Competiciones de análisis de datos como herramienta docente en el ámbito de la estadística

Miguel A. Martínez Beneito[1], Carmen Armero Cervera[2], Paloma Botella Rocamora[3], David Conesa Guillén[4], Anabel Forte Deltell[5], Carmen Íñiguez Hernández[6], Antonio López Quílez[7], Francisco José Santonja Gómez[8], Óscar Zurriaga Lloréns[9].

[1] ***Departament d'Estadística i Investigaciò Operativa, Universitat de València, Dr Moliner, 50, 46100, Burjassot, València, Spain, e-mail: miguel.a.martinez@uv.es.***
[2] ***Departament d'Estadística i Investigaciò Operativa, Universitat de València, Dr Moliner, 50, 46100, Burjassot, València, Spain, e-mail: carmen.armero@uv.es.***
[3] ***Dirección General de Salud Pública y Adicciones, Conselleria de Sanitat, Av. Cataluña, 21, 46020, València, Spain, e-mail: botella_pal@uv.es.***
[4] ***Departament d'Estadística i Investigaciò Operativa, Universitat de València, Dr Moliner, 50, 46100, Burjassot, València, Spain, e-mail: conesa@uv.es.***
[5] ***Departament d'Estadística i Investigaciò Operativa, Universitat de València, Dr Moliner, 50, 46100, Burjassot, València, Spain, e-mail: anabel.forte@uv.es.***
[6] ***Departament d'Estadística i Investigaciò Operativa, Universitat de València, Dr Moliner, 50, 46100, Burjassot, València, Spain, e-mail: carmen.iniguez@uv.es.***
[7] ***Departament d'Estadística i Investigaciò Operativa, Universitat de València, Dr Moliner, 50, 46100, Burjassot, València, Spain, e-mail: antonio.lopez@uv.es.***
[8] ***Departament d'Estadística i Investigaciò Operativa, Universitat de València, Dr Moliner, 50, 46100, Burjassot, València, Spain, e-mail: francisco.santonja@uv.es.***
[9] ***Departamento de Medicina Preventiva y Salud Pública, Ciencias de la Alimentación, Toxicología y Medicina Legal, Universitat de València, Avda. Vicent Andrés Estellés, s/n, 46100, Burjassot, València, Spain, e-mail: oscar.zurriaga@uv.es.***

Data analysis competitions as a teaching tool for statisticians

RESUMEN

Las competiciones de análisis de datos se han venido popularizando durante los últimos años como herramientas de aprendizaje y mejora de competencias estadístico-informáticas de sus participantes, entre otros beneficios. Dichas competiciones retan a sus participantes a conseguir modelos estadísticos con propiedades predictivas tan acertadas como sea posible. El presente trabajo plantea el uso de dichas competiciones en el ámbito académico, concretamente dentro de la asignatura de Modelos Lineales del Máster en Bioestadística de la Universitat de València, como herramienta docente. Estas

competiciones desarrollan competencias que habitualmente no resulta tan sencillo de trabajar con herramientas docentes más tradicionales. Este trabajo expone la experiencia docente de dicha aplicación durante el curso académico 2020-21.

Palabras clave: Desarrollo de competencias laborales, Estadística, Modelos predictivos.

ABSTRACT

For the last few years, data analysis competitions have become popular learning tools that allow improving the statistics and information technology skills of their participants. Such competitions challenge their participants to build the statistical model with best predictive features. This work proposes the use of such competitions as teaching tools in an academic context, specifically within the Linear Models course of the Master of Biostatistics of the University of Valencia. These competitions develop skills that are not so easy to develop with traditional teaching tools. This work describes such teaching experience developed during the academic course 2020-21.

Keywords: Labour skills development, Predictive models, Statistics.

INTRODUCCIÓN

Dentro de la comunidad de usuarios del análisis de datos, de un tiempo a esta parte, se han hecho muy populares las competiciones de resolución de problemas estadísticos, o de análisis de datos si se quiere, siendo Kaggle seguramente el más popular de estos servicios. En estas competiciones se propone un problema con una importante componente estadística y equipos de analistas de datos, muchos del entorno académico, compiten para obtener la mejor solución al problema planteado. La propuesta de soluciones en este ámbito requiere el uso de todos los conocimientos y competencias que los grupos pudieran haber desarrollado, para poder presentar soluciones imaginativas, realmente competitivas, en relación a las propuestas del resto de grupos. Estas soluciones invitan a la combinación de distintas estrategias/posibilidades de análisis que, en el caso de estudiantes, no se ceñirían exclusivamente al contenido de una única asignatura sino que admiten la utilización de toda la formación estadística que estos/as pudieran haber adquirido hasta esa fecha. En este sentido el uso de este tipo de herramientas como recurso docente presenta, sin duda, potencial interés.

