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

https://doi.org/10.23913/ride.v12i24.1196

Artículos científicos

Modelos predictivos progresivos del rendimiento académico de estudiantes universitarios

Progressive Predictive Models of the Academic Performance of University Students

Modelos preditivos progressivos de desempenho acadêmico de estudantes universitários

Andrés Rico Páez
Instituto Politécnico Nacional, México
aricop.ipn@gmail.com
https://orcid.org/0000-0002-6450-318X

Resumen

El objetivo de esta investigación fue desarrollar modelos predictivos progresivos del rendimiento académico de estudiantes universitarios de México y evaluarlos para distintas técnicas de aprendizaje automático. En este estudio se recopilaron calificaciones de actividades académicas de 260 estudiantes universitarios para crear modelos de predicción de los resultados académicos mediante técnicas de aprendizaje automático. Se construyeron los modelos en diferentes etapas a lo largo del curso y se evaluaron empleando la exactitud en la predicción de 112 estudiantes de un curso posterior. Se observó una exactitud de hasta 70.5 % en un tiempo de 21 % del total de la duración del curso. Este tipo de metodología puede ser replicada para diferentes tipos de cursos debido a que el registro de calificaciones es común en casi todos ellos. Además, esta metodología es flexible en cuanto a la elección de la etapa temporal en la cual realizar las predicciones, sin perder el compromiso con la exactitud. Así, se puede efectuar en etapas tempranas para detectar problemas con el rendimiento académico y evitar, en la medida de lo posible, la reprobación y deserción de estudiantes.

Palabras clave: aprendizaje automático, modelo matemático, prevención, rendimiento académico.

Abstract

The objective of this research was to develop progressive predictive models of the academic performance of university students in Mexico and evaluate them for different machine learning techniques. In this study, grades of academic activities of 260 university students were collected to create prediction models of academic results using machine learning techniques. The models were built at different stages throughout the course and were evaluated using the accuracy of the predictions by applying it to the prediction of 112 students in a subsequent course. An accuracy of up to 70.5 % was observed in a time of 21 % of the total duration of the course. This type of methodology can be replicated for different types of courses because the recording of grades is common in almost all courses. In addition, this methodology is flexible in terms of choosing the time stage in which to make the predictions, maintaining a compromise between the accuracy of the predictions and that they be made at the earliest possible stage to detect problems with academic performance, avoiding, in as far as possible, the failure and desertion of students.

Keywords: machine learning, mathematical model, prevention, academic performance.

Resumo

O objetivo desta pesquisa foi desenvolver modelos preditivos progressivos do desempenho acadêmico de estudantes universitários no México e avaliá-los para diferentes técnicas de aprendizado de máquina. Neste estudo, foram coletadas notas de atividades acadêmicas de 260 estudantes universitários para criar modelos de previsão de resultados acadêmicos usando técnicas de aprendizado de máquina. Os modelos foram construídos em diferentes etapas ao longo do curso e testados usando a precisão de previsão de 112 alunos de um curso subsequente. Foi observada acurácia de até 70,5% em um tempo de 21% da duração total do curso. Esse tipo de metodologia pode ser replicada para diferentes tipos de cursos, pois o registro de notas é comum a quase todos eles. Além disso, essa metodologia é flexível quanto à escolha do momento de realização das previsões, sem perder o compromisso com a precisão. Assim, pode ser feito precocemente para detectar problemas com o desempenho acadêmico e evitar, na medida do possível, a reprovação e a evasão dos alunos.

Palavras-chave: aprendizado de máquina, modelo matemático, prevenção, desempenho acadêmico.

Fecha Recepción: Noviembre 2021 **Fecha Aceptación:** Mayo 2022

Introduction

Currently, technological development has caused, in the educational area, to emerge large amounts of data referring to students, teachers and other members of the educational process. Commonly, this data is generated for certain purposes and is not analyzed, mainly because it is unknown how to do it. However, there may be potentially useful information that can benefit the educational process in reducing student dropout and improving academic performance. Student dropout is a problem that is associated with many variables and that harms all actors in the educational process (Rivera, 2021). Academic performance is the sum of different and complex factors of the student environment (Garbanzo, 2007) and, in educational institutions, it is one of the main indicators of educational quality. In this way, for educational institutions it is important to collect data to analyze them and find information that can improve their educational system (Bakhshinategh, et al., 2018). In recent years, various studies have used data analysis to predict school performance (Kalaivani, Priyadharshini, & Selva, 2017; La Red et al., 2015).

