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

Introduction

Background

Information and communication technologies (ICT) have experienced rapid growth in recent years due to the diverse applications that have been generated in a large number of sectors of human activity, such as the Internet, databases, cellular telephony, among many others, in a way that allowed to develop what is known as the "information society". This technological development has led to an increase in the amount of information to be stored. Most of this information is generated for specific purposes and, subsequently, it is not analyzed, although it may contain some type of hidden and potentially useful information. This is due, in most cases, to the ignorance of how to analyze it to extract some type of knowledge. The analysis of information with classical statistical tools is a rather complex task, which has motivated the use of data mining techniques for this type of problems, mainly in business or commercial areas (Han, 2012). Data mining is the process of extracting useful and understandable knowledge, previously unknown, from stored data (Hernández, Ramírez and Ferri, 2004, Witten, Frank and Hall, 2005). This process of analysis works at the level of knowledge with the purpose of finding patterns and relationships, as well as predictive models that provide knowledge patterns for decision making. Data mining uses various methods such as artificial intelligence, graphic computing or mass processing of information sets and as raw materials databases (Han, 2012).

Data mining, applied to education or educational data mining, emerges as a paradigm oriented to design, tasks, methods and algorithms with the aim of exploring the data of the educational environment (Peña, 2014). The purpose of educational data mining is to discover knowledge and patterns within student data (Luan, 2002). These patterns characterize student behavior based on their achievements, evaluations and mastery of knowledge content (Ballesteros and Sánchez, 2013).

Due to the above, there is a tendency to use data mining in the area of education (Romero and Ventura, 2010, 2012, Peña, 2014). However, this application of data mining is recent in Latin American countries (Estrada, Zamarripa, Zúñiga and Martínez, 2016), so there are several open problems in the use and development of this type of techniques.

Objective of the investigation

Currently, the main problems of educational institutions are high rates of failure and school dropout (Vera, Ramos, Sotelo, Echeverría and Serrano, 2012, Martinez, Hernandez, Carillo, Romualdo and Hernandez, 2013). In the case of Mexico, the Organization for Economic Cooperation and Development (OECD) points out that there is a problem of desertion because it ranks first among the 35 member countries in the number of school dropouts. One of the main factors of dropout is the low academic performance obtained by students in one or more subjects, which tend to fail after exhausting the opportunities for approval in ordinary and extraordinary periods, a situation that leads to the abandonment of their preparation.

To reduce these serious and complex problems of student desertion in educational institutions, successful data mining techniques have been applied to create predictive models of academic performance (Xing, Guo, Petakovic and Goggins, 2015). The results obtained with this type of techniques have been promising and show how some factors or characteristics of students can affect academic performance (Márquez, Romero and Ventura, 2012). However, in the educational environment of Mexico, data mining techniques applied to the creation of academic performance prediction models are still underdeveloped.

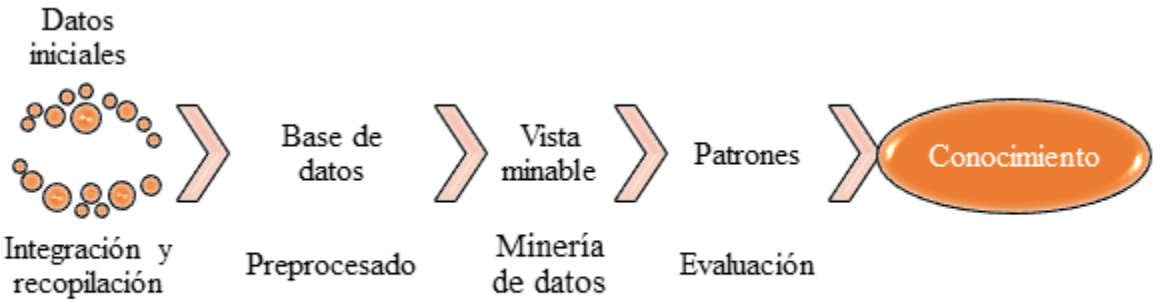
In educational institutions, such as the National Polytechnic Institute (IPN), the design and application of predictive models is required because it offers the possibility of proposing strategic prevention programs for underperforming students, detecting students at risk of dropping out and identifying characteristics. of students that allow you to obtain good academic performance, among many other potential benefits.

