International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 / https://doi.org/10.14445/22315381/IJETT-V70I4P219

Volume 70 Issue 4, 223-235, April 2022

Original Article

Prediction of Student Performance using Genetically Optimized Feature Selection with Multiclass Classification

Safira Begum[1], Sunita S Padmannavar[2]

[1]*Department of Master of Computer Applications, Visvesvarya Technological University – RRC, Belagavi, Karnataka, India.*
[2]*Department of Master of Computer Applications, Gogte Institute of Technology, Belagavi, Karnataka, India.*

[1]safirabgm@gmail.com

Received: 23 February 2022 Revised: 02 April 2022 Accepted: 05 April 2022 Published: 26 April 2022

Abstract *- Educational data mining is the key aspect of improving students' performance in education. the academic performance of students or instructors can be predicted by using the techniques and algorithms in educational data mining and data mining. the paper proposed a machine learning approach to predict the academic performance of secondary school students in Mathematics and Portuguese lessons. the proposed algorithm primarily applies the normalization and z-score normalization in the pre-processing stage to solve the unbalanced class distribution problem. Then, feature selection processes are performed using a Genetic algorithm. Students' success in Mathematics and Portuguese lessons is estimated by the k-nearest neighbour (KNN), linear discriminant analysis (LDA) and support vector machine (SVM) classifications. the experimental results compare the accuracy, precision, F-score, and sensitivity values of the abovementioned methods.*

Keywords *- Educational Data Mining, K-Nearest Neighbour, Linear Discriminant Analysis, Machine Learning, Support Vector Machine.*

1. Introduction

In recent years, education has changed, as a result of the available technological advance that has led to the instrumentation of the educational sector, both in teaching software, in the digital administration of academic records by the managers of the institutions, as well as in the use of the internet for learning, especially due to the popularization of e-learning. All these factors have driven exponential growth in the volume of educational data, and to analyze a large amount of data, it is essential to have computing resources. Otherwise, the task becomes impractical [1].

In this way, data mining techniques are gaining more and more important in the educational sector, as they are a way of monitoring, analyzing and evaluating the learning process [2]. Probably, data mining techniques can provide educational policymakers with models to support their goals of improving the efficiency and performance of teaching.[2]. in addition, various machine learning approaches can be seen as the basis for a systemic change, capable of positively impacting the solutions of specific problems in educational institutions, for example, enabling solutions that involve the personalization of educational environments or providing efficient structure for the decision-making process in the educational environment [1] [2] [3] [4].

Educational data mining (EDM) stands out in this scenario, which uses (DMDM) techniques to extract relevant information from diverse educational data sets. According to the international educational data mining society, this area can be defined as follows:

It is an emerging discipline concerned with developing methods to explore unique and increasingly large-scale data from educational contexts and uses these methods to better understand students and the settings in which they learn [5].

In other words, DM refers to a set of computational techniques to extract information from large masses of data. When the analyzed data comes from educational contexts, it is called EDM [3]. Likewise, the authors of [6] define EDM as an area dedicated to developing methods to explore data from educational environments and use them to understand the teaching and learning processes better. in this sense, the authors of [7] claim that EDM is the research area that aims to improve and mature techniques to investigate data sets obtained in educational settings. According to the authors, the nature of these data is more diverse than that observed in the data traditionally used for mining operations, which require adjustments and new approaches. At the same time, the diversity of these data represents the potential for an important resource to improve education [6] [7] [8].

Thus, methods and tools are needed to assist in the task of validating, interpreting and linking these data to obtain useful and relevant information, which, according to the authors, has become the focus of DMDM methods for revealing behavioural patterns intuition leads to improved products and services [6].

Data mining is used in various fields of expert systems. Currently, with the demand for distance learning and computer courses, various researchers in the field of computer science in education (especially the application of artificial intelligence in the educational model) have explored the data mining to study scientific problems in education (for example, what factors influence education or how to create a more efficient education system?). in this regard, a new field of research called Educational Data Mining has gained attention.

