Diálogos sobre Educación
TEMAS ACTUALES EN INVESTIGACIÓN EDUCATIVA

año 10 | número 19 | julio-diciembre 2019 | ISSN 2007-2171
Recibido: 12 de mayo de 2018 | Aceptado: 27 de marzo de 2019

Construcción e implementación de un modelo para predecir el rendimiento académico de estudiantes universitarios mediante el algoritmo *Naïve Bayes*

Construction and implementation of a model to predict the academic performance of university students using the Naïve Bayes algorithm

DOI: 10.32870/dse.v0i19.509

Andrés Rico Páez*
Nora Diana Gaytán Ramírez**
Daniel Sánchez Guzmán***

Resumen

Una de las aplicaciones más utilizadas de la minería educativa de datos es la predicción del rendimiento académico. El objetivo de este trabajo es presentar la construcción, evaluación y aplicación de un modelo predictivo del rendimiento académico de estudiantes universitarios por medio de la técnica de minería de datos conocida como algoritmo *Naïve Bayes*. En este trabajo se recabaron datos de 122 estudiantes como entrenamiento para el algoritmo y se aplicó el modelo para predecir el rendimiento académico de 71 estudiantes. Los resultados muestran que el modelo predictivo, además de obtener predicciones del rendimiento académico, también identifica los factores que más influyen en él. Este tipo de estudios permite a los profesores diseñar estrategias de prevención e identificar estudiantes que son vulnerables a reprobar.

Palabras clave: predicción – rendimiento académico – minería de datos – modelo predictivo – algoritmo *Naïve Bayes*.

Abstract

One of the most widely used applications of educational data mining is predicting academic performance. The aim of this paper is to present the construction, evaluation and implementation of a predictive model of the academic performance of university students by means of the data mining technique known as the Naïve Bayes algorithm. We collected data from 122 students as training for the algorithm and applied the

* Maestro en Ciencias en la especialidad de Ingeniería Eléctrica con opción en Comunicaciones. Estudiante del posgrado en Tecnología Avanzada. Instituto Politécnico Nacional. México. aricop.ipn@gmail.com

** Ingeniero en Comunicaciones y Electrónica. Estudiante del posgrado en Tecnología Avanzada en el Centro de Investigación de Ciencia Aplicada y Tecnología Avanzada, Instituto Politécnico Nacional. México. nora_diana@hotmail.com

*** Doctor en Tecnología Avanzada. Profesor en el Centro de Investigación de Ciencia Aplicada y Tecnología Avanzada, Unidad Legaria del Instituto Politécnico Nacional. México. dsanchez@ipn.mx

model to predict the academic performance of 71 students. The results show that, in addition to obtaining predictions of academic performance, the predictive model also identifies the factors that influence it the most. This type of study allows teachers to design prevention strategies and identify students who are vulnerable to failure.

Keywords: prediction – academic performance – data mining – predictive model – Naïve Bayes algorithm.

Antecedentes

Existe un crecimiento en el uso y manejo de las Tecnologías de la Información y la Comunicación (TIC) en diversas áreas, debido al surgimiento de aplicaciones generadas por la actividad humana tales como la internet, los sistemas de comunicaciones móviles celulares, las redes inalámbricas, entre muchas otras. Este avance tecnológico ha propiciado un incremento en la cantidad de información a almacenar, la cual es generada para objetivos específicos y, una vez cumplidos, es ignorada o eliminada. No obstante, dicha información puede ser analizada para extraer algún tipo de conocimiento para generar un beneficio. Actualmente, este tipo de análisis se realiza por medio de técnicas de minería de datos, principalmente, en áreas comerciales y empresariales (Han, 2012). Debido a su éxito, se ha comenzado a utilizar estas técnicas en ambientes educativos con el objetivo de descubrir información útil que pueda beneficiar a los procesos de enseñanza y aprendizaje de diversas instituciones educativas. De esta manera, existe una tendencia al uso de técnicas de minería de datos en ambientes educativos (Romero y Ventura, 2010; 2012; Peña, 2014). En países de Latinoamérica, es un área reciente (Estrada, Zamarripa, Zúñiga y Martínez, 2016), por lo que existen diferentes líneas de investigación para su uso y desarrollo. Una de estas líneas de investigación es la predicción del rendimiento académico de estudiantes. El rendimiento académico se refiere al nivel de logro o éxito que se puede alcanzar en una o varias asignaturas (Reynoso y Méndez, 2018) y es parte esencial en las universidades, debido a que es uno de sus principales criterios de calidad académica (Shahiri, Husain y Rashid, 2015).

La capacidad de predecir el rendimiento académico ofrece beneficios al profesor, a los estudiantes y a la institución educativa, tales como: plantear programas de prevención estratégicos para estudiantes con bajo rendimiento, detectar estudiantes en peligro de deserción, identificar características de los estudiantes que les permiten obtener un buen rendimiento académico, entre muchas otras.

Perspectiva teórica

La reprobación estudiantil de uno o varios cursos es un insuficiente rendimiento cuantitativo y/o cualitativo de las capacidades de un estudiante para conseguir los parámetros mínimos

establecidos por una institución educativa, lo cual limita o detiene el avance de un estudiante en su vida académica (Rodallegas *et al.*, 2010). La reprobación puede ocasionar la deserción escolar, es decir, el abandono definitivo y sin causa justificada de la institución educativa por parte del estudiante sin haber terminado su etapa educativa que esté cursando. En el caso de las universidades, las principales problemáticas son los índices de reprobación y la deserción escolar, lo cual provoca bajos índices de eficiencia terminal (Vera, Ramos, Sotelo, Echeverría y Serrano, 2012). Con el objetivo de abatir estos problemas educativos y brindar una buena formación académica, las instituciones educativas están interesadas en retener a sus estudiantes. No obstante, a pesar de los programas que se han implementado para evitar el fracaso escolar, disminuir la reprobación y evitar el abandono, no es sencillo debido a que es un problema multifactorial.

