International Journal of Computer Applications (0975 - 8887)
Volume 176 - No.34, June 2020

A Combined Model based on Clustering and Regression to Predicting School Dropout in Higher Education Institution

Marilia N. C. A. Lima, Wedson L. Soares, Iago R. R. Silva, Roberta A. de A. Fagundes
Department of Computer Engineering
University of Pernambuco, Brazil

ABSTRACT

School dropout is a frequent problem in Brazil that has a professional and personal impact. The governing authorities seek to reduce this problem in education. Thus, the identification of the factors that cause the dropout rate and its prediction in higher education institutions are difficult tasks. Therefore, three combined models are proposed that use groupings and regression predict school dropouts in Higher Education Institutions (HEIs) in Brazil. The proposed models make the combination of algorithms, K-means with Linear Regression (LR), K-means with Robust Regression (RR), and K-means with Support Vector Regression (SVR). Four classic algorithms for evaluating our combined models (SVR, Bagging, LR, RR) are selected for comparison. The methodology utilized in this work was the Cross-Industry Standard Process for Data Mining (CRISP-DM). A comparative analysis performed with classic algorithms presents the efficiency and reliability of the proposed models for the school dropout problem.

General Terms

Data Mining, Regression, Education, School dropout

Keywords

Education Data Mining, Combined Model, Clustering, Regression, Predicting, School dropout.

1. INTRODUCTION

School dropout is one of the most frequent problems in the area of education. It has an impact whether in management or socioeconomic factors in the lives of young people [28]. From the manager's point of view, they have looked for methods to identify the factors that cause the problem. Identifying these factors, it's possible to predict this problem based on previous knowledge. The techniques of machine learning and data mining can help in decision making. Then it's possible to take action on this problem before the dropout happens.
Educational Data Mining (EDM) helps the prediction of matters related there is the educational area [22] [11]. Among the EDM techniques, there is clustering and regression. Clustering techniques have the goal of clustering where the data have the most significant similarity [14]. Already Regression techniques analyzing the relationship between variables so that an independent variable (X) explains a predictive variable (Y) [18]. The main difference between the clustering and regression techniques is the learning process. The regression has a response variable (Y) associated with the independent variables (X) (supervised learning) while the clustering performs unsupervised learning already which does not have a Y related to its attributes [12].
There are different types of clustering algorithms such as partitions and hierarchies. In partitioned algorithms, the K-means is widely used and easy to implement [13]. The regression techniques are divided into parametric and non-parametric. Among non-parametric regression techniques, there is the Support Vector Regression (SVR) [27]. The SVR algorithm presented good results for EDM problems compared to other non-parametric regression algorithms [8].
In data mining is also used ensemble. Ensemble aims to integrate basic models to generate a final output [17]. When the ensemble used, it is sought to increase the generalization power of the model. Ensemble technologies used in classification and regression problems.Among these techniques, the Bagging is used. Bagging algorithm performs the data parsing through bootstrap techniques and combines the generated models obtaining a final result Among these techniques, the Bagging is used. Bagging algorithm performs the data parsing through bootstrap techniques and combines the generated models obtaining a final result [4].
Some of these learning techniques in their basic version present some fragility. For example, when the data aren't scattered a linear regression technique will be very suitable for the problem. In this way, the development of a combined approach has been used in the resolution of several problems. Combined approaches used the strengths of each technique [2] [21]. In work [23], the authors used clustering and regression models for the prediction of death and length of stay in the intensive care unit. The authors concluded that clustering before regression analysis improved prediction accuracy.
In this context, this article aims to develop a combined model that uses clustering and regression in the context of data mi-ning to predict school dropout in HEI in Brazil. The proposed models make the combination of K-means with regression techniques (LR,RR and SVR).The classic algorithms (SVR, Bagging, LR, RR) are used to compare with the proposed combined models.
The contributions of this work are the development of the models that combines clustering and regression in the context of data mining for school dropout prediction. Moreover, compari-

son with parametric and non-parametric regression techniques in the context of data mining, choosing the factors that contribute to the dropout school. The combined models are built with the intention:

—Decrease the variances of the data, then create homogeneous regions;

—Use classic algorithms in these homogeneous regions to provides an efficient prediction model.

The rest of this work is organized as follow. Section 2 presents the background on machine learning approaches on school dropout. Section 3 presents in detail the proposed model. Section 4 describes the methodology used in this work. Section 5 presents the results obtained, and a comparison of our model with another techniques. Finally, section 6 presents the principal contributions and limitations of this work.

2. RELATED WORKS

In this section, a description is presented of some related work in the EDM area for predicting abandonment, in addition to showing the difference of our work to the others.

