



## Genetically Optimized Ensemble Classifiers for Multiclass Student Performance Prediction

Safira Begum<sup>1\*</sup>      Sunita S Padmannavar<sup>2</sup>

<sup>1</sup>*Visvesvarya Technological University – RRC, India*

<sup>2</sup>*Gogte Institute of Technology, India*

\* Corresponding author's Email: safirabgm@gmail.com

---

**Abstract:** The knowledge obtained from data can be useful for the improvement of education systems, giving rise to a research space called Educational Data Mining (EDM). EDM covers the development of methods to explore information collected from educational environments, allowing to understand students more effectively and adequately, providing better educational benefits to them. Machine learning (ML) technologies are growing considerably in recent years. The field of data mining in education provides researchers and educators with metrics of success, failure, dropout, and more, allowing students to guess. The main reason for dropping out of school is not studying. Several researchers have proposed various educational data mining techniques to predict student performance and analyzed the techniques found in educational datasets. This paper proposes a student predictive model with the use of ensemble classifiers. Initially data is pre-processed and an analysis of the correlation between the entrance attributes was carried out to identify the existence of possible redundancies between them, resulting from a very high positive correlation. The filtered attribute is trained and tested with Boosting, Bagging and Random subspace classifiers. Further to improve the accuracy of predictive model genetic algorithm is applied on three classifiers. Genetic Algorithm is an approach used to find optimized solution to search problems and it intend to increase the probability of solving the problem. The process of optimization involves selection of the best option from the available set of options to achieve the desired goal. Selection is done such that the efficiency can be maximized and error can be minimized. An analysis of the correlation between the entrance attributes was carried out to identify the existence of possible redundancies between them, resulting from a very high positive correlation. There is significant improvement in classifier accuracy, when tested mathematic and Portuguese data i.e. 3 % and 11% respectively.

**Keywords:** Machine learning, Educational data mining, Boosting, Bagging, Random subspace.

---

### 1. Introduction

Educational Data Mining (EDM) deals with educational data in Data mining. It is a research area that has emerged in recent years and is used by researchers in different areas, such as education, computer science, intelligent tutoring systems, statistics and psychology in the analysis of large data sets to solve educational research problems [1]. EDM is concerned with the development of methods to explore information collected from educational environments, allowing the understanding of students more effectively and adequately, providing better educational benefits to them [2].

This research deals with the use of Machine Learning, as a tool to help school managers in checking and combating school dropout. From the analysis of historical data stored in databases, the aim was to build analytical models that can learn from data, identify patterns and help in decision making. Following are the major steps involve in data mining (DM).

Step I - Pre-processing: Data obtained from educational systems must first be pre-processed and transformed into a format suitable for data mining. The major tasks in this step includes: clean-up, attribute selection, and data integration.

Step II - Data Mining: This is the main phase and in this context, data mining techniques are applied to

preprocessed data. Here are some examples of data mining techniques: visualization, regression, classification, clustering, association rules, and more.

Stage III - Post-processing: at this stage, the results or models obtained are interpreted and used in the decision-making process regarding the educational environment.

The tasks in an EDM process that are responsible for analyzing data generated from interactions between students in learning environments can answer questions related to how to improve student learning [3]. Thus, it is possible to understand him and how he learns, in addition to other factors that influence learning. It is also possible to identify which instructional approach (individual or collaborative learning) provides the best educational benefits to students, if the student is unmotivated or confused. With this knowledge, it is possible to create personalized environments and teaching methods that offer better learning conditions [4].

The exploitation of data mining techniques in educational systems with the objective of improving the teaching-learning process is seen as a formative assessment technique. Data mining methods can be utilized to achieve information that assists educators in creating and maintaining courses in educational systems and taking action and decisions.

In this work, the aim of the prediction is to approximate anonymous values of a variable that defines the student. In education, this value can inform about academic performance, knowledge about a certain subject, grades in assessments, achievements in subjects, among others.

Section 3 discusses the methodology for this study. Section 4 initially describes the datasets and their preprocessing, secondly, students create a partial model based on their course enrollment and activity characteristics, and finally discusses the results. The interpretability and utility of the resulting partial models is analyzed in Section 5. Finally, Section 6 concludes the article and suggests future work.

## 2. Literature review

According to [5] it is possible to make dropout predictions even before students start their courses, that is, the student is enrolled, but has not yet started their classes. The author tests four classification algorithms using socioeconomic and pre-university data, such as student performance in high school. Model accuracy ranges from 91.98 % to 67.74%. In the research carried out by [6] the authors focused on predicting the pass rate of students in the first two years of undergraduate courses. For this, they used a

Multilayer Perceptron (MLP) neural network with three layers. The authors mixed demographic and academic variables creating a dataset (DataSet) with the following variables: ethnicity, gender, intention to enroll in college, age, school of origin in high school etc. The study shows that they had an average hit rate of 72%. Authors of [7] conducted a survey of students at the beginning of graduation in the Economics course. Only students who were in the first two years were considered. The objective of this work was to test and compare the following classification models: Decision Tree, Logistic Regression and Neural Networks. The SEMMA methodology (Sampling, Exploring, Modifying, Modeling and Assessment) [8] was used. For the construction of the dataset, the authors chose to use the variables related to the candidate's enrollment and attributes related to the study process. The results showed that the neural network model showed a better performance compared to the others, indicating that 36% of students will be able to drop out. In [9] only explored attributes extracted from students' academic records. Five classification algorithms and data from six courses at the Federal University of Rio de Janeiro were used. This approach led them to obtain an accuracy of at least 87.0% for each course and a true positive rate ranging from 66.08%. Authors of [9] aimed to predict student dropout through a dynamic model by Classification and Regression Trees (CART). The authors suggest adding other factors such as: age, ethnicity, work situation, study environment and type of education to improve performance. Motivated by the self-evasion rate, the authors of [10] carried out an experimental study in search of a method of predicting dropout among freshmen in the engineering course. The authors used the following models based on WEKA classifiers: Decision Trees and a Bayesian classifier. For the construction of the dataset, variables related to performance in high school and university data were considered. The results showed that one of the applied Decision Tree models reached an accuracy of 68% when analyzing only the pre-university data. When verifying the complete set of data, the method obtained accuracy between 75% and 80% in identifying evasion.

Several authors carry out analyzes and propose tools in the context of DM in education. Authors of [11] compare DM techniques in classifying students based on usage data and course final grades on the moodle learning management platform. A decision support tool was also developed for the use of instructors based on the analyzed techniques. Authors of [12] proposed a DM model to analyze the performance of higher education students, using end-

of-semester performance data in a decision tree classification task. Authors of [13] proposed a framework to direct support actions for freshman students, who need support according to their high school academic characteristics. Authors of [14] compared several classifier algorithms from the data mining tool Weka, in an attempt to predict the risk of student dropout based on the performance of the first semester. Data from students of the Electrical Engineering course at the Eindhoven University of Technology were used. The decision tree classifier presented the best results, with an accuracy between 75 and 80%.

