Article **Journal of Technical Education**
June 2022, Vol.6 No.16 1-5

Predictive model for the analysis of academic performance and preventing student dropout using machine learning techniques

Modelo predictivo para el análisis del rendimiento académico y prevenir la deserción estudiantil utilizando técnicas de aprendizaje automático

LÓPEZ-GARCÍA, Lourdes†*, LINO-RAMÍREZ, Carlos, ZAMUDIO-RODRÍGUEZ, Víctor Manuel and DEL VALLE- HERNÁNDEZ, Josué

Tecnológico Nacional de México, Campus León, División de Estudios de Posgrado e Investigación, León, Guanajuato, México

ID 1st Author: *Lourdes, López-García* / **ORC ID:** 0000-0002-9817-5022, **CVU CONACYT ID:** 901504

ID 1st Co-author: *Carlos, Lino-Ramírez* / **ORC ID:** 0000-0002-6415-8435, **CVU CONACYT ID:** 395781

ID 2nd Co-author: *Víctor Manuel, Zamudio-Rodriguez* / **ORC ID:** 0000-0002-9246-7999, **CVU CONACYT ID:** 70912

ID 3rd Co-author: *Josué, Del Valle- Hernández* / **ORC ID**: 0000-0002-3373-3310, **CVU CONACYT ID:** 167917

DOI: 10.35429/JOTE.2022.16.6.1.5 Received January 10, 2022; Accepted June 30, 2022

Abstract

School dropout is one of the biggest problems in the country of Mexico, there are several factors that cause it, so it is necessary to propose strategies and lines of action to reduce it. This document analyzes a database with the demographic and social characteristics of high school students, which were collected through the application of school questionnaires and reports, in order to detect the factors that cause students to drop out of school, as well as to identify in time the students who need personalized counseling to offer them educational guidance and prevent them from dropping out of school, this analysis was implemented through machine learning techniques by developing a predictive model with the gradient descent algorithm, from the results to check the forecast errors by applying the mean square error metric, to estimate the possible prediction errors of the model, it is expected to have a great social impact by applying these machine learning techniques in educational community achieving that students can strengthen their comprehensive training, in addition to guiding their talents and interests.

Predictive model, Gradient descent, School dropout

Resumen

La deserción escolar es una de las problemáticas más grandes en el país de México, son diversos los factores que la causan por lo que es necesario proponer estrategias y líneas de acción para abatirla. En el presente documento se analiza una base de datos con las características demográficas y sociales de los alumnos de secundaria, que se recopilaron con la aplicación de cuestionarios escolares e informes, con la finalidad de detectar los factores que son causantes de la deserción escolar, así mismo identificar a tiempo a los alumnos que necesitan asesorías personalizadas para ofrecerles orientación educativa y evitar que abandonen sus estudios, dicho análisis se implementó mediante técnicas de aprendizaje automático elaborando un modelo predictivo con el algoritmo de descenso de gradiente, a partir de los resultados poder comprobar los errores de pronóstico aplicando la métrica de error cuadrático medio, para estimar los posibles errores de predicción del modelo, se espera tener un gran impacto social al aplicar estas técnicas de aprendizaje automático en comunidad educativa logrando que los alumnos puedan fortalecer su formación integral, además de orientar sus talentos e intereses.

Modelo predictivo, Descenso de gradiente, Deserción escolar

Citation: LÓPEZ-GARCÍA, Lourdes, LINO-RAMÍREZ, Carlos, ZAMUDIO-RODRÍGUEZ, Víctor Manuel and DEL VALLE- HERNÁNDEZ, Josué. Predictive model for the analysis of academic performance and preventing student dropout using machine learning techniques. Journal of Technical Education. 2022. 6-16:1-5.

* Correspondence to the Author (E-mail: m20241262@leon.tecnm.mx)
† Researcher contributing as first author.

©ECORFAN-Republic of Peru

Introduction

This document begins with a summary of the topic to be developed, followed by the introduction containing an explanation of the topic in general, the problem to be solved and the central hypothesis, giving a solution to the problem in question, followed by the method and technique used for the development of the project, then the conclusion of the document was written thinking about the limitations of the work done and a proposal for improvement for future work. Finally, the references contain the bibliographic data implemented for the elaboration of the research.

One of the objectives of the analysis of educational data is to find patterns and predictions that allow characterizing the academic development of students, however, it is required the collection of data on the characteristics of students taking into account the context, in order to achieve a better understanding of the results obtained. Some of these characteristics are socioeconomic factors, family and school data of the student (Rico Páez & Gaytán Ramírez, 2022).

It is possible "if machine learning techniques are implemented for the development of a model that allows the prediction of the academic performance of secondary education students, as an effective tool for educational institutions can prevent school dropout", since knowing those students who need counseling and educational guidance, can prevent them from abandoning their studies and significantly reduce school dropout.

For the above described in this work, we analyzed the data of high school students with social demographic characteristics that were collected through reports and school questionnaires, using *Python* programming language, and also applied machine learning techniques to develop a predictive model using the *gradient descent algorithm* and from the data we estimated the probabilities and distributions of the objects of interest, so as to compare the forecast errors by applying the *mean square error metric* to estimate the possible prediction errors of the model.