There are two ways to predict school dropout. The first is u-sing supervised classification techniques, and the second is using regression techniques. Classification techniques based on using databases that contain X and Y variables. The X are the factors of the problem labeled data, in this case, the socioeconomic or manager factors. The Y is the target, in this case, if the student has dropped out or no. Similarly is the regression, that analyses the relationship of variables explains (X), the predictive variable (Y).

In [19], a structure is proposed to manage the prediction of student performance using Learning Analytics (integrates data analysis and data mining techniques). For this, the authors use linear regression techniques to predict student performance. The results show that students with excellent performance in mathematics courses are more likely to acquire good results in other computer science courses.

The works of Sara et al. [25], Marquez et al. [16], and Sansone et al. [24] use classification approaches to predict the school dropout. The goal is to detect patterns in labeled databases and classify new unknown cases for school dropout in different databases. The experiments were performed with classic machine learning algorithms to predict the school dropout, e.g., Random Forest, Support Vector Machines (SVM), and K-Nearest Neighbors (K-NN). After that, the works got the accuracy results upon 90%, and it's good results for classification tasks.

As a regressive model, the work in [10] They proposed a dropout prediction using logistic regression. The perspective is longitudinal; that is, it takes into account data that depend on time. Data like age, attendance, gender, and test score used. The goal was to perform the early dropout prediction on students. The model can warn a manager about the critical situation of the student for decision making. The error rate calculated with the predictions was 0.12.

The work in [15] used fuzzy regression discontinuity for dropout prediction. The authors predicted using administrative data. The case study was to predict the dropout of students in higher education. The conclusion the authors get is that students suffering from high school dropout can pass an admission examination at the academy. But, they can not complete the course.

Another work [8] presents a model for predicting school dropout using regression models based on SVR and quantile regression. For the accomplishment of the work, it follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology [5]. The database used by INEP, which contains data from Brazilian students. After applying the regression techniques, the absolute mean error metric is evaluated. With the analysis of the results, the authors concluded that obtained with SVR were more significant, with an error equals 0.015665.

The work in [7] presents an approach of prediction of school dropout and disapproval using linear and robust regression models. The work also follows the CRISP-DM methodology. This work uses the educational bases of elementary school students from the State of Pernambuco (Brazil) provided by INEP. With the application of the prediction model using the previously cited regressors, the absolute mean error metric was used for evaluation. The authors concluded that the robust regression, with an error of 0.0306, obtained better results in the estimates.

In the work of [26], ensemble models were used to estimate dropout in students, the authors proposed bagging models with linear regression, bagging with robust regression, bagging with ridge regression and compared with a stacking model. It was concluded that the bagging models had lower prediction errors than the stacking model. Thus, the problem of dropping out of school has good accuracy in estimating using a bagging approach.

The previous works are limited to apply unique classic algorithms for the prediction of dropout. Therefore, it is proposed to develop a combined model. This model is based on cluster and regression applied to that problem. Thus, the model provides the identification of homogeneous regions through the clustering algorithm. The regression algorithm in these homogeneous regions can provide an efficient forecasting model.

3. PROPOSED MODEL

In this section, the Proposal Model (PM) developed in the present work presented. The PM was combined with two EDM techniques and was composed of five steps, according to Fig. 1.

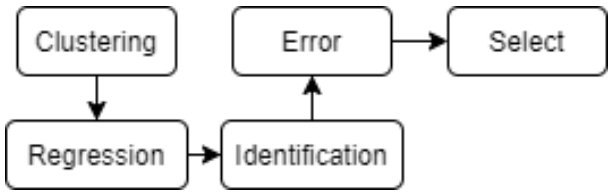


Fig. 1: Steps of Proposed Model used to conduct work

—**Clustering:** In step, the clustering process performed according to the amount of cluster (K) defined as a parameter for the K-means algorithm. The clustering process is performed of value two until the maximum value is set.

—**Regression:** In this step, for each formed cluster, a regression model is constructed; that is, each group has a specific regression model. This process happens for up to the maximum value of K defined.

—**Identification:** In this step, to test the PM, it is necessary to identify to which cluster the test data is more similar. To do this, it uses the smallest Euclidean distance of the test data concerning the centroid of a given cluster. Thus, groups of test data formed about the clusters built with the training data. The Euclidean distance is calculated by equation 1.

Where the points are n-dimensional $P = (p_1, p2, ..., p_n)$ and $Q = (q_1, q2, ..., q_n)$

$$distance = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2} \quad (1)$$

—**Error:** In this step, to calculate the error of the PM. Thus, it used the mean square error (MSE) and Mean Absolute Error (MAE). To evaluate our PM use the following equation 3:

$$MSE_{PM} = \frac{\sum e_t}{c_t} \quad (2)$$

$$MAE_{PM} = \frac{\sum e_t}{c_t} \quad (3)$$

Where:
MSE_{PM} is the general MSE of the PM,MAE_{PM} is the general MAE of the PM, e_t is the error of each cluster in the test group, and c_t is the number of test groups.

