IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 25, Issue 6, Ser. 1 (Nov. – Dec. 2023), PP 06-10
www.iosrjournals.org

Academic Dataset To Predict Dropout FromHigher Education

Cindy Espinoza Aguirre[1,2*]
[1]*Departamento de TI, Universidad San Francisco de Quito, Av. Diego de Robles, Quito, 170901, Ecuador.*
[2*]*Computer Science and Engineering Department, UniversidadCarlos III de Madrid, Avda Universidad 30, Leganes, 28911, Madrid, Spain.*

Abstract
In this research been applied predictive analytics techniques, and a KNN model has been generated to evaluate the pre- diction percentage. The KNN (K-Nearest Neighbors) algorithm was applied to predict the probability of dropout in each stu- dent and its accuracy was evaluated using a confusion matrix. Also, information was collected on the 12 relevant academic vari- ables for the study, including academic performance, the number of subjects taken, the number of electives, and selection, To pre- dict dropout, the number of neighbors was set to 4, since this refers to the number of closest data points that will be used to determine the classification of an unknown point. and compulsory subjects, and career changes that were excluded from the analysis. A training sample was used to train the model and a test sam- ple to validate its accuracy. The accuracy achieved is 0.97 how- ever, cross-validation techniques should be applied in order to assess whether the model has not fallen into overfitting or underfitting. Based on the analysis generated, an important finding is a rela- tionship between the number of compulsory subjects and its high relationship with university dropouts. In another manner the vari- ables a number of elective, selective, and compulsory subjects present a medium correlation with dropout, there is no differ- ence applied in their correlation, however, it has been shown that grades have a high correlation in relation to dropout.

Keywords: *dropout, higher education, machine learning, data mining,predictive patterns*

Date of Submission: 24-10-2023 Date of Acceptance: 04-11-2023

I. Introduction

Open education can reduce dropout since it provides the following character- istics: accessibility, flexibility, collaboration and participation, innovation, and access to specific skills for the new specialized professional profiles required by companies. By making education more accessible and relevant to a broader population of students, barriers and obstacles that can lead to school dropout can be reduced and students' academic and career success can be increased. Traditional education does not promote well-being over academic rigor and demands

Rigid curricular design, unable to adapt to the era of knowledge and arti- ficial intelligence. Open education will be an alternative that allows students without a specific education to achieve their academic purposes since they will be able to access knowledge and education without borders, inclusive, qual- ity, and easily accessible. Traditional university curricular design promotes desertion. Among the articles that show a failure in the incorporation of data analytics in administrative processes are: The expanse of data available today offers institutions new opportunities to assess, measure, and document learning [1] [2].

Universities do not consider the application of artificial intelligence tech- niques as a business strategy, missing opportunities for improvement. In general, the curricular design is not adaptable or changing, this behavior must be evaluated since at present the only constant is change.Apply data min- ing to analyze the university dropout rate of traditional education students. Finding the relationship between academic variables and curricular design and university desertion. Identify the relationship between university dropout and traditional curricular design using systematic review techniques. Evaluate the indicator of student retention in a population of traditional education students[3].

Hence the structure of the paper is as follows: section 1 Introduction, section 2 State of the Question, section 3 Methodology, section 4 Academic dataset, section 5 Correlation with student dropout, section 6 Results, section 7 Conclusion, section 8 future works and finally, section 9 figure, section 10

discussion and finally references.

II. State of the Question

For this article, a student may be considered a dropout when he has not completed his academic credits or has dropped out. In this study, a KNN (K- Nearest Neighbors) algorithm was used to analyze the relationship between university dropout and other relevant academic variables in a sample of 1622 university students. The Pearson correlation coefficient was used to evalu- ate the correlation between variables with the purpose of evidencing whether the students who took elective, elective, or compulsory subjects present any behavior that promotes student retention.

