Maximizing Student Retention using Supervised Models Informed by Student Counseling Data

John Anderson Rodriguez Ramirez[1,2,*], Olmer García-Bedoya[2] and Ixent Galpin[2]

[1]*Universidad de Los Andes, Bogota, Colombia*

[2]*Universidad de Bogota Jorge Tadeo Lozano, Bogota, Colombia*

Abstract

Student retention is one main challenge faced by higher education institutions in Colombia. Over recent years, there has been an increasing trend of students dropping out of university for various academic, social, economic, institutional, or personal reasons. This has significant impacts on private higher education institutions, whose main livelihood depends on student numbers, with some institutions in a critical situation. In this paper, we design and implement a data analytics model that predicts the risk of dropout that students may present. The sources of information used to construct this model contain academic and socioeconomic variables, covering past academic periods to understand the problem in detail. Machine Learning approaches are used to determine the best model using the CRISP-DM methodology. This model is deployed via a decision-making tool for the academic counseling office, whose main objective is to maximize student retention.

Keywords

Student Retention, Machine Learning, Counseling, Higher Education

1. Introduction

Student dropout is one of the biggest threats that Higher Education Institutions (HEIs) are increasingly facing in Colombia. According to studies and analyses carried out by the Ministry of Education, the figures increase year-on-year, and there are many variables that are relevant during analyses that enable decision-making that aim for student retention to be maximized for an HEI [1].

At the national level, the Ministry of Education has a Cox Model to monitor this phenomenon of student desertion called SPADIES (abbreviation in Spanish for *System for the Prevention of Desertion in Higher Education Institutions*), which allows the state and evolution of the academic performance of the students to be known, by carrying out an analysis of both academic and socio-economic variables prior to the student's admission to the University. According to statistics provided by the Ministry of Education, for 2018 the dropout rate for university degree courses

ICAIW 2022: Workshops at the 5th International Conference on Applied Informatics 2022, October 27–29, 2022, Arequipa, Peru

[*] Corresponding author

ja.rodriguez@uniandes.edu.co (J. A. Rodriguez Ramirez); olmer.garciab@utadeo.edu.co (O. García-Bedoya); ixent@utadeo.edu.co (I. Galpin)

0000-0003-3903-813X (J. A. Rodriguez Ramirez); 0000-0002-6964-3034 (O. García-Bedoya); 0000-0001-7020-6328 (I. Galpin)

in Colombia was 8.79%. This is slightly lower compared to the figures for HEI technological and technical courses, whose dropout rates were 10.75% and 17.41% respectively. Nevertheless, the figures are worthy of concern. It is important to bear in mind that the conditions faced by each student vary depending on the institution in which a student is enrolled. Furthermore, diverse variables may influence student decisions with regard to dropping out of an HEI program, and these may be individual, academic, socio-economic, and institutional [2].

We work with an anonymized data set made up of eleven variables and 77,073 records, which correspond to information stored in a Customer Relationship Management (CRM) system held by an HEI in Colombia. The data set spans six semesters and contains data that was not previously stored in a database and is deemed to contain information useful for predicting student dropout. Such data includes information about attendance at student counseling sessions at support centers. Such is expected to contribute to a great extent within the models to allow the student academic situations for be better understood.

In this work, two approaches for predicting student retention are proposed, making use of supervised Machine Learning techniques. For the first model, we work with the complete data set which has a history of six semesters. For the second, only the last two semesters are used. Five different supervised learning techniques are employed: Naive Bayes, Decision Trees, Random Forests, Logistic Regression, and XGBoost. We evaluate model performance to determine the combination that delivers the best precision in the predictions made.

This paper is structured as follows: Section 2 presents related work. The remainder of the paper broadly follows the steps of the CRISP-DM [3] methodology commonly used in data analytics projects. We present the steps corresponding to business and data understanding in Section 3. Section 4 describes the data preparation undertaken prior to the modeling presented in Section 5. We present the results of our evaluation in Section 6. Section 7 presents a discussion, including considerations for a future deployment, and Section 8 concludes.

2. Related Work

Since the earliest works that focus on the phenomenon of student dropout, there has been a broad consensus on defining it as the voluntary abandonment of studies by students. However, different entities interpret this concept depending on their perspective or needs [14]. Conversely, other work argues that it is paramount to differentiate between voluntary and involuntary dropping out, where the former is understood as the renunciation of an HEI program for personal, social, or economic reasons, which is reported to the institution. The latter is directly related to institutional decisions due to poor performance or lack of student discipline [15].

Numerous studies have been carried out on this subject, and different approaches that each one can give to the subject matter. However, most of this work body of work points to common groups of variables that intervene and are decisive in student desertion, which are individual, academic, institutional, and socio-economic [2]. However, a study carried out by the Nueva Granada Military University in Colombia considers other factors that affect university student dropout [16], such as demotivation, the influence of minority groups, crushes, problems with relatives, etc.