—**Select:** In step, a search for the lowest error for each K that PM used in the clustering process performed. Finally, the value of K is returned that created the best similarity region of the data concerning the lowest prediction error obtained.

4. METHODOLOGY

In this section, the process of CRISP-DM is used [5]. The steps of this methodology are business understanding, data understanding, data preparation, modeling, evaluation, deployment. Fig. 2 shows the steps of CRISP-DM. This metho-dology is used in the area of data mining and has been bringing satisfactory results. The process of CRISP-DM is interactive and iterative. Thus, the activities of each step of this methodology described in the following subsections.

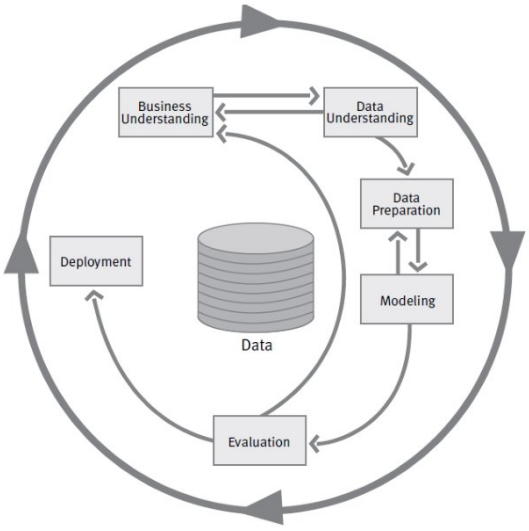


Fig. 2: Steps of CRISP-DM methodology [5]

4.1 Business Understanding

This step consists in understanding the problem to be studied. This work will consider the possible factors of HEI that influence school dropout. The dataset used was that of INEP [1]. The data are referring to the undergraduate courses of HEI obtained from the census of higher education and the data of flow indicators of higher education of the year 2015.

4.2 Data Understanding

Two datasets are used to form a single one are the dataset of the Higher Education Census (HEC) and Indicators of Higher Education Flow (IHEF). The dataset of the HEC contains information about HEI. Already the dataset IHEF attributes statistical value to the quality of teaching.

4.3 Data Preparation

In this step, the data are prepared for use. In the present work, is use the following preparations:

—Join databases;

—Checking for categoric values;

—Filtering the data;

—Checking for missing or blank values;

—Data normalization (between 0.15 and 0.85); and

—Selection of variables (stepwise).

For the process of joining the two datasets, the dependent variable is the dataset IHEF, while the explanatory variables in the dataset HEC that contains the information about undergraduate courses. The datasets were merged so that only one referring to the year 2015, with the suitable variables for this study. The two datasets were combined according to the variable that represented the unique identification code of each course. The unique course code is present in both databases. In filtering the data, a stratified sampling of 10 % of the data set is used and obtained 134877 instances. Is Checked if categorical data existed and converted to numeric.
The variable that refers to the school dropout contained 115445, not existing values; these values are excluded. Thus, the dataset has 19432 instances and 45 attributes. The other characteristics of the base also contained not exist values and replaced by the median.
The data are normalize between 0.15 and 0.85 based on [8] [6] and use the stepwise selection method to select the variables for the study. Table 1 shows the 13 variables selected.

Table 1. : Variables selected for the study and prediction of the techniques used for comparison with the Proposed Model

Ref	Variable	Description
CAS	CO_ALUNO_SITUACAO	The situation of the student's in the course (active, locked enrollment, Unlinked from the course, Transferred to another course of the same IES, Formed, deceased)
IITE	IN_ING_TRANSF_EXOFFICIO	Informs if the student gets on the course utilizing Ex-officio Transfer
IAA	IN_APOIO_ALIMENTACAO	Informs if the student get food support
IABP	IN_APOIO_BOLSA_PERMANENCIA	Informs if the student get financial aid
IAMD	IN_APOIO_MATERIAL_DIDATICO	Informs if the student gets support for the acquisition of academic material
IIT	IN_INGRESSO_TOTAL	Informs if the student is in the course
QI	QT_INGRESSANTE	Number of new students
QP	QT_PERMANENCIA	Number of students who persists in the course
QF	QT_FALECIDO	Number of deceased students
IAP	IAP	Indicator the accumulate permanence
ICA	ICA	Indicator the accumulate conclusion
IMC	IN_MATUTINO_CURSO	Study morning period
INC	IN_NOTURNO_CURSO	Study night period
TDA	TDA	School dropout values

By analyzing Table 1, the variables related to student financial support were selected (IAA, IABP, IAMD). Financial assistance

may be a differential for students of socioeconomic vulnerability. It can be an essential factor so that the student can have permanence in the course. Thus, school dropout often is associated with socioeconomic disadvantages.
Table 2 shows the values of Pearson's correlations for the variables selected for the study. The highest correlations were the IAP and ICA variables of -0.44130240 and -0.76434023, respectively. These indicators are related directly to the variable TDA. Thus, the higher the permanence of the student, the less the school dropout. The variable IMC is related to dropping out of school, since usually people who study in the night period work in parallel to complete the course. The full-time courses may show a more significant drop as students can not afford to stay in university.

