Decision Tree-Based Model to Determine Student's Dropout Factors in a Mexican Higher Education Institution

María del Pilar Meza Bucio, Gustavo Gutiérrez-Carreón

Universidad Michoacana de San Nicolas de Hidalgo, Michoacán, Mexico

{maria.meza,gustavo.gutierrez}@umich.mx

Abstract. The current context in which higher education institutions operating in Mexico is adverse and multifactorial. In this work, the data obtained from a survey applied to 1,582 students are analyzed to determine the main factors that influence school dropout in a pre-COVID19 stage. With this information, an analysis of the decision tree was developed, detecting the main routes that influence school dropout. This study can be useful both to the public and to the instances involved in decision-making, to try to create an environment conducive to allow students to continue with their university education.

Keywords: learning analytics, higher education, decision making.

1 Introduction

Data analytics is a technique used in many fields of research, one of them is Education, being very useful for finding patterns in large data sets, and with it optimize and predict results that allow improving the management and administration of Higher Education Institutions (HEIs), since it is very useful for the detection and prevention of specific problems and decisions makes. Higher Education Institutions (HEIs) are in a complex landscape, where their resources are limited, and situations are changing.

According to a diagnosis of the year 2019 made by the Subsecretaria de Educacion Superior (SES), the Asociacion Nacional de Universidades e Instituciones de Educacion Superior (ANUIES), and the Asociacion Mexicana de Organos de Control y Vigilancia en Instituciones de Educacion Superior, A.C. (AMOCVIES), there are nine entities in Mexico whose universities have been in economic crisis for several years: Morelos, Oaxaca, Zacatecas, Chiapas, Estado de Mexico, Tabasco, Michoacan, Nayarit, and Sinaloa. The magnitude of the accumulated deficit of the nine universities is equivalent to 71% of their ordinary public subsidy (23,461 million pesos), with a range ranging from 29% to 190%. Under this unfavorable economic and social environment, there is a need to detect transversal problems such as school dropout and the multiple causes that originate it, through various techniques.

One of them could be using learning analytics. Elias [1] mentions that learning analytics refers to the collection and analysis of data about learners and their environments to understand and improve learning outcomes.

The Facultad de Contaduria y Ciencias Administrativas (FCCA) of the Universidad Michoacana de San Nicolas de Hidalgo (UMSNH) can enroll approximately 5,000 students from many states of the country in different degrees and modalities offered. However, the data reported by the UMSNH in the last five years of the number of students who start their studies decreases significantly with the number of students who graduate. The main objective of this work is to try to find the combination of factors that generates a greater risk of dropout out in students.

2 Related Work

In [2], a predictive model is proposed using data mining techniques through a Web interface that facilitates the identification of students vulnerable to school dropout at the Universidad Tecnologica de Izucar de Matamoros (UTIM), in Mexico. In [3], a classification model is designed to detect early dropout in the Facultad de Ingenieria of the Universidad La Salle, through the application of the CRISP-D methodology (Cross Industry Standard Process for Data Mining).

In this work, a review of the literature from 1982 to 2017 was made, in which the applications of machine learning and data mining are analyzed to board the problem with methods such as decision tree, neural networks, vector support machines (SVM), naive bayes, uniform random, k-nearest neighbor (KNN), logistic regression (LR), among others, with which prediction rules are generated based on a group of management indicators that can be used in the design of educational policies to determine the reasons for some inefficiencies of the HEIs.

The work [4] carries out a master's thesis work, where the objective of the research was based on using multivariate statistical techniques: SVM, Discriminant Analysis (DA), KNN, and LR to classify undergraduate students at the Universidad Nacional de Colombia located in two towns of Medellin (with the possibility or not of dropout), based on the information that was available on the variables defined and identified as determinants of student dropout University.

For this study, the information provided by the students who entered at the Universidad Nacional de Colombia from the first semester of 2009 to the first semester of 2016 was used, their corresponding academic performance in each enrolled period and the identification of which of them lost the quality of student at the university due to poor performance and which continued with their studies, which allowed to have a percentage of data that were used for the training of the models and the rest of the data as validation. The results allowed us to identify the technique to obtain the model with a lower percentage of error and greater sensitivity, and that could be used to make predictions of dropout in new individuals from the information of the selected variables.

