Ensemble Regression Models Applied to Dropout in Higher Education

Paulo M. da Silva[†], Marília N. C. A. Lima[*], Wedson L. Soares[*], Iago R. R. Silva[*],
Roberta A. de A. Fagundes[*], Fernando F. de Souza[†]
[*]Department of Computer Engineering, University of Pernambuco, Brazil
[†] Federal University of Pernambuco, Brazil
pms3@cin.ufpe.br, mncal@ecomp.poli.br, wedsonlino27@gmail.com,
irrs@ecomp.poli.br, roberta.fagundes@upe.br, fdfd@cin.ufpe.br

***Abstract*—Context: School dropout is a significant challenge for the Brazilian education system. Several factors need to be corrected, and others eliminated so that students can to have access to higher education and guarantee the completion of their courses. Motivation: finding the best model to predict a specific problem is not a simple task. It's because the phenomena involved are not known, or are sophisticated modeling. Thus, combining models often produces better accuracy than individual models. Different models use this combination approach and have been applied in the context of Data Mining (MD), for prediction and classification. Objective: we propose in this study three different models to predict school dropout. These are based on Ensemble Regression. We apply the models in the context of the Brazilian Higher Education Institutions. Besides, it may help in the identification of the factors associated with dropout. For this, we used two techniques for the attribute selection: Stepwise and Pearson correlation. That techniques determine the factors related to dropout. Methodology: we used the data from the Census and Flow Indicators Higher Education. The methodology is based on CRISP-DM to understand, prepare, and model the data. We used predictive bagging methods to make a model to predict dropout. Results: the ensemble regression models proposed obtained better performance compared model literature. The ensemble model based on bagging of linear regression had a smaller prediction error. Besides, the models proposed in this study will help the educational administrators and policymakers working within the educational sector in the development of new policies that are relevant to student retention. But, the global implications of this research to practice is its ability to help in early identifying factories associated with students at risk of dropout of High Education.**

***Keywords*—Ensemble Models, Bagging, Data Mining, Prediction, Machine Learning,**

I. Introduction

It's common to relate the school dropout to the loss of students who start but do not complete their courses. Dropout is a complex phenomenon associated with non-fulfillment of expectations. It's a reflection of many causes that need to be understood. These causes are associated in the socio-economic, political, and cultural context, in the educational system and educational institutions. Dropout is an exclusion process determined by factors and internal and external variables to educational institutions. It's a phenomenon perceived both in public educational institutions and in private institutions.

In Brazil, according to data from the Census of Higher Education, dropout rates in higher education show worrying rates. Data show that 49% of students entering higher education in 2010 dropped out of courses within five years. In private institutions, the dropout reached 53%, and in public institutions, it reached 47%. They also reached 38% in the state and 43% in the federals [1].

In this context, it's essential to identify in advance the factors associated with dropout. For this, we may use Educational Data Mining (EDM) techniques. These techniques are capable of obtaining information and organizing such information into useful knowledge. EDM requires adaptations of existing methods and the development of new technologies. The nature of the data analyzed in EDM is more diverse. That's in comparison to the data used in classical Data Mining (DM) approaches. This diversity in the data represents a potential for implementation of critical resources to aid in the improvement of Education [2].

One of the main areas of development is the Machine Learning (ML) area. ML is often confused with EDM itself. That's because both share concepts and are commonly used together. It's common to see these correlated areas in the development of models that help the discovery of new patterns in data sets. ML provides many of its concepts and techniques to the area of EDM.

Among the ML techniques used in EDM for knowledge discovery in data sets, we highlight the regression models. These analyze the relationships between data variables. The regression algorithms estimate the value of a numerical dependent variable (Y) that makes use of one or more independent variables (X) [3], [4]. The function can use one or many variables to explain the prediction of the output variable. This procedure requires that a loss function be minimized over the joint distribution of all values(Y, X) [5].

The ensemble approach is used to increase the accuracy of predictive models. With the generation of combined models, it's expected that when some of these models get poor performance. The system may reduce the error by using many models [6], [7].