Table 2. : Pearson's Correlation value of the selected variables in relation to the school dropout rate (TDA) variable

Ref Variables	Correlation Value
CAS	-0.01703789
IITE	-0.00345445
IAA	-0.01434576
IABP	0.01882726
IAMD	0.02179449
IIT	0.01618965
QI	0.05162753
QP	-0.17793200
QF	-0.01577863
IAP	-0.44130240
ICA	-0.76434023
IMC	0.07534873
INC	0.17525950
TDA	1

4.4 Modeling

In this step, it uses the techniques of the SVR, LR, RR, Bagging (LR and RR, SVR), and our PM (LR, SVR, and RR). In the experiments, 30 iterations are performed, and the method of partition used will be the holdout (75% training and 25% test) [20].
To understand the equations described in the modeling, the following variables are used: y_i is the actual value $\hat{y}_i$ is the value estimated by the models, i being the value of each instance up to n. The β estimates are determined by minimizing an objective function for all β, and ϵ the error associated with the model.

—**LR** - the relationship between variables is a function linear (Multiple Linear Regression). Equation 4 presents the multiple linear regression model [18]. Where,α represents the intercept of the axis with the y, and $\beta = (\beta_1, ..., \beta_p)$ it is a vector of parameters that represents the variation of y as a function of the variation of x. And $x = (x_1, ..., xi)$ a vector of explanatory variables. Finally, $\epsilon = (\epsilon_1, ..., \epsilon_n)$ a vector error message.

$$Y_i = \alpha + \beta_1 x_{1i} + ... + \beta_p x_{pi} + \epsilon_i \ (i = 1, 2, ..., n) \tag{4}$$

—**RR**- consists of finding estimators that are more efficient when there are small *outliers* in the sample distribution [9].Equation 5 presents the robust regression model. . Where the ρ function provides the contribution of each residue to the objective function

$$y_i = \sum_{i=1}^{n} \rho(e_i) = \sum_{i=1}^{n} \rho(y_i - x_i'\beta) \tag{5}$$

—**SVR** - this algorithm generates the maximum support number of vectors with small error values to separate the data at the highest margin. It follows the same proposal as the SVM (Support Vector Machine), but for regression. The algorithm uses a kernel function [6]. Equation 6 presents the Kernel function. Where the γ parameter controls function flexibility kernel.

$$exp\left(-\gamma \left\| x_i - x_j \right\|^2\right) \tag{6}$$

—**Bagging** - also called the Bootstrap Aggregation, was developed by [3]. The Bagging aims to create several sets of training data using the bootstrap technique, later constructing regression models. Finally, it generalizes the model by a mean [3]. Equation 7 presents the generalization of this ensemble.. Where, $fbagging(x)$ is the prediction of the combined model for the instant x, M is the amount of model regressors and, $\hat{f}(x)$ is the prediction given by the regressor the sample.

$$fbagging(x) = \frac{1}{M} \sum_{i=1}^{M} \hat{f}(x) \tag{7}$$

4.5 Evaluation

In this step, the results obtained are evaluated, verifying their relation with the objective of the work. To assess the accuracy of the prediction models, using two standard performance measures: Mean Absolute Error (MAE) and Mean Squared Error(MSE) as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \hat{y}_i \right| \tag{8}$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \tag{9}$$

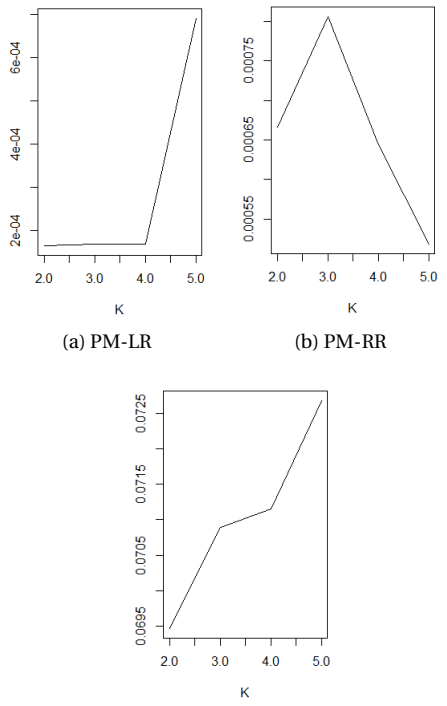
4.6 Deployment

In this work, the steps of the CRISP-DM methodology are considered. Thus, it is obtained knowledge and elaborate strategies for the scenario of the problem study.

