Revista Iberoamericana para la Investigación y el Desarrollo Educativo
ISSN 2007 - 7467

Diseño de un modelo para automatizar la predicción del rendimiento académico en estudiantes del IPN

Design of a model to automate the prediction of academic performance in students of IPN

Projeto de modelo para automatizar a previsão do desempenho acadêmico em estudantes do IPN

Andrés Rico Páez
Instituto Politécnico Nacional, México
aricop.ipn@gmail.com
ORCID ID: 0000-0002-6450-318X

Daniel Sánchez Guzmán
Instituto Politécnico Nacional, México
dsanchez@ipn.mx
ORCID ID: 0000-0001-9322-2734

Resumen

La minería de datos educativa permite extraer conocimiento útil y comprensible a partir de datos académicos para la solución de problemas acerca de diversos procesos de enseñanza y de aprendizaje. Una de las aplicaciones más populares de la minería de datos educativa es la predicción del rendimiento académico. El principal objetivo de este trabajo fue diseñar y automatizar un modelo predictivo del rendimiento académico de estudiantes del Instituto Politécnico Nacional (IPN).

Para la construcción del modelo, se analizaron las calificaciones de actividades académicas y la calificación final de 94 estudiantes inscritos en una carrera de ingeniería perteneciente al IPN. Este modelo se aplicó a 86 estudiantes para predecir su rendimiento académico. Posteriormente, se compararon estas predicciones con los resultados reales obtenidos por los estudiantes al final del curso. Se obtuvieron exactitudes de las predicciones de la aprobación

del curso de hasta 73%, únicamente con cinco atributos correspondientes a las calificaciones de las actividades académicas iniciales del mismo. Además, se construyó una plataforma que facilita la implementación del modelo para predecir automáticamente el desempeño académico de nuevos estudiantes. También se identificaron las principales actividades académicas que influyen en el desempeño académico a través del valor de las probabilidades del modelo. En particular, los resultados muestran que las actividades 3, 4 y 5 fueron las que influyeron de manera más significativa en la predicción de aprobación de los estudiantes que participaron en este estudio. El desarrollo de este tipo de modelos permite a las instituciones educativas predecir el rendimiento académico de sus estudiantes e identificar los principales factores que influyen en él.

Palabras clave: algoritmo Naïve Bayes, minería de datos, modelo predictivo, probabilidades, rendimiento académico.

Abstract

Educational data mining allows extracting useful and understandable knowledge from academic data to solve problems about various teaching and learning processes. One of the most popular applications of educational data mining is the prediction of academic performance. The main objective of this work was to design and automate a predictive model of the academic performance of students of the National Polytechnic Institute (IPN).

For the construction of the model, the qualifications of five academic activities and the final grade of 94 students enrolled in an Engineering career belonging to the IPN were analyzed. This model was applied to 86 students to predict their academic performance. Subsequently, these predictions were compared with the actual results obtained by the students at the end of the course. Accuracy was obtained from the predictions of the course approval of up to 73% and only with five attributes corresponding to the qualifications of the initial academic activities. In addition, a platform was built that facilitates the construction and use of the model to automatically predict the academic performance of new students. Also, the main academic activities that influenced academic performance were identified through the value of the probabilities of the model. In particular, the results showed that activities 3, 4 and 5 were those that most significantly influenced the prediction of approval of the students who

participated in this study. The development of this type of models allows educational institutions to predict the academic performance of their students and identify the main factors that influence it.

Keywords: Naïve Bayes algorithm, data mining, predictive model, , probabilities, academic performance.

Resumo

A mineração de dados educacionais permite extrair conhecimento útil e compreensível de dados acadêmicos para resolver problemas sobre vários processos de ensino e aprendizagem. Uma das aplicações mais populares da mineração de dados educacionais é a previsão do desempenho acadêmico. O objetivo principal deste trabalho foi projetar e automatizar um modelo preditivo de desempenho acadêmico dos estudantes do Instituto Nacional Politécnico (IPN).