The present investigation was carried out to answer the following questions: How to design and automate a predictive model of academic performance and what is its accuracy in the predictions of IPN students? How to identify, from the model, the main academic factors that influence more significantly the academic performance of the students who participated in this study? Therefore, the objective of this research was to design and automate a predictive model of the academic performance of students of the IPN and evaluate it with respect to the accuracy of predictions, in addition to identifying the main academic factors that affect the academic performance of students. from this model. The methodological basis and data mining technique used for this purpose was presented in the following section.

Discovery of knowledge in databases and Naïve Bayes algorithm

The methodology used in the present investigation was based on the complete process of application of data mining techniques known as knowledge discovery in databases (Espinosa, Farías and Verduzco, 2016), which places data mining as one of the phases of it. This process is shown in figure 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).

The first phase of the knowledge discovery process in databases is the integration and collection of data, in which the sources of information are determined and the way to obtain them to form the database to be used. The next phase is known as preprocessing, which consists of the selection, cleaning and transformation of the data to form the subset of data

to be mined or view minable. Subsequently, there is the data mining phase, in which the type of task to be performed and the algorithm to be implemented are defined. Finally, there is the evaluation phase, where the validity and reliability of the extracted knowledge is determined.

Of the types of data mining tasks, the predictive ones are the most popular and most widely used in educational data mining (Romero and Ventura, 2010, 2012; Peña, 2014) because it allows the detection of academic problems with anticipation and apply the necessary measures.

Within the predictive tasks of data mining is the classification, which is to label each record or instance of a training database as part of a class represented by the value of an attribute called classifier attribute or class of the instance . The other attributes are used to predict the class. The objective is to predict the class of new instances (test data) of which the class is unknown. In this way, there is a set of attributes $\{A_1,\ldots,A_n\}$ and a class variable C_i, belonging to a set $\Omega_C=\{C_1,\ldots,C_k\}$. The a posteriori probability of the class variable C_i, given a set of attributes, it is calculated from Bayes' theorem in the following way:

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

In the classification, it is necessary to identify the most probable value and return it as a result. In Bayes' theorem, the most probable hypothesis is that with maximum a posteriori probability. In this way, the value of the most likely class is:

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

The algorithm known as Naïve Bayes (Hernández et al., 2004; Witten et al., 2005) assumes that all attributes are independent once the value of the class is known. The accuracy of the classification (percentage of records correctly classified among the total of classified records) with the Naïve Bayes algorithm is similar or superior to other data mining techniques (Michie, Spiegelhalter and Taylor, 1994, Kotsiantis, Pierrakeas and Pintelas , 2003). Because of this, the Naïve Bayes algorithm is the data mining technique used in this research.

Based on this assumption of independence, the value of the class to be returned is:

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

The classification with this algorithm consists of two parts. The first is the construction of the model and the second is the evaluation of the model based on the classification of the new data.

For the construction of the model, probabilities are estimated a priori and a posteriori. The a priori probabilities $P(C_i)$ are estimated by dividing the number instances of the class C_i of the training data among the total of them. The estimation of the posterior probabilities$P(A_j|C_i)$ of each discrete attribute can be calculated from the frequency of occurrence in the training database by means of the number of favorable cases among the number of total cases. In this work, to solve the case in which $P(A_j|C_i)=0$, the estimate based on the law of succession of Laplace is used (Hernández et al., 2004), which consists of obtaining the number of favorable cases plus one divided by the number of total cases plus the number of possible values.