The ensemble regression models proposed reduce the error and or variance of the individual models. It is done by combining several of these models to create a combined

2643-6264/19/$31.00 ©2019 IEEE
DOI 10.1109/BRACIS.2019.00030

model that achieves better performance. The approach used was the CRISP-DM methodology (Cross Industry Standard Process for Data Mining) [8] applied in the context of EDM. Besides, we apply the models developed to predicting problems of student retention. We tried to estimate the variables associated with dropout through the proposed ensemble regression models.

The result of this work was empirical research. We compared the accuracy performance and efficiency of regression ensemble models. The proposed combined models predict student dropout with high accuracy and performance efficiency based on factor analysis. These factors, we considered the demographic, academic, and socio-economic information. When combined, this information will be based on the proposition of the predictive models for the dropout of a student.

This paper is organized as follows. Section 2 presents related works. Section 3 presents the theoretical framework. Section 4 presents the proposed model. Section 5 presents the results of the experiments. Final considerations and future work are in section 6.

II. Related Works

At work [9], the authors aim to forecast the dropout. They predict undergraduate courses face-to-face. It is done to visualize perspectives that allow an intense action of intervention, mitigating the process of dropout. The authors used machine learning techniques and supervised classification task. The proposed method uses students' personal, academic, social and economic information to construct the forecast models. Besides, combining several data mining models to optimize the outcome of the process. This study had a predicted evasion rate of 74%. They are contributing to the evaluation of models that allow identifying the main attributes that help in predicting the factors associated with dropout.

The work [10] addresses evasion in undergraduate courses at a private higher education institution. The paper aims to identify and evaluate the variables that interfere with dropout. With this, it is possible to act proactively and preventive in this context. The study used multivariate logistic regression analysis techniques. The proposed method makes a delineation of the profile of the student with dropout propensity to construct a statistical model that can predict. As a result, the prediction of dropout obtained a rate of 62%. The most significant variables were the avoidance of social, economic, academic performance and professional choice factors.

In [11] the authors propose the prevention of school dropout from the description of an educational data analysis. The proposed case study focuses on the detection of the abandon of students of the course of Systems Engineering (SE). Academic data form expanded and enriched through a feature engineering process. To identify the predictors of the form avoidance used the classifiers Decision Tree, Logistic Regression and Naive Bayes. The results say that the decision tree technique reached the best prediction rate. This technique obtained a 94% accuracy rate in dropout prediction. The variables that affect evasion are several semesters, the average of the course, and the accumulated performance. The results of this research serve to propose discussions with the SE faculty to improve and implement it in a productive environment.

In the work [12], the authors propose to identify school dropout patterns from several data series. The work used data from socioeconomic, academic, disciplinary and institutional types. The authors use data mining techniques, such as decision tree-based classification. The proposed method uses data from students entering the university from 2004 to 2011 to discover the socioeconomic and school dropout profiles. A prediction rate of 80% was obtained. The results show that the most significant data were: university enrollment rate, being single and living with the mother, low academic performance, teaching staff, and initial training. The identification of these factors contributes to the adoption of strategies by the university. It can minimize the academic and financial damage caused by the evasion.

In [13] examine the probability of prediction of student performance in the first semester of the Computer Science course at the Federal University of Paraíba (UFPB). The paper uses the entry notes in the institution as a basis. With the use of classifiers, authors achieved hit rates of up to 75%. They argue that identifying student performance in the first-period subjects is useful in combating the dropout produced by failure in the new disciplines.

In [5] the author evaluates an approach based on boosting-based ensemble approach, forward stage-wise additive modelling (FSAM), o improve some widely used base regressors prediction ability. They used 10 regression algorithms in four different types to make predictions on 10 diverse data from different scientific areas. Then they compared the experimental results in terms of correlation coefficient, mean absolute error, and root mean squared error metrics. Furthermore, they made use of scatter plots to demonstrate the effect of ensemble modelling on the prediction accuracy of evaluated algorithms.

In the work of [6] the ensemble method was used to construct combined regression models. The methodology of bagging and boosting ensembles with 10 sub-learners in each one. They made comparisons between the performances of bagging and boosting ensembles in 25 sub-learners on standard, as results the reference data sets and the proposed set of ensemble gave better accuracy.

