International Journal of Computer Applications (0975 - 8887)
Volume 174 - No.22, February 2021

Extreme Learning Machine Models for Predicting Student Performance

Wedson L. Soares
Department of Computer Engineering
University of Pernambuco
Pernambuco, Brazil

Roberta A. de A. Fagundes
Department of Computer Engineering
University of Pernambuco
Pernambuco, Brazil

ABSTRACT

Predicting the individual performance of each student can provide valuable information as to which students are at greatest risk of failure or dropout, and consequently highlight which characteristics negatively influence the student's academic life. Data mining provides the tools necessary to address this educational data in the search for knowledge and patterns that can be obtained. Therefore, this work uses an educational database obtained at the UCI machine learning repository related to students grades in Portuguese and proposes models using extreme learning machine networks, ensemble learning and optimization by particle swarm in order to predict students' grades. In addition two simulated data sets were also used to verify the consistency of the results obtained through the proposed regression models. After obtaining the error value for each proposed model, hypothesis tests were performed to ascertain the veracity of the results. The results indicate a better performance of the model that combines the ensemble learning, particle swarm optimization and extreme learning machine networks.

General Terms

Data Mining, Regression, Education, Student Performance

Keywords

Educational Data Mining, Extreme Learning Machine, Ensemble Learning, Regression, Particle Swarm Optimization

1. INTRODUCTION

Several factors can negatively influence a student's academic life and consequently his performance. The ability to individually predict the performance of each student can provide useful information about which individuals are at risk of failure or dropout [20]. In this way, predicting the performance of students at an institution can serve as a guide, allowing a visualization of factors related to the teaching and learning process and how they influence student performance. Therefore, it is possible to efficiently identify and analyze which factors need to be monitored in order for the performance rate to approach what is desired. It is estimated that the amount of information in the world doubles every 20 months [9], this accumulation of data at a fast pace makes it increasingly necessary the emergence of new theories and tools capable of assisting in the task of extracting useful information from these growing volumes of digital data [8]. The science of extracting knowledge from large databases is known as data mining (MD), and for that, knowledge and techniques from different areas are used, such as statistics, machine learning, pattern recognition and artificial intelligence [11].

MD has been applied in a large number of fields, including computer science, engineering, mathematics, physics, neuroscience and cognitive science [11]. In recent years, there has been a growing interest in using data mining to investigate scientific issues within educational research, educational data mining. Educational data mining (EDM) is defined as the scientific area focused on the development and application of MD methods to the specific type of data that come from educational environments to address important educational issues, and using these methods to understand the factors that influence the teaching process student learning [2]. Among some of the techniques used in the MD stage, neural networks (NN) stand out for their ability to approximate complex functions and their ability to provide models for a wide range of difficult to handle natural and artificial phenomena [14].

A negative point of NN's is the lack of faster training algorithms. Traditional learning methods are usually much slower than necessary, and seek to adjust the parameters of the network to achieve the modeling of the curve that best defines the problem in question, as an example it is possible to mention the algorithms based on gradient descent that are quite widespread for training neural networks, however they tend to be very slow and easily converge to local minimums, [14]. On the other hand, [13] demonstrates that a neural network with only one hidden layer can be able to learn N distinct observations, working with almost any nonlinear activation function, depending only on a number of N neurons in the hidden layer . To train this type of network, it was developed by [14] a much faster training algorithm, called the extreme learning machine (ELM), which operates by inverting output matrices from the hidden layer of the neural network and performs training thousands of times faster than traditional methods and obtaining a better generalization.