Related Works

The factors that influence university dropout are multi-systemic, and the strategies used by the different higher education institutions to increase the retention of their students take preponderance (Henríquez Cabezas & Vargas Escobar, 2022).

In recent years, *Artificial Intelligence* (AI) techniques such as *Machine Learning* (ML) and *Deep Learning* (DL) have had a positive impact on the advancement of different fields of knowledge, including education. Education is an important driver of all societies, enabling individuals to be more productive and solve problems more effectively by generally applying creative approaches. In education, the aforementioned ML techniques have been used for different tasks including dropout prediction and student performance support (Cruz et al., 2022).

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 (Rico Páez & Sánchez Guzmán, 2018).

Data mining is the process of extracting useful and understandable knowledge, previously unknown, from stored data. Such analysis process works at the knowledge level with the purpose of finding patterns and relationships, as well as predictive models that provide knowledge patterns for decision making (Rico Páez & Sánchez Guzmán, 2018).

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 data in the educational environment. Therefore, it is proposed that educational data mining aims to discover knowledge and patterns within student data. These patterns characterize student behavior based on their achievements, assessments, and mastery of knowledge content (Rico Páez & Sánchez Guzmán, 2018).

ISSN-2523-2460
ECORFAN® All rights reserved.

Methodology

Tools used

The project was coded with *Python* programming language, which is the most efficient for data science. In the data processing section, the pandas libraries were imported, Numpy specifically for linear algebra, Seaborn for data visualization and finally Matplotlib for the creation of predictive model graphics, together with Scikit-learn, which is a collection of algorithms.

Data set

The logistic regression model was developed from a dataset of student achievement in secondary education (Cortez & Silva, 2008). The data attributes include social demographic characteristics and are related to two schools, were collected through school reports and questionnaires.

Two sets of data are provided about achievement in two different subjects: mathematics and Spanish. The target attribute for the data analysis is: 'final grade' as it corresponds to the final average for the year.

Data processing

The variable 'school' was used to separate the data set into two strata and perform a sampling, which can be seen in Table 1:

	School	Sex	Ege	Addres	Family	Stat
0	GP	F	18	U	GT3	A
1	GP	F	17	U	GT3	T
2	GP	F	15	U	LE3	T
3	GP	F	15	U	GT3	T
4	GP	F	16	U	GT3	T

Table 1 Data reading
Source: Own elaboration

The technique of cross validation with random permutation or *ShuffleSplit* was implemented for statistical analysis and to obtain other measures of estimated performance, such as mean and variance, so as to know the performance of the data and the training/testing rates, likewise for the creation of the machine learning model, the variable of 'school' was used, with stratified sampling where the data set was divided into 80% test and 20% training. The purpose of implementing the *ShuffleSplit* cross validation was to return the stratified random folds, these folds are made preserving the percentage of samples for each class, as follows: Table 2:

	School	Sex	Age	Address	Family
464	MS	M	16	R	GT3
595	MS	M	18	U	LE3
268	GP	M	17	R	LE3
346	GP	M	17	U	LE3
528	MS	F	17	R	GT3

Table 2 Training set
Source: Own elaboration

The main measures of central tendency and some basic descriptive statistics were returned for each variable in the data sheet, as shown in Table 3 of the data set with the respective percentiles (division of an ordered series of data).

	0	1	2
count	519.000000	519.000000	519.000000
mean	16.739884	2.549133	2.344894
Std	1.223865	1.117168	1.102178
min	15.000000	0.000000	0.000000
25%	16.000000	2.000000	1.000000
50%	17.000000	2.000000	2.000000
75%	18.000000	4.000000	3.000000
max	22.000000	4.000000	4.000000

Table 3 Descriptive statistics
Source: Own elaboration

Correlation matrix

For the analysis of the data, potential relationships between the variables being analyzed were explored, using the statistical measure called *Pearson's correlation*, which indicates the magnitude and direction of the relationship that could exist between two variables, and for this purpose the correlation matrix was calculated.

ISSN-2523-2460
ECORFAN® All rights reserved.

Standardization of nominal variables and binary variables

The data set contains nominal variables and binary variables, for this reason it was necessary to separate them using the *OneHotEncoder* functions for the numerical variables and *OrdinalEncoder* for the categorical variables, and *StandardScaler* to standardize these variables, in order to avoid any error in the predictive model, as shown in Table 4:

	0	1	2
count	519.000000	519.000000	519.000000
mean	0.348748	0.416185	0.707129
Std	0.477034	0.493401	0.455519
min	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000
50%	0.000000	0.000000	1.000000
75%	1.000000	1.000000	1.000000
max	1.000000	1.000000	1.000000