EDM is defined as a field of research whose main focus is the development of methods for studying datasets collected in educational institutions. in this way, students can better understand how they learn, the role of the learning context, and other factors that influence learning. for example, you can determine in which situations a pedagogical approach (such as individual or collaborative learning) offers students the greatest pedagogical benefit.

The EDM query process is similar to the DMDM query process in that the workflow of a typical data mining process includes the following steps: 1) data collection; 2) the function of extracting and cleaning data (pre-processing and transformation) so that the data can be processed; 3) Analytical Processing and Algorithms: Development of efficient analytical methods to extract relevant knowledge and information from processed data; and the author also suggests that the results need to be analyzed and/or interpreted, so it is up to the researcher to verify the best way to carry out this analysis. the sequence of steps in the process proposed by the authors of [9] is shown in Figure 1, in which it can be seen that the DMDM process can be iterative.

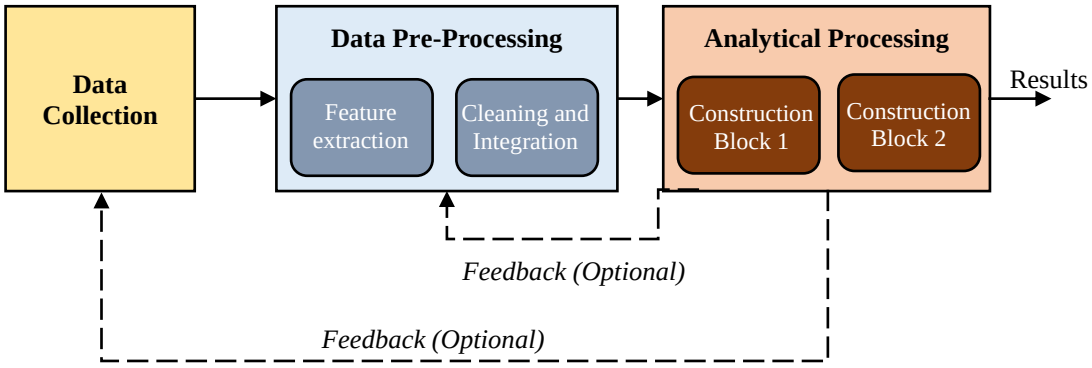


Fig. 1 Data Mining Process [9]

Some more prominent techniques can be used regarding the analytical processing step, such as descriptive and inferential statistics, Machine Learning (ML), or Deep Learning (DL). Since 2006, this technique has attracted a lot of attention. It has been successfully applied in many areas such as pattern, speech and image recognition [10], machine health monitoring [11], machine learning, intrusion detection, natural language processing, and medical prediction.

However, according to [12], DLP applications in the educational context are relatively scarce, at least so far, compared to ML, which is more established as a Data Mining technique. DL can be used in feature extraction, pattern recognition and classification, so it is an approach capable of solving problems in the educational field [12]. in a recent systematic literature mapping, authors of [13] identified that of the 158 articles mapped, only 9 used DL as the EDM technique. the body of studies on ML in the EDM is very extensive, to the detriment of DL, which is growing in several other areas. Still, it is not yet consolidated in the context of educational research. Thus, it is important to carry out studies like this one that denote the effectiveness of deep learning and how it can be useful in research in this area, in which its use can lead to a focus on improving behaviour modelling and student performance, expanding the horizons of studies in educational data mining.

Most importantly, EDM plays a key role in predicting student performance, which seeks to identify how the student will perform during the course to intervene if necessary and thus improve their learning process [13]. in the mapping mentioned above, authors of [13] pointed out 18 studies that address this topic, being the second most investigated among researchers (second only to behaviour analysis). This is because student performance is a mandatory aspect of educational institutions since one of the criteria for schools and universities to be considered of best

track record of academic achievement [14]. Therefore, the authors of [14] state that predicting student achievement is very useful for teachers and students in improving the teaching and learning process.