Some problems accompany learning algorithms that have only one hypothesis as their output, such as the chance of being trapped in local minimums, or the need to better map the search space to obtain the best solution based only on a small set training. In this way, there may be several solutions that satisfy the stipulated conditions, but only one will serve as a return for the learning algorithm [7]. Ensemble learning operates through the training of several instances of the same basic learning algorithm, in order to solve

stability problems and improve the final result, such as boostrap aggregation, which consists of training several instances of the same learning algorithm with different samples from the training set [3]. Additionally, when working with ELM Networks, it is necessary to keep in mind the random factor induced by the generation of the weights and thresholds of the hidden layer, a way to get around this negative point, is by adjusting the network parameters so that it is possible to define these settings in an optimized way, for example, the number of neurons in the hidden layer, thus leaving the random factor present in the ELM network less impacting on the results. Therefore, it is possible to model an optimization problem that aims to adjust the parameters of the ELM network according to some metric to be used as a cost, for example the absolute mean error (MAE). When working with optimization problems, the Particle Swarm Optimization (PSO) [15] is noteworthy for its ability to be applied and to be able to solve most problems of this type, the algorithm operates by initializing a certain number of particles in a search space with defined limits so that the particles will move in function the one that has the set that best optimizes the chosen function.

1.1 Motivation

The search for ways to use DM applied to educational environments in order to identify and predict factors that influence students' teaching and learning processes has been increasing. Educational data mining (MDE) focuses on developing methods to explore data from educational contexts [17] and thus address important issues, such as repetition, failure, school dropout and student performance analysis. Based on this, predicting the performance of these students allows better targeting of educational resources, assessing how they are applied. Allowing it to be possible to identify and analyze which factors influence their teaching and learning process, thereby ensuring an improvement in the distribution of resources and student performance.
The present work seeks to combine the ELM neural networks with the ensemble learning method, to build regression models that benefit from the positive characteristics of both techniques, as well as using the PSO to reduce the impact of the ELM algorithm randomness. , optimally defining the number of estimators used in the ensemble and adjusting the best parameters of the ELM network to obtain better predictions at the end. This models models were applied to simulated databases in order to verify the consistency of the results, after this these model (s) were applied to the context of educational data from high school, specifically from the Portuguese subject, to predict student performance through the final grade, assisting in decision making and the development of actions by part of the coordination of the educational environment. The results demonstrate an efficiency in the prediction, as well as a superiority on the part of the models based on ensemble used in conjunction with the PSO.

2. RELATED WORKS

In this section, works related to the use of ELM, ensemble learning and PSO applied to educational problems, such as student performance, school dropout, failure rate and approval, are presented.
[16] It seeks to examine the accuracy of ensemble methods to predict the performance of students in an undergraduate engineering course that lasts 4 years. Several ensemble algorithms were used, such as Boosting, Bagging and Random Forest and these were compared with base algorithms, without using ensemble methods. At the end, the authors observed that the ensemble methods obtained better final results, when compared to the base algorithms.
[6] Develops and applies regression methods through ensemble to data from educational systems to try to predict the dropout rate in Brazilian universities. The results demonstrate an improvement when using the ensemble methods, which can help the educational system administrators in decision making
[5] Seeks to predict dropout rates for online courses using a combination of ELM Networks and decision trees, due to ELM's fast training. In the end it was possible to observe improvements in the final results of the proposed model when compared to traditional machine learning methods.
[12] Uses classification by swarming of particles in the field of EDM to classify questions on cognitive levels. When compared with seven other machine learning algorithms, it is possible to observe gains in the results from the use of the particle cluster classification algorithm, reinforcing the improvement that can be obtained through this type of algorithm.
[1] Uses the particle swarm optimization algorithm based on discrete spaces to propose a classification method that can be used to predict student final results. When compared with other classification algorithms, there was a considerable gain in the accuracy of the model.
[21] It seeks to predict the performance of higher education students by predicting their grades, for this the author uses PSO to reduce the dimensionality of the data set before applying classification methods. After carrying out the experiments, it was observed that the proposed model can help improve the quality of education and decision-making in the educational system.
The differential of the work developed here in relation to those mentioned in this section, is in the unified use of ELM Networks, ensemble and PSO methods. Since the negative points of ELM as random weight generation can be complemented through the use of ensemble learning that operates through the training of several instances of the same algorithm in order to reduce instability, at the same time the high training speed of ELM allows multiple training through the ensemble to be done more efficiently. Additionally an adjustment of the parameters through the PSO to define the number of units of the hidden layer of the ELM as well as the number of estimators in the ensemble seeks to obtain an optimized model and consequently an improvement in the final result.