For the evaluation of the model, the a priori and a posteriori probabilities are used to classify a new record, the probabilities of the attributes of said record are determined and the formula (3) is applied to determine which class corresponds.

Methodology

Integration and collection

The source of training data were the grades of the first five academic activities and the final grade of a Differential Equations course of students enrolled in an engineering career belonging to the IPN. They were data of five groups of students forming a total of 94 records. With similar amounts of records in Kotsiantis et al. (2003) and Mueen, Zafar and Manzoor (2016), it was observed that the Naïve Bayes algorithm provided better performance in prediction accuracy, compared to other data mining techniques.

Preprocessing

The grades of the first five activities of the course were represented with the attributes act1, act2, act3, act4 and act5. Subsequently, these values were defined as Approved, "A", (6.0-10.0); Failed, "R", (0.0-5.9); and No Presento, "NP". The final grade is the classifier attribute defined as "approves" and may have the values of "YES" or "NO". Table 1 shows the possible values of these attributes.

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.

Phase of data mining

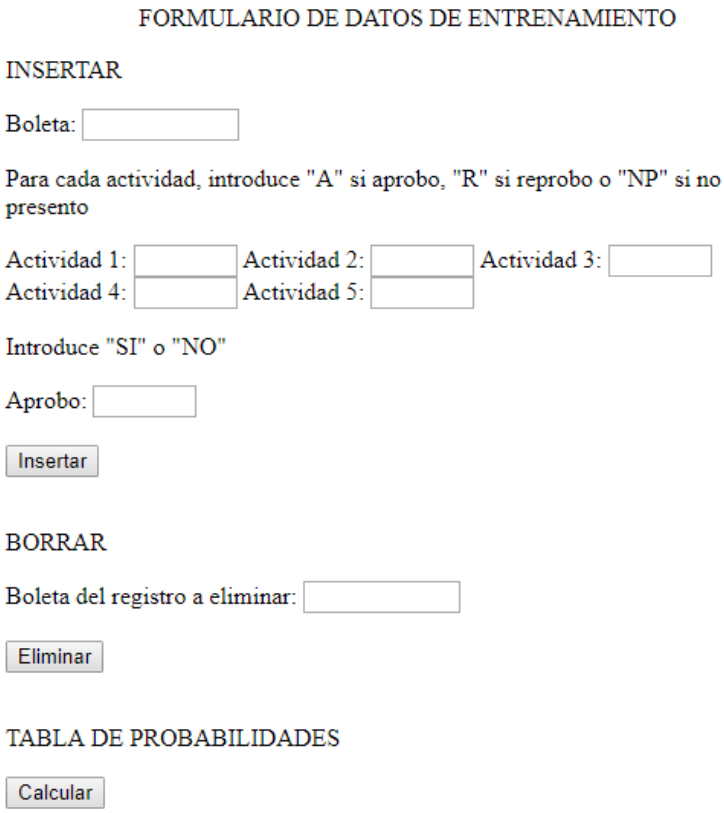
In this work, the predictive task used was the classification and the technique used was the Naïve Bayes algorithm. The predictive model was constructed by calculating the a priori and a posteriori probabilities of the attributes described in the previous sections. For this, there are computer tools that help to obtain prediction models, as was done in Jaramillo and Paz (2015) and Pacheco y Fernández (2015). However, most of these tools require expert users in the area to be able to use them properly. Unlike these works and similar to Valero, Salvador and García (2010), in this work a platform was developed in which the Naïve Bayes

algorithm was programmed in HTML5 (HyperText Markup Language, version 5) and PHP (Hypertext Pre-Processor) with the aim of publishing it in a future on a website as a support to teachers. The main cost was the rent of an Internet server, however, there are servers that allow free hosting of databases that are not very large, such as those used in this work. Optionally, if the amount of data to be used is larger, you can rent a server that, depending on the benefits and amount of data to store, its price can vary between $ 50 (fifty pesos 00/100 MN) and $ 1,500 (one thousand five hundred pesos) 00/100 MN) monthly.

