A Score approach to identify the risk of students dropout: an experiment with Information Systems Course

Robinson Crusoé da Cruz
robinsoncruz@uniaraxa.edu.br
Centro Universitário do Planalto de Araxá
Araxá, MG, Brazil

Renato Correa Juliano
renatocorrea@uniaraxa.edu.br
Centro Universitário do Planalto de Araxá
Araxá, MG, Brazil

Francisco Carlos M. Souza
franciscosouza@utfpr.edu.br
Universidade Tecnológica Federal do Paraná
Dois Vizinhos, PR, Brazil

Alinne C. Correa Souza
alinnesouza@utfpr.edu.br
Universidade Tecnológica Federal do Paraná
Dois Vizinhos, PR, Brazil

ABSTRACT

Context: The student dropout in Higher Education contributes to much social, economic, and academic loss. Students have different reasons for dropping out but the main ones are related to difficulty in learning the content, the structure proposed by the course, and the lack of financial resources.
Problem: Besides understanding the motive for students completely abandoned their studies, the most important problem is identifying which groups of students are at risk of dropping out. However, studies focus essentially on categorical indicators, i.e., binary results that denote whether a student is or is not in the risk group. This type of analysis is important, but, it does not present the variation in the student's performance during their academic life.
Solution: Creating a score using machine learning techniques (KNN) can provide an instrument to measure how close the student is or not to the dropout group.
Theory: We used Organizational knowledge creation to make available and expand knowledge about dropping and to provide inputs for the creation of a knowledge system.
Method: The experimental study is quantitative and it was performed from the execution of KNN and its validation from statistical analyses.
Results: With equation developed and accuracy of 87% with KNN, was possible to develop a drop out risk score with values between 0 and 1,000, where the closer to 0, the greater the probability of the student to drop out.
Contributions and Impact in the IS area: The main contribution of the paper is to provide a new method to assist in the analysis of higher education dropout in Information System and other courses.

CCS CONCEPTS

• **Information systems → Information systems applications**; **Data mining**; **Data analytics**; • **Education**;

KEYWORDS

Dropout from Higher Education, Score, Machine Learning, KNN

ACM Reference Format:
Robinson Crusoé da Cruz, Renato Correa Juliano, Francisco Carlos M. Souza, and Alinne C. Correa Souza. 2023. A Score approach to identify the risk of students dropout: an experiment with Information Systems Course. In *XIX Brazilian Symposium on Information Systems (SBSI '23), May 29–June 01, 2023, Maceió, Brazil.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3592813.3592896

1 INTRODUCTION

Students who begin their studies in Higher Education and does not finish have caused much social, economic, and academic loss. The main reasons for dropping out are the students' difficulty in following the content, the structure proposed by the course, and the lack of financial resources [5]. Dropout from Higher Education has been widely discussed by Higher Education Institutions (HEIs) and by government agencies in order to define strategies to minimize this problem [5]. In 2010, the Ministry of Education (MEC) performed an investigation with higher education students from both the public and private institutions during the period from 2010 to 2014 to verify the permanence of these students in the same entrance course. According to the results, 49% of students who started in Higher Education in 2010 dropped out of their courses and academic performance is among the main reasons that led them to drop out [11]. Dropping out of the Bachelor of Information Systems course is considered high, as shown in study [16], with the main reasons being lack of interest in the course, course structure and financial difficulties.

Some studies [9], [2] were conducted to identify the reasons for the dropout of Higher Education. In some of these studies, the Machine Learning technique was applied to discover factors that may contribute to drop out and the creation of classifiers. The study [9] aimed to identify groups at risk of dropout. For this, data mining techniques with J48 classification algorithm were applied to data from students from twelve courses at the Federal University of Rio Grande (FURG) - Brazil, obtaining an accuracy of 90%. In contrast, in the investigation by [2], the logistic regression technique was used to identify groups at risk of dropping out. For this, academic and demographic data of 32,538 students at the University of Washington in the USA were analyzed, obtaining an accuracy of 66%. The main goal of this kind of approach is to identify risk groups, with

SBSI '23, May 29–June 01, 2023, Maceió, Brazil

ACM ISBN 979-8-4007-0759-9/23/05…$15.00
https://doi.org/10.1145/3592813.3592896

binary classifications, that is, they present results if the student is or is not in the dropout risk group.

Studies and applications of machine learning algorithms are promising. However, they have limitations to identify students situations of a student in relation to the risk group. For example, when applying a classifier, a student can be classified in the group with no risk of dropout. On the other hand, using a score system, in addition to classifying the student as belonging to the group with no risk of dropping out, it is possible to define whether or not he is close to the risk group.

