Educational Data Mining: The Case of Department of Mathematics and Computing in the Period 2009-2018

M. Sitoe, O. Zacarias

Open Science Index, Computer and Information Engineering Vol:18, No:5, 2024 publications.waset.org/10013637.pdf

Abstract**—**University education is influenced by several factors that range from the adoption of strategies to strengthen the whole process to the academic performance improvement of the students themselves. This work uses data mining techniques to develop a predictive model to identify students with a tendency to evasion and retention. To this end, a database of real students' data from the Department of University Admission (DAU) and the Department of Mathematics and Informatics (DMI) was used. The data comprised 388 undergraduate students admitted in the years 2009 to 2014. The Weka tool was used for model building, using three different techniques, namely: K-nearest neighbor, random forest, and logistic regression. To allow for training on multiple train-test splits, a cross-validation approach was employed with a varying number of folds. To reduce bias variance and improve the performance of the models, ensemble methods of Bagging and Stacking were used. After comparing the results obtained by the three classifiers, Logistic Regression using Bagging with seven folds obtained the best performance, showing results above 90% in all evaluated metrics: accuracy, rate of true positives, and precision. Retention is the most common tendency.

Keywords**—**Evasion and retention, cross validation, bagging, stacking.

I. Introduction

THE area of Information and Communication Technologies (ICTs), since its emergence, has undergone a constant evolution adapting to new trends, mainly those imposed by the advance of the information society that involve acquisition, storage, processing and distribution of information. The ICTs have evolved from a traditional orientation, which consisted of supporting administrative activities, to the current strategic decision-making support systems [1]. Currently, the task of ICTs has become, in addition to what was initially established, also the production of useful information for decision-making.

In recent decades, as the cost of hardware has fallen, it has become possible to store ever-increasing amounts of data. New and more complex storage structures were developed, such as: databases, Data *Warehouses,* Virtual Libraries, Web, and others [2]. This new reality allowed almost all electronic devices to be considered as sources of data capture, which originated a new paradigm in the field of Data Science that is Big Data. This new paradigm consists of extracting information from large volumes of data and converting it into a competitive advantage, so that organizations obtain even greater market opportunities [3]. In view of this scenario, new concepts have emerged that aim to propose technologies and treatments in situations where traditional data exploration and analysis techniques are not sufficient or adequate, as in the case of Data Mining [4].

In organizations, large volumes of data are produced and stored in various operational databases every second. Normally, these exceed the analysis capacity by means of traditional analysis techniques, from human (manual) to technological as in the case of Structured Query Language (SQL) techniques. With these limitations, valuable information for decision making is hidden in the high volume of data or even incorrect data are applied for decision making. This situation led to the emergence of the process known as Knowledge Discovery in Databases *(*KDD) [5].

Poor performance of students in higher education institutions in Mozambique, has become a problem for both the institutions and the body that oversees this sector, namely, the Ministry of Science, Technology and Higher Education (MCT). Moreover, dropouts are also a problem that is generally found in these institutions. The dropout and retention are mainly related to student's lack of financial resources and poor management of their expectations in relation to the courses or institutions [6]. In addition, if considering less favored students from social and financial perspective, retention and dropout levels may increase [7]. At Eduardo Mondlane University (UEM), all these situations are present at students' lives and represent a huge challenge to academic decision makers in their work of keeping balance and ensure educational equity and fairness to all admitted students.

II. Problem Definition

The DMI of the UEM offers four undergraduate courses, namely: Mathematics, Informatics, Statistics and Geographic Information Sciences. A brief analysis of the number of student's admissions and the graduation rates indicates that many of the students enrolled in years 2009 and 2014 have have either dropped out or are retained in their courses with successive unsatisfactory results. This preliminary analysis of collected data, showed that from 2009 to 2014, there was an increase in the number of admissions, being this in the order of 11.9%, but a decrease in the number of graduations in the order of 7.2% for 2008 and 5.8% for 2014 [8].

Taking all this information into consideration, the present study proposes the application of Data Mining techniques, using the undergraduate data, collected within courses offered by the DMI at UEM, in the period 2009-2018.

M. Sitoe is a master student and O. Zacarias is senior lecturer at the Department of Mathematics and Informatics at the Faculty of Science, Eduardo Mondlane University, Mozambique (e-mail: sitoem05@gmail.com).