There are related works in which the use of techniques of expert systems and data mining allow to establish prediction models, with which we help the person in charge in the taking of these. The paper [5] shows the results of research whose purpose is to

evaluate the technical efficiency of HEIs in Colombia between the years 2011-2013, through the application of data envelope analysis and data mining techniques.

In [6], a model is proposed to detect possible dropouts in Higher Education in a public university, where they suggest two proposals for the quantification of dropout: The first, is established as the proportion of students who graduate in a certain time that corresponds to the duration of the career; and the second is simply the number of students who dropout.

To reduce dropout, these investigations propose to improve the mechanisms of early detection of potential deserters. To elaborate their research, they used some methods: logistic regression, k-nearest neighbors, decision trees including random forests, Bayesian networks, neural networks, among others.

3 Method

For this project, variables were chosen that identified each student of the Facultad de Contaduria y Ciencias Administrativas of the UMSNH, studies previously prepared by researchers from other universities were taken as a reference, as is the case of the Universidad Complutense de Madrid, where they carried out a study to determine academic success/failure, using the techniques of multiple linear regression and logistic regression [7].

In this part of the work, the variables that are considered to have an impact on the dropout rate of the students of the bachelor's Degrees in Facultad de Contaduria y Ciencias Administrativas of the UMSNH will be described. In addition, the conceptual model is presented that will allow us to understand the interaction of the variables to later develop the Data Analysis Model.

[8] Take into consideration making a rigorous separation of the types of dropouts for their study. The authors explain that student dropout can be understood from two points of view: temporal and space. As a temporary concept, they identify three types of dropouts:

1. Premature dropout: when a student leaves a program before been accepted.
2. Early dropout: when the program is abandoned during the first four semesters.
3. Late dropout: understood as abandonment from the fifth semester onwards.

Secondly, as a space concept, reference is made to the fact that a student:

1. Change programs within the same institution.
2. Change educational institutions.
3. Leave the educational system, where there is the possibility of re-entry in the future, either to the same or to another campus in the country.

For the analysis that will be carried out in this work, the dropout will be taken as a decision to abandon the academic program, which may become temporary or definitive. With the data that we have from the Control Escolar of the UMSNH, it is difficult to track whether the students who have enrolled have temporarily dropped out, therefore, it must be conceived of dropout in general and therefore a survey will be applied in which the reasons for dropout will be detected to analyze the data.

Table 1. Entry registration against graduation of students from five universities in Mexico. The school year 2010-2011 to 2015-2016.

	UNAM		IPN		UAM		UV		UMSNH	
School year	**MI**	**ME**	**MI**	**ME**	**MI**	**ME**	**MI**	**ME**	**MI**	**ME**
2010-2011	180,763	58,584	95,743	13,478	40,712	4,854	56,582	5,025	37,264	5,637
2011-2012	187,195	58,855	95,743	13,820	41,325	4,448	58,497	5,625	39,646	5,942
2012-2013	190,707	60,748	98,624	13,077	42,242	4,674	58,212	4,767	38,561	5,706
2013-2014	196,565	60,749	100,854	12,915	43,762	5,334	58,995	5,129	31,439	5,561
2014-2015	201,206	63,346	104,125	13,630	44,301	5,063	59,284	4,779	37,139	4,355
2015-2016	204,940	Nd	104,409	12,684	44,712	5,147	59,583	5,643	nd	nd

It is not the first time that there is talk of dropout in the UMSNH, it is a very perceptible problem, especially in the schools or Faculties with an admission of low students, that is, with less than 200 new students. [9]

She analyzed the dropout of students at the Universidad Michoacana de San Nicolas de Hidalgo (UMSNH), and a comparison with the Universidad Nacional Autonoma de Mexico (UNAM), the Instituto Politecnico Nacional (IPN), the Universidad Autonoma Metropolitana (UAM), the Universidad de Veracruz (UV). The data that analyzed the entrance enrollment against the graduation enrollment. The following figure shows the results.