Para a construção do modelo, foram analisados os graus de atividades acadêmicas e a nota final de 94 alunos matriculados em uma carreira de engenharia pertencente ao IPN. Este modelo foi aplicado a 86 estudantes para prever seu desempenho acadêmico. Posteriormente, essas previsões foram comparadas com os resultados reais obtidos pelos alunos no final do curso. A precisão foi obtida a partir das previsões da aprovação do curso de até 73%, com apenas cinco atributos correspondentes aos graus das atividades acadêmicas iniciais. Além disso, foi criada uma plataforma para facilitar a implementação do modelo para prever automaticamente o desempenho acadêmico de novos alunos. As principais atividades acadêmicas que influenciam o desempenho acadêmico também foram identificadas através do valor das probabilidades do modelo. Em particular, os resultados mostram que as atividades 3, 4 e 5 foram as que mais influenciaram significativamente a previsão de aprovação dos alunos que participaram deste estudo. O desenvolvimento deste tipo de modelos permite que as instituições educacionais prevejam o desempenho acadêmico de seus alunos e identifiquem os principais fatores que a influenciam.

Palavras-chave: algoritmo Naïve Bayes, mineração de dados, modelo preditivo, probabilidades, desempenho acadêmico.

Revista Iberoamericana para la Investigación y el Desarrollo Educativo
ISSN 2007 - 7467

Fecha Recepción: Marzo 2017 **Fecha Aceptación:** Agosto 2017

Introducción

Antecedentes

Las tecnologías de la información y comunicación (TIC) han tenido un rápido crecimiento en los últimos años debido a las diversas aplicaciones que se han generado en un gran número de sectores de la actividad humana, tales como el Internet, las bases de datos, la telefonía celular, entre muchas otras, de manera que han permitido desarrollar lo que se conoce como "sociedad de la información". Este desarrollo tecnológico ha originado un incremento en la cantidad de información a almacenar. La mayoría de esta información se genera con propósitos específicos y, posteriormente, no se analiza, aunque pudiera contener algún tipo de información oculta y potencialmente útil. Esto se debe, en la mayoría de los casos, al desconocimiento de cómo analizarla para extraer algún tipo de conocimiento. El análisis de información con herramientas estadísticas clásicas es una tarea bastante compleja, lo cual ha motivado al empleo de técnicas de minería de datos para este tipo de problemáticas, principalmente, en áreas del tipo empresarial o comercial (Han, 2012). La minería de datos es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde datos almacenados (Hernández, Ramírez y Ferri, 2004; Witten, Frank y Hall, 2005). Dicho proceso de análisis trabaja a nivel de conocimiento con el propósito de encontrar patrones y relaciones, así como también modelos predictivos que proporcionen patrones de conocimiento para la toma de decisiones. La minería de datos utiliza diversos métodos como la inteligencia artificial, la computación gráfica o el procesamiento masivo de conjuntos de información y como materia prima las bases de datos (Han, 2012).

La minería de datos, aplicada a la educación o minería de datos educativa, surge como un paradigma orientado al diseño, tareas, métodos y algoritmos con el objetivo de explorar los datos del ambiente educativo (Peña, 2014). La minería de datos educativa tiene como propósito descubrir conocimiento y patrones dentro de datos de estudiantes (Luan, 2002).

Estos patrones caracterizan el comportamiento de los estudiantes con base en sus logros, evaluaciones y dominio del contenido de conocimiento (Ballesteros y Sánchez, 2013).

Por lo anterior, existe una tendencia al uso de minería de datos en el área de la educación (Romero y Ventura, 2010, 2012; Peña, 2014). No obstante, esta aplicación de la minería de datos es reciente en países de Latinoamérica (Estrada, Zamarripa, Zúñiga y Martínez, 2016), por lo que existen varios problemas abiertos en el uso y desarrollo de este tipo de técnicas.

Objetivo de la investigación

En la actualidad, las principales problemáticas de las instituciones educativas son los altos índices de reprobación y de deserción escolar (Vera, Ramos, Sotelo, Echeverría y Serrano, 2012; Martínez, Hernández, Carillo, Romualdo y Hernández, 2013). En el caso de México, la Organización para la Cooperación y el Desarrollo Económicos (OCDE) señala que existe un problema de deserción debido a que ocupa el primer lugar, entre los 35 países miembros, en el número de desertores escolares. Uno de los principales factores de deserción escolar es el bajo rendimiento académico obtenido por los estudiantes en alguna o varias asignaturas, las cuales tienden a reprobar después de agotar las oportunidades de aprobación en periodos ordinarios y extraordinarios, situación que conduce al abandono de su preparación.

Para reducir estos graves y complejos problemas de deserción de estudiantes en las instituciones educativas, se han aplicado técnicas de minería de datos con éxito para crear modelos predictivos del rendimiento académico (Xing, Guo, Petakovic y Goggins, 2015). Los resultados obtenidos con este tipo de técnicas han sido prometedores y demuestran cómo algunos factores o características de los estudiantes pueden afectar el rendimiento académico (Márquez, Romero y Ventura, 2012). Sin embargo, en el entorno educativo de México, las técnicas de minería de datos aplicadas a la creación de modelos de predicción de rendimiento académico todavía se encuentran poco desarrolladas.

