno a una cohorte determinada. Ante este

supuesto, para analizar el impacto de la condición sobre el aprendizaje podría resultar interesante controlar por el grado de compromiso (*engagement*) como *proxy* de distintas posibles variables latentes, principalmente de la motivación. Sin embargo, como hemos visto en la revisión bibliográfica, todo apunta a que parte del efecto que podrían tener las preguntas embebidas sobre el aprendizaje se podría deber precisamente al efecto de éstas sobre la motivación y, por ello, controlar por el compromiso podría, teóricamente, tener el efecto de cancelar parte de esa la influencia. En cualquier caso, al no disponer de datos precisos sobre las interacciones de los alumnos con la plataforma, se descarta ningún tipo de análisis en este sentido.

- Los resultados sobre el aprendizaje se analizan realizando varias comparaciones entre los resultados, tanto de los resultados globales como de cada test, y, para ello, el procedimiento a aplicar será el mismo que en la comparación inicial de ambos grupos: se usará un test-t independiente, si se cumplen las asunciones de normalidad (test de Shapiro-Wilk) y homocedasticidad (test de Levene), o un test de Mann-Whitney-Wilcoxon en caso de que las asunciones no se cumplan. La medida del efecto se calculará mediante el coeficiente de correlación de Pearson, r.

En todos los casos se presentarán los intervalos de confianza al 95 % y se toma un nivel de significación del 5 % para todos los test. El análisis se lleva a cabo mediante el software de estadística R.

4.2.2. Comparando el punto de partida de los dos grupos

Antes de realizar la comparación de los resultados entre los dos grupos, se realiza una comparación sobre los test previos del curso: «calentando motores», que recordamos que es un test muy, muy sencillo que tiene como objetivo que los alumnos autoevalúen si tienen los conocimientos mínimos de matemáticas que les permita seguir el curso con solvencia, y «conocimientos previos» que trata de comprobar el grado de conocimiento previo que tienen los alumnos en la materia. El resultado esperado, producto de la asignación aleatoria a los grupos, es que no haya una diferencia significativa entre ellos.

Resultado del test «calentando motores»

Como se ha descrito en capítulos anteriores, el test «calentando motores» es un test que consta solamente de 5 preguntas, con cuatro respuestas cada una, contando cada respuesta correcta dos puntos, y que puede ser contestado todas las veces que el alumno lo considere oportuno, ya que el objetivo es que se asegure de que recuerda los (pocos) conceptos matemáticos necesarios para seguir el curso.

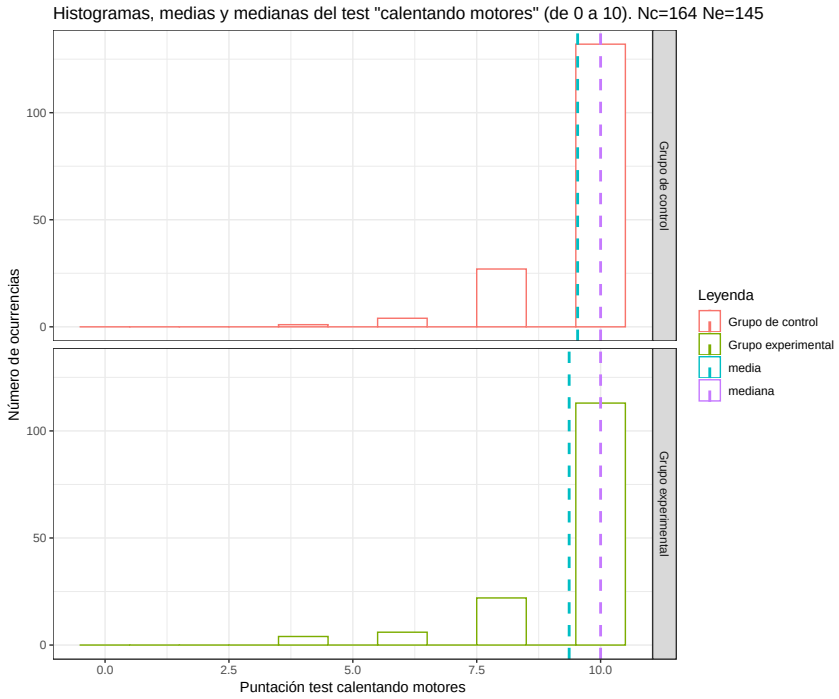


Figura 4.4.: Distribución de resultados en el test «calentando motores»

En la figura 4.4 se representa la distribución de los resultados de la última contestación al test de cada alumno, para cada cohorte. A simple vista podemos ver como la gran mayoría de alumnos es capaz de contestar correctamente a todas las preguntas –comportamiento esperado debido a la simplicidad de la prueba y a la posibilidad de realizar múltiples intentos– y que los resultados de ambos grupos parecen muy similares.

A pesar de que el histograma es claro, se refrenda la no-normalidad de la distribución mediante sendos test de Shapiro-Wilk (tanto control como experimental $p < 0{,}0001$) por lo que para comparar las medias se opta por un test no paramétrico (Mann-Whitney-Wilcoxon). Para la realización del test Mann-Whitney-Wilcoxon se utiliza la función al efecto que proporciona el paquete «coin» de R ya que parece tener un mejor manejo de los empates respecto al paquete estándar. El resultado es el esperado de no significancia del test ($z = 0{,}704$; $p = 0{,}485$) lo que refuerza la idea inicial de que ambos grupos son equivalentes.

Resultado del test de conocimientos previos

El test de conocimientos previos trata de medir los conocimientos que ya tiene el alumno sobre la materia al iniciar el curso. Se trata de un test de diez preguntas, con cuatro respuestas cada una, que solamente se puede contestar una vez. Al va-

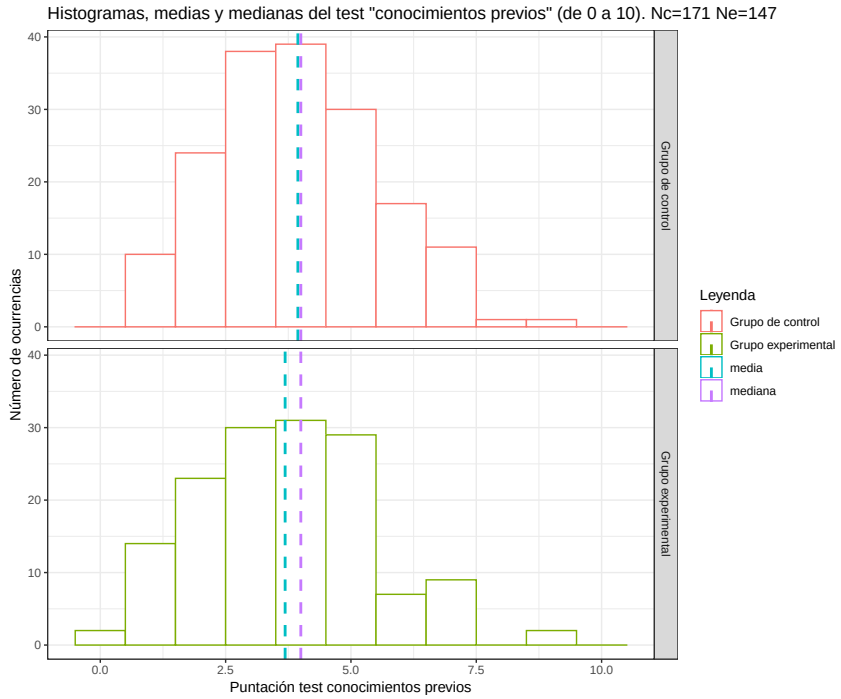


Figura 4.5.: Distribución de resultados en el test de conocimientos previos

lorar los resultados hay que tener en cuenta que se pidió a los participantes que contestaran a las preguntas aunque no conocieran la respuesta por lo que, si a esta situación le añadimos algo de sentido común, parece bastante plausible que los alumnos contesten correctamente a dos o tres preguntas sin contar con un conocimiento previo real. Como ejemplo, la persona externa que realizó el *beta testing*, con un perfil de ingeniero industrial, pero que no contaba con ningún tipo de conocimientos previos en aprendizaje automático, obtuvo una puntuación de 5 en el test. En la figura 4.5 podemos comprobar cómo la mayoría de los resultados están entre unas notas de 2 y 5 y cómo los resultados entre ambos grupos son bastante similares.

En el diagrama de violines (fig. 4.6) podemos ver la densidad de las distribuciones con los datos de medias, medianas y cuartiles. La mayor diferencia observable es que la distancia del tercer cuartil a la mediana es mayor en el grupo experimental que en el de control. Al aplicar el test Shapiro-Wilk a ambas distribuciones, se obtiene que no pasan el test de normalidad ($p_{control} < 0{,}0001$; $p_{experimental} = 0{,}00019$) por lo que se aplica el test Mann-Whitney-Wilcoxon para comparar los resultados de los dos grupos, obteniendo que los grupos no difieren de forma significativa ($z = 1{,}3087$; $p = 0{,}191$).

Con estos resultados podemos tener mayor seguridad de que el proceso de asigna-

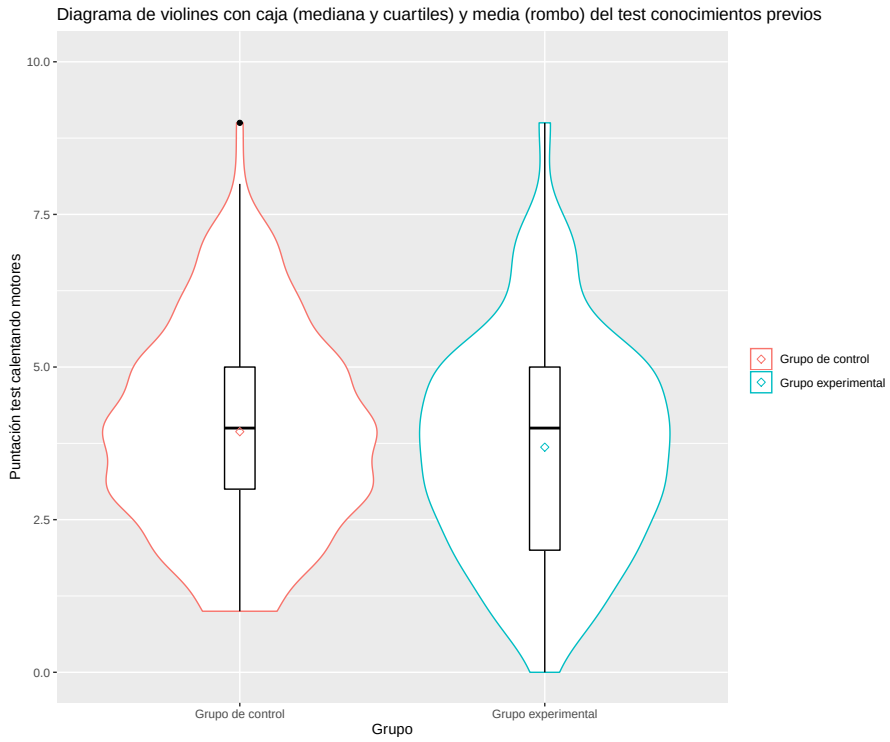


Figura 4.6.: Diagrama de violines y cajas del test de conocimientos previos

ción aleatoria ha sido correcto y de que ambos grupos son, en general, comparables.

4.2.3. El efecto sobre la tasa de abandono

En este apartado se analiza el efecto que ha podido tener la interpolación de preguntas de refuerzo en los vídeos formativos del MOOC sobre la tasa de abandono global del curso.

La figura 4.7 muestra el porcentaje de alumnos de cada grupo que contesta a cada test a lo largo del curso. De las 395 y 317 personas que forman inicialmente el grupo de control y el experimental respectivamente, tan solo 171 del grupo de control y 147 del experimental realizan el test de conocimientos previos. Esto es una muestra de la diferencia que hay entre el número de personas que se registran a un MOOC y el número de personas que finalmente se disponen a realizar el curso, al menos como el profesor espera ya que la bibliografía ya nos avisa de las distintas formas que pueden tener los alumnos de aproximarse a un MOOC (entre otras: sólo revisar el material o solamente ver los vídeos).

Entre el test de conocimientos previos y el test de la lección 2 se produce una muy importante caída de alumnos. Esta fuerte caída pude tener distintas causas, las más importantes de las cuales pueden ser: (1) que la lección 1 no tiene un test evaluado y por tanto es lógico que la caída sea mayor; (2) que en ese primer contacto con el curso es cuando es posible que el alumno se dé cuenta que el curso no es lo que esperaba; y (3) que la lección 2 es la que más carga matemática tiene con diferencia del curso (tratando conceptos como la regresión lineal y logística), pudiendo provocar un abandono mayor. Hubiera sido interesante disponer de un test al final de la primera lección, para poder separar el efecto sobre la tasa de abandono de esa segunda lección, que puede haber sido una barrera importante para muchos alumnos, y poder analizar si existe alguna diferencia en la progresión de la tasa de abandono de los dos grupos hasta el final de la primera lección y entre el final de la primera lección y el final de la segunda –se recuerda que la primera lección es idéntica para ambos grupos–.

Entre el resto de los test la caída de alumnos es más paulatina y da la impresión de que la tendencia de caída de alumnos es muy similar para ambos grupos, aunque algo mayor para el grupo de control que para el experimental. En la lección 4 («redes neuronales y aprendizaje profundo») la caída de participantes del grupo de control se reduce notablemente. Hipotetizando sobre las posibles causas, es posible que la mayor caída anterior de alumnos provoque que hayan llegado a esa lección alumnos más motivados; que el tema que trata esa lección sea más interesante para los alumnos, aunque en este caso el efecto debería apreciarse también sobre el grupo experimental y no es el caso, o que el intercalar preguntas esté consiguiendo retrasar el momento del abandono del curso, es decir, que a pesar de la gran caída

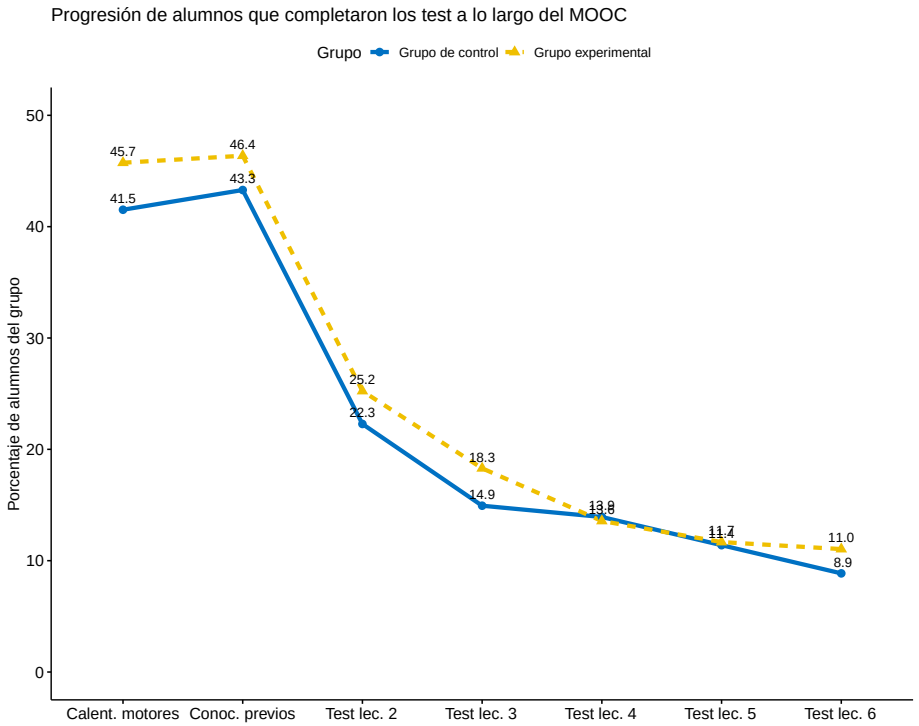


Figura 4.7.: Porcentaje de alumnos de cada grupo que completaron los test según avanzó el curso

de alumnos que hay en los dos grupos entre las lecciones 1 y 3, que el intercalar preguntas en los vídeos pueda estar consiguiendo retrasar ese abandono, aunque por no demasiado tiempo.

Si se comparan los valores globales de permanencia en el curso, se puede ver cómo, a pesar de que el curso se diseñó explícitamente con el objetivo de maximizar el número de alumnos que lo finalizaran, las tasas de finalización son bastante bajas (ver tabla 4.1). Esas tasas son bastante mayores si contamos solamente a aquellas personas que han rellenado alguno de los test iniciales del curso («calentando motores» y «conocimientos previos») ascendiendo a un $17{,}2\,\%$ para el grupo de control y a un $21{,}5\,\%$ para el grupo experimental, si bien hay que reseñar que 3 personas del grupo de control han finalizado el curso sin haber rellenado los test previos y no han formado parte del cómputo de esa tasa.

A pesar de que en las muestras concretas disponibles parece haber un mejor comportamiento (mayor tasa de permanencia) para el grupo experimental, si comparamos los dos grupos mediante un test Chi-cuadrado obtenemos que no hay una asociación significativa entre la posición de las preguntas de refuerzo –y su correspondiente *feedback*– y la tasa de permanencia en el curso ($\chi^2(1) = 0{,}94$; $p =$

Variable	Gr. control	Gr. experimental	Total
Total personas	395	317	712
Finalizan el curso	35	35	70
Aprueban el curso	33	34	67
Tasa de finalización	8,9 %	11 %	9,83 %

Tabla 4.1.: Comparativa de finalización del curso entre grupos

$0{,}3315$; $odds\,ratio = 1{,}28$) lo cual no nos permite extraer conclusiones con respecto a la Hipótesis 1, esto es, que intercalar preguntas tenga el efecto de conseguir una mayor tasa de finalización del curso.

4.2.4. El efecto sobre el aprendizaje

La segunda hipótesis que ha motivado el experimento ha sido que se espera que utilizar preguntas intercaladas en los vídeos mejore los resultados en el aprendizaje frente a realizar las mismas preguntas al final del vídeo. En este apartado se estudia si efectivamente se produce este efecto. Primero se comparan los resultados de los dos grupos experimentales en el total de los cinco test evaluados del curso, diferenciando la puntuación obtenida sobre las preguntas «de recuerdo» y sobre las preguntas «de transferencia». Posteriormente se detallan los resultados de cada uno de los test sobre los dos niveles de aprendizaje.

