Actas de las Jenui, vol. 1. 2016.
Páginas: 201-208

Predicción de la aceptación o rechazo de las calificaciones finales propuestas por el alumnado usando técnicas de Minería de Datos

Javier Fuentes-Alventosa, Cristóbal Romero, Carlos García-Martínez, Sebastián Ventura
Departamento de Informática y Análisis Numérico
Universidad de Córdoba
Córdoba
{javier.fuentes,cromero,cgarcia,sventura}@uco.es

Resumen

Una posible alternativa o complemento a las técnicas clásicas de evaluación es la utilización de técnicas de autoevaluación (*self-grading* o *self-assessment*) que es un proceso en el que es el propio estudiante el que juzga los logros conseguidos respecto a una tarea o actividad determinada. Sin embargo, antes de poder incluirla en un programa educativo, es necesario evaluar la fiabilidad del proceso y compararlo con los métodos tradicionales que actualmente utiliza el profesorado. Siguiendo esta idea, el presente trabajo propone una metodología basada en la minería de datos y la autoevaluación con el fin de validar la auto-calificación de los estudiantes. Nuestro objetivo es predecir si el instructor podría aceptar o no las calificaciones propuestas por los estudiantes en un curso. Nuestra experiencia empírica sobre 53 estudiantes universitarios de informática sugiere que son necesarios 1) una distinción en el tipo de errores que puede cometer el clasificador, pues no es lo mismo no aceptar la nota propuesta de un estudiante que realmente puede sacar dicha nota, que aceptar la de otro que no estaría capacitado para obtenerla; 2) una vigilancia sobre dichos errores; y 3) considerar medidas de rendimiento del estudiante durante el curso, sobre las cuales las predicciones parecen tener una gran dependencia.

Abstract

An alternative or complement to classic assessment methodologies is the use of *self-grading* or *self-assessment*, which is a process where the own student judges and evaluates his/her achievements with regards to particular tasks or activities. However, it is necessary to evaluate the reliability of the process before considering it in a learning program, and to compare it with traditional methods that teachers apply at the moment. According to this idea, this work proposes a methodology based on data-mining and *self-assessment* with the aim of validating the mark that students propose. Our objective is to predict whether the teacher might or might not accept these marks. Our empirical experience on 53 university students suggests that it is needed 1) a distinction between the type of errors of the possible built classifier, given that it is usually preferable not to accept the mark proposed by a student that can get it, than to accept the mark proposed by a student that can not get it; 2) an inspection on particular cases where students might deceive the system; and 3) to consider student performance measures along the course, by which predictions have shown to be strongly determined.

Palabras clave

Autoevaluación, auto-calificación, clasificación sensible al coste, discretización de atributos.

1. Introducción

La evaluación y calificación del alumnado en el contexto del Espacio Europeo de Educación Superior es un proceso complejo y susceptible de innumerables críticas [16, 19]. Como principal exponente, el examen final es uno de los métodos mayormente utilizados y criticados [11]. Uno de los argumentos es que la realización de un examen no refleja una situación real en el que un estudiante aplique los conocimientos y competencias que adquirió durante el curso. Por otro lado, el acto de asignar una calificación u otra a un estudiante no carece de subjetividad [15, 22]; cada instructor utiliza diferentes criterios, y pone un énfasis diferente en cada uno de ellos, a la hora de diseñar y evaluar un examen o trabajo de evaluación. Esto ha hecho que varios investigadores detecten que un porcentaje significativo de estudiantes empieza a interesarse más por

la superación de las materias, que son requisito para la obtención de un título, que por la adquisición de conocimientos y competencias [16, 19].

La autoevaluación (*self-grading* o *self-assessment*) es un proceso en el que es el propio estudiante el que juzga los logros conseguidos respecto a una tarea o actividad determinada [4, 7, 10]. En la actualidad existen numerosos trabajos que estudian su validez, tanto como instrumento formativo como herramienta para la asignación de calificaciones a estudiantes [5, 25, 24]. Entre otros, Brew [6] argumenta que la autoevaluación y la evaluación por pares son esenciales para la culminación del aprendizaje y pieza necesaria en el aprendizaje permanente (*life-long learning*); Roman y Soriano [20] lo relacionan además con el aprendizaje autónomo; y Kulkarni y otros [17] analizan su uso en cursos online (*massive open online courses*) y describen diferentes experimentos con el objetivo de mejorar su precisión.

Además del interés en fomentar una toma de consciencia y una educación en la responsabilidad que promueva el aprendizaje autónomo en el alumnado [6, 20], en este trabajo nos preguntamos si el uso de técnicas de autoevaluación podría descargar al profesorado de la tarea de asignación de calificaciones en clases numerosas [23], pues podría reducir la carga en técnicas de evaluación tradicionales, y orientar el esfuerzo en la reeducación del alumnado en la auto-crítica y la producción de autoevaluaciones objetivas y responsables. En última instancia, el profesorado podría incluso aceptar la notas propuestas por el propio alumnado, mediante un juicio que podría considerar otras actividades de evaluación complementarias realizadas durante el curso, pudiendo incluso evitar la realización del examen final.