In genetic algorithm, natural genetic mechanism is simulated to find the optimal global solution [15]. The genetic algorithm adaptively manages the search process so as to automatically accumulate and acquires information about the search space through the search process to accomplish the optimal solution. Authors of [16] proposed the GA ensemble algorithm to measure the weight of each base classifier.

Drawbacks of previous research works:

- The research work carried out by the authors of [6] used MLP neural network and achieved an accuracy of 72% which is very low. If the weight and bias values of the neural network are not optimized then certainly it will provide low computational efficiency due to more features which is a major limitation of neural network.
- The outcome from of the authors of [10] also does not meet the expectations since it is showing the maximum accuracy of 80% utilizing Bayesian classifier which is also considered as low.
- Similarly, the authors of [14] exhibited the lower accuracy.
- Rest of the papers explained in literature review do not show the numeric value of accuracy in their respective papers.

This research work utilizes the Genetic algorithm due to its simplified architecture in machine learning and data mining applications.

### 3. Proposed methodology

#### 3.1 Student performance prediction

Various data mining techniques can be used in creating models that make inferences about student performance. A model can be understood as an abstraction that makes predictions about the future on past or present information. A model comprises a machine learning algorithm that learns about characteristics of objects in a training dataset to make

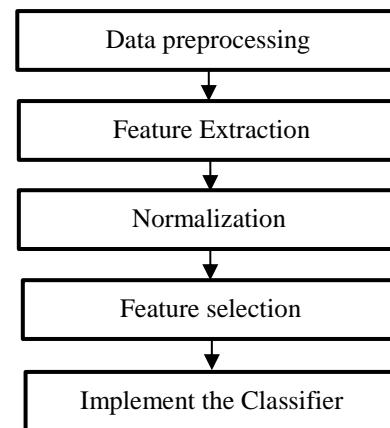


Figure. 1 Flow diagram of a predictive model

estimations about other objects. Predictive models are basically allocated into two sub-categories: regression and pattern classification. The regression analyzes the relationship between two variables in order to find trends between these variables. In pattern classification, data are assigned to classes according to data information previously reported to the machine learning algorithm.

From data input, the model can infer on this data and produce a prediction as output. The prediction obtained by the model, among other things, depends on the quality and quantity of data available. A model is built based on the information available in the database and the appropriate selection of data mining techniques.

The generated model needs to be validated to verify that it meets the conditions imposed to solve the problem. Different models can solve the problem, the one that presents the best results in the validation step is chosen. In the process of obtaining the model, two phases are performed: in the first, a set of training data will be used by the algorithm to build the descriptive model of the data. In the second phase, new data, called test data, are used in order to test the model generated in the first phase. The analysis of the results is made by the person responsible for the execution of the KDD processes [17], in order to verify if the model meets the purpose.

### 3.2 Data acquisition and pre-processing

#### 3.2.1. Data pre-processing:

Data pre-processing is an essential step in Data Mining, Data Mining (DM), which addresses several important issues in the data mining process. Data Knowledge Discovery Knowledge Database Discovery (KDD). Data from real world databases usually have inconsistencies, errors, missing values, or are simply not suitable for use in DM processes.

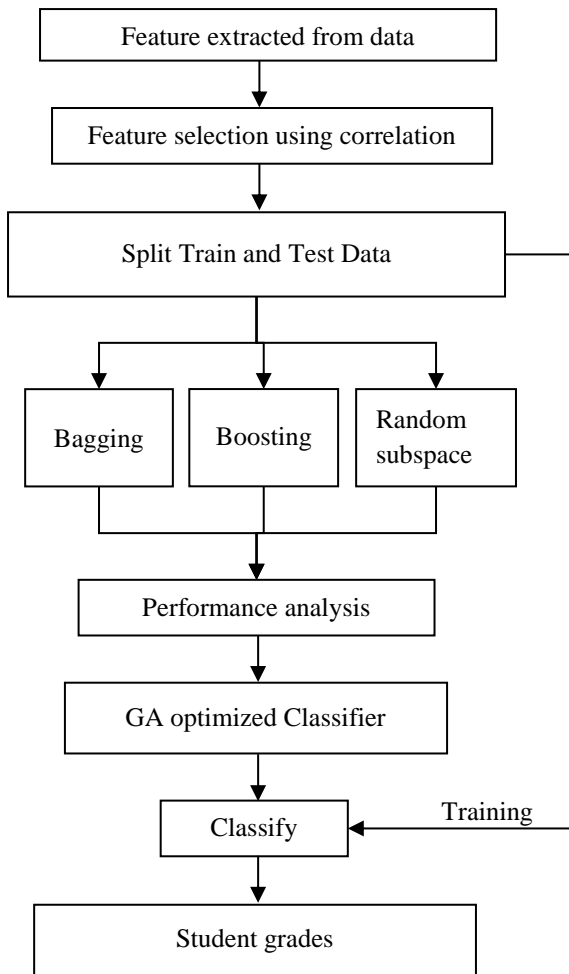


Figure. 2 Flow diagram of proposed research

### 3.2.2. Data standardization:

Normalization is a data manipulation technique that aims to make the magnitude of the attributes of the feature vectors on the same scale. Normalization techniques should not change the way the data is distributed, so the implicit information of each attribute is kept. It is common for normalized data to have intervals between 0 and 1, or -1 and 1.

Normalization makes the data more suitable for the training process of a neural network, as it avoids the saturation of neurons, speeds up the training time of a neural network, in addition to improving the accuracy and efficiency of mining algorithms that involve distance measurements.

### 3.2.3. Data normalization with z-score:

In z-score normalization, each attribute  $x$  of a feature vector is normalized based on its mean and standard deviation, as defined in Eq. (1):

$$x = \frac{x - \mu_x}{\sigma_x} \quad (1)$$

Where  $\mu_x$  is the mean of the attribute values and  $\sigma_x$  is the standard deviation.

### 3.3 Feature extraction

We collected data for this study from various sources in higher education institutions in the UCI repository [18].

In building a predictive model based on this example, we faced three main problems: data inconsistency, imbalance, and overlap. For students who have spent at least one semester in a college program, there will be several data characteristics that will help build highly accurate predictive models.

Feature extraction is an important step in classification because the effectiveness of a learning model depends on input variables (substantial features) that describe student characteristics and can be used to predict student performance.

This data refers to the performance of secondary school students in two Portuguese schools. Attribute data (including student grades, demographic, social, and academic characteristics) were collected using school reports and questionnaires. Two sets of performance data are provided in two different subjects: Mathematics (Math) and Portuguese (Po). In [19], both datasets were modeled using binary / five-level classification and regression assignment. Important note: The target G3 attribute has a strong correlation with the G2 and G1 attributes. This is because G3 is the final value (issued in the 3rd period) and G1 and G2 are the values of the 1st and 2nd periods. It is more difficult to predict G3 without G2 and G1, but these predictions are much more useful.