Arguably all of these variables are significant. However, it is also necessary to evaluate the

Table 1
Summary of related work on the student dropout phenomenon

Source	Variables considered	Techniques used	Time window considered
[4]	Socio-economic, academic, disciplinary and institutional variables data about the students	Classification, Clustering, Association Rules	3 years (2004–2006), years until 2011 as future work
[5]	Pre-university variables (sociodemographic and academic) such as school grades, admission grade, subjects taken, subjects failed	Logistic Association, Logistic Regression	Dataset of 13 first-semester cohorts corresponding to new students enrolled 2006–2013
[6]	Academic, socio-economic and institutional variables	Descriptive statistics	Semesters from 2003–2006
[7]	Gender, age, council tax band, whether the student was working when presenting University entrance exam, whether the family owns a home, educational level of the mother, academic program	Descriptive statistics	2009–2013
[8]	Academic program, enrolled courses, average grade, current semester, gender	Naive Bayes, Maximum Entropy, kNN	Unspecified
[9]	Degree of financial independence, fees balance with HEI, disciplinary data, number of subjects enrolled, passed, failed, failed twice, failed thrice, etc.	Accuracy Score, Decision Trees, Random Forests, XGBoost	2016 – 2019 (Approximately 12,000 records)
[10]	Age, gender, disability, ethnicity, immigration status, school of origin, parental level of education, type of home, scholarships, family financial support, working conditions, extracurricular activities, university entry grade, high school GPA, number of subjects taken/failed, attendance	KNN, Decision Trees, Random Forest, Support Vector Machine, Neural Networks	Unspecified
[11]	Age, gender, place of origin, employment status, How HEI program was chosen, admission grade, income source, University financing method, academic performance	Random forest, Decision Trees Nearest Neighbors Algorithm, Logistic Function Algorithm, Multilayer Perceptron	2014-1–2018-1 giving a total 12,698 records
[12]	academic performance and assistance to classes	Random forest, Decision Trees, Logistic Function Algorithm, Naives Bayes	2017-1–2021-1
[13]	Unspecified	Classification, estimation, cluster and association rules	3,204 students admitted 1999–2004

conditions of the students depending on the entity in which they are pursuing their course of study. For example, students experience different situations depending on whether the HEI is public or private, especially with regard to social issues. For example, in a study carried out at a private University in Medellín, Colombia, it was determined that most of the students who drop out do so in their first semester. Furthermore, it was found that 35% of students drop out in their fourth and seventh semesters (assuming a course of study comprising nine or ten semesters) [17]. These finds provide important insights for the work carried out in this paper.

Table 1 presents a comparison between different studies carried out that focus on the student dropout phenomenon. Despite being from different countries, institutions, and periods of time, it is possible to observe the coincidences between the study variables and the methods used. This highlights the importance of addressing this phenomenon globally, and also, how data and technology have been contributing more with their advances and models to the understanding of this phenomenon.

3. Business and Data Understanding

The initial objective was to recognize each factor involved in risk analysis within the academic stage. Within the institution, an approach to Student Success is managed through five strategies to support the students to achieve their objectives and goals within the university. These strategies include work with teachers, the culture of student success, support focused on first-year students, policies of the institution, and support for decision-making based on technology.

The university focuses its efforts on the type of population. One of the populations is first-year students, that is, students who are in their first or second semester. This is because studies within the same university have shown that the issues of adaptation to change and the transition from school to university are factors that significantly affect the academic performance of students and lead to possible early dropouts. The second population for which special care and analysis is taken within the entity is students at academic risk, those whose cumulative general average is less than 3.4/5.0. Due to their poor academic performance, these students fall into this zone in which, if their average does not improve in the following semester, they can be suspended and possibly excluded from the university. These groups should have to take care of this work.

Finally, we should be able to find and detect these cases in the early stages. Reaching these students and providing all the support offered on time, among which there are two important ones within the study: academic counseling and attendance at support centers, can help significantly reduce student dropout.

3.1. Data Set Description

The university has a CRM (Customer Relationship Manager) tool for its Student Success project. This software is used to obtain a broad picture of the current academic status of the students by making use of alerts and scores that are automatically calculated by the CRM and that later allow different actions or tasks to be carried out for the benefit of the students. In addition to the CRM information fed synchronously every day by the university's primary source of information, the users (professors and academic coordinators) have the opportunity to create new records of information, such as counseling. Additionally, to provide an overview of the academic status, other imports of information from different sources have also been added, which are not handled by the academic software of the university and are useful in the calculations and evaluations carried out by the tool.

The CRM data is the source of the data set used in the work. A history of six academic periods, from 2018 to 2020 (each year with two periods), from which it was possible to extract a total of 77,073 records, was obtained. Each record comprises eleven variables about each student, corresponding to a particular academic period. The data set has a total of eleven variables which are described as follows:

- **FACULTY:** The faculty to which the student belongs in his first program. There may be students undertaking more than one program, but for this model, only the first will be used.
- **PERIOD:** This variable identifies the academic period from which the student's data is taken.
- **CODE:** A unique identifier within the university for each student. This enables records from different academic periods to be cross-referenced.
- **ACADEMIC_STATUS:** This variable shows the Academic Status with which each student ends the semester. Although there are more than eight states, for this study we group them into two states. These are *Normal* for students that do not exhibit any risk of dropping out, and *At Risk* otherwise.
- **SPADIES:** This field stores the information on the score assigned to students when they enter the university, which defines their dropout risk in scores from 0 to 5. It is a platform developed to analyze and monitor student dropout from higher education in Colombia. Through the analysis of different risk variables, it is possible to know statistical data and trends to identify the factors or reasons why more dropouts are presented at universities. Worked for different populations that are involved in the entire educational process, such as students, directors, teachers, researchers, and government agencies, and that, by allowing students to know the level of risk, facilitates institutions the work of consultation and evaluation of strategies to reduce student dropout. The model variables for calculating the risk level of each student are:
  - Gender (Male, Female)
  - Student Age
  - Parents' income
  - Education level of the mother
  - SABER 11 test result
  - Number of classmates with whom you enter the same University
  - Number of siblings
  - Credits enrolled for your first semester
- **FUNDING:** This variable shows whether or not the student had financing during the academic period. Financing is understood as any economic support that she may have received officially to pay some percentage of her tuition. The variable is taken as a binary field where 1 represents yes and 0 represents no.
- **COUNSELING:** This variable contains the information about whether or not the student had academic counseling with her advisor professor in each period or semester. The variable is taken as a binary field.