In this context, this study aims to predict student performance; for this, a common dataset from the UCI machine learning repository is utilized; with this, it was possible to compare techniques already consolidated in the scope of the EDM with the machine learning technique. the approach used was supervised learning for classification, in which students' grades are predicted, but these were divided into four categories, their numerical values not being used. With these predictions, it was also possible to verify if the attributes that make up the database are sufficient to generate effective models in predicting student performance and evaluate genetic algorithm-based feature selection to ensure redundancy management and proper training in supervised classification. Finally, this study intends to make a document that explains how to carry out the educational data mining process available to those interested in the area.

To this end, this document is arranged as follows: section 2 represents the literature review; Section 3 deals with the main aspects of the pre-processing and feature selection techniques, as well as the architecture of KNN, LDA and SVM classifiers used in this study; section 4 discuss the results section; finally, section 5 describes the authors' conclusions with the development of this study.

2. Literature Review

Research in the field of EDM has been going on for many years. the authors of [15] improved the prediction accuracy of student data using an ensemble model using various classification algorithms. They also identified association rules that affect student performance using rule-based methods [15]. Authors of [16] analyze the effect of data pre-processing on classification algorithms on a student performance dataset with uneven classroom distribution. in this direction, the unbalanced class distribution problem is solved by applying undersampling and oversampling techniques for support vector machines, decision trees and naive Bayes classification algorithms. in the experimental results, higher accuracy values are achieved with the SMOTE algorithm belonging to the oversampling class [16]. Authors of [17] analyze the success of various classification algorithms using Weka open source software on a student performance dataset. in the experiments, they observed that some features affect the student performance more during the pre-processing data stage and the accuracy performance of the classification algorithms increases after this stage [8]. Authors of [18] use decision tree-based classification algorithms to predict students' performance [18]. Authors of [19] estimate student performance on two different student datasets using linear regression, decision tree and naive Bayes algorithms. the experimental results show that the accuracy values of the classification algorithms increase after the feature selection process for both data sets [19]. Iterative classifier, OneR, LogitBoost and artificial neural network methods are applied to predict the student's performance level on the student achievement dataset. They obtain better performance values with the OneR method than others [20]. Authors of [21] evaluate the performance of classification algorithms for the student data set by applying the feature selection method to the decision tree. the random forest algorithm obtains the best accuracy values [21]. Authors of [22] predict the students' success in the courses with the Naive Bayes classification algorithm on the student performance dataset [22]. Authors [23] analyzed decision trees and support vector machine algorithms with 10-fold cross-validation on the student dataset. to increase the performance of the algorithms, they performed hyperparameter optimization with the grid search algorithm [23]. Their study [24] applies three different approaches to the student data set: binary classification, five-level classification, and regression. in these approaches, k-means, nearest neighbour, support vector machines and naive Bayes algorithms are used for comparison [24]. Authors of [25] used a support vector machine, multilayer perceptron and random forest algorithm to predict students' willingness to attend a higher education program. in the experimental results, they concluded that the random forest algorithm is more successful when compared to other algorithms [25]. Authors of [26] used classification algorithms such as naive Bayes to predict student performance. They observed that the schools and working hours of the students had a significant impact on the final grade [26]. Authors [27] tried to determine the most appropriate model to predict student performance using k-nearest neighbour and decision tree algorithms. the experimental results obtained better performance values with the decision tree algorithm [27]. Authors of [28] analyzed the effect of the fast correlation-based filtering method on the support vector machine classification algorithm on three different student performance datasets. the experimental results concluded that the feature selection process improves the student academic performance prediction model [28]. Authors of [29] examined the effects of students' living conditions and social environment on their Turkish, Mathematics lessons and their general success averages at the end of the semester. in this direction, the performance evaluation of students (0-100) was estimated by regression methods, and the course grade was estimated by classification methods based on the 5-point scale. in the experiments, it was determined that the logistic classification algorithm, in which the random forest regression method in estimating the course score and the correlation-based feature subset method in the estimation of course grades, showed good results [29].