The prediction of academic performance can be done at different levels of detail, for example, to predict results of tasks, exams or an entire course (Asif et al., 2017). In any case, the prediction of academic performance is desirable because it allows early identification of students at risk of failing and allows some type of intervention to prevent them from abandoning their studies and encourage their retention in school.

Machine learning techniques are those that learn a model from a data set, and are currently being used to build prediction models of student learning outcomes (Xing et al., 2015). In other words, these techniques allow the construction of models that learn from data from educational environments and then predict results from new data (Contreras, Fuentes and Rodríguez, 2020). Altujjar et al. (2016) built a predictive model using the machine learning technique known as the ID3 decision tree algorithm to predict underachievement in college students. Hussain et al. (2018) made predictions of learning outcomes with different Bayesian network algorithms and decision trees based on socioeconomic and demographic data of university students. Usman et al. (2019) used decision tree, naive Bayes and k-nearest neighbors techniques to predict academic performance based on their interaction with an educational platform on the Internet. Contreras et al. (2020) implemented machine learning techniques such as the k nearest neighbors to predict the performance of industrial engineering students.

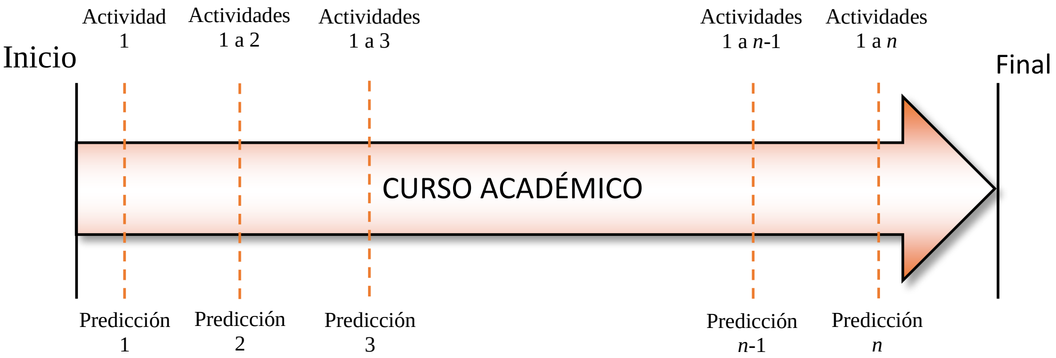
In the literature there are several studies about the prediction of academic performance that have provided useful information for the understanding and planning of educational processes. However, low student grades continue to be a problem in educational institutions, mainly universities. This can be reflected in the large number of university students who fail and drop out, which is more frequent in the first years of study (Silva, 2011). In this way, there is a need to develop methodologies for the prediction of learning outcomes that allow the analysis and use of school information to improve educational quality. However, in our country, there are few works that make use of machine learning techniques in the construction of predictive models of academic performance (Ayala, López and Menéndez, 2021; Juárez, Cortés and Coronilla, 2014; Valero, Vargas and García, 2010).

In this research, the following questions are raised: how to carry out progressive prediction models of the academic performance of students of a university in Mexico? And how to evaluate the models made using the proposed machine learning techniques? In this way, the objective of this research is to develop progressive prediction models of the academic performance of students from a Mexican university and evaluate them for different machine learning techniques.

Methodology

This paper proposes a methodology that consists of using the grades obtained in academic activities by students during a course at a public university in Mexico. The methodology consists of making a prediction 1 with a predictive model built from activity 1, then a prediction 2 is made with a predictive model made with activities from 1 to 2, this process is repeated until activities 1 are used. up to n This methodology is illustrated in Figure 1. Therefore, as the course progresses, more data is added to the predictive models.

Figura 1. Metodología propuesta para construir modelos predictivos progresivos del rendimiento académico en un curso



Fuente: Elaboración propia

260 students from a public university in Mexico participated in this study and 14 academic activities carried out during a course were used. These activities are carried out in similar time spaces throughout the course. Each activity is considered approved (A) if its grade is between 6 and 10; otherwise, it is considered failed (R), and the case in which the student does not present the activity (NP) is also considered. In this way, a table with 260 records and 15 columns (attributes) is built, which will be used to build the predictive models. A sample of this is presented in Table 1. Each activity is represented as “act” and the number of the activity. In addition, the approval of the student is represented with the attribute "approves", which can have the values of "Yes" or "No".