En las instituciones educativas, como el Instituto Politécnico Nacional (IPN), se requiere del diseño y aplicación de modelos predictivos debido a que ofrece la posibilidad de proponer programas de prevención estratégicos para estudiantes con bajo rendimiento,

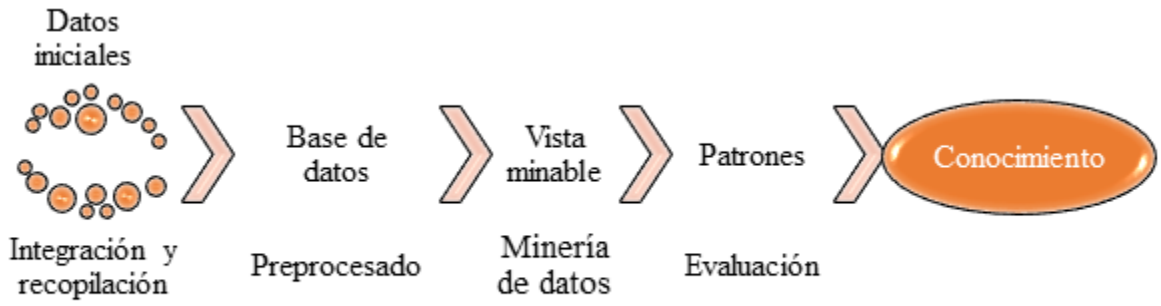
detectar estudiantes en peligro de deserción e identificar características de los estudiantes que le permiten obtener un buen desempeño académico, entre muchos otros beneficios potenciales.

La presente investigación se realizó para contestar las siguientes preguntas: ¿Cómo diseñar y automatizar un modelo predictivo del rendimiento académico y cuál es su exactitud en las predicciones de estudiantes del IPN? ¿Cómo identificar, a partir del modelo, los principales factores académicos que influyen de manera más significativa en el rendimiento académico de los estudiantes que participaron en este estudio? Por lo tanto, el objetivo de esta investigación fue diseñar y automatizar un modelo predictivo del rendimiento académico de estudiantes del IPN y evaluarlo con respecto a la exactitud de las predicciones, además de identificar los principales factores académicos que inciden en el desempeño académico de los estudiantes a partir de este modelo. La base metodológica y la técnica de minería de datos utilizada para este objetivo se presentó en la siguiente sección.

Descubrimiento de conocimiento en bases de datos y algoritmo Naïve Bayes

La metodología empleada en la presente investigación se basó en el proceso completo de aplicación de técnicas de minería de datos conocido como descubrimiento de conocimiento en bases de datos (Espinosa, Farías y Verduzco, 2016), que coloca a la minería de datos como una de las fases del mismo. Este proceso se muestra 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 (Espinosa *et al.*, 2016).

La primera fase del proceso de descubrimiento de conocimiento en bases de datos es la integración y recopilación de datos, en las que se determinan las fuentes de información y la manera de conseguirlas para formar la base de datos a utilizar. La siguiente fase es conocida como preprocesamiento, que consiste en la selección, limpieza y transformación de los datos para formar el subconjunto de datos que se va a minar o la vista minable. Posteriormente, se encuentra la fase de minería de datos, en la que se define el tipo de tarea a realizar y el algoritmo a implementar. Finalmente, está la fase de evaluación, en donde se determina la validez y confiabilidad del conocimiento extraído.

De los tipos de tareas de la minería de datos, las predictivas son de las más populares y de uso más extendido en la minería de datos educativa (Romero y Ventura, 2010, 2012; Peña, 2014) debido a que permite detectar problemas académicos con anticipación y aplicar las medidas necesarias.

Dentro de las tareas predictivas de la minería de datos se encuentra la clasificación, que consiste en etiquetar cada registro o instancia de una base de datos de entrenamiento como parte de una clase representada mediante el valor de un atributo llamado atributo clasificador o clase de la instancia. Los demás atributos se utilizan para predecir la clase. El objetivo es predecir la clase de nuevas instancias (datos de prueba) de la que se desconoce la clase. De esta manera, existe un conjunto de atributos $\{A_1,\ldots,A_n\}$ 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:

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

En la clasificación, 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*. De esta forma, el valor de la clase más probable es:

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

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. La exactitud de la clasificación (porcentaje de registros clasificados correctamente entre el total de registros clasificados) 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). Debido a esto, el algoritmo Naïve Bayes es la técnica de minería de datos utilizada en esta investigación.