El Máster en Bioestadística de la Universitat de València goza ya de una trayectoria de más de 10 años y en él se forman anualmente estudiantes con un claro interés por el análisis de datos. Dicho Máster cuenta con una asignatura dedicada a los modelos (de regresión) lineales, en la que se introduce a los/as estudiantes a la modelización estadística, cuyo contenido es posteriormente ampliado/generalizado a un contexto más amplio en posteriores asignaturas. El contenido de dicha asignatura incluye una introducción a los modelos lineales, de manera más tradicional, aunque la asignatura concluye con un par de temas en el que se aborda el uso de este tipo de modelos cuando se dispone de un gran número de covariables, digamos que centenas de ellas. Evidentemente, cuando se dispone de dicho número de covariables las posibilidades de modelización y abordajes distintos son enormes. Por tanto este contexto resulta muy indicado para la propuesta de competiciones de análisis de datos, dada la diversidad de enfoques que podrían seguirse para resolver el problema en cuestión, casi tantos como alumnos/as de la asignatura.

El presente trabajo propone el uso de las competiciones de análisis de datos como herramienta docente. En particular, proponemos su uso dentro de la asignatura de Modelos Linales del Máster en Bioestadística de la Universitat de València y describimos el resultado de dicha aplicación durante el curso académico 2020-21.

METODOLOGÍA

El temario de la asignatura de Modelos Lineales del Máster en Bioestadística de la Universitat de València, contiene la típica introducción a los modelos de regresión Gaussianos, conteniendo dicho temario: inferencia estadística, regresión lineal simple y múltiple, predictores lineales categóricos, interacción entre variables y selección de modelos. En la parte final de la asignatura se introducen también algunos temas más avanzados directamente relacionados con los modelos de regresión con un gran número (cientos o miles) de covariables. Evidentemente, el aumento de la complejidad de estos modelos respecto a los estudiados en la parte inicial del curso es bastante considerable, pasando de modelos con una o unas pocas covariables, a modelos en los que el aumento de la dimensionalidad, y los problemas concretos que ello introduce, supone un reto en sí mismo. Los/as alumnos/as a lo largo del curso se familiarizan con los modelos de regresión con unas pocas covariables, digamos menos de 10. A continuación, y como extensión natural de los temas ya estudiados, se presenta el caso con muchas covariables, donde se requieren técnicas específicas que puedan ofrecer soluciones razonables en este entorno de trabajo tan complejo.

Es en el contexto de los modelos de regresión con un gran número de variables donde se enmarcaría la tarea propuesta como competición de análisis de datos a los/as alumnos/as del Máster en Bioestadística. De manera resumida, el problema que se les plantea vendría a ser lo siguiente: Los datos a modelizar son los de un archivo de sonido del balido de una oveja, a los que hemos añadido cierto ruido aleatorio. Dicho banco de datos, o más bien vector de valores, consta de 19764 observaciones que, teniendo en cuenta que el archivo de sonido ha sido codificado a 8000 herzios, corresponde a un balido de 2.47(=19764/8000) segundos. Se les proporciona el archivo mencionado a los/as estudiantes del curso con el objetivo de que modelicen la media del proceso, haciendo uso de lo aprendido en modelos lineales, de forma que se pueda recuperar la onda/datos del balido de la manera más fiable posible y filtrando el ruido que el archivo original pudiera presentar. Como covariables se les indica que utilicen una base de 876 funciones de Fourier de distintas frecuencias, que serían las variables potencialmente empleables en sus modelos de regresión lineal. Evidentemente, el número de covariables a manejar en este problema es bastante elevado, lo que hace necesario el uso de técnicas de regresión con un gran número de covariables. La propuesta de análisis que se presenta a los/as alumnos/as únicamente detalla el objetivo del estudio, pero no las técnicas de análisis que han de emplear ni cómo han de acometerlo. Es más, en este sentido no tienen limitación ninguna más allá de sus conocimientos estadísticos, tienen total libertad para utilizar cualquier recurso o conocimiento con el fin de alcanzar su objetivo de la forma más acertada posible. La Figura 1 muestra, para los 200 valores iniciales que componen la onda sonora, tanto los datos con ruido que se les proporciona a los/as estudiantes, línea negra, como los datos originales de donde proviene la versión adulterada de los datos, línea roja. El objetivo del trabajo es recomponer, de la manera más precisa posible, la línea roja del gráfico, de la que no disponen información directa sino sólo indirecta a traves de los valores proporcionados por la línea negra.