Todas las comparaciones que se realizan en el presente apartado siguen la misma estructura. Primero se presenta la diferencia de las medias de la puntuación de ambos grupos. El motivo de presentar las medias a pesar de que, en principio, como se verá, podría tener más sentido presentar la mediana, es que en ambos grupos la gran mayoría de los alumnos obtiene notas muy altas en todos los test, como podrá observarse en las distintas figuras, y la diferencia entre los grupos se observa en las notas bajas. Estos resultados son esperados y son el resultado de que los alumnos pueden consultar el material mientras hacen los test y que los test se han diseñado para no presentar una alta dificultad que pudiera desanimar al alumno (el objetivo, tanto desde el punto de vista exclusivo del MOOC, al tratarse de un curso que se encuentra entre la divulgación y una introducción técnica, como desde el punto de vista del experimento, fue desde un principio que la mayor parte posible de alumnos llegaran al final del curso).

Como se verá, los resultados arrojan una diferencia significativa en los resultados de recuerdo del total de los test, pero, contrariamente a lo esperado, este efecto no se repite en los resultados de transferencia puesto que, aunque se obtiene una ligera

mejora en las medias, no se alcanza la significancia estadística. Un patrón que se repite en todos los casos, de recuerdo y de transferencia, tanto globales como de cada uno de los test por separado, es que para el grupo experimental las puntuaciones medias de todos los test son superiores y que se reduce la desviación estándar mediante un menor número de notas bajas en el grupo experimental como se podrá observar en los histogramas.

4.2.4.1. El efecto sobre los resultados totales

Para analizar los resultados del total de los test se han tenido en cuenta aquellos alumnos que han contestado a todos y cada uno de los test. Como se ha detallado anteriormente, esto implica que el tamaño de la muestra se ha reducido de manera importante hasta un número de 35 personas en cada condición. Esta pérdida de personas puede tener distintas implicaciones. Descartado, en principio, el *sesgo de selección* –o al menos un efecto importante del mismo ya que la diferencia en los datos de finalización del curso de ambas cohortes no ha presentado significancia estadística–, existen otras posibles implicaciones a tener en cuenta como (1) que aquellos alumnos que finalizan el curso son a priori los más motivados o (2) que la reducción en el tamaño muestral puede dar problemas a la hora de detectar efectos que no sean fuertes.

Resultado total sobre las preguntas de recuerdo

La figura 4.8 presenta las medias totales (entre 0 y 30, máximo de 6 puntos en cada uno de los 5 test) y el error estándar de las dos condiciones experimentales.

El grupo experimental obtiene una media 2,3 puntos superior. En los histogramas de la figura 4.9 parece observarse como el grupo experimental tiene un menor número de resultados bajos, con el grupo experimental obteniendo un sólo caso por debajo de los 23 puntos, por 9 casos en el caso del grupo de control. El diagrama de violines (figura 4.10) aclara aún más esta observación con la distancia en la mitad inferior de la distribución siendo mucho menor en el grupo experimental que en el de control. Esta observación se repetirá, aunque no con tanta fuerza, cuando se analicen cada uno de los test por separado.

El resultado de la comparación de los resultados de los dos grupos mediante un test estadístico (Mann-Whitney-Wilcoxon, al no superarse los test de Shapiro-Wilk, $p_{control} = 0{,}006$; $p_{experimental} = 0{,}0001$) es que hay una diferencia significativa de tamaño medio entre el grupo experimental y de control ($z = -3{,}09$; $p = 0{,}0017$; $dif\,medianas = -2$; $CI\,95\,\% = (-3; -1)$; $r = -0{,}36$).

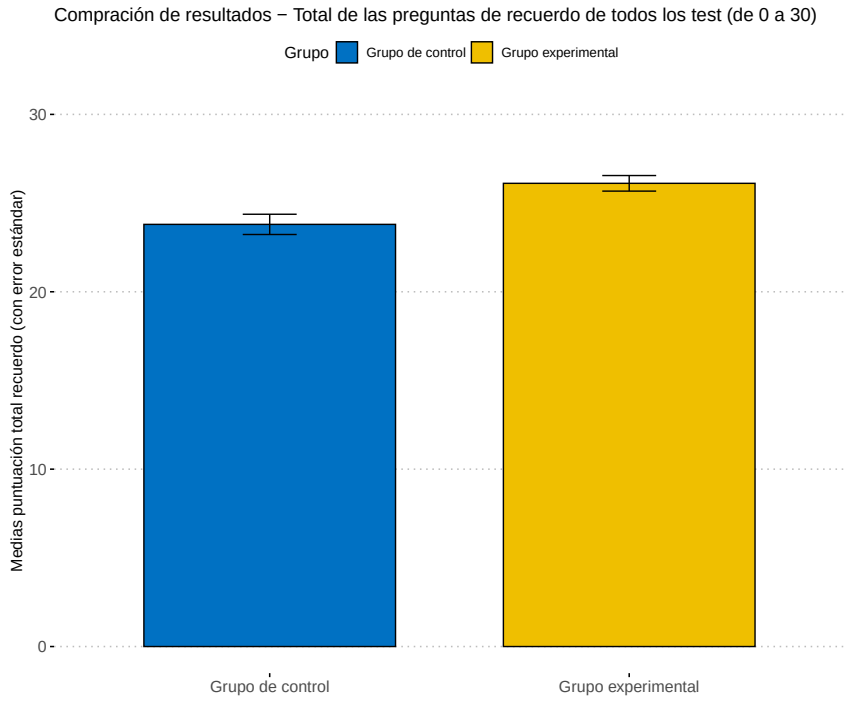


Figura 4.8.: Comparación de los resultados (medias) globales de recuerdo.

Variable	**Gr. control**	**Gr. experimental**
N	35	35
Media (*DE*)	23,8 (3,38)	26,11(2,6)
Mediana (*RI*)	25 (3,5)	26 (3)
Min-max (*amplitud*)	13 – 29 (16)	16 – 30 (14)
Error estándar	0,57	0,44

Tabla 4.2.: Variables descriptivas de la distribución de los resultados totales de recuerdo: grupo de control vs grupo experimental

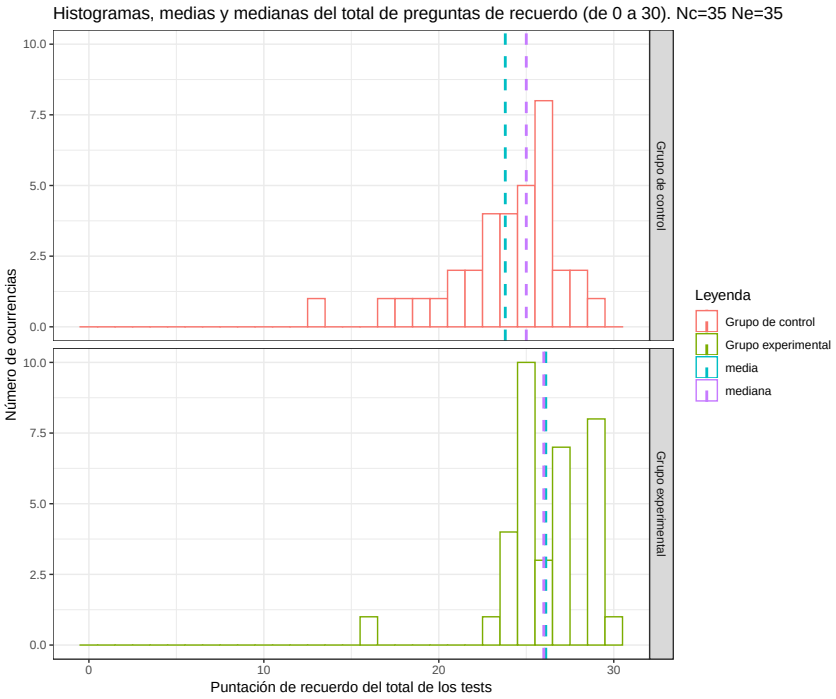


Figura 4.9.: Distribución de los resultados globales de recuerdo.

Resultado total sobre las preguntas de transferencia.

En la media del total de las preguntas de trasferencia de todos los test el grupo experimental obtiene un resultado ligeramente superior, de poco más de un punto (sobre 20), y una desviación estándar más reducida. La mediana es idéntica y el rango intercuatílico un punto inferior para el grupo experimental.

De la misma manera que en el caso de las preguntas de recuerdo, podemos observar tanto en los histogramas (fig. 4.12), como en el diagrama de violines (fig. 4.13) un menor número de casos con puntuación baja en el grupo experimental que en el grupo de control, agrupándose los datos, en este caso, más cerca de la media y consiguiendo reducir la amplitud de la mitad inferior de los valores.

Al comprobar la asunción de normalidad mediante el test de Shapiro-Wilk se obtiene que $p_{control} = 0{,}064$; y que $p_{experimental} = 0{,}003$ por lo que se podría considerar que la muestra del grupo de control cumple con la asunción de normalidad, sin embargo, la distribución del grupo experimental no la cumple por lo que, como se viene realizando, se aplica el test Mann-Whitney-Wilcoxon con el siguiente resultado: ($z = -1{,}09$; $p = 0{,}2798$; $dif\,medianas = -1$; $CI\,95\,\% = (-2; 1)$; $r = -0{,}13$), por lo que la diferencia no es significativa.

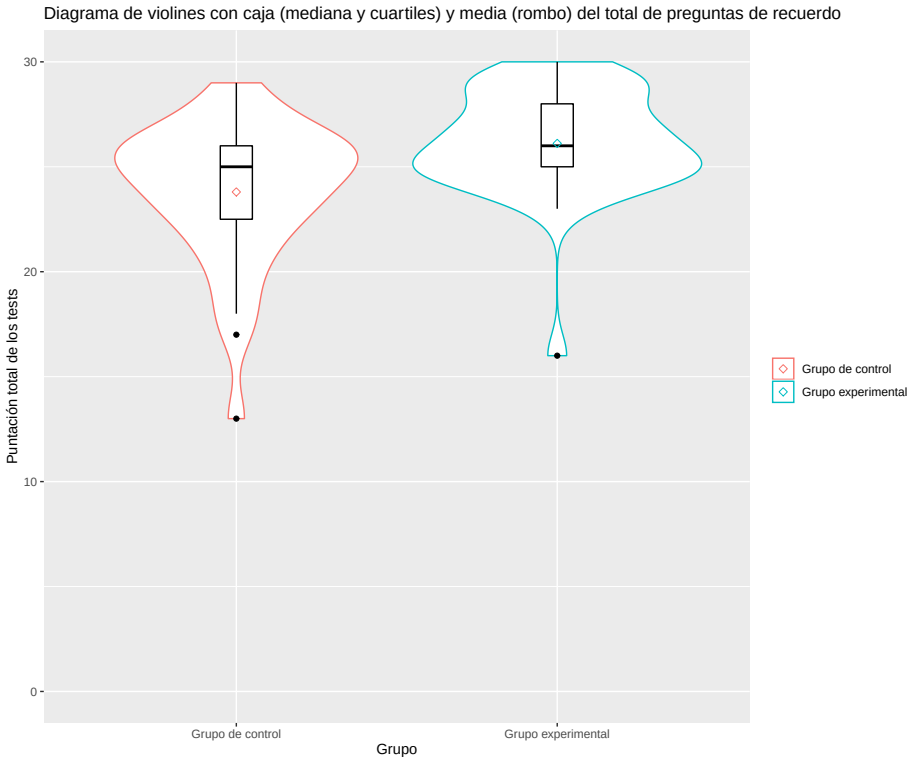


Figura 4.10.: Diagrama de violines y cajas de los resultados globales de recuerdo.

Variable	**Gr. control**	**Gr. experimental**
N	35	35
Media (DE)	14,31(3,55)	15,37 (2,79)
Mediana (RI)	15 (4)	15 (3)
Min-max (amplitud)	6 – 20 (14)	5 – 20 (15)
Error estándar	0,6	0,47

Tabla 4.3.: Variables descriptivas de la distribución de resultados totales de transferencia: grupo de control vs grupo experimental

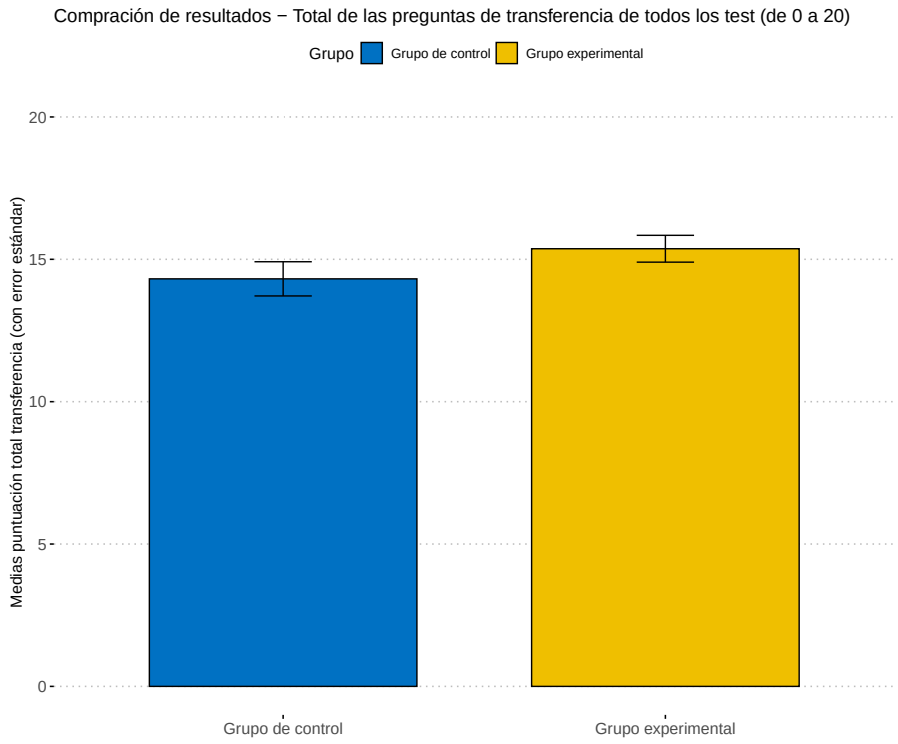


Figura 4.11.: Comparación de los resultados (medias) globales de transferencia.

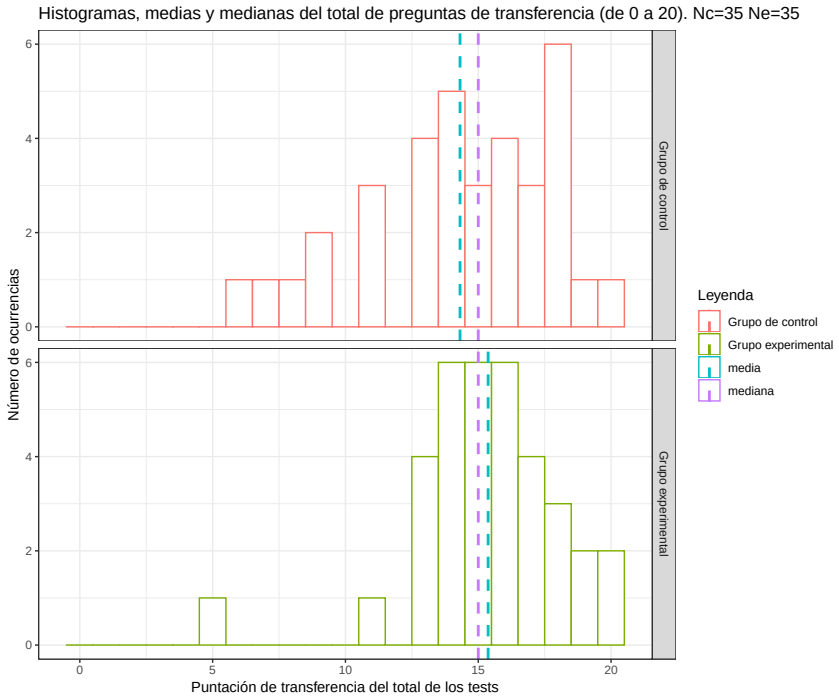


Figura 4.12.: Distribución de los resultados globales de transferencia.

4.2.4.2. El efecto en cada test

Una vez expuestos los resultados globales, del conjunto de todos los test, que la posición de las preguntas de refuerzo tienen sobre el aprendizaje, tanto en su nivel inferior –mejora de la capacidad retentiva–, mediante las preguntas que se han denominado «preguntas de recuerdo», como en su nivel superior –aplicación de los conocimientos a contextos diferentes–, mediante las «preguntas de transferencia», es el momento de analizar los resultados que obtiene cada uno de los grupos en cada uno de los test, en estos dos niveles del aprendizaje.

Preguntas de recuerdo.

El resultado específico sobre las preguntas de recuerdo en cada test se puede ver en la figura 4.14. Los resultados son siempre favorables al grupo experimental, aunque en todos los casos la diferencia es bastante exigua, excepto en el test de la última lección, la lección 6. Esta lección es un tanto distinta de las demás al tratar temas relacionados con los aspectos éticos y sociales de la aplicación de las técnicas del aprendizaje automático, por lo que es posible que su especificidad en este sentido haya influido en que la diferencia haya sido mayor. En cuanto al resto de lecciones, en la lección 2 y en la lección 5 parece que la diferencia ha sido algo superior a la obtenida para las lecciones 3 y 4. Una posibilidad es que se deba a que si se catalogan

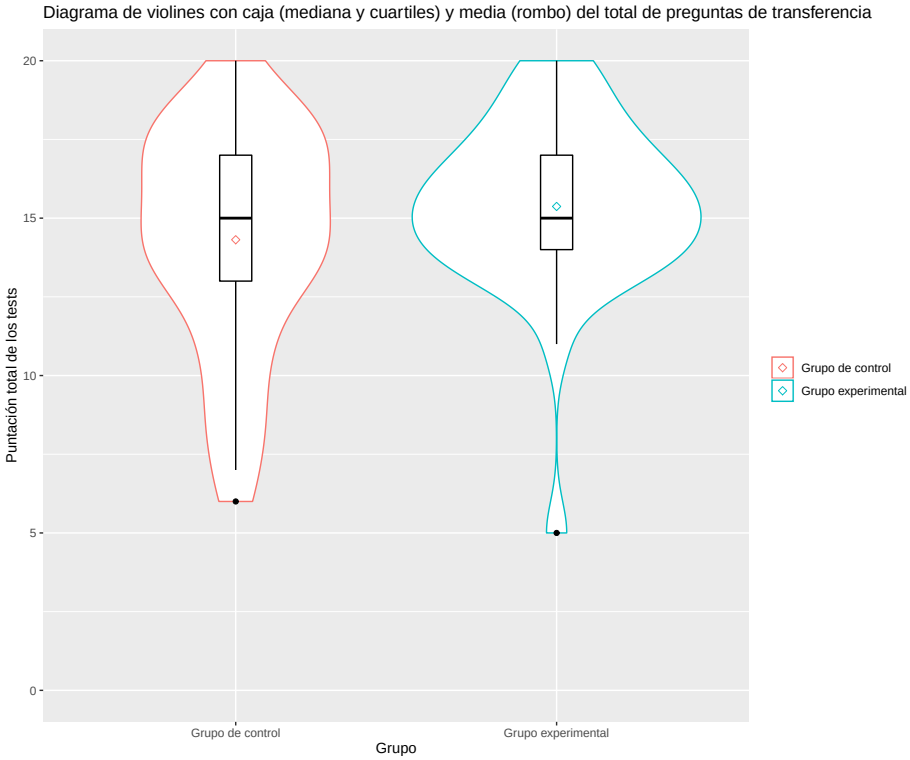


Figura 4.13.: Diagrama de violines y cajas de los resultados globales de transferencia.