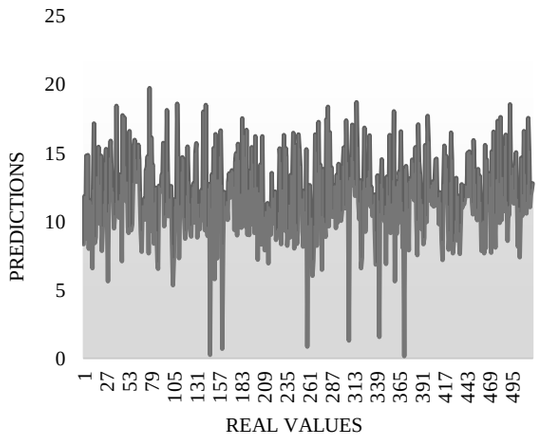
Table 4 Descriptive statistics of the standardized variables
Source: Own elaboration

Gradient descent

For the coding of the *gradient descent algorithm* we used the parameters (m and b) that refer to unknown constants called coefficients, it is the quotient between the interaction obtained from both variables and the sum of quadratic of the values of the dependent variable, in such a way that these parameters minimize the *cost function*, which was calculated with the best possible vector directly for the global minimum, the *cost function* is used to calculate the loss based on the predictions made. The mathematical formula for calculating the *cost function* is:

$$y = f(x) = wx + b \tag{1}$$

Figure 1 shows the results obtained in the execution of the *algorithm*:



Graphic 1 Gradient descent algorithm predictions
Source: Own elaboration

Evaluation of the predictive model by applying the root mean square error (RMSE) metric.

Subsequently after obtaining the model predictions, the root mean square error technique was used to measure the differences between the predicted and observed values of the model, as well as to fit the data in order to compare the forecast errors of the downward gradient model.

The root *mean square error* or root mean square deviation (RMSD) is the square root of the average of the squared errors. RMSD is a measure of accuracy for comparing forecast errors of different models for a particular data set. the formula for calculating it is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2} \tag{2}$$

To evaluate the performance and accuracy of the *algorithm*, the *linear regression* function was implemented with the *sklearn* library, this step is particularly important to compare how well the techniques used work with the data analyzed.

Results

Interestingly, when applying *Pearson's* correlation technique it was detected that the variables in the data set that have the highest correlation with the interest of high school students to continue studying high school are the school they attend, gender, the place where they live, whether they have access to the Internet and the mother's job.

ISSN-2523-2460
ECORFAN® All rights reserved.

As for the evaluation of the results of the algorithm, as shown in Table 5 "*Mean square error*", when comparing the mean square error metric of both *linear regression* and *gradient descent* models, it can be seen that both are almost identical, but it was easier and faster to implement the linear regression technique.

Gradient descent	Linear regression
1.1704544935329404	1.1704456199971924

Table 5 Root mean square error
Source: Own elaboration

Graph 1 shows the predictions obtained as a result of the algorithm execution, and it can be seen that the predictive model was adjusted to the data set since the RMSE has a smaller value, in other words, it quantified them, so that closer values were obtained between the predicted and observed values.

Obtaining a root mean square error value of 1.170 means that the *algorithm* was very accurate since a low RMSE value indicates a better fit.

Conclusions

Speaking of technical limitations, when testing several *algorithms* to select the one that would work best, it takes too much time to train the *algorithm* and a large amount of data was used, which can cause errors, and finally the model parameters are difficult to interpret.

As a proposal to improve the present project, it is necessary to include more student information from different fields; therefore, it is expected to collect a larger data set with the *database* of middle and high school students in order to later apply the model using the new information. In addition, other methods of comparison of data mining techniques can be applied to identify signs of student dropout, based on academic performance, and to perform a classification of students based on their school performance.

Acknowledgments

We would like to thank the Tecnológico Nacional de México / Instituto Tecnológico de León for the support provided to this research.

References

Cortez, P., & Silva, A. M. G. (2008). Using Data Mining to Predict Secondary School Student Performance. In A. Brito, & J. Teixeira (Eds.), Proceedings of 5th Annual Future Business Technology Conference, Porto, 5-12. Obtenido de: https://hdl.handle.net/1822/8024

Cruz, E., González, M., & Rangel, J. C. (2022). Técnicas de machine learning aplicadas a la evaluación del rendimiento ya la predicción de la deserción de estudiantes universitarios, una revisión. Prisma Tecnológico, 13(1), 77-87. https://doi.org/10.33412/pri.v13.1.3039

Henríquez Cabezas, N., & Vargas Escobar, D. (2022). Modelos predictivos de rendimiento y deserción académica en estudiantes de primer año de una universidad pública chilena. Revista de estudios y experiencias en educación, 21(45), 299-316. https://doi.org/10.21703/0718-5162.v21.n45.2022.015

Páez, A. R., & Ramírez, N. D. G. (2022). Modelos predictivos del rendimiento académico a partir de características de estudiantes de ingeniería. IE Revista de Investigación Educativa de la REDIECH, 13, e1426-e1426. https://doi.org/10.33010/ie_rie_rediech.v13i0.1426

Rico Páez, A., & Sánchez Guzmán, D. (2018). Diseño de un modelo para automatizar la predicción del rendimiento académico en estudiantes del IPN. RIDE Revista Iberoamericana Para La Investigación Y El Desarrollo Educativo, 8(16), 246 - 266. https://doi.org/10.23913/ride.v8i16.340.

ECORFAN® All rights reserved.