- **CREDITS:** This value shows the number of credits each student has enrolled in a particular academic period.
- **SUPPORT_CENTER_ATTENDANCE** This field records whether or not the students have attended any of the Support Center support that the university has, these places offer tutorials, accompaniment, and assistance, for workshops and preparation of midterms. The variable is taken as a binary field.
- **WITHDRAWALS:** This variable shows whether or not the student has withdrawn subjects in the academic period, regardless of whether it was one or several, the variable is taken as a binary field where 1 represents Yes and 0 No.
- **GRADE_AVERAGE:** This variable refers to the average obtained by the student in the semester.

3.2. Descriptive Analysis

In Figure 1, an analysis is carried out between the variables COUNSELING, ACADEMIC_STATUS and SPADIES. On the x-axis shows the different States of ACADEMIC_STATUS, on the y-axis you can see the number of students that have been assisted via counseling. Each color represents a score of the SPADIES variable. With these results, we can conclude that the students who have a SPADIES equal to 5 (purple), are the ones who make the most use of the Counseling resource at the University, which is very much in line with the strategies of the University since according to this model they are who have the highest risk of desertion and are the ones who most seek support.

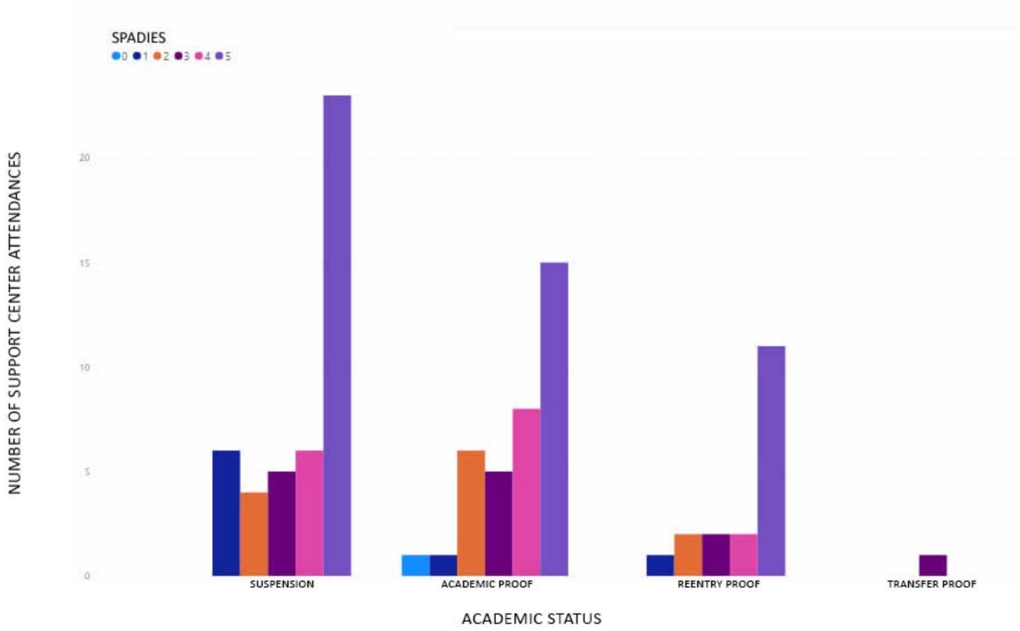

Figure 1: Attendance at counseling sessions vs. academic status vs. SPADIES

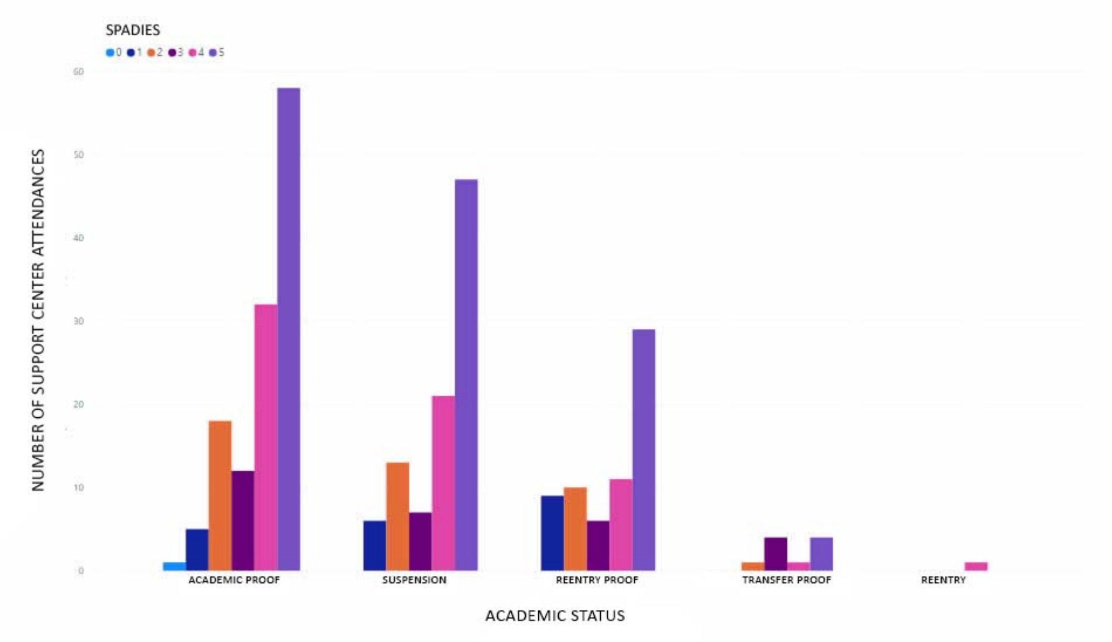

Figure 2: Attendance at counseling sessions vs. academic status vs. SPADIES

In Figure 2, the same comparison is made, between `SPADIES`, `ACADEMIC_STATUS`, but now with respect to the variable of assistance to the University Support Centers. Here we can find trends similar to those observed with the Councils, where the students who have their `SPADIES` score equal to 5 are also the most recurrent in these spaces. This again shows us the use of support resources by the students with a higher risk of desertion according to `SPADIES`.

In Figure 3, you can see a comparison between the number of credits enrolled by students and the number of visits for counseling. A trend is evident among students who enroll between 15 and 24 credits, who are the ones who attend counseling the most. This group is above average attendance to counseling sessions, which is highlighted in the graph with the red line and which corresponds to a value of 139 counseling sessions. With this, we can assume that students with low or very high academic loads in each semester, do not attend counseling sessions as often, as those who are around the average academic load at the University which is in the 17 credits. The highest points of attendance occur with students who have 18, 19 and 20 credits enrolled, which already exceed the normal academic load at the University, and this support is deemed necessary in order to successfully complete their semester. We found a similar behavior when we make the comparison between the withdrawals of subjects made by the students with respect to the number of registered credits. This can be observed in Figure 4, were between 13 and 22 credits the higher number of withdrawals. It is something to understand since the higher the academic load, the more students may require support or decide to withdraw from the courses instead of failing them.

In Figure 5, the number of counseling sessions is compared with the total number of records per faculty, finding that the number of students who use this service in each faculty is low with