### 3.4 Selection of attributes

An analysis of the correlation between the entrance attributes was carried out to identify the existence of possible redundancies between them, resulting from a very high positive correlation.

In Fig. 3, the scatter plot of the Mathematics Average and Portuguese Average attributes in the entrance exam is presented with a correlation coefficient  $r = 0.62$ , which indicates a moderate positive correlation.

In Fig. 3, for the General Average and Average Mathematics attributes obtained in the entrance exam,  $r = 0.74$ , which indicates a high positive correlation. Fig. 4, for the General Average and Portuguese Average attributes in the entrance exam, we have  $r = 0.76$ , which represents a high positive correlation. None of the possible pairs of attributes has a very high positive correlation (greater than 0.5), indicating their individual relevance in the data classification task. In

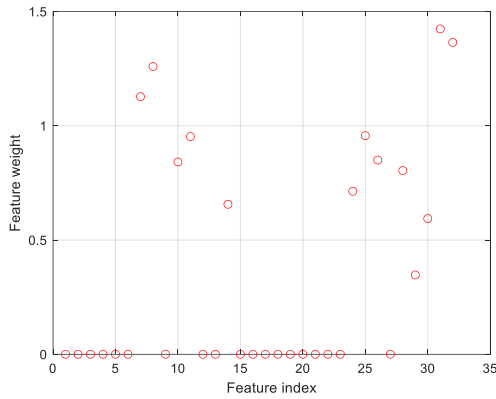


Figure. 3 Selected feature for mathematics

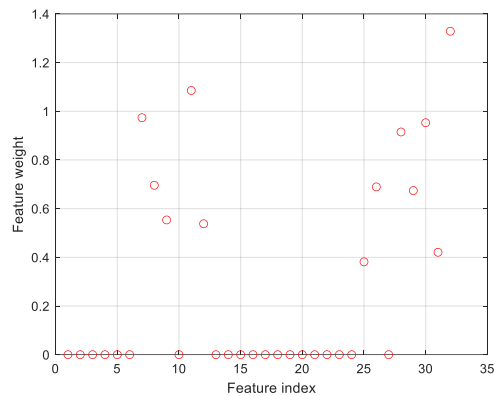


Figure. 4 Selected features for portuguese

addition, tests were performed with the exclusion of each of the variables, where the relevance of the other attributes in the classification task was observed.

- With less data to be processed, there is an increase in processing speed so DM algorithms can learn faster;
- Increases the predictive performance of the DM algorithm, so the model better generalizes from the data;
- Make the results simpler so that they are easy to understand and use;

### 3.5 Classification models

The search algorithm ends with a method ensemble model that provides the best sorting accuracy among all the combinations of prediction methods and associated hyperparameters that we have tested. The prediction engine of the identified ensemble model is based on a voting scheme that accepts prediction results from most constituent procedures. The main ensemble methods are: pockets, momentum and random subspaces. Predictive models often use supervised learning to estimate the value of an unknown or future dependent variable based on the properties of its independent variable.

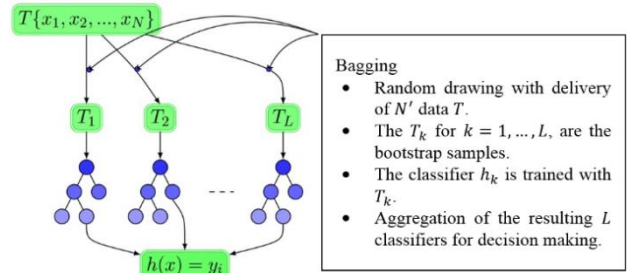


Figure. 5 Illustration of the principle of bagging for a set of decision trees [20]

#### 3.5.1. Bagging - principles of operation:

The Bagging method [20] in its "classical" version is part of the category of data manipulation methods in the taxonomy of the induction principles of set of classifiers that we presented in the previous section. It applies the principle of Bootstrap or Bootstrapping [20] to the aggregation of classifiers; hence its name Bagging for Bootstrap Aggregating.

Let us assume that a set  $T = \{x_1, x_2, x_3, \dots, x_N\}$  of  $N$  observed data of our population, and that we are interested in a statistic written  $S(T)$ .

The bootstrap will consist in forming  $L$  samples  $T_k^* = (x_1^*, x_2^*, x_3^*, \dots, x_{N'}^*)$  for  $k = 1, \dots, L$ , where each  $T_k^*$  is constituted by random pulling with  $N'$  data given in  $T$ . These  $L$  samples are usually called bootstrap samples.

We can then calculate  $S(T_k^*)$  for each bootstrap sample, and thus obtain  $L$  estimates of our statistics. So instead of having a single estimate for a sample, we have an empirical distribution of our statistics. We can then calculate the empirical average from all these values (see Eq. (2)), which will then give us a more precise estimate of the statistic, or else its standard error to measure its dispersion (see Eq. (3)).

$$S_{boot} = \sum_{k=1}^L S(T_k^*) / L \tag{2}$$

$$\widehat{se}_{boot} = \sqrt{\frac{\sum_{k=1}^L (S(T_k^*) - S_{boot})^2}{(L-1)}} \tag{3}$$

It is simply a question of considering that the statistic that one seeks to study is a learning algorithm noted  $h(x)$  and to then apply the principle of bootstrap as we have just explained it. Thus each elementary classifier  $h_k(x)$  of the set will be trained on one of the  $L$  bootstrap samples so that they are all trained on a different learning set. Fig. 5 illustrates the Bagging process applied to a set of decision trees.

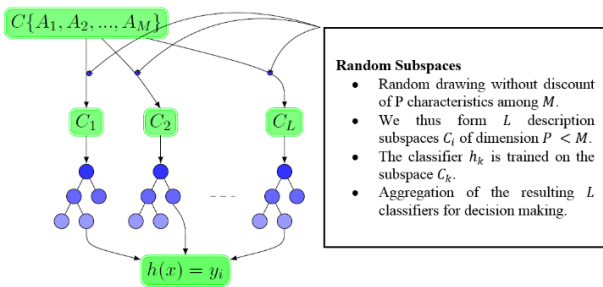


Figure. 6 Illustration of the random subspaces principle for a set of decision trees [21]

### 3.5.2. Random subspaces:

The Random Subspaces Method (RSM) is quite similar in Bagging's idea [21]. This time, however, it is no longer a question of manipulating the learning data, but of manipulating the characteristics. The basic principle is the training of each essential classifier on a random subspace of the description space. Each of these random subspaces is of the same dimension P, with P < M, where M shows the dimension of the original description space.

More concretely, for the induction of an elementary classifier  $h_k$ , it is:

- Perform a random draw without discounting P characteristics among the M available characteristics.
- To project all learning data into this new feature subspace.
- To train the classifier  $h_k$  on these projections of the learning data.