As seen in Table 1, 6 school cycles were analyzed. The column with MI heading corresponds to the registration of start or new entry and the column with the heading ME, refers to the registration of graduation. Data labeled nd indicates that there is no data. With this study it is appreciated it is not a problem exclusive to the UMSNH or the State Universities, it is a national-level problem.

To elaborate the table 2, the total number of new and graduate students was added to determine the average of each of them and be able to obtain the average graduation rate of the last four school cycles, which corresponds to 51%, in other words, half of the students who enter complete their university studies. Another interesting fact is the dropout rate, which is 49%.

Subsequently, a research instrument was designed in the form of a survey. Which was developed in the Google Forms application that allows you to prepare questionnaires, store the results to be able to consult them, and generate some graphs.

The questions that were asked were: age, sex, state, and hometown, marital status, family income, bachelor's degree, modality, semester, the status of the student, which can be regular students (who do not repeat subjects for the second time) and irregular student (who does repeat or recurs one or more subjects). These questions were applied to identify the interviewees, however, for the elaboration of the decision tree model, this information was not used.

A dropout factor was determined, from the related studies, a series of qualitative variables are identified that can influence the student to drop out. They are asked about personal aspects that could influence the decision to drop out of college. The questions were: due to lack of time, fail one or more subjects, due to personal problems, because you did not like the career, due to work, due to family impediment, health problems, or

Table 2. Figures of Entry and Exit of the Facultad de Contaduria y Ciencias Administrativas of the UMSNH. Last 4 School cycles (2015-2016 to 2018-2019).

Total: graduates	Total: entry	Average graduation	Average entry	Average graduation rate	Average dropout
2,832	5,563	708	1,392	51%	49%

Table 3. Question Options

No. option	Question
Opc1	Lack of time
Opc2	Fail to pass one or more subjects
Opc3	Due to personal problems
Opc4	Because you did not like the career
Opc5	By work
Opc6	Family impairment
Opc7	Health problem
Opc8	Economic problems
Opc9	Because you have another academic option (change university or a different bachelor)

economic problems, when you have another academic option (change degree or university).

4 Results

In the 2019-2020 school year in which the survey was applied to determine the causes of dropout in FCCA students, 3,290 were enrolled, and the survey was applied to 1,582 students. According to the determination of the sample size made, it was determined that the sample size is adequate. To perform the data analysis, the R Studio tool was used [10]. R Studio is the primary integrated development environment for R.

It is available in open source and commercial editions on the desktop (Windows, Mac, and Linux) and from a web browser to a Linux server running R Studio Server or R Studio Server Pro. In the survey applied to FCCA students, the question where the dropout factor is determined, quotes, "If you had to leave the University, what would be the reason? (You can select one or more)". The options are shown in the following Table 3:

One of the main contributions of this work is the decision tree model from a predictive method, shown in Figure 1, which from a certain behavior of the data tries to delimit the path through which it is crossed to reach a certain point, in addition, each option has a certain score, the answers selected will depend on determining which final score the respondent will obtain. The model logically predicts if a certain condition is met, the probability of deserting that would be obtained.

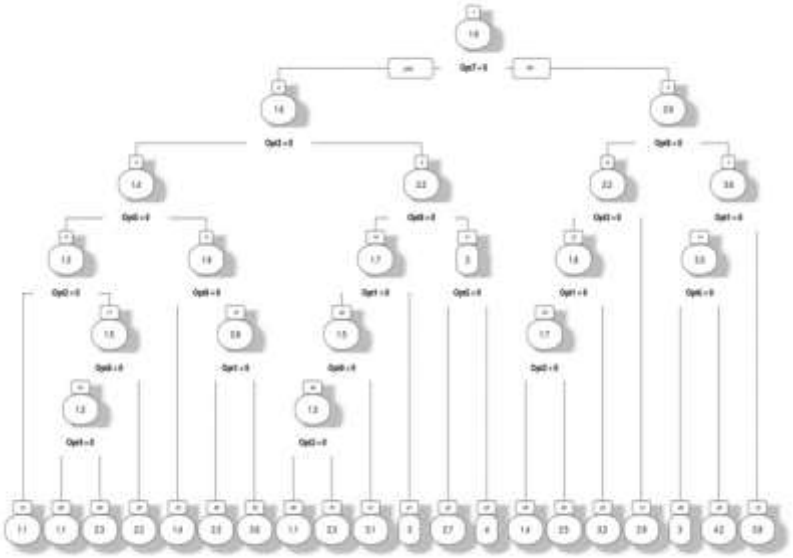

Fig. 1. Predictive Decision Tree Model to determine the dropout factor in the FCCA of the UMSNH.