5. RESULTS AND DISCUSSION

This section presents the results obtained and a comparison of the proposed models with other techniques. The techniques used for this work applying the Kernel function "RBF". The value parameter of K equals 5 for PM. This number of K represents a maximum partitioning that the K-means algorithm used to carry out the clustering process. Three different models are used to analyze the factors influencing school dropout and also to check PM performance
Fig. 3 shows the number of K in relation to the mean of the prediction error of the 30 iterations. It demonstrates that the similarity that the clustering process accomplishes makes it possible to have more accurate regression models for the problem of school dropout. Thus, it was possible to observe that there is a creation of regions of acceptance for the data since the same group data are more similar.
The formation of groups with the educational context makes it clear that the diversification of clusters (higher number of clusters) has a lower accuracy of the PM. In this way, there will be less

data for each model to be trained and to have better generalization. Therefore, the accurate and early identification of student dropout is more efficient when used by fewer clusters.



(a) PM-LR (b) PM-RR

(c) PM-SVR

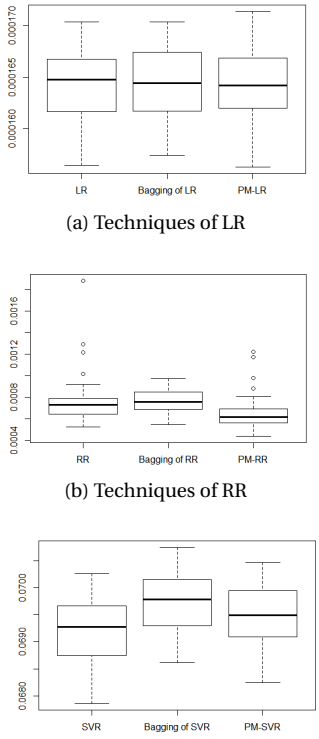
Fig. 3: Number of K in relation Mean MAE of PM

Table 3 shows the error averages and between parentheses the standard deviation of the 30 runs. The error values are for the two metrics used.

Table 3. : Mean and Standard Deviation of the MAE and MSE errors of the techniques used in the study

Technique	MSE	MAE
LR	$1.08X10^{-7}$ ($2.65X10^{-9}$)	0.00016 ($3.81X10^{-6}$)
RR	$1.54X10-5$ ($4.58X10^{-6}$)	0.00079 (0.000269)
SVR	0.00708 (0.000119)	0.0692 (0.00066)
Bagging of RL	$1.084X10^{-7}$ ($2.639X10^{-9}$)	0.00016 ($3.77X10^{-6}$)
Bagging of RR	$1.397X10^{-5}$ ($4.141X10^{-6}$)	0.00076(0.00011)
Bagging of SVR	0.0071 (0.00011)	0.0697 (0.00063)
PM -LR	$1.0887X10^{-7}$ ($2.806X10^{-0}$)	**0.000164** ($4.0005X10^{-6}$)
PM-RR	$4.9945X10^{-6}$ ($1.7807X10^{-6}$)	**0.000518** ($8.6289X10^{-5}$)
PM -SVR	0.00714 (0.00012)	0.06946 (0.0006)

The experiments carried out with the use of SVR presented more significant errors than the linear regression techniques. Thus, it can justify that the data are more suitable for using parametric techniques.
Fig. 4 shows the boxplot the models relation the LR, RR, and SVR, respectively. Visually analyzing the PM-RR has a lower variability than RR and Bagging-RR. The PM-LR presents a similar variability to the LR and Bagging-LR techniques. The SVR algorithm presented a lower variability than PM-SVR. In the educational context for predicting student dropout, these results show that models that use LR are better estimated. Therefore, its use can follow up to make more strategic decisions for students to stay at the university.



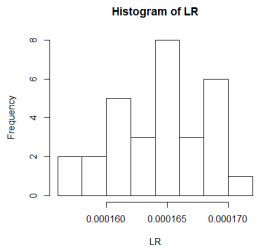
(a) Techniques of LR

(b) Techniques of RR

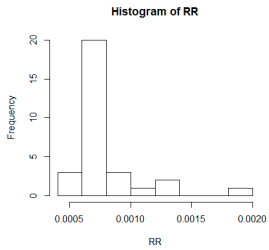
(c) Techniques of SVR

Fig. 4: Box-plots of the MAE errors of the techniques used in the study comparing with PM

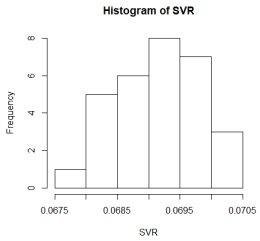
The hypothesis test was conducted to statistically verify if PM presents less errors than compared techniques (H_1). Thus, the null hypothesis is that PM has same errors with the compared techniques(H_0). Equation 10 presents the hypotheses elaborated. Where μ_1 is the PM model to be compared with the other techniques studied (μ_2) individually.
First, the error histograms of the techniques used in the study are analyzed. Figures 5 and 6 show the histograms of the MAE and MSE errors, respectively. The visual analysis of these graphs shows that the errors do not follow a normal curve and can then be categorized as non-normal. To confirm the visual analysis, the Kolmogorov Smirnov test was performed to verify the normality of the data. All the data does not follow a normal distribution, the Wilcoxon test with 95% confidence is used.