La clasificación es una de las tareas más antiguas y más conocidas de la minería de datos en la educación (EDM; *Educational Data Mining*) [21]. El objetivo de la predicción es deducir un atributo categórico o algún aspecto particular de los datos (variable predicha) a partir de alguna combinación de otros aspectos de los mismos (variables de predicción). Una amplia gama de algoritmos, técnicas y enfoques de clasificación se han aplicado en EDM.

Para analizar la validación del proceso de aceptacion de la nota propuesta por el alumnado, en este experimento se parte de un conjunto de datos y se analiza si se puede predecir la decisión del profesorado de aceptar o no dicha nota mediante técnicas de clasificación. En particular, se propuso a un total de 53 estudiantes de la asignatura Sistemas Inteligentes, obligatoria de 2º del Grado de Ingeniería Informática de la Universidad de Córdoba, que elaborasen un auto-informe crítico acerca de los conocimientos y competencias adquiridas en la asignatura y que se propusiesen una calificación para ésta. El profesorado posteriormente aceptaba o rechazaba la calificación propuesta por el alumnado en base al juicio que hacía sobre dicho informe y sus conocimientos acerca de la participación del alumnado en la asignatura, entre otros, la puntuación en tres actividades realizadas durante el curso que comprendían cuestiones de teoría y problemas a resolver. En el experimento llevado a cabo, el alumnado debía completar el auto-informe unos días antes de la realización del examen final de la asignatura. Este proceso les podría dar la posibilidad de no tener que realizar el examen, si el docente aceptase la nota propuesta.

Es importante tener en cuenta que no estamos interesados en la tarea de calibración, que se refiere a cómo de preciso pueden los individuos predecir qué tan bien lo hacen en una tarea antes de comenzar y lo bien que juzgan su rendimiento después de la finalización [3]. Al contrario, sólo estamos interesados en predecir la decisión del instructor de aceptar o rechazar las notas propuestas por los estudiantes.

En esta primera prueba experimental, hemos utilizado 37 algoritmos de clasificación proporcionados por el software de minería de datos Weka [12, 14] para predecir la decisión del docente. Hemos evaluado además la utilización de mayor cantidad de información (además de la simple nota propuesta) y diferentes técnicas de pre-procesado y clasificación para intentar mejorar la predicción, como son la utilización de notas parciales, el cuestionario de autoevaluación, la discretización de los atributos y la clasificación basada en costos [1, 2, 9, 13]. Los resultados sugieren que son necesarios: 1) una distinción en el tipo de errores que puede cometer el clasificador, pues no es lo mismo no aceptar la nota propuesta de un estudiante que realmente puede sacar dicha nota, que aceptar la de otro que no estaría capacitado para obtenerla; 2) una vigilancia sobre dichos errores; y 3) considerar medidas de rendimiento del estudiante durante el curso, sobre las cuales las predicciones parecen tener una gran dependencia.

El trabajo se organiza de la siguiente manera: La Sección 2 muestra la metodología empleada en este trabajo; la Sección 3 presenta los conjuntos de datos utilizados en la experimentación; la Sección 4 describe los experimentos, los resultados, y algunos modelos de ejemplo obtenidos; y, por último, las conclusiones y los trabajos futuros se describen en la Sección 5.

2. Metodología

La metodología que hemos utilizado en este estudio es la siguiente:

- Durante el curso, los estudiantes son evaluados mediante un conjunto de cuestiones y problemas a resolver.

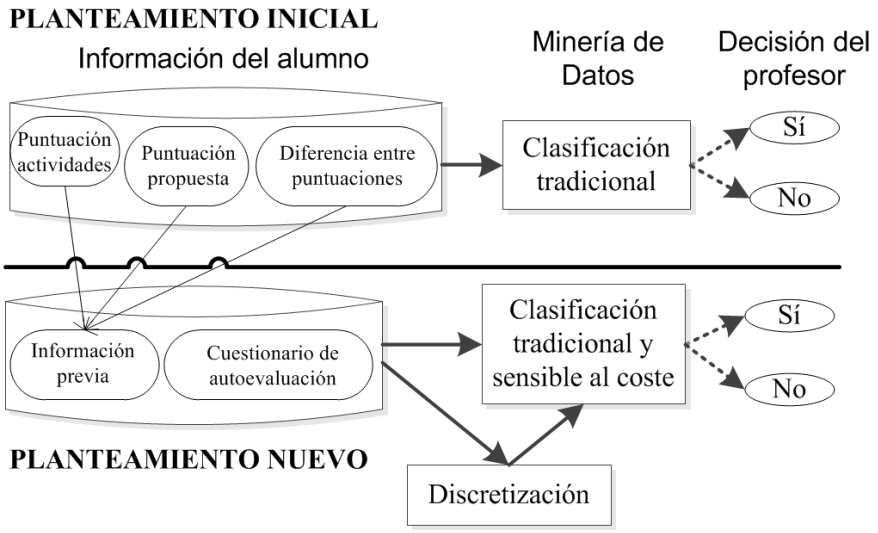

Figura 1: Métodos para predecir la decisión del instructor.