Con base en esta suposición de independencia, el valor de la clase a devolver es:

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

La clasificación con este algoritmo consta de dos partes. La primera es la construcción del modelo y la segunda es la evaluación del modelo a partir de la clasificación de los nuevos datos.

Para la construcción del modelo, se estiman las probabilidades *a priori* y *a posteriori*. Las probabilidades *a priori* $P(C_i)$ se estiman dividiendo el número instancias 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 puede calcular a partir de la frecuencia de aparición en la base de datos de entrenamiento por medio del número de casos favorables entre el número de casos totales. En este trabajo, para solucionar el caso en el que $P(A_j|C_i)=0$, se utiliza la estimación basada en la ley de sucesión de Laplace (Hernández *et al.*, 2004), que consiste en obtener el número de casos favorables más uno dividido entre el número de casos totales más el número de valores posibles.

Para la evaluación del modelo se utilizan las probabilidades *a priori* y *a posteriori* para clasificar un nuevo registro, se determinan las probabilidades de los atributos de dicho registro y se aplica la fórmula (3) para determinar a qué clase corresponde.

Metodología

Integración y recopilación

La fuente de datos de entrenamiento fueron las calificaciones de las primeras cinco actividades académicas y la calificación final de un curso de Ecuaciones Diferenciales de estudiantes inscritos en una carrera de ingeniería perteneciente al IPN. Fueron datos de cinco grupos de estudiantes formando un total de 94 registros. Con cantidades similares de registros en Kotsiantis *et al.* (2003) y Mueen, Zafar y Manzoor (2016), se observó que el algoritmo Naïve Bayes proporcionaba mejor rendimiento en cuanto a exactitud de las predicciones, en comparación con otras técnicas de minería de datos.

Preprocesamiento

Las calificaciones de las primeras cinco actividades del curso se representaron con los atributos act1, act2, act3, act4 y act5. Posteriormente, estos valores se definieron como Aprobada, “A”, (6.0-10.0); Reprobada, “R”, (0.0-5.9); y No Presento, “NP”. La calificación final es el atributo clasificador definido como “aprueba” y puede tener los valores de “SÍ” o “NO”. En la tabla 1 se presentan los valores posibles de estos atributos.

Tabla 1. Valores posibles de los atributos.

Atributos	Valores posibles
act1, act2, act3, act4, act5	A (Aprobada), R (Reprobada), NP (No Presento)
aprueba	SÍ, NO

Fuente: elaboración propia.