Given the presented scenario, we propose the application of combined regression models in the context of EDM to identify the factors related to dropout. For this, we used two databases of Brazilian Higher Education: the Census and Flow Indicators of the classroom courses of the public and private universities of the year 2013. [1].

III. Theoretical Background

In this section, we present the concepts related to combined models, which are the prediction model applied in this study.

A. Combined Models

In ML, set learning consists of methods and aware that integrates several basic models to generate the final output. It gained great popularity due to its excellent generalization performance. The combined models generally result in better accuracy than the individuals composing them [14].

A combined model is a technique resulting from the integration of two or more similar or different type algorithms. It allows us to create a more robust system that incorporates the result of all the [15] techniques. The idea is that this will make the result more robust, accurate and less prone to bias.

This approach has been used in the literature for both regression and classification problems. The combined models are capable of increasing the generalization capacity and, so, overall system performance.

B. Construction of a Combined Model

The construction of a combined model aims at a set of models F_0, where $F_0 = f_i, i = 1, ..., M$ where M is the total number of models generated. If the models are made using the same algorithm, the set is called homogeneous and is the most widespread in the literature [16]. Different sets are obtained when more than one learning algorithm is used. It's expected that these approaches constitute models with diversity and thus the process of modeling of sets can be manipulated through training data, techniques or set of parameters.

Data diversity allows the generation of many data sets from the original data set to train different predictors. The data sets must be different from one another so that there are several decisions from the results of the trained predictors.

C. Bootstrap Aggregation

Breiman developed the Bootstrap Aggregation method, also called Bagging [17]. Bagging produces several different training sets with data replacement. Then builds a model for each of the sets using the same machine learning algorithm. The model predictions are combined through their mean for regression problems or through voting for classification problems.

In the case of the regression equation, assume a set of training data $D_train = (x_1; y_1), ..., (x_n; y_n)$. The instances are extracted from a probability distribution $P(x, y)$. Bagging works by combining the prediction of a collection of regressors, which each of these regressors is constructed by applying a fixed learning algorithm in a Bootstrap sample different from the original D_train training data. The forecast for the set is the average of the individual forecasts of the M regressors generated. The representation of Bagging is described in Equation 1.

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

where $f_bagging(x)$ is the prediction of the combined model for the instant x; M is the number of regressors of the model; $f_i(x)$ is the prediction given by the $i-th$ regressor constructed under the $i-th$ bootstrap sample of training data.

One of the factors that motivate the use of this method is the simplicity of implementation. It has been proven to improve the predictive capacity for regression or classification algorithms [18].

IV. Proposed Model applying education data

We present in this section the proposition of three models based on an ensemble of regressors applied to a set of educational data. For this, we used the CRISP-DM methodology to proposition of these models, according to the following phases.

A. Business Understanding

We made a literature review, verifying in the literature all the material already elaborated on dropout as well as its causes and factors related to educational scenarios, data mining and regression models. The main factors associated with circumvention were listed below. These factors are showed in the works [19], [20] and [21]. Among the several factors found in these studies are the lack of motivation, personal and socioeconomic problems, dissatisfaction with the course/institution, learning problems associated with teaching methodologies and evaluation processes, restrictions on the labor market, lack of knowledge of the course and level of the previous study. Besides to the methods for measuring the works [22] and [23]. We applied the proposed models in data from the Higher Education Census along with data from the Higher Education Flow Indicators. The target was the face-to-face courses of Brazilian public and private universities. We aim to identify and predict their occurrence from the factors associated with evasion in these institutions.

B. Data Understanding

The educational indicators used by the study come from the Higher Education Census Database and the Higher Education Flow Indicators. Both are available by the National Institute of Educational Studies and Research (INEP) in the year 2013. To extract the data from these databases, the factors proposed in the literature were considered as factors associated with dropout. For example, demographic, academic, and socioeconomic attributes.

One limitation is related to the data provided by the Census. These deal with the absence of other information, which could further complement the extracted attributes. For example aspects related to the student, infrastructure of the institutions, and teaching, among others.