This paper presents a Score approach to analyze the student's dropout risk in the Information system course of higher education. The Score approach was developed using a real database with general and socioeconomic students' data from a non-profit University (Foundation). Finally, an experiment was conducted to evaluate the effectiveness of the proposed approach. Overall, the contributions of the present work can be summarized in the following points: *i)* use of the K-Nearest Neighbors (KNN) algorithm to classify students as dropouts and non-dropouts; *ii)* an effective score approach to analyze the student's dropout risk and *iii)* an empirical assessment is conducted on 1,496 samples which represent the students to evaluate the effectiveness of our approach.

The remainder of this paper is organized as follows: Section 2 presents the related works. Section 3 details the proposed approach. Section 4 reports research design of experiments conducted. Section 5 presents the results obtained and discusses the benefits, relevance, and limitations of the proposed approach. Finally, in Section 6 the conclusions and future directions are discussed.

2 RELATED WORK

Some works have discussed the classifiers and methods development to analyze the degree of students dropout in Higher Education. In general, these works propose the binary classifiers development to define whether or not a student is in the dropout risk group. The works conducted by Saraiva et al. [15] performs a study of grouping socioeconomic and academic data from students of a Computer Technician Course to analyze the influence of data on dropout. The study was based on the CRISP-DM Methodology [17], common in data mining projects, and for the grouping application was used the k-means method. The results indicate that the family income, study shift, student's gender, age group and level of education, can influence the student's academic performance.

In [6], a study was carried out with data from a public database made available by the National Institute of Educational Studies and Research Anísio Teixeira (INEP), using five Machine Learning techniques: Naive Bayes, K-Nearest Neighbors, Decision Trees, Random Forest, and Neural Networks. As part of the results, the authors performed a comparison between the techniques, and Random Forest performed the best overall. In [7] two databases were analyzed: (i) academic containing student information; (ii) demographic and socioeconomic data for Brazil, using IBGE (Brazilian Institute of Geography and Statistics) indicators. In this research, the accuracy was analyzed according to the areas of the courses, obtaining an overall accuracy of 70%.

Lanes and Alcântara [9] performed an analysis of data from the Federal University of Rio Grande (FURG) and applied the J48 algorithm to analyze dropout. This analysis was conducted with students who had at least one year of course and with demographic and academic data. The results achieved an accuracy of 90% in the classification of the dropout risk group. The research of Júnior et al. [8], a study was carried out to select variables to help predict higher education dropout. The variables of this study have similarities with those used in this approach.

In [2], the authors analyzed the academic and demographic data of 32,538 students at the University of Washington in the USA. The results point to an accuracy of 66% using Logistic Regression to identify groups at risk of dropout. Finally, the work [4] analyzed data of 11,036 students and 415,327 student performance results from eight courses at the Federal University of Pernambuco (UFPE) between 1998 and 2018. A regression method and other techniques were applied to produce a model that aims to predict and assist in the dropouts analysis from courses. The results indicate the use of socioeconomic, cultural, and behavioral data.

Table 1 presents a summary of the related works presented. The first column lists these works, the algorithms and/or techniques used (second column), general accuracy achieved by classifiers (third column), the period the research was carried out (fourth column), samples (fifth column), and the number of courses (sixth column). We used "Not Informed (NI)" and "Not Applied (NA)" for information not provided or not applied by the studies.

The related works essentially focus on categorical indicators, which aim to denote whether the student is or is not in the risk group. The authors, in general, indicate points to improve the identification risk of students dropouts, such as handling the imbalance between classes and the use of new variables and methods. From this perspective, the present work does not aim only at classification, but it proposes a new approach that classifies the student according to his Score.

3 STUDY STRUCTURE

In this section, we outline the details of the Score approach to identify the students dropout risk in higher education. As shown in Figure 1, our approach uses a real database with general and socioeconomic students' data. The process consists of four steps: *(1)* Data Definition; *(2)* ETL; *(3)* Classifier Analysis; and *(4)* Score development.

3.1 Step 1 - Data Definition

We used data from a Private Non-Profit University. This type of University consists of a non-governmental organization, maintained by a Foundation, constituted by traditional confessional, philanthropic, and community universities, where it is expected that the resources obtained through monthly fees are reverted to the Institution [14].

The dataset is composed of academic semesters of the Bachelor of Information Systems course, between the second semester of 2013 and the first semester of 2021, with a total of 1,496 records distributed in variables as shown in Table 2. The first column displays the 19 academic variables; the second column details the variable's description, and the third column indicates which values the variables take.