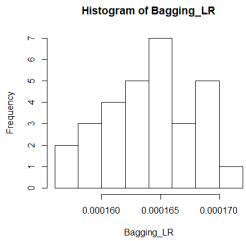
(a) LR



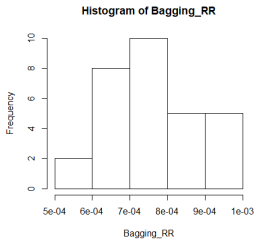
(b) RR



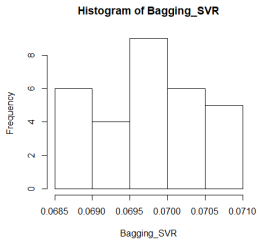
(c) SVR



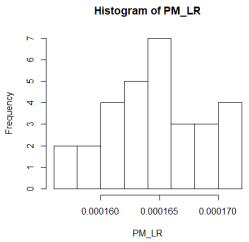
(d) Bagging of LR



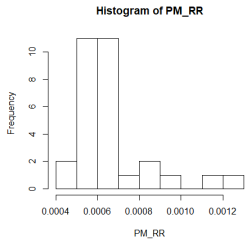
(e) Bagging of RR



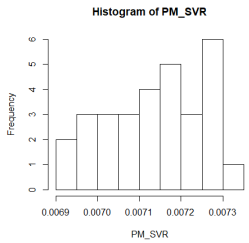
(f) Bagging of SVR



(g) PM-LR

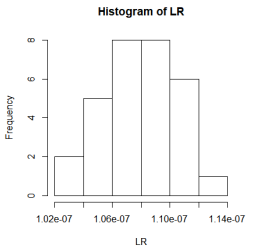


(h) PM-RR

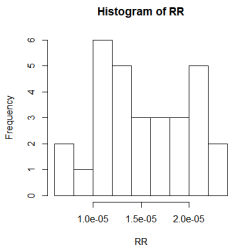


(i) PM-SVR

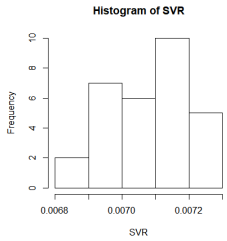
Fig. 5: Histogram of MAE errors of techniques used at work



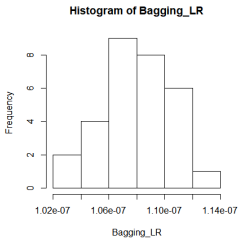
(a) LR



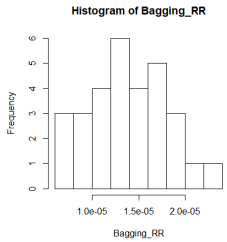
(b) RR



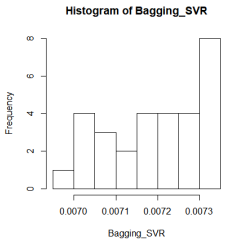
(c) SVR



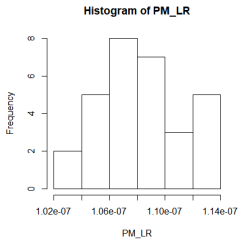
(d) Bagging of LR



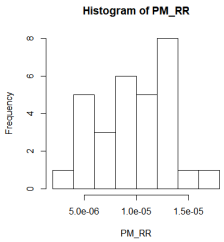
(e) Bagging of RR



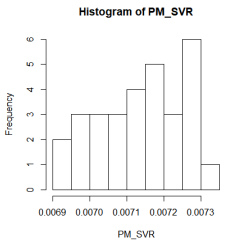
(f) Bagging of SVR



(g) PM-LR



(h) PM-RR



(i) PM-SVR

Fig. 6: Histogram of MSE errors of techniques used at work

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{cases} \quad (10)$$

Table 4 shows the p-value values Wilcoxon test of the PM-LR compared to the other techniques. The PM-LR obtained statistical evidence of lower errors than all other techniques, except LR and Bagging of LR.

Table 4. : P-value of the Wilcoxon hypothesis test comparing the PM-LR with the other techniques.