### 3.5.3. Boosting:

The name Boosting [22] refers to a learning principle rather than an EoC method and is therefore a family of several algorithms. The basic principle is to gradually specialize the classifiers of the set iteratively, and then combine each of the classifiers obtained at each iteration. Typically, it is at the iteration k to concentrate the learning of the classifier  $h_k$  on the errors of classifiers  $h_{k-1}, h_{k-2}, \dots, h_1$ , obtained at the previous iterations. In the Boosting principle, this objective is achieved by a weighting of the learning data. The weighting is then updated on the basis of these prediction errors, so as to increase the weight of the learning data that has been misclassified by this classifier, while simultaneously decreasing the weights of the well-ranked data. Thus, classifiers are gradually specialized to concentrate on learning previously poorly classified data.

#### Algorithm 1: AdaBoost.M1

**Input:**  $\mathcal{L}$  a weak learner.

**Input:** L the number of classifiers in the final set.

**Input:** T a set of N learning data

1:  $D_1(x_i) = \frac{1}{N}, i = 1, \dots, N$      **Initialization of weights (equal probability)**

2: for  $t = 1, \dots, L$  do

3:  $h_t = \mathcal{L}(D_t)$      **Learning  $h_t$**

4:  $\hat{\epsilon}_t = \sum_{i: h_t(x_i) \neq y_i} D_t(x_i)$      **Calculation of the weighted error of  $h_t$**

5: if  $\hat{\epsilon}_t = \frac{1}{2}$  then

6: stop the loop

7:  $\beta_t = \frac{\hat{\epsilon}_t}{(1-\hat{\epsilon}_t)}$      **Calculation of the weighting coefficient of  $h_t$**

8: for  $i = 1, \dots, N$  do

9: if  $h_t(x_i) = y_i$  then

10:  $D_{t+1}(x_i) = \frac{D_t(x_i)}{Z_t} \times \beta_t$

11: Else

12:  $D_{t+1}(x_i) = \frac{D_t(x_i)}{Z_t}$

13:  $h_c(x) = \arg \max_{y \in Y} \sum_{t=1}^L \log \frac{1}{\beta_t} \times \mathbb{I}_{h_t(x)=y}$

Main idea: Boosting is based on a deterministic principle for the creation of diversity in sets, whereas random forests by definition do so via the principles of randomization.

### 3.5.4. Genetically optimized classifiers:

The methodology of genetic algorithm is as follows [23]:

- A random initial population is created.
- Fitness of the initial population is evaluated.
- Reproduction includes selection, crossover and mutation. Selection is done such that the individual with best fitness level from the current population is selected. Now the best individuals reproduce to form new population. Mutation is done at a definite point in the new created individual.
- The process is repeated until a maximum number of generation has evolved or a desired solution is obtained.

By introducing weights  $w_{cj}$ , base classifier scores are  $p_{cj}$  are represented by Eq. (4)

$$p_{cj}(x_t) = w_{cj} \sum_{i=1}^n \alpha_i y_i k(x_i, x_t) + b \quad (4)$$

Where c represents  $c^{th}$  ensemble classifier, j symbolize  $j^{th}$  base classifier of the  $c^{th}$  ensemble classifier, and  $x_t$  is the test sample.

Each ensemble classifier has the best score from the base set of classifiers, and the output of each ensemble classifier corresponds to the class. Therefore, we define the best score for the ensemble

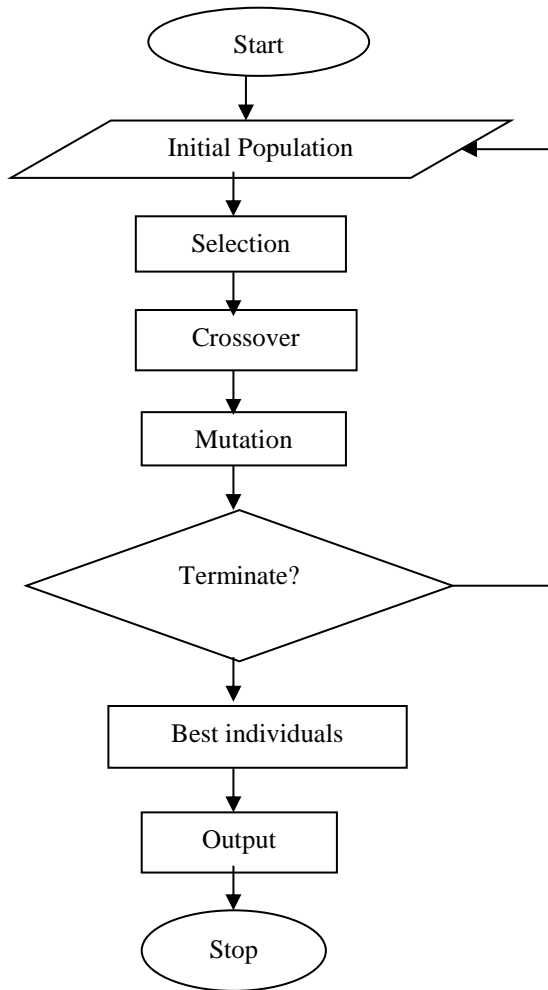


Figure. 7 Genetic algorithm flow chart

classifier  $c^{th}$  as  $p_{c_{best}}(x_t)$ . The final solution  $p_{\max(x_t)}$  is determined by Eq. (5).

$$p_{\max(x_t)} = \text{argmax}_c^n P_{c_{best}(x_t)} \quad (5)$$

Where  $n$  represents the number of classes. After entering the weights,  $w_{cj}$ . GA optimizes the weight matrix to improve classification accuracy. GA is utilized to find the ideal weight matrix to maximize the overall accuracy of the trained classifier, and the fitness function must be used to find the optimal individual. In other words, the best individual is the optimal weight matrix we can find. Therefore, Eq. (6) defines the fitness function:

$$L = \frac{1}{K} \sum_i^k \theta \quad (6)$$

Where  $\theta$  indicates the correctness of classification and defined as:

$$\theta = \begin{cases} 1 & y_i = p_{\max}(x_t) \\ 0 & y_i \neq p_{\max}(x_t) \end{cases} (x_t) \quad (7)$$

Table 1. Evaluation parameters

<b>TP (True Positive)</b>	Indicated the number of records that were classified as "correctly classified"
<b>TN (True Negative)</b>	Indicated the number of records that were classified as "not classified correctly."
<b>FP (False Positive)</b>	Indicated the number of records that were classified as "incorrectly classified"
<b>FN (False Negative)</b>	Indicated the number of records that were classified as "not Classified incorrectly"