Open Science Index, Computer and Information Engineering Vol:18, No:5, 2024 publications.waset.org/10013637.pdf

III. Educational Data Mining

Data Mining (DM) is an emerging research area that seeks to explore and analyze large datasets. It aims to identify patterns and relationships present in the data, in order to provide valuable information and knowledge for decision making in various areas of business, social context, industry, etc. When applied with data generated in an educational context, it is called Education Data Mining (EDM) [9]. In the education framework, it addresses issues related to school management, student performance assessment and teaching personalization.

As for prediction of academic performance, several studies have used Machine Learning (ML) algorithms to predict academic success of students based on variables such as demographic data, school history and behavior in Virtual Learning Environments (VLE). For instance, in a recent study on "Implementing AutoML in Educational Data Mining for Prediction Tasks" [9], concluded that the models were able to accurately predict students' academic performance, with a hit rate of around 80%.

Reference [10] shows that the meaningful determinants for success in training at HEIs, are subdivided into four groups:

- Sociodemographic determinants
- Behavioral determinants
- Psychological determinants
- Factors related to the educational institution

IV. Methods

In Educational Data Mining, ML algorithms are widely used to perform tasks such as predicting academic performance, analyzing learning patterns and for personalizing teaching. Among the most commonly used algorithms, we found random forest, K-nearest neighbors, and logistic regression. They are also compared in terms of their performance and applicability.

In this study, we also use Random Forest, K-Nearest Neighbor and Logistic Regression algorithms, with however, varying K-fold parameters to 3, 5 and 7 folds. With this procedure, we aimed to investigate the influence and performance of this variation on the results of the models. To reduce the rate of False Positive (FP) values and increment the model's performance, we used ensemble methods such as bagging and stacking.

V. Data Pre-Processing

The data preprocessing phase was achieved following part of the model [11], as shown in the Fig 1, where the corresponding steps are explained in A and B.

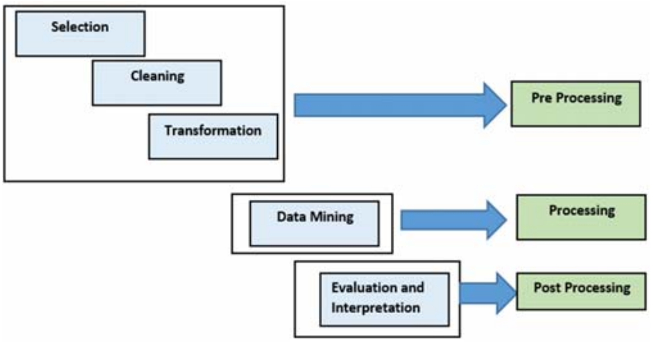
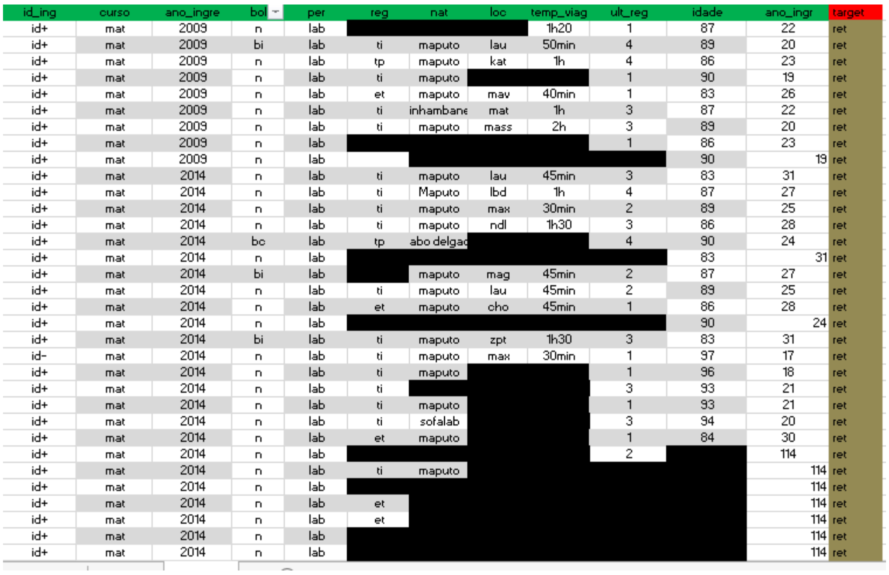


Fig. 1 Data preprocessing framework

A. Data Selection and Cleaning

388 lines (instances) were selected, 173 from 2009 and 215 lines from 2014, respectively.