Tabla 1. Muestra de datos para los modelos predictivos del rendimiento académico

act1	act2	act3	act4	act5	act6	act7	act8	act9	act10	act11	act12	act13	act14	aprueba
R	A	R	A	A	R	A	A	NP	R	NP	A	R	R	Sí
NP	NP	NP	NP	NP	NP	NP	NP	NP	NP	NP	NP	NP	NP	No
R	A	R	A	A	R	A	A	A	R	A	R	A	R	No
NP	NP	R	A	A	R	A	R	A	R	A	R	R	R	Sí
R	NP	R	A	A	R	A	R	A	A	A	A	R	R	Sí
NP	NP	A	NP	A	NP	NP	R	A	R	A	A	R	R	Sí
R	R	R	NP	A	R	A	A	A	A	A	R	R	R	Sí
NP	NP	NP	NP	NP	R	NP	NP	NP	NP	NP	NP	NP	NP	No
R	NP	A	A	NP	R	A	NP	NP	R	A	NP	NP	NP	No
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:

Fuente: Elaboración propia

Results

A predictive model is built using a set of data known as training data and a machine learning technique. In this study, the training data is the table with 260 records and the machine learning techniques used are naive Bayes, k nearest neighbors and C4.5 decision tree (Hernández, Ramírez and Ferri, 2004). All the data analyzes shown in this article were carried out with the support of the free software Weka (Witten, Frank and Hall, 2005).

In this research, 14 academic activities of students were collected, for which 14 data groups were constructed through table 1. The first group consists of the data from the first column of the table of 260 records (activity 1) and from the attribute column “approves”. The second group contains the data from the first two columns of the table (activities 1 to 2) and from the attribute column “approves”. So on, until all the columns of the table are included (14 activities and the attribute “approves”). With each of the 14 groups of training data, a predictive model is built for each of the machine learning techniques. In this way, a predictive model is built, firstly, with activity 1, then with activities 1 and 2, and so on, until the 14 academic activities are used. That is, the predictive models progressively increase academic activities in their training data.

In this research, the accuracy of the predictions is used as a metric for evaluating the performance of the predictive models, which is defined as the number of predictions that were correct divided by the total predictions (Durairaj and Vijitha, 2014).

The k nearest neighbors technique uses the parameter k, one way to select it is by choosing the one that achieves a higher value in the accuracy of the predictions. For this, cross-validation is used, which consists of randomly partitioning the data into a fixed number of partitions; a partition is reserved to make the predictions and the rest to build the predictive model, this action is repeated leaving a different partition to make the predictions. Accuracy is calculated by averaging the accuracies obtained with each partition. For these experiments, a cross-validation with 10 partitions was used, since it has been used in similar works (Márquez et al., 2012; Mueen, Zafar and Manzoor, 2016). Each data group has 260 records, so for each of the 14 groups the accuracy is calculated for various values of k (1, 2, 3…, 260) and the value of k is chosen for each group where the highest accuracy is obtained, as shown in Table 2.

Tabla 2. Valores de *k* donde se obtiene la mayor exactitud con la validación cruzada mediante la técnica *k* vecinos más cercanos

Cantidad de actividades académicas	Valor de *k* donde se obtuvo la mayor exactitud
Actividad 1	1
Actividades 1 a la 2	3
Actividades 1 a la 3	11
Actividades 1 a la 4	68
Actividades 1 a la 5	90
Actividades 1 a la 6	74
Actividades 1 a la 7	36
Actividades 1 a la 8	18
Actividades 1 a la 9	87
Actividades 1 a la 10	6
Actividades 1 a la 11	155
Actividades 1 a la 12	164
Actividades 1 a la 13	115

