Juguemos a científicos de datos

Tarea 2 de Innovación Docente e Iniciación a la Investigación Educativa (Matemáticas)

Antonio García Sevilla

15 de junio de 2018

Índice

1. Introducción

1.1. Motivación y contexto

En la vida moderna, la cantidad de datos a nuestra disposición es ingente. Como miembros de una sociedad cada vez más global e interconectada, las decisiones que nos afectan ya no existen a nivel de aldea o grupo, sino más bien a nivel de sociedad, o país. En televisión, periódicos y análisis político toda esta información viene ya procesada y resumida, en la forma de análisis estadístico.

A pesar de ello, los alumnos actuales tienen muy poca exposición a este tipo de conocimiento. La estadística se ve a nivel teórico, proporcionando las herramientas de probabilidad y modelado necesarias, pero sólo a nivel teórico. Exponer a los alumnos a un trabajo práctico, en el que utilizar estas herramientas para un objetivo, les ayudará a comprender las características reales de la estadística, su utilidad, y cómo leer e interpretar las gráficas y resúmenes que tan a menudo aparecen en los medios.

1.2. Planteamiento

Para motivar a los alumnos a realizar esta práctica, similar al trabajo real de un estadista o analista de datos, se utilizará el método de la "gamificación", en el que se plantea la actividad a desarrollar como un juego con puntuaciones y recompensas a los objetivos parciales.

La idea será desarrollar, a lo largo de una cantidad de tiempo determinada, una "competición" estadística. El profesor irá generando datos (observaciones) de una población que tendrá ya generada. Estos datos tendrán ruido, sesgo, y serán incompletos. Según vayan siendo publicados por el profesor, los alumnos tendrán que elaborar modelos estadísticos que les permitan predecir datos futuros, u otros observables claves en los distintos momentos de evaluación. Según el acierto de sus modelos, los alumnos irán obteniendo puntuación, y a los que tengan más puntos al final de la actividad o en momentos clave, se les darán pequeñas recompensas simbólicas.

Como resultado final de la actividad, además de las predicciones y observaciones realizadas, los alumnos tendrán que exponer el modelo que han desarrollado, explicando sus teorías sobre la naturaleza de la población original mediante tests, gráficas, y otras herramientas de estadística descriptiva.

2. Objetivos

2.1. Resultados que se persiguen

Conseguir que los alumnos participen en una actividad práctica estadística, entendiendo lo que conlleva y adquiriendo intuiciones sutiles sobre la naturaleza de los datos y la tarea del analista.

2.2. Competencias que desarrollarían los estudiantes destinatarios de la propuesta

1. Tener una idea más acertada de la naturaleza de los datos y observaciones, el ruido, sesgo y otras características.
2. Capacidad de análisis de datos en un entorno realista, desde la fase de exploración a la de desarrollo de modelos.
3. Capacidad crítica sobre la predicción y utilización de modelos.
4. Capacidad de síntesis y desarrollo de estadística descriptiva.

3. Contenidos matemáticos relevantes al trabajo

3.1. Probabilidad básica

Como fundamento teórico, se repasarán los conceptos de variable aleatoria y las funciones de distribución y densidad. El alumno se espera que ya tenga un conocimiento elemental de probabilidad y muestreo, en el sentido de experimentos en espacios de probabilidad finitos y estimación de probabilidad frecuentista.

Como herramienta para el trabajo posterior, se estudiarán los efectos de distintas operaciones matemáticas (suma, producto...) sobre la distribución de las variables aleatorias. Se verán los conceptos básicos de esperanza y varianza, así como otras medidas de distribución como pueden ser momentos de orden superior, o estadísticos descriptivos como la mediana y los percentiles.

3.2. Distribuciones habituales

Se hará un repaso de funciones de distribución comunes, incluyendo su formulación matemática, momentos y características descriptivas más importantes. Se espera del alumno no que las memorice, sino que confeccione una pequeña guía de campo que le sea de utilidad durante la ejecución de la prueba.

3.3. Estadística descriptiva

Se verán herramientas descriptivas útiles para el tratamiento de datos: histogramas, diagramas de cajas, nubes de puntos (diagramas de dispersión). Se estudiarán los conceptos de covarianza y correlación, y se presentarán los conceptos avanzados de las leyes de potencias y el análisis de la varianza.

