Data Mining to Identify Risk Factors Associated with University Students Dropout

Jesús Silva[1](✉), Alex Castro Sarmiento[2],
Nicolás María Santodomingo[2], Norka Márquez Blanco[3],
Wilmer Cadavid Basto[4], Hugo Hernández P[4], Jorge Navarro Beltrán[4],
Juan de la Hoz Hernández[4], and Ligia Romero[2]

[1] Universidad Peruana de Ciencias Aplicadas, Lima, Peru
jesussilvaUPC@gmail.com
[2] Universidad de la Costa, St. 58 #66, Barranquilla, Atlántico, Colombia
{acastrol0,nmarial,lromeroll}@cuc.edu.co
[3] Universidad Libre Seccional Barranquilla, Atlántico, Colombia
norka.marquezb@unilibre.edu.co
[4] Corporación Universitaria Latinoamericana, Barranquilla, Colombia
{wcadavid,hhernandez,jjnavarro,jdelahoz}@ul.edu.co

Abstract. This paper presents the identification of university students dropout patterns by means of data mining techniques. The database consists of a series of questionnaires and interviews to students from several universities in Colombia. The information was processed by the Weka software following the Knowledge Extraction Process methodology with the purpose of facilitating the interpretation of results and finding useful knowledge about the students. The partial results of data mining processing on the information about the generations of students of Industrial Engineering from 2016 to 2018 are analyzed and discussed, finding relationships between family, economic, and academic issues that indicate a probable desertion risk in students with common behaviors. These relationships provide enough and appropriate information for the decision-making process in the treatment of university dropout.

Keywords: Knowledge extraction process · Tutoring · Decision making · Data mining

1 Introduction

A recent study by the World Bank reveals that Colombia is the second country in the region with the highest university dropout rate. The list is led by Bolivia, Ecuador is third, and Panama is in fourth place. The research shows that 42% of people in Colombia, between the ages of 18 and 24 who enter the higher education system, drop out [1, 2].

According to UNESCO [3], one of the reasons for this behavior may be the rising costs of higher education in the country, which averages US $ 5,000 per student, becoming one of the most expensive educations in the continent, after Mexico and Chile. However, Colombia is also the country that offers the graduate the best return on investment in university studies [4].

Y. Tan and Y. Shi (Eds.): DMBD 2019, CCIS 1071, pp. 44–52, 2019.
https://doi.org/10.1007/978-981-32-9563-6_5

"This conclusion was obtained through an analytical model called Mincer regression, which shows how incomplete higher education generates 40% return on investment to the graduate, compared to not studying, while completing university education generates the professional in Colombia up to 140% return on investment, compared to another educational level" [5].

According to several authors [6–10], the student dropout profiles obtained through data mining techniques involving classification, association, and grouping, indicate that these techniques can generate models consistent with the observed reality and the theoretical support based only on the databases of different universities. This work aims to study a database regarding students' dropout in a private university in Colombia, using data mining techniques in order to identify patterns that allow the development of retention strategies.

2 Theoretical Review

2.1 Techniques to Reduce Student's Dropout

SINNETIC developed a statistical analysis of the scientific literature with more than 1,726 research papers associated with the subject to find best practices that reduce university dropout [11].

2.1.1 Centralized Evaluation

It consists of generating standardized evaluation resources (scientifically developed exams) applied by the higher education institution, making 70% or more of the grades depend on this evaluation format, leaving the teacher only 30% of evaluation capacity with mechanisms such as homework, short tests, etc [12].

In universities that have applied this mechanism, student's dropout has been reduced by up to 34%, with institutions from India and Spain showing the best figures in this regard [13].

2.1.2 Recreate Proficiency Markets Among Teachers

This measure implies that a subject matter can be taught by two or more teachers, generating competitiveness and recreating an incentive scheme where the teachers receive an economical income that can be improved by higher level of enrolled to their courses. To avoid an indulgent position from the teacher, the evaluation is centralized in the university and are elements such as reputation, good comments from the previous cohorts which determine the flow of students enrolled by subject, semester to semester [14, 15].

In institutions where this measure has been applied, dropout reduction reaches levels of 32% throughout the training cycle, although this measure has a greater impact on early dropout, understood as desertion during the first two years, when the results of reduction is around 44% [16].