3. Proposed Methodology

3.1 The flow of Proposed Student Performance Prediction

EDM transforms the data collected in the educational environment into useful information. EDM is often used to predict student performance. to develop EDM, feature selection is important to find optimal attributes and classification is used to predict. in this study, using the proposed feature selection approach based on genetic algorithms, the performance of high school students in language lessons was assessed using the SVM, K-NN, and LDA classification methods. to enhance the prediction accuracy of classification, the problem of an unbalanced distribution of classes is firstly solved. the methods mentioned above are described in the research subsection.

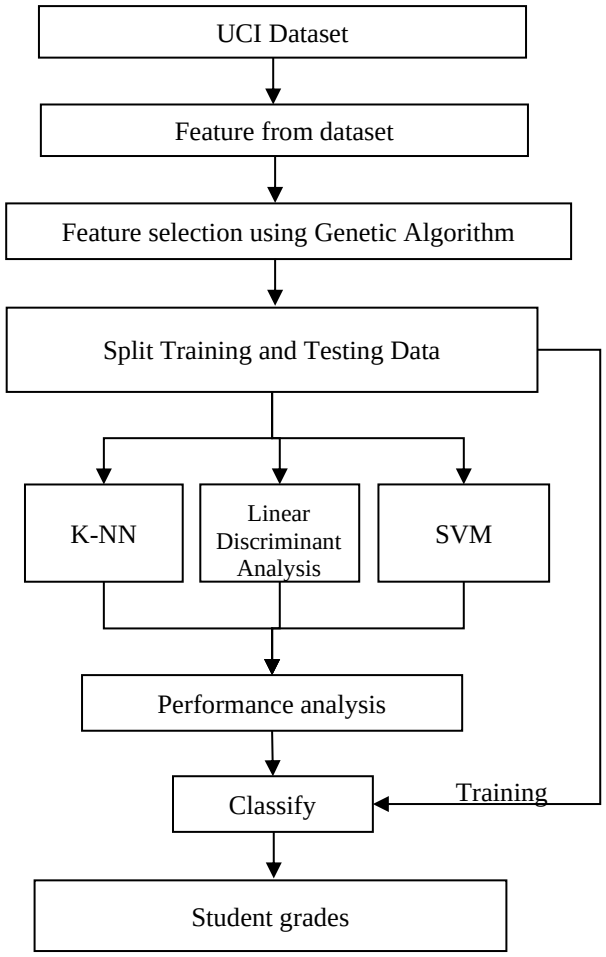


Fig. 2 proposed research outline

3.2 Data Acquisition and Pre-Processing

3.2.1 Data Pre-Processing

Data from real-world databases usually have inconsistencies, errors, missing values, or are simply unsuitable for DMDM processes.

3.2.2 Data Normalization

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.

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 equation (1):

$$x = \frac{x - \mu_x}{\sigma_x} \tag{1}$$

Where μ_x is the mean of the attribute values , and σ_x is the standard deviation.

3.3 Extraction of feature

The collected data for this study came from various sources in higher education institutions in the UCI repository [30]. Based on this example, building a predictive model faced three main problems: data inconsistency, imbalance, and overlap. for students who have spent at least one semester in a college program, several data characteristics 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. It can be used to predict student performance.

This data contains the performance results of two Portuguese secondary school students. Attribute data (including student grades, demographics, social and academic characteristics) was collected through newsletters and surveys. Two sets of performance data are provided for two different subjects: Mathematics (Math) and Portuguese (Po). in [31], both datasets were structured using binary or five label regression and classification. the G3 is the target attribute, closely related to attributes G2, G1. G3 is the end value (given in the 3rd period), and G1 and G2 are the values of the 1st and 2nd periods. Without G2 and G1, predicting G3 is more difficult, but these estimates are much more useful.