Techniques	p-value of MAE errors	p-value of MSE errors
PM-LR X LR	0.5322	0.7167
PM-LR X RR	$2.2X10^{-16}$	$1.459X10^{-11}$
PM-LR X SVR	$2.2X10^{-16}$	$1.46X10^{-11}$
PM-LR X Bagging of LR	0.5612	0.6989
PM-LR X Bagging of RR	$2.2X10^{-16}$	$1.456X10^{-11}$
PM-LR X Bagging of SVR	$2.2X10^{-16}$	$1.46X10^{-11}$
PM-LR X PM-RR	$2.2X10^{-16}$	$1.459X10^{-11}$
PM-LR X PM-SVR	$2.2X10^{-16}$	$1.46X10^{-11}$

Table 5 shows the p-value values Wilcoxon test of the PM-RR compared to the other techniques. The PM-RR obtained statistical evidence of lower errors than RR, SVR, Bagging of RR, Bagging of SVR, and PM-SVR. Therefore, only for models that use RL that the PM-RR did not have evidence of minor errors.

Table 5. : P-value of the Wilcoxon hypothesis test comparing the PM-RR with the other techniques.

Techniques	p-value of MAE errors	p-value of MSE errors
PM-RR X LR	1	1
PM-RR X RR	0.0013	$1.853X10^{-5}$
PM-RR X SVR	$2.2X10^{-16}$	$1.509X10^{-11}$
PM-RR X Bagging of LR	1	1
PM-RR X Bagging of RR	0.0002881	0.0003098
PM-RR X Bagging of SVR	$2.2X10^{-16}$	$1.509X10^{-11}$
PM-RR X PM-LR	1	1
PM-RR X PM-SVR	$2.2X10^{-16}$	$1.509X10^{-11}$

Table 6 shows the p-value values Wilcoxon test of the PM-SVR compared to the other techniques. The PM-SVR obtained statistical evidence of lower errors than SVR and Bagging of SVR. For the other techniques, the PM-SVR did not obtain evidence of minor errors. However, this model is more efficient when compared to models that use the SVR.

Table 6. : P-value of the Wilcoxon hypothesis test comparing the PM-SVR with the other techniques.

Techniques	p-value of MAE errors	p-value of MSE errors
PM-SVR X LR	1	1
PM-SVR X RR	1	1
PM-SVR X SVR	$2.2X10^{-16}$	0.9715
PM-SVR X Bagging of LR	1	1
PM-SVR X Bagging of RR	1	1
PM-SVR X Bagging of SVR	$2.2X10^{-16}$	0.03161
PM-SVR X PM-LR	1	1
PM-SVR X PM-RR	1	1

Comparing the results with the model proposed in [26], the bagging approach with robust regression, it was is observed that statistically smaller errors the PM-LR about Bagging-RR. For this reason, the PM model should present relevant results for the literature in the context of the students' school dropout process. Thus, with the results obtained by the proposed model, it is observed a higher representation capacity because the precision in the prediction is more significant than some other algorithms studied.

Based on the results, student dropout to be correlated with financial. It supports the hypothesis that students with socioeconomic vulnerability find it more challenging to finish university. Predicting student dropout in school is an essential issue in education because it concerns institutions over the entire world. Thus, this study helps to spot out those factors that cause school dropouts. Although articles have explained school dropout, few studies have attempted to predict school dropout using regression techniques. In this study, to automatically identify the relevant factors for this problem (Stepwise), machine learning techniques are used. With the appropriate combined model, it is possible to predict student dropout using the automatically selected variables that are likely to be useful in forecasting. This was achieved using regressors (combination) that tend to do better than a single individual regressor. These models produced by the combination of k-means and regression achieved better satisfactory results.

6. CONCLUSION

In this work, a method is proposed to predict school dropouts. The method consists of combined clustering and regression algorithms that apply a K-means procedure to identify homogeneous regions in terms of similarities between training instances and, secondly, trains an LR, RR, and SVR models in each region. In the prediction phase, a part of interest is selected dynamically as the most similar region to the query instance. The results prove satisfactory of the PM for the problem, which obtained functional generalization capacity.

The public datasets provided by INEP is used, and through the stepwise, selected the factors of HEI that influence school dropout. The factors choose related to financial support for students. Dropout is rarely a sudden decision on the part of the student. Thus, the selection of these factors influences the creation of public policies that seek to reduce the school dropout, since it is a frequent problem at all levels of education.

Predicting educational dropout is a significant challenge administrator and educators. The use of de EDM has good results. However, it is essential to notice that most of the current research on the application of EDM use classification. But, in the paper, is use the technic regression..

In future works, it is intended to add sociodemographic attributes to verify the factors that influence students' dropout, taking into account the students' social question—seeking to contribute to the process of teaching and learning of students in HEI. It is still intended to use other clustering algorithms in addition to different regression models (quantile regression, kernel regression) verifying the application of our combined PM.