2.1.3 Reduction of Regulations

More than 13 studies that report the number of standards, the linguistic complexity of the policies, the number of statutes or articles within documents such as research regulations, teaching regulations, etc., tend to be positively correlated with student's dropout. This correlation oscillates around 0.43 and 0.67 on a scale ranging from −1 to 1 [17].

Universities that reduce the regulatory burden and simplify communication through strategies based on behavioral economics have managed to reduce student´s dropout by up to 23% [7].

2.1.4 Differential Registration Schemes with Fiduciary Incentives

This mechanism consists in raising the barriers of dropout by means of economic incentives that consist of charging during the first four semesters more than 70% of the cost of the whole program under the promise that 20% of extra charge in the first semesters will go to a savings program (fiducia, CDT, etc.) in order to generate interests for reducing the costs of finishing a career. The student who drops out before completing 50% of the program will lose the interests and over cost. Evidence of this mechanism can be found in only two studies showing reductions of 21% in students' dropout [18].

2.2 Data Mining and Classifiers

The process of extracting knowledge from large volumes of data has been recognized by many researchers as a key research topic in database systems, and by many industrial companies as an important area and opportunity for greater profits [15]. Fayyad et al. define it as "The non-trivial process of identifying valid, new, and potentially useful patterns from the data that are fundamentally understandable to the user".

The Knowledge Discovery in Databases (KDD) is basically an automatic process in which discovery and analysis are combined. The process consists of extracting patterns in the form of rules or functions from the data, for the user to analyze them. This task usually involves preprocessing data, doing data mining, and presenting results [17–19]. The KDD process is interactive and iterative involving several steps with the intervention of the user in making many decisions and is summarized in the following five stages.

Selection Stage. In the Selection Stage, once the relevant and priority knowledge is identified and the goals of the KDD process are defined from the point of view of the final user, an objective data set is created selecting the whole data set or a representative sample on which the discovery process will be carried out [20].

Preprocessing/Cleaning Stage. In the Preprocessing/Cleaning stage (Data Cleaning), the data quality is analyzed, and basic operations are applied such as the removal of noisy data. Strategies are selected for the management of unknown data (missing and empty), null data, duplicated data, and statistical techniques for its replacement [21].

Transformation/Reduction Stage. In the data transformation/reduction stage, useful features are searched to represent data depending on the goal of the process. Methods for dimensions reduction or transformation are used to decrease the effective number of variables under consideration or to find invariant representations of the data [16]. Dimensions reduction methods can simplify a table in a database in a horizontal or vertical way. The horizontal reduction involves the elimination of identical tuples as a result of the substitution of any attribute value for another of high level in a defined hierarchy of categorical values or by the discretization of continuous values. Vertical reduction involves the elimination of attributes that are insignificant or redundant with respect to the problem. Reduction techniques such as aggregations, data compression, histograms, segmentation, entropy-based discretization, sampling, etc. [19] are used.

Data Mining Stage. Data mining is the most important stage of the KDD process [20]. The objective of this stage is the search, extraction, and discovery of unsuspected and relevant patterns. Data mining consists of different tasks, each of which can be considered as a type of problem to be solved by a data mining algorithm as Adamo and Hernández et al. assert, where the main tasks are Classification, Association, and Clustering [21, 22].

Data Interpretation/Evaluation Stage. In the interpretation/evaluation stage, the discovered patterns are interpreted, and it is possible to return to the previous stages for subsequent iterations. This stage can include the visualization of the extracted patterns, the removal of redundant or irrelevant patterns, and the translation of useful patterns in terms that are understandable to the user. On the other hand, the discovered knowledge is consolidated to be integrated into another system for subsequent actions, or simply to document and report it to the interested parties, as well as to verify and solve potential conflicts with previously discovered knowledge [23].

3 Materials and Methods

3.1 Database

The data used in this study was obtained from a student tracking system aligned with the policies of the Colombian Institute for the Promotion of Higher Education. The sample is made up of 985 industrial engineering students from a private university in Colombia during the time period 2016–2018.

The questionnaires and tests included in the system are briefly described below [24].