Studies	Algorithms	Accuracy	Period	Sample	N. courses
[15]	k-means	NA	2010 to 2018	500	1
[6]	5 Algorithms	79% (avg)	NI	376,746	NI
[7]	Random Florest	70.00%	from 2010	51175/31482	59
[9]	J48	90.70%	2012 to 2017	916	12
[8]	J48	83 to 87%	6 semesters	NI	NI
[2]	Regression	66.59%	1998 to 2006	32538	NI
[4]	Regression	NA	1998 to 2018	11036	NI
This study	KNN	87%	2013 to 2021	1496	1

Table 1: Overview of related works.

Variable	Description	Values
AcademicPeriod	Period the student was studying (occupying vacancy)	1 to 8
AgeDays	Age the student was at the end of the semester	6,000 to 36,500
FirstTwoMonths	Average Grade obtained in the first two months	0 to 10.00
SecTwoMonths	Average Grade obtained in the second two months of class semester	0 to 10.00
ACQG	Undergraduate Quality Control Assessment Grade	0 to 2.00
AverageFinalExam	Average Grade obtained in the final exam	0 to 10.00
AverageFinalGrade	Average Student's Final Grade	0 to 10.00
PerceAbsence	Student absence percentage	0 to 1
CountSubTak	Number of subjects the student was taking in the semester	0 to 10
CountSubAp	Number of subjects that the student was approved in the semester	0 to 10
CountSubFail	Number of subjects that the student was fail in the semester	0 to 10
TotalWorkLoadSub	Total course load of subjects	20 to 800
TotalWorkSubTheor	Total workload of theoretical subjects.	0 to 800
TotalWorkSubPract	Total workload of practical subjects.	0 to 800
CountMatCalcSub	Number of math and calculus subjects the student was taking	0 to 10
LiveCity	defines if the student lives in the city where the University is located	1=Yes e 2=No
Gender	Gender	0=Female, 1=Male e 2=Not defined/declared
HighSchoolType	Type of High School the student attended	1=Public e 2=Private
MaritalStatus	Student's marital status	1=Single, 2=Married, 3=Divorced e 0=Others
Situation/Class	Student status during or at the end of the semester	1=Non-dropout e 2=Dropout

Table 2: Academics variables analyzed.

3.2 Step 2 - ETL

The data extracted from the Education Institution's database that can identify students has been changed or removed. The available variables were mapped and classified to assist in the KNN application. For example, the student's marital status was classified with values between 0 and 3, and the student's age was converted to days as shown in Table 2. Then the data was extracted and stored in the dataset, and finally, the data was validated with the official source.

3.3 Step 3 - Classifier Analysis

In this step, the KNN algorithm [1] was applied to analyze the classification of dropout and non-dropout students using the 19 variables presented in Table 2. Some variables such as $CountSubTak$

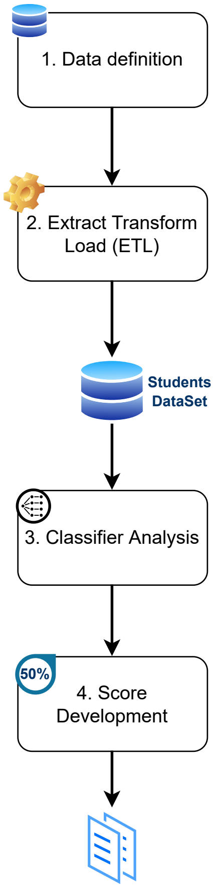


Figure 1: Study Steps

and $CountMatCalcSub$ represent a sum or average of the outcomes achieved by students in the disciplines of the semester.

The $CountSubTak$ variable represents the number of disciplines that students attended in the semester, while the $CountMatCalcSub$ variable represents the number of disciplines with math content. In addition, some variables were added to analyze the student's profile, for example, gender, if the student lives in the city where the university is located, and the student's age at the end of the semester.

The choice of the KNN algorithm was essential for the SCORE approach development because this algorithm is based on the distance between neighbors (k). Therefore, it is expected that if a student has a high number of neighbors with a risk of dropping out (class 2), then, the higher the probability that will belong to this class. On the other hand, if there is a low number of class 2 neighbors, it may have a low risk of evasion despite belonging to class 2. The KNN classifier was applied using the Python programming language with the scikit-learn library. In the analysis, 20% of the sample was separated for testing and 80% for training. The GridSearchCV available in scikit-learn was used to assist in choosing the better hyperparameters.