Cleaning and harmonization were carried out, removing lines with missing data or data that could somehow confuse or lead to an algorithm modeling error. This process resulted in the elimination of a total of 148 lines, reducing the size of the initial dataset by 36.02%.



id_ing	curso	ano_ingre	bol	per	reg	nat	loc	temp_viag	ult_reg	idade	ano_ingr	target
id+	mat	2009	n	lab				1h20	1	87	22	ret
id+	mat	2009	bi	lab	ti	maputo	lau	50min	4	89	20	ret
id+	mat	2009	n	lab	tp	maputo	kat	1h	4	86	23	ret
id+	mat	2009	n	lab	ti	maputo			1	90	19	ret
id+	mat	2009	n	lab	et	maputo	mav	40min	1	83	26	ret
id+	mat	2009	n	lab	ti	inhambane	mat	1h	3	87	22	ret
id+	mat	2009	n	lab	ti	maputo	mass	2h	3	89	20	ret
id+	mat	2009	n	lab					1	86	23	ret
id+	mat	2009	n	lab						90	19	ret
id+	mat	2014	n	lab	ti	maputo	lau	45min	3	83	31	ret
id+	mat	2014	n	lab	ti	Maputo	lbd	1h	4	87	27	ret
id+	mat	2014	n	lab	ti	maputo	max	30min	2	89	25	ret
id+	mat	2014	n	lab	ti	maputo	ndl	1h30	3	86	28	ret
id+	mat	2014	bc	lab	tp	abo delgad			4	90	24	ret
id+	mat	2014	n	lab						83	31	ret
id+	mat	2014	bi	lab		maputo	mag	45min	2	87	27	ret
id+	mat	2014	n	lab	ti	maputo	lau	45min	2	89	25	ret
id+	mat	2014	n	lab	et	maputo	cho	45min	1	86	28	ret
id+	mat	2014	n	lab						90	24	ret
id+	mat	2014	bi	lab	ti	maputo	zpt	1h30	3	83	31	ret
id-	mat	2014	n	lab	ti	maputo	max	30min	1	97	17	ret
id+	mat	2014	n	lab	ti	maputo			1	96	18	ret
id+	mat	2014	n	lab	ti				3	93	21	ret
id+	mat	2014	n	lab	ti	maputo			1	93	21	ret
id+	mat	2014	n	lab	ti	sofalab			3	94	20	ret
id+	mat	2014	n	lab	et	maputo			1	84	30	ret
id+	mat	2014	n	lab					2		114	ret
id+	mat	2014	n	lab	ti	maputo					114	ret
id+	mat	2014	n	lab							114	ret
id+	mat	2014	n	lab	et						114	ret
id+	mat	2014	n	lab	et						114	ret
id+	mat	2014	n	lab							114	ret
id+	mat	2014	n	lab							114	ret

Fig. 2 Dataset Cleaning

Open Science Index, Computer and Information Engineering Vol:18, No:5, 2024 publications.waset.org/10013637.pdf

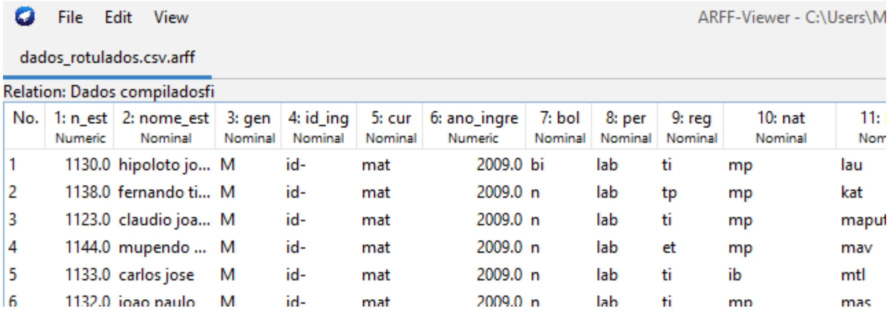
B. Data Transformation

In this step, the data obtained from the previous step were converted into nominal and non-nominal variables. The "objective class" was also defined, and is here called "target".