1. Questions about family, socioeconomic, and academic topics of the students are included in order to know their background. The information is stored in the system database and can be modified in each application of the questionnaires if there is any change in the student's situation.
2. Questions are included to know specifically if the students work, the kind of work, who or how they pay their expenses, who they keep informed of their studies, and their general health conditions.
3. The tests of self-esteem, assertiveness, learning styles, and study skills are questionnaires of closed questions to recognize problems related to organization of activities, study techniques, and motivation to study.
4. Ratings. Optionally, there is a section of grades where students capture the grade obtained by each learning unit of the different educational programs they attend.

3.2 Methods

The data analysis was carried out using two important data mining tools: clustering and association rules. Particularly, the K-means clustering algorithm and the A priori algorithm were used for the association rules. To carry out the data analysis process, the KDD method was used, consisting of the following steps [14, 20, 21, 25]:

1. Selection of data and analysis of their properties. Once the questionnaires are applied to the students and the information is in the database, the relevant data are defined to analyze the available information such as academic, family, and socio-economic background. In this step for association rules, specific data from the database were used, consisting in sport practice, economic problems, whether they work or not, high school average, interruptions in their studies, and the career study field. In the case of grouping, the data used consisted of information about the reasons to choose the career, medical treatments, economic dependencies (if they are married or have children), and knowledge of programs such as scholarships.
2. Data pre-processing. In this stage, the database is cleaned when looking for inconsistent, out of range, missing data, or empty fields to be later integrated and used in the analysis with data mining. The files created in this stage have the extension in .arff format for the analysis with Weka software, which is an application designed for the analysis of databases applying different algorithms for both supervised and unsupervised learning to obtain statistics and trend patterns according to the research objectives.
3. Application of data mining techniques. Once the files are obtained in .arff format, they are loaded to the Weka tool, and the option by which the information will be processed is selected. In the case of association rules, the *associate* option is chosen and then the a priori algorithm is selected and parameters such as the number of rules, the output, the minimum coverage, among others, are adjusted to later start the execution and obtain the patterns.

   In the case of clustering, a similar procedure is followed, selecting the *cluster* option, then the algorithm to be used, k-means in this case, and the parameters are adjusted considering the number of clusters in which the information is to be grouped and the iterations that it will have.
4. Interpretation and evaluation based on the results obtained from the previous phase. The “patterns” obtained from Weka are analyzed and evaluated if they are useful. This task is performed by the analyst.

4 Results

This section presents the main results obtained from the analysis of the system data studied through data mining techniques. It is divided into two subsections, one for the discussion of results obtained with association rules and the other one for the results obtained from the analysis by means of grouping tools.

4.1 Association Rules

In order to generate strong rules that exceed the support and minimum trust, the minimum support was established at 3% and confidence at 80%. 1,957 rules were generated, from which, the rules were chosen with 100% confidence. The most representative association rules are the following:

Rule 1. 100% of students who drop out are single, their grade average is less than 2.4, they have failed courses in the first semesters (1 to 4) and they failed these courses only once.
Rule 2. 100% of the students who drop out have completed their secondary studies in a public school, they are single, their average grade is less than 2.4, they have failed courses in the first semesters (1 to 4) and all of them failed these courses only once.
Rule 3. 100% of students who drop out are single men, their average grade is less than 2.4, and they are from a private university.

According to the above results, among the factors associated with student's dropout are: being single, having a low average, having failed courses in the first semesters, and coming from a public school.

For the analysis by means of association rules, the information of the database that corresponds to the family, economic, and academic background sections was used. Based on this information, the following conclusions were reached.

1. Students who usually practice some type of sport like soccer have economic problems and those who do not practice any sport activity do not present economic problems.
2. Most of the students who drop out have economic problems, which is why they should look for a way to earn an income and decide to go to work and later they decide to abandon their studies to continue obtaining the income.
3. Students who drop out have got a favorable average in their previous studies, which indicates that the cause of their desertion is not directly related to academic situations.
4. Most of the students who drop out have never interrupted their studies at primary and secondary levels.
5. Most of the students entering the engineering career have a misconception of the approach to the moment they decide to enroll.

Based on the results obtained with the use of association rules, a general view was obtained about the situation of the students, the student's community, the industrial engineering career, and the probable causes for leaving studying. For example, the contrast of opinions between students (men and women) and the relationship that exists between students who practice sports and their economic situation, gives a reference of how students can act and what they think. These factors can influence the low academic performance of students and can have an impact on the decision of dropping out. However, these results are preliminary, and a more extensive study is necessary in order to obtain more information to enhance the decision making processes in the academic community.