3.4 Step 4 - Score Development

In this step, the Score approach was developed based on the inputs generated in the previous step, using the better hyperparameter result for the KNN classifier. The proposal was to use the k value (neighbors) of the best test result of the classifier. This value was used as a denominator to define the ranges and limits of each score range. Figure 2 shows the proposal and structure of the Score approach that depends directly on the k value.

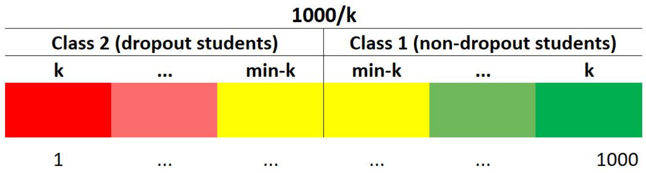


Figure 2: Score conceptual proposal.

The interval is divided between the values 1 and 1000, and the amount of groups depends on the k value, where *min-k* represents the shortest of k value that defines the sample classification in class 1 or 2.

4 EXPERIMENTAL STUDY

We conducted an experiment to determine if the KNN classifier application and the Score proposed are effective to identify the risk of students dropout in higher education. In this study the guideline recommended by Wohlin et. al [18] were used. These guideline provides guidance on conducting the experiment, which must include research questions, variables, in addition to presenting the steps taken during the conduction of the experiment. For achieving the goal, we raise the following Research Questions (RQs):

RQ_1: **How effective is the KNN classifier application to identify the risk of students dropout in higher education?** This RQ aims to analyze the classification of dropout and non-dropout students.

RQ_2: **How effective is the Score approach to identify the risk of students dropout in higher education?** This RQ aims to analyze how close the students are or are not into the risk group through the Score approach, which uses the better hyperparameter result for the KNN classifier.

The assessment of the approach was carried out through two experiments ($E = e_1; e_2$). The first experiment (e_1) is regarding the KNN classifier and the second Score approach. The classes used in these experiments were defined as: *1-non-dropout* and *2-dropout*. To answer the RQs, we conducted the experiments in four steps:

(1) Definition of the students' data;
(2) Performing the statistical analysis to identify factors and variables that could have a higher or lower influence on the results;
(3) Application of the KNN classifier to analyze the classification of dropout and non-dropout students;
(4) Application of Score approach using the better result from KNN classifier.

5 RESULTS AND DISCUSSIONS

This section presents the experimental results. We focus on the analysis of results concerning the effectiveness of the KNN classifier and our proposed Score approach.

5.1 Statistical analysis

Figure 3 shows the dropout in academic semesters. We analyzed that in the pandemic period, between the years 2020 and 2021, there was an insignificant increase in dropouts. Therefore, the COVID-19 pandemic didn't influence the students' dropouts this Institution.

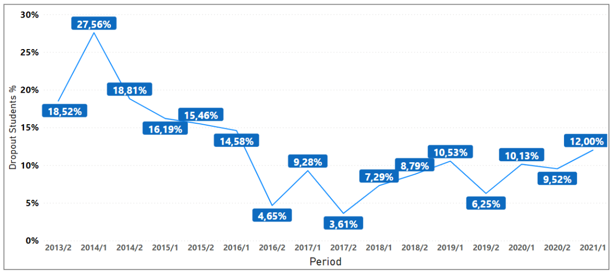


Figure 3: Distribution of dropout students according to period.

We observed that 70.43% of dropouts are concentrated in the first three academic semesters, as shown in Figure 4. Dropping out in the first semester indicates that the student did not identify with the course or there are other problems. This result indicates that the $AcademicPeriod$ variable from Table 2 will be important when applying the classifier.

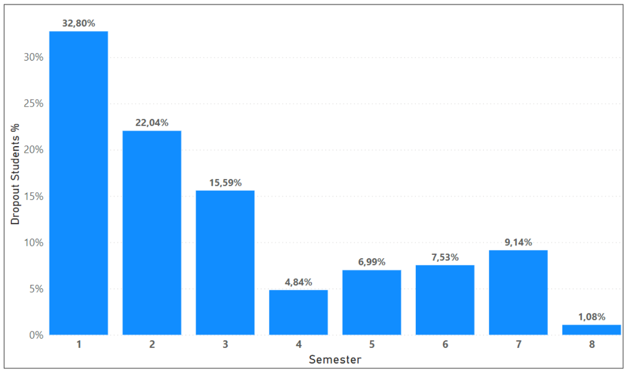


Figure 4: Distribution of dropout students according to academic semester.