C. Data Preparation

We made a combination of the two databases (Census and Flow Indicators). We checked for missing data in the school dropout variable. We excluded the instances for the blank values for these variables. We filled by median values the missing data of the other variables for full data set. We also used the Stepwise method to select variables. Thus, Table I shows the relation of the number of instances before and after the pre-processing performed.

TABLE I: Dimensions of Dataset

Before pre-processing		**After pre-processing**	
Nº Variables	Nº Instances	Nº Variables	Nº Instances
50	133528	14	22972

We present in Table II the description of the 14 variables selected by the Stepwise method. In these variables we use correlation to get the highest correlations for the construction of another scenario. These variables indicate the factors related to demographic, academic and socioeconomic information. When combined this information will be based on the proposition of the predictive models efficient. Thus, we elaborated two scenarios for the study. In the first scenario, we used the variables selected by Stepwise. In the second, we used the four variables with the highest correlation with the school dropout rate in higher education institutions in Brazil. The variables of higher correlation are: TAP (-0.6254), TCA (-0.2522), INC(0.1566) and QP (-0.1308).

TABLE II: Description of variables.

Variables	**Description**
CAS	The situation of the student in the course (active, locked, Unlinked from the course, Transferred to another course of the same IES, Formed, deceased)
IABP	Informs if the student receives some remuneration for the stay in the institution of higher education
QI	Number of new students
QP	Number of students remaining in the course
TAP	indicator of the student's stay
TCA	completion indicator of the student's
INC	Studies in night time
CCRA	Color / Race of the student (white, black, brown, yellow, indigenous, undeclared, not informed)
IABT	Informs if the student receives remuneration for activity developed within the institution of higher education
ICE	Informs if the student does non-compulsory extracurricular activity
ICEX	Informs if the student participates in extracurricular extension
QCC	Number of graduates in the course
IAE	Informs if the student participates in some extracurricular activity (internship, extension, monitoring and research)
QC	Number of graduates
TDA (Y)	Dropout rate

D. Modeling

We used the bagging [17] method in the modeling phase to generate a bootstrap sample set of the original data. This dataset will generate a set of models using a simple learning algorithm by combining their means. It's according to Equation 1 described in section 3.

$fbagging(x)$ is the prediction of the ensemble for the instant x, M is amount of model regressors and, $\hat{f}(x)$ is the prediction given by the ith regressor the sample. Hence, based on function $\hat{f}(x)$ proposed tree models, as follow:

- *ProposedModel1:* ensemble Bagging with linear regression (here called $PM1$). The linear regression is defined as, $y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i = x_i' \beta + \epsilon_i$
  With the ith of the n observations. Giving as estimator b to β the fit model is:
  $$\hat{f}(x) = \hat{y}_i = \alpha + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik} + \epsilon_i = x_i' b \quad (2)$$
  And the residual given by:
  $$e_i = y_i - \hat{y}_i \quad (3)$$
  Where y_i is the real value and $\hat{y}_i$ is the estimated value. As estimates b are determined by minimizing an objective function for all b.
- *ProposedModel2:* ensemble Bagging with robust regression (here called $PM2$). The robust regression is defined based on the Equation 2 and 3 as,
  $$\hat{f}(x) = \sum_{i=1}^{n} \rho(e_i) = \sum_{i=1}^{n} \rho(y_i - x_i' b) \quad (4)$$
  where the funtion ρ provides the contribution of each residue to the objective function.
- *ProposedModel3:* o ensemble Bagging with ridge regression (here called $PM3$). the ridge regression is defined as,
  $$\hat{f}(x) = (Y - X\beta)^T (Y - X\beta) + \lambda \beta^T \beta \quad (5)$$
  where, $\lambda \geq 0$ is a tuning parameter for the penalty, which is determined separately.
- *LiteratureModel4:* the techniques used in the work of [24] for the construction of stacking had as a meta-predictor the Ridge regression. The regressions for composing the ensemble are linear regression, lasso regression, bagging (using decision tree), boosting, random forest, vector regression support and k-nearest neighbors (here called $LM4$).