3.4. Estimación y test

Finalmente, se hará uso de los conceptos de estimación e hipótesis, se tratarán los distintos tipos de error, y se explicará el concepto de test de hipótesis. Se verá el concepto de significatividad en profundidad y se estudiarán distintos tipos de test y como se relacionan con la distribución de la muestra y de la población original.

3.5. Predicción y clasificación

Finalmente, se verá la herramienta de regresión como herramienta de predicción, y se estudiará la regresión lineal y polinomial para su uso en el ejercicio. También se presentará el concepto de clasificación, frontera de decisión, error acumulado, y como introducción a los métodos no paramétricos, la clasificación en k vecinos más cercanos.

4. Metodología

4.1. Marco de trabajo

La "gamificación" es un concepto que propugna el convertir tareas de la vida cotidiana (ya sea en el trabajo, la escuela, etc.) en una suerte de juego, con objetivos, puntuaciones y recompensas. Esta técnica ayuda a mejorar la motivación, estimulando la parte competitiva y deseosa del éxito de los seres humanos. El juego es una parte fundamental de la infancia y adolescencia, y esto permite que si se usa con éxito, esta técnica haga a los alumnos más interesados en el tema de estudio, y por tanto mejorar su rendimiento.

En el caso de esta propuesta educativa, el juego va acompañado por un aprendizaje por simulación. Al convertir la tarea del estadístico, una realidad cada vez de mayor importancia, en un "juego" en el que se busca la victoria, los alumnos se ven motivados para realmente entender, y por tanto aprender (no sólo memorizar) los conceptos fundamentales que subyacente al tratamiento y estudio formales de los datos.

Debido a la naturaleza del ejercicio, es imposible o muy complicado realizarlo con métodos tradicionales. El uso de herramientas TIC como *R*[1] u *octave*[2] para el análisis de datos, y *moodle*[3] o algún otro entorno virtual para orquestar la competición, supone una motivación adicional para los alumnos, un aprendizaje más realista en el contexto social moderno, y una propuesta innovadora que puede ser útil en otros campos de la enseñanza.

Por supuesto, esto implica la necesidad de contar con medios TIC suficientes para la realización del ejercicio. Muchas de las herramientas necesarias están disponibles en abierto, pero eso no quita que sea necesaria su instalación y mantenimiento, por no hablar de la infraestructura informática necesaria por parte tanto del profesor como de los alumnos. Sin embargo, parece razonable que el centro educativo disponga de estos recursos y pueda ponerlos a disposición de este ejercicio.

[1] `https://www.r-project.org/`
[2] `https://www.gnu.org/software/octave/`
[3] `https://moodle.org/`

4.2. Plan de ejecución

1. Preparación previa Esta propuesta requiere de una preparación previa significativa. El profesor debe preparar una serie de poblaciones, que no pueden ser aleatorias sino pensadas para despertar en los alumnos las curiosidades y problemas que les hagan avanzar en la comprensión del problema. Una vez preparadas las poblaciones, la secuencia de extracción de las muestras también debe ser planificada.

Por ejemplo, se puede preparar una distribución bimodal, pero sesgar la probabilidad de muestreo claramente hacia uno de los picos de la distribución. Esto hará a los alumnos plantear una hipótesis inicial, que verán errónea cuando se descubra el resto de la población, y así comprenderán lo importante de la exhaustividad del muestreo y la validez de la reformulación de hipótesis.

También es recomendable que los datos no sean números crudos, sino que se correspondan a medidas (ficticias) de objetos reales que estimulen el interés de los alumnos. Por ejemplo, estadísticos de jugadores de fútbol de un equipo imaginario, o cualquier tipo de elementos que estén relacionados con algún campo de interés de los estudiantes.

Una vez preparado el ejercicio, es aconsejable que el docente pruebe a realizarlo él mismo, con objeto de comprobar que los contenidos a utilizar sean necesarios y que la tarea no sea tan complicada que su dificultad desmotive a los alumnos.

2. Repaso de teoría Los contenidos matemáticos relevantes, si los alumnos no son del todo familiares con ellos, deben ser repasados. Se puede hacer especial hincapié en los conceptos y herramientas que el profesor, habiendo diseñado ya el ejercicio, sabe que van a ser más importantes.