3.4 Feature Selection using Genetic Algorithm

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. This paper uses a Genetic algorithm for the selection of features. the genetic algorithm proposed by John Holland in 1975 is a well-known evolutionary technology. Genetic algorithms draw inspiration from biological methods such as Mendel's laws and the theory of

evolution proposed by Charles Darwin. It is the process of finding a solution to a particular problem that living beings emulate in their evolution. for example, it uses the same words as classical biology and genetics, genes, chromosomes, individuals, populations and generations [32].

- A Gene: It is a set of symbols representing the value of a variable. in most cases, a gene is represented by a single symbol (a bit, an integer, a real, or a character).
- A Chromosome: It is a set of genes present in a given order in a way that takes into account the constraints of the problem to be treated. for example, in the commercial traveller problem, the size of the chromosome is equal to the number of cities to travel to. Its content represents the order of travel in different cities. in addition, care must be taken that a city (represented by a number or a character, for example) should not appear in the chromosome more than once. Figure 3 illustrates a general diagram of the steps in the search process of the genetic algorithm.
- an Individual: is composed of one or more chromosomes. It represents a possible solution to the problem dealt with.
- A Population: is represented by a set of individuals (i.e. the set of solutions to the problem).
- A Generation: is a succession of iterations composed of a set of operations allowing the passage from one population to another.

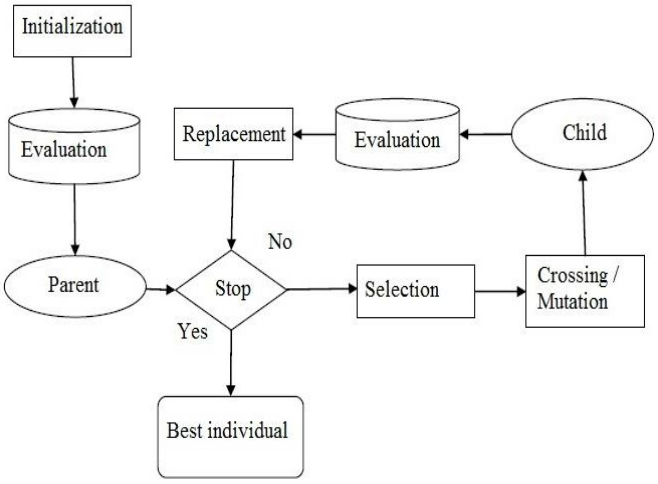


Fig. 3 Approach of a genetic algorithm [32]

The search process of the genetic algorithm is based on the following operators [33]:

- An Operator Coding Individuals: It allows the representation of chromosomes representing individuals.
- An Operator of Initialization of the Population: It allows the production of the individuals of the initial population. Although this operator intervenes only once and at the beginning of the research, it plays a non-negligible role in converging towards the global optimum. the choice of the initial population can search for the optimal solution to the problem treated easier and faster.
- A Selection Operator: It helps promote the reproduction of the individuals who have the best fitness (i.e. the best qualities).
- A Crossing Operator: It allows the exchange of genes between parents (two parents in general) to create one or two children trying to combine the good characteristics of the parents. This operator aims to create new individuals by exploiting the research space.
- A Mutation Operator: It consists in modifying some genes of the chromosomes of the individuals to integrate more diversity within the process of the research.
- An Evaluation Operator: It enhances the quality of individuals based on the objective function (the fitness function) that calculates the quality of each individual.
- In addition to different operators to guide the search by the genetic algorithm, the latter requires several basic parameters, on which depend the various operators mentioned above. These parameters must be fixed in advance, and they play a very important role in the algorithm's performance. Let's talk about: the size of the population, the probability of crossing, the probability of mutation, and the maximum number of generations.
- The Size of the Population: It represents the number of individuals in the population. If it is too big, the search process requires a high search cost, whether memory space or computation time. However, if it is too small, the algorithm risks falling in the case of premature convergence because of the lack of diversity in the population. It is, therefore, preferable to choose an average size taking into account the instance of the problem to be treated [33].
- The Probability of Crossing: It represents the probability of exchanging wealth (i.e. genes) between two or more individuals. the larger it is, the more it allows the generation of new children who can be better than their parents.
- The Probability of Mutation: It is generally weak to avoid the possibilities of radical modifications of the solutions, especially solutions of good qualities, which require only little improvement to pass to the optimal solutions.
- The Maximum Generation Number: This parameter can act as a stop criterion. It can build an obstacle for

the algorithm. It can prevent different operators from finding the best solution if it is too small, as it can generate a prohibitive calculation time in the case where it is too big. Thus, the choice of its value can be based on preliminary tests. the algorithm represents all the steps of the genetic algorithm [33].