The final dataset used following the Weka format is given in Fig. 3.

TABLE I
DATASET ATTRIBUTES

Attribute	Codification	Description	possible values	Type
Student number	n_est	Represents the student number	Due to ethical issues, the values of these two fields were not presented	Numeric
Name of the student	name_est	Represents the student's name		Nominal
Gender	gen	Student's gender	"M" or "F"	Nominal
Age at which you joined	id_ing	Entry age in relation to the normal age established in the ENS	"+d" or "-d"	Nominal
Course you joined	cure	Course in which admitted	"mat", "est", "inf" and "cig"	Nominal
year of entry	Year_ingr	year of entry	"2009" or "2014"	Numeric
Type of Scholarship	ball	Whether the student is a scholarship holder or not	"bc", "bi" and "n"	Nominal
Period	Per	Study period.	"lab" or "pl"	Nominal
Regime	Register	Time dedicated to studies	"ti", "et", "et/ti" and "tp"	Nominal
Naturalness	nat	student's origin	infinite	name
Location	location	Location in relation to college	See the attachment	
Time of travel	T_viag	Travel time from usual location to college.	infinite	Nominal
Last records/DMI	ult_reg		infinite	String
Year of birth	born	Year of birth	infinite	Numeric
Target variable	target	Student status.	"ret" or "ev"	Nominal



Relation: Dados compiladosfi

No.	1: n_est Numeric	2: nome_est Nominal	3: gen Nominal	4: id_ing Nominal	5: cur Nominal	6: ano_ingre Numeric	7: bol Nominal	8: per Nominal	9: reg Nominal	10: nat Nominal	11: Nom
1	1130.0	hipoloto jo...	M	id-	mat	2009.0	bi	lab	ti	mp	lau
2	1138.0	fernando ti...	M	id-	mat	2009.0	n	lab	tp	mp	kat
3	1123.0	claudio joa...	M	id-	mat	2009.0	n	lab	ti	mp	maput
4	1144.0	mupendo ...	M	id-	mat	2009.0	n	lab	et	mp	mav
5	1133.0	carlos jose	M	id-	mat	2009.0	n	lab	ti	ib	mtl
6	1132.0	joao paulo	M	id-	mat	2009.0	n	lab	ti	mp	mas

Fig. 3 Final dataset used in the data mining phase

VI. RESULTS AND DISCUSSION

We started with three experiments subdivided into three groups, where the number of folds, the parameter k, was assigned different values, using the index 1 for the KNN and 10 iterations in the Logistic Regression. In order to analyze the performance of different classifiers in the context of a classification problem, we use the accuracy metric (ACC) and the precisions (PRE) as the basis for the analysis. Additionally, the Time to Build the Model (TBM corresponding to TPCM in Portuguese) was taken into consideration, which represents the computational cost of the model. At Subections *A* are the results obtained by the classifiers in both, non-ensemble and ensemble methods.

A. Results with no Ensemble Methods

In this section, we present the results obtained without utilizing Ensemble Methods. By focusing solely on individual models and their performance, we aim to provide a comprehensive understanding of the inherent capabilities and limitations of each technique employed in our study.

B. Results with Ensemble Methods

The analysis of results from group A at Subsection *A* shows that the LR had good performance, while the KNN and the RF presented a lag, therefore, the application of the ensemble methods was pursued. Moreover, inspired by set theory, the ensemble methods present themselves as a combination of several classification models, thus bringing the possibility of increasing the performance of the classifiers by reducing variance, as each new model is influenced by the performance of the previous one.

The study presented in [12] have also tested Bagging, Boosting and Stacking, with their results showing better performance with Bagging, with regard to accuracy and smaller errors for FP. These results are also obtained in this research. On the other hand, [13] obtained better performance with Stacking in 9 out of 10 tests performed. Taking these two studies into account, Bagging and Stacking were used to increase model performance in this study. Thus, for models presented in part B, ensemble methods were applied to improve the performance of the classifiers, with the aim of minimizing the False Negative Rate (FP) observed in the experiments of group A.