## 4. Simulation results and discussion

### 4.1 Performance evaluation parameters

The confusion matrix, composed of the first four values: True positive, false negative, false positive and True negative. The matrix was very useful, mainly for two reasons: first because its data described the result of the classification of each record, and second because it is through it that the other metrics were obtained.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (10)$$

$$Specificity = \frac{TN}{TN+FN} \quad (11)$$

$$Error Rate = \frac{FP+FN}{TP+TN+FP+FN} \quad (12)$$

$$False Positive Rate (FPR) = \frac{FP}{FP+TN} \quad (13)$$

$$F - Score = \frac{2TP}{2TP+FP+FN} \quad (14)$$

$$Matthews Correlation Coefficient (MCC) = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}} \quad (15)$$

$$Kappa Statistics = \frac{2(TP \times TN - FN \times FP)}{(TP+FP) \times (FP+TN) + (TN+FN) \times (FN+TN)} \quad (16)$$

Table 2. Genetic algorithm parameters

<b>Population Size</b>	30
<b>No of Iteration</b>	50
<b>No of variables</b>	2
<b>Lower bound</b>	10
<b>Upper bound</b>	120

### 4.2 Dataset

In Portugal, secondary education consists of 3 years of education, followed by 9 years of primary education and then higher education. Most students enter the public education system free of charge. The database was created from two sources: certificates based on sheets of paper and with multiple attributes (for example, three-period estimates and dropouts); and the questionnaire used to fill in the above information. The data was combined into two datasets for Mathematics (with 395 examples) and Portuguese (649 datasets). During preprocessing, some features are discarded due to lack of distinguished values. The data attributes include school name, age, gender, travel time, distance from school to home, hobbies, health data, etc. They were collected using school reports and questionnaires. Details on attributes are provided in the UCI repository [18].

### 4.3 Validation strategies

Predictive accuracy informs the classification model's ability to predict correctly the class of an unknown object. This estimate allows the validation of the generated model, as well as the comparison of performances between different classifiers on the same test base.

Cross Validation: In this work, to obtain the predictive accuracy of classification models, cross-validation techniques were used. To perform cross-validation, the database is divided into two parts,

- Training: Used to train the model. Training allows you to observe object characteristics and the relationships between input and output data to create a representation of this data. Mathematics and Portuguese datasets are taken for training with split of 60:40 ratio. 60 % of data are used for training purpose.
- Testing: It is usually a set smaller than the one used in training. The test suite allows validation of the generated model. Rest of 40 % data from Mathematics and Portuguese datasets are used for testing purpose.

Cross-validation ensures that data used for training and testing are different. Thus, ensuring greater accuracy in the generalizability of the predictive model. K-fold cross-validation divides a dataset  $D$  into  $K$  parts of equal sizes, called folds and represented by  $D_1, D_2, \dots, D_k$ . Each part  $D_i$  is, in turn, set aside to be the test set, and the other parts together will be the training set  $D - D_i$ . After training the  $M_i$  model with  $D - D_i$ , its performance will be evaluated with the test set  $D_i$  to obtain the  $i^{th}$  estimate  $\theta_i$ . The

accuracy will be the average of the percentages of correct answers in the  $K$  iterations and can be estimated by:

$$\widehat{\mu}_\theta = \frac{1}{k} \sum_{i=1}^k \theta_i \tag{17}$$

And its variance is given by:

$$\widehat{\sigma}_\theta^2 = \frac{1}{k} \sum_{i=1}^k (\theta_i - \widehat{\mu}_\theta)^2 \tag{18}$$

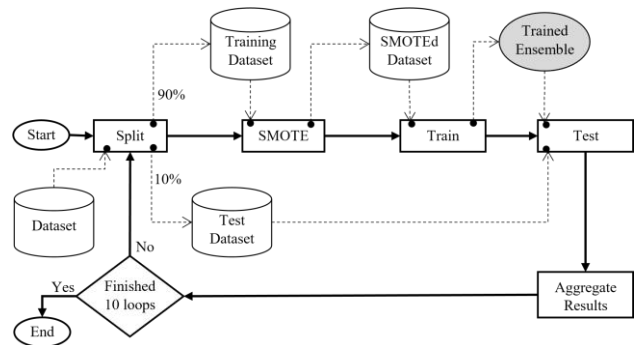
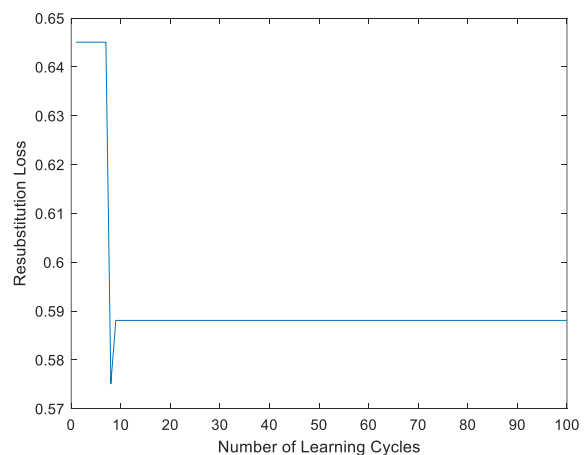
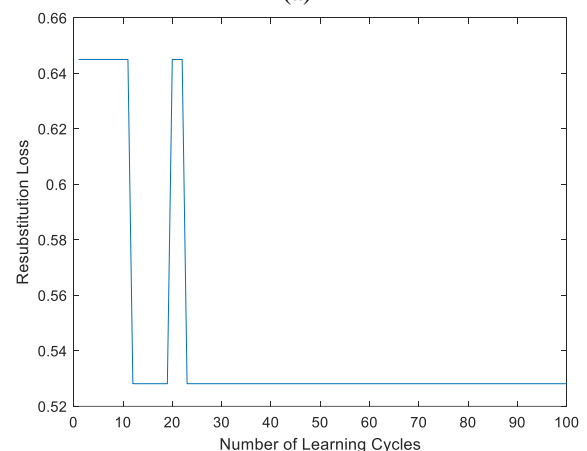


Figure. 8 10-folds cross validation of the model accuracy [24]



(a)



(b)

Figure. 9 Resubstitution plot: (a) Maths training data and (b) Portuguese training data