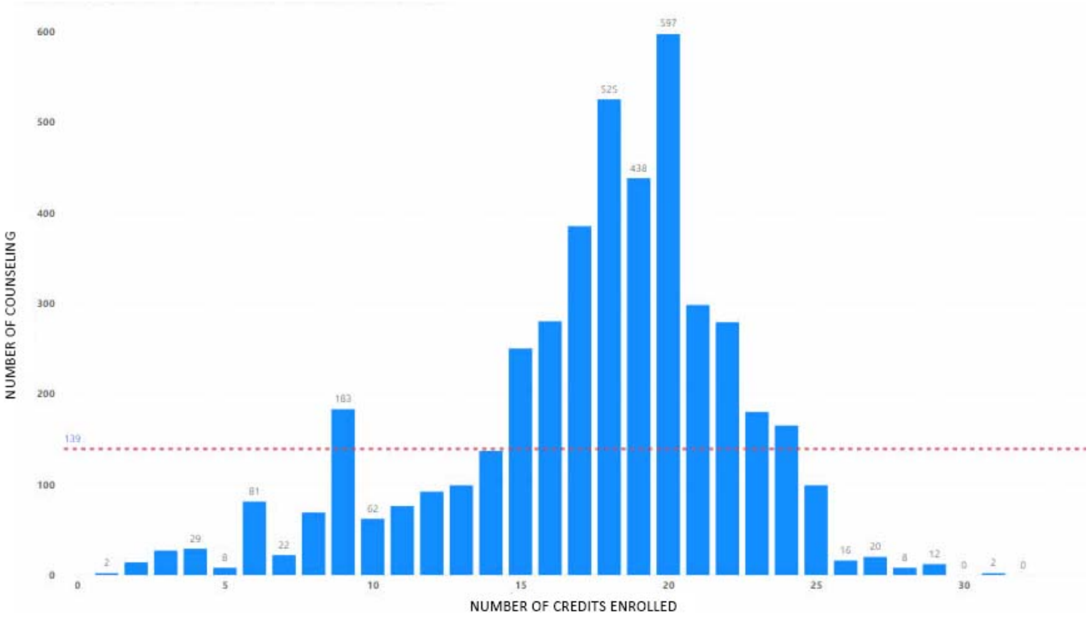

Figure 3: Number of counseling session vs. number of credits enrolled

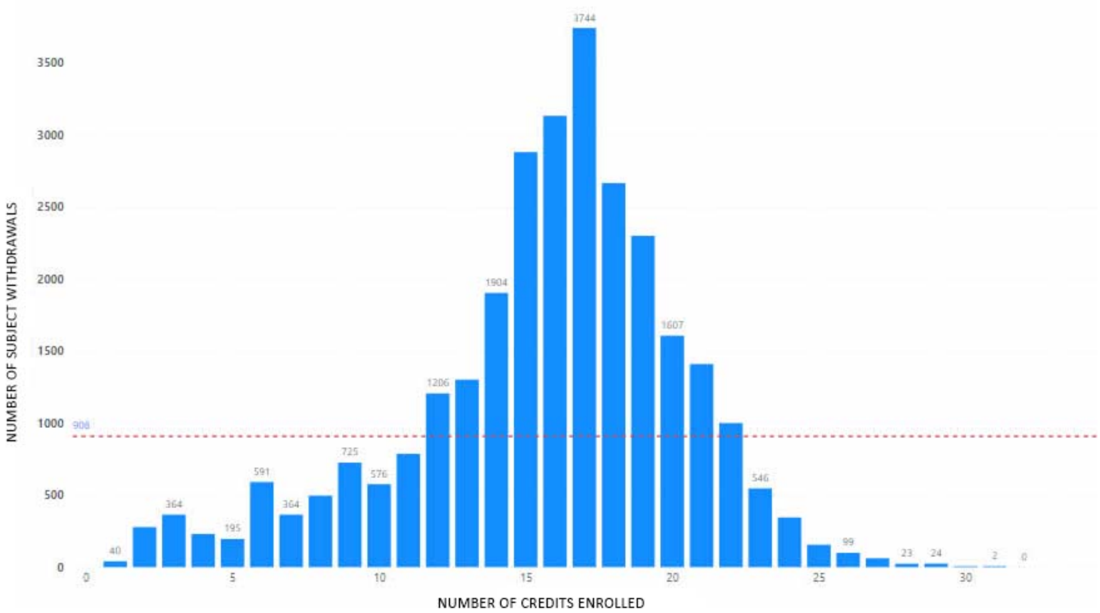

Figure 4: Number of subject withdrawals vs. number of credits enrolled

respect to the total number of records. Recall that each of the records corresponds to a student in a certain academic period, therefore, the same student may appear differently in each of the

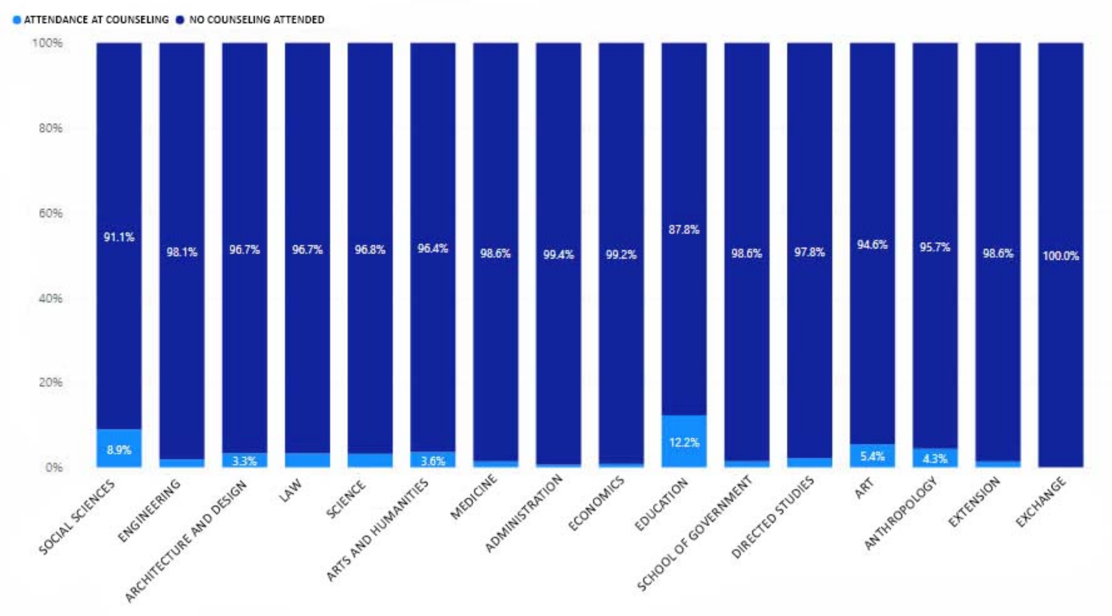

Figure 5: Counsoling percentaje – Faculty

six periods evaluated. Both faculties that register the highest number of counseling sessions are Education with 12.2%, which represents 67 records of its 547 Total records. Social Sciences exhibits 8.9% attendance, which represents 1,374 of its 15,402 records.

The descriptive statistics show that the Faculty of Engineering shows the largest number of records with 25,677 records. The SPADIES variable is only found for 62,386 records of the 77,073 total and according to the distribution given to students, handles a range of 0 for lower and 5 for higher risk of dropout. The average obtained is 3,064 and a standard deviation of 1,466. Additionally, the Registered Credits have an average of 16.73, a consistent value since it is below 17 credits, which is what the University considers a student with a high academic load. Regarding the variable of the Semester Average, the average of 3.466 and the standard deviation of 1.51 shows us the range in which most of the averages are found for the students' semester. To better understand these statistics, Figure 6 includes all the quantitative fields. In the result, some atypical data are observed in the Credits and Average variables. In the case of Credits, there are cases of students who enrolled more than 28 credits in any of the periods, which is a fairly high number, taking into account that in all programs except Medicine, all students with more than 17 credits are deemed to have a high load.