4.2 Clustering or Grouping

Once the application of association rules was completed, the k-means clustering technique was used on a broader set of data since new information was included from the application of the questionnaire and self-esteem, assertiveness, and learning styles tests, besides the registration of new grades. From the application of this algorithm, the following results were obtained:

1. From a total of 940 students, 321 of them think they had liked to study another career. From these 321 students, 150 prefer not to continue studying and start working to obtain an income, while the rest prefer to change to another career they like.
2. From the 70% of students who take medication, their parents have higher educational studies, so they have medical attention, but it must be followed up to see if it is cause of desertion. In this case, the interest is because 100% of students who take medication only 70% have medical care, while the other 30% do not have it which, in some cases, is not enough or their parents do not know about the problem.
3. Students who have someone who depends economically on them -married or working- are strong candidates to drop out due to their economic and family situation.
4. All the students surveyed say they know the field of knowledge of the career, and 93% of them have not been enrolled in another engineering. However, only 3% of the students have no reason to finish their studies, so the career was their first choice and they do not have another engineering in mind. In this way, the causes of desertion are due to social, economic, and personal causes, or that they do not have enough commitment to complete their studies.
5. Students who have another career in mind have difficulty in the school situation since they are in classes only because they have no other option at the moment but when their opportunity arrives they will decide to leave the career and enroll in the career they really want to study.
6. The economic situation of the student causes desertion when they do not know what scholarships the student can access either by the institution or by a government agency. This is the job of tutors or the staff of academic services of the institution, however, this issue is sometimes left aside, and not enough attention is paid to the publication of calls among students.
7. From the sick students, it is observed that the institutional insurance helps to cover the cases, and not having this benefit can make difficult to cover their needs, which leads to some students being studying for keeping the insurance and not because they really want to do it. This is reflected in the qualifications at the end of each semester.

From the results obtained in this analysis, as more data is obtained from the application of the questionnaires, the analysis becomes deeper and more precise information is obtained about the needs of the students. It represents a benefit for the decision-making processes that allow the permanence of the students in the institution and can complete the career in a satisfactory way.

5 Conclusions

The research focuses on the capture of information on students of industrial engineering in a private university in Colombia. The data are stored in the database to be later analyzed by data mining techniques to obtain common behavior patterns among students. The resulting information can help identify problems in the student community that affect their performance and cause problems both for the students who drop out and for the institution in the search for terminal efficiency.

The main contribution of the work discussed in this paper is that, in addition to systematizing the tutoring process, it supports the analysis of the data through data mining tools and pattern recognition. With the analysis of the information that was carried out, preliminary results were obtained on the factors that can influence the dropout and underperformance of the students of industrial engineering in Colombia. Among these factors are the social, economic, and academic ones.

For future researches, it is necessary to try other techniques such as expert system based on artificial neural networks to improve the results obtained so far.

References

1. Caicedo, E.J.C., Guerrero, S., López, D.: Propuesta para la construcción de un índice socioeconómico para los estudiantes que presentan las pruebas Saber Pro. Comunicaciones en Estadística **9**(1), 93–106 (2016)
2. Torres-Samuel, M., Vásquez, C., Viloria, A., Lis-Gutiérrez, J.P., Borrero, T.C., Varela, N.: Web visibility profiles of top100 latin american universities. In: Tan, Y., Shi, Y., Tang, Q. (eds.) DMBD 2018. LNCS, vol. 10943, pp. 254–262. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93803-5_24
3. Zhang, G.P.: Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing **50**(1), 159–175 (2003)
4. Duan, L., Xu, L., Liu, Y., Lee, J.: Cluster-based outlier detection. Ann. Oper. Res. **168**(1), 151–168 (2009)
5. Haykin, S.: Neural Networks a Comprehensive Foundation, 2nd edn. Macmillan College Publishing, Inc. USA (1999). ISBN 9780023527616
6. Haykin, S.: Neural Networks and Learning Machines. Prentice Hall International, New Jersey (2009)
7. Abhay, K.A., Badal, N.A.: Novel approach for intelligent distribution of data warehouses. Egypt. Inf. J. **17**(1), 147–159 (2015)
8. Aguado-López, E., Rogel-Salazar, R., Becerril-García, A., Baca-Zapata, G.: Presencia de universidades en la Red: La brecha digital entre Estados Unidos y el resto del mundo. Revista de Universidad y Sociedad del Conocimiento **6**(1), 1–17 (2009)
9. Bontempi, G., Ben Taieb, S., Le Borgne, Y.-A.: Machine learning strategies for time series forecasting. In: Aufaure, M.-A., Zimányi, E. (eds.) eBISS 2012. LNBIP, vol. 138, pp. 62–77. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36318-4_3
10. Isasi, P., Galván, I.: Redes de Neuronas Artificiales. Un enfoque Práctico. Pearson (2004). ISBN 8420540250
11. Kulkarni, S., Haidar, I.: Forecasting model for crude oil price using artificial neural networks and commodity future prices. Int. J. Comput. Sci. Inf. Secur. **2**(1), 81–89 (2009)