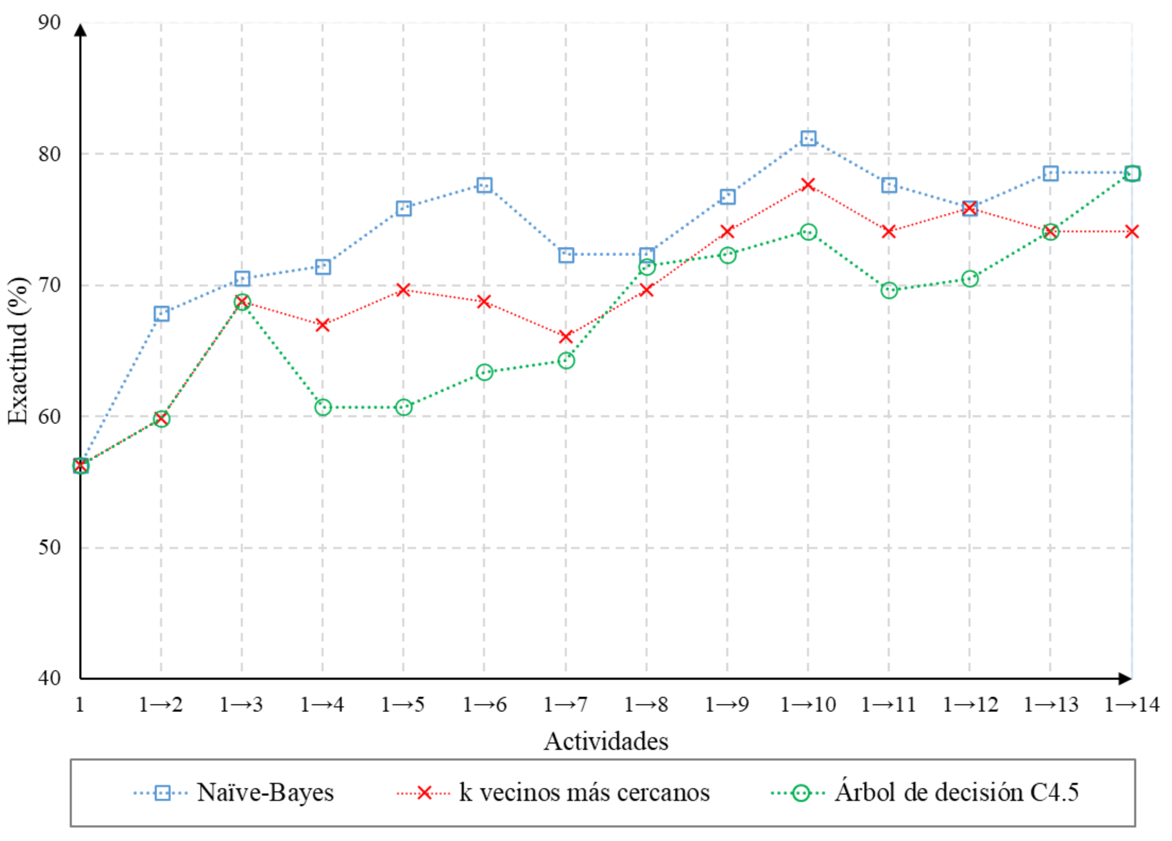
Actividades 1 a la 14	105

Fuente: Elaboración propia

With the calculations above, the three machine learning techniques are applied to the 14 sets of training data to build predictive models of academic performance. These models were used to predict the academic performance of 112 college students taking the same course with the same number of activities, but one semester after the semester in which the training data was collected. This type of data is known as test data, that is, it is data that is different from the training data and to which the predictions will be made.

After making the predictions, the results at the end of the course obtained by the 112 students were collected and the predictions that were correct were counted. In this way, the accuracy was calculated on the test data from the predictive models built with each of the 14 training data sets and with each of the three machine learning techniques. It should be noted that the predictive models for each of the 14 groups represent progressive predictive models of academic performance because they change as the number of activities increases throughout the course. Figure 2 presents the accuracy of progressive predictive models with different machine learning techniques. The number of academic activities used to build the predictive models is indicated on the abscissa axis, for example, when activity 1 to 3 was used, it was represented with the notation *1→3*.