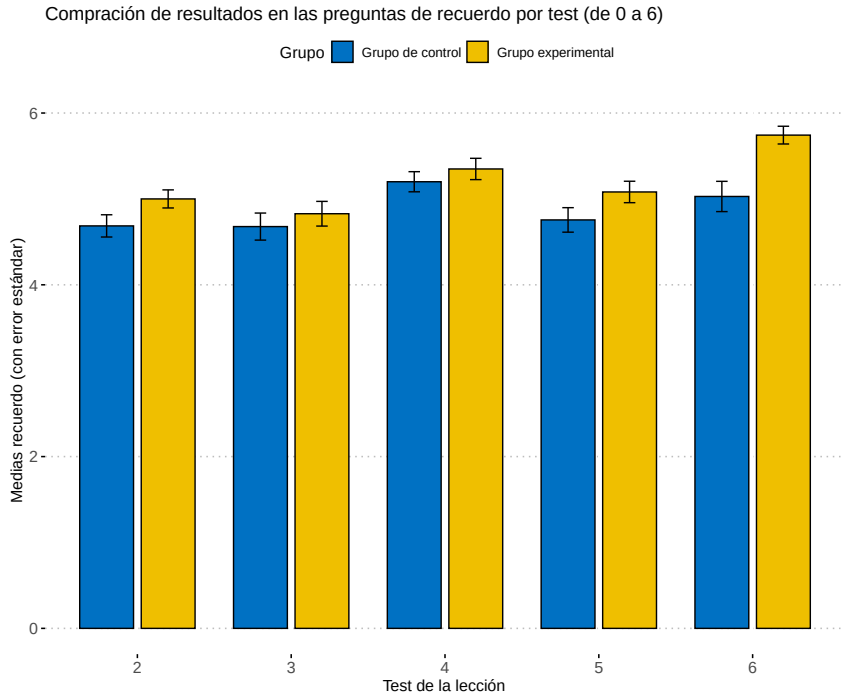


Figura 4.14.: Comparación de los resultados (medias) de recuerdo en cada test.

las lecciones por dificultad, las lecciones 2 y 5 se podría considerar que tienen una mayor dificultad que las lecciones 3 y 4: la lección 2 por su carga matemática y la lección 5 por ser la de contenido más abstracto y posiblemente más avanzado, al tratar conceptos como el sobreajuste, sesgo y la varianza de los modelos.

Un aspecto más a resaltar es que, como ya se ha comentado al valorar los resultados agregados de todos los test, no sólo la media es mayor en todos los casos en el grupo experimental, sino que la desviación estándar también es menor que en el grupo de control. Este aspecto, teniendo en cuenta que la mayor parte de los alumnos obtiene muy buenos resultados debido a la sencillez de los test (la mediana no baja de 5 sobre 6 para ninguno de los grupos en ninguno de los test), indica que en el grupo experimental se ha dado, de manera consistente, un menor número de resultados bajos que en el grupo experimental.

Se ha aplicado un test Mann-Whitney-Wilcoxon (en ningún caso se cumple con la asunción de normalidad) para comparar los resultados en cada uno de los test y el resultado se presenta en la tabla 4.5. Salvo en el caso del test de la lección 6 ($p = 0{,}0003$, $r = -0{,}27$), ninguno de los test resulta en significancia estadística. La diferencia entre las distribuciones de los resultados de las dos cohortes es muy llamativa para este test, especialmente si tenemos en cuenta que para el resto de test las diferencias han sido muy sutiles (fig. 4.15). Por ello, para descartar un posible

	Grupo	*N*	*Media* (*DE*)	*Mediana* (*RI*)	*Min-max* (*amplitud*)	*Err. est.*
Lec. 2	**Con.**	86	4,69 (1,2)	5 (2)	1 – 6 (5)	0,13
	Exp.	80	5 (0,94)	5 (2)	2 – 6 (4)	0,11
Lec. 3	**Con.**	59	4,68 (1,21)	5 (2)	1 – 6 (5)	0,16
	Exp.	58	4,83(1,09)	5 (2)	2 – 6 (4)	0,14
Lec. 4	**Con.**	55	5,2 (0,87)	5 (1)	3 – 6 (3)	0,12
	Exp.	43	5,35 (0,81)	6 (1)	3 – 6 (3)	0,12
Lec. 5	**Con.**	45	4,76 (0,96)	5 (1)	2 – 6 (4)	0,14
	Exp.	37	5,08 (0,76)	5 (1)	3 – 6 (3)	0,12
Lec. 6	**Con.**	35	5,03 (1,04)	5 (1,5)	2 – 6 (4)	0,18
	Exp.	35	5,74 (0,61)	6 (0)	3 – 6 (3)	0,1

Tabla 4.4.: Variables descriptivas de la distribución de resultados de las preguntas de recuerdo en los distintos test.

error de procedimiento, se ha revisado específicamente que tanto la configuración del MOOC en esa lección para ambos grupos, como la extracción y preprocesado de datos ha sido correcta.

El efecto sobre las preguntas de transferencia

La figura 4.16 muestra las medias de cada grupo para cada uno de los test en las preguntas de transferencia. Cuatro preguntas de cada test estaban diseñadas para medir el efecto sobre este tipo de aprendizaje por lo que la puntuación para cada test está comprendida entre 0 y 4. El grupo experimental mejora en todos los casos los resultados del grupo de control, aunque por muy escaso margen. A su vez, continúa detectándose una reducción de la desviación estándar en el grupo experimental con respecto al de control (tabla 4.6), salvo para el test de la lección 6, en el que la desviación estándar es muy similar (1,08 del grupo de control y 1,1 del grupo experimental). Los resultados de ambos grupos en la lección 6 son sensiblemente inferiores al resto de las lecciones, probablemente debido a que las preguntas de transferencia de esta lección tenían una dificultad mayor que en otras lecciones.

En ningún caso la comparación de los grupos mediante Mann-Whitney-Wilcoxon es significativa (detalles en la tabla 4.7) por lo que no es posible tener ninguna seguridad de que el mejor resultado de la cohorte experimental no sea simplemente efecto del azar. Sin embargo, el que continúe la tendencia detectada para las preguntas de recuerdo, aunque a una menor intensidad, hace pensar que existe la posibilidad

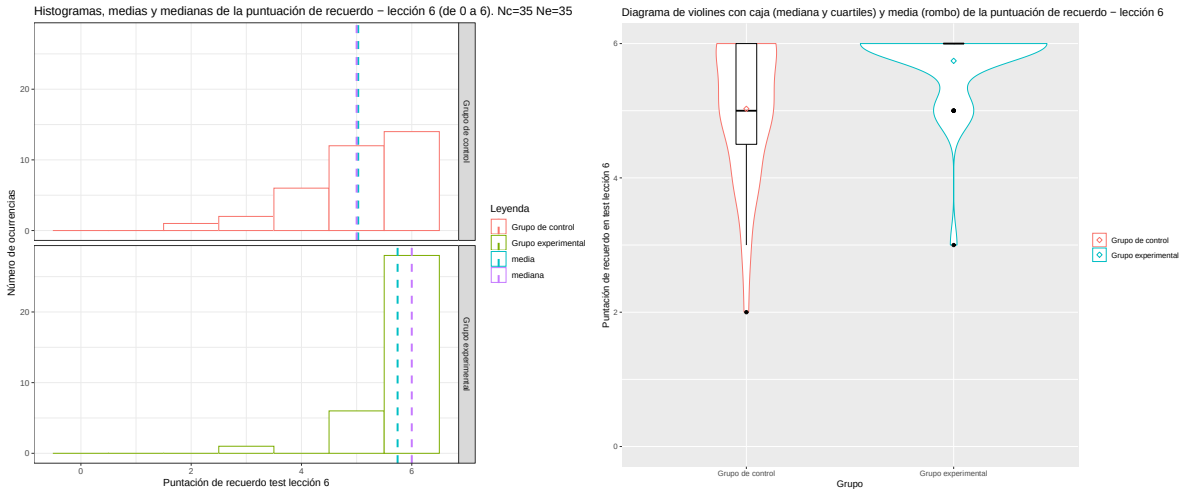


Figura 4.15.: Distribución de los resultados de recuerdo del test de la lección 6.

	z	p	*diferencia medianas*	*Int. conf 95 %*	r
Test lección 2	$-1{,}56$	$0{,}119$	0	$(-1\,;\,0)$	$-0{,}12$
Test lección 3	$-0{,}51$	$0{,}6144$	0	$(-1\,;\,0)$	$-0{,}04$
Test lección 4	$-0{,}89$	$0{,}3815$	0	$(0\,;\,0)$	$-0{,}07$
Test lección 5	$-1{,}65$	$0{,}102$	0	$(-1\,;\,0)$	$-0{,}13$
Test lección 6	$-3{,}53$	$0{,}0003$*	-1	$(-1\,;\,0)$	$-0{,}27$

*diferencia significativa $(p < 0{,}05)$

Tabla 4.5.: Resultados del test Mann-Whitney-Wilcoxon en los cinco test del curso - preguntas de recuerdo

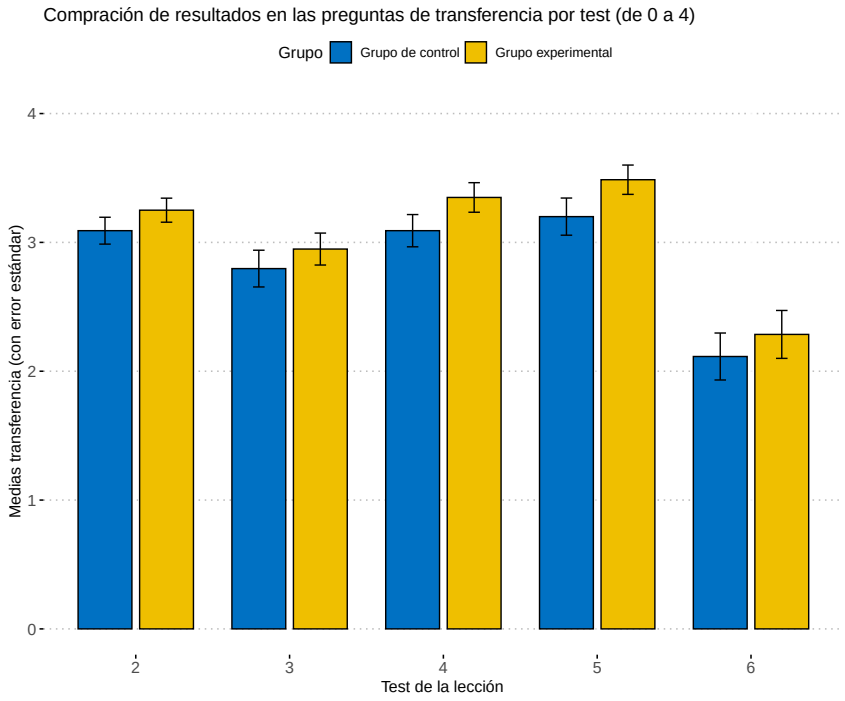


Figura 4.16.: Comparación de los resultados (medias) de transferencia en cada test.

	Grupo	*N*	*Media (DE)*	*Mediana (RI)*	*Min-max (amplitud)*	*Err. est.*
Lec. 2	**Con.**	88	3,09 (0,98)	3 (1)	0 – 4 (4)	0,1
	Exp.	80	3,25 (0,83)	3 (1)	0 – 4 (4)	0,1
Lec. 3	**Con.**	59	2,8 (1,1)	3 (2)	0 – 4(4)	0,14
	Exp.	58	2,95(0,94)	3 (2)	0 – 4(4)	0,12
Lec. 4	**Con.**	55	3,09 (0,93)	3 (1)	1 – 4 (3)	0,13
	Exp.	43	3,35 (0,75)	3 (1)	1 – 4 (3)	0,11
Lec. 5	**Con.**	45	3,2 (0,97)	4 (1)	1 – 4 (3)	0,14
	Exp.	37	3,49 (0,69)	4 (1)	2 – 4 (2)	0,12
Lec. 6	**Con.**	35	2,11 (1,08)	2 (2)	0 – 4 (4)	0,18
	Exp.	35	2,29 (1,1)	2 (1,5)	0 – 4 (4)	0,19

Tabla 4.6.: Variables descriptivas de la distribución de resultados de las preguntas de transferencia en los distintos test.

	z	p	*diferencia medianas*	*Int. conf 95 %*	r
Test lección 2	$-0{,}87$	$0{,}3831$	0	$(0\,;\,0)$	$-0{,}07$
Test lección 3	$-0{,}61$	$0{,}5442$	0	$(0\,;\,0)$	$-0{,}05$
Test lección 4	$-1{,}30$	$0{,}1971$	0	$(-1\,;\,0)$	$-0{,}10$
Test lección 5	$-1{,}18$	$0{,}2392$	0	$(0\,;\,0)$	$-0{,}09$
Test lección 6	$-0{,}45$	$0{,}6571$	0	$(-1\,;\,0)$	$-0{,}03$

Tabla 4.7.: Resultados del test Mann-Whitney-Wilcoxon en los cinco test del curso - preguntas de transferencia

de que los instrumentos utilizados, con tan sólo cuatro preguntas de transferencia por test, pueden no haber tenido la suficiente sensibilidad para detectar un mayor efecto. A esto hay que añadir un bajo tamaño muestral para los últimos test y que habitualmente los efectos de transferencia son más complejos de sacar a la luz.

4.2.5. El efecto sobre la impresión subjetiva del curso.

Al finalizar el MOOC se solicitó a los alumnos que rellenaran una encuesta de satisfacción para valorar tanto la impresión general, como por grupo experimental, que ha generado el curso. El cuestionario consta de 35 preguntas, divididas en 7 bloques de 5 preguntas: 1 bloque para cada una de las 6 lecciones, y 1 bloque sobre el curso en general.

Cada uno de los bloques constaba de las siguientes cuestiones:

- Valora en general tu experiencia con la lección X (o con el curso en general)
- Valora tu experiencia con el contenido escrito de la lección X (o con el curso en general)
- Valora tu experiencia con el contenido audiovisual de la lección X (o con el curso en general)
- Valora tu experiencia con las preguntas de refuerzo de la lección X (o con el curso en general)
- Valora tu experiencia con el cuestionario de evaluación de la lección X (o con el curso en general)

Para cada una de las preguntas se permitían cinco posibles respuestas: «muy insatisfactoria», «insatisfactoria», «neutral», «satisfactoria», «muy satisfactoria», a cada una de las cuales se les ha asignado un valor numérico: 0, 1, 2, 3 y 4 respectivamente.

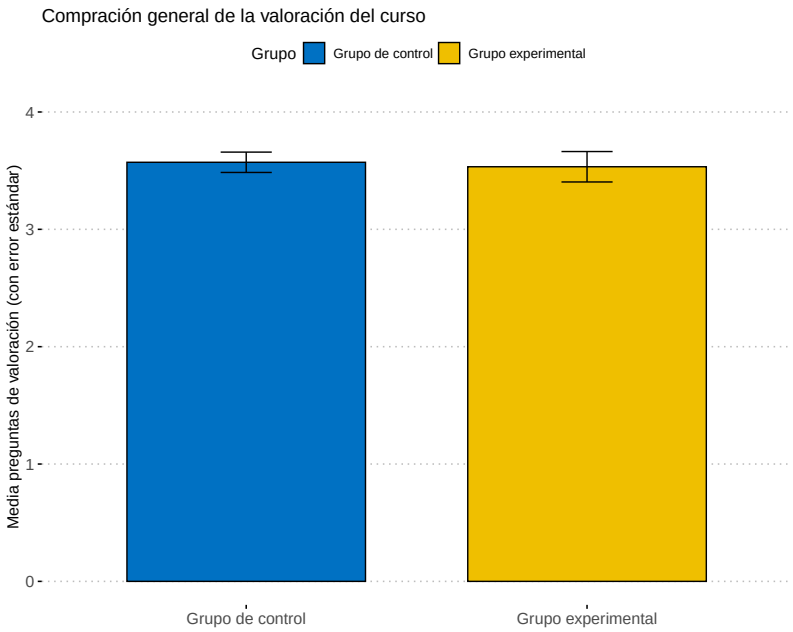


Figura 4.17.: Comparación de los resultados (medias) del total de preguntas de valoración.

El cuestionario fue contestado por 29 personas del grupo de control y por 27 personas del grupo experimental y los resultados de las medias de cada grupo para las 35 preguntas están representadas en la figura 4.17. Como se puede ver, la satisfacción de los alumnos que han contestado a la encuesta es bastante alta. La media de la valoración total es muy ligeramente superior para el grupo de control (detalles disponibles en la tabla 4.8).

Al analizar la gráfica de violines (fig. 4.18), que muestra la densidad de las dos muestras, podemos comprobar como los resultados son bastante parecidos, con la salvedad de la existencia de un claro valor atípico en el grupo experimental. En efecto, si se calcula la media de los valores eliminando el 10 % superior y el 10 % inferior, el resultado cambia y pasa a ser ligeramente superior para el grupo experimental. En cualquier caso, una comparación formal de los grupos, con todos los valores, mediante el test Mann-Whitney-Wilcoxon, debido a que ni los valores del grupo de control, ni los valores del experimental superan la asunción de normalidad (Shapiro-Wilk: $p_{control} < 0{,}0001$; y $p_{experimental} < 0{,}0001$), obtiene, como era esperable a la vista de los datos descriptivos, que la diferencia no es significativa: $z = 0{,}35$; $p = 0{,}7282$; $dif\,medianas = 0$; $CI\,95\,\% = (-0{,}20; 0{,}14)$; $r = 0{,}05$.