En la actualidad, para reducir la reprobación y deserción de estudiantes en las instituciones educativas, se han aplicado con éxito técnicas de minería de datos para crear modelos predictivos del rendimiento académico (Márquez, Romero y Ventura, 2012). Algunos atributos de estudiantes que se han utilizado en la literatura para predicción del rendimiento académico son las evaluaciones internas, el promedio actual y otros factores socioeconómicos (Shahiri *et al.*, 2015).

De manera general, la minería de datos es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, de datos almacenados (Hernández, Ramírez y Ferri, 2004; Witten, Frank y Hall, 2005). Este proceso, aplicado a la educación, se conoce como minería educativa de datos, la cual surge como un paradigma orientado al diseño, tareas, métodos y algoritmos, con el objetivo de explorar los datos del ambiente educativo (Peña, 2014). Tiene como propósito descubrir conocimiento y patrones dentro datos educativos para realizar predicciones del comportamiento de estudiantes (Luan, 2002), es decir, permite encontrar y analizar patrones que caractericen los comportamientos con base en sus logros, evaluaciones y dominio del contenido de conocimiento que tienen los alumnos en los diversos mecanismos de enseñanza y aprendizaje de las diversas instituciones públicas y privadas (Ballesteros y Sánchez, 2013).

De forma más amplia, el proceso completo de aplicación de técnicas de minería de datos es conocido como descubrimiento de conocimiento en bases de datos (*Knowledge Discovery in Databases*, KDD) (Hernández *et al.*, 2004), el cual coloca a la minería de datos como una de las fases del mismo. Este proceso se muestra esquemáticamente en la figura 1.

Figura 1. Proceso de descubrimiento de conocimiento en bases de datos

Fuente: elaboración propia a partir de la metodología mostrada en Hernández *et al.* (2004).

La primera fase del proceso KDD es la integración y recopilación de datos, en la cual se determinan las fuentes de información y la manera de conseguirlas para formar la base de datos a utilizar. La siguiente fase es el procesamiento, en donde se realiza la selección de los datos más relevantes para el estudio y la limpieza de datos para eliminar problemas tales como datos erróneos o faltantes. En la transformación de los datos se convierten a una representación que sea manejable para la técnica de minería de datos seleccionada (Hernández *et al.*, 2004). Posteriormente, se encuentra la fase de minería de datos, en la que define el algoritmo de minería de datos a implementar. Finalmente está la fase de evaluación, en la cual se determina la validez y confiabilidad de los patrones obtenidos que representan el conocimiento extraído.

Entre las tareas de minería educativa de datos, las predictivas son de las más populares y de uso más extendido (Romero y Ventura, 2010, 2012; Peña, 2014) debido a que permiten detectar problemas académicos con anticipación y tomar las decisiones más adecuadas al problema.

Una de las tareas predictivas es la clasificación, la cual consiste en predecir la clase de registros de datos de prueba por medio de registros de datos de entrenamiento. De manera más específica, existe un conjunto de atributos $\{A_1,...,A_n\}$ y una variable de clase C_i perteneciente a un conjunto $\Omega_C=\{C_1,\ldots,C_k\}$. La probabilidad *a posteriori* de la variable de clase C_i dado un conjunto de atributos, se calcula a partir del teorema de Bayes de la siguiente forma:

$$P(C_i|A_1,\ldots,A_n)=[P(A_1,\ldots,A_n|C_i)P(C_i)]/P(A_1,\ldots,A_n) \quad (1)$$

Para clasificar un registro, es necesario identificar el valor más probable y devolverlo como resultado. En el teorema de Bayes, la hipótesis más probable es aquella con máxima probabilidad *a posteriori* (MAP). De esta forma, el valor de la clase más probable es:

$$C_{MAP}=\arg\max_{C_i E \Omega_C} P(C_i|A_1,\ldots,A_n)$$

$$=\arg\max_{C_i E \Omega_C} [P(A_1,\ldots,A_n|C_i)P(C_i)]/P(A_1,\ldots,A_n) \quad (2)$$

$$=\arg\max_{C_i E \Omega_C} P(A_1,\ldots,A_n|C_i)P(C_i)$$

El algoritmo conocido como *Naïve Bayes* (Hernández *et al.*, 2004; Witten *et al.*, 2005) supone que todos los atributos son independientes una vez conocido el valor de la clase. Con base en esta suposición, el valor de la clase a devolver es:

$$C_{MAP}=\arg\max_{C_i E \Omega_C} P(C_i) \prod^{n}_{j=1} P(A_j|C_i) \quad (3)$$

Cabe mencionar que la exactitud de las predicciones correctamente clasificadas con el algoritmo *Naïve Bayes* es semejante o superior al de otras técnicas de minería de datos (Michie, Spiegelhalter y Taylor, 1994; Kotsiantis, Pierrakeas y Pintelas, 2003).

A partir de las fórmulas anteriores, se construye el modelo predictivo mediante la estimación de las probabilidades *a priori* y *a posteriori*. Las probabilidades *a priori* $P(C_i)$ se estiman dividiendo el número de registros de la clase C_i de los datos de entrenamiento entre el total de los mismos. La estimación de las probabilidades *a posteriori* $P(A_j|C_i)$ de cada atributo discreto se calculan dividiendo el número de casos de aparición del evento entre el número de casos totales. En este trabajo, para solucionar el caso en el que las probabilidades $P(A_j|C_i)$ sean igual a cero, se utiliza la estimación basada en la ley de sucesión de Laplace (Hernández *et al.*, 2004), la cual consiste en obtener el número de casos favorables más uno, dividido entre el número de casos totales más el número de valores posibles.

Una vez construido el modelo, se puede clasificar un nuevo registro calculando las probabilidades de sus atributos y aplicando la fórmula 3 para determinar a qué clase corresponde.

Planteamiento del problema

Un problema que se ha identificado es la necesidad de modelos predictivos del rendimiento académico en estudiantes universitarios con técnicas de minería de datos para diseñar programas de prevención de reprobación y deserción escolar. Esta necesidad es evidente en México debido al poco desarrollo de este tipo de modelos, a pesar del potencial beneficio en la mejora del rendimiento académico de estudiantes.