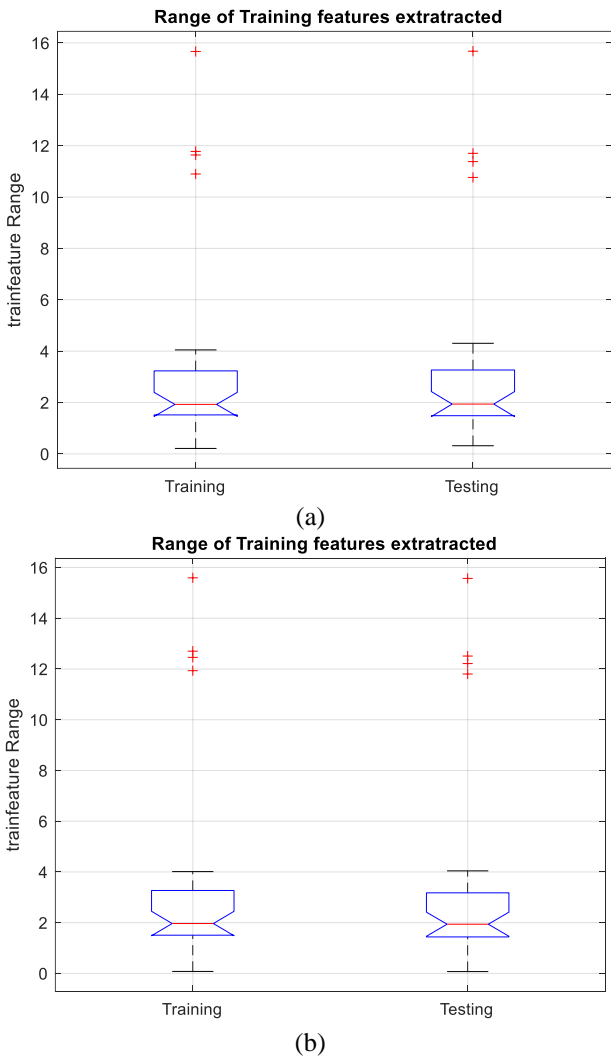


Figure. 10 Notch plot of training features: (a) Maths dataset and (b) Portuguese datasets

The literature informs that K equal to 10 is the most adequate to obtain a good estimation of classifier accuracy [24]. A disadvantage of this technique is the excessive computational cost, since K trainings are performed to evaluate the model [24].

A decrease in re-substitution loss might indicate that the software trained the ensemble sensibly.

The pre-test data for the control group report that the Q1o lower quartile is 1.5 points and Q3 or upper quartile of 2.8 points out of the 4 possible, generating an Interquartile Range RI of 1.8 points, which indicates that 50% of the student data is between 1.8 and 2.8, the median had a value of 2. The feature in the proposed system is less skewed. Skewness indicates that the data may not be normally distributed. Therefore, the filtered features have a stable data distribution from the classifier as the training sample.

The true positive rate (which represents the proportion of individuals in a class that were classified correctly), there are 21 outcomes to predict.

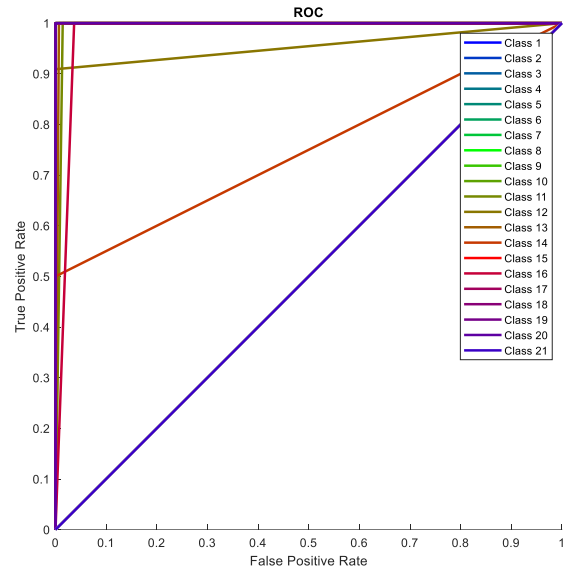


Figure. 11 ROC curve of boosting for maths dataset

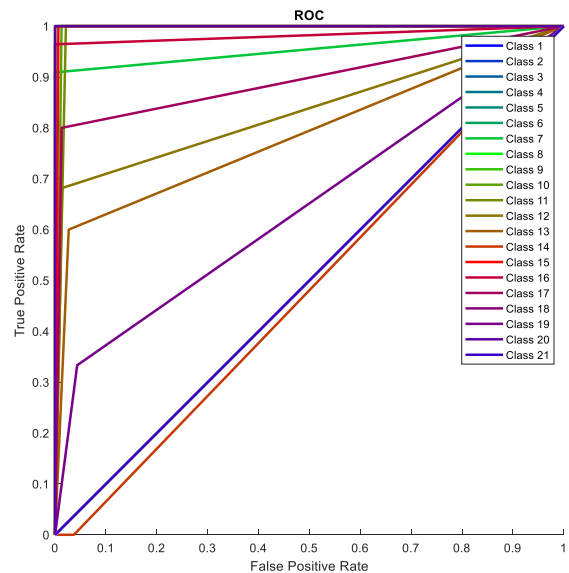


Figure. 12 ROC curve of bagging for maths dataset

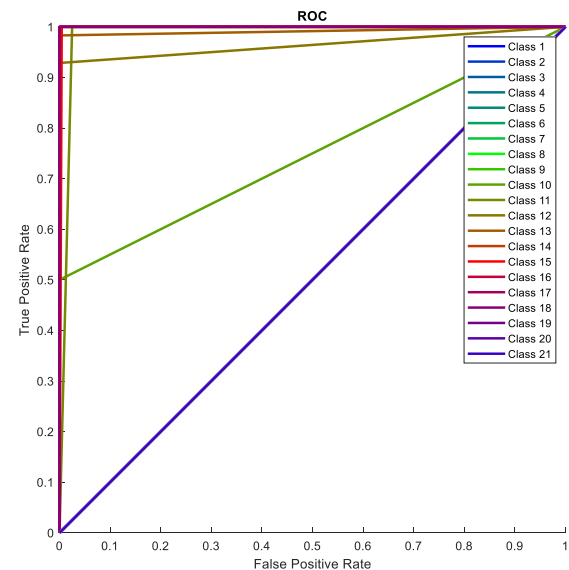


Figure. 13 ROC curve of boosting for portuguese dataset

Table 3. Comparative analysis of normal classifiers and GA optimized classifier on mathematics dataset

Parameters	Boosting	Bagging	Random subspace	Optimized Boosting	Optimized Bagging	Optimized Random subspace
Accuracy	0.9811	0.8321	0.7671	0.9768	0.8622	0.8459
Error	0.0189	0.2006	0.2329	0.0871	0.1822	0.1566
Sensitivity	0.9890	0.7241	0.8259	0.8761	0.7361	0.8717
Specificity	0.9984	0.9833	0.9028	0.9927	0.9861	0.9518
Precision	0.9545	0.7170	0.7461	0.8757	0.7281	0.8338
FPR	0.0016	0.0167	0.0972	0.0073	0.0139	0.0482
F1-score	0.9650	0.7079	0.7378	0.8758	0.7205	0.8352
MCC	0.9668	0.6984	0.6612	0.8685	0.7137	0.7954
Kappa	0.8671	0.1294	0.4778	0.8457	0.0041	0.5913

Table 4. Comparative analysis of normal classifiers and GA optimized classifier on portuguese