To find out correlations between variables, the Pearson and Spearman methods were employed. The highest correlation is found with the Average and Credits variables with values of 0.2337 with Pearson and 0.2108 with Spearman, which considering the ranges is not a significant value. In the case of Credits and Withdrawals, a value of -0.1567 is found with Pearson and -0.2364 with Spearman, which indicates a negative correlation. The others are below 0.01 and do not show a strong correlation between variables. In general terms and according to the results obtained by these two methods, we can conclude that a strong correlation between variables

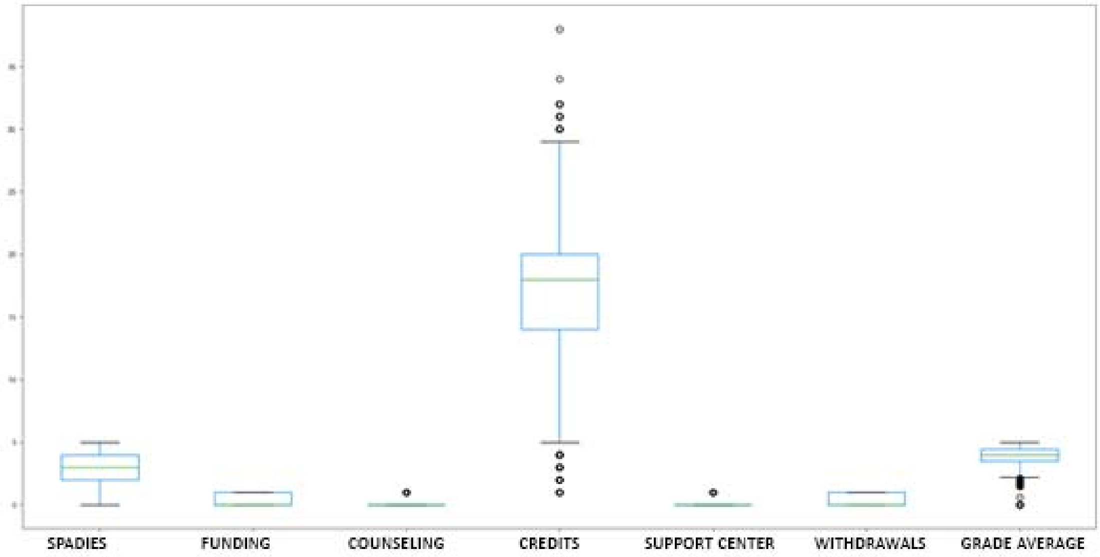

Figure 6: Box graph of the data

cannot be found, which is why it is important to take them all into account for the study.

4. Data Preparation

The data set contains a list of student records, with their respective academic programs, including whether they are majors or minors. In order to facilitate the work and the analysis, it was necessary, as a first step, to work only with the first program of each student, in order to avoid duplicate information in the records. The programs are also grouped by faculty. For example, the nine engineering programs were grouped together, as done with other Faculties such as Sciences. This reduced the group of more than 30 programs to 17 faculties.

Whilst reviewing the data, missing data was detected in the GRADE_AVERAGE and SPADIES columns. In order to preserve the maximum number of records possible, these missing values were imputed using the average values of each field. This was done with the intention of keeping as many records as possible, so as to improve model performance. The ACADEMIC_STATUS variable contains all the possible statuses that a student can have at the University. After discussions with the management of the counseling support centers, it was agreed to only have two states: `Normal` and `At Risk`. For all models proposed in Section 5, the variables CODE is discarded as it is the primary key and does not provide useful information for the model predictions.

5. Modeling

Two approaches were explored. For the first approach, we worked using the complete history available for a student, whereas, for the second approach, only the last two semester's worth of data are considered.

5.1. Approach 1: Using the full history (all semesters)

For our prediction models, the variable `ACADEMIC_STATUS` is determined to be the output (predicted) variable. The models are trained using 31 input variables. The first attempt was to attempt a Logistic Regression model. This model predicts 100% of cases of `Normal` status (i.e., 37,613 true negatives) correctly. However, it is not able to predict any students who could remain in the `At Risk` state, meaning that there are in effect 924 false negatives. As such, it is concluded that it is not a viable model for use. To remedy this, data balancing is carried out. The number of cases that present a Risk of Academic Dropout is a minority, only 2.34% of the record set. Taking this into account, the Imbalanced [18] technique and the artificial samples strategy, also known as Oversampling, are employed. This goal is to create new synthetic samples for the minority variable, which in this case is that of students at risk of dropping out. With an adjustment of 0.8, a more balanced data set is achieved (see Figure 7), with which it is already possible to start working with the proposed models.