La presente investigación se realizó para contestar las siguientes preguntas: ¿Cómo construir un modelo predictivo del rendimiento académico en estudiantes universitarios mediante el algoritmo *Naïve Bayes* y cuál es su exactitud en las predicciones? y ¿Cómo implementar el modelo predictivo construido utilizando lenguajes de programación adecuados para su uso en la web? De esta manera, el objetivo de esta investigación es construir un modelo para predecir el rendimiento académico de estudiantes universitarios mediante el algoritmo *Naïve Bayes*, obtener su exactitud en las predicciones y aplicarlo mediante lenguajes de programación para su uso en la web.

Metodología

En esta sección se presenta la primera fase del proceso KDD, la cual consiste en la integración y recopilación de datos.

Participantes

La muestra de datos corresponde a estudiantes inscritos en una carrera de ingeniería perteneciente al Instituto Politécnico Nacional en la Ciudad de México. Los datos recopilados fueron la aprobación y reprobación de los estudiantes en un curso de matemáticas. También se recopilaron datos personales y escolares que han sido utilizados en la literatura para la predicción del rendimiento académico, tales como ingreso familiar, escolaridad del padre, entre otros (Shahiri *et al*., 2015). En total participaron 122 estudiantes en este estudio. En Kotsiantis *et al*. (2003), y Mueen, Zafar y Manzoor (2016), se ha observado que, con cantidades similares, el algoritmo *Naïve Bayes* ha obtenido una exactitud mayor en comparación con otras técnicas de minería de datos.

Instrumento

La información de la aprobación y reprobación de los estudiantes participantes en este estudio fue proporcionada por los docentes del curso. Los demás datos fueron recopilados por medio de una encuesta, la cual estuvo formada por una hoja tamaño carta en la que se especifica el propósito y los datos a recabar. De manera oral, se indicó a los participantes que la información recabada tiene fines estadísticos y de investigación. En la figura 2 se muestra el formato del instrumento utilizado para recabar la información.

Figura 2. Formato del instrumento utilizado para recabar la información

Esta encuesta tiene el propósito de analizar algunas características de los alumnos que inciden en su desempeño académico. La información obtenida es anónima y será utilizada sólo con fines estadísticos y de investigación.

Matrícula: _____

Escolaridad del padre:

Primaria__ Secundaria__ Bachillerato__ Universidad__ Posgrado__

Escolaridad de la madre:

Primaria__ Secundaria__ Bachillerato__ Universidad__ Posgrado__

Ingresos familiares:

Menos de $5,000_____ De $5,000 a $10,000_____ Más de $10,000_____

Promedio final obtenido en el bachillerato: _____

Cantidad de materias reprobadas actualmente: _____

Promedio actual: _____

¿Cómo prefieres estudiar? Solo_____ En dúo_____ En grupo_____

¿Cómo prefieres realizar actividades en clase? Solo__ En dúo__ En grupo__

¿Qué tan frecuentemente estudias?:

Continuamente___

Una semana antes de la evaluación___

Un día antes de la evaluación___

FIN DE LA ENCUESTA

¡Muchas gracias!

Fuente: elaboración propia.

Una vez que se recopilaron los datos se formó una tabla con 122 registros. En la figura 3 se presenta una muestra de los datos de esta tabla.

Figura 3. Muestra de la tabla de datos recopilados

escpadre	escmadre	ingresof	promant	matrep	promactual	prefestudio	prefact	frestudio	aprueba
secundaria	primaria	5000-10000	7.6	1	7.96	solo	grupo	continuo	SI
media superior	primaria	5000-10000	7.8	1	7.5	duo	solo	continuo	SI
superior	secundaria	<5000	8	3	6.9	solo	solo	semana antes	SI
superior	secundaria	5000-10000	6.95	2	7.2	solo	solo	semana antes	NO
superior	superior	5000-10000	8	1	7.5	grupo	grupo	semana antes	SI
media superior	media superior	5000-10000	7.1	3	5.86	solo	grupo	continuo	NO
superior	superior	>10000	8.5	0	7.1	solo	grupo	semana antes	SI
primaria	primaria	<5000	8.1	1	6.4	grupo	grupo	continuo	SI
media superior	media superior	5000-10000	9.65	3	6.42	duo	duo	continuo	SI
superior	superior	5000-10000	8.8	2	7.2	solo	grupo	semana antes	NO
secundaria	media superior	<5000	8.2	2	6.2	grupo	duo	continuo	NO
superior	media superior	5000-10000	8	2	6.7	duo	duo	semana antes	NO
posgrado	secundaria	5000-10000	7.3	1	7.33	grupo	grupo	semana antes	NO
media superior	media superior	<5000	8.3	1	6	solo	grupo	continuo	NO
superior	secundaria	5000-10000	7.8	2	5.4	grupo	grupo	semana antes	NO
posgrado	superior	5000-10000	7.5	0	7.4	solo	solo	semana antes	SI
media superior	primaria	5000-10000	8	1	7	grupo	duo	semana antes	NO
secundaria	media superior	<5000	7.8	1	6.4	duo	solo	continuo	SI
media superior	media superior	5000-10000	7.81	0	8.33	duo	grupo	semana antes	SI
superior	superior	>10000	7.81	1	8.1	solo	grupo	dia antes	SI
secundaria	secundaria	<5000	7	1	7.5	solo	duo	semana antes	SI
primaria	primaria	5000-10000	9.3	0	7.83	solo	duo	continuo	SI
secundaria	secundaria	5000-10000	7.1	1	7.2	solo	grupo	semana antes	SI
superior	superior	>10000	8.96	0	8.83	solo	duo	continuo	SI
posgrado	superior	>10000	7.3	1	7.4	solo	solo	continuo	SI
secundaria	media superior	<5000	8.7	1	7.5	solo	grupo	semana antes	SI

Fuente: elaboración propia.

En los datos recopilados no se tuvieron datos faltantes o erróneos, por lo que solo es necesaria la transformación de datos para el algoritmo de predicción utilizado. Esta transformación se explica en la siguiente sección.