Open Science Index, Computer and Information Engineering Vol:18, No:5, 2024 publications.waset.org/10013637.pdf

TABLE II
Results of Experiment I: k = 3

	ACC	TP Rate	FP Rate	PRE	Recall	F-Measur	MCC	ROC Area	PRC Area	TPCM
RF	0.878543	0.879	0.701	0.852	0.879	0.860	0.244	0.883	0.917	0.08
KNN	0.959514	0.960	0.130	0.960	0.960	0.960	0.805	0.975	0.976	0,000
LR	0.995951	0.996	0.001	0.996	0.996	0.996	0.980	0.999	0.999	0,000

TABLE III
Results of Experiment II: k = 5

	ACC	TP Rate	FP Rate	PRE	Recall	F-Measur	MCC	ROC Area	PRC Area	TPCM
RF	0.874494	0.874	0.732	0.843	0.874	0.854	0.202	0.869	0.916	0.01
KNN	0.959514	0.960	0.130	0.961	0.960	0.960	0.805	0.975	0.976	0.010.
LR	0.995951	0.996	0.001	0.996	0.996	0.996	0.980	1,000	1,000	0.06

TABLE IV
Results of Experiment II: k = 5

	ACC	TP Rate	FP Rate	PRE	Recall	F-Measu	MCC	ROC Area	PRC Area	TPCM
RF	0.88664	0.887	0.700	0.860	0.887	0.866	0.276	0.909	0.928	0.23
KNN	0.975709	0.976	0.034	0.973	0.973	0.976	0.889	0.936	0.980	0.00
LR	0.991903	0.992	0.001	0.992	0.992	0.992	0.962	0.998	0.998	0.09

TABLE V
Results of Experiment I: Bagging with k = 3

	ACC	TP Rate	FP Rate	PRE	Recall	F-Measure	MCC	ROC Area	PRC Area	TPCM
RF	0.874494	0.874	0.764	0.838	0.874	0.849	0.170	0.883	0.918	0.63
KNN	0.951417	0.951	0.131	0.955	0.951	0.953	0.53	0.988	0.985	0.02
LR	1.00	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0.69

TABLE VI
Results of Experiment II: Bagging with k = 5

	ACC	TP Rate	FP Rate	PRE	Recall	F-Measur	MCC	ROC Area	PRC Area	TPCM
RF	0.87449	0.874	0.746	0.838	0.874	0.849	0.170	0.897	0.925	0.29
KNN	0.97166	0.972	0.035	0.975	0.972	0.973	0.873	0.992	0.994	0.00
LR	1.00	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0.29

TABLE VII
Results of Experiment III: Bagging with k = 7

	ACC	TP Rate	FP Rate	PRE	Recall	F-Measur	MCC	ROC Area	PRC Area	TPCM
RF	0.88664	0.879	0.763	0.842	0.879	0.852	0.186	0.916	0.932	0.16
KNN	0.975709	0.976	0.034	0.978	0.976	0.976	0.889	0.991	0.988	0.00
RL	1.00	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0.23

TABLE VIII
Results of Experiment I: Stacking with k = 3

	ACC	TP Rate	FP Rate	PRE	Recall	F-Measu	MCC	ROC Area	PRC Area	TPCM
RF	0.88664	0.887	0.887	?	0.887	?	?	0.488	0.797	0.01
KNN	0.88664	0.887	0.887	?	0.887	?	?	0.512	0.801	0
RL	0.995951	0.996	0.001	0.996	0.996	0.996	0.980	1,000	1,000	0.02

TABLE IX
Results of Experiment II: Stacking with k = 5

	ACC	TP Rate	FP Rate	PRE	Recall	F-Meas	MCC	ROC Area	PRC Area	TPCM
RF	0.887	0.887	?	0.887	0.887	?	?	0.477	0.794	0.02
KNN	0.88664	0.887	0.887	?	0.887	?	?	0.487	0.796	0
RL	0.995951	0.996	0.032	0.996	0.996	0.996	0.980	0.965	0.979	0.32

TABLE X
Results of Experiment III: Stacking with k = 7

	ACC	TP Rate	FP Rate	PRE	Recall	F-Measu	MCC	ROC Area	PRC Area	TPCM
RF	0.88664	0.887	0.887	?	0.887	?	?	0.497	0.798	0
KNN	0.88664	0.887	0.887	?	0.887	?	?	0.487	0.796	0
RL	1.00	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0.02

Open Science Index, Computer and Information Engineering Vol:18, No:5, 2024 publications.waset.org/10013637.pdf

The model built using the LR algorithm has the highest accuracy score in all experiments of part B, with values greater than 99%. KNN ranked second, with accuracy close to 96%, while RF had the lowest score, with less than 90% accuracy. As for TPCM, RL was the fastest in building the model (for k = 7 and TPCM = 0.23 sec), followed by KNN and, finally, the RF model took the longest time to be built. Thus, Bagging provided significant improvements in the results, reducing the TPCM and improving the accuracy of the classifiers.