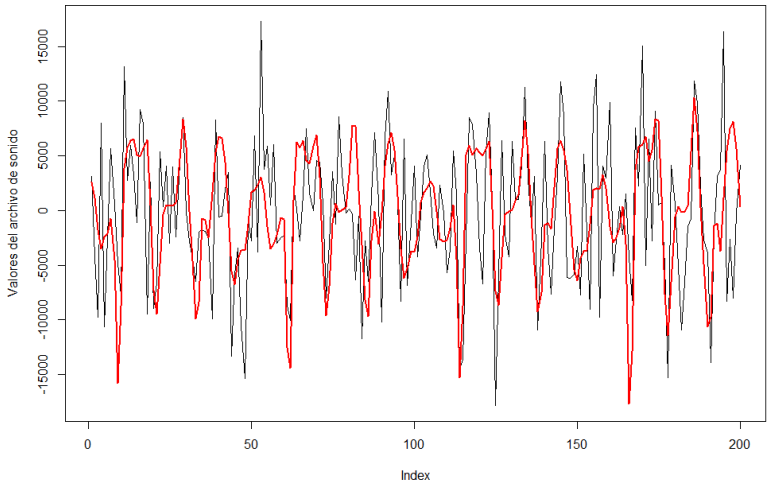

Figura 1: Onda sonora facilitada a los/as alumnos/as (con ruido), línea negra, y onda original (sin ruido), línea roja, que ha de ser reconstruída.

El problema se plantea a los/as estudiantes en los siguientes términos: Se les presenta los 19764 valores correspondientes al archivo de sonido distorsionado, explicándoles que corresponden al alarido de un dinosaurio (el *cabritus hawaianus*) que se acaba de encontrar en una excavación arqueológica en Hawai. Las ondas del alarido han podido ser recuperadas, con algún deterioro (el ruido aleatorio introducido en el archivo de sonido), ya que el animal murió sepultado en una erupción de lava de un volcán y las ondas de su alarido, su último alarido, quedaron impregnadas en la lava. En base a ese hallazgo, y la reconstrucción de la onda de sonido almacenada en la lava solidificada, se supone que seremos capaces, por primera vez, de escuchar el sonido emitido por un dinosaurio, lo que tiene entusiasmada a la Comunidad Científica.
Uno de los problemas a los que se enfrentan habitualmente las competiciones de análisis de datos es el potencial sobreajuste de los modelos estadísticos desarrollados por sus participantes. Concretamente, si los concursantes dispusieran de los datos del alarido que han de reproducir, sin ningún ruido adicional, podrían dedicarse simplemente a ajustar infinidad de modelos, por descabellados que fueran, y determinar aquel de todos ellos que proporcione un mejor ajuste. Dicho modelo podría ofrecer tal ajuste simplemente por azar, no porque sea particularmente bueno en términos explicativos, sino porque en caso de ajustar muchos modelos algunos de los no tan buenos podrían ofrecer un buen ajuste para los datos concretos de los que se dispone, pero no para otros datos adicionales. Este efecto se podría deber al sobreajuste que proporciona dicho modelo de los datos empleados para ajustarlo, lo que hace que su comportamiento para otros datos análogos no sea tan bueno. Este efecto será más probable cuanto más flexibles sean los modelos que se desarrollen, ya que de esta manera la posibilidad de sobreajuste es más clara. Evidentemente, la competición de datos propuesta no querría promover ni premiar ese tipo de procedimientos, producto de la fuerza bruta (número de modelos desarrollados), sino todo lo contrario, ya que dichos modelos no tienen porqué ser buenos para predecir cualquier otro conjunto de datos que pudiéramos analizar.
Para evitar el tipo de situaciones que acabamos de describir se creó una plataforma web