Desarrollo

En esta sección se realiza la transformación de datos recopilados y la implementación del modelo predictivo mediante lenguajes de programación.

En la figura 3 se observa que los datos recopilados contienen diez atributos (columnas), de los cuales el atributo "aprueba" define la etiqueta de la clase. Los datos se transformaron a valores nominales para su mejor manejo en la fase de minería de datos. En la tabla 1 se muestran todos los atributos utilizados y sus posibles valores nominales.

Tabla 1. Atributos de los estudiantes con sus posibles valores

Atributos	Valores posibles
Escolaridad del padre (escpadre)	Educación primaria y secundaria (básica), educación media superior (msup), educación superior o mayor (sup)
Escolaridad de la madre (escmadre)	Educación primaria y secundaria (básica), educación media superior (msup), educación superior o mayor (sup))
Ingreso familiar (ingresof)	Menos de $5000 (bajo), entre $5000 y $10000 (medio), más de $10000 (alto)
Promedio de media superior (promant)	Entre 0 y 7.4 (bajo), entre 7.5 y 8.4 (medio), entre 8.5 y 10 (alto)
Materias reprobadas actualmente (matrep)	Cero materias (cero), una materia (una), dos o más materias (dosomas)
Promedio actual (promactual)	Entre 0 y 7.4 (bajo), entre 7.5 y 8.4 (medio), entre 8.5 y 10 (alto)
Preferencia de estudio (prefestudio)	Estudiar solo (solo), estudiar en dúo (duo), estudiar en grupo (grupo)
Preferencia para realizar actividades (prefact)	Realizar actividades solo (solo), realizar actividades en dúo (duo), realizar actividades en grupo (grupo)
Frecuencia de estudio (frestudio)	Estudiar de forma continua (continuo), estudiar una semana antes del examen (semantes), estudiar un día antes del examen (diantes)
Aprobación del curso (aprueba)	SÍ, NO

Fuente: elaboración propia.

En la figura 4, se presenta una muestra de los datos transformados a valores nominales.

Figura 4. Muestra de la tabla de datos transformados a valores nominales

escpadre	escmadre	ingresof	promant	matrep	promactual	prefestudio	prefact	frestudio	aprueba
basica	basica	medio	medio	una	medio	solo	grupo	continuo	SI
msup	basica	medio	medio	una	medio	duo	solo	continuo	SI
sup	basica	bajo	medio	dosomas	bajo	solo	solo	semantes	SI
sup	basica	medio	bajo	dosomas	bajo	solo	solo	semantes	NO
sup	sup	medio	medio	una	medio	grupo	grupo	semantes	SI
msup	msup	medio	bajo	dosomas	bajo	solo	grupo	continuo	NO
sup	sup	alto	alto	cero	bajo	solo	grupo	semantes	SI
basica	basica	bajo	medio	una	bajo	grupo	grupo	continuo	SI
msup	msup	medio	alto	dosomas	bajo	duo	duo	continuo	SI
sup	sup	medio	alto	dosomas	bajo	solo	grupo	semantes	NO
basica	msup	bajo	medio	dosomas	bajo	grupo	duo	continuo	NO
sup	msup	medio	medio	dosomas	bajo	duo	duo	semantes	NO
sup	basica	medio	bajo	una	bajo	grupo	grupo	semantes	NO
msup	msup	bajo	medio	una	bajo	solo	grupo	continuo	NO
sup	basica	medio	medio	dosomas	bajo	grupo	grupo	semantes	NO
sup	sup	medio	medio	cero	bajo	solo	solo	semantes	SI
msup	basica	medio	medio	una	bajo	grupo	duo	semantes	NO
basica	msup	bajo	medio	una	bajo	duo	solo	continuo	SI
msup	msup	medio	medio	cero	medio	duo	grupo	semantes	SI
sup	sup	alto	medio	una	medio	solo	grupo	diantes	SI
basica	basica	bajo	bajo	una	medio	solo	duo	semantes	SI
basica	basica	medio	alto	cero	medio	solo	duo	continuo	SI
basica	basica	medio	bajo	una	bajo	solo	grupo	semantes	SI
sup	sup	alto	alto	cero	alto	solo	duo	continuo	SI
sup	sup	alto	bajo	una	bajo	solo	solo	continuo	SI
basica	msup	bajo	alto	una	medio	solo	grupo	semantes	SI

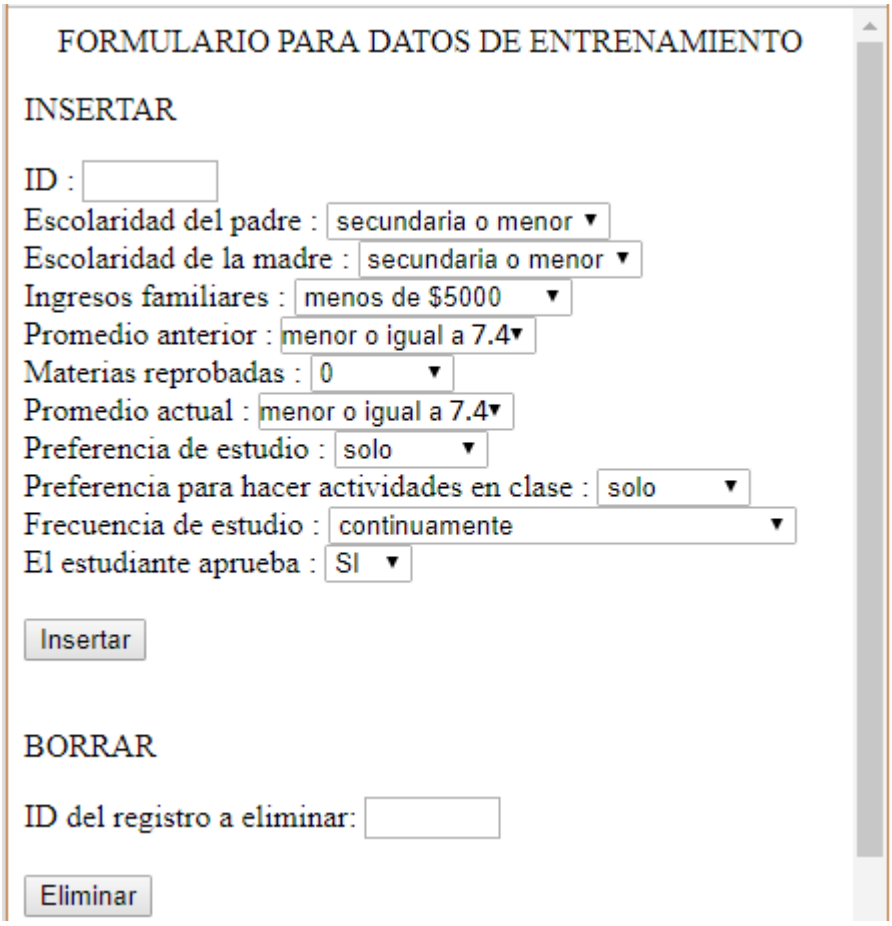
Fuente: elaboración propia.