Parameters	Boosting	Bagging	Random subspace	Optimized Boosting	Optimized Bagging	Optimized Random subspace
Accuracy	0.9682	0.7343	0.4593	0.9682	0.9009	0.4603
Error	0.0530	0.2804	0.5407	0.0530	0.1204	0.5373
Sensitivity	0.8914	0.6900	0.7363	0.8914	0.8130	0.6714
Specificity	0.9941	0.9668	0.7937	0.9941	0.9861	0.8204
Precision	0.8794	0.6271	0.4548	0.8794	0.8105	0.4329
FPR	0.0059	0.0332	0.2063	0.0059	0.0139	0.1796
F1-score	0.8819	0.6372	0.3945	0.8819	0.8085	0.4237
MCC	0.8777	0.6157	0.3025	0.8777	0.7961	0.3297
Kappa	0.8267	0.3138	0.1768	0.8267	0.4587	0.3039

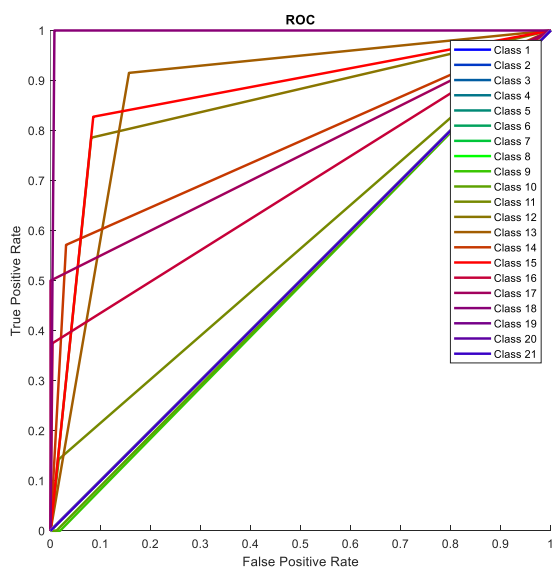


Figure. 14 ROC curve of bagging for portuguese dataset

Observing the results shown in Table 3, which represents mathematic data, it can be seen that the

accuracy of the algorithms was never less than 80%. The algorithm that achieved greater precision was the Boosting, with an average of 98.11%. It can also be observed that students in grade A (passed in 3 subjects) are more difficult to be classified than students in grade B (failed in at least 1 subject). It is believed that this is due to the fact that the student's good performance in the entrance exam is not the only influencing factor in the student's success in the first period of the course, despite being an extremely important characteristic in their success. It can be seen there is a significant improvement in Bagging and Random subspace classifier when optimized with GA. The observed accuracy of Bagging and Random subspace is 83.21% and 76.71% whereas GA optimized yields 86.22% and 84.59% respectively.

Observing the results in above table, which represents Portuguese data, it can be seen that the accuracy of the algorithms is low because of data attributes. The algorithm that achieved greater precision was the Boosting, with an average of

Table 5. Performance analysis of existing and proposed GA optimised ensemble classifiers with different classifiers for UCI maths data

Methods	Accuracy	Selected Features
Logistic Regression (SVM)[27]	62.05%	“sex, Fedu, sex, Pstatus, sex, Mjob, sex, study-time, age, reason, Medu, sex, guardian, sex, Fjob, famsize, address, travel-time, sex, sex”
RFBT-RF (SVM)[26]	79.74%	“internet, higher, Fjob, Pstatus, nursery, activities, famsup, sex, Mjob, famsize, address, schools-up, Medu, Fedu, age, travel-time, paid, reason, Guardian, failures, study-time”
RFBT-RF(NBC)[26]	88.60%	
SVM [25]	89.7%	N-A
Proposed Boosting	98.11%	“internet, higher, Fjob, Pstatus, nursery, activities, sex, Mjob, famsize, address, schools-up, Medu, Fedu, age, travel-time, paid, reason, failures, study-time”
Proposed Bagging	86.22%	
Proposed Random subspace	84.11%	

96.82%. It can be seen there is a significant improvement in Bagging and Random subspace classifier when optimized with GA. The observed accuracy of Bagging and Random subspace is 73.43% and 45.93% whereas GA optimised yields 90.09% and 46.03% respectively.

To compare the results obtained with the methods proposed in this work with the results obtained in other works, tests were carried out with the methods of classification used in some works. This comparative study was carried out as described. The proposed methods for Mathematics and Portuguese datasets outperform than Relief-F and Budget Tree-Random Forest as improvement of 11 % of accuracy. Whereas Logistic Regression based method gives 62.05% and 67.69 % in mathematics and Portuguese dataset respectively, which is very low as compared to proposed method.

Table 6. Performance analysis of existing and proposed GA optimised ensemble classifiers with different classifiers for UCI portuguese data

Methods	Accuracy	Selected Features
Logistic Regression (SVM)[27]	67.69%	“sex, sex, traveltime, sex,sex, address, age, studytime, sex, sex, sex, famsize, Mjob, sex, guardian”
RFBT-RF (SVM)[26]	66.92%	“freetime, famrel, school, romantic, guardian, higher, studytime, famsup, internet, age, nursery, Medu, Fedu, paid, activities, Mjob, Fjob, address, Pstatus, schoolsup, famsize, reason, failures, sex, traveltime”
RFBT-RF(NBC)[26]	81.53%	“sex, age, address, famsize, Pstatus, Medu, Fedu, Mjob, Fjob, travellin, studytime, famsup, internet”
Proposed Boosting	96.82%	
Proposed Bagging	90.09%	
Proposed Random subspace	46.03%	

### 5. Conclusion

It is believed that the results obtained in this study can help educators, since it is possible to obtain estimates of student performance, and then serve as a basis for planning strategies and policies aimed at reducing the number of failures, reducing, as Consequently, the evasion of students from the Mathematic course and Portuguese language. The reported work in this paper contributes in the field of predicting student academic success. Specifically, it relies on GA optimized Ensemble Boosting, Bagging and Random subspace are ensemble classifier are explored with two different dataset and achieve accuracy is 98.11%,83.21% and 71.76 % respectively in Mathematics and 96.82%,73.435% and 45.93% respectively for Portuguese data. Further genetic algorithm is applied to improve the classification accuracy and significant improvement is observed in Bagging and Random subspace classifiers on both the dataset. For mathematics observed accuracy there is 86.22% and 84.59% whereas 90.09% and 46.03 % in Portuguese dataset respectively. Hence it is concluded significant improvement in accuracy for an optimal student performance prediction model. As

future work, it is considered that a more detailed analysis in the process of selecting the input attributes is essential, investigating the existence of other variables that may influence student performance in the first period of the course. Studying the relationship between the performance of students in the first period and their performance in subsequent periods is also necessary. Furthermore, it is believed that it is possible to carry out similar research for students from different courses, verifying whether the results are similar for other subjects present in the entrance exam.

### Conflicts of Interest

The authors declare no conflict of interest.

### Author Contributions

“Conceptualization, S.B. and S.P.; methodology, S.B. and S.P.; software, S.B.; validation, S.B.; data curation, S.B.; writing—original draft preparation, S.B.; review and editing, S.P”.