Before analyzing the relationship between curricular design and student dropout, it is important to understand that these terms are used as indica- tors of the quality of education, increase the reputation of the institution, save resources, and even define the quality of life of a population [4]. Open educa- tion promotes the construction of knowledge in a democratic and decentralized manner, in addition to providing the opportunity to promote quality educa- tion. There are several studies that demonstrate the advantages of education on the reduction of university desertion, among the most important are: [5] [6] [7] [8]

III. Methodology

For this research, the academic data lake worked from previous investigations has been taken, using an adaptation of the best practices KDD, CRISP- DM (Cross Industry Standard Process for Data Mining), SEMMA (Sample, Explore, Modify, Model, Assess), TDSP (Team Data Science Process), These are just some of the more common methodologies used in data mining. [9] Each of these methodologies has its own strengths and weaknesses, therefore in previous studies, the collection of each of the best practices applied to the problem of university dropout was detailed in order to generate data sets readyfor use and application.

Therefore, the difference of this research lies in the analysis of academic data and its relationship with the behavior of dropout students. Hence, Pearson has been applied. In this sense, Pearson's correlation is a measure of the linear relationship between two continuous variables. In this sense, This measure ranges from -1 to 1, where -1 indicates perfect negative correlation, 0 indicates no correlation, and 1 indicates perfect positive correlation. This research has been classified into three levels: high >= -0.5 or 0.5, medium = -0.5 or 0.5, and low <= -0.5 or 0.5. In this way, it will be possible to identify the variables that correlate with student desertion [10].

In addition, the KNN algorithm has been used because the KNN algorithm is often used in classification problems where objects need to be labeled into different categories. KNN can be a good option in this situation because it is able to classify objects based on their similarity to other known objects withoutmaking assumptions about the data distribution. Also, The KNN algorithm

IV. Academic DataSet

It is vital to study the academic variables in the student dropout problem because they show the factors that contribute to a student's decision to drop out. The academic variables include academic performance, the number of credits approved, the number of elective subjects, the number of selective subjects, and the number of compulsory subjects, among others.

These variables can indicate underlying issues such as excessive academic load or problems with course content. They can also point out external prob- lems that may be affecting the student. Therefore, by studying academic variables, patterns, and trends can be identified that allow educational institu- tions to take steps to address these issues and improve student retention and success. The 12 variables used in the academic data set are detailed below. (refer to Tables 1)

Table 1: Detail Academic information

Data Set	Num.	Variable name	Description
Academy	1	a1	Periodo
Academy	2	a2	Semester Age
Academy	3	a3	Final Note
Academy	4	a4	Semester number
Academy	5	a5	Mandatory subject
Academy	6	a6	Selective subject
Academy	4	a4	Total credits
Academy	5	a5	Number mandatory subject
Academy	6	a6	Number elective subject
Academy	4	a4	Number selective subject
Academy	5	a5	Last semester

Academy	5	a5	Student dropout

[1]The data set that has been used is listed: academic information, academic performance

V. Correlation with student dropout

Pearson's correlation coefficient would be applied to correlate variables when one wants to investigate the relationship between two quantitative and linear variables. If a linear relationship between two variables is suspected, the Pear- son correlation coefficient can provide a quantitative estimate of the strength and direction of the relationship.

As can be seen in figure 1 , the variables that present a high correlation are: final note, semester number, selective subject, total credits, number mandatorysubject, number elective subject