en la que, bajo usuario y contraseña, los/as estudiantes pueden mandar las distintas estimaciones de la onda sonora correspondiente a los modelos estadísticos que hayan ajustado. La plataforma compara los valores predichos de la onda sonora por cada uno de los modelos con los valores de la onda verdadera, sin ruido, para así evaluar la bondad del ajuste llevado a cabo para cada uno de los modelos estadísticos desarrollados. Dicha plataforma permite, para cada usuario/a, comparar el ajuste obtenido con los datos reales en un número limitado de ocasiones, en concreto no más de 25. De esta manera, los/as usuarios/as de la aplicación no pueden dedicarse a probar un número indiscriminado de modelos, sin control, con el objetivo de obtener el mejor ajuste "por fuerza bruta", dada la estricta limitación del número de modelos que se les permite contrastar.

RESULTADOS

Los resultados de la experiencia llevada a cabo en este trabajo han sido muy positivos. La reacción inicial de los/as estudiantes cuando se les plantea el problema es en un primer lugar de asombro, ya que la mayor parte de la asignatura la han dedicado a modelos lineales sencillos de sólo unos pocas covariables. Sin embargo, una vez sobrepuestos a dicha sensación inicial, y ya de manera más fría, la respuesta de los/as alumnos/as es muy positiva. En general, casi todos los/as alumnos/as despliegan gran parte de las herramientas que han podido aprender a lo largo de la asignatura, con lo que buena parte de los objetivos de la actividad quedan sobradamente cumplidos. Sin embargo, al menos 2 de los grupos que entregaron su tarea utilizaron herramientas más allá de las aprendidas a lo largo del curso, por lo que la aportación extra de creatividad que se pretendía desarrollar como objetivo adicional también parece haberse conseguido, al menos en ocasiones concretas. Ambos grupos emplearon ideas de model averaging en las que se promediaban las predicciones de los distintos modelos que se habían desarrollado, al menos aquellos que proporcionaban mejores resultados. La idea formal de model averaging no les había sido introducida a los/as alumnos/as hasta esa fecha en el Máster en Bioestadística, por lo que su uso, de manera intuitiva o quizás algo rudimentaria, fue una grata noticia y un ejemplo de cómo la competición propuesta es un elemento de motivación para los/as alumnos/as introduzcan dosis de creatividad en sus análisis de datos.

En segundo lugar, más allá del alto desempeño observado en la mayoría de los trabajos, también fue una grata sorpresa observar la reacción de los/as alumnos/as al planteamiento, un tanto teatralizado, o si se quiere jocoso, de la actividad con la intención de incrementar la motivación de los/as estudiantes. Así, en la memoria final de la tarea, 2 de los grupos emplearon un tono similar al empleado en su planteamiento, incluso poniendo nombre (evidentemente de manera figurada) a la isla concreta donde se encontró el fosil (*Ni'ihau*), o a la tribu local que colaboró en el hallazgo (*Pu'uwai*). La figura 2 muestra parte de la memoria de uno de los/as alumnos/as en la que se recrea, según su punto de vista, lo que podría haber sido el hallazgo arqueológico que motiva la tarea que se les había encargado.

También, quizás como anécdota, uno de los grupos dio exactamente con el archivo de sonido original que había sido utilizado para generar los datos que se les había proporcionado para llevar a cabo la práctica. Se dieron cuenta de que dicho archivo, contenido en una librería del paquete estadístico R, correspondía al original ya que lo subieron a la aplicación como si fuera el resultado de cualquiera de los modelos que hubieran corrido y vieron que les daba un error de predicción de 0 cuando lo comparaban con la

Ilustración 1. Fósil del ejemplar 42-OG9 excavado en Ni'ihau.
Obsérvese en la zona lumbar las marcas de lo que podría haber sido un Auana.