E. Evaluation

At this stage, we evaluate the developed models, to meet the defined objectives. Thus, we chose two errors for this evaluation: absolute mean error (MAE) and mean square error (MSE). We show that in equations 6 and 7.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \hat{y}_i \right| \quad (6)$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \quad (7)$$

Where n is the dataset size, y_j is a real variable value, and y_i is the estimated variable by the model.

With the sample of 500 iterations, we calculate the standard deviation (SD) of the error. Besides, we performed statistical tests, such as the Kolmogorov-Smirnov and Wilcoxon tests. We also use boxplots and relative gain (RG) charts also to evaluate the performance of the models.

F. Application

The last step of the CRISP-DM methodology implies in the diffusion and application of the model constructed in the real data. After the execution of all the steps have been subsidies for the resolution of the proposed problem. Thus, the strategies used in this paper will be described and analyzed in the following results section.

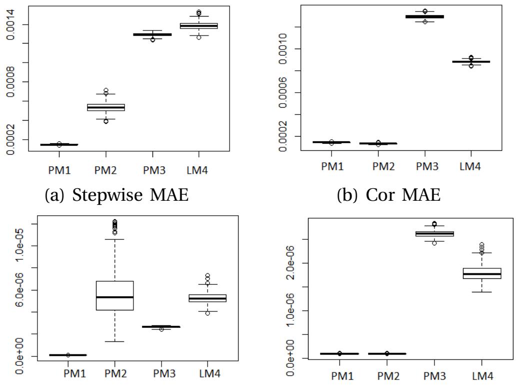
V. Results

In this section, we present the results obtained with the experiments in Table III, Figure 1 and TableIV. We proposed three models and scenarios of the selected variables with the Stepwise Method and the Pearson Correlation based on the metrics MAE and MSE. Table III displays the mean values of the MAE and MSE of the standard deviation (SD) for the proposed models after the 500 iterations. It's worth noting that $PM1$ obtained a lower average value for the two metrics in the scenario using the Stepwise Method. While in the Pearson Correlation scenario $PM1$ and $PM2$ presented similar results.

TABLE III: Results for Mean Error (SD).

Tech - nique	**Metrics**	$PM1$	$PM2$	$PM3$	$LM4$
Step-wise	MSE	$3.14x10^{-3}$ $(2.75x10^{-4})$	$1.39x10^{-2}$ $(5.31x10^{-3})$	$6.54x10^{-3}$ $(2.33x10^{-3})$	$1.31x10^{-2}$ $(4.57x10^{-3})$
	MAE	$1.47x10^{-4}$ $(7.89x10^{-3})$	$5.3x10^{-4}$ $(3.46x10^{-2})$	$1.29x10^{-3}$ $(1.20x10^{-2})$	$1.4x10^{-3}$ $(2.69x10^{2})$
Cor	MSE	$2.266x10^{-2}$ $(3.3x10-4)$	$9.29x10^{-3}$ $(2.6x10^{-4})$	$8.054x10^{-3}$ $(1.04x10^{-4})$	$4.43x10^{-3}$ $(1.13x10^{-3})$
	MAE	$1.21x10^{-1}$ $(1x10^{-3})$	$7.38x10^{-2}$ $(1.15x10^{-3})$	$7.25x10^{-2}$ $(5.6x10^{-4})$	$9x10^{-4}$ $(9.33x10^{-3})$

For the graphical representation and analysis of the two scenarios, Figure 1 shows the boxplots generated by the 500 iterations. We can see that the graphs in Figure 1 show that there was no significant difference in the median of errors between $PM1$ and $PM2$ for the selected variables with correlation. It also identifies the presence of outliers in $PM2$ for the Stepwise scenario. Besides, $PM2$ presents greater variability than the other models for this scenario.



(a) Stepwise MAE

(b) Cor MAE

(c) Stepwise MSE

(d) Cor MSE

Fig. 1: Boxplot about ensemble models.