- Antes de la fecha del examen final, a todos los estudiantes se les pide que elaboren un informe acerca de su rendimiento en la asignatura, y que además propongan una calificación, que es la que piensan que merecen en el curso.
- A continuación, el instructor acepta o rechaza la nota propuesta de cada estudiante como la nota final de la asignatura. De esta manera, solamente los estudiantes cuya calificación fue rechazada por el instructor tendrá que hacer el examen final.
- Por último, intentaremos predecir la decisión del instructor de aceptar o rechazar la puntuación propuesta por los estudiantes. Utilizaremos dos procedimientos, referenciados como *inicial* y *nuevo*, según se muestra en la Figura 1.
  El planteamiento inicial sólo utiliza tres variables numéricas: la puntuación obtenida por los estudiantes en las actividades del curso, las calificaciones propuestas por los estudiantes y la diferencia entre estos dos resultados anteriores. A continuación, se aplican algoritmos de clasificación tradicionales [1, 13] para predecir la decisión del instructor sobre la aceptación o no de las calificaciones propuestas por los estudiantes.
  El nuevo enfoque utiliza las tres variables anteriores, así como un cuestionario de autoevaluación para la obtención de otra fuente de información sobre los estudiantes [26]. Este cuestionario fue preparado en la Universidad de Ohio (EE.UU.) y la prueba se realizó entre cien estudiantes: 50 buenos y 50 mediocres [18]. A continuación, aplica algoritmos de clasificación tradicionales, de clasificación sensible al costo y discretización de atributos [1, 2, 9, 13]. La clasificación sensible al coste se utiliza para la obtención de un mejor rendimiento que con la clasificación tradicional. La discretización de atributos se utiliza con el objetivo de obtener modelos más comprensibles con datos categóricos, en lugar de datos numéricos.

3. Conjunto de Datos

Hemos utilizado un conjunto de datos recogidos de estudiantes de universidad del segundo año de Informática en 2012-13. Durante el curso, el instructor dio a los estudiantes la opción de auto-calificarse. De los 86 estudiantes matriculados en el curso, 53 aceptaron la propuesta, aproximadamente el 60 %. De cada uno de estos 53 estudiantes, recopilamos los siguientes cuatro atributos de entrada:

- **Puntuación de las actividades.** Esta es la puntuación media obtenida por el alumnado en tres actividades realizadas durante el curso. Las actividades comprendían cuestiones de teoría y problemas a resolver. La puntuación obtenida en cada una de las actividades, valorada como un número entre 0 y 10 puntos, se daba a conocer al propio alumnado. Debemos destacar que la evaluación de dichas actividades requirió un esfuerzo importante por el profesorado, que pudo ser equivalente a la corrección de los exámenes del alumnado.
- **Puntuación propuesta.** Esta es la puntuación definitiva que los estudiantes consideran que se merecen. Es un número entre 0 y 10.
- **Diferencia entre las puntuaciones.** Es la diferencia entre los dos valores anteriores. Se trata de un valor positivo o negativo (entre -10 y 10) obtenido

automáticamente como la puntuación de las actividades menos la puntuación propuesta.

- **Puntuación del cuestionario de autoevaluación.** Esta es la puntuación obtenida en un cuestionario de autoevaluación. Utilizamos un cuestionario desarrollado en la Universidad de Ohio (EE.UU.) [18, 26]. Éste contiene 50 preguntas de tipo sí/no para determinar si un estudiante es un estudiante bueno o malo. Los estudiantes completaron el cuestionario dos semanas antes de la fecha del examen final. La Universidad de Ohio también nos proporcionó una plantilla con las respuestas de los estudiantes buenos. Con esta plantilla se calculó una puntuación para cada estudiante como el número de respuestas iguales a los de los buenos estudiantes. Entonces, normalizamos este valor en un número entre 0 y 10.

El atributo de salida o la clase para la predicción en nuestro problema es la decisión del instructor. Se trata de un valor binario: SÍ o NO, que indica si el instructor acepta o rechaza las puntuaciones propuestas por los estudiantes. El instructor nos proporcionó este valor para cada uno de los 53 estudiantes: 37 SI (70 %) y 16 no (30 %).