De las cinco preguntas realizadas, tienen especial interés para valorar la experiencia subjetiva a la luz del experimento realizado, tres de ellas: la valoración general de cada tema, la valoración de los medios audiovisuales, ya que la experiencia con la

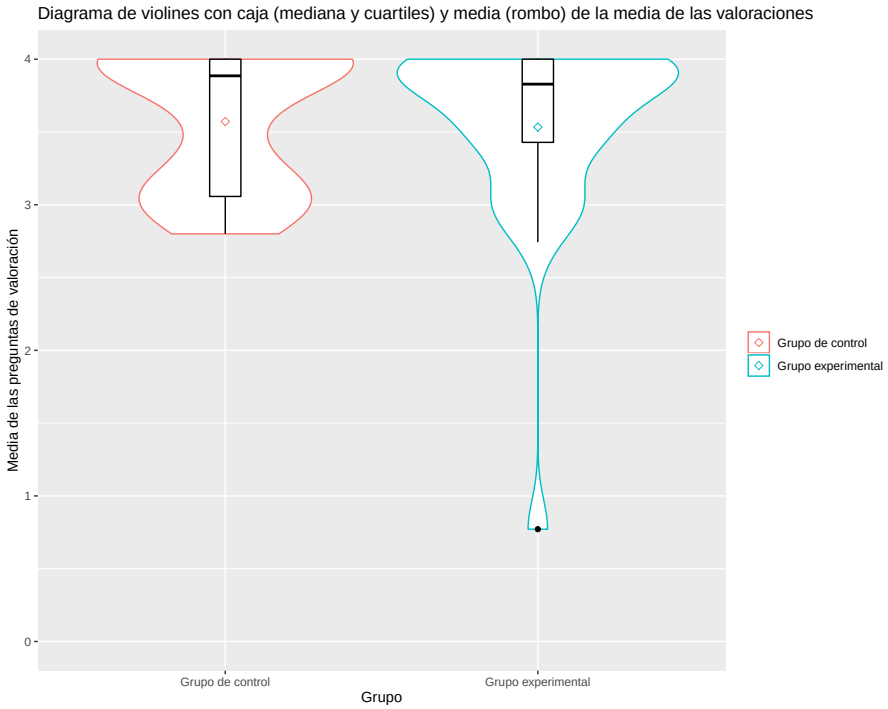


Figura 4.18.: Diagrama de violines y cajas de las medias de todas las preguntas de valoración.

Variable	**Gr. control**	**Gr. experimental**
N	29	27
Media (DE)	3,57 (0,47)	3,53 (0,67)
Media recortada (10 %)	3,6	3,65
Mediana (RI)	3,89 (0,94)	3,83 (0,57)
Min-max (amplitud)	2,8 – 4 (1,2)	0,77 – 4 (3,23)
Error estándar	0,09	0,13

Tabla 4.8.: Variables descriptivas de la distribución de la media de las peguntas de valoración del curso: grupo de control vs grupo experimental

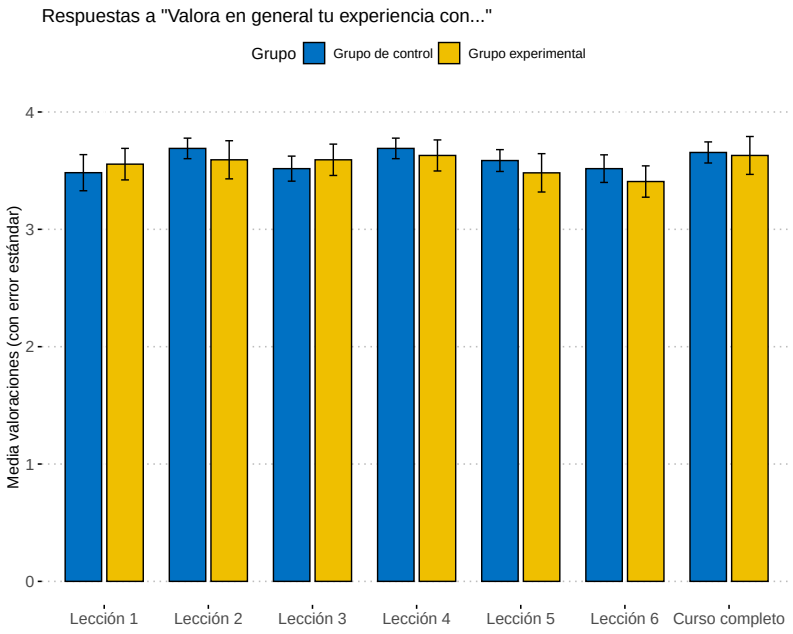


Figura 4.19.: Comparación de los resultados (medias) de las respuestas a la pregunta sobre la valoración general de cada lección y del curso completo.

mayoría de los vídeos es diferente en cada grupo, y la valoración con las preguntas de refuerzo. Si se comparan las respuestas de los componentes de cada cohorte por cada lección a la cuestión de valorar la experiencia en general con cada lección y con el curso en general, obtenemos el resultado de la figura 4.19.

En general los resultados son ligeramente inferiores para el grupo experimental (Mann-Whitney-Wilcoxon: $z = -0{,}38$; $p = 0{,}7088$; $dif\,medianas = 0$; $CI\,95\,\% = (0; 0)$; $r = 0{,}05$), salvo para la lección 1, que conviene recordar que es idéntica para ambos grupos, y para la lección 3.

Respecto a las preguntas que solicitaban la valoración de los vídeos y la valoración de las preguntas de refuerzo, los resultados se pueden observar en las figuras 4.20 y 4.21. En ninguno de los dos casos se observa una diferencia significativa, siendo los resultados del test Mann-Whitney-Wilcoxon a la valoración de los vídeos $z = -0{,}82$; $p = 0{,}4129$; $dif\,medianas = 0$; $CI\,95\,\% = (0; 0)$; $r = -0{,}11$ y a la valoración de las preguntas de refuerzo $z = -0{,}43$; $p = 0{,}6666$; $dif\,medianas = 0$; $CI\,95\,\% = (0; 0)$; $r = -0{,}06$

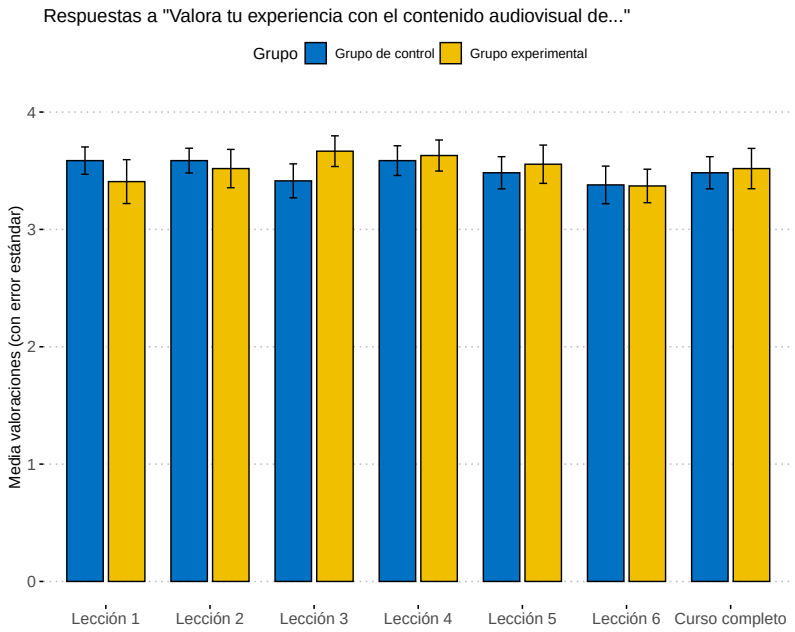


Figura 4.20.: Comparación de los resultados (medias) de las respuestas a la pregunta sobre la valoración del contenido audiovisual de cada lección y del curso completo.

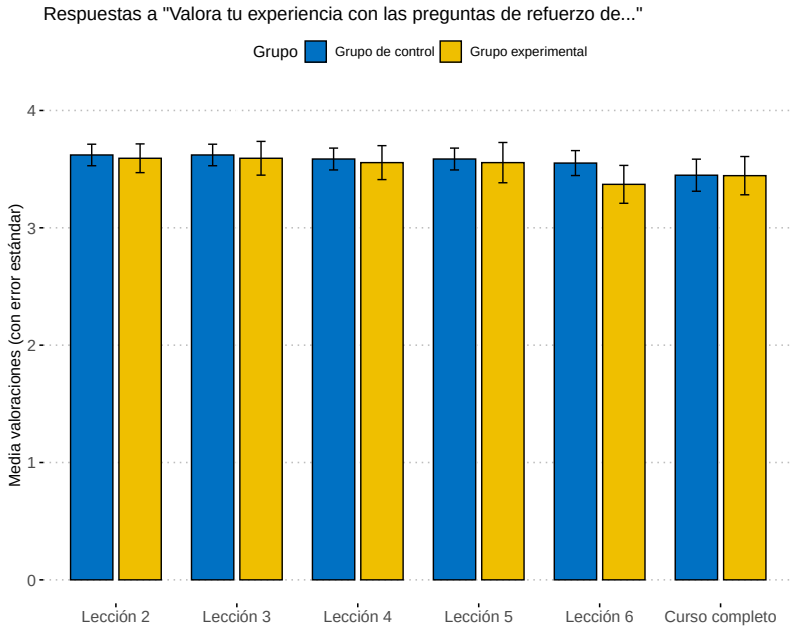


Figura 4.21.: Comparación de los resultados (medias) de las respuestas a la pregunta sobre la valoración de las preguntas de refuerzo de cada lección y del curso completo.

4.3. Discusión de los resultados

Aunque los resultados del experimento arrojan que comparativamente finalizan el curso más personas del grupo experimental que el de control (*odds ratio*= 1,28), la diferencia no es significativa por lo cual son resultados que son perfectamente plausibles sin que el intercalar preguntas en los vídeos tenga efecto en la tasa de abandono de los alumnos. Esto es, no es posible extraer ningún argumento a favor de la primera hipótesis planteada (intercalar preguntas de refuerzo favorece la disminución de la tasa de abandono global del curso). A este respecto, y para poder extraer conclusiones más precisas, hubiera sido interesante poder contar con información de las interacciones de los alumnos con los vídeos, para poder saber si existe o no diferencia a nivel de vídeo (tasa de abandono de los vídeos), es decir, si el compromiso (*engagement*) a nivel de vídeo no se hace extensivo al curso –por ejemplo debido a que el MOOC tiene mucho más contenido que los meros vídeos– o si la diferencia a nivel de vídeo tampoco existe y, por tanto, contrariamente a lo observado en otros estudios (Cummins et al. (2016); Kovacs (2016)), el efecto sobre la motivación de las *in-video quizzes* no es el suficiente para provocar una diferencia significativa. Un indicio que puede hacer pensar que el efecto que pudieran tener las preguntas de refuerzo sobre la motivación queda diluido en el global del MOOC, es que, contrariamente a lo esperado, tampoco se observa ninguna diferencia en la valoración del curso, que es bastante buena para los dos grupos.

Respecto a los resultados sobre el aprendizaje, se obtiene una diferencia significativa favorable al grupo experimental sobre las preguntas orientadas a evaluar la retención, lo cual es un argumento favorable para pensar que el uso de las preguntas de refuerzo interpoladas en los vídeos tiene un efecto positivo sobre este tipo de aprendizaje respecto a la formulación de esas preguntas al finalizar el vídeo. Sin embargo, los resultados obtenidos sobre las preguntas orientadas a evaluar la transferencia no son los esperados dado que, aunque son ligeramente superiores para el grupo experimental, la diferencia no es significativa. Es posible que, dado el tamaño muestral, no se haya contado con el suficiente poder estadístico para detectar el efecto, aunque, aun así, el resultado no hubiera sido el esperado ya que, derivado de la mejora en la capacidad de autorregulación y de la reducción de la carga cognitiva, aspectos que en principio deberían facilitar la construcción de un modelo mental coherente y correcto, se esperaba que el efecto fuera superior sobre la capacidad de transferencia que sobre la de retención.

A este respecto se pueden hacer dos observaciones. La primera, que es posible que o bien las preguntas de refuerzo o las planteadas en los test no hayan sido las idóneas para conseguir detectar el efecto. La segunda, que, en este experimento, incluso para el grupo de control, al no ser los vídeos excesivamente extensos, la evaluación formativa está relativamente cercana a su concepto relacionado, por lo que es posible que no exista diferencia o que ésta sea menor de lo esperado. Esto implicaría

que tampoco la segmentación del vídeo, provocada por intercalar preguntas en el mismo, ha tenido efecto sobre los resultados de transferencia. Pero, si esto fuera así, y no existe una diferencia en transferencia al intercalar las preguntas de refuerzo, ¿por qué sí se encuentra una diferencia significativa en los resultados de retención cuando teóricamente la condición debería afectar más a la capacidad de transferencia? Sólo podemos conjeturar con que, como se ha mencionado anteriormente, las preguntas intercaladas estuvieran beneficiando el compromiso con los vídeos, y de paso, la actitud del alumno. Esta mejora en la actitud, al tener el alumno todos los materiales disponibles para consultarlos mientras realiza el test, es lo que podría provocar que el grupo experimental obtenga un mejor resultado sobre las preguntas pensadas para evaluar la retención –que de ser éste el caso, podrían no estar midiendo exactamente eso–. Como se ha dicho anteriormente, sin contar con los *logs* de interacciones de edX, no es posible tratar de aclarar este aspecto.

Una última observación a resaltar es que la gran mayoría de los alumnos obtiene muy buena puntuación en los test, probablemente debido a que pueden consultar el material que deseen y a que los test no son excesivamente complejos. De hecho, probablemente hubiera sido más óptimo contar con test más complicados, que no provocaran que las medianas estuvieran tan cerca del valor máximo posible (25 sobre 30 en las preguntas de recuerdo y 15 sobre 20 en las de transferencia) para evitar limitar la distribución en sus valores superiores y facilitar el que hubiera mayores diferencias entre las mejores notas. Probablemente debido a este efecto, las diferencias que se detectan son, sobre todo, en los valores inferiores donde se observa que el grupo experimental obtiene, de forma consistente, tanto en recuerdo como en transferencia, no sólo una puntuación media superior sino también reducir la desviación estándar, fundamentalmente consiguiendo que haya un menor número de calificaciones bajas.

El hecho de que se detecte la misma tendencia de forma consistente en todos los test, tanto en las preguntas de recuerdo como en las preguntas de transferencia, aunque con un efecto menos intenso en las últimas, hace que se valore como más probable que, o bien, efectivamente, el efecto es menor para transferencia y no se ha podido detectar, o bien los instrumentos no han sido los idóneos para conseguir capturar el efecto.

5. Conclusiones y líneas futuras

En este capítulo final se abordan las limitaciones del presente trabajo, se plantean posibles líneas de trabajo a futuro y se extraen conclusiones.

5.1. Limitaciones del trabajo

Probablemente la principal limitación de este trabajo sea el no haber podido utilizar para el grupo experimental preguntas realmente embebidas en el vídeo, al no haber sido posible la instalación del módulo necesario en la plataforma Open edX. En su lugar se decidió intercalar las preguntas entre segmentos del vídeo. El resultado puede parecer bastante similar, pero es posible que el alumno perciba diferencias entre estas dos formas de presentar las preguntas al ser, de alguna manera, más «automática» la presentación de la pregunta y la vuelta a la reproducción del vídeo cuando se trata de preguntas realmente embebidas. Futuros trabajos deben contrastar los resultados obtenidos en este experimento mediante el uso de preguntas embebidas y no simplemente intercaladas.

Una segunda limitación tiene que ver con el tamaño muestral. Hubiera sido óptimo poder contar con un mayor número de personas que finalizaran el curso para así tener más capacidad para detectar efectos de pequeño tamaño, como es posible que haya sucedido con los resultados en transferencia. Relacionado con esto, es posible que el estudio adolezca de pocas preguntas de transferencia y que hubiera sido interesante contar al menos con el mismo número de preguntas de transferencia que de recuerdo.

Aunque no estaba previsto entre los análisis iniciales, contar con las interacciones de los alumnos con los vídeos, podría haber ayudado, en este momento, o en una futura investigación, a clarificar algunos de los resultados obtenidos. En concreto, podría clarificar si existe diferencia en la tasa de abandono de los vídeos entre el grupo de control y el grupo experimental o, lo que es lo mismo, si existe diferencia en el compromiso de los alumnos con respecto a los vídeos para dilucidar si, efectivamente, las preguntas intercaladas tienen un efecto beneficioso sobre la motivación del alumno.

El uso de un MOOC real, con completa libertad para el alumno para consultar los materiales al realizar los test, puede interferir con el objetivo de medir el aprendiza-

je, especialmente en cuanto a los resultados en retención. Futuros trabajos deberán tratar de limitar el acceso a los materiales al realizar los test para evitar esa interferencia.

Por último, los resultados obtenidos en este trabajo deben ser replicados en otros contextos, con MOOCs que versen sobre otras materias y con vídeos de distintas características, antes de poder extraer ninguna conclusión clara sobre el beneficio de intercalar o embeber la evaluación formativa en los vídeo pódcast.

5.2. Trabajo futuro

Obviando como trabajo futuro el solventar las limitaciones expuestas en la sección 5.1 y el clarificar si usar preguntas embebidas tiene efectos beneficiosos sobre la aplicación de los conceptos aprendidos a otros contextos, entre los posibles trabajos a futuro estarían algunos análisis que han quedado fuera del alcance de este trabajo y que podrían haber sido interesantes como, por ejemplo: tratar de comparar los resultados entre los grupos solamente en aquellas preguntas de transferencia que tratan el mismo concepto que las preguntas de refuerzo; estudiar si existe diferencia en el número de veces que se contestan las preguntas de refuerzo entre las cohortes; o analizar si existe relación entre el uso de las preguntas de refuerzo y el resultado en los test.

Dado que no existe en la literatura, hasta donde se tiene conocimiento, ningún artículo en una publicación revisada por pares que aborde el tema tratado mediante un experimento puramente aleatorio, está previsto utilizar los resultados obtenidos en el presente trabajo para escribir uno y presentarlo a una revista del Q1 para optar a su publicación. Por otro lado, también está previsto que, dentro de las actividades de investigación del grupo de innovación docente miniXmodular, se utilice el texto ilustrado asociado al MOOC para hacer pruebas en un proyecto, presentado al Vicerrectorado de Digitalización e Innovación, orientado a construir un aplicativo consiste en un conversor de formatos orientado principalmente a transformar material docente digital y accesible generado mediante programas de autoría (MS Word, OpenOffice, LibreOffice o LaTeX) en formatos empaquetados (ePub 3) o en versiones compatibles con plataformas web (edX). Por último, está previsto que el MOOC desarrollado continúe estando disponible para futuras actuaciones formativas y como elemento de promoción de los másteres de la UNED.