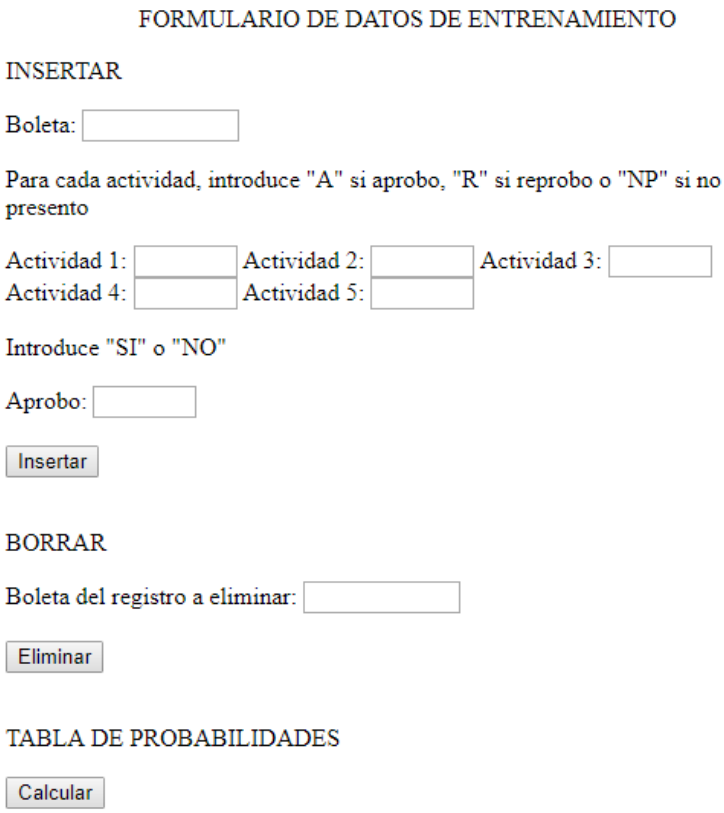
Fase de minería de datos

En este trabajo, la tarea predictiva empleada fue la clasificación y la técnica utilizada fue el algoritmo de Naïve Bayes. El modelo predictivo se construyó por medio del cálculo de probabilidades *a priori* y *a posteriori* de los atributos descrito en las secciones anteriores. Para esto, existen herramientas informáticas que ayudan a obtener modelos de predicción, tal y como se realizó en Jaramillo y Paz (2015) y Pacheco y Fernández (2015). Sin embargo, la mayoría de estas herramientas necesitan de usuarios expertos en el área para poder usarlas de manera adecuada. A diferencia de estos trabajos y 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 objetivo de publicarla en un futuro en un sitio web como un apoyo a profesores. El principal costo fue la renta de un servidor de Internet, sin embargo, existen servidores que permiten alojar gratuitamente bases de datos que no sean muy grandes, como las empleadas en este trabajo. Opcionalmente, si la cantidad de datos a utilizar es más grande se puede rentar un servidor que, dependiendo de los beneficios y cantidad de datos a almacenar, su precio puede variar entre $50 (cincuenta pesos 00/100 M.N.) y $1,500 (mil quinientos pesos 00/100 M.N.) mensuales.

La plataforma desarrollada permite introducir las actividades académicas de un número variable de estudiantes y de cualquier área. De esta manera, dicha plataforma no solo se puede utilizar en la institución del IPN en la que se recabaron los datos, sino en cualquier institución educativa. De esta manera, los profesores que no tengan conocimientos profundos en minería de datos pueden realizar predicciones del rendimiento académico de sus estudiantes por medio de sus actividades académicas. Esta plataforma ofrece la posibilidad de introducir los valores de las actividades académicas de los estudiantes como entrenamiento y calcular automáticamente las probabilidades para construir el modelo predictivo. La interfaz gráfica para introducir los datos de entrenamiento se muestra en la figura 2.

Revista Iberoamericana para la Investigación y el Desarrollo Educativo
ISSN 2007 - 7467

Figura 2. Interfaz gráfica para introducir los datos de entrenamiento.



Fuente: elaboración propia.

Por medio de la plataforma implementada se calcularon las probabilidades *a priori* y *a posteriori* de los atributos, las cuales se muestran en la tabla 2.

Revista Iberoamericana para la Investigación y el Desarrollo Educativo
ISSN 2007 - 7467

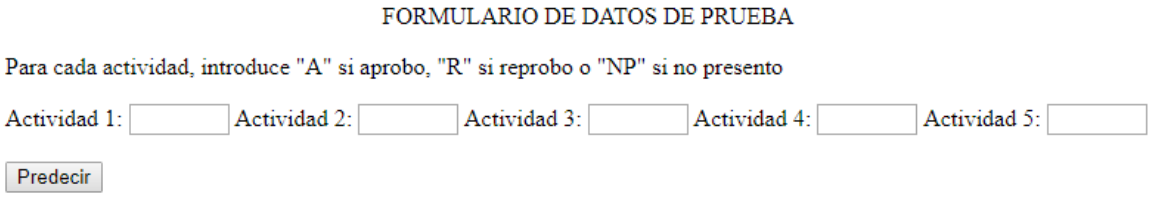
Tabla 2. Probabilidades estimadas de los datos de entrenamiento.

Atributos	**Probabilidades *a posteriori***					
	P(A/SI)	P(R/SI)	P(NP/SI)	P(A/NO)	P(R/NO)	P(NP/NO)
act1	0.3673	0.4898	0.1428	0.2745	0.3725	0.3529
act2	0.4898	0.2857	0.2245	0.5294	0.1568	0.3137
act3	0.6122	0.3061	0.0816	0.4313	0.2941	0.2745
act4	0.6530	0.1836	0.1632	0.3725	0.2549	0.3725
act5	0.6734	0.2040	0.1224	0.4313	0.1568	0.4117
	Probabilidades *a priori*					
	P(aprueba=SI)			P(aprueba=NO)		
Aprueba	0.4894			0.5106		

Fuente: elaboración propia.

A partir de estas probabilidades, se puede predecir la aprobación de un nuevo estudiante (datos de prueba) aplicando la fórmula (3). Esto se puede realizar de forma automática a través de la plataforma construida. La interfaz gráfica para introducir los datos de prueba se presenta en la figura 3.