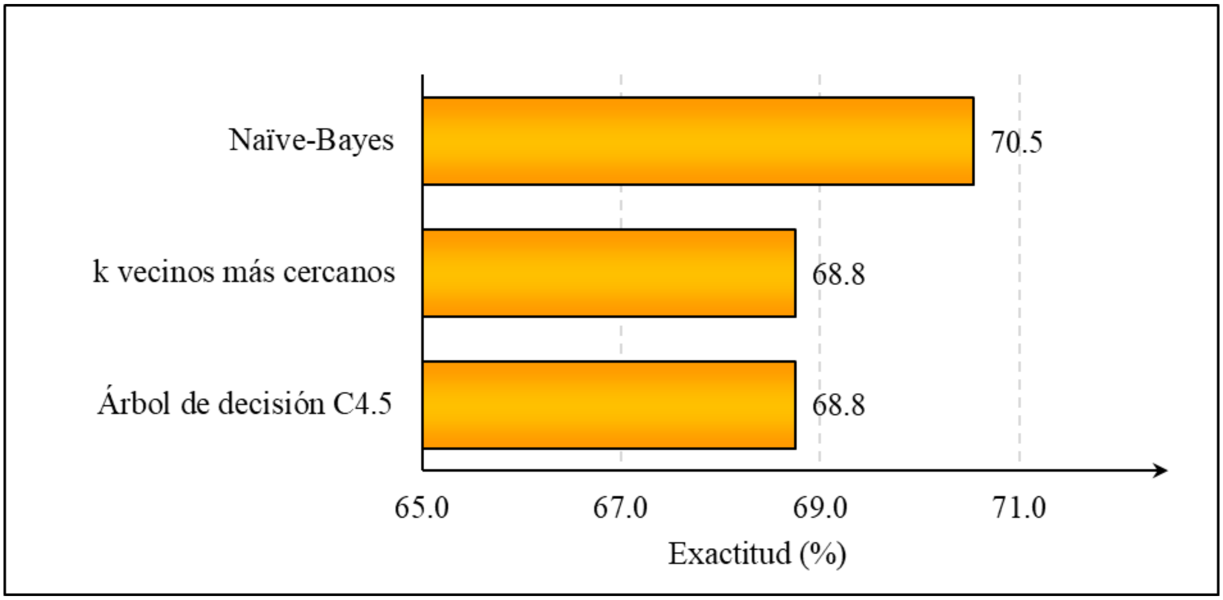
Figura 2. Exactitud de los modelos predictivos progresivos del rendimiento



Fuente: Elaboración propia

Figure 2 shows that there is a certain tendency that the greater the number of activities that the predictive models have, the greater the accuracy of the predictions, regardless of which machine learning technique is used. In other words, predictive models tend to have greater prediction accuracy as the course progresses and more information is added. Also, it can be noted that the progressive predictive models of academic performance carried out with the naive Bayes technique have a better performance in terms of the accuracy of the predictions. The highest accuracy value is 81.25% and is achieved with the naive Bayes technique and when activities 1 to 10 are used. However, activity 10 is near the end of the course, so the prediction of a student near the end of a course limits the actions that can be taken to avoid failure. In this way, it is necessary to maintain a compromise between the early identification of the student's academic performance and the accuracy of the predictions. A value of interest is when activities 1 to 3 are used to create predictive models because with the three techniques they generate an accuracy value of around 70% (figure 3) and it is achieved with the first three activities. In other words, from the beginning of the course to the third activity, around 21% of the course duration has passed (100% x 3/14), which allows early identification of students in danger of failing.

Figura 3. Exactitud de los modelos predictivos del rendimiento académico con las primeras tres actividades académicas



Fuente: Elaboración propia

Discussion

In the results obtained, it can be seen how the accuracy of the predictive models has a growing trend as the number of activities increases, due to the fact that more information is added to the models as the course progresses, and although on some occasions the accuracy is maintains or decreases, overall it tends to increase. The progressive predictive models carried out with the naive Bayes technique have a better performance in terms of the accuracy of the predictions because in most cases the accuracy is superior to the other two techniques. This agrees with what was observed by Osmanbegović and Suljić (2012) and Mueen et al. (2016), who, with similar amounts of data, obtained higher accuracy values with this technique compared to others.

In progressive predictive models, it is not necessary to reach the end of the course or even the last academic activities to obtain acceptable predictions to identify possible students who will fail the course. In this sense, it is observed that only with the first three academic activities predictive models of up to 70.5% are elaborated with the naive Bayes technique.

This is important because it not only allows students in danger of failing to be detected, but also gives educational institutions and teachers a reasonable period of time to carry out the necessary interventions on specific students.