Toda esta información fue anonimizada mediante la eliminación de los nombres de los estudiantes para proteger su privacidad. Por último, se creó un archivo .ARFF (Attribute-Relation File Format) que contiene toda la información anterior para ser analizada con el software de minería de datos Weka [12, 14].

4. Experimentos y resultados

Hemos llevado a cabo varios experimentos con el fin de poner a prueba nuestra propuesta de metodología para la predicción de la decisión del instructor. En estos experimentos, hemos utilizado 37 algoritmos de clasificación proporcionadas por Weka 3.7.10 [12, 14], categorizados en los siguientes 6 grupos:

- **Bayes:** NaiveBayes, NaiveBayesSimple y NaiveBayesUpdateable.
- **Function:** Logistic, MultilayerPerceptron, RBFNetwork, SimpleLogistic, SMO y VotedPerceptron.
- **Lazy:** IB1, IBk, KStar y LWL.
- **Misc:** HyperPipies y VFI.
- **Rules:** ConjunctiveRule, DecisionTable, DTNB, JRip, NNge, OneR, PART, Ridor y ZeroR.
- **Trees:** ADTree, BFTree, DecisionStump, FT, J48, J48graft, LADTree, LMT, NBTree, RandomForest, RandomTree, REPTree y SimpleCart.

Hemos ejecutado todos los algoritmos usando validación-cruzada de 10 particiones y sus parámetros por defecto. Se utilizaron tres medidas de desempeño de clasificación para comparar los resultados de los algoritmos:

	Enfoque Inicial	Enfoque Nuevo
Precisión	$73,89 \pm 5,33$	$77,46 \pm 6,67$
	R-=84	R+=619
Tasa de TP	$84,73 \pm 7,64$	$86,85 \pm 7,64$
	R-=341	R+=362
Tasa de TN	$48,81 \pm 8,79$	$55,74 \pm 4,41$
	R-=219,5	R+=483,5

Cuadro 1: Medidas de clasificación de rendimiento medio para el enfoque inicial y el nuevo usando datos numéricos.

- **Precisión:** Se trata de la relación global de estudiantes clasificados correctamente.
- **Tasa de verdaderos positivos (tasa TP) o sensibilidad:** Es la proporción de positivos reales (estudiantes a los que sí se les aceptó la nota propuesta) que se predijeron como positivos.
- **Tasa de verdaderos negativos (tasa TN) o especificidad:** Es la proporción de negativos reales (estudiantes a los que no se les aceptó la nota propuesta) que se predijeron como negativos.

Se utilizó el test de Wilcoxon [27] para evaluar estadísticamente la diferencia de rendimiento entre varios enfoques. Esta prueba calcula la diferencia entre los resultados del mismo algoritmo en dos escenarios o enfoques diferentes (inicial y nuevo). Aquellos casos en los que los resultados obtenidos fueron mejores mediante el enfoque nuevo se asocian al valor R+, de acuerdo con la magnitud de la diferencia; y aquellos en los que los mejores resultados pertenecen al enfoque inicial, al valor R-. La hipótesis nula, que asume que no hay diferencia alguna entre los resultados de ambos enfoques, se rechaza cuando alguno de los valores, R- o R+, es inferior al valor crítico. En el primer caso, R- es menor que el valor crítico, la prueba encuentra diferencias significativas a favor de la nueva metodología. Por el contrario, si el valor de R+ es inferior al valor crítico, la prueba mostraría que el enfoque inicial proporciona mejores resultados. Si ninguno de los valores es inferior al valor crítico, la hipótesis nula no puede ser rechazada, y la prueba no encuentra diferencias significativas.

En el primer experimento, hemos comparado el enfoque inicial frente al enfoque nuevo utilizando técnicas de clasificación tradicionales y las puntuaciones numéricas originales. Por lo tanto, los resultados de los algoritmos de clasificación, con y sin la puntuación del cuestionario de autoevaluación, son comparados para probar que vienen de diferentes distribuciones.

El Cuadro1 muestra los valores medios y las desviaciones estándar ($Media \pm Desv$) y la suma de los

rangos con signo (R+ o R-) para las tres evaluaciones, sobre todos los algoritmos de clasificación utilizados. Observamos en dicho cuadro que los valores medios de la precisión, tasa TP y tasa TN se han incrementado en el nuevo enfoque. Hemos aplicado un test de Wilcoxon con un factor significativo del 95 %, con los 37 algoritmos (el correspondiente valor crítico es 221). La prueba detecta diferencias significativas a favor de la nueva metodología para la precisión y la tasa de predicción verdadera negativa (debido a que los valores del R- son inferiores al valor crítico), y no encuentra diferencias para la tasa de predicción verdadera positiva. Estos resultados nos muestran que el uso de la puntuación del cuestionario de autoevaluación ha mejorado la precisión y la especificidad de los algoritmos sin afectar a la sensibilidad.