3. THEORICAL BACKGROUND

In this section, the main concepts and methods used in the present work. Thus, will be presented the concepts of extreme learning machine, ensemble learning and particle swarm optimization.

3.1 Extreme Learning Machine

Extreme learning machine (ELM) is a machine learning algorithm developed by [14] to carry out the training of hidden layer feedforward neural networks (SLFN) that can be thousands of times faster than traditional literature methods such as the backpropagation algorithm while obtaining better generalization capabilities. The algorithm is based on the fact that an SLFN is able to approximate any function using random values both for the weights of the input layer and for each activation threshold of the hidden layer of the network.
The essence of ELM is based on the fact that the hidden neuron layer does not need to be adjusted iteratively and in addition, the training error $||\mathbf{H}\beta - \mathbf{y}||$ and the weight standard $||\beta||$ is minimized. given a set of N observations, $(x_i, y_i), i \leq N$. with $x_i \in \mathbf{R}^p$ and

$\mathbf{y} \in \mathbf{R}$. An SLFN with a m number of neurons in the hidden layer can be expressed by the following sum:

$$\sum_{i=1}^{m} \beta_i f(w_i x_j + b_i), 1 \leq j \leq N, \quad (1)$$

Where β_i is the output weights, f is an activation function, w_i is the input weights and b_i is the activation threshold. Assuming that the model describes the data perfectly, the relationship can be written in matrix form as $\mathbf{H}\beta = \mathbf{y}$, with

$$\mathbf{H} = \begin{bmatrix} f(\mathbf{w}_1 \cdot \mathbf{x}_1 + \mathbf{b}_1) & \cdots & f(\mathbf{w}_m \cdot \mathbf{x}_1 + \mathbf{b}_m) \\ \vdots & \ddots & \vdots \\ f(\mathbf{w}_1 \cdot \mathbf{x}_n + \mathbf{b}_1) & \cdots & f(\mathbf{w}_m \cdot \mathbf{x}_n + \mathbf{b}_m) \end{bmatrix}, \quad (2)$$

$\beta = (\beta_1, ..., \beta_m)^T$ and $\mathbf{y} = (y_1, ..., y_n)^T$. ELM's approach is based on randomly initializing w_i and b_i, and compute the output weights $\beta = \mathbf{H}^{\dagger}\mathbf{y}$ using a Moore-Penrose pseudo-inverse [21].

3.2 Ensemble Learning

This collection of methods combines the output of predefined machine learning techniques from a given group of techniques to obtain lower prediction errors (for regression tasks) or minimize error rates (for classification tasks). In this context, these machine learning algorithms used in groups are called weak learners because the ensemble methodology serves precisely to improve the predictions of these algorithms [10]. Among the methods of ensemble, it is possible to mention boostrap aggregation (bagging), boosting and stacking [10]. These ensemble methods operate by resampling the training set that is passed on to the machine learning algorithms that are being used as weak learners in order to improve the final result.

3.2.1 Bagging. In the algorithm called boostrap aggregation (bagging) each of the machine learning predictors used as weak learners are trained independently in resampled training sets, which are selected at random from the original training set. Therefore, bagging is dedicated to algorithms that suffer from instability problems, where any small change in the training set can result in changes in the output of the algorithm in question. The training of each predictor can occur in parallel, since each of these predictors is trained independently [10].
To address a regression problem, assume a set of training data $T_{train} = (x_1; y_1), ..., (x_n, y_n)$, whose instances are extracted through a probability distribution $P(x, y)$. Bagging operates by combining a certain number of regressors, each of these regressors is constructed using a fixed learning algorithm that is applied to samples other than the original T_{train}. The final forecast is obtained by means of the individual average of the m regressors used in the process. The representation of Bagging is described in Equation 3.

$$f_{bagging}(x) = \frac{1}{M} \sum_{M}^{i=1} \hat{f}(x) \quad (3)$$

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 [6].