The developed platform allows to introduce the academic activities of a variable number of students and of any area. In this way, this platform can not only be used in the IPN institution in which the data was collected, but in any educational institution. In this way, teachers who do not have deep knowledge in data mining can make predictions of their students' academic performance through their academic activities. This platform offers the possibility of introducing the values of the students' academic activities as training and automatically calculating the probabilities to build the predictive model. The graphical interface to enter the training data is shown in figure 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.

By means of the implemented platform, the a priori and a posteriori probabilities of the attributes were calculated, which are shown in table 2.

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

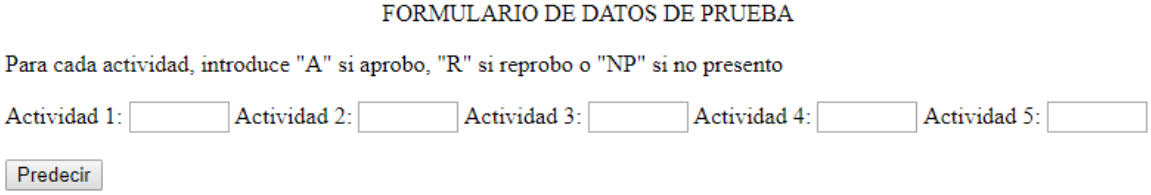
Tabla 2. Probabilidades estimadas de los datos de entrenamiento.

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.

From these probabilities, the approval of a new student (test data) can be predicted by applying formula (3). This can be done automatically through the built platform. The graphic interface to enter the test data is presented in Figure 3.

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



Fuente: elaboración propia.

Evaluation

The evaluation of the constructed predictive model is done by calculating the accuracy of the correct predictions. For this, the method known as cross-validation was used (Hernández et al., 2004). This method consists of randomly dividing the training data into a fixed number of groups. In this case, it was divided into two groups of equitable data. Then, a model was constructed with the first set that was used to predict the results in the second set and its accuracy was calculated. Subsequently, a model was constructed with the second set that was used to predict the results of the first set and the accuracy was calculated. Finally, the accuracy of the constructed model was calculated by averaging the accuracies calculated previously.

In this way, the 94 records of the five groups of students were divided into two sets of 47 records. Each set was constructed randomly, ensuring that they had similar amounts of samples from the five groups of students. The accuracy of the first set, obtained with training data of the second set, was 59.57% and that of the second set, obtained with training data of the first set, was 68.09%. Therefore, the accuracy of the built model was 63.83%.

It must be borne in mind that the accuracy obtained from the evaluation of a model does not guarantee that it is reflected in the real world. It only indicates that, if the new data to be predicted have a behavior similar to the training data, then the accuracy will be similar to that of the model.

Results and Discussion

The constructed predictive model was applied to four groups of university students of the IPN enrolled in the Differential Equations course of the following semester from which the training data had been obtained. There were a total of 86 test data records.

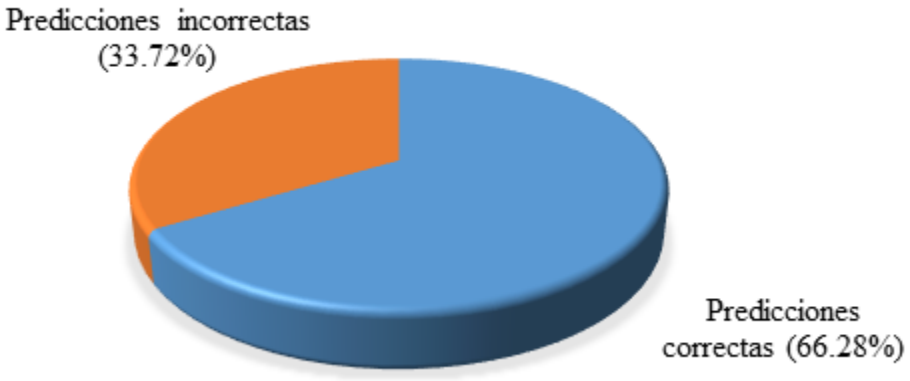
First, the predictive model was used to obtain predictions of the four test groups. Subsequently, these predictions were compared with the actual results obtained by the students at the end of the course. Table 3 shows the number of correct and incorrect predictions and the accuracy of the predictions of each of the test groups. The total accuracy of all test records is shown in Figure 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.