Once the data has been balanced, the training process is carried out with various supervised learning techniques (Logistic Regression, Naive Bayes, Decision Trees, Random Forests, and XGBOOST) using cross-validation. The Accuracy metric is used, recommended when the data has been properly balanced. The results show that the Decision Tree and Random Forest models are the ones that deliver the best results with 0.97 and 0.98 respectively, and exceed the estimated percentage for the desired model. The XGBOOST model delivers a result of 0.85 does not exceed the expected percentage. The Naive Bayes and Logistic Regression models are not as effective,

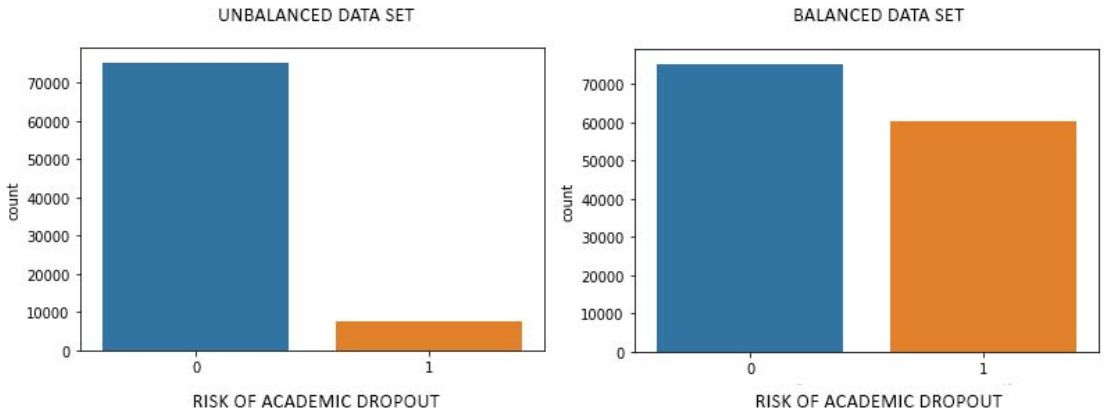

Figure 7: Balanced vs. Unbalanced data set

since their results only reach a 0.45 and 0.7 level of precision, respectively.

5.2. Approach 2: Using only the last two semesters

For this model, the records are grouped by Student Code, to subsequently concatenate them into a single record. Students with less than two academic periods in their history are discarded. In the grouping process, it is necessary to eliminate those fields that are the same in all academic periods and that will not be useful in the evaluation of the models, for example, `FACULTY`, `SPADIES`, `PERIOD` and `CODE`.

Only 2.17% of records have the `Academic Status` as `At Risk`, a minority of the data with with respect to the 97.8% that is `Normal`. As such, the process of balance with Oversampling set to 0.8 is carried out. It is found that the accuracy of the Decision Trees is 0.99, Random Forests is 0.99 and XGBOOST is 0.97. In the case of the models with Naive Bayes and Logistic Regression, they are again the ones with the lowest prediction scores with 0.51 and 0.84 levels of accuracy respectively. However, it is interesting to observe that the second approach improves the results significantly.

6. Evaluation

For the evaluation process of the models, we evaluate the test data and take into account as the Recall of the model we aim to able to identify cases of students who are in an `ACADEMIC_STATE` of being `At Risk`. As such, we are interested in the model being able to minimize cases where the minority classes are missed.

6.1. Approach 1: Using the full history (all semesters)

The evaluation of Approach 1, carried out with the set of data destined for the test, shows us a total of 42,900 records of which a Recall of 1 can be observed, where the model only exhibits 18 errors, i.e., there are 18 students for whom an `ACADEMIC_STATUS` of `Normal` was predicted, when in fact it was `At Risk`. There are also 20,296 cases where the status is correctly predicted. Likewise, a Recall of 0.96 is obtained in terms of the predictions of the `Normal` status, with 21,699 hits and 887 misses. Comparing the results with the test model carried out, the improvement is visible going from a Recall from 0 to 1 for the prediction cases of students who will possibly remain in a state of `At Risk`.

6.2. Approach 2: Using only the last two semesters

The evaluation of Approach 2, also carried out with the set of data destined for the test, shows us a total of 9,642 records, with which a Recall of 1 is obtained, for the prediction of students who will remain in some State of Academic Risk. With this, we see that the model is capable of correctly predicting 100% of the cases. For the case of Normal states, the recall is 0.99, which shows a total of 5,329 predictions made correctly and only 40 failures. Compared to the test model, we can see an improvement in results when going from a recall value of 0.25 to 1. With this, we obtain a model that notably improves the prediction of cases of interest, which are the students who will be in a possible state of academic risk.

7. Discussion

The analysis of the variables is a fundamental part of the learning process pertaining to the academic results obtained by the students. During this phase, it is possible to know the relationships between them, trends in the results, the importance of the fulfillment of the academic objectives, and above all, the role that each plays within the entire academic process.

The final results of the descriptive analysis present the relevance of the `COUNSELING` and `SUPPORT_CENTER_ATTENDANCE` variables. They show a clear trend where students who attend or take these services at the University, during that period, obtain better results on averages, and in turn, avoid reaching the `At Risk` state.

The data shows that the students whose `SPADIES` score is equal to 5, that is, those who according to this model have a higher risk of dropping out, are the ones who mostly attend and make use of the academic counseling and support center services. This in turn leads them to obtain better averages in the semesters in which they attend.

Due to the COVID-19 pandemic, the University has made changes to the `ACADEMIC_STATUS` assigned to students during the semesters where all academic activity was online. As such, it has not been possible to test the models with real data. It is therefore recommended that they start being used as from the second semester of 2022, in which new academic risk statuses will be assigned. From then on, it will be possible to test the models.