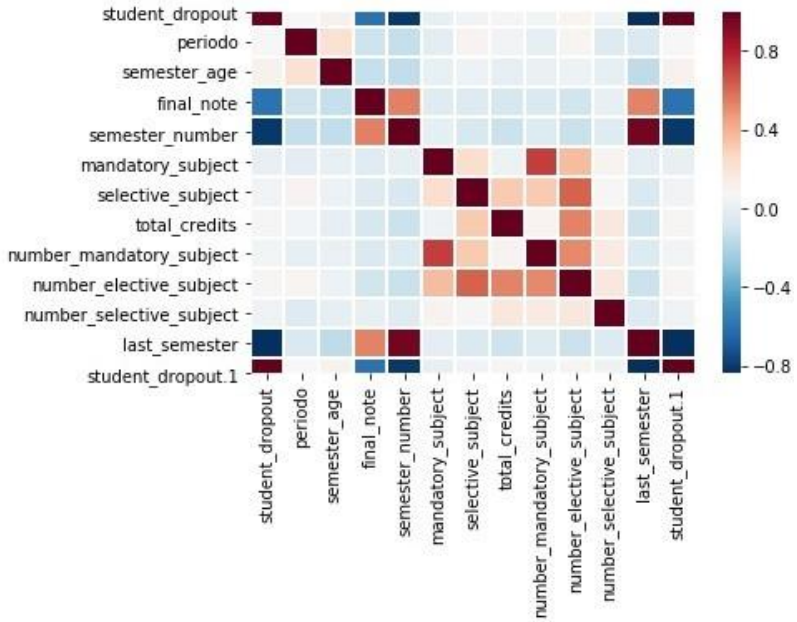

Fig. 1: Heat map obtained with the selected variables, using Pearson

VI. Results

In the case of college dropout, KNN could be used to identify patterns in student data and predict whether or not a student will drop out of college. The algorithm would rank students based on their similarity to students who have dropped out in the past [11] [10].

In addition, KNN is a non-parametric algorithm, which means that it does not assume a particular distribution of the data, which makes it useful when all the variables that influence university dropout are not known.

In the KNN classification 2 chart, each point in the plane represents an instance in the test data set, and its color indicates its actual class. The lighter- colored areas represent the class predicted by the KNN model. For the model we used (K=3) parameterized in the model, therefore the decision borders can be more or less smooth

However, it is important to note that KNN may not be the best approach to all college dropout issues. Choosing the appropriate algorithm based on the complexity of the data set and the characteristics of the students. Once the KNN model is run, the following metrics were obtained.(refer to Tables 2).