Figura 3. Interfaz gráfica para introducir los datos de prueba. Fuente: elaboración propia.



Fuente: elaboración propia.

Evaluación

La evaluación del modelo predictivo construido se realiza por medio del cálculo de la exactitud de las predicciones correctas. Para esto, se utilizó el método conocido como validación cruzada (Hernández *et al.*, 2004). Este método consiste en dividir aleatoriamente los datos de entrenamiento en un número fijo de grupos. En este caso, se dividió en dos grupos de datos equitativos. Después, se construyó un modelo con el primer conjunto que se usó para predecir los resultados en el segundo conjunto y se calculó su exactitud. Posteriormente, se construyó un modelo con el segundo conjunto que se usó para predecir los resultados del primer conjunto y se calculó la exactitud. Finalmente, se calculó la exactitud del modelo construido promediando las exactitudes calculadas anteriormente.

De esta manera, se dividieron los 94 registros de los cinco grupos de estudiantes en dos conjuntos de 47 registros. Cada conjunto se construyó de forma aleatoria, procurando que tuvieran cantidades similares de muestras de los cinco grupos de estudiantes. La exactitud del primer conjunto, obtenida con datos de entrenamiento del segundo conjunto, fue de 59.57% y la del segundo conjunto, obtenida con datos de entrenamiento del primer conjunto, fue de 68.09%. Por lo tanto, la exactitud del modelo construido fue de 63.83%.

Se debe tener en cuenta que la exactitud obtenida de la evaluación de un modelo no garantiza que se refleje en el mundo real. Únicamente indica que, si los nuevos datos a predecir tienen un comportamiento similar a los datos de entrenamiento, entonces la exactitud será similar a la del modelo.

Revista Iberoamericana para la Investigación y el Desarrollo Educativo
ISSN 2007 - 7467

Resultados y discusión

El modelo predictivo construido se aplicó a cuatro grupos de estudiantes universitarios del IPN inscritos en el curso de Ecuaciones Diferenciales del semestre siguiente al que se habían obtenido los datos de entrenamiento. Fueron un total de 86 registros de datos de prueba.

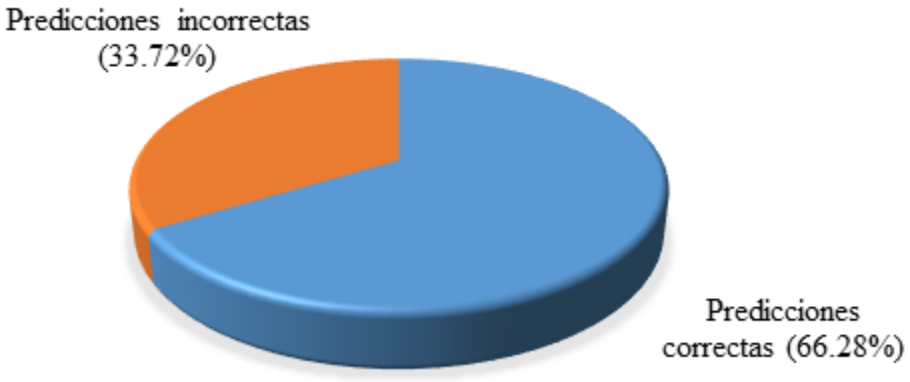
Primeramente, se utilizó el modelo predictivo para obtener predicciones de los cuatro grupos de prueba. Posteriormente, se compararon estas predicciones con los resultados reales obtenidos por los estudiantes al final del curso. En la tabla 3 se presenta el número de predicciones correctas e incorrectas y la exactitud de las predicciones de cada uno de los grupos de prueba. La exactitud total de todos los registros de prueba se muestra en la figura 4.

Tabla 3. Exactitud de las predicciones de los grupos de prueba.

Grupos de prueba	Cantidad de estudiantes	Predicciones correctas	Predicciones incorrectas	Exactitud
1	28	17	11	60.71 %
2	15	11	4	73.33 %
3	26	19	7	73.08 %
4	17	10	7	58.82 %

Fuente: elaboración propia.

Figura 4. Exactitud total de las predicciones de los grupos de prueba.



Fuente: elaboración propia.

Los resultados obtenidos muestran cómo a partir de algunas actividades académicas iniciales del curso se puede 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. Para automatizar el modelo predictivo, se construyó una plataforma que permitió ingresar los datos de los estudiantes para construir el modelo y, posteriormente, introducir los datos de estudiantes de los cuales se va a predecir su rendimiento académico.