3. Explicación de la tarea y primeros pasos Cuando los conceptos estén claros, se puede proceder a detallar la tarea. Es recomendable que el docente haga mención de la naturaleza de los datos que los alumnos van a ir recibiendo, y del objetivo de cada sesión. Si este motivo va ligado a una evaluación extra, en forma de premio o refuerzo positivo, aumentará el interés de los alumnos sin causarles miedo a la dificultad.

En opinión del autor, probablemente sea mejor no darles a los alumnos ninguna indicación a priori sobre cómo realizar la actividad. Éstos ya estarán sobre la pista debido al repaso previo, pero darles libertad para utilizar cualquier método que crean conveniente estimulará su creatividad. Por supuesto, en cualquier momento que un alumno se halle en apuros podrá solicitar tutoría y el docente podrá apoyar el aprendizaje mediante pistas o repaso de los contenidos necesarios.

Quizá también sea útil realizar las primeras extracciones de muestras y su análisis en equipo, con todo el grupo presente, y una guía más explícita del profesor. Esto ayudará a los alumnos a entender qué se espera de ellos, y también les pondrá en el buen camino de cómo realizar la actividad.

4. Ejercicio independiente Tras esta fase inicial, se esperará que los alumnos trabajen de forma independiente. Para fomentar el aprendizaje a largo plazo, las sesiones deberán estar espaciadas. Por ejemplo, los lunes se liberará un nuevo dato o conjunto de datos por parte del profesor. El viernes de esa semana, o dos semanas después, los alumnos deberán entregar sus predicciones o modelos propuestos. Cada cierto tiempo, quizá incluso cada sesión, el profesor podrá examinar los resultados hasta entonces, hacer comentarios que guíen el desarrollo de la actividad o analizar errores comunes. También se podrán repartir los premios, aunque sea sólo en la forma de un reconocimiento verbal, de los alumnos que hayan tenido mayor rendimiento en distintos aspectos, para reforzar el aspecto "gamificado" de la actividad.

5. Resultados y análisis Al concluir la actividad, se realizará una sesión extra de análisis de ella. Primero los alumnos expondrán brevemente el modelo que han desarrollado y sus conclusiones. Entonces, el profesor enseñará la población completa, las muestras extraídas, y analizará el acierto, los errores, y los modelos de los alumnos. Esta sesión servirá como cierre y evaluación de la actividad, pero el análisis riguroso del proceder de la actividad también será una lección valiosa para los alumnos. El profesor también podrá notar cuestiones de interés sobre el desarrollo o aspectos de mejora para ocasiones posteriores.

5. Conclusiones

La sociedad moderna está inundada de datos. Las empresas, a través de su publicidad, nos aseguran de que 9 de cada 10 doctores recomiendan su pasta de dientes, o de que 8 de cada 10 personas que comen el alimento de turno adelgazan. Estudios científicos financiados por las partes interesadas corroboran todas las hipótesis imaginables, y los medios continuamente emiten datos estadísticos sobre economía o política.

En este contexto, es fundamental que los alumnos comprendan la potencia y los peligros de la estadística, desarrollen un conocimiento profundo de estas herramientas, y sea capaces de realizar, con pensamiento crítico y riguroso, un análisis de los datos que reciben a diario.

Sin embargo, aleccionar a los alumnos sobre estas cuestiones es poco productivo. El estudio de la estadística y la probabilidad, materias arduas por su rigor y complejidad, tampoco es fácil de trasladar a situaciones cotidianas en las que aparentemente los datos son sencillos, números aislados, definitivos y sin contexto.

Es por esto que la mejor manera de que los alumnos comprendan realmente la naturaleza del análisis de datos es que lo practiquen ellos mismos. En esta propuesta de innovación, al ser los alumnos parte interesada del proceso, podrán aprender a desarrollar un pensamiento analítico que les permita más efectivamente en el futuro moverse en este mundo que es un mar de datos. Esta actividad está planteada como un juego, para mejorar la motivación y el aprendizaje por simulación de los alumnos, pero trata de una cuestión muy seria e imparte un conocimiento procedimental que, en opinión del autor, es importantísimo hoy en día y no suficientemente enseñado con los métodos tradicionales.