3.3 Particle Swarm Optimization

The particle swarm optimization (PSO) algorithm was first devised by Eberhart and Kennedy [15] consists of a population of particles that are initialized in a given search space and have their speed (acceleration) information changed at each instant of time towards their best values of pbest (best value for the function particle cost) and gbest (best value for the cost function in the population). The acceleration is weighted by a random term, with separate random numbers being generated for acceleration towards the locations of pbest and gbest [19].
the speed and position of the particle are updated using respectively the Equations (4) and (5)

$$v_{id} = w * v_{id} + c_1 * rand * (p_{id} - x_{id}) + c_2 * Rand * (p_{gd} - x_{id}) \quad (4)$$

$$x_{id} = x_{id} + v_{id} \quad (5)$$

Where w is the inertia value [18] $c1$ and $c2$ are two positive constants and $rand()$ and $Rand()$ are two random functions in the range of [0,1].
The i-th particle is represented as $x_i = (x_{i1}, x_{i2}...x_{iD})$. The best position ever found by the i-th particle (position that best optimizes the cost function in question) is represented by $p_i = (p_{i1}, p_{i2}...p_{iD})$. The index that represents the particle with the best value for the cost function is represented by the symbol g and the velocity rate for the particle i is represented by the term $v_i = (v_{i1}, v_{i2}...v_{iD})$.

4. METHODOLOGICAL FLOW

The methodology used in this work was based on [4] and will be presented in details this section. The methodology consists of four phases: description of the dataset, modeling, calculate evaluation and analysis of the results, presented in Figure 1.

(5) (1).png

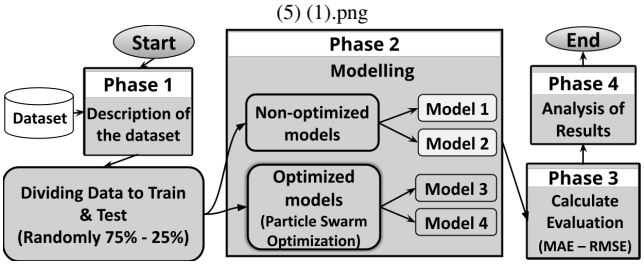

Fig. 1: Methodology: Flow of developed tasks

The activities developed in each of these stages are described in detail below. The phase 01 describes the educational data sets (subsection 4.1), phase 02 presents the four models applied for performance student's (subsection 4.2), phase 03 show evaluation criterion of the four models (section 5) and phase 04 show the analysis and discussion of results in this work (section 6).

4.1 Description of the data sets

In this work the models were first applied to two simulated data sets to verify the consistency of the results obtained trough each model. These data sets were created using the following linear regression function

$$y = \beta_0 + \beta_1 x + e \quad (6)$$

β_0 and β_1 were generated using a random uniform distribution between the range [0,1], x was generated using a random uniform distribution between the range [-1,1] and the e was generated using a normal distribution with mean = 0 and standard deviation = 1. The y is used as the variable to be predicted, and x is used as input to the regression models, the β_0 and β_1 are the values to be estimated.
One of the data sets were populated with using the rule defined by the Equation 6 this data set is refered in this work as data set without noise. In the other data set, noise was also added to the y value in order to increase the complexity of prediction, this noise was defined using the following equation

$$\hat{n} = std * 5 \tag{7}$$

where std represents the standard deviation obtained from the data set without noise y values. For the new data set, $\hat{n}$ was added to 5% of the y values, in this way generating distortion in the y values and increasing the complexity of the prediction. In this work this data set is refered as data set with noise
The educational data set used in this work was obtained from the UCI machine learning and has as a problem the prediction of the Portuguese grades of high school students. The data includes student grades, as well as demographic and school-related social characteristics. And it was collected using school reports and questionnaires, in this work the database referring to the performance of students in the matter of Portuguese was used. The base initially had 32 columns, after the treatment to remove the binary, categorical and textual variables, the base was left with 14 columns. The columns chosen for the final database, as well as their meaning, are described in Table 1.