En el segundo experimento analizamos si la discretización de los datos puede ofrecer una mejor predicción de la decisión del profesorado. La discretización es el proceso de conversión o partición de atributos continuos a atributos discretizados o nominales. Hay un gran número de técnicas de discretización supervisadas y no supervisadas [8]. Nosotros hemos utilizado una técnica de discretización manual. Por un lado, las puntuaciones de las actividades, la puntuación propuesta y la del cuestionario de autoevaluación se han discretizado manualmente de acuerdo a la escala de calificación española:

- **Suspenso** para puntuaciones inferiores a 5,
- **Aprobado** para puntuaciones entre 5 y 6,9,
- **Notable** entre 7 y 8,9, y
- **Sobresaliente** entre 9 y 10.

Por otro lado, la diferencia entre las puntuaciones de las actividades y la propuesta se ha discretizado manualmente de la siguiente manera:

- **Positiva** para valores mayores que 0,
- **Ninguna** cuando la diferencia es igual a 0, y
- **Negativa** para valores inferiores a 0.

El Cuadro 2 muestra los valores medios, desviaciones estándar y la suma de los pruebas de clasificación para las tres medidas de la ejecución de la evaluación obtenidos por todos los algoritmos de clasificación sobre los enfoques numéricos y discretizado. El Cuadro 3 muestra además los resultados de cada algoritmo de clasificación, junto con la media por familia y los mejores resultados en negrita.

Se observa en el Cuadro 2 que los valores promedio de las tres medidas de rendimiento se han incrementado al utilizar los datos discretizados. Hemos aplicado el test de Wilcoxon de nuevo, con el mismo factor significativo y valor crítico como antes (95 % y 221). Se encuentran diferencias significativas a favor de la explotación de los datos discretizados en lugar del numérico original para la precisión y la tasa de predicción

	E. Nuevo Num.	E. Nuevo Disc.
Precisión	$77,46 \pm 6,67$	$81,28 \pm 5,33$
	R-=60	R-=643
Tasa de TP	$86,85 \pm 7,64$	$92,03 \pm 4,89$
	R-=77,5	R+=625,5
Tasa de TN	$55,74 \pm 4,41$	$56,41 \pm 7,67$
	R-=339	R+=364

Cuadro 2: Medidas de clasificación de rendimiento medio para el nuevo enfoque usando datos numéricos y discretizados.

verdadera positiva, y no se encuentran diferencias para la tasa de predicción verdadera negativa. Estos resultados nos muestran que el uso de los datos discretizados en nuestro problema ha mejorado la precisión y la sensibilidad de las predicciones en general, independientemente del algoritmo de clasificación utilizado, sin afectar a la especificidad. Además, de los resultados del Cuadro 3 observamos que el clasificador que mejor precisión consigue, además de mejor tasa TN, es el `MultilayerPerceptron`, el cuál implementa una red neuronal. Otros clasificadores consiguen la misma o incluso superior tasa TP, pero no detectan igual de bien al alumnado para el cuál no se acepta la nota de autoevaluación.

En el tercer experimento, hemos utilizado la clasificación sensible al coste [9]. A menudo ocurre que, optimizar la tasa de clasificación, sin tener en cuenta el costo asociado al tipo de error cometido, puede conducir a resultados no suficientemente óptimos. Esto es común cuando estamos interesados en clasificar correctamente una de las clases, por ejemplo, la minoritaria. De hecho, en nuestro problema particular, estamos mucho más interesados en la clasificación correcta del NO (la clase minoritaria) que del SI (la clase de la mayoritaria). Estos costes pueden ser incorporados en el algoritmo (a través del metaclasificador `CostSensitiveClassifier` de Weka) y consideradas en la clasificación. En el caso de las dos clases, los costos se ponen en una matriz de 2×2 en la que los elementos de la diagonal representan los dos tipos de clasificaciones correctas y los elementos fuera de la diagonal representan los dos tipos de errores.

$$\begin{pmatrix} 0 & 1 \\ N & 0 \end{pmatrix}$$

Esta matriz indica que es N veces más importante clasificar correctamente los patrones de la segunda clase, los estudiantes a los que no se le aceptó la nota (NO), que a los de la primera, los estudiantes a los que sí se les aceptó (SÍ).