Figura 2: Fragmento de una de las memorias entregadas por los/as alumnos/as donde se recrea el hallazgo arqueológico del *cabritus hawaianus*.

onda sonora original. Evidentemente, dicha solución no fue aportada como su propuesta final ya que no utilizaba el contenido introducido en el curso de Modelos Lineales, pero nuevamente hicieron una descripción en tono distendido de su hallazgo: "Para acabar, decidimos que era más que probable que los datos originales se encontraran ya en internet, debido a que es posible que los/as estudiantes del máster en bioestadística no fueramos los únicos encargados de desencriptar este tan importante hallazgo. Después de navegar por un buen puñado de documentaciones de paquetes de R llegamos a un paquete, que contenía los datos de un artículo que hablaba de algo similar al cabritus hawaianus. Estos datos, parece que guardaban el secreto tan bien guardado del alarido original...".

CONCLUSIONES

La principal impresión de nuestra experiencia es, en términos generales, muy positiva. Creemos que la tarea que se encomienda a los/as alumnos/as les permite desarrollar una competencia laboral que no resulta tan fácil de fomentar con herramientas docentes más tradicionales, como es el uso y combinación de técnicas y conocimientos, sin limitación, de todo aquello estadísticamente relevante que hayan podido aprender hasta la fecha. Habitualmente, las tareas que deben realizar los/as alumnos/as se ciñen al contenido introducido en una asignatura concreta, a diferencia del ámbito laboral en el que tendrán que echar mano de todo aquello que conozcan sin limitarse al contenido estanco de ciertas asignaturas. Esta tarea les permite trabajar dicha competencia que les puede ser de gran utilidad de cara a su perfil laboral.

La respuesta por parte del alumnado ha sobrepasado las expectivas que se tenían inicialmente. Por un lado, el uso y combinación de técnicas estadísticas, en ciertas ocasiones, ha ido más allá de lo que se preveía inicialmente, utilizando incluso nociones estadísticas de combinación de modelos que no se les había introducido a los/as alumnos/as hasta la fecha. Por tanto, el aporte de creatividad de ciertos alumnos/as les ha hecho "descubrir", aunque sea de manera rudimentaria, conceptos-herramientas esta-

dísticas con las que posiblemente se toparán en un futuro en el desempeño de su vida laboral. Gracias a esta tarea han sido capaces de desarrollar por sí solos estas ideas que posiblemente más tarde conocerán laboralmente.

Por último, la respuesta jocosa de parte de los/as alumnos/as en sus memorias finales también ha sido un factor inesperado y muy grato. El talante de dicha respuesta muestra la motivación de los/as alumnos/as por la tarea que se les propone, llevando dicha motivación incluso más allá de lo que originalmente se les propone. Quizás una moraleja importante de esta conclusión sería que los/as estudiantes devuelven, como un espejo, lo que se les ofrece, por lo que el esfuerzo de motivación extra del alumnado que se puede incorporar a muchas tareas, sin duda tiene un retorno directo en su respuesta. En base a nuestra experiencia ese esfuerzo parece merecer la pena.

A día de hoy, dada la satisfacción que percibimos con la actividad planteada, planeamos extrapolarla a la misma materia, de contenidos similares, del Máster en Ciencia de Datos de la Universitat de Valencia. La realización de esta misma actividad en este entorno distinto permitirá conocer y evaluar las particularidades que presentan los enfoques de los/as alumnos/as de ambas titulaciones y mejorar, en cada uno de ellas, los enfoques que presenten mayor debilidad. De la misma manera, también nos planteamos el desarrollo de una competición similar de ánálisis de datos, pero en este caso en el módulo de Modelización Avanzada del propio Máster de Bioestadística de la Universitat de Valencia, en este caso con una propuesta de actividad adaptada al contenido de dicho módulo, evidentemente. Ambas actividades se pondrán en marcha en el curso académico 2021/22, para lo que se ha pedido financiación en la convocatoria de Proyectos de Innovación Docente de esta universidad, en su convocatoria de 2021. Además, en dicha edición de este mismo proyecto, se contemplará la evaluación de cada una de las actividades planteadas, aspecto que por cuestiones técnicas no ha podido ser abordado de manera oportuna en la presente edición del proyecto.

AGRADECIMIENTOS

Los autores de este trabajo hacen constar, y agradecen, su financiación en la convocatoria de “Ajudes per al desenvolupament de projectes d’innovació educativa per al curs 2020-2021"de la Universitat de València, proyecto código: “UV-SFPIE_PID20-1354272".