Figure 5 presents the variables distribution regarding the final average of the students' grades ($AverageFinalGrade$) and Frequency ($PerceAbsence$). According to the results, the students grades belonging to class 1 have a higher concentration above grade 6, while the grades of class 2 are dispersed, despite having a proportion close to zero.

We analyzed some Spearman correlations between students grades, as shown in Figure 6. According to the results, we noticed a high correlation between the grades of the first two months ($FirstTwoMonths$) and the last two months ($SecTwoMonths$). This correlation indicates that the higher grade from the two first months, the higher grade from the two last months.

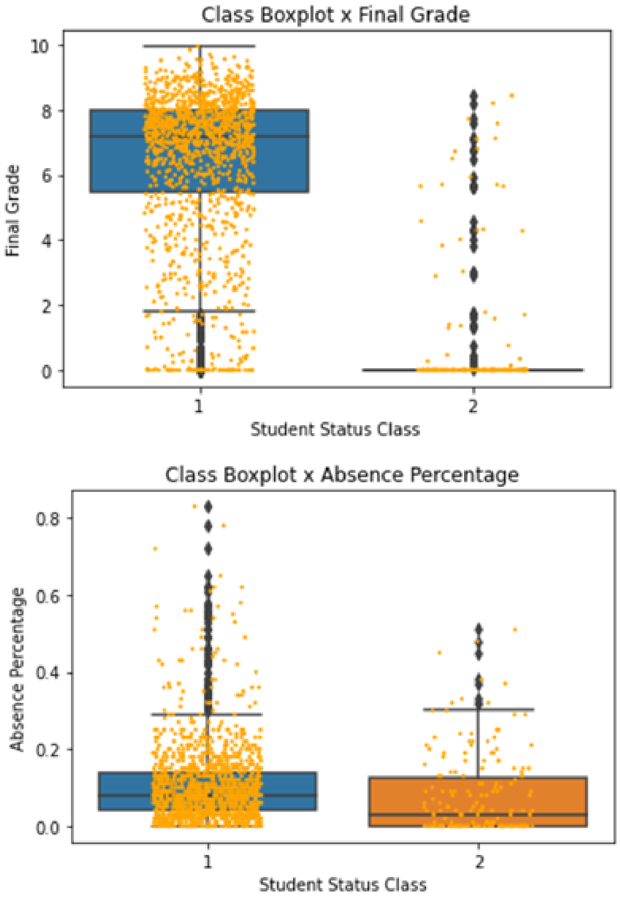


Figure 5: Distribution of the final average of the students' grade and absence percentage.

5.2 Effectiveness of KNN Classifier (RQ_1)

This RQ aims to evaluate the KNN classifier's effectiveness to identify the risk of students' dropout and non-dropout. To answer this RQ, we analyze a sample of 1.496 records, being 1.310 from class 1 (Non-dropout) and 186 from class 2 (Dropout). For this, we performed six tests in which the variables were normalized before applying the classifier (Table 3). Despite the similarity between the results, the test 4 can be considered the better result, as it has the reduction of components (PCA) with 5 neighbors *(k).*

The total sample has 87.57% (1.310) of records that belong to class 1 (Non-Dropouts). It is important to highlight that a high imbalance between classes can interfere with the results of the classifiers (Table 4)[13]. For this, we applied the NearMiss method [10] to ensure the sample balancing of class 1 (Non-Dropouts). The test 10 result after balancing, as shown in Table 4, indicates a similarity with the results of Test 4 (Table 3). However, there was a reduction in the final accuracy of test 10 of Table 4.

	Count	Class Sample				Class Precision			class fi-score	
Test/Result	**Variables**	**1**	**2**	**PCA**	**K**	**1**	**2**	**Accuracy**	**1**	**2**
1	19	1310	186	No	5	0.97	0.88	0,96	0,98	0,81
2	19	1310	186	No	7	0.96	0.90	0,96	0.98	0.81
3	19	1310	186	No	9	0.96	0.87	0.95	0.97	0.79
4	19	1310	186	9	5	0.96	0.90	0.96	0.98	0.81
5	19	1310	186	9	7	0.96	0.87	0.95	0.97	0.79
6	19	1310	186	9	9	0.96	0.84	0.95	0.97	0.76

Table 3: Unbalanced classes tests.