Las Figuras 2 y 3 muestran los valores promedio, sobre todos los clasificadores utilizados, de precisión, tasa de TP y tasa de TN para diferentes valores de N (1 ,2 ,3 ,4 y 5) cuando se utilizan el nuevo enfoque con

Algoritmo	Precisión	Tasa TP	Tasa TN
NaiveBayes	84,9057	91,8919	68,7500
NaiveBayesSimple	84,9057	91,8919	68,7500
NaiveBayesUpdateable	84,9057	91,8919	68,7500
Media Bayes	84,9057	91,8919	68,7500
Logistic	83,0189	94,5946	56,2500
MultilayerPerceptron	**90,5660**	**97,2973**	**75,0000**
RBFNetwork	81,1321	86,4865	68,7500
SimpleLogistic	83,0189	91,8919	62,5000
SMO	77,3585	91,8919	43,7500
VotedPerceptron	81,1321	94,5946	50,0000
Media Function	82,7044	92,7928	59,3750
IB1	75,4717	81,0811	62,5000
IBk	**88,6792**	**97,2973**	**68,7500**
KStar	**88,6792**	**97,2973**	**68,7500**
LWL	81,1321	91,8919	56,2500
Media Lazy	83,4906	91,8919	64,0625
HyperPipes	71,6981	**100,0000**	6,2500
VFI	**77,3585**	83,7838	**62,5000**
Media Misc	74,5283	91,8919	34,375
ConjunctiveRule	73,5849	81,0811	56,2500
DecisionTable	81,1321	91,8919	56,2500
DTNB	77,3585	86,4865	56,2500
JRip	79,2453	91,8919	50,0000
NNge	81,1321	86,4865	**68,7500**
OneR	**83,0189**	91,8919	62,5000
PART	79,2453	91,8919	50,0000
Ridor	**83,0189**	97,2973	50,0000
ZeroR	69,8113	**100,0000**	0,0000
Media Rules	78,6164	90,9910	50,0000
ADTree	**84,9057**	**97,2973**	56,2500
BFTree	**84,9057**	**97,2973**	56,2500
DecisionStump	77,3585	83,7838	**62,5000**
FT	83,0189	91,8919	**62,5000**
J48	81,1321	91,8919	56,2500
J48graft	81,1321	91,8919	56,2500
LADTree	**84,9057**	**97,2973**	56,2500
LMT	83,0189	91,8919	**62,5000**
NBTree	83,0189	91,8919	**62,5000**
RandomForest	83,0189	91,8919	**62,5000**
RandomTree	**84,9057**	**97,2973**	56,2500
REPTree	77,3585	86,4865	56,2500
SimpleCart	77,3585	91,8919	43,7500
Media Trees	82,0029	92,5156	57,6923

Cuadro 3: Resultados de cada clasificador con el nuevo enfoque y datos discretizados

datos numéricos y datos discretizados, respectivamente. Podemos observar en ambas figuras que cuando se incrementa el costo del error sobre estudiantes para los que NO se aceptó la nota, aumenta la tasa de TN. Sin embargo, también disminuye la precisión y la tasa de TP, más aún para el enfoque con datos discretizados. Por lo tanto, es necesario seleccionar un valor de N, para que la tasa de TN pueda mejorar sin afectar excesivamente a la precisión y a la tasa de TP. Por ejemplo en nuestro caso, podemos ver que en ambas figuras N=3 podría ser una buena opción, pues es donde las tasas de acierto quedan balanceadas. Para el caso concreto del clasificador `MultilayerPerceptron` con datos numéricos, esta asignación producía una precisión de 86,79, tasa TP de 86,49 y tasa TN de 86,79.

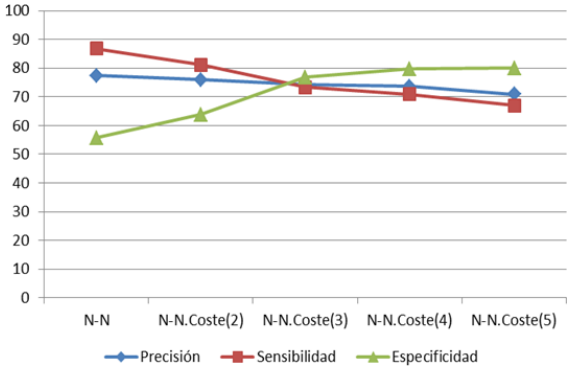

Figura 2: Rendimiento de la clasificación sensible al costo con datos numéricos.

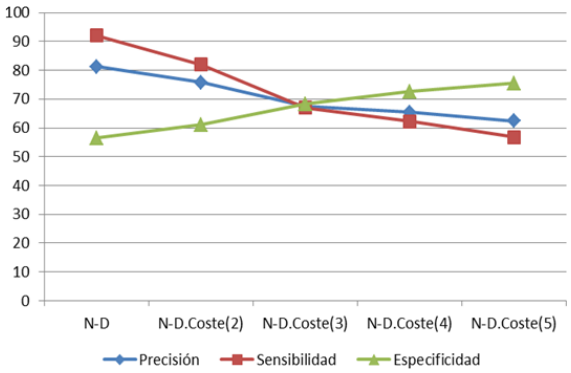

Figura 3: Rendimiento de la clasificación sensible al costo con datos discretizados.