Una vez transformados los datos, se les aplica la técnica de minería de datos para construir el modelo predictivo.

En este trabajo, la tarea predictiva empleada es la clasificación, y la técnica utilizada es el algoritmo *Naïve Bayes*. El modelo predictivo se construye por medio del cálculo de las probabilidades *a priori* y *a posteriori,* descrito en las secciones anteriores. Por ejemplo, el número de registros en donde el atributo "aprueba" es igual a "SÍ" es de 65 de 122 posibles, por lo que la probabilidad *a priori* P(aprueba=SÍ) es igual a 65/122 = 0.5328. Otro ejemplo, es el número de registros en donde el atributo "escpadre" es igual a "básica" dado que "aprueba" es igual a "SÍ", tomando en cuenta la ley de sucesión de Laplace, es de 28 de 68 posibles; de esta manera, la probabilidad *a posteriori* P(escpadre=basica/aprueba=SÍ) es igual a 0.4118. La estimación de este tipo de probabilidades es laboriosa debido a la cantidad de datos a analizar, por lo que

existen diferentes herramientas informáticas que permiten aplicar técnicas de minería de datos para crear modelos predictivos, tal y como se ha hecho en Jaramillo y Paz (2015), y Pacheco y Fernández (2015). No obstante, este tipo de herramientas requieren que los usuarios tengan conocimientos especializados, tanto en el uso de la herramienta como en minería de datos. Por lo que, a diferencia de estos trabajos y de manera similar a Valero, Salvador y García, (2010), en este trabajo se realizó una plataforma en la que se programó el algoritmo *Naïve Bayes* en HTML5 (*HyperText Markup Language*, versión 5) y PHP (*Hypertext Pre-Processor*) con el propósito de publicarlo en un sitio web como un apoyo a profesores, no solo de la institución donde fueron recabados los datos sino de cualquier institución educativa. Esta plataforma permite introducir datos de entrenamiento de un número variable de estudiantes y de cualquier área de estudios para calcular automáticamente las probabilidades *a priori y a posteriori*. La interfaz gráfica para introducir los datos de entrenamiento se muestra en la figura 4.

Figura 4. Interfaz gráfica de la plataforma para introducir los datos de entrenamiento



Fuente: elaboración propia.

Por medio de esta plataforma se obtuvieron todas las probabilidades *a priori* y *a posteriori* que componen el modelo predictivo, las cuales se presentan en la tabla 2.

Tabla 2. Probabilidades estimadas con algoritmo *Naïve Bayes*

Probabilidades	***clase = SÍ***	***clase = NO***
P(aprueba=clase)	0.5328	0.4672
P(escpadre=basica/aprueba=clase)	0.4118	0.2667
P(escpadre=msup/aprueba=clase)	0.3235	0.3333
P(escpadre=sup/aprueba=clase)	0.2647	0.4
P(escmadre=basica/aprueba=clase)	0.4118	0.3333
P(escmadre=msup/aprueba=clase)	0.3529	0.4167
P(escmadre=sup/aprueba=clase)	0.2353	0.25
P(ingresof=bajo/aprueba=clase)	0.3382	0.35
P(ingresof=medio/aprueba=clase)	0.5147	0.45
P(ingresof=alto/aprueba=clase)	0.1471	0.2
P(promant=bajo/aprueba=clase)	0.2794	0.4167
P(promant=medio/aprueba=clase)	0.5147	0.4167
P(promant=alto/aprueba=clase)	0.2059	0.1666
P(matrep=cero/aprueba=clase)	0.5	0.3333
P(matrep=una/aprueba=clase)	0.4118	0.25
P(matrep=dosomas/aprueba=clase)	0.0882	0.4167
P(promactual=bajo/aprueba=clase)	0.4853	**0.7833**
P(promactual=medio/aprueba=clase)	0.3971	0.15
P(promactual=alto/aprueba=clase)	0.1176	0.0667
P(prefestudio=solo/aprueba=clase)	**0.5294**	0.4167
P(prefestudio=duo/aprueba=clase)	0.2206	0.2833
P(prefestudio=grupo/aprueba=clase)	0.25	0.3
P(prefact=solo/aprueba=clase)	0.2647	0.2
P(prefact=duo/aprueba=clase)	0.3235	0.3
P(prefact=grupo/aprueba=clase)	0.4118	0.5
P(frestudio=continuo/aprueba=clase)	0.3676	0.4
P(frestudio=semantes/aprueba=clase)	0.5	0.5167
P(frestudio=diantes/aprueba=clase)	0.1324	0.0833

Fuente: elaboración propia.