	Count	Class Sample				Class Precision			class fi-score	
Test/Result	**Variables**	**1**	**2**	**PCA**	**K**	**1**	**2**	**Accuracy**	**1**	**2**
7	19	186	186	No	5	0.83	0.91	0.87	0.88	0.86
8	19	186	186	No	7	0.83	0.88	0.85	0.86	0.85
9	19	186	186	No	9	0.83	0.88	0.85	0.86	0.85
10	19	186	186	14	5	0.83	0.91	0.87	0.88	0.86
11	19	186	186	14	7	0.83	0.88	0.85	0.86	0.85
12	19	186	186	14	9	0.83	0.88	0.85	0.86	0.85

Table 4: Balanced classes tests

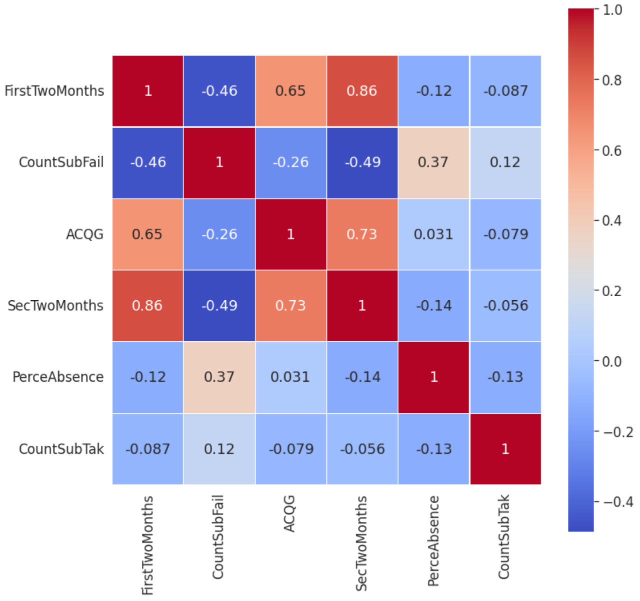


Figure 6: Correlation between some variables.

Therefore, the results achieved indicate that test 10 (Table 4) contains the better result whether we consider the importance of between classes balancing, dimension reduction (PCA), and the number of neighbors (k).

5.3 Effectiveness of Score Approach (RQ_2)

This RQ aims to analyze the score approach effectiveness. For this, we used test 10, which was the better result achieved in RQ_1 as shown in Table 4. Table 5 presents details of the classifier KNN results according to the hyperparameters, presented in Table 4, referring to test 10 ($PCA = 14$ and $k = 5$). This result was selected for the Score approach development due to balancing the $f1-score$ value has better accuracy of classes and values for k and PCA.

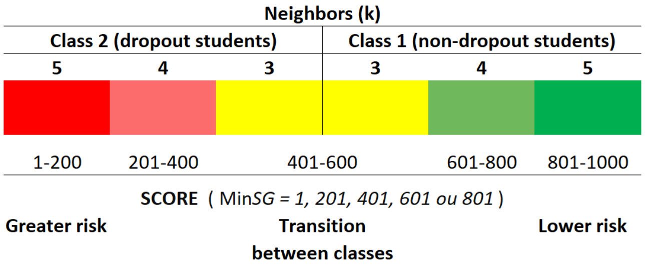
Class			Average of variables		
Ranked on KNN	Correct	Count	AcademicPeriod	CountSubTak	CountSubAp
2	2	**30**	2.80	5.73	0.23
2	1	**3**	2.66	6.00	0.33
1	1	**35**	3.42	4.97	4.00
1	2	**7**	3.43	4.42	1.86

Table 5: Confusion matrix and variables.

According to Table 5, dropout students have a lower number of approvals in subjects (CountSubAp), as shown in line 1. On the other hand, it is noted that, on average, there were more subjects, when compared to the group of non-dropouts, as shown in line 3.

Figure 7 presents the Score approach which uses the values of k as parameters to define the *Score* limits and values. The Score values can range between 1 (highest risk of dropout) and 1,000 (lowest risk of dropout). As the better result for the classifier was $k = 5$, the Score was divided into five intervals with 200 points ($1000/k$). If a given student has $k = 3$ from class 1 or 2, it will be classified in the range of 401 to 600. This convergence point indicates that the student is in the transition between classes.

Regarding class 2 (dropout), if a student is classified in class 2 and has five neighbors (k) of this class, thus this student belongs to class 2 and has a high risk of dropout. On the other hand, if a student is classified in class 2 and has three neighbors (k) from this class, then, they have a lower risk of dropout because there is a proximity to 2 students from class 1 (non-dropout). This same analysis can be applied to students who are classified in class 1 (non-dropout) because they have a value of $k = 3$. However, these students have a greater risk of dropping out because they are in the transition between classes 1 and 2.