12. Mazón, J.N., Trujillo, J., Serrano, M., Piattini, M.: Designing data warehouses: from business requirement analysis to multidimensional modeling. In: Proceedings of the 1st International Workshop on Requirements Engineering for Business Need and IT Alignment, Paris, France (2005)
13. Jain, A.K., Mao, J., Mohiuddin, K.M.: Artificial neural networks: a tutorial. IEEE Comput. **29**(3), 1–32 (1996)
14. Kuan, C.M.: Artificial neural networks. In: Durlauf, S.N., Blume, L.E. (eds.) The New Palgrave Dictionary of Economics. Palgrave Macmillan, Basingstoke (2008)
15. Mombeini, H., Yazdani-Chamzini, A.: Modelling gold price via artificial neural network. J. Econ. Bus. Manag. **3**(7), 699–703 (2015)
16. Parthasarathy, S., Zaki, M.J., Ogihara, M.: Parallel data mining for association rules on shared-memory systems. Knowl. Inf. Syst. Int. J. **3**(1), 1–29 (2001)
17. Sekmen, F., Kurkcu, M.: An early warning system for Turkey: the forecasting of economic crisis by using the artificial neural networks. Asian Econ. Financ. Rev. **4**(1), 529–543 (2014)
18. Sevim, C., Oztekin, A., Bali, O., Gumus, S., Guresen, E.: Developing an early warning system to predict currency crises. Eur. J. Oper. Res. **237**(1), 1095–1104 (2014)
19. Singhal, D., Swarup, K.S.: Electricity price forecasting using artificial neural networks. IJEPE **33**(1), 550–555 (2011)
20. Vasquez, C., Torres, M., Viloria, A.: Public policies in science and technology in Latin American countries with universities in the top 100 of web ranking. J. Eng. Appl. Sci. **12** (11), 2963–2965 (2017)
21. Vásquez, C., et al.: Cluster of the latin american universities top100 according to webometrics 2017. In: Tan, Y., Shi, Y., Tang, Q. (eds.) DMBD 2018. LNCS, vol. 10943, pp. 276–283. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93803-5_26
22. Viloria, A., Lis-Gutiérrez, J.P., Gaitán-Angulo, M., Godoy, A.R.M., Moreno, G.C., Kamatkar, S.J.: Methodology for the design of a student pattern recognition tool to facilitate the teaching – learning process through knowledge data discovery (big data). In: Tan, Y., Shi, Y., Tang, Q. (eds.) DMBD 2018. LNCS, vol. 10943, pp. 670–679. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93803-5_63
23. Prodromidis, A., Chan, P.K., Stolfo, S.J.: Meta learning in distributed data mining systems: Issues and approaches. In: Kargupta, H., Chan, P. (eds.) Book on Advances in Distributed and Parallel Knowledge Discovery. AAAI/MIT Press (2000)
24. Savasere, A., Omiecinski, E., Navathe, S.: An efficient algorithm for data mining association rules in large databases. In: Proceedings of 21st Very Large Data Base Conference, vol. 5, no. 1, pp. 432– 444 (1995)
25. Stolfo, S., Prodromidis, A.L., Tselepis, S., Lee, W., Fan, D.W.: Java agents for metalearning over distributed databases. In: Proceedings of 3rd International Conference on Knowledge Discovery and Data Mining, vol. 5, no. 2, pp. 74–81 (1997)