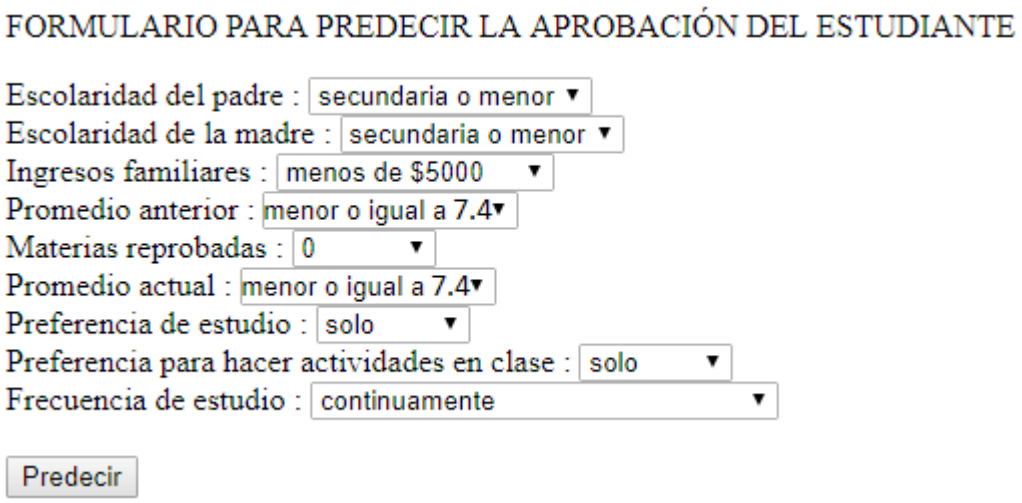
De la tabla 2 se observa que la probabilidad *a posteriori* P(prefestudio=solo/aprueba=SÍ) = 0.5294 es la más alta de la clase aprueba=SÍ y la probabilidad *a posteriori* P(promactual=bajo/

aprueba=NO) = 0.7833 es la más alta de la clase aprueba=NO. En el algoritmo *Naïve Bayes*, las probabilidades *a posteriori* se multiplican para predecir el rendimiento académico, por lo que las probabilidades *a posteriori* con valores mayores son las que más influyen en la aprobación y en la reprobación. De esta manera, en el modelo predictivo construido, el factor que más influye en que un estudiante repruebe el curso es si tiene un promedio actual bajo (≤7.4). De la misma forma, el factor que más influye en que un estudiante apruebe el curso es si el estudiante prefiere estudiar solo.

Resultados

Una vez construido el modelo, se puede predecir la aprobación de un nuevo estudiante a partir de las probabilidades de la tabla 2 que correspondan a los atributos de dicho estudiante y aplicando la fórmula 3. Esto se puede hacer de forma automática por medio de la plataforma construida, ya que permite calcular la predicción de un nuevo registro, una vez estimadas las probabilidades del modelo. La interfaz gráfica para introducir los datos de prueba se muestra en la figura 5.

Figura 5. Interfaz gráfica de la plataforma para introducir los datos de prueba



Fuente: elaboración propia.

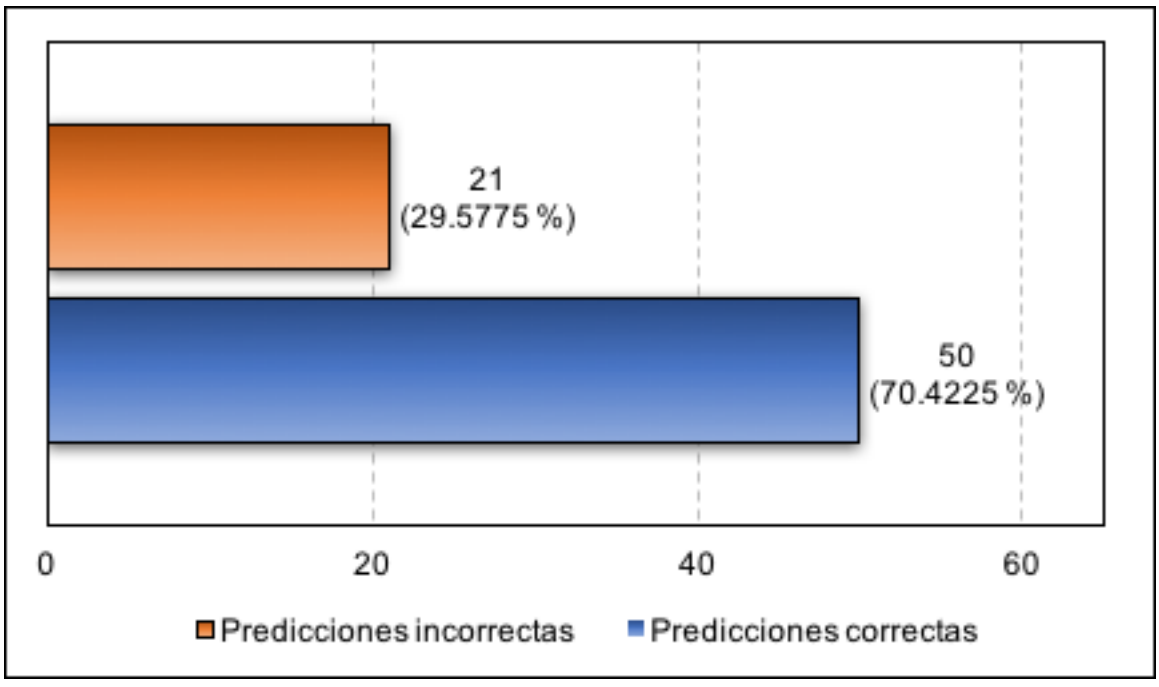
Para evaluar el modelo predictivo construido se calcula la exactitud de las predicciones (porcentaje de la cantidad de registros con predicciones correctas entre el total de registros), utilizando el método conocido como validación cruzada (Hernández *et al.*, 2004). Este método consiste en dividir aleatoriamente el total de los datos de entrenamiento en un número fijo de grupos; en este caso, se dividió en dos conjuntos equitativos. Se construye un modelo con

el primer conjunto y se usa para predecir los resultados del segundo conjunto y se calcula su exactitud. Después, se construye un modelo con el segundo conjunto y se usa para predecir los resultados del segundo conjunto y se calcula su exactitud. Finalmente, se calcula la exactitud del modelo construido promediando las exactitudes calculadas anteriormente.

De esta manera, se dividieron en forma aleatoria los 122 registros de estudiantes en dos conjuntos de 61 registros, en donde un conjunto fue de entrenamiento y el otro de prueba para calcular su exactitud y viceversa. El promedio de estas exactitudes en porcentaje fue de 61.4754%, el cual representa la exactitud estimada del modelo predictivo.