Como posibles líneas futuras de investigación, se proponen las siguientes. Primero, estudiar el impacto que tiene en el aprendizaje y en la motivación el tiempo que transcurre entre la explicación de un concepto y la práctica de la evaluación formativa, en un entorno de enseñanza a través de medios virtuales. Existen indicios que indican que son más útiles aquellos test que van más cerca del contenido (Dempster

(1991)) pero no se ha encontrado referencia alguna a este aspecto en los distintos trabajos sobre la evaluación formativa y el *feedback* analizados en el capítulo 2. Estos trabajos se centran más en el tiempo que pasa desde que se realiza la evaluación formativa hasta que el alumno recibe el *feedback*, pero no entre que se explica un concepto y se realiza la evaluación. Segundo, ahondar en qué tipo de *feedback* tiene un mayor beneficio al usarlo en preguntas embebidas. Por ejemplo, estudiar si con un *feedback* elaborado más complejo que el utilizado en este experimento es posible obtener mejores resultados. Y, tercero, y relacionada con la propuesta anterior, una potencial y apasionante línea de investigación estriba en analizar el impacto de una evaluación formativa –incluido el *feedback*– que incorpore características adaptativas, de forma que se adapte al ritmo, conocimientos previos y estilo de aprendizaje del alumno.

5.3. Conclusiones

En esta memoria se ha presentado un estudio del efecto que tiene sobre el aprendizaje el uso de preguntas de evaluación formativa en pódcast de vídeo en el contexto de un MOOC. En concreto, se ha realizado una comparación entre el uso de lo que se ha denominado «preguntas de refuerzo» (preguntas de evaluación formativa con su correspondiente *feedback*) presentadas nada más finalizar en el vídeo la explicación de un concepto (*in-video quiz*) respecto a presentación de la misma pregunta al finalizar el vídeo (*post-video quiz*). La comparación se ha realizado mediante un experimento aleatorio para la realización del cual se ha diseñado y construido un MOOC real que trata de trata de explicar las bases del aprendizaje automático para todo tipo de públicos.

Con anterioridad se había estudiado cómo afectaba el uso de preguntas embebidas al compromiso con los vídeos o analizado el comportamiento del alumno con respecto a las mismas (Cummins et al. (2016); Kovacs (2016); Brinton et al. (2016)) y, a pesar de que ambas modalidades –preguntas embebidas o preguntas posvídeo– son bastante comunes en los MOOCs comerciales, solamente se había localizado un estudio que tratara de comparar el uso de *in-video quizzes* con el uso de *post-video quizzes* (Vural (2013)), que cuenta con una configuración cuasiexperimental. Por ello, en este trabajo se ha tratado de cubrir ese hueco en la investigación comparando esas dos condiciones mediante un experimento puramente aleatorio.

Los resultados del experimento indican que el uso de preguntas intercaladas en el vídeo es beneficioso sobre la capacidad de retención. También se han obtenido mejores resultados para el grupo experimental en transferencia, mejorando las medias en todos los test, y comparativamente más personas del grupo experimental finalizaron el curso, pero ambos resultados carecen de significancia estadística. Dados

los resultados obtenidos y el bajo coste de implementación de la medida, se recomienda el uso en los MOOCs de preguntas embebidas en los vídeos en lugar de esperar a que finalice el vídeo para realizarlas.

Con respecto al uso de la plataforma Open edX para la realización de experimentos, la conclusión es que su uso es perfectamente factible y puede ser muy útil para realizar observaciones en un entorno de aprendizaje real. La configuración para albergar los grupos experimentales y para generar el contenido no es compleja, pero, en cualquier caso, se recomienda cerciorarse de que la plataforma cuenta con el soporte para el experimento concreto que se desee realizar mediante pruebas previas en una instalación propia de Open edX, o bien, mejor aún, habilitando un entorno de pruebas al efecto que cuente con las mismas características que el entorno de «producción». Como propuestas de mejora, si la plataforma se generaliza como medio para la realización de experimentos, es de esperar que sean necesarias distintas configuraciones específicas que no tengan un soporte nativo por lo que sería necesario tener en cuenta este aspecto a la hora de dar soporte al investigador. Más importante aún es el conseguir la disponibilidad de los datos completos de los *logs* de edX para poder realizar trabajos de análisis y minería avanzados.

A. Ejemplo del material de una lección del curso

Como muestra del contenido desarrollado en para el MOOC, en este apéndice se presenta el material correspondiente a la lección 5 del curso, titulada «Selección de modelo y análisis del error», en la que se tratan conceptos como el conjunto de validación, el uso de éste para realizar una selección del modelo más apropiado, el sobreajuste o el compromiso sesgo-varianza.

Como se verá, además del vídeo que acompaña esta lección, el alumno tenía a su disposición un texto de acompañamiento que refuerza y amplía los contenidos que cubre el vídeo docente. El orden en la sección del vídeo que se presenta en este apéndice es el que tuvo a su disposición la cohorte experimental. Se recuerda que la cohorte de control accedía al vídeo completo, sin segmentar, y que se le presentaban las preguntas de refuerzo al finalizar éste. Asimismo, se recuerda que ambos grupos solamente diferían en este aspecto del curso y que el resto de materiales eran los mismos, en concreto el texto de acompañamiento y los foros del curso.

Al final de este apéndice se detallan las referencias utilizadas para el desarrollo del contenido del curso.

A.1. Lección 5 del MOOC: «Selección de modelo y análisis del error»

A.1.1. Introducción

Una vez que ya se ha tratado qué es un modelo, cómo se entrena y que hay modelos más sencillos –lineales– y otros que conforman fronteras de decisión tremendamente complejas, en este capítulo se presentan de forma intuitiva algunos conceptos fundamentales en el mundo del aprendizaje automático. Son conceptos que es importante conocer, aunque no sea en profundidad, si se va a tener relación con algún proyecto de desarrollo o uso de un modelo de aprendizaje automático y que deben guiar el proceso de desarrollo de un modelo de aprendizaje.

A.1.2. Selección de modelo

Ya se ha tratado la cuestión de la importancia de conservar unos datos separados, que no se usen para el entrenamiento del modelo, a fin de poder probar el rendimiento del modelo sobre datos «limpios» y así comprobar que el modelo tiene capacidad de generalizar. Pero ¿qué ocurre si durante el proceso de desarrollo de un modelo se desea probar distintos algoritmos de aprendizaje –por ejemplo: una regresión lineal, un árbol de decisión y una red neuronal– o, más aún, si quiero probar distintas configuraciones de esos algoritmos o de los parámetros que gobiernan su comportamiento –los llamados **hiperparámetros**–? (por ejemplo, en una red neuronal, variar el número de neuronas de las distintas capas o probar con distinto número de capas).

Éste es en realidad el escenario más común: el desarrollo de un modelo de aprendizaje es un proceso iterativo en el que se generan unas características para utilizar en el entrenamiento, se escoge un algoritmo con una configuración, se entrena un modelo, se prueba sobre datos que no se hayan utilizado en el entrenamiento y el resultado obtenido sirve al equipo que está desarrollando el modelo para aprender del resultado obtenido y volver a empezar. El proceso se repite hasta que se consiga un resultado suficientemente satisfactorio para el objetivo buscado. A este proceso se le denomina *selección del modelo*.

Si en todas las pruebas que se realizan para buscar el modelo más apto –buscando que el rendimiento sobre el problema a resolver sea el mejor posible– se utiliza el conjunto de prueba para valorar el resultado, se estarán «ensuciando» esos datos de prueba en cuanto a que el equipo de desarrollo estará aprendiendo sobre el problema que tiene entre manos y de alguna manera dirigiendo la solución para que funcione bien sobre los datos que tiene disponibles para probar. Esto provoca que no sea posible tener total seguridad de que el rendimiento del modelo elegido se mantendrá cuando se aplique sobre datos nuevos.

Cuando las pruebas realizadas alcanzan un número importante es recomendable utilizar un tercer grupo de datos, que puede recibir varios nombres, siendo uno de los más comunes el de **conjunto de validación**. También es llamado conjunto de desarrollo haciendo referencia a que los resultados sobre este conjunto guían el proceso de desarrollo del modelo.

Existen técnicas para evitar el uso de un conjunto de validación si no se dispone de suficientes datos. La más común es seguramente la *validación cruzada*, que veremos en un momento. En cualquier caso, si el número de datos es suficientemente grande, es preferible usar un conjunto de datos separado para realizar la tarea de validación y desde luego, como norma general, es mejor no utilizar el conjunto de pruebas.

Por tanto, salvo que se use validación cruzada, la primera tarea a realizar al comenzar el desarrollo de un modelo, antes de cualquier entrenamiento, debe ser la sepa-

ración de los datos en los **tres conjuntos** comentados: **entrenamiento, validación y prueba**.

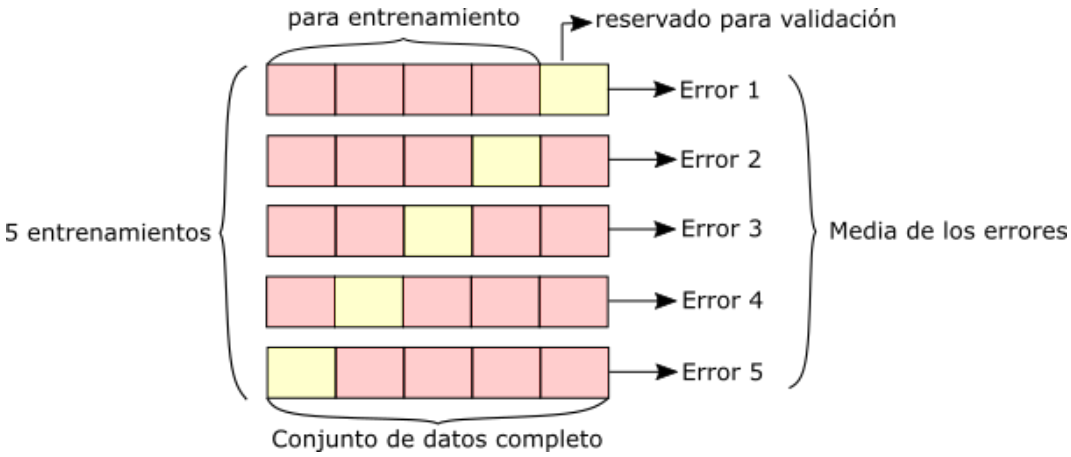
¿Cuántos datos o ejemplos es necesario separar en los conjuntos de validación y prueba? En realidad, depende de los datos disponibles. Si se dispone de un número de ejemplos que ronda las pocas decenas de miles o menos, algo parecido a la regla del $70 - 30\,\%$ comentada en lecciones anteriores es bastante razonable. Esto podría convertirse en un $60 - 20 - 20\,\%$ cuando la división se realiza en tres conjuntos. A medida que el conjunto de datos total es mayor que esas pocas decenas de miles, ese porcentaje destinado a validación y pruebas podrá ir descendiendo. Por ejemplo, con 10000 ejemplos para el conjunto de validación y otros 10000 para el de pruebas debería ser más que de sobra en casi todos los casos y lo normal es ver conjuntos de validación y pruebas mucho más pequeños. Cuanto más datos se reserven para los conjuntos de validación y prueba, más seguridad tendremos de que el error obtenido será muy parecido al error con datos futuros (siempre que los datos disponibles provengan de la misma distribución –estén producidos por el mismo «motor»– que los datos sobre los que usaremos el modelo, como veremos posteriormente).

A.1.3. Validación cruzada

Puede darse el caso de que el volumen de datos disponible no aconseje separar un conjunto de datos para validación ya que, como se ha dicho, no sería posible utilizar esos datos para el entrenamiento. Para esos casos tenemos disponible una técnica denominada validación cruzada que nos permite la posibilidad de no perder datos para el entrenamiento del modelo a cambio de necesitar un mayor coste computacional (tiempo de entrenamiento).

La idea subyacente es que se repiten ciclos de entrenamiento y validación con distintas particiones del conjunto de entrenamiento de partida. La forma más común de validación cruzada es la validación cruzada «k-veces» (en inglés *k-folds*) que consiste en dividir el conjunto de datos en k partes y en realizar k entrenamientos, reservando cada vez una de las partes para la validación y, por tanto, sin usar en el entrenamiento. Finalmente se cogen las k medidas de error y se realiza la media. Es muy común usar un valor de k de 5 o 10.

En la siguiente imagen se puede ver un esquema para el caso de una validación cruzada de tipo «5-veces».



A.1.4. Vídeo - Sesgo, varianza y análisis del error

Tramo 1 del vídeo «Sesgo, varianza y análisis del error» (https://youtu.be/sFL6or58-VA, entre los minutos 00:00 y 10:08).

Pregunta de refuerzo 1:

Cuestión sobre sesgo/varianza y sobreajuste 1

1 point possible (ungraded)

Se disponen de 50 ejemplos con 5 características para entrenar un clasificador. Para el entrenamiento se decide usar una red neuronal con 10 capas ocultas de 10 neuronas cada una. Con el modelo generado por esa red se obtiene un resultado perfecto sobre los datos de entrenamiento (no comete ningún error), pero al probar el modelo sobre el conjunto de validación el error que se obtiene es bastante abultado.

¿Qué cree que está ocurriendo?

- ○ El modelo generado sobreajusta a los datos de entrenamiento debido al pequeño número de datos de entrenamiento disponible
- ○ El modelo generado sobreajusta a los datos de entrenamiento debido a la complejidad de las soluciones que es capaz de generar la red neuronal
- ○ El modelo generado tiene mucha varianza
- ○ Las tres respuestas anteriores son correctas

Enviar

Tramo 2 del vídeo «Sesgo, varianza y análisis del error» (https://youtu.be/sFL6or58-VA, entre los minutos 10:09 y 15:45)

Vídeo - Sesgo, varianza y análisis del error - Tramo 2

Pregunta de refuerzo 2:

Cuestión sobre sesgo/varianza y sobreajuste 2

1 point possible (ungraded)

Se sabe que para un problema concreto el error mínimo teórico que es posible obtener es del 1%. Se dispone de un clasificador para ese mismo problema que obtiene un 35% de error sobre los datos de entrenamiento y un 40% sobre los datos de validación,

¿cuál es el principal problema que presenta el clasificador y cuál puede ser un camino que tome el equipo de desarrollo para solucionarlo?

- ○ Existe un problema de sesgo y el equipo podría probar con modelos más simples para tratar de solucionarlo
- ○ Existe un problema de sesgo y el equipo podría probar con modelos más complejos para tratar de solucionarlo
- ○ Existe un problema de varianza y el equipo podría probar con modelos más complejos para tratar de solucionarlo
- ○ Existe un problema de varianza y el equipo podría probar con modelos más simples para tratar de solucionarlo

Enviar

Tramo 3 del vídeo «Sesgo, varianza y análisis del error» (https://youtu.be/sFL6or58-VA, entre los minutos 15:47 y el final del vídeo)

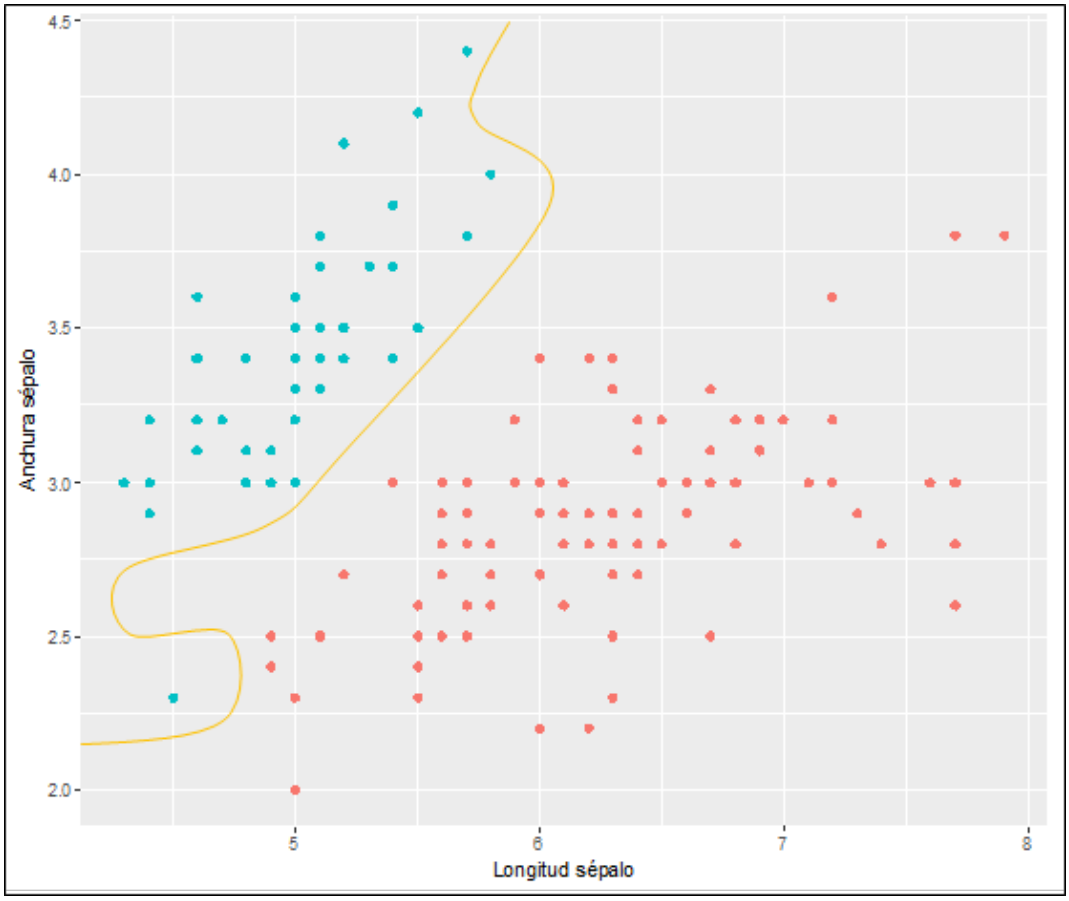
Vídeo - Sesgo, varianza y análisis del error - Tramo 3

A.1.5. El sobreajuste

En algún momento ya hemos tratado sobre la capacidad de generalización de los modelos. Es la capacidad deseada de que mantengan su rendimiento sobre datos con los que no han sido entrenados, que es en realidad el objetivo final deseado. Para comprobar que los modelos obtenidos son capaces de generalizar se usa la comparación entre el error que cometen sobre los datos de entrenamiento y el error sobre los datos de validación. Si la diferencia en el porcentaje de error no es muy grande podemos decir que el modelo generaliza correctamente y por tanto concluir que ha aprendido.