Una vez construido el modelo, se pueden identificar los principales factores (actividades académicas) que influyen de manera más significativa en el rendimiento académico de los estudiantes que participaron en este estudio por medio del cálculo de las probabilidades *a posteriori*. Lo anterior debido a que las probabilidades que tengan mayor valor influirán de forma significativa en la decisión de aprobación o reprobación del estudiante de acuerdo con el modelo desarrollado con el algoritmo Naïve Bayes.

En la tabla 2, se puede observar que las probabilidades *a priori* para el atributo "aprueba" son similares, sin embargo, la probabilidad P(aprueba=NO) resulta mayor, debido a que, en los datos de entrenamiento, existe una mayor cantidad de estudiantes reprobados que aprobados, lo cual resulta típico en cursos universitarios de Matemáticas. Las

probabilidades *a posteriori* P(A/SI) de las actividades 3, 4 y 5 fueron las que tuvieron un valor mayor. Por lo tanto, estas actividades académicas son las que más influyeron en la predicción de aprobación de los estudiantes que participaron en este estudio. De esta manera, los estudiantes que aprueban las actividades 3, 4 y 5 tienen una mayor probabilidad de aprobar el curso. Esto no ocurre con las actividades 1 y 2, lo cual puede atribuirse a diversos factores, por ejemplo, que los estudiantes en estas primeras actividades apenas se están adaptando al curso o que no entraron desde el principio del curso por retrasos administrativos en su inscripción.

En la tabla 3 se muestra que la exactitud de las predicciones es más alta en unos grupos de prueba que en otros debido a diversos factores del estudiante no tomados en cuenta en el modelo construido, por ejemplo, promedio actual, cantidad de materias reprobadas o escolaridad de los padres, entre otros.

La exactitud total de todos los datos de prueba (66.28%) fue muy parecida a la estimada en la validación cruzada (63.83%), por lo que los datos de prueba tuvieron un comportamiento similar a los datos de entrenamiento.

En trabajos que han utilizado el algoritmo Naïve Bayes, se han obtenido valores de exactitud parecidos. En Jishan, Rashu, Haque y Rahman (2015), participaron 181 estudiantes; el valor más alto de exactitud se obtuvo utilizando seis atributos y fue de 75%. En Mueen *et al.* (2016) participaron 60 estudiantes y el valor más alto de exactitud se obtuvo con 38 atributos y fue de 86%. En este trabajo, se obtuvieron exactitudes de las predicciones de la aprobación del curso de hasta 73%, únicamente con cinco atributos correspondientes a las calificaciones de las actividades académicas iniciales del mismo. También se identificaron los principales factores que influyeron en el rendimiento académico de la muestra de datos analizados, similar a como se hizo en Mueen *et al.* (2016). Además, a diferencia de los trabajos mencionados, se construyó una plataforma que permite la automatización de las predicciones del rendimiento académico para facilitar su uso por profesores de instituciones educativas.

Este tipo de modelos ofrece la posibilidad a las instituciones educativas de diseñar estrategias de prevención de reprobación e identificar los factores más relevantes que influyen en el rendimiento académico de sus estudiantes.

Conclusiones

El objetivo de este trabajo fue el diseño y automatización de un modelo predictivo del rendimiento académico de estudiantes del IPN. Este modelo se construyó por medio del algoritmo Naïve Bayes y se automatizó utilizando lenguajes de programación adecuados para su posterior publicación en un sitio web y, de este modo, se vuelva accesible a cualquier tipo de profesor y no solo a expertos en el área de minería de datos.

El modelo se evalúo con respecto a la exactitud de las predicciones, obteniendo valores de hasta 73%, teniendo en cuenta que solo se utilizaron pocas actividades iniciales realizadas por los estudiantes.

En la construcción del modelo predictivo, se calcularon las probabilidades *a priori* y *a posteriori* de los atributos. Por medio de estas, se identificaron las principales actividades que inciden en el rendimiento académico del conjunto de datos académicos analizados. De esta manera, el modelo construido permite obtener predicciones del rendimiento académico e identificar las principales actividades académicas que inciden en él.