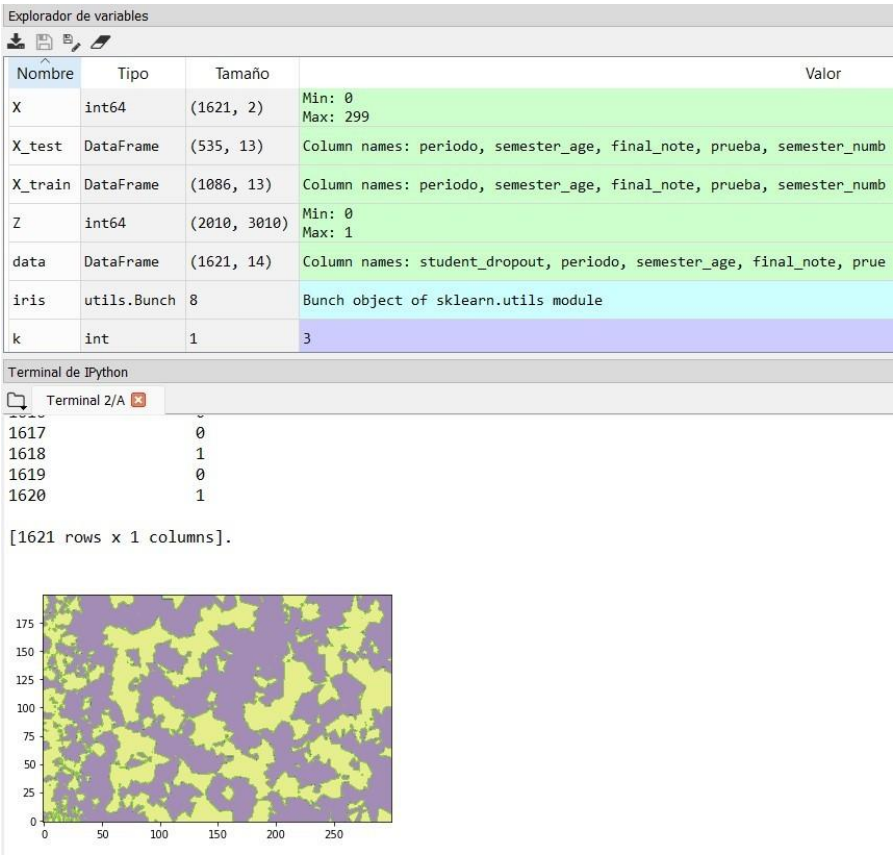

Fig. 2: Dropout Analysis Abstract Model

Table 2: Analyze the results of machine learning

Model	Total Variable	% Prediction	Accuracy	Loss
KNN	12	92%	0.978	0.1052

Note:

[1]Applying machine learning, a loss of 0.10 and 0.9178 of accuracy

VII. Conclusion

It is worth mentioning that one of the limitations of this work was not con- sidering enough academic variables that affect university dropout, which may limit the accuracy of the model.

The results obtained from the correlation show a medium and high level for the academic variables: final note, semester number, selective subject, total

Future works and Finally

For future work in order to corroborate the advantages of education over stu- dent retention, all relevant variables should be considered, properly adjusting the parameters of the model. However, despite having applied machine learning techniques to a traditional population, it is essential to evaluate the behavior of students who are trained under an open education modality in order to find metrics related to permanence and dropout.

References

[1]. Alexander, B., Ashford-Rowe, K., Barajas-Murphy, N., Dobbin, G., Knott, J., Mccormack, M., Pomerantz, J., Seilhamer, R., Weber, N.: EDUCAUSE Horizon Report: 2019 Higher Education Edition

[2]. Quiroz-Fabra, J., Bran-Piedrahita, L., Valencia-Arias, A.: Tendencias Investigativas Alrededor De La DeserciOn Estudiantil En Contextos De Edu- CaciOn Superior: Un AnÁlisis BibliomÉtrico (30), 83–100. Https://Doi.Org/ 10.18601/16577175.N30.06. Number: 30. Accessed 2023-05-12

[3]. Cajahuanca, J.E.V., Raymundo, F.N., Franco, A.C.L., Flores, J.D.J.: DeserciOn Universitaria: EvaluaciOn De Diferentes Algoritmos De Machine Learning Para Su PrediccIOn XXVIII(3). Accessed 2023-05-15

[4]. Esteban GarcÍa, M., Bernardo GutiÉrrez, A.B., RodrÍguez-Muñiz, L.J.: Permanencia En La Universidad: La Importancia De Un Buen Comienzo 44(1), 1–6. Https://Doi.Org/10.1016/J.Aula.2015.04.001. Publisher: Else- Vier. Accessed 2023-05-15

[5]. Vinuesa, T.S., Zermeño, M.G.G., LeOn, N., De Marzo: Doctorado En EducaciOn Y TIC (E-Learning) Universitat

Oberta De Catalunya (UOC)
[6]. Quintero-Guasca, R.E., Avellaneda-Nieves, M., Cristancho-Garc´Ia, M., SÁnchez-Medina, I.I., Quintero-Guasca, R.E., Avellaneda-Nieves, M., Cristancho-GarcÍa, M., SÁnchez-Medina, I.I.: Permanencia Estudiantil En Programas De Posgrado E-Learning: Un Caso De Estudio 14(3), 17–24. Https://Doi.Org/10.4067/S0718-50062021000300017. Publisher: Centro DeInformaciÓn TecnolÓgica. Accessed 2023-05-15
[7]. Quintero-Guasca, R.E., Avellaneda-Nieves, M., Cristancho-Garc´Ia, M., SÁnchez-Medina, I.I., Quintero-Guasca, R.E., Avellaneda-Nieves, M., Cristancho-GarcÍa, M., SÁnchez-Medina, I.I.: Permanencia Estudiantil En Programas De Posgrado E-Learning: Un Caso De Estudio 14(3), 17–24. Https://Doi.Org/10.4067/S0718-50062021000300017. Publisher: Centro DeInformaciÓn TecnolÓgica. Accessed 2023-05-15