In the literature, studies have been observed that make predictions of academic performance using machine learning techniques with academic activities. Sharma and Vishwakarma (2017) conducted a study in which 70 students participated. They divided the data into 50 records to build the predictive model and 20 records to predict academic performance. Thus, they obtained an accuracy of 90% using all academic activities up to the middle of the course, that is, this accuracy value was obtained at 50% of the duration of the course. Del Campo et al. (2017) used 124 student records and obtained, through cross-validation, a maximum accuracy of 71.57% at 50% of the time of starting the course. Alcaraz et al. (2020) used 78 student records and, with cross-validation, obtained a maximum accuracy of 79.5% at 25% of the time of starting the course. In these works only the training data were used to calculate the accuracy. Unlike these works, in this research the models were used on test data, that is, from students of a course after the course from which the training data was collected. It is worth mentioning that, although the accuracy of the predictions in these works was greater than that obtained in this article (70%), in this investigation said accuracy was achieved in 21% of the time of having started the course, that is, obtained this accuracy at an earlier stage than in the other works. Also, this research shows the behavior of the accuracy of the predictions at different stages of the course through the representation of academic activities, which allows selecting a number of activities that maintain a compromise between a high accuracy value and an early realization of the predictions.

In this study, the causes that led to students failing could have been a poor design of academic activities, teacher performance, learning style, among others, because school failure depends on many factors (Antelm, Cacheiro and Gil, 2015). Therefore, it is recommended that the teacher carry out activities that facilitate the students' acquisition of knowledge and skills that allow them to build their learning process to improve their academic performance.

Conclusions

In this research, a methodology was proposed that uses academic activities during a course to make predictions of the academic performance of students. To make these predictions, it was shown how to develop progressive prediction models of the academic performance of students from a Mexican university. These models were built using 260 student records and employing naive Bayes, k-nearest neighbors, and C4.5 decision tree machine learning techniques. To evaluate these models, the accuracy of the predictions obtained by applying these models in the prediction of the academic performance of 112 students was used. Accuracy was obtained for different amounts of academic activities during the course.

It was observed that the accuracy had a tendency to increase as the course progressed, and that using the first three activities an accuracy of up to 70.5% can be obtained in 21% of the time of having started the course and it was done at an earlier stage than the first three activities. in other similar articles reviewed in the literature. It should be noted that the record of grades for academic activities is common to all courses, so it is data that can be easily collected, which allows this methodology to be applied to a large number of courses in different areas. Similarly, the methodology is flexible in terms of selecting the time stage in which the predictions are made based on the accuracy that is considered acceptable.

Future lines of research

Despite the advances achieved in this research, it is convenient to mention the possible explorations that can be carried out in this area. Firstly, it may be of interest to carry out studies using more complex techniques than those presented in this research, such as the majority vote technique, which involves voting between machine learning techniques, such as those shown in this work, and making a decision based on what the majority decides. Also, attributes referring to qualifications of academic activities have been used, however, demographic or socioeconomic attributes that may influence academic performance can also be used.

References

Alcaraz, R., Martínez, A., Zangróniz, R. y Rieta, J. J. (2020). Predicción temprana del fracaso en una asignatura de electrónica con técnicas de aprendizaje automático. Ponencia presentada en el XIV Congreso de Tecnologías Aplicadas a la Enseñanza de la Electrónica. Porto, del 8 al 10 de julio de 2020. https://dialnet.unirioja.es/servlet/articulo?codigo=7980475.

Altujjar, Y., Altamimi, W., Al-Turaiki, I. y Al-Razgan, M. (2016). Predicting Critical Courses Affecting Students Performance: A Case Study. *Procedia Computer Science 82*, 65-71. https://doi.org/10.1016/j.procs.2016.04.010.

Antelm, A. M., Cacheiro, M. L. y Gil, A. J. (2015). Análisis del fracaso escolar desde la perspectiva del alumnado y su relación con el estilo de aprendizaje. *Educación y Educadores, 18*(3), 471-489. https://www.redalyc.org/articulo.oa?id=83443150006.

Asif, R., Merceron, A., Ali, S. A. y Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education, 113*, 177-194. https://doi.org/10.1016/j.compedu.2017.05.007.

Ayala, E., López, R. E. y Menéndez, V. H. (2021). Modelos predictivos de riesgo académico en carreras de computación con minería de datos educativos. *Revista de Educación a Distancia (RED), 21*(66). https://doi.org/10.6018/red.463561.

Bakhshinategh, B., Zaiane, O. R., ElAtia, S. y Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies, 23*(1), 537-553. https://doi.org/10.1007/s10639-017-9616-z.