We performed the Kolmogorov-Smirnov test to verify if the error vector of the 500 iterations follows a normal distribution. The result shows that the data does not follow this type of distribution. Thus, we used the Wilcoxon test to perform the hypothesis test with 5% significance. The elaborate alternative hypothesis is that $PM1$ performs better than $PM2$, $PM3$, and $LM4$.

We can prove statistically with a confidence level of 95% that $PM1$ presents smaller prediction errors about all the models used for the scenario of the variables selected with Stepwise. The $p-value$ obtained was $2.47x10^{-7}$. Using the Correlation method, $PM1$ got smaller errors than $PM3$ and $LM4$ since the value of p-value was $2.47x10^{-7}$. While for $PM1$ ratio it was lower than $PM2$, analyzing the MSE and MAE metrics the p-value values were $7.916x10^{-2}$ and $7.916x10^{-2}$, respectively.Thus, in relation model $PM1$ with $PM2$ there is no statistical evidence to accept the null hypothesis.

Table IV presents the RG (in module) of $PM1$ about the other models used in this work. We can verify that the gain obtained was very significant. We demonstrate that $PM1$ is more efficient than other models. We demonstrated that performance prediction proposed models improvement can be achieved using bagging ensemble with the resultant effect of increase in accuracy, reduced error rate as well as increase in predictive efficiency. It can ratify the mean values obtained in Table III.

TABLE IV: Result of RG

Tech - nique	**Metrics**	$PM1$ **X** $LM4$	$PM2$ **X** $LM4$	$PM3$ **X** $LM4$
Step-wise	MSE	96.441%	6.565885%	49.96957%
	MAE	89.325%	61.46408%	6.516776%
Cor	MSE	94.741%	94.73033%	46.8963%
	MAE	83.431%	84.73009%	46.59059%

Thus, we can see that $PM1$ performed better in school dropout prediction than $LM4$ using both tested scenarios. Besides, we propose the use of other ensemble regression models using Bagging method that also obtained significant results for the problem of evasion. However, we proved that the proposed models based on ensemble regression based on bagging method got results superior to the literature model (Ensemble Stacking $LM4$).

VI. Conclusion

In this paper, we proposed three ensemble regression models based on bagging methods. The aim is for the prediction of school dropout in Higher Education Institutions in Brazil. We used the factors present in the Census and Flow Indicators of Higher Education. This information was demographic, academic, and socioeconomic. We used the Stepwise method and Pearson correlation to select the variables. The main contributions of this paper are:

1) We showed that there is a correlation between the identified variables and that performance prediction

model improvements can be achieved using ensemble regression model based on bagging. This could increase in accuracy, reduced error rate, and increase in predictive efficiency.

2) We showed through the experiments that $PM1$ got better results compared to the literature model ($LM4$). It's an ensemble linear regression model based on the bagging method. As well, about ensemble robust regression model based on bagging method ($PM2$) and ensemble ridge regression model based on bagging method ($PM3$). Ensemble regression models based on bagging method proposed are to fit several independent models and average their predictions to get a model with a lower variance.
3) The proposed ensemble regression models ($PM1$, $PM2$, and $PM3$) allow identifying the main factors associated with the dropout process. The factors identified were: a) accumulated permanence indicator (TAP); b) cumulative completion indicator (ACT); c) Study in the Evening Period (INC); d) Number of students who remain in the course (QP), who describe some factors that explain the movement of students in Higher Education Institutions from their entry to the conclusion of the course. According to [20], the dropout does not occur in the last years, but at the beginning of higher education. To help educational managers in the direction of decisions that lead to the adoption of public policies and educational and organizational strategies that allow better conditions for the maintenance of students in their institutions until the conclusion of higher education. As well, contribute to the proposal of actions to measure and mitigate the occurrence of evasion in the Brazilian Higher Education Institutions.

The proposed ensemble regression models could be used as a tool for accurate prediction of student performance. They reduced errors as well as early identification of student dropout. But, the findings in the study emphasize the need for more detail and generalized research in this area. This should include the inclusion of more variables and combinations of variables from other sources before analyzing using ensemble techniques.