Las evaluaciones de actividades iniciales de un curso son una práctica habitual por la mayoría de los profesores de las instituciones educativas. De esta manera, la metodología realizada puede ser replicada por los estos para construir modelos predictivos para sus propios estudiantes, y, así, tener la oportunidad de diseñar estrategias de prevención y disminuir las estrategias de recuperación que impliquen que el alumno repruebe alguna evaluación parcial para realizar algún tipo de intervención. Las estrategias de recuperación son una práctica frecuente en la mayoría de las instituciones educativas.

Referencias

Ballesteros, A., y Sánchez, D. (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. Recuperado de http://www.lajpe.org/dec13/22-LAJPE_814_bis_Alejandro_Ballesteros.pdf

Espinosa, M., Farías, N., y Verduzco, J. A. (2016). Análisis de los Datos Históricos de la Programación de Cursos en los CECATI del Estado de Colima. *Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, *6*(12), 114-134. Recuperado de http://www.ride.org.mx/index.php/RIDE/article/view/192/842

Estrada, R. I., Zamarripa, R. A., Zúñiga, P. G., y Martínez I. (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

Jaramillo, A., y Paz H. (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. Recuperado de http://www.rte.espol.edu.ec/index.php/tecnologica/article/view/351/229

Jishan, S., Rashu, R., Haque, N., y Rahman, R. (2015). Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. *Decision Analytics*, *2*(1), 1-25. doi: 10.1186/s40165-014-0010-2

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

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

Kotsiantis, S. B., Pierrakeas, C. J., y Pintelas, P. E. (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. 2774. Knowledge-Based Intelligent Information and Engineering Systems* (pp. 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., Romero, C., y Ventura, S. (2012). Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos. *IEEE-RITA*, *7*(3), 109-117. Recuperado de http://rita.det.uvigo.es/201208/uploads/IEEE-RITA.2012.V7.N3.A1.pdf

Martínez, A., Hernández, L. I., Carillo, D., Romualdo, Z., y Hernández, C. P. (2013). Factores asociados a la reprobación estudiantil en la Universidad de la Sierra Sur, Oaxaca. *Temas de Ciencia y Tecnología*, *17*(51), 25-33. Recuperado de http://www.utm.mx/edi_anteriores/temas51/T51_1Ensayo3-FactAsocReprobacion.pdf

Michie, D., Spiegelhalter D., y Taylor, C. (1994). *Machine learning, neural and statistical classification*. Nueva Jersey, Estados Unidos: Prentice Hall.

Mueen, A., Zafar, B., y Manzoor U. (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 Fernández, Y. (2015). Aplicación de técnicas de descubrimiento de conocimientos en el proceso de caracterización estudiantil. *Ciencias de la Información*, *46*(3), 25-30. Recuperado de: 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

Romero, C., y Ventura, S. (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

Romero, C., y Ventura, S. (2012). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12-27. doi: 10.1002/widm.1075

Valero, S., Salvador, A., y García, M. (2010). Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k 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. (pp. 33-39). Mérida, México. Recuperado de http://www.utim.edu.mx/~svalero/docs/e1.pdf

Vera, J. A., Ramos, D. Y., Sotelo, M. A., Echeverría, S., y Serrano, D. M. (2012). Factores asociados al rezago en estudiantes de una institución de educación superior en México. *Revista Iberoamericana de Educación Superior*, *3*(7), 41–56. doi: 10.22201/iisue.20072872e.2012.7.81

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

Xing, W., Guo, R., Petakovic, E., y Goggins, S. (2015). Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. *Computers in Human Behavior, 47,* 168-181. doi: 10.1016/j.chb.2014.09.034

Rol de Contribución	Autor(es)
Conceptualización	Andrés Rico Páez «igual» Daniel Sánchez Guzmán «igual»
Metodología	Andrés Rico Páez «principal» Daniel Sánchez Guzmán «que apoya»
Software	Andrés Rico Páez «principal» Daniel Sánchez Guzmán «que apoya»
Validación	Andrés Rico Páez
Análisis Formal	Andrés Rico Páez «igual» Daniel Sánchez Guzmán «igual»
Investigación	Andrés Rico Páez
Recursos	Andrés Rico Páez
Curación de datos	Andrés Rico Páez
Escritura - Preparación del borrador original	Andrés Rico Páez «principal» Daniel Sánchez Guzmán «que apoya»
Escritura - Revisión y edición	Andrés Rico Páez «igual» Daniel Sánchez Guzmán «igual»
Visualización	Andrés Rico Páez
Supervisión	Andrés Rico Páez
Administración de Proyectos	Andrés Rico Páez «principal» Daniel Sánchez Guzmán «que apoya»
Adquisición de fondos	Andrés Rico Páez