Finalmente, se muestra un ejemplo de un modelo obtenido con uno de los algoritmos de clasificación del segundo experimento en la Figura 4 (`J48`, con precisión 81 %, TP de 92 % y TN de 56 %) :

Diferencia entre las puntuaciones = positiva o ninguna: sí
Diferencia entre las puntuaciones = negativa
| Puntuación Propuesta = suspenso o suficiente: sí
| Puntuación Propuesta = notable
| | Puntuación de Autoevaluación = notable o sobresaliente: SI
| | Puntuación de Autoevaluación = suspenso o suficiente: NO
| Puntuación Propuesta = sobresaliente: NO

Figura 4: Ejemplo de árbol de decisión obtenido.

Hemos seleccionado la salida de dicho algoritmo (`J48`) y experimento, por ser muy fácilmente interpretable, ya que es un algoritmo bien conocido de clasificación de caja blanca, a diferencia otros de caja negra como las redes neuronales. En particular, en el ámbito educativo, la comprensión de los modelos obtenidos es una cuestión importante, ya que debe ser interpretable por los usuarios que no son expertos en minería de datos, tales como instructores, los autores de los cursos y otras partes interesadas. De esta manera, se puede utilizar directamente en los procesos de toma de decisiones. De hecho, se prefieren los modelos de minería

de datos de caja blanca basados en reglas a los modelos de caja negra como el Bayesiano y redes neuronales artificiales, aunque sean normalmente más precisos [21].

La Figura 4 predice que si la diferencia entre las puntuaciones es positiva (es decir, la puntuación de las actividades es superior a la puntuación propuesta) o cero (ambas puntuaciones son iguales), entonces la decisión del instructor es SÍ (es decir, se acepta la puntuación propuesta como la nota final en el curso). Esto resalta la importancia de la correspondencia entre las evaluaciones realizadas durante el curso y la nota propuesta para la predicción del sistema. De lo contrario, si la diferencia entre las puntuaciones es negativa (es decir, la puntuación propuesta es superior a la puntuación de las actividades), entonces la decisión del instructor depende de la propia puntuación propuesta y el resultado del test de autoevaluación: SÍ, si la puntuación propuesta es suspenso o suficiente; NO, si la puntuación propuesta es sobresaliente; y en el caso de proponer notable, SÍ, si el resultado en el test de autoevaluación es notable o sobresaliente y NO, si el resultado es suspenso o suficiente. Ante esta interpretación de la salida de este clasificador, y en el caso de que éste se llevase a explotación, advertimos los siguientes casos en los que existe cierta probabilidad de que alumnado consiga evitar realizar el examen final, y sea calificado con una nota que, por alguna razón que escapase al sistema, pudiese no ser aceptada por el profesorado (recordamos que la tasa TN de este clasificador era inferior al 60 %):

- Proponer un suficiente, dado que el sistema siempre responde SÍ en este caso, con independencia de que el estudiante sea capaz de conseguir dicha nota en el examen.
- Proponer un notable y saber qué responder en el test de autoevaluación. La implantación de este sistema con este clasificador corre el riesgo de que el alumnado aprenda qué debe responder en el test de autoevaluación para poder proponer un notable y que esta nota sea aceptada.

5. Conclusiones y líneas futuras

En este trabajo, se ha propuesto una metodología para la validación de la auto-calificación del estudiante mediante el uso de minería de datos y la autoevaluación. Hemos aplicado diferentes métodos de clasificación para predecir si el instructor aceptará o rechazará la calificación propuesta por el estudiante. El objetivo es evaluar su uso como un esquema de ahorro de tiempo, ya que sólo los estudiantes rechazados (cuyas puntuaciones el instructor no acepta) tendrían que presentarse al examen final al término del curso. En cuanto al rendimiento de la predicción de la decisión del instructor, los resultados obtenidos muestran que (con nuestra base de datos) el uso de la autoevaluación, la discretización del atributo y la clasificación sensible al coste mejoró la precisión, sensibilidad y especificidad. En cuanto al ahorro de tiempo, nuestros experimentos han mostrado una importante dependencia de las predicciones con respecto a la correspondencia entre la nota propuesta y la evaluación de los estudiantes durante el curso, por lo que el esfuerzo requerido por el profesorado pudo haberse más bien distribuido a lo largo de éste, en vez de reducirse. En cuanto a la fiabilidad del proceso, es importante resaltar la posibilidad de que el alumnado aprenda a engañar al sistema para que acepte una calificación que pudiese no ser capaz de obtener en el examen, por lo que sería deseable establecer cierta vigilancia para casos concretos.