The results found in this study also show that increasing the parameter k (folds) influences the performance of the model, with higher values of k resulting in better results. In addition, Bagging strategy contributed towards improving the classifiers and error reduction. Therefore, the results obtained by Logistic Regression combined with the Bagging ensemble method was found to be the best model for the posed problem, showing the best results in terms of accuracy and model construction time.

A deeper analysis of the results shows that retention is the most widespread phenomenon among DMI students, with around 80% of the sample presenting a situation of subjects in arrears and 20% of these, being the non-completion of the final work of the course.

VII. Conclusion

This work aimed to develop a predictive model to identify the profile of DMI students, with a tendency to retention and/or dropout, using data mining techniques. For this, a number of factors were considered, both socio-demographic and related to the internal rules of the DMI and University. These data were obtained and compiled from DAU and DMI records.

The application of data mining allowed the identification of patterns in the students' attributes, which were used to model the profile of students with a tendency to retention and dropout. Three classification algorithms were applied and evaluated using cross-validation as a validation technique and accuracy, precision, renewal and ROC curve as evaluation metrics, with more emphasis on accuracy, as it evaluated the percentage of correct answers between the number of hits and total entries.

The KDD process was used for the selection and extraction of useful knowledge for the design of the predicative model. However, the process was not linear, due to gaps in the selected data, which were overcome by cleaning and normalizing the data.

The results showed that the variation of the k parameter influences the performance of the models. Furthermore, they show that the ensemble methods significantly improve the performance of the models.

References

[1] M. A. Caldeira and A. M. Guerreiro, "Information's Systems". 2nd ed. Lisbon, FCA. 2004.

[2] C. Camilo and J. Silva, "Data Mining: Concepts. Tasks, Methods and Tools". Goiânia: Ufg, 2009.

[3] A. Mcafee and E. Brynjolfsson. "Big Data: The Management Revolution" Harvard Business Review Brasil, São Paulo, v.21, October-October 2012. Available at: https://hbrbr.uol.com.br/edicoes-anteriores/outubro-2012/ Accessed on: 03/22/2019.

[4] L. P. Fávero. "KDD and Data Mining: Concepts Only". Retrieved from IT FORUM: https://itforum.com.br/colunas/kdd-e-data-mining-mais-do-que-çou-conceitos/. April, 2019

[5] O. N. P. Cardoso and R.T.M. Machado, "*Knowledge management using data mining: a case study at the Federal University of Lavras*". Journal of Public Administration, v. 42, no. 3, p. 495-528. 2008.

[6] S. Filho and R.L. Roberto, "*Evasion in Brazilian higher education*". Research Notebooks. 2007.

[7] M. L. Gisi, "Higher Education in Brazil and the unequal character of access and permanence". Educational Dialogue, Curitiba, v. 6, no. 17, p. 97-11. 2006.

[8] Peuem. "*Eduardo Mondlane University strategic plan*". Maputo". UEM. 2017.

[9] M. Tsiakmaki, G. Kotsiantis and S. Ragos, "*Implementing AutoML in Educational Data Mining for Prediction Task. Applied Science*". Patras. 2019.

[10] H. L. B. Da Rocha and J. O. Dos Santos. "School failure: Limits to citizenship". Brazilian Journal of Education and Health, 5(4), 36-42. Retrieved from: http://www.gvaa.com.br/revista/index.php/REBES/article/view/4117 (Links). 2016.

[11] U. M. Fayyad, G. Piatetsky-Shapiro and P. Smyth. "From data mining to knowledge discovery in databases". Artificial Intelligence Magazine, v. 17, no. 3, p. 37-54.1996. 1996.

[12] T. Lobato and E. Carvalho, E. "Proposal for an Ensemble Model for Credit Scoring". Brazilian Journal of Development. 2021

[13] R. Rossi, and F. Perreira. "Study of Ensemble Techniques for Data Classification". Mato Grosso do Sul. 2017.