As future work, use other ensemble models, as: *Stacking* e o *Boosting*. Besides, applying these ensemble models in higher education institutions that have distance education courses (EAD). It's to investigate factors other than those present in face-to-face teaching that influence the evasion process. We can do it by collecting new attributes. These are only to this modality of teaching such as student/course interaction, student/tutor, student/discipline. Besides the integration of the same to this teaching modality. The inclusion of different factors and/or variables related to aspects of learning, social, and infrastructure, to identify their relationship with the dropout process.

References

[1] "Inep," http://download.inep.gov.br/, accessed Feb 3, 2018.

[2] C. ROMERO, C.; VENTURA, "Educational data mining: a review of the state of the art," 2010, pp. 601–618.

[3] A. D. Himanshu Singh and V. Pudi, "Pager: parameterless,accurate, generic, eficient knn-based regression," 2010, pp. 168–176.

[4] E. A. P. Douglas C Montgomery and G. G. Vining., "Introduction to linear regression analysis," in *2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI)*.

[5] A. Ozcift, "Forward stage-wise ensemble regression algorithm to improve base regressors prediction ability: an empirical study," *Expert Systems*, pp. 1–8, 2014.

[6] S. B. Kotsiantis, "Combining bagging and additive regression," *Int J ComputMath Sci*, pp. 61–67, 2007.

[7] L. I. Kuncheva, "ombining pattern classfiers: methods and algorithms," *Int J ComputMath Sci*, 2004.

[8] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "Crisp-dm 1.0 step-by-step data mining guide," 2000.

[9] G. Kantorski, E. G. Flores, J. Schmitt, I. Hoffmann, and F. Barbosa, "Prediçao da evasao em cursos de graduaçao em instituiçoes públicas," in *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, vol. 27, no. 1, 2016, p. 906.

[10] R. Fritsch, C. S. da Rocha, and R. F. Vitelli, "A evasão nos cursos de graduação em uma instituição de ensino superior privada," *Revista Educação em Questão*, vol. 52, no. 38, pp. 81–108, 2015.

[11] B. Perez, C. Castellanos, and D. Correal, "Applying data mining techniques to predict student dropout: A case study," in *2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI)*. IEEE, 2018, pp. 1–6.

[12] R. T. Pereira and J. C. Zambrano, "Application of decision trees for detection of student dropout profiles," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2017, pp. 528–531.

[13] D. M. De Brito, I. A. de Almeida Júnior, E. V. Queiroga, and T. G. do Rêgo, "Predição de desempenho de alunos do primeiro período baseado nas notas de ingresso utilizando métodos de aprendizagem de máquina," in *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, vol. 25, no. 1, 2014, p. 882.

[14] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, 2000, pp. 1–15.

[15] R. Polikar, "Ensemble methods in machine learning," vol. 6, no. 3, 2006, pp. 21–45.

[16] ——, "Ensemble approaches for regression: A survey," *ACM Computing Surveys (CSUR)*, vol. 45, pp. 1–10, 2012.

[17] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[18] P. Bühlmann and B. Yu, "Analyzing bagging," *The Annals of Statistics*, vol. 30.

[19] W. G. Spady, "Dropouts from higher education: An interdisciplinary review and synthesis," *Interchange*, vol. 1, no. 1, pp. 64–85, 1970.

[20] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," *Review of educational research*, vol. 45, no. 1, pp. 89–125, 1975.

[21] J. P. Bean, "Dropouts and turnover: The synthesis and test of a causal model of student attrition," *Research in higher education*, vol. 12, no. 2, pp. 155–187, 1980.

[22] R. L. L. Silva Filho, P. R. Motejunas, O. Hipólito, and M. B. C. M. Lobo, "A evasão no ensino superior brasileiro." *Cadernos de pesquisa*, vol. 37, no. 132, pp. 641–659, 2007.

[23] M. Lobo, "Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções," *Associação Brasileira de Mantenedoras de Ensino Superior. Cadernos*, vol. 25, 2012.

[24] J. Beemer, K. Spoon, L. He, J. Fan, and R. A. Levine, "Ensemble learning for estimating individualized treatment effects in student success studies," *International Journal of Artificial Intelligence in Education*, vol. 28, no. 3, pp. 315–335, 2018.