Para comprobar la exactitud estimada del modelo predictivo construido, se aplicó el modelo para predecir el rendimiento académico de 71 estudiantes del mismo curso, pero del siguiente semestre. Primeramente, se utilizó el modelo predictivo construido y la plataforma realizada para obtener predicciones de los estudiantes. Posteriormente, se compararon estas predicciones con los resultados reales obtenidos por los estudiantes al final del curso. En la figura 6 se presenta el número de predicciones correctas e incorrectas. Además, se observa que la exactitud de las predicciones fue de 70.4225 %, la cual es superior a la estimada con el método de validación cruzada.

Figura 6. Cantidad de predicciones correctas e incorrectas de los datos de prueba



Fuente: elaboración propia.

Discusión y conclusiones

El objetivo de este trabajo ha sido presentar la construcción de un modelo predictivo del rendimiento académico de estudiantes universitarios, obtener la exactitud en sus predicciones e implementarlo en una plataforma educativa, en la cual se utilizó el algoritmo *Naïve Bayes*. Este algoritmo ha sido utilizado en la literatura para predecir el rendimiento académico (Shahiri *et al.*, 2015), además, ha presentado una buena exactitud en las predicciones con una cantidad pequeña de datos de entrenamiento (Kavipriya, 2016). Por lo que este algoritmo es útil en ambientes educativos que generen pocos datos, tal es el caso de grupos de estudiantes de clases presenciales. Los resultados de este estudio muestran cómo, a partir de algunos datos académicos y personales de estudiantes al inicio de un curso, permite predecir con cierto porcentaje de exactitud el rendimiento académico de estudiante al final del curso. El diseño del modelo predictivo se realizó a partir del proceso de descubrimiento de extracción de conocimiento de bases de datos. El modelo predictivo se construyó en una plataforma que permite ingresar los datos de los estudiantes para construir el modelo y, posteriormente, introducir los datos de estudiantes a los cuales se les va a predecir su rendimiento académico. La plataforma construida acepta diferentes cantidades de registros de estudiantes de cualquier área, con los atributos mostrados en este análisis. Tiene la opción de introducir los datos de entrenamiento para calcular las probabilidades del modelo y de introducir los datos de prueba para realizar las predicciones del rendimiento académico. Esto brinda la posibilidad de ser utilizada por profesores que no tengan conocimientos de minería de datos.

El algoritmo *Naïve Bayes* predice si un estudiante aprueba o no con base en el producto de sus probabilidades, como lo indica la ecuación 3. Los valores más grandes de probabilidad son los que influyen de forma más significativa en la decisión de la predicción de aprobación o reprobación. El primer renglón de la tabla 2 representa las probabilidades *a priori* de aprobación y reprobación. En este trabajo, la probabilidad P(aprueba=SÍ) es mayor que P(aprueba=SÍ) debido a que la cantidad de estudiantes aprobados es mayor que la de reprobados en los datos de entrenamiento. A partir del siguiente renglón de la tabla 2, se encuentran las probabilidades *a posteriori*. Se observa que la probabilidad *a posteriori* más alta de la tercera columna (probabilidad que más influye en la reprobación) es P(promactual=bajo/aprueba=NO) = 0.7833, la cual ocurre cuando el promedio actual del estudiante es bajo (≤7.4). De la misma manera, la probabilidad *a posteriori* más alta de la segunda columna (probabilidad que más influye en la aprobación) es P(prefestudio=solo/aprueba=SÍ) = 0.5294, la cual sucede cuando el estudiante prefiere estudiar solo. El hecho de que los estudiantes que tienen un promedio actual bajo tengan mayor probabilidad de reprobar el curso, se debe a que el promedio actual es uno de los principales indicadores del rendimiento académico de estudiantes (Shahiri *et al.*, 2015). En cuanto al hecho de que los estudiantes prefieran estudiar solos y tengan mayor probabilidad de aprobar el curso, tiene que

ver con el poco tiempo que tienen para hacerlo o la falta de hábito de hacerlo en equipo, como se menciona en Álvarez, Gugelmeier y Hermida (2013). De esta manera, la construcción del modelo predictivo no solo permite obtener predicciones del rendimiento académico, además, permite identificar los factores que más influyen en la aprobación y reprobación de estudiantes específicos. Esto es de interés para profesores e instituciones educativas debido a que permite tomar las medidas necesarias con los estudiantes para evitar la reprobación y deserción escolar.

En la figura 3, se observa que la exactitud estimada del modelo (61.4754 %) es menor a la exactitud obtenida al aplicar el modelo predictivo en nuevos datos (70.4225 %), lo que significa que los nuevos datos tuvieron un comportamiento ligeramente diferente al de los datos de entrenamiento.

Se debe notar que pocos trabajos instrumentan una plataforma para automatizar la predicción del rendimiento académico como el de Rico y Sánchez (2018), quienes utilizaron cinco atributos estudiantiles consiguiendo una exactitud de 62.3 % a partir de 94 estudiantes que participaron en el estudio. A diferencia de este trabajo, en esta investigación se instrumentó una plataforma automática de predicción que utiliza nueve atributos estudiantiles y se obtuvo una exactitud de 70.42 por ciento.

Finalmente, se debe resaltar que el modelo construido permite a los profesores identificar desde el inicio de sus cursos qué factores influyen en el rendimiento académico y qué estudiantes tienen mayor probabilidad de aprobar y reprobar. Así, los profesores tienen la oportunidad de diseñar estrategias de prevención y disminuir las estrategias de recuperación que impliquen que el estudiante repruebe alguna evaluación parcial, para que ellos puedan realizar algún tipo de intervención. Las estrategias de recuperación son una práctica frecuente en la mayoría de las instituciones educativas y es necesario cambiarlas por estrategias de prevención para mejorar los procesos de enseñanza y de aprendizaje.