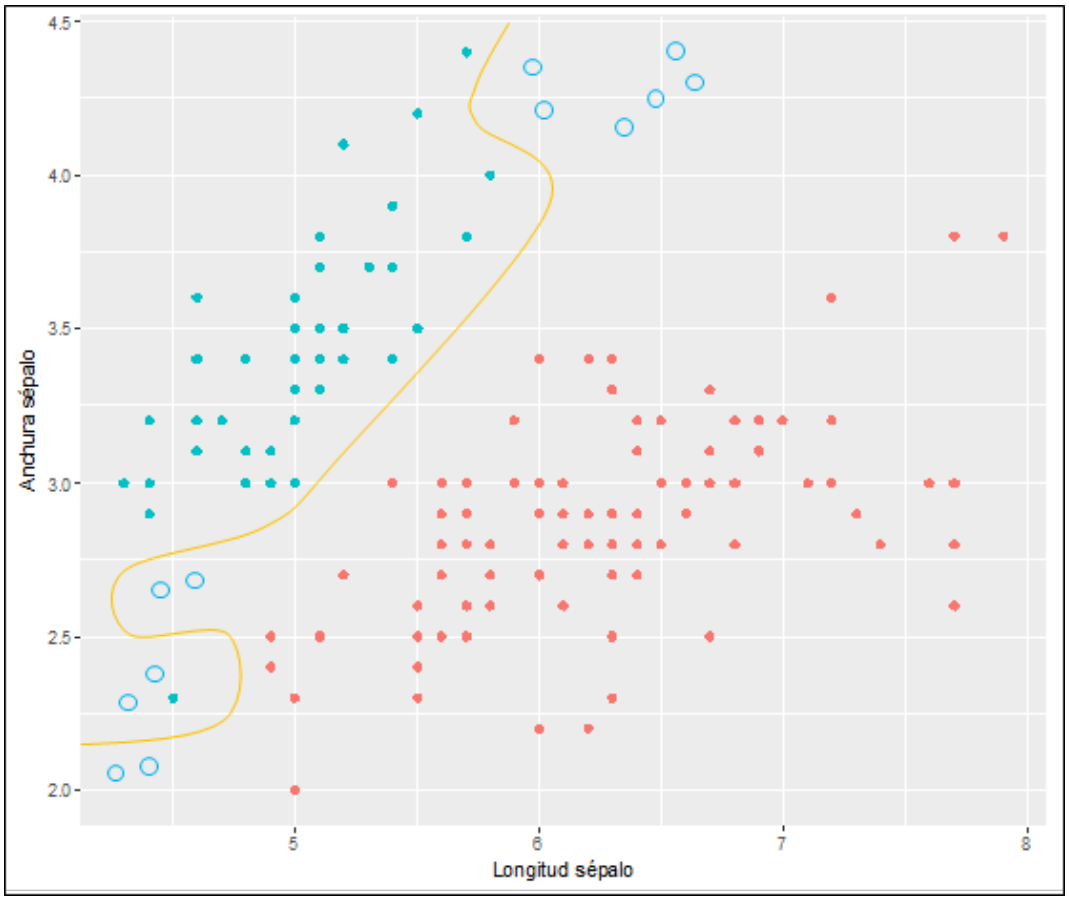
Sin embargo, es muy común encontrarse con que el modelo entrenado obtiene muy poco error sobre los datos de entrenamiento y que posteriormente, el error se dispara sobre los datos de validación. Es el problema conocido como **sobreajuste**.

Veamos un ejemplo que ya hemos usado anteriormente:



En esta gráfica se ve cómo la frontera de separación de las clases –línea amarilla– es, intuitivamente, más compleja de lo que parecería necesario y que ajusta demasiado a los datos realizando unas curvas realmente llamativas. La realidad es que **hemos**

usado un modelo que es capaz de generar hipótesis –la línea amarilla– cuya complejidad no puede ser «domada» por los datos disponibles. Por ello, clasifica muy bien los datos de entrenamiento (comete un error muy pequeño –o en este caso–), pero en cuanto se encuentra con datos nuevos no lo hace igual de bien:



Por ello se dice que el modelo resultante **sobreajusta** a los datos disponibles (hablando en lenguaje coloquial podríamos decir que «se pasa de lista»).

Recordemos que, a priori, no se dispone de la capacidad de conocer todos los datos posibles de las clases que se desean clasificar (de lo contrario no tendríamos ningún problema entre manos) y que se trata de generar un modelo que extraiga los aspectos generales comunes en los datos disponibles sin dejarse engañar por los casos concretos y extraños (que se denominan comúnmente «ruido»).

Es un problema muy fácil de ver en las gráficas, pero cuando ya no se trabaja en dos/tres dimensiones, es necesario fijarse en los datos de error obtenidos sobre los datos de entrenamiento y validación. Continuando con la analogía de los problemas de matemáticas y el examen, cuando se sobreajusta el alumno habría memorizado los problemas vistos en clase sin haber comprendido y aprendido el mecanismo

subyacente que le permitiría resolver los problemas nuevos planteados en el examen –siempre que sean del mismo tipo que los vistos en clase–.

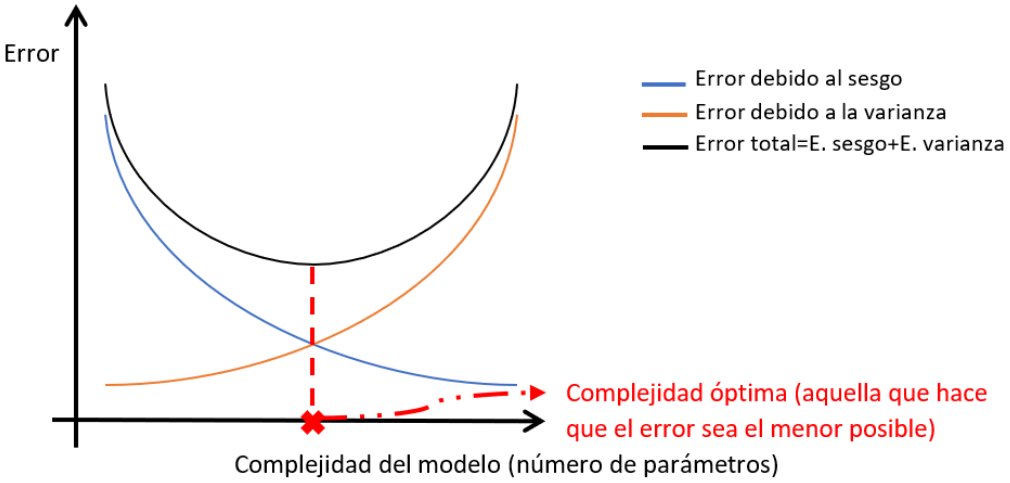
A.1.6. El compromiso entre aproximación y generalización

Aunque pueda resultar contraintuitivo, en muchas ocasiones es recomendable permitir que el modelo de aprendizaje cometa fallos sobre los datos con los que se entrena –limitar la capacidad de aproximación del modelo– a fin de obtener una solución que no ajuste tanto a los datos disponibles pero que, a cambio, tenga una mejor capacidad de generalización.

Es decir, se trata de restringir la capacidad expresiva de las hipótesis que se contemplan como posibles soluciones –restringir lo que los modelos son capaces de aprender– a una complejidad que pueda ser «domada» por los datos disponibles para evitar que se produzca un sobreajuste. La complejidad de las hipótesis que es capaz de generar un modelo de aprendizaje depende, en muchos casos, del número de parámetros que tenga el modelo –por ejemplo, en las redes neuronales, el umbral de cada neurona de la red y el peso de cada uno de los enlaces entre las diferentes neuronas–.

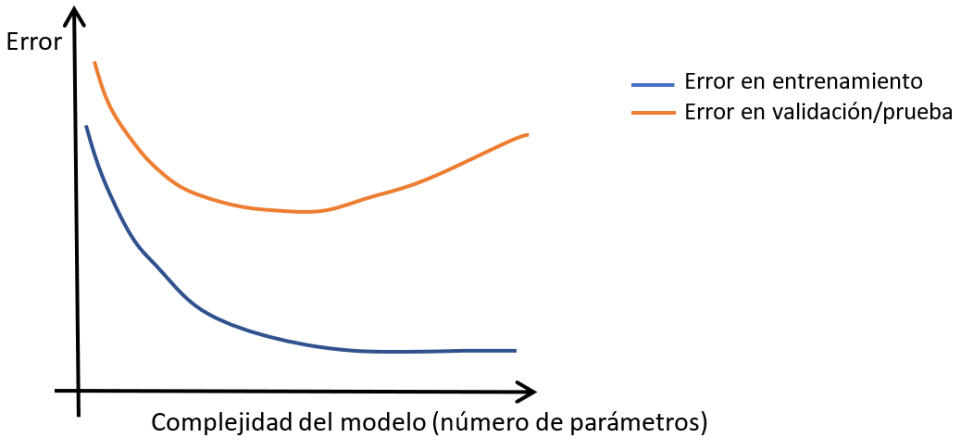
A la sensibilidad que tiene el modelo con respecto a los datos con los que es entrenado en la generación de una determinada solución, se le denomina técnicamente *varianza*. El *sesgo*, en cambio, mide la capacidad del modelo de aproximar a la solución ideal (aquella que separaría de forma perfecta todos y cada uno de los casos que nos puedan llegar en el futuro y que es, lógicamente, desconocida). El error total que comete un modelo es la suma del sesgo y de la varianza.

El compromiso entre aproximación y generalización se da porque, típicamente, si se busca reducir la varianza (utilizar modelos más simples, que dependan menos de los datos concretos con los que se entrenan), el sesgo aumentará mientras que, al contrario, si se busca reducir el sesgo (utilizar un modelo más complejo que se pueda acercar más a la solución ideal), la varianza aumentará.



Es decir, el sesgo y la varianza capturan el compromiso existente entre capacidad de aproximación y capacidad de generalización del modelo.

El error sobre los datos de entrenamiento y validación con respecto a la complejidad del modelo sigue, típicamente, un comportamiento parecido a lo siguiente:



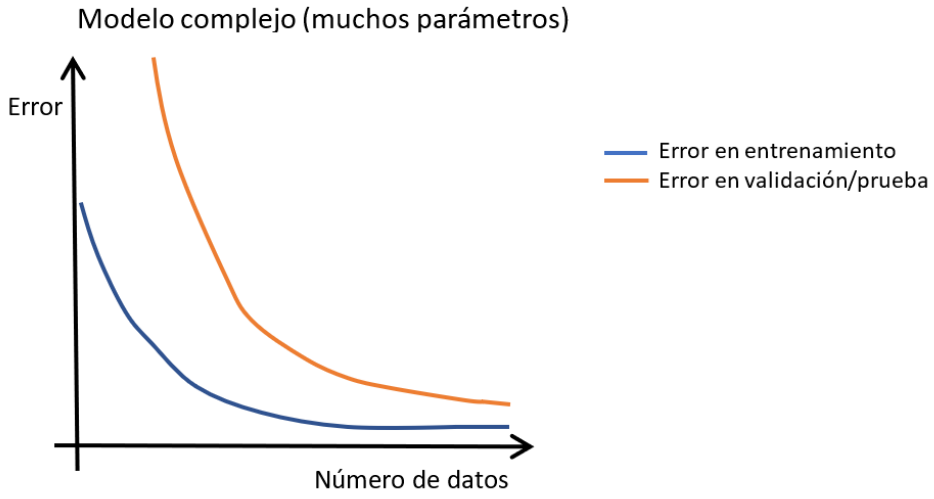
Adaptado de: Abu-Mostafa, Y. S., Magdon-Ismail, M., & Lin, H. T. (2012). *Learning from data* (Vol. 4). New York, NY, USA:: AMLBook.

El error sobre los datos de entrenamiento mejora conforme se aumenta la complejidad del modelo hasta llegar a un punto donde no mejora más (a menudo, cuando ya no comete ningún error) mientras que el error sobre los datos que el modelo no ve siempre suele ser mayor que el error en entrenamiento y, en un principio mejora conforme se usan modelos más complejos, pero llega un punto donde la tendencia

cambia y el error comienza a aumentar. **El objetivo es encontrar ese modelo cuya complejidad hace lo más pequeño posible el error en validación.**

Cuantos más datos se tengan y cuanto mejor calidad tengan –menos datos erróneos– modelos más complejos se podrán entrenar sin sufrir sobreajuste (antes de que el error en validación comience a ascender).

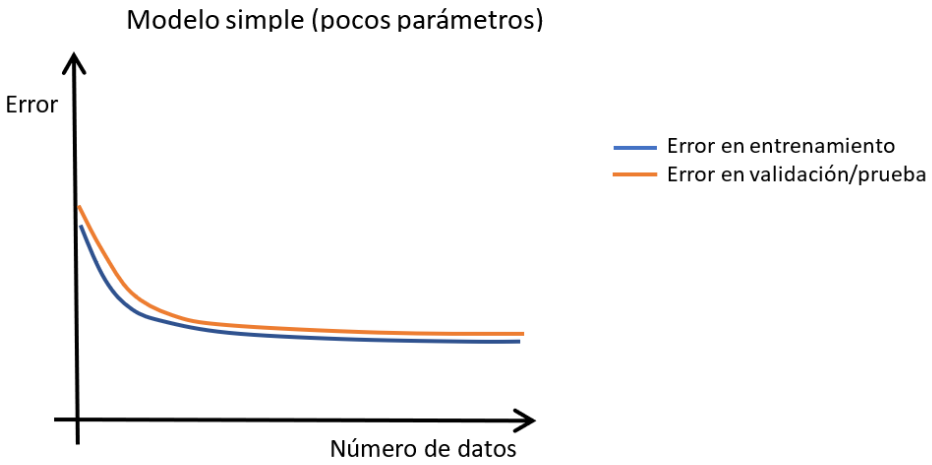
Cuando usamos un modelo complejo el error evoluciona con respecto al número de datos disponible de forma parecida a lo siguiente:



Adaptado de: Abu-Mostafa, Y. S., Magdon-Ismail, M., & Lin, H. T. (2012). *Learning from data* (Vol. 4). New York, NY, USA:: AMLBook.

Se puede ver que el error en entrenamiento desciende mucho más rápidamente que el de validación conforme se tienen más datos. Esto es algo típico en modelos medianamente complejos. Cuando la diferencia entre el error de entrenamiento y validación es alta, sabemos que tenemos problemas de generalización y que **necesitamos más datos o usar un modelo más simple**, es decir **reducir la varianza** (por ejemplo, si estamos usando una red neuronal, ir a por un modelo lineal).

Cuando el modelo que usamos es sencillo (como decíamos, por ejemplo, un modelo lineal), normalmente el modelo generaliza bien y tanto el error de entrenamiento y de validación descienden de forma similar cuantos más datos se tienen.



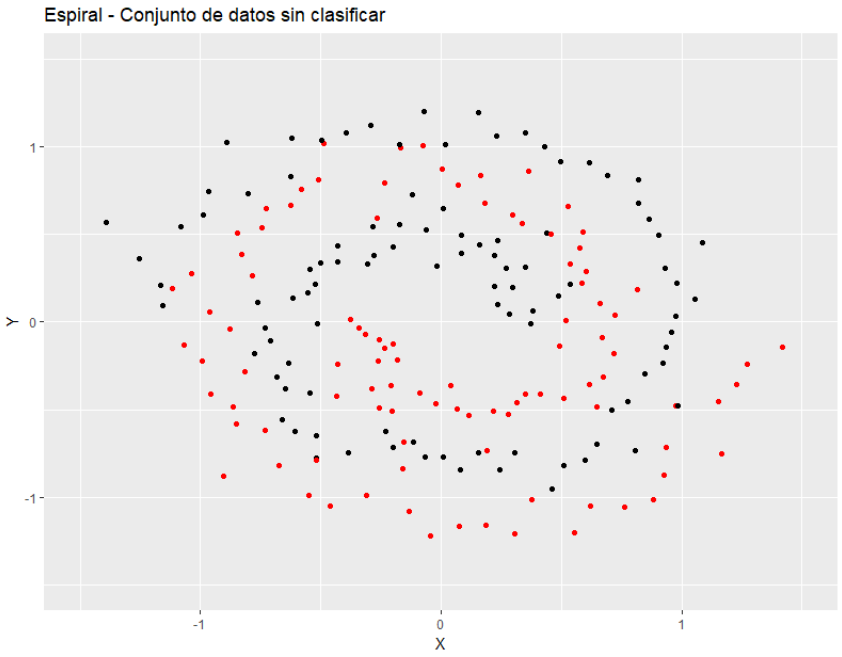
Adaptado de: Abu-Mostafa, Y. S., Magdon-Ismail, M., & Lin, H. T. (2012). *Learning from data* (Vol. 4). New York, NY, USA:: AMLBook.

Sin embargo, el mejor error de entrenamiento que se obtendrá será, en general, peor que el error obtenido con modelos más complejos. Si el resultado no es satisfactorio y se necesita un mejor rendimiento para la utilidad que se le vaya a dar al modelo, **es necesario plantearse el uso de modelos más complejos**, es decir, **reducir el sesgo**.