Neighbors (k)

Class 2 (dropout students)			Class 1 (non-dropout students)		
5	4	3	3	4	5
🟥	🟥	🟨	🟨	🟩	🟩
1-200	201-400	401-600		601-800	801-1000

SCORE (Min*SG = 1, 201, 401, 601 ou 801*)

Greater risk Transition between classes Lower risk

Figure 7: Score with k value identified.

The results presented in Figure 7, indicate that it is possible to classify the student according to his class and score. However, in the experiments, variables were identified that can be used to help in the Score development. In Table 5 for example, we noted that students classified in class 2 (dropout) have a lower number of approvals and belong to the first periods of the course. With these indications, it was necessary to consider the variables' importance presented the Table 2 to refine the Score between the intervals. For example, a student may belong to class 1, with a value of $k = 5$, with a classification in a range between 801 and 1000. Furthermore, with the approach presented in Figure 7 it is not possible to define the exact value of the *Score*. However, with the *Score* Adjustment Equation 1, it is expected that it will be possible to set this value.

$$ScoreFinal = MinSG + \sum_{i=1}^{n} (v_i * p_i * 10) \tag{1}$$

where n represents the number of variables analyzed; v the variable normalized between 0 and 1; p the weight of the variable between 0 and 19.90, according to its importance within the context; $MinSG$ the smallest[1]; and score of the student according to the class and number of k. Regardless of the number of variables (v), the sum of the weights (p) will be <= 19.90, e.g., the total weight will be divided into all the variables used.

Hypothetical example of a student: consider that only the variables $CountSubAp$ and $AcademicPeriod$, presented in Table 5 are used to adjust the Score of a student and have normalized values 1 and 0.5, weights 14.90 and 5 (Total=19.90), respectively. After applying the KNN classifier, the student is classified in class 1 (non-dropout), as this student has 4 neighbors (k) of class 1. Initially, this student would be classified in class 1 (non-dropout) and Score with a range between 601 and 800. With this result, the value of $MinSG$ will be 601 and the calculation can be represented by:

$$ScoreFinal = 601 + ((1 * 14.90 * 10) + (0.5 * 5 * 10)) \tag{2}$$

The $ScoreFinal$ value will be 775. Thus the student belongs to class 1 (non-dropout), and according to the result of Equation 2, this student is close to the group of students with a low risk of dropout (801 to 1,000). As the weight of each variable will be between 0 and 1, in the $CountSubAp$ variable the student obtained 100% (1), and in the $AcademicPeriod$ 50% (0.5). Therefore, the student has a high rate of approvals and does not belong to the first academics period.

The partial result of the approach, shown in Figure 7, indicates that it is possible to classify the student and define their Score group within an interval. However, to refine its classification, further studies would be needed to identify the variables and their respective weights, and thus calculate the exact result, as proposed by the Equation 1. The results can be used as an indication of a possible dropout because even if the student has data that can classify him as at risk of dropout, it is not possible to say that this student will be a dropout. Regarding the reuse of this approach in other scenarios, we noted that it will be necessary to adjust the groups according to the value of k, the variable's definition, and their respective weights. Therefore, this shows that the approach is promising and can be adapted to other courses.

In addition, it is expected that the results of this study can be applied together with BI (Business Intelligence) projects, for example, the study presented in [12], which aimed to create a BI to assist in the understanding of dropout Phenomenon in Information Systems Courses and others Courses.

6 CONCLUSION

The research identified in the literature focuses essentially on categorical indicators, that is, binary results that denote that the student is or is not in the risk group. Although this analysis is important, it does not show the variation in student performance during their academic life

The preliminary results achieved indicate that the creation of a Score is promising and can be an alternative, because in addition to classifying a student according to his class, it is possible to analyze how close he is or is not to the risk group. Although the results obtained with the Score model are partial, they indicate that when applying the adjustments using the Equation 1, significant results can be obtained when considering the weight of the variables. For example, if a determined student has a high frequency in the subjects, this variable would be fundamentals to increase his *Score*.

[1] For example, if the student is classified with $k = 4$ and belongs to class 2 (Dropout), the value of $MinSG$ will be 201.

As Brazil went through a period of Pandemic, this brought changes in learning, evaluation, income, and psychological aspects. So, these factors may have influenced the dropout of students, and as a consequence, the results achieved. However, part of the sample belongs to a period before the Pandemic, and in the future, new experiments will be applied, and if necessary, the appropriate adjustments will be made. Finally, the results presented in Figure 4 indicate that the pandemic did not influence in the student's dropout of the Information Systems course.

This study complement the results of the studies presented in [3]. The results were identical, however the focus of this study has a larger range data and with a proposal directed to the Information Systems course. These results that can possibly be extended to other higher education courses through individual analysis.