### Reference

- [1] A. Ali and J. Kalita, “Intelligent Tutoring systems: A comprehensive Historical Survey with Recent Developments”, *University of Colorado Colorado Springs, College of Engineering and Applied Sciences, Department of Computer Science ArXiv Preprint ArXiv:1812.09628*, pp. 1-31, 2018.
- [2] C. Romero and S. Ventura, “Educational Data Mining and Learning Analytics: An Updated Survey”, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 10, No. 3, e1355, pp. 1-21, 2020.
- [3] I. G. Ndukwe, B. K. Daniel, and R. J. Butson, “Data Science Approach for Simulating Educational Data: Towards the Development of Teaching Outcome Model (TOM)”, *Big Data and Cognitive Computing*, Vol. 2, No. 3, p. 24, 2018.
- [4] G. Mahajan and B. Saini, “Educational Data Mining: A State-Of-The-Art Survey on Tools and Techniques used in EDM”, *International Journal of Computer Applications & Information Technology*, Vol. 12, No. 1, pp. 310-316, 2020.
- [5] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, “Analyzing Undergraduate Students' Performance Using Educational Data Mining”, *Computers & Education*, Vol. 113, pp. 177-194, 2017.
- [6] A. R. Iyanda, O. D. Ninan, A. O. Ajayi, and O. G. Anyabolu, “Predicting Student Academic Performance in Computer Science Courses: A Comparison of Neural Network Models”, *International Journal of Modern Education & Computer Science*, Vol. 10, No. 6, pp. 1-9, 2018.
- [7] Y. Nieto, V. G. Díaz, C. Montenegro, C. C. González, and R. G. Crespo, “Usage of Machine Learning for Strategic Decision Making at Higher Educational Institutions”, *IEEE Access*, Vol. 7, pp. 75007-75017, 2019.
- [8] S. A. M. Selamat, S. Prakoonwit, R. Sahandi, W. Khan, and M. Ramachandran, “Big data analytics—A Review of Data-Mining Models for Small and Medium Enterprises in the Transportation Sector”, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 8, No. 3, p. e1238, 2018.
- [9] F. Jimenez, A. Paoletti, G. Sanchez, and G. Sciavicco, “Predicting the Risk of Academic Dropout with Temporal Multi-Objective Optimization”, *IEEE Transactions on Learning Technologies*, Vol. 12, No. 2, pp. 225-236, 2019.
- [10] S. Gnanambal, M. Thangaraj, V. T. Meenatchi, and V. Gayathri, “Classification Algorithms with Attribute Selection: An Evaluation Study Using WEKA”, *International Journal of Advanced Networking and Applications*, Vol. 9, No. 6, pp. 3640-3644, 2018.
- [11] N. Ademi, S. Loshkovska, and S. Kalajdziski, “Prediction of Student Success Through Analysis of Moodle Logs: Case Study”, In: *Proc. of International Conf. on ICT Innovations*, pp. 27-40, 2019.
- [12] A. S. Olaniyi, S. Y. Kayode, H. M. Abiola, S. I. T. Tosin, and A. N. Babatunde, “Student's Performance Analysis using Decision Tree Algorithms”, *Annals. Computer Science Series*, Vol. 15, No. 1, pp. 55-62, 2017.
- [13] M. S. Rodríguez, C. Tinajero, and M. F. Páramo, “Pre-Entry Characteristics, Perceived Social Support, Adjustment and Academic Achievement in First-Year Spanish University Students: A Path Model”, *The Journal of Psychology*, Vol. 151, No. 8, pp. 722-738, 2017.
- [14] N. E. R. Maya, C. L. Álvarez, O. M. Tzuc, and B. A. S. Carranza, “Modeling Students' Dropout in Mexican Universities”, *Research in Computing Science*, Vol. 139, pp. 63-175, 2017.
- [15] Z. Wang and A. Sobey, “A Comparative Review between Genetic Algorithm Use in Composite Optimisation and the State-Of-The-Art in Evolutionary Computation”, *Composite Structures, Elsevier*, Vol. 233, p. 111739, 2020.
- [16] H. Lu, H. Gao, M. Ye, and X. Wang, “A Hybrid Ensemble Algorithm Combining AdaBoost and Genetic Algorithm For Cancer Classification

- With Gene Expression Data”, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 18, No. 3, pp. 863-870, 2019.
- [17] E. Espitia and A. F. Montilla, “Applying CRISP-DM in a KDD Process for the Analysis of Student Attrition”, In: *Proc. of Colombian Conf. on Computing*, pp. 386-401, 2018.
- [18] Dataset, Online Available at: <https://archive.ics.uci.edu/ml/datasets/studentperformance>
- [19] P. Cortez and A. M. G. Silva, “Using Data Mining to Predict Secondary School Student Performance”, *IEEE Transactions on Learning Technologies*, Vol. 1, No. 2, pp. 1-8, 2008.
- [20] S. González, S. García, J. D. Ser, L. Rokach, and F. Herrera, “A Practical Tutorial on Bagging and Boosting based Ensembles for Machine Learning: Algorithms, Software Tools, Performance Study, Practical Perspectives and Opportunities”, *Information Fusion*, Vol. 64, pp. 205-237, 2020.
- [21] M. Hosni and A. Idri, “Software Effort Estimation using Classical Analogy Ensembles Based on Random Subspace”, In: *Proc. of Symposium on Applied Computing*, pp. 1251-1258, 2017.
- [22] K. Berkling and K. Neubehler, “Boosting Student Performance with Peer Reviews; Integration and Analysis of Peer Reviews in A Gamified Software Engineering Classroom”, In: *Proc. of IEEE Global Engineering Education Conference*, pp. 253-262, 2019.
- [23] S. Fletcher, B. Verma, Z. M. Jan, and M. Zhang, “The Optimized Selection of Base-Classifiers For Ensemble Classification using a Multi-Objective Genetic Algorithm”, In: *Proc. of International Joint Conf. on Neural Networks*, pp. 1-8, 2018.
- [24] Y. Yang and J. O. Pedersen, “A Comparative Study on Feature Selection in Text Categorization”, *The International Conference on Machine Learning*, Vol. 97, pp. 412-420, p. 35, 1997.
- [25] H. Pallathadka, A. Wenda, E. R. Asís, M. A. López, J. F. Albornoz, and K. Phasinam, “Classification and Prediction of Student Performance Data Using Various Machine Learning Algorithms”, *Materials Today: Proceedings, Elsevier*, pp. 1-4, 2021.
- [26] K. Deepika and N. Sathyanarayana, “Relief-F and Budget Tree Random Forest Based Feature Selection for Student Academic Performance Prediction”, *International Journal of Intelligent Engineering and Systems*, Vol. 12, No. 1, pp. 30-39, 2019.
- [27] C. Mason, J. Twomey, D. Wright, and L. Whitman, “Predicting Engineering Student Attrition Risk using a Probabilistic Neural Network and Comparing Results with a Backpropagation Neural Network and Logistic Regression”, *Research in Higher Education*, Vol. 59, No. 3, pp. 382-400, 2018.