The results obtained show how, from some initial academic activities of the course, the student's academic performance at the end of the course can be predicted with a certain percentage of accuracy. The design of the predictive model was made from the discovery process of knowledge extraction of databases. In order to automate the predictive model, a platform was created that allowed to enter the data of the students to build the model and, later, to enter the student data of which their academic performance will be predicted.

Once the model is constructed, the main factors (academic activities) that significantly influence the academic performance of the students who participated in this study can be identified by calculating the posterior probabilities. This is because the probabilities that have greater value will significantly influence the decision of approval or failure of the student according to the model developed with the Naïve Bayes algorithm.

In Table 2, it can be seen that the a priori probabilities for the attribute "approves" are similar, however, the probability P (approves = NO) is greater, because in the training data, there is a greater amount of failed students who pass, which is typical of university courses in Mathematics. The posterior probabilities P (A / SI) of activities 3, 4 and 5 were those that had a higher value. Therefore, these academic activities are those that most influenced the prediction of approval of the students who participated in this study. In this way, students who pass activities 3, 4 and 5 are more likely to pass the course. This does not happen with activities 1 and 2, which can be attributed to several factors, for example, that students in these early activities are just adapting to the course or did not enter from the beginning of the course due to administrative delays in their enrollment.

Table 3 shows that the accuracy of the predictions is higher in some test groups than in others due to various factors of the student not taken into account in the built model, for example, current average, number of failed subjects or schooling of parents, among others.

The total accuracy of all the test data (66.28%) was very similar to that estimated in the cross validation (63.83%), so the test data behaved similarly to the training data.

In works that have used the Naïve Bayes algorithm, similar accuracy values have been obtained. In Jishan, Rashu, Haque and Rahman (2015), 181 students participated; the highest accuracy value was obtained using six attributes and was 75%. In Mueen et al. (2016) 60 students participated and the highest accuracy value was obtained with 38 attributes and was 86%. In this work, accuracies of the predictions of the approval of the course of up to 73% were obtained, with only five attributes corresponding to the grades of the initial academic activities of the same. We also identified the main factors that influenced the academic performance of the sample of data analyzed, similar to how it was done in Mueen et al. (2016). In addition, unlike the mentioned works, a platform was created that allows the

automation of predictions of academic performance to facilitate its use by teachers of educational institutions.

This type of models offers the possibility to educational institutions to design strategies for prevention of failure and identify the most relevant factors that influence the academic performance of their students.

Conclusions

The objective of this work was the design and automation of a predictive model of the academic performance of IPN students. This model was built using the Naïve Bayes algorithm and was automated using appropriate programming languages for subsequent publication on a website and, thus, becomes accessible to any type of teacher and not only experts in the mining area. of data.

The model was evaluated with respect to the accuracy of the predictions, obtaining values of up to 73%, taking into account that only a few initial activities carried out by the students were used.

In the construction of the predictive model, the a priori and a posteriori probabilities of the attributes were calculated. Through these, the main activities that affect the academic performance of the academic data set analyzed were identified. In this way, the built model allows to obtain predictions of academic performance and identify the main academic activities that affect it.

Evaluations of the initial activities of a course are a common practice for most teachers in educational institutions. In this way, the methodology can be replicated by the latter to build predictive models for their own students, and, thus, have the opportunity to design prevention strategies and decrease recovery strategies that imply that the student rejects any partial evaluation for perform some type of intervention. Recovery strategies are a common practice in most educational institutions.

References

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