Finally, as future work are to define of the variables and their respective weights, validate and adjust the Equation 1 and, validate this approach with student results in the second half of 2022.

REFERENCES

[1] David W. Aha, Dennis Kibler, and Marc K. Albert. 1991. Instance-based learning algorithms. *Machine Learning* 6, 1 (1991), 37–66. https://doi.org/10.1007/BF00153759

[2] Lo. Aulck, N. Velagapudi, J. Blumenstock, and J. West. 2016. Predicting student dropout in higher education. *arXiv preprint arXiv:1606.06364* (2016).

[3] Robinson Crusoé da Cruz, Renato Correa Juliano, Alinne Cristinne Correa Souza, and Francisco Carlos Monteiro Souza. 2022. Desenvolvimento de um Score para análise de risco de evasão de estudantes do Ensino Superior baseado em Aprendizado de Máquina. *Anais do Computer on the Beach* 13 (2022), 142–148.

[4] H. da Silva and P. Adeodato. 2012. A data mining approach for preventing undergraduate students retention. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*. 1–8. https://doi.org/10.1109/IJCNN.2012.6252437

[5] Delsi Fries Davok and Rosilane Pontes Bernard. 2016. Avaliação dos índices de evasão nos cursos de graduação da Universidade do Estado de Santa Catarina - UDESC. (jul 2016).

[6] Leonardo de Almeida Teodoro and Marco André Abud Kappel. 2020. Aplicação de Técnicas de Aprendizado de Máquina para Predição de Risco de Evasão Escolar em Instituições Públicas de Ensino Superior no Brasil. *Revista Brasileira de Informática na Educação* 28 (2020), 838–863.

[7] Bruno Claudino Pereira de Brito, Rafael Ferreira Leite de Mello, and Gabriel Alves. 2020. Identificação de Atributos Relevantes na Evasão no Ensino Superior Público Brasileiro. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. SBC, 1032–1041.

[8] J. Júnior, R. Noronha, and C. Kaestner. 2017. Criação e Seleção de Atributos Aplicados na Previsão da Evasão de Curso em Alunos de Graduação. In *Anais do Computer on the Beach*. 61–70. https://doi.org/10.14210/cotb.v0n0.p061-070

[9] M. Lanes and C. Alcântara. 2018. Predição de Alunos com Risco de Evasão: estudo de caso usando mineração de dados. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, Vol. 29. 1921.

[10] Inderjeet Mani and I Zhang. 2003. kNN approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets*, Vol. 126. ICML United States.

[11] MEC. 2016. *Altos índices de desistência na graduação revelam fragilidade do ensino médio, avalia ministro*. http://portal.mec.gov.br/ultimas-noticias/212-educacao-superior-1690610854/40111-altos-indices-de-evasao-na-graduacao-revelam-fragilidade-do-ensino-medio-avalia-ministro

[12] André Menolli, Flávio Horita, José Jorge L Dias, and Ricardo Coelho. 2020. BI–based Methodology for Analyzing Higher Education: A Case Study of Dropout Phenomenon in Information Systems Courses. In *XVI Brazilian Symposium on Information Systems*. 1–8.

[13] D Ramyachitra and P Manikandan. 2014. Imbalanced dataset classification and solutions: a review. *International Journal of Computing and Business Research (IJCBR)* 5, 4 (2014), 1–29.

[14] Helena Sampaio. 2000. Ensino superior no Brasil: o setor privado. *Cadernos de Pesquisa* (2000), 213–213.

[15] Daniel Victor Saraiva, Silas SL Pereira, Reinaldo B Braga, and Carina T de Oliveira. 2021. Análise de Agrupamentos para Caracterização de Indicadores de Evasão. In *Anais do XXIX Workshop sobre Educação em Computação*. SBC, 238–247.

[16] Juliana Saraiva, Vanessa Dantas, and Amanda Rodrigues. 2019. Compreendendo a Evasão em uma Década no Curso Sistemas de Informação à luz de fatores humanos e sociais. In *Anais do IV Workshop sobre Aspectos Sociais, Humanos e Econômicos de Software*. SBC, 21–30.

[17] Rüdiger Wirth and Jochen Hipp. 2000. 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*, Vol. 1. Manchester, 29–40.

[18] C. Wohlin, P. Runeson, M. Host, M. C. Ohlsson, B. Regnell, and A. Wesslen. 2012. *Experimentation in Software Engineering: An Introduction* (1st. ed.). Springer-Verlag Berlin Heidelberg.