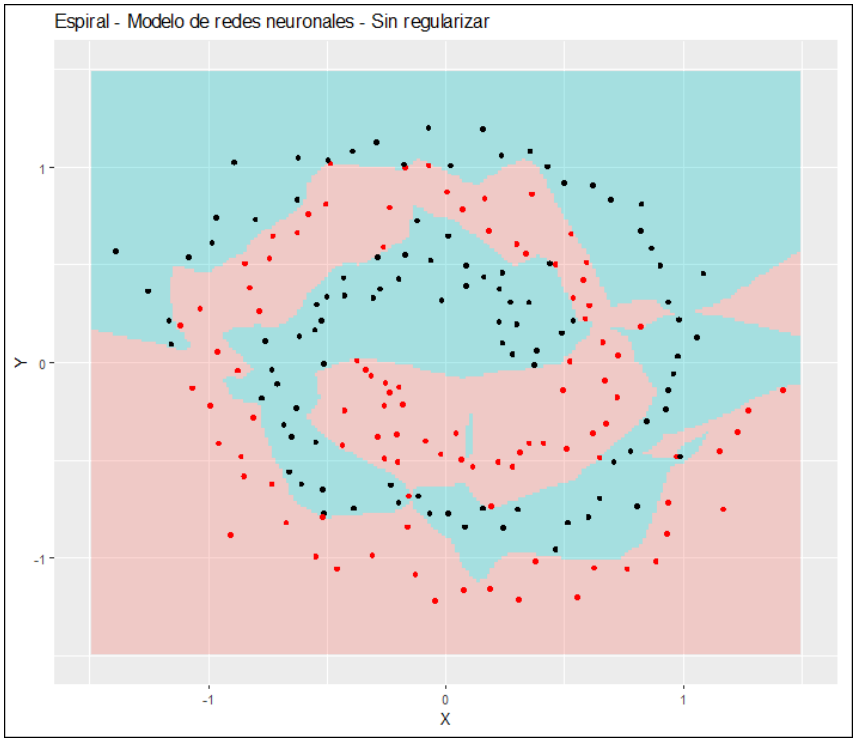
A.1.7. Regularización

Existe una técnica, muy utilizada, para forzar a modelos complejos a que prefieran las hipótesis más simples sobre las más complejas: **la regularización**. No se profundizará en esta técnica, pero es importante conocer que existe. Fundamentalmente, se trata de inducir a que el modelo no trate de ajustar a todo lo que vea y obligarle a quedarse con lo más general de los datos reduciendo de esta manera la posibilidad de sobreajuste a los datos. Se ve mejor el efecto con un ejemplo:

Si tenemos un conjunto de datos como el de la figura, con puntos de dos clases, roja y negra:

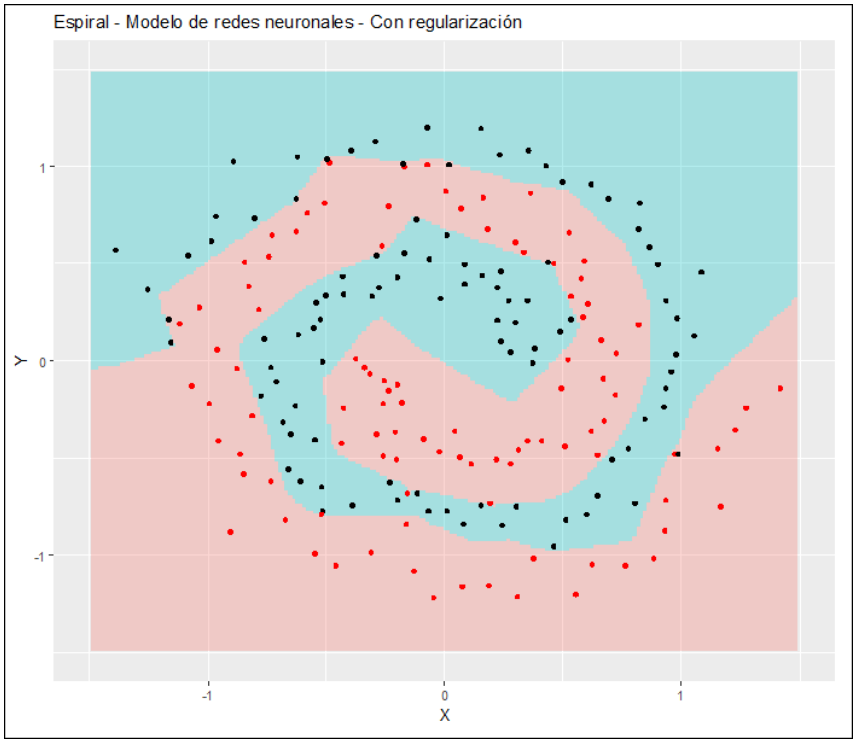


y tratamos de clasificarlo mediante una red neuronal diseñada con un número de neuronas deliberadamente exagerado para provocar que sobreajuste a los datos, el resultado es el siguiente:



Se puede apreciar cómo, en varias ocasiones, el límite entre las regiones se desvía notablemente solo para «capturar» dentro de su región a uno de los puntos que les corresponde.

Si entrenamos la misma red neuronal aplicando regularización (obligando al modelo a generar hipótesis más simples), el resultado es el siguiente:



Se puede comprobar como ahora el límite entre las dos regiones varía de forma menos drástica y que, de hecho, se parece bastante más a una espiral, que es en realidad el «motor» que se ha utilizado para generar los datos y, por tanto, lo que deseamos aproximar mediante el modelo generado por la red neuronal.

A.1.8. Análisis del error

Una vez visto, de forma muy somera y superficial, el compromiso que se debe tener siempre presente entre la capacidad de aproximación y de generalización de un modelo con respecto a los datos con los que se cuenta, en este apartado se verá cómo es posible usar ese conocimiento para dirigir el proceso de desarrollo de un modelo de aprendizaje.

Esta sección y la siguiente están basadas en las recomendaciones del profesor Ng, recogidas en su libro *Machine learning yearning* y en sus MOOCs. Se puede encontrar más información sobre ellos en la sección de referencias del curso.

Como ya se ha explicado, el desarrollo de un modelo es un proceso iterativo, donde lo fundamental es tener la capacidad de construir un modelo de forma lo más rápida posible, evaluar cómo se comporta el modelo, y, a partir de esa información, iterar y construir nuevos modelos. Usaremos el análisis del error para saber en qué dirección debemos movernos a la hora de iterar.

Para poder ser capaces de realizar un análisis del error nos falta conocer un dato: cuál es el menor error que se puede cometer a la hora de abordar el problema que se trata de solucionar. Es decir, si estamos tratando de clasificar todas las imágenes en que aparecen árboles, saber cuál es el mejor resultado que es posible obtener. Esto puede sonar a tontería y podemos pensar que una persona siempre podrá clasificar correctamente las imágenes que tienen árboles y separarlas de las que no, y sin duda los seres humanos realizamos muy bien esta tarea, pero debemos pensar en qué pasaría si las imágenes son de muy mala calidad, están muy mal iluminadas, etc. Existe un límite teórico que delimita lo bien que se puede realizar una determinada tarea de clasificación sobre unos datos con unas características concretas, que se denomina **tasa de error de Bayes** (*Bayes error rate*). Nada, ni nadie, puede conseguir un error menor a la tasa de error de Bayes.

Sin embargo, aunque ese error de Bayes se puede estimar, no se puede conocer, por lo que en muchas ocasiones se utiliza una manera muy sencilla de estimar un valor cercano a ese error: **cómo de bien es capaz el ser humano de realizar la tarea para la que queremos entrenar un clasificador** (error humano). Mientras no se haya entrenado un clasificador que cometa un error menor que el que cometen los seres humanos, este valor es muy útil como estimación del error de Bayes para ayudar al equipo de desarrollo a dirigir el proceso de entrenamiento de un modelo de aprendizaje.

Una aclaración: cuanto menor sea el error que consiga un humano, más cercano será ese valor al error de Bayes, que es el que de verdad nos interesa, y por eso, se debe tomar como referencia aquel menor error del que se tenga constancia (ojo, sobre datos que tengan las mismas características que tienen los datos sobre los que queremos que trabaje el modelo que estamos desarrollando) es decir, si tenemos un especialista que logra un error menor que una persona común, ese es el resultado que nos interesa y si contamos con un grupo de expertos que todavía lo hacen mejor, nos decantaremos por tomar ese valor como referencia del error de Bayes.

A partir de que ya disponemos de una referencia con la que comparar el error de entrenamiento, vamos a ver unos casos concretos:

	% error de clasificación	
	Escenario 1	**Escenario 2**
Error humano	1 %	1 %
Error en entrenamiento	4 %	2 %
Error en validación	5 %	5 %

En el escenario 1, la diferencia entre el error humano y el error de entrenamiento es bastante mayor que entre el error de entrenamiento y el error de validación. Esto significa que el equipo debería centrarse en **disminuir el sesgo** buscando un modelo que sea capaz de ajustar mejor a los datos de entrenamiento (típicamente, un modelo más complejo).

En el escenario 2 la situación es la contraria, la diferencia entre el error de entrenamiento y de validación es mayor que la diferencia entre el error humano y el error de entrenamiento. Ya sabemos que cuando existe mucha diferencia entre el error que comete el modelo sobre los datos de entrenamiento y el error sobre los datos de validación, es que tenemos un problema de generalización, y **deberíamos fijar los esfuerzos en la reducción de la varianza del modelo**: en general, usar un modelo más sencillo u obtener más datos con los que entrenar el modelo y así conseguir –siguiendo con la analogía– «domar» su complejidad.

A.1.9. El objetivo al que apunta el modelo

Para finalizar el capítulo, se trata el aspecto de definir, de forma que se facilite el proceso de construcción del modelo, un objetivo que guíe el proceso de desarrollo.

Hay dos aspectos que influyen en cuál es el objetivo hacia el cual apuntamos nuestro modelo durante el proceso de construcción: los datos con los que se cuenta (los datos utilizados para entrenar y seleccionar el mejor modelo), y la métrica (exactitud, precisión, sensibilidad, etc.) que se defina como más importante para el uso que se le quiera dar al modelo. Vamos a dejar para el siguiente capítulo el tema de los datos y centrémonos en las métricas.

Para hacer más fácil para el equipo de desarrollo la evaluación de los distintos modelos obtenidos y la comparación entre ellos, es muy recomendable analizar, antes de empezar el desarrollo, qué tipo de valor preferimos que el proceso de desarrollo trate de optimizar. Es decir: ¿Se desea obtener el modelo con un menor número de errores de clasificación? ¿o se prefiere aquel que minimice los falsos positivos? ¿O es quizás el modelo deseado aquel con menor tasa de error, siempre que los falsos negativos no excedan de un porcentaje determinado?

Definir qué se desea ayudará al equipo de desarrollo a saber cómo dirigir su exploración de la búsqueda del mejor modelo para el uso que se pretende. Es recomendable definir una única métrica a optimizar puesto que facilita el proceso de comparación de modelos. Si se necesita relacionar varios valores, se definirá la métrica a optimizar bajo una serie de condiciones (restricciones), por ejemplo: «Queremos el modelo con menor tasa de error de clasificación siempre que los falsos positivos no excedan del 5 %»

A.1.10. Test de evaluación de la lección 5

Pregunta 1
¿Cómo se denomina al proceso por el cual se realizan las distintas elecciones en busca del modelo que mejor se comporte para el problema que se trata de resolver?

- ○ Configuración del modelo
- ○ Selección del modelo
- ◉ Adaptación del modelo ✖
- ○ Refinado del modelo

Pregunta 2
A la capacidad de un modelo de mantener un rendimiento parecido al obtenido en el entrenamiento sobre datos con los que no se ha entrenado, se le denomina:

- ○ Capacidad de abstracción
- ◉ Capacidad de extensión ✖
- ○ Capacidad de generalización
- ○ Capacidad de predicción

Pregunta 3
¿Existe siempre la posibilidad de reducir el sesgo y la varianza a la vez?

- ○ Sí, si se usa el modelo adecuado
- ○ No, típicamente, al intentar reducir uno, tiende a aumentar el otro
- ◉ Sí, por definición siempre que se reduce el sesgo se reduce la varianza ✖
- ○ No, solamente es posible si se usan redes neuronales profundas

Pregunta 4
La sensibilidad que tiene un modelo de aprendizaje con respecto a los datos con que es entrenado la mide:

- ○ La varianza
- ○ El sesgo
- ○ La precisión
- ◉ El valor F1 ✖

Pregunta 5
Si la diferencia entre los valores de error en entrenamiento y validación es alta, ¿qué NO tendría, en principio, demasiado sentido hacer?

- ○ Probar con un modelo de aprendizaje más complejo
- ○ Tratar de conseguir más datos etiquetados con los que entrenar el modelo
- ◉ Probar con un modelo de aprendizaje más sencillo ✖
- ○ Usar técnicas de regularización

Pregunta 6
¿En qué consiste la validación cruzada K-veces?

- ○ Dividir aleatoriamente el conjunto de datos en K partes, entrenar el modelo con K-1 partes, hacer la validación sobre la parte restante calculando el error de predicción, repetir este proceso considerando en cada ocasión una parte distinta de validación y hacer la media de los K errores de predicción
- ◉ Dividir aleatoriamente el conjunto de datos en K partes, entrenar el modelo en una de esas partes, hacer la validación sobre las K-1 partes restantes calculando el error de predicción, repetir este proceso considerando en cada ocasión una parte distinta de entrenamiento y hacer la media de los K errores de predicción ✖
- ○ Tomar aleatoriamente dos partes del conjunto de datos, por ejemplo 80% y 20%, y considerar la parte más grande como conjunto de datos de entrenamiento y validar sobre la parte más pequeña restante, calculando el error de predicción
- ○ Es un método supervisado en el que se buscan los K vecinos más próximos según una distancia prefijada

Pregunta 7
Una empresa contrata un servicio de desarrollo de un modelo predictivo. Cuentan con 50 000 ejemplos de X->Y que le entregan a la empresa especializada para que pueda trabajar en desarrollar el modelo. El modelo final entregado alcanza un 92% de exactitud. En la memoria descriptiva de los trabajos que entrega la empresa se puede ver que han dividido los datos en dos conjuntos: uno para entrenar los modelos, con 40 000 ejemplos, y otro con 10 000 ejemplos para probar los distintos modelos. La empresa ha entrenado distintos modelos con distintas configuraciones que luego ha probado sobre el conjunto de 10 000 ejemplos, seleccionando el modelo que mejor se ha comportado sobre ese conjunto, que es el que obtiene el 92% de exactitud

En este contexto, ¿cuál de las siguientes afirmaciones piensa que es correcta?

- ○ Son, sin ninguna duda, demasiados ejemplos para el conjunto de prueba
- ○ Son, sin ninguna duda, muy pocos ejemplos para el conjunto de prueba
- ◉ Hubiera sido más correcto reservar un conjunto de datos más y usarlo sólo para valorar el rendimiento final del modelo ✔
- ○ Es totalmente seguro que al poner en funcionamiento el modelo su rendimiento será considerablemente menor al 92% esperado

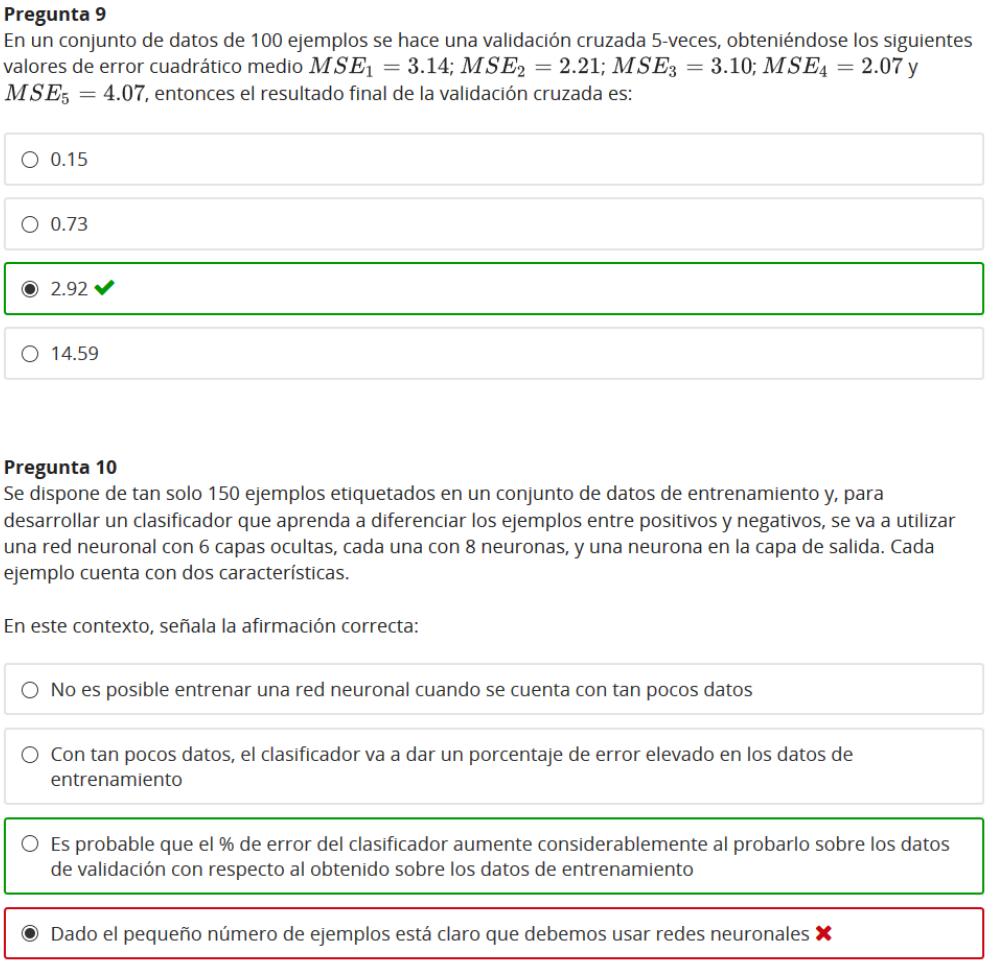
Pregunta 8
Al entrenar un modelo, el error de clasificación obtenido (porcentaje de ejemplos mal clasificados) es el siguiente:

Error en el conjunto de entrenamiento: 3%

Error en el conjunto de validación: 7%

¿Qué sugieren estos datos, sabiendo que un grupo de expertos obtiene en esa tarea un error del 2%?

- ○ Podría ser buena idea usar regularización
- ◉ Esos valores indican que tenemos un sesgo alto ✖
- ○ No es posible mejorar ese resultado, está muy cerca de lo que logran los expertos
- ○ Podría ser buena idea pasar algunos datos del conjunto de datos de entrenamiento al de validación



Pregunta 9
En un conjunto de datos de 100 ejemplos se hace una validación cruzada 5-veces, obteniéndose los siguientes valores de error cuadrático medio $MSE_1 = 3.14$; $MSE_2 = 2.21$; $MSE_3 = 3.10$; $MSE_4 = 2.07$ y $MSE_5 = 4.07$, entonces el resultado final de la validación cruzada es:

- ○ 0.15
- ○ 0.73
- ◉ 2.92 ✔
- ○ 14.59

Pregunta 10
Se dispone de tan solo 150 ejemplos etiquetados en un conjunto de datos de entrenamiento y, para desarrollar un clasificador que aprenda a diferenciar los ejemplos entre positivos y negativos, se va a utilizar una red neuronal con 6 capas ocultas, cada una con 8 neuronas, y una neurona en la capa de salida. Cada ejemplo cuenta con dos características.