En el futuro nos gustaría hacer más experimentos con diferentes cursos y un mayor número de alumnos matriculados en el grado. Además, pretendemos combinar las técnicas de autoevaluación con técnicas de evaluación por pares, de forma que se puedan reducir los errores que pudiera cometer el sistema.

6. Agradecimientos

Este trabajo fue apoyado por el Ministerio de Ciencia y Tecnología con el proyecto TIN2014-55252-P y por la Universidad de Córdoba con el proyecto de innovación docente 2015-2-5021.

Referencias

[1] C. C. Aggarwal. Data Mining - The Textbook. Springer International Publishing, 2015.

[2] C. C. Aggarwal. Data classification - Algorithms and applications. CRC Press, 2015.

[3] L. Bol, R. Riggs, D. J. Hacker y D. Dickerson. The calibration accuracy of middle school students in math classes. Journal of research in education, 21(2):81-96, 2012.

[4] D. Boud. The role of *self-assessment* in student grading. Assessment and Evaluation in Higher Education, 14:20-30, 1989.

[5] D. Boud. Enhancing learning through self-assessment. Routledge, 2013

[6] A. Brew. Towards autonomous assessment: using self-assessment and peer assessment. En Assessment matters in higher education: Choosing and using diverse approaches, páginas 159-171, 1999.

[7] J. Davis y D. Rand. *Self-grading* versus instructor grading. Journal of Educational Research, 73(4):207-217, 2001.

[8] J. Dougherty, R. Kohavi y M. Sahami. Supervised and unsupervised discretization of continuous features. En actas del 12th International Conference on Machine Learning, páginas 194–202, 1995.

[9] C. Elkan. The foundations of cost-sensitive learning. En International Joint Conf. on Artificial Intelligence, páginas 1-6, 2001.

[10] N. Falchikov y D. Boud. Student self-assessment in higher education: A meta-analysis. Review of Educational Research, 59(4), 395-430, 1989.

[11] N. Falchikov. Improving assessment through student involvement. Practical solutions for aiding learning in higher and further education. RoutledgeFalmer, 2005

[12] E. Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I.H. Witten y L. Trigg. Weka-a machine learning workbench for data mining. En Data mining and knowledge discovery handbook, páginas 1269-1277, 2009.

[13] J. Han, M. Kamber y J. Pei Data Mining: Concepts and Techniques, 3^{rd} ed. Morgan Kaufmann Publishers, 2011.

[14] G. Holmes, A. Donkin y I.H. Witten. Weka: A machine learning workbench. En actas del Second IEEE Australian and New Zealand Conference on Intelligent Information Systems, páginas 357-361, 1994.

[15] V.E. Johnson. Beyond Grade Inflation: Grading Problems in Higher Education (review). The Review of Higher Education, 30(1):76-77, 2006

[16] A. Kohn The case against grades. Educational Leadership, 69(3):28-33, 2011

[17] C. Kulkarni, K.P. Wei, H. Le, D. Chia, K. Papadopoulos, J. Cheng, D. Koller y S.R. Klemmer. Peer and Self Assessment in Massive Online Classes En Design Thinking Research, Springer International Publishing, páginas 131-168, 2015

[18] G.F. Madaus y D.L. Stufflebeam (Eds). Educational Evaluation: Classic Works of Ralph W. Tyler. Springer Netherlands, 1989

[19] L. Rediehs. Trust and distrust: The problem with traditional grading. Página web accesible el 17 de mayo de 2016 en `http://it.stlawu.edu/~lrediehs/grading\_files/problem.htm`

[20] S.N. Roman y S.A. Soriano. Autonomous Learning and Self-Assessment through the European Language Portfolio (ELP): A Pilot Study on Primary Education. The Journal of Language Teaching and Learning, 5(2):37-53, 2015

[21] C. Romero y S. Ventura. Data Mining in Education. Wire Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3:12-27, 2013.

[22] M.D. Shermis, J. Bursten, D. Higgins y K. Zechner Automated essay scoring: Writing assessment and instruction. En International encyclopedia of education, páginas 75-80, 2010.

[23] B. Strong, M. Davis y V. Hawks. *Self-grading* in large general education classes. Colleague Teaching, 52(2):52-57, 2004.

[24] K. Tan. Qualitatively different way of experiencing students self-assessment. Higher Education Research and Development. 52:52-57, 2008.

[25] M. Taras. Student self-assessment: what have we learned and what are the challenges. RELIEVE-Revista Electrónica de Investigación y Evaluación Educativa 21(1), art. ME8:1-14, 2015

[26] Test de Autoevaluación para Estudiantes Secundarios y Universitarios. Disponible en `http://www.tecnicas-de-estudio.org/general/sabes-estudiar.htm`

[27] F. Wilcoxon. Individual comparisons by ranking methods. Biometrics 1:80-83, 1945.