8. Conclusions and Future Work

Through the use of Data Mining and Machine Learning, it is possible to carry out a detailed study of the academic situation of higher education students. We show that it is possible to analyze, explore and execute possible strategies that allow higher education institutions to make decisions that allow timely attention to the needs of students, and in turn reduce student desertion. This has been an issue and an important focus of HEIs throughout the world.

Making use of the eleven variables used for this work, promising results are obtained for the models used. There is potential room for improvement if other relevant variables that have an important bearing on the academic situation of the students are employed.

The two approaches described in the project show good prediction results for students who will possibly remain at with `At Risk` status. However, the second approach shows better results at a general level, that is, the grouping of the data, and use of the history of the students, allowing predictions to be made in a more accurate way.

With the results obtained and using the generated models, it is possible for the University to carry out a more detailed follow-up of the students who may possibly be in a state of academic risk, in order to reduce the dropout figures that afflict students and HEIs. As such, the results can be implemented within the CRM, to strengthen the mechanisms of analysis, detection, and attention to students at risk that are currently being supported.

As future work, it is possible to include data from sections or courses, along with their partial grades, in order to generate a new model that can be run in the middle of the academic semester, and that can perhaps generate additional value or even higher precision with the knowledge of which students have or possibly may be at academic risk at the end of the semester. Furthermore,

it is important to continue collecting information as far as possible, of each and every one of the activities carried out by students within the University that may be related to their academic performance. This can be included in future work so that the impact that each piece of data may have on the results obtained can be evaluated. As the information increases and there are more records and variables, it is possible to improve the analysis, and propose new models and new ideas to reduce students dropping out.

References

[1] G. Parody, N. Ariza, D. Basto, Guía para la implementación de educación superior del modelo de gestión de permanencia y graduación estudiantil en instituciones, 2015.

[2] E. Castaño, S. Gallón, K. Gómez, J. Vásquez, et al., Análisis de los factores asociados a la deserción estudiantil en la educación superior: un estudio de caso, Revista de educación (2008).

[3] R. Wirth, J. Hipp, Crisp-dm: Towards a standard process model for data mining, in: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, volume 1, Manchester, 2000, pp. 29–39.

[4] J. A. Jiménez Toledo, S. R. Timarán Pereira, Caracterización de la deserción estudiantil en educación superior con minería de datos, Revista Tecnológica-ESPOL 28 (2015).

[5] X. F. Orrantia, E. Silva, Deserción estudiantil universitaria en el primer semestre. el caso de una institución de educación superior ecuatoriana, Deserción, calidad y reforma universitaria. Apuntes para el debate (2014).

[6] R. Zarate Rueda, E. Mantilla Pinilla, La deserción estudiantil uis, una mirada desde la responsabilidad social universitaria, Zona próxima (2014) 121–134.

[7] L. G. Isaza, C. D. Lubert, D. M. Montoya, Caracterización de la deserción estudiantil en la universidad de caldas el período 2009-2013. análisis a partir del sistema para la prevención de la deserción de la educación superior–spadies, Latinoamericana de Estudios Educativos 12 (2016) 132–158.

[8] L. Pájaro Fuentes, Sistema predictivo basado en aprendizaje automático para la deserción estudiantil en instituciones de educación superior, 2016.

[9] J. Garcia Franco, Implementación de un modelo computacional basado en reglas de clasificación supervisadas para la predicción de la deserción estudiantil en la universidad peruana unión filial juliaca, 2019.

[10] S. Quishpe-Morales, D. Pillo-Guanoluisa, I. Revelo-Portilla, L. Guerra-Torrealba, Modelo de predicción de la deserción universitaria mediante analítica de datos: Estrategia para la sustentabilidad, Revista Ibérica de Sistemas e Tecnologias de Informação (2020) 38–47.

[11] D. I. Candia Oviedo, Predicción del rendimiento académico de los estudiantes de la unsaac a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático, 2019.

[12] D. Bustamante, O. Garcia-Bedoya, Predictive academic performance model to support,

prevent and decrease the university dropout rate, in: Applied Informatics, 2021, pp. 222–236.

[13] M. F. Haderne, Uso de tecnologías de la información para detectar posibles deserciones universitarias, in: VII Congreso de Tecnología en Educación y Educación en Tecnología, 2012, pp. 1–9.

[14] E. Castaño, S. Gallón, K. Gómez, J. Vásquez, et al., Deserción estudiantil universitaria: una aplicación de modelos de duración, Lecturas de economía (2004) 39–65.

[15] E. Himmel, Modelo de análisis de la deserción estudiantil en la educación superior, Calidad en la Educación (2002) 91–108.

[16] A. S. Escarria, Deserción universitaria en colombia, Academia y virtualidad 3 (2010) 50–60.

[17] B. A. Castro-Montoya, C. M. Lopera-Gómez, R. D. Manrique-Hernández, D. Gonzalez-Gómez, Modelo de riesgos competitivos para deserción y graduación en estudiantes universitarios de programas de pregrado de una universidad privada de Medellin (Colombia), Formación universitaria 14 (2021) 81–98.

[18] G. Lemaître, F. Nogueira, C. K. Aridas, Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning, Journal of Machine Learning Research 18 (2017) 1–5.