En este contexto, señala la afirmación correcta:

- ○ No es posible entrenar una red neuronal cuando se cuenta con tan pocos datos
- ○ Con tan pocos datos, el clasificador va a dar un porcentaje de error elevado en los datos de entrenamiento
- ○ Es probable que el % de error del clasificador aumente considerablemente al probarlo sobre los datos de validación con respecto al obtenido sobre los datos de entrenamiento
- ◉ Dado el pequeño número de ejemplos está claro que debemos usar redes neuronales ✖

A.2. Referencias utilizadas para desarrollar el contenido del curso

ABU-MOSTAFA, Yaser S.; MAGDON-ISMAIL, Malik; LIN, Hsuan-Tien. *Learning from data*. New York, NY, USA:: AMLBook, 2012.

ACEMOGLU, Daron; RESTREPO, Pascual. *Artificial intelligence, automation and work*. National Bureau of Economic Research, 2018.

ANGWIN, Julia, et al. *Machine bias*. ProPublica, mayo, 2016, vol. 23.

ARTICLE 29 DATA PROTECTION WORKING PARTY, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679* (WP 251), 2018

AVATI, Anand, et al. *Improving palliative care with deep learning*. BMC medical informatics and decision making, 2018, vol. 18, no 4, p. 122.

DOMINGOS, Pedro. *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Penguin Books, 2017.

ESTEVA, Andre, et al. *Dermatologist-level classification of skin cancer with deep neural networks*. Nature, 2017, vol. 542, no 7639, p. 115.

FELLER, Avi, et al. *A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear*. The Washington Post, 2016.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. *Deep learning*. MIT press, 2016.

HARDT, Moritz, et al. *Equality of opportunity in supervised learning. Advances in neural information processing systems*. 2016. p. 3315-3323.

INFORMATION COMMISSIONER'S OFFICE, UK. *Guide to the General Data Protection Regulation* (GDPR). URL: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/ (ultimo acceso abril de 2019)

MITCHELL, Tom M.; *Machine Learning*. Mcgraw-hill, 1997.

NG, Andrew. *Machine Learning* (MOOC). Coursera. 2012.
URL: https://www.coursera.org/learn/machine-learning (último acceso abril de 2019).

NG, Andrew. *Structuring Machine Learning Projects* (MOOC). Coursera. 2017. URL: https://www.coursera.org/learn/machine-learning-projects (último acceso abril, 2019).

NG, Andrew. *Machine Learning Yearning* (draft version). 2018.
URL: http://www.mlyearning.org (último acceso abril de 2019).

NIELSEN, Michael A. *Neural Networks and Deep Learning*. Determination Press, 2015.
URL: http://neuralnetworksanddeeplearning.com/ (último acceso abril de 2019).

PARLAMENTO, EU. *Reglamento (UE) 2016/679* (Reglamento general de protección de datos), 2016

PEHRSSON, Emily. *The Meaning of the GDPR Article 22*. Stanford-Vienna European Union Law Working Papers, No. 31. 2018.

SENADO, U. S. *A review of the data broker industry: Collection, use, and sale of consumer data for marketing purposes*. US Senate Committee on Commerce Science and Transportation. 2013.

SILVER, David, et al. *Mastering the game of Go with deep neural networks and tree search*. Nature, 2016, vol. 529, no 7587, p. 484.

WHITE HOUSE, U.S.A.. *Artificial intelligence, automation, and the economy*. Executive office of the President. 2016

Referencias

Attali, Yigal y van der Kleij, Fabienne. Effects of feedback elaboration and feedback timing during computer-based practice in mathematics problem solving. *Computers & Education*, 110:154–169, 2017.

Baddeley, Alan. Working memory: looking back and looking forward. *Nature reviews neuroscience*, 4(10):829, 2003.

Bareinboim, Elias; Tian, Jin; y Pearl, Judea. Recovering from selection bias in causal and statistical inference. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

Barrouillet, Pierre y Camos, Valérie. The time-based resource-sharing model of working memory. *The cognitive neuroscience of working memory*, 455:59–80, 2007.

Basu, Satabdi; Biswas, Gautam; y Kinnebrew, John S. Learner modeling for adaptive scaffolding in a computational thinking-based science learning environment. *User Modeling and User-Adapted Interaction*, 27(1):5–53, 2017.

Black, Paul y Wiliam, Dylan. Assessment and classroom learning. *Assessment in Education: principles, policy & practice*, 5(1):7–74, 1998.

Boucheix, Jean Michel y Guignard, Hélène. What animated illustrations conditions can improve technical document comprehension in young students? format, signaling and control of the presentation. *European Journal of Psychology of Education*, 20(4):369–388, 2005.

Brinton, Christopher G; Buccapatnam, Swapna; Chiang, Mung; y Poor, H Vincent. Mining mooc clickstreams: Video-watching behavior vs. in-video quiz performance. *IEEE Transactions on Signal Processing*, 64(14):3677–3692, 2016.

Brusilovsky, Peter y Millán, Eva. User models for adaptive hypermedia and adaptive educational systems. In P. Brusilovsky, W. Nejdl A. Kobsa, editor, *The adaptive web*, pages 3–53. Springer, 2007.

Challen, Geoffrey y Seltzer, Margo. Enabling mooc collaborations through modularity. In *Proceedings of the 2014 Learning with MOOCs Practitioner's Workshop*, 2014.

Chen, Chih Ming y Wu, Chung Hsin. Effects of different video lecture types on sustained attention, emotion, cognitive load, and learning performance. *Computers & Education*, 80:108–121, 2015.

Chuang, Isaac y Ho, Andrew. Harvardx and mitx: Four years of open online courses–fall 2012-summer 2016. *Available at SSRN 2889436*, 2016.

Clariana, Roy B; Wagner, Daren; y Murphy, Lucia C Roher. Applying a connectionist description of feedback timing. *Educational Technology Research and Development*, 48(3):5–22, 2000.

Cross, Andrew; Bayyapunedi, Mydhili; Cutrell, Edward; Agarwal, Anant; y Thies, William. Typerighting: combining the benefits of handwriting and typeface in online educational videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 793–796. ACM, 2013.

Cummins, Stephen; Beresford, Alastair R; y Rice, Andrew. Investigating engagement with in-video quiz questions in a programming course. *IEEE Transactions on Learning Technologies*, 9(1):57–66, 2016.

de la Fuente Sánchez, Damián; Solís, Montserrat Hernández; y Martos, Inmaculada Pra. El mini video como recurso didáctico en el aprendizaje de materias cuantitativas. *RIED. Revista Iberoamericana de educación a Distancia*, 16(2):177–192, 2013.

Deci, Edward L; Koestner, Richard; y Ryan, Richard M. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin*, 125(6):627, 1999.

Dempster, Frank N. Synthesis of research on reviews and tests. *Educational leadership*, 48(7):71–76, 1991.

Faber, Janke M; Luyten, Hans; y Visscher, Adrie J. The effects of a digital formative assessment tool on mathematics achievement and student motivation: Results of a randomized experiment. *Computers & education*, 106:83–96, 2017.

Griffin, Darren K; Mitchell, David; y Thompson, Simon J. Podcasting by synchronising powerpoint and voice: What are the pedagogical benefits? *Computers & Education*, 53(2):532–539, 2009.

Guo, Philip J; Kim, Juho; y Rubin, Rob. How video production affects student engagement: An empirical study of mooc videos. In *Proceedings of the first ACM conference on Learning@ scale conference*, pages 41–50. ACM, 2014.

Hasler, Béatrice Susanne; Kersten, Bernd; y Sweller, John. Learner control, cognitive load and instructional animation. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 21(6):713–729, 2007.

Hattie, John y Timperley, Helen. The power of feedback. *Review of educational research*, 77(1):81–112, 2007.

Hernán, MA y Robins, JM. *Causal Inference*. Boca Raton: Chapman & Hall/CRC, forthcoming, 2019.

Hidi, Suzanne. Interest and its contribution as a mental resource for learning. *Review of Educational research*, 60(4):549–571, 1990.

Homer, Bruce D; Plass, Jan L; y Blake, Linda. The effects of video on cognitive load and social presence in multimedia-learning. *Computers in Human Behavior*, 24(3): 786–797, 2008.

Hong, Jianzhong; Pi, Zhongling; y Yang, Jiumin. Learning declarative and procedural knowledge via video lectures: cognitive load and learning effectiveness. *Innovations in Education and Teaching International*, 55(1):74–81, 2018.

Ibrahim, Mohamed. Implications of designing instructional video using cognitive theory of multimedia learning. *Critical questions in education*, 3(2):83–104, 2012.

Johnson Glenberg, Mina C. Embedded formative e-assessment: who benefits, who falters. *Educational Media International*, 47(2):153–171, 2010.

Jordan, Katy. Massive open online course completion rates revisited: Assessment, length and attrition. *The International Review of Research in Open and Distributed Learning*, 16(3), 2015.

Kay, Robin H. Exploring the use of video podcasts in education: A comprehensive review of the literature. *Computers in Human Behavior*, 28(3):820–831, 2012.

Kizilcec, René F; Piech, Chris; y Schneider, Emily. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the third international conference on learning analytics and knowledge*, pages 170–179. ACM, 2013.

Kizilcec, René F; Papadopoulos, Kathryn; y Sritanyaratana, Lalida. Showing face in video instruction: effects on information retention, visual attention, and affect. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2095–2102. ACM, 2014.

Kluger, Avraham N y DeNisi, Angelo. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin*, 119(2):254, 1996.

Korving, H; Hernández, M; y De Groot, E. Look at me and pay attention! a study on the relation between visibility and attention in weblectures. *Computers & Education*, 94:151–161, 2016.

Kovacs, Geza. Effects of in-video quizzes on mooc lecture viewing. In *Proceedings of the third (2016) ACM conference on Learning@ Scale*, pages 31–40. ACM, 2016.

Kozma, Robert B. Implications of instructional psychology for the design of educational television. *ECTJ*, 34(1):11–19, 1986.

Lazowski, Rory A y Hulleman, Chris S. Motivation interventions in education: A meta-analytic review. *Review of Educational research*, 86(2):602–640, 2015.

Letón, Emilio y Molanes López, Elisa M. Two new concepts in video podcasts. In *Proceedings of the 6th International Conference on Computer Supported Education-Volume 2*, pages 292–297. SCITEPRESS-Science and Technology Publications, Lda, 2014.

Letón, Emilio; Gómez del Río, Isabel; Quintana Frías, Ignacio; y Molanes López, Elisa M. Clasificación de las distintas modalidades de grabación y su relación con los mini-vídeos docentes modulares. In *XVI Congreso Internacional de Tecnologías para la Educación y el Conocimiento*, 2012.

Letón, Emilio; Molanes López, Elisa M; Luque, Manuel; y Conejo, Ricardo. Video podcast and illustrated text feedback in a web-based formative assessment environment. *Computer Applications in Engineering Education*, 2017.

Luzón, José María y Letón, Emilio. Use of animated text to improve the learning of basic mathematics. *Computers & Education*, 88:119–128, 2015.

Mautone, Patricia D y Mayer, Richard E. Signaling as a cognitive guide in multimedia learning. *Journal of educational Psychology*, 93(2):377, 2001.

Mayer, Richard E. Cognitive theory of multimedia learning. *The Cambridge handbook of multimedia learning*, 41:31–48, 2005.

Mayer, Richard E y Chandler, Paul. When learning is just a click away: Does simple user interaction foster deeper understanding of multimedia messages? *Journal of educational psychology*, 93(2):390, 2001.

Mayer, Richard E y Moreno, Roxana. Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist*, 38(1):43–52, 2003.

Merkt, Martin; Weigand, Sonja; Heier, Anke; y Schwan, Stephan. Learning with videos vs. learning with print: The role of interactive features. *Learning and Instruction*, 21(6):687–704, 2011.

Miller, Sandra. *Formative Computer-based Assessments: The potentials and pitfalls of two formative computer-based assessments used in professional learning programs*. PhD thesis, Queen's University, 2009.

Moreno, Roxana. Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional science*, 32(1-2):99–113, 2004.

Moreno, Roxana. Optimising learning from animations by minimising cognitive load: Cognitive and affective consequences of signalling and segmentation methods. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 21(6):765–781, 2007.

Muñoz Merino, Pedro J; Ruipérez Valiente, José A; Delgado Kloos, Carlos; Auger, Maria A; Briz, Susana; de Castro, Vanessa; y Santalla, Silvia N. Flipping the class-

room to improve learning with moocs technology. *Computer Applications in Engineering Education*, 25(1):15–25, 2017.

Narciss, Susanne. Feedback strategies for interactive learning tasks. *Handbook of research on educational communications and technology*, 3:125–144, 2008.

Narciss, Susanne y Huth, Katja. How to design informative tutoring feedback for multimedia learning. *Instructional design for multimedia learning*, 181195, 2004.

Narciss, Susanne y Huth, Katja. Fostering achievement and motivation with bug-related tutoring feedback in a computer-based training for written subtraction. *Learning and Instruction*, 16(4):310–322, 2006.

Paas, Fred; Van Gog, Tamara; y Sweller, John. Cognitive load theory: New conceptualizations, specifications, and integrated research perspectives. *Educational psychology review*, 22(2):115–121, 2010.

Petrović, Juraj; Pale, Predrag; y Jeren, Branko. Online formative assessments in a digital signal processing course: Effects of feedback type and content difficulty on students learning achievements. *Education and Information Technologies*, 22(6): 3047–3061, 2017.

Rodriguez Ascaso, Alejandro; Letón, Emilio; Muñoz Carenas, Jaime; y Finat, Cecile. Accessible mathematics videos for non-disabled students in primary education. *PloS one*, 13(11):e0208117, 2018.

Roediger III, Henry L y Butler, Andrew C. The critical role of retrieval practice in long-term retention. *Trends in cognitive sciences*, 15(1):20–27, 2011.

Roediger III, Henry L y Karpicke, Jeffrey D. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science*, 17(3):249–255, 2006.

Ryan, Richard M. Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of personality and social psychology*, 43(3): 450–461, 1982.

Ryan, Richard M y Deci, Edward L. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, 55(1):68, 2000.

Schittek Janda, Martin; Tani Botticelli, Antonella; Mattheos, Nikos; Nebel, Daniel; Wagner, Anders; Nattestad, Anders; y Attström, Rolf. Computer-mediated instructional video: a randomised controlled trial comparing a sequential and a segmented instructional video in surgical hand wash. *European Journal of Dental Education*, 9(2):53–58, 2005.

Schrader, Peter G y Rapp, Eric Eugene. Does multimedia theory apply to all students? the impact of multimedia presentations on science learning. *Journal of Learning and Teaching in Digital Age (JOLTIDA)*, 1(1):32–46, 2016.

Schwan, Stephan y Riempp, Roland. The cognitive benefits of interactive videos: learning to tie nautical knots. *Learning and instruction*, 14(3):293–305, 2004.

Shute, Valerie J. Focus on formative feedback. *Review of educational research*, 78(1): 153–189, 2008.

Spanjers, Ingrid AE; Van Gog, Tamara; y van Merriënboer, Jeroen JG. A theoretical analysis of how segmentation of dynamic visualizations optimizes students' learning. *Educational Psychology Review*, 22(4):411–423, 2010.

Spanjers, Ingrid AE; Wouters, Pieter; Van Gog, Tamara; y Van Merrienboer, Jeroen JG. An expertise reversal effect of segmentation in learning from animated worked-out examples. *Computers in Human Behavior*, 27(1):46–52, 2011.

Sweller, John. Evolution of human cognitive architecture. *Psychology of learning and motivation*, 43:216–266, 2003.

Sweller, John. Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational psychology review*, 22(2):123–138, 2010.

Sweller, John; Van Merrienboer, Jeroen JG; y Paas, Fred GWC. Cognitive architecture and instructional design. *Educational psychology review*, 10(3):251–296, 1998.

Tacoma, SG; Drijvers, PHM; y Boon, PBJ. Using student models to generate feedback in a university course on statistical sampling. In *Proceedings of the Tenth Congress of the European Society for Research in Mathematics Education (CERME10, February 1-5, 2017)*, pages 844–851. DCU Institute of Education and ERME, 2017.

van der Kleij, Fabienne M; Eggen, Theo JHM; Timmers, Caroline F; y Veldkamp, Bernard P. Effects of feedback in a computer-based assessment for learning. *Computers & Education*, 58(1):263–272, 2012.

van der Kleij, Fabienne M; Feskens, Remco CW; y Eggen, Theo JHM. Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of educational research*, 85(4):475–511, 2015.

van Gog, Tamara; Verveer, Ilse; y Verveer, Lise. Learning from video modeling examples: Effects of seeing the human model's face. *Computers & Education*, 72: 323–327, 2014.

van Wermeskerken, Margot y van Gog, Tamara. Seeing the instructor's face and gaze in demonstration video examples affects attention allocation but not learning. *Computers & Education*, 113:98–107, 2017.

Vural, Omer Faruk. The impact of a question-embedded video-based learning tool on e-learning. *Educational Sciences: Theory and Practice*, 13(2):1315–1323, 2013.

Wachtler, Josef; Hubmann, Michael; Zöhrer, Helmut; y Ebner, Martin. An analysis of the use and effect of questions in interactive learning-videos. *Smart Learning Environments*, 3(1):13, 2016.

Wang, Jiahui y Antonenko, Pavlo D. Instructor presence in instructional video: Effects on visual attention, recall, and perceived learning. *Computers in human behavior*, 71:79–89, 2017.

Wetzel, C Douglas; Radtke, Paul H; y Stern, Hervey W. *Instructional effectiveness of video media.* Lawrence Erlbaum Associates, Inc, 1994.

Wieling, MB y Hofman, WHA. The impact of online video lecture recordings and automated feedback on student performance. *Computers & Education*, 54(4):992–998, 2010.

Winne, PH y Butler, DL. Student cognition in learning from teaching. Husen, T., Postlewaite, T. (Eds.) International Encyclopaedia of Education (2nd ed., pp 5738-5745), 1994.

Xie, Heping; Wang, Fuxing; Hao, Yanbin; Chen, Jiaxue; An, Jing; Wang, Yuxin; y Liu, Huashan. The more total cognitive load is reduced by cues, the better retention and transfer of multimedia learning: A meta-analysis and two meta-regression analyses. *PloS one*, 12(8):e0183884, 2017.

Zhang, Dongsong; Zhou, Lina; Briggs, Robert O; y Nunamaker Jr, Jay F. Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness. *Information & management*, 43(1):15–27, 2006.

Índice de figuras

Índice de tablas