Table 1. : Data description

Column name	Description
Age	student's age
Medu	mother's education
Fedu	Father's education
traveltime	home to school travel time
studytime	weekly study time
failures	number of past class failures
famrel	quality of family relationships
freetime	free time after school
goout	going out with friends
Dalc	workday alcohol consumption
Walc	weekend alcohol consumption
health	current health status
absences	number of school absences
G3	final grade

In this way, the variable to be predicted is column G3 that houses the information of the students' final grade, and all others are used as input for the model. First, the database is divided into training and testing, from which it will be used as input for 4 different learning models, these models are described in detail in the following section.

4.2 Modeling

The models used here can be divided into two groups, the first group, to which Models 1 and 2 belong, is the group of models not optimized through the use of PSO. Group 2 is made up of Models 3 and 4, and deals with models optimized through the use of PSO. More details on the construction of each model are described below:

4.2.1 Non optimized models. This subsection describes the non optimized models used in this work, the Table 2 shows the parameters used in the ELM for both models.

—Model 1 : Consisting only of the basic ELM Network with manually adjusted parameters, the database is used as input and at the end errors related to the forecast of the final grade (G3) are obtained as output.

—Model 2 : In this model an ensemble of the Bagging type is used, which has the basic ELM as the base algorithm, from then on the model will operate by training an N number of these algorithms and at the end the prediction errors will be obtained. In this model, the parameters of the ELM network used are the same as in Table 2, and in addition the number of estimators of the ensemble is defined, in this case 50 were used.

Table 2. : Model 1 and Model 2 parameters

Parameter	Value used
Alpha	0.2
Number of hidden neurons	20
Activation function	Sigmoid

4.2.2 Optimized models. This subsection describes the optimized models used in this work, the Table 3 shows the parameters used for Model 3, and Table 4 shows the parameters used for model 4.

—Model 3 : In this model, an ELM network is used in conjunction with the PSO, optimizing the parameter values in order to obtain improvements in the final results. In this model, each PSO particle coordinates an ELM network, which in turn is used to calculate the absolute mean error to be minimized, which defines the direction and acceleration of the particle population. In Table 3 it is possible to observe which parameters were chosen to be adjusted in this model, as well as a description of each one:

Table 3. : Model 3 parameters

Parameter	Description
Alpha	Mixture coefficient for distances and scalar product input activations.
Number of hidden neurons	Number of units to generate the hidden layer.

—Model 4 : This is the proposed model to unify the ensemble with the PSO, it uses a Bagging ensemble that uses the ELM network as a base estimator. Each PSO particle is responsible for obtaining the absolute mean bagging error at each iteration of the algorithm, seeking to minimize this value, obtained at the end of each PSO iteration. Table 4 describes the parameters adjusted in this model, and a description of each is also made:

Table 4. : Parameters to optimize

Parameter	Description
Alpha	Mixture coefficient for distances and scalar product input activations.
Number of hidden neurons	Number of units to generate the hidden layer.
Number of estimators	Number of base estimators used to build the ensemble

4.2.3 Implementation of the PSO. Therefore, first the PSO is initialized with 30 particles having a position array composed of the values that denote the values of the parameters to be adjusted. Table 5 explains the intervals defined for random initialization of each of the parameter values for models 3 and 4.

Table 5. : Initialization Intervals of PSO

Parameter	Initialization interval
Alpha	0.01 to 1
Number of hidden neurons	10 to 500
Number of estimators	25 to 300

In both model 3 and model 4, the cost function to be optimized was the absolute mean error. The pseudocode that defines the PSO its specified as follows.