Contreras, L. E., Fuentes, H. J. y Rodríguez, J. I. (2020). Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático. *Formación Universitaria, 13*(5), 233-246. https://doi.org/10.4067/S0718-50062020000500233.

Del Campo, J., Ramos, G., Morales, R. y Baena, M. (2017). Minería de datos educativos para la predicción personalizada del rendimiento académico. Conferencia Internacional de Procesamiento de la Información. Villa Clara, del 23 al 27 de octubre de 2017. https://riuma.uma.es/xmlui/handle/10630/15477.

Durairaj, M. y Vijitha, C. (2014). Educational Data mining for Prediction of Student Performance Using Clustering Algorithms. *International Journal of Computer*

Science and Information Technologies, *5*(4), 5987-5991. http://www.ijcsit.com/docs/Volume%205/vol5issue04/ijcsit20140504249.pdf.

Garbanzo, G. M. (2007). Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública. *Educación, 31*(1), 43-63. https://www.redalyc.org/pdf/440/44031103.pdf.

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

Hussain, S., Dahan, N. A., Ba-Alwi, F. M. y Ribata, N. (2018). Educational Data Mining and Analysis of Students' Academic Performance Using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*, *9*(2), 447-459. https://doi.org/10.11591/ijeecs.v9.i2.pp447-459.

Juárez, A., Cortés, J. y Coronilla, U. (2014). Aplicación de la inteligencia artificial en la sistematización de procesos educativos. Caso: Sistema de detección de riesgo escolar en ESCOM. *Revista Iberoamericana de Producción Académica y Gestión Educativa*, *1*(1), 140-163. https://www.pag.org.mx/index.php/PAG/article/view/92/140.

Kalaivani, S., Priyadharshini, B. y Selva, B. (2017). Analyzing Student's Academic Performance Based on Data Mining Approach. *International Journal of Innovative Research in Computer Science & Technology*, *5*(1), 194-197. https://doi.org/DOI:10.21276/ijircst.2017.5.1.4.

La Red, D., Karanik, M., Giovannini, M. y Pinto, N. (2015). Perfiles de rendimiento académico: un modelo basado en minería de datos. *Campus Virtuales*, *4*(1), 12-30. https://redined.educacion.gob.es/xmlui/bitstream/handle/11162/120661/1.pdf?sequence=1&isAllowed=y.

Márquez, C., Cano, A., Romero, C. y Ventura, S. (2012). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied Intelligence*, *38*(3), 315-330. https://doi.org/10.1007/s10489-012-0374-8.

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*, *8*(11), 36-42. https://doi.org/10.5815/ijmecs.2016.11.05.

Osmanbegović, E. y Suljić, M. (2012). Data Mining Approach for Predicting Student Performance. *Journal of Economics and Business, 10*(1), 3-12. https://www.econstor.eu/handle/10419/193806.

Rivera, K. (2021). Modelo predictivo para la detección temprana de estudiantes con alto riesgo de deserción académica. *Revista Innovación y Software, 2*(2), 6-13. https://revistas.ulasalle.edu.pe/innosoft/article/view/40/37.

Sharma, G. y Vishwakarma, S. K. (2017). Analysis and Prediction of Student's Academic Performance in University Courses. *International Journal of Computer Applications, 160*(4), 40-44. https://doi.org/10.5120/IJCA2017913045.

Silva, M. (2011). El primer año universitario. Un tramo crítico para el éxito académico. *Perfiles Educativos*, *33*(especial), 102-114. http://www.scielo.org.mx/pdf/peredu/v33nspe/v33nspea10.pdf.

Usman, U. I., Salisu, A., Barroon, A. I. E. y Yusuf, A. (2019). A Comparative Study of Base Classifiers in Predicting Students' Performance based on Interaction with LMS Platform. *FUDMA Journal of Sciences, 3*(1), 231-239.

Valero, S., Vargas, 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 Prieto, M. E., Dodero, J. M. y Villegas, D. O. (eds.), *Recursos digitales para la educación y la cultura volumen Kaambal*, (pp. 33-39). Mérida, México: Universidad Tecnológica Metropolitana. http://fcaenlinea.unam.mx/anexos/1566/1566_u6_act1b.pdf.

Witten, I., Frank, E. y Hall, M. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Massachusetts, United States: 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. https://doi.org/10.1016/j.chb.2014.09.034.