Referencias

Álvarez, L., V. Gugelmeier y L. Hermida (2013). "¿Cómo aprenden los estudiantes de odontología que cursan el último año de la carrera?". *Odontoestomatología,* 15(21), 4-11. http://www.scielo.edu.uy/pdf/ode/v15n21/v15n21a02.pdf

Ballesteros, A. y D. Sánchez, (2013). "Minería de datos educativa: Una herramienta para la investigación de patrones de aprendizaje sobre un contexto educativo". *Revista Latinoamericana de Física Educativa, 7*(4), 662-668. http://www.lajpe.org/dec13/22-LAJPE_814_bis_Alejandro_Ballesteros.pdf

Estrada, R. I., R. A. Zamarripa, P. G. Zúñiga e I. Martínez (2016). "Aportaciones desde la minería de datos al proceso de captación de matrícula en instituciones de educación superior particulares". *Revista Electrónica Educare,* 20(3), 1-21. doi:10.15359/ree.20-3.11

Han, J. (2012). *Data Mining: Concepts and Techniques*. Waltham: Morgan Kaufmann Publishers.

Hernández, J., M. Ramírez y C. Ferri (2004). *Introducción a la minería de datos.* Madrid: Pearson.

Jaramillo, A. y H. Paz (2015). "Aplicación de técnicas de minería de datos para determinar las interacciones de los estudiantes en un entorno virtual de aprendizaje. *Revista Tecnológica ESPOL,* 28(1), 64-90. http://www.rte.espol.edu.ec/index.php/tecnologica/article/view/351/229

Kavipriya, P. (2016). "A Review on Predicting Students' Academic Performance Earlier, Using Data Mining Techniques". *International Journal of Advanced Research in Computer Science and Software Engineering,* 6(12), 101-105. http://ijarcsse.com/Before_August_2017/docs/papers/Volume_6/12_December2016/V6I12-0129.pdf

Kotsiantis, S. B., C. J. Pierrakeas y P. E. Pintelas (2003). "Preventing Student Dropout in Distance Learning Using Machine Learning Techniques". En V. Palade, R. J. Howlett y L. Jain (eds.). *Lecture Notes in Computer Science, vol. 2,774. Knowledge-Based Intelligent Information and Engineering Systems,* 267–274. Heidelberg, Alemania: Springer-Verlag. doi:10.1007/978-3-540-45226-3_37

Luan, J. (2002). "Data Mining and its Applications in Higher Education". *New Directions for Institutional Research*, 113, 17-36. doi:10.1002/ir.35

Márquez, C., C. Romero y S. Ventura (2012). "Predicción del fracaso escolar mediante técnicas de minería de datos". *IEEE-RITA*, 7(3), 109-117. http://rita.det.uvigo.es/201208/uploads/IEEE-RITA.2012.V7.N3.A1.pdf

Michie, D., D. Spiegelhalter y C. Taylor (1994). *Machine Learning, Neural and Statistical Classification*. Nueva Jersey: Prentice Hall.

Mueen, A., B. Zafar y U. Manzoor (2016). "Modeling and Predicting Students' Academic Performance Using Data Mining Techniques". *International Journal of Modern Education and Computer Science,* 11, 36-42. doi:10.5815/ijmecs.2016.11.05

Pacheco, A. y Y. Fernández (2015). "Aplicación de técnicas de descubrimiento de conocimientos en el proceso de caracterización estudiantil". *Ciencias de la Información,* 46(3), 25-30. http://www.redalyc.org/articulo.oa?id=181443340004

Peña, A. (2014). "Review: Educational Data Mining: A Survey and a Data Mining Based Analysis of Recent Works". *Expert Systems with Applications,* 41(4),1432-1462. doi:10.1016/j.eswa.2013.08.042

Reynoso, O. y T. E. Méndez (2018). "¿Es posible predecir el rendimiento académico? La regulación de la conducta como un indicador del rendimiento académico en estudiantes de educación superior". *Diálogos sobre Educación,* 9(16). http://dialogossobreeducacion.cucsh.udg.mx/index.php/DSE/article/view/397

Rico, A. y D. Sánchez (2018a). "Diseño de un modelo para automatizar la predicción del rendimiento académico en estudiantes del IPN". *Revista Iberoamericana para la Investigación y el Desarrollo Educativo, 8*(16), 246-266. doi:10.23913/ride.v8i16.340

Rodallegas, E., A. Torres, B. Gaona, E. Gastelloú, R. Lezama y S. Valero (2010). "Modelo predictivo para la determinación de causas de reprobación escolar mediante minería de datos". En M. E. Prieto, J. M. Dodero y D. O. Villegas (Eds.). *Lecture Notes in Computer Science: vol. Kaambal. Recursos digitales para la educación y la cultura,* 48-55. Mérida, México. http://ccita2011.its-motul.edu.mx/documentos/Recursos_digitales.pdf

Romero, C. y S. Ventura (2010). "Educational Data Mining: A Review of the State of the art". *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* 40(6), 601-618. doi:10.1109/TSMCC.2010.2053532

_____ (2012). "Data Mining in Education". *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3*(1), 12-27. doi:10.1002/widm.1075

Shahiri, A., W. Husain y N. Rashid (2015). "A Review on Predicting Student's Performance Using Data Mining Techniques". *Procedia Computer Science*, 72, 414-422. doi:10.1016/j.procs.2015.12.157

Valero, S., A. Salvador y M. García (2010). "Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los vecinos más cercanos". En M. E. Prieto, J. M. Dodero y D. O. Villegas (eds.). *Lecture Notes in Computer Science: vol. Kaambal. Recursos digitales para la educación y la cultura,* 33-39. Mérida, México. http://www.utim.edu.mx/~svalero/docs/e1.pdf

Vera, J. A., D. Y. Ramos, M. A. Sotelo, S. Echeverría y D. M. Serrano (2012). "Factores asociados al rezago en estudiantes de una institución de educación superior en México". *Revista Iberoamericana de Educación Superior,* 3(7), 41–56. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S2007-28722012000200003

Witten, I., E. Frank y M. Hall (2005). *Data Mining: Practical Machine Learning Tools and Techniques.* Massachusetts: Morgan Kaufmann Publishers.