Algorithm 1 Particle Swarm Optimization

```
Initialize population P in the defined limits
for i = 0 to MaxIter do
    for each particle p in P do
        cost = cost(p)
        if cost ¡ bestPersonalCost then
            bestPersonalCost = cost
        end if
    end for
    bestGlobalcost = bestPersonalCost in P
    for each particle p in P do
        Update particle speed and position
    end for
end for=0
```

Each particle has information related to its position that has a vector composed of a number N of values that depends on the number of parameters to be adjusted. In this case, model 3 has a vector composed of 2 positions, while model 4 has a vector composed of 3 positions. This vector is then iteratively adjusted for each particle using equation (4), 30 particles and 15 PSO iterations were used. After that, these values are used to train an ELM Network (model 3) and a bagging type assembly that uses ELM networks as a base algorithm (model 4). Then the value of the absolute mean error is obtained, which in turn is used as a cost function, and provides the value of the best overall cost to guide the adjustment of the speed and position of the particles. The Process that starts in the division of the database in training and testing, and ends with obtaining the predictions of each one, is then executed 300 times to obtain the average of the final errors of each cycle, so that it is possible to guarantee robustness and confidence in the final results.

5. EVALUATION

In this section, the metrics used in the work will be described, and the results obtained in each of the proposed models will also be displayed, as well as the results of the hypothesis tests to confirm and confirm the results obtained. The metrics used are described below, as follows: Mean absolute error and the root of the mean square error.

5.1 Mean Absolute Error (MAE)

it is a measure that defines the distance between the predicted values and the real values. The mean absolute error is defined by the equation

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

5.2 Root of mean squared error (RMSE)

It is a measure that defines the by obtaining the square root of the mean square difference between the estimated values and the predicted value, this metric can be defined through the equation of the mean square error, and can be defined using the equation

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \tag{9}$$

6. RESULTS AND DISCUSSION

The Tables 6 and 7 describes respectively the prediction errors obtained for each of the established regression models, for the simulated databases without noise and with noise.

Table 6. : Prediction errors - Simulated dataset: Without noise

Model	MAE	RMSE
Model 1 – Basic ELM	0.077	0.097
Model 2 - Bagging Ensemble	0.076	0.096
Model 3 – Basic ELM with PSO	0.074	0.095
Model 4 – Bagging ELM with PSO	0.075	0.095

Table 7. : Prediction errors - Simulated dataset: With noise

Model	MAE	RMSE
Model 1 - Basic ELM	0.1892	0.3646
Model 2 - Bagging ensemble	0.1836	0.3592
Model 3 – Basic ELM with PSO	0.1734	0.3499
Model 4 – Bagging ELM with PSO	0.1746	0.3537

It is possible to observe that, for the two sets of simulated data, the models that use PSO for parameter optimization obtained superior performance, because in both cases these methods obtained smaller prediction errors than the non-optimized models. The results obtained on the simulated data sets provides a guide about the consistency of the errors when using the optimized models.
Table 8 describes the error results for each of the models used in this work. As stated earlier, the models were applied to the educational

database, and aimed to predict the G3 variable that refers to the students' final grade.

Table 8. : Models prediction errors - Education Database

Model	MAE	RMSE
Model 1 – Basic ELM (no PSO)	2.21	3.03
Model 2 – Bagging ensemble (no PSO)	2.10	2.91
Model 3 – Basic ELM with PSO	2.06	2.88
Model 4 – Bagging ELM with PSO	2.05	2.85

For a better visualization, Figure 2 shows the regression curves using the predicted values in the y axis, and the real values in the x axis.