Different personal situations can be represented in the students of the F.C.C.A of the Decision Tree Model, some of it could be the following: In the model indicates that, if the student does not have health problems, option 7, but if he has economic problems, option 8 and lack of time, option 1, the probability of deserting is the highest, it would have a score of 5.8. , now, if a student has health problems, option 7, personal problems, option 3, economic problems, option 8, and lack of time, option 1, their score is 3.

Another personal situation that would not considerably affect the decision to dropout would be that, if someone has health problems, option 7, personal problems, option 3, and option 2, which fails one or more subjects, the score will be 1.1. The options that a student can choose are very diverse, therefore, the scores obtained depending on the personal situation of each student.

5 Conclusions

This paper addressed issues that are currently of great concern in different areas of administrative, educational, and technological research, to try to solve common problems. While it is true that most Higher Education Institutions have serious problems, the directors of the Facultad de Contaduria y Ciencias Administrativas of the UMSNH must make decisions with the help of adequate research and technological tools such as data mining through the predictive model of decision trees where it hierarchizes the most significant characteristics that affect the level of student dropout.

Although funding problems may prevail at the UMSNH, late dropout rates must be avoided [11] to increase enrollment and decrease their dropout rates, otherwise, there will be negative consequences, which can impact not only on the aspect of FCCA enrollment but also on university enrollment and even more on the population in general.

References

1. Elias, T.: Learning analytics. Learning, pp. 1–22 (2011)
2. Orea, S. V., Vargas, A. S., Alonso, M. G.: Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos, 779(73), pp. 33 (2005)
3. Felizzola, H. A., Arias, Y. A. J., Pedroza, F. V., Pastrana, A. M. C.: Modelo de predicción para la deserción temprana en la Facultad de Ingeniería de la Universidad de la Salle. Encuentro Internacional de Educación en Ingeniería (2018)
4. Madrid-Echeverry, J. I.: Propuesta de un modelo estadístico para caracterizar y predecir la deserción estudiantil Universitaria. Escuela de Ingeniería de la Organización (2017)
5. Cadavid, D. V., Mendoza, A. M., Rodríguez, E. C.: Eficiencia en las instituciones de educación superior públicas colombianas: una aplicación del análisis envolvente de datos. Civilizar: Ciencias Sociales y Humanas, 16(30), pp. 105–118 (2016)
6. Noboa, C., Ordóñez, M., Magallanes, J.: Statistical Learning to Detect Potential Dropouts in Higher Education: A Public University Case Study. Learning Analytics for Latin America, 2231, pp. 12–21 (2018)
7. Jiménez, M. V. G., Izquierdo, J. M. A., Blanco, A. J.: La predicción del rendimiento académico: regresión lineal versus regresión logística. Psicothema, 12(Su2), pp. 248–525 (2000)
8. Vásquez-Velásquez, J., Castaño-Vélez, E. A., Gallón-Gómez, S. A., Gómez- Portilla, K.: Determinantes de la deserción estudiantil en la Universidad de Antioquia (2003)
9. Rodriguez, M. G. O.: Deserción de estudiantes de licenciatura de la UMSNH Análisis y propuesta de solución. Economía y Sociedad, 22(38), pp. 15–32 (2018)
10. Gandrud, C.: Reproducible research with R and R studio. CRC Press (2013)
11. Vélez, E. C., Gómez, S. G., Portilla, K. G., Velásquez, J. V.: Análisis de los factores asociados a la deserción y graduación estudiantil universitaria. Lecturas de economía, 65, pp. 9–36 (2006)