7. ACKNOWLEDGMENTS

This research was supported by Coordination for the Improvement of Higher Education Personnel (CAPES) - Finance Code 001.

8. REFERENCES

[1] Inep. http://www.inep.gov.br/, accessed August 8, 2018.

[2] Ibrahim Berkan Aydilek and Ahmet Arslan. A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. *Information Sciences*, 233:25–35, 2013.

[3] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[4] Peter Bühlmann, Bin Yu, et al. Analyzing bagging. *The Annals of Statistics*, 30(4):927–961, 2002.

[5] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. Crisp-dm 1.0 step-by-step data mining guide. 2000.

[6] Manoel Alves de Almeida Neto, Roberta Andrade A de de Fagundes, and Carmelo JA Bastos-Filho. Using multi-objective algorithms for optimizing support vector regression parameters. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

[7] Rafaella Leandra Souza do Nascimento, Geraldo Gomes da Cruz Junior, and Roberta Andrade de Araújo Fagundes. Mineração de dados educacionais: Um estudo sobre indicadores da educação em bases de dados do inep. *RENOTE*, 16(1).

[8] Rafaella Leandra Souza do Nascimento, Ricardo Batista das Neves Junior, Manoel Alves de Almeida Neto, and Roberta Andrade de Araújo Fagundes. Educational data mining: An application of regressors in predicting school dropout. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 246–257. Springer, 2018.

[9] Roberta Andrade de Araújo FAGUNDES and Francisco José de Azevêdo CYSNEIROS. Métodos de regressão robusta e kernel para dados intervalares. 2013.

[10] Bobby J Franklin and Stephen B Trouard. An analysis of dropout predictors within a state high school graduation panel. *Schooling*, 5:1–8, 2014.

[11] Sharad Gangele, Kirti Soni, and Sunil Patil. Data mining approach towards students behavior assessment methods for higher studies. *International Journal of Computer Applications*, 181(30):11–14, 2018.

[12] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Unsupervised learning. In *The elements of statistical learning*, pages 485–585. Springer, 2009.

[13] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[14] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[15] Christopher Jepsen, Peter Mueser, and Kenneth Troske. Second chance for high school dropouts? a regression discontinuity analysis of postsecondary educational returns to the ged. *Journal of Labor Economics*, 35(S1):S273–S304, 2017.

[16] Carlos Márquez-Vera, Alberto Cano, Cristobal Romero, Amin Yousef Mohammad Noaman, Habib Mousa Fardoun, and Sebastian Ventura. Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, 33(1):107–124, 2016.

[17] Joao Mendes-Moreira, Carlos Soares, Alípio Mário Jorge, and Jorge Freire De Sousa. Ensemble approaches for regression: A survey. *ACM Computing Surveys (CSUR)*, 45(1):10, 2012.

[18] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*, volume 821. John Wiley & Sons, 2012.

[19] OD Oyerinde and PA Chia. Predicting students' academic performances–a learning analytics approach using multiple linear regression. 2017.

[20] Tapio Pahikkala, Hanna Suominen, Jorma Boberg, and Tapio Salakoski. Efficient hold-out for subset of regressors. In *International Conference on Adaptive and Natural Computing Algorithms*, pages 350–359. Springer, 2009.

[21] MR Pooja and MP Pushpalatha. A hybrid decision support system for the identification of asthmatic subjects in a cross-sectional study. In *Emerging Research in Electronics, Computer Science and Technology (ICERECT), 2015 International Conference on*, pages 288–293. IEEE, 2015.

[22] Cristóbal Romero and Sebastián Ventura. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, 2010.

[23] Mahsa Rouzbahman, Aleksandra Jovicic, and Mark Chignell. Can cluster-boosted regression improve prediction of death and length of stay in the icu? *IEEE journal of biomedical and health informatics*, 21(3):851–858, 2017.

[24] Dario Sansone. Beyond early warning indicators: High school dropout and machine learning. *Social Science Research Network*, 2017.

[25] Nicolae-Bogdan Sara, Rasmus Halland, Christian Igel, and Stephen Alstrup. High-school dropout prediction using machine learning: A danish large-scale study. In *ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 319–324, 2015.

[26] Paulo Silva, Rafaella Leandra Souza do Nascimento, Marilia Lima, Roberta Fagundes, and Fernando da Fonseca de Souza. Modelos de regressão aplicados a predição do desempenho escolar de estudantes do ensino fundamental. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 30, page 1621, 2019.

[27] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[28] Silzá Tramontina, Silvia Martins, Mariana B Michalowski, Carla R Ketzer, Mariana Eizirik, Joseph Biederman, and Luis A Rohde. School dropout and conduct disorder in brazilian elementary school students. *The Canadian Journal of Psychiatry*, 46(10):941–947, 2001.