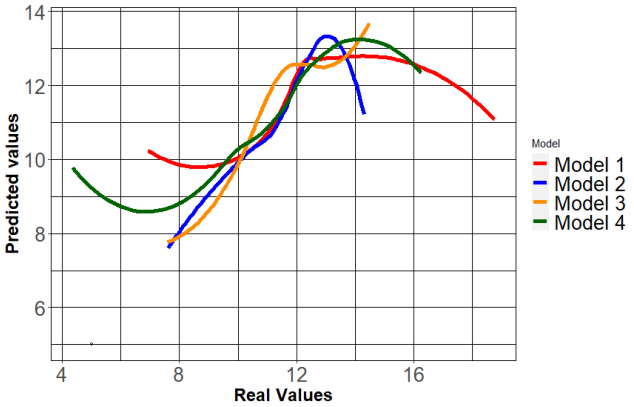

Fig. 2: Regression curves for each model

The curves shows the behavior of each model predictions in comparison with the real values. Since the perfect match occurs when the curve describes its trajetory creating the same angles for both axis, its possible to visualize that the Model 4 keeps a more constant trajetory without many abrupt curves, showing a higher stability when compared with the other models.
When observing the values it's possible to notice that between the non optimized models, the Model 2 was superior since it obtained lower prediction errors in both MAE and RMSE, reinforcing that the bagging ensemble can improve the robustness and final results when compared with Model 1, that uses only ELM.
Analyzing the optimized models, it's possible to notice again that the model which combines the bagging ensemble and the PSO was superior to the Model 3, that uses only the ELM optimized with PSO. The lower errors for both MAE and RMSE obtained through Model 4 gives more confidence to the use of the bagging ensemble and reinforces its's power in reducing the errors.
When observing the results of all models together it's possible to notice that the proposed Model 4 is able to achieve better results than all of the other models, since it has achieved the lower errors for MAE and RMSE among all the models. Thus, from the moment that bagging and PSO are used together, it is possible to emphasize the effectiveness of these methods in obtaining better performance in the educational context.
To obtain a higher degree of reliability, statistical tests were performed with the errors obtained, and then hypothesis tests were formulated to confirm the results obtained. First, the normality of the data was analysed using the shapiro-wilk test, after performing this test, it was observed that the data did not follow a normal distribution, so the Wilcoxon hypothesis test was chosen using a 5 % significance level . To verify the veracity of the results of Model 4, it was compared with all other models, using the hypothesis test described by

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

Where μ_1 represents the error values of the Model 4, and μ_2 represents the values of the model being tested. H_0 represents the null hypothesis that the values of model μ_1 are greater than or equal to that of model μ_2 and H_1 represents the alternative hypothesis that model μ_1 has results less than μ_2. Table 9 presents the results of the p-value for the Wilcoxon tests. With 95% confidence, it is possible to reject the null hypothesis in all cases, so it is possible to conclude that the difference between the results obtained by the models are statistically significant.

Table 9. : p-value for each test

Compared models	Metric	p-value
Mode 4 - Model 1	MAE	4.82×10^{-51}
Mode 4 - Model 2	MAE	8.71×10^{-50}
Mode 4 - Model 3	MAE	2.59×10^{-05}
Mode 4 - Model 1	RMSE	8.78×10^{-51}
Mode 4 - Model 2	RMSE	5.54×10^{-40}
Mode 4 - Model 3	RMSE	1.30×10^{-08}

7. CONCLUSION

The use of EDM to approach data from educational contexts is revealed to be a powerful tool, because through robust algorithms and refined techniques it is possible to model the problem in order to obtain information that can help in decision making, as well as to identify possible points that need more attention in order to mitigate a target problem in question. In this work the problem is based on the prediction of the students' grades in the Portuguese discipline of high school and thus to evaluate the performance focusing on developing approaches to predict student performance and proposing a model combining ELM neural networks, the ensemble bagging method and the PSO algorithm in order to improve the results obtained.
The models were applied to simulated data sets to confirm the consistency of the results. Then the models were applied to the educational data and after obtaining the results of each model, it was possible to ascertain in fact an improvement in the final results when compared with the other models proposed. This advantage was statistically proven through the hypothesis test that confirmed the veracity of the proposed model and provides confidence so that it is possible to affirm that the proposed model is robust and produces quality results to deal with student performance prediction problems.
Addressing the problem of student performance prediction has many benefits, as it enables better decision making and allocation of resources in a safe, intelligent and reliable way in the educational environment. For future work, it is planned to apply the models developed here to large educational databases, to analyze student performance and help in policy formulation and decision making. Additionally, it is also planned to make changes in points of the models, such as the replacement of the basic ELM with a more robust

ELM model and testing the results with others ensemble methods, to seek both greater stability and better results.

8. ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

9. REFERENCES

[1] Hind Almayan and Waheeda Al Mayyan. Improving accuracy of students' final grade prediction model using pso. In *2016 6th International Conference on Information Communication and Management (ICICM)*, pages 35–39. IEEE, 2016.

[2] RSJD Baker et al. Data mining for education. *International encyclopedia of education*, 7(3):112–118, 2010.

[3] Peter Bühlmann. Bagging, boosting and ensemble methods. In *Handbook of Computational Statistics*, pages 985–1022. Springer, 2012.

[4] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rudiger Wirth, et al. Crisp-dm 1.0: Step-by-step data mining guide. *SPSS inc*, 9:13, 2000.

[5] Jing Chen, Jun Feng, Xia Sun, Nannan Wu, Zhengzheng Yang, and Sushing Chen. Mooc dropout prediction using a hybrid algorithm based on decision tree and extreme learning machine. *Mathematical Problems in Engineering*, 2019, 2019.

[6] Paulo M da Silva, Marilia NCA Lima, Wedson L Soares, Iago RR Silva, A de A Roberta, and Fernado F de Souza. Ensemble regression models applied to dropout in higher education. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 120–125. IEEE, 2019.

[7] Thomas G Dietterich et al. Ensemble learning. *The handbook of brain theory and neural networks*, 2:110–125, 2002.

[8] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37, 1996.

[9] William J Frawley, Gregory Piatetsky-Shapiro, and Christopher J Matheus. Knowledge discovery in databases: An overview. *AI magazine*, 13(3):57–57, 1992.

[10] Magdalena Graczyk, Tadeusz Lasota, Bogdan Trawiński, and Krzysztof Trawiński. Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal. In *Asian conference on intelligent information and database systems*, pages 340–350. Springer, 2010.

[11] David J Hand, Heikki Mannila, and Padhraic Smyth. *Principles of data mining (adaptive computation and machine learning)*. MIT Press, 2001.

[12] Seyed MH Hasheminejad and M Sarvmili. S3pso: Students' performance prediction based on particle swarm optimization. *Journal of AI and Data Mining*, 7(1):77–96, 2019.

[13] Guang-Bin Huang and Haroon A Babri. Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions. *IEEE transactions on neural networks*, 9(1):224–229, 1998.

[14] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

[15] James Kennedy and Russell Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95-International Conference on Neural Networks*, volume 4, pages 1942–1948. IEEE, 1995.

[16] Mrinal Pandey and S Taruna. A comparative study of ensemble methods for students' performance modeling. *International Journal of Computer Applications*, 103(8), 2014.

[17] 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.

[18] Yuhui Shi and Russell C Eberhart. Parameter selection in particle swarm optimization. In *International conference on evolutionary programming*, pages 591–600. Springer, 1998.

[19] Yuhui Shi et al. Particle swarm optimization: developments, applications and resources. In *Proceedings of the 2001 congress on evolutionary computation (IEEE Cat. No. 01TH8546)*, volume 1, pages 81–86. IEEE, 2001.

[20] Edward Wakelam, Amanda Jefferies, Neil Davey, and Yi Sun. The potential for student performance prediction in small cohorts with minimal available attributes. *British Journal of Educational Technology*, 51(2):347–370, 2020.

[21] Qi Yu, Yoan Miche, Emil Eirola, Mark Van Heeswijk, Eric SéVerin, and Amaury Lendasse. Regularized extreme learning machine for regression with missing data. *Neurocomputing*, 102:45–51, 2013.