Development of Genetic Algorithm
Start
Generate an initial random population
Repeat
Fitness assessment
Selection
Reproduction
Crossover
Mutation
Until a termination criterion is reached
End

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 the proposed model has tested. Following are the classification methods that are used in this paper.

3.5.1 K-Nearest Neighbor (K-NN)

The KNN is certainly easy to implement learning/classification algorithms. Within the learning algorithms, this falls into the class of supervised algorithms. the vast majority of classification methods begin by inducing the classificatory model to deduce the new cases' class. However, in KNN, both tasks are linked (transduction) [34].

The idea behind this paradigm is that new cases are most often placed in a class belonging to their K nearest neighbours. Using KNN with categorical data, the algorithm will return the category to which the unknown case must belong. If used with continuous data, the algorithm will return the mean of the neighbour values. the use of KNN in classification is based on the vote (majority) to decide the most appropriate value. the vote can be with or without weights. Something to keep in mind in the case of binary classification is that it is interested in an odd k value to avoid ties. Let's see better how it works with an example:

In figure 4, it can see that there are 10 examples belonging to two different classes, 5 to class A in yellow and 6 to class B in purple. Each example has two attributes that help classify it, X_1 and X_2. As can be seen, the distance between the examples of each class is generally less than the distance between the examples of the opposite class, so they appear to be grouped according to their class. for the classification of a new example, represented by the red star in the centre. If $k = 3$, the new example will be classified with class B since 2 of those 3 closest examples belong to that class. However, if $k = 6$, the example will fall into class A since 4 of the 6 examples belong to that class. Therefore, the k is a determining parameter, although not the only one.

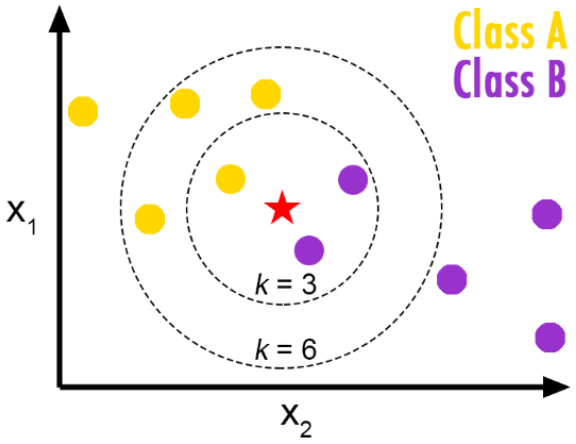


Fig. 4 Visual example of KNN with two variables and two classes [34]

The algorithm uses the distances between cases (examples, vectors) to calculate the closest ones. the choice of the distance metric is also critical to the algorithm's performance. Among the most common metrics it can be found, example, the Euclidean distance and the Manhattan distance:

$$(\mathrm{DP}, q) = \sqrt{\sum_{i-1}^{N} (q_i - p_i)^2} \tag{2}$$

$$(\mathrm{DP}, q) = \sum_{i-1}^{N} |q_i - p_i| \tag{3}$$

to deal with distributions (e.g. histograms), it is usual to use the chi-square distance. in the formula, n is the number of bins, $x1_i$ is the value of the first bin, and $y1_i$ is the value of the second.

$$\sum_{i=1}^{n} \frac{(x1_i - y_{1i})^2}{(x1_i + y_{1i})} \tag{4}$$

Or

$$\frac{1}{2} \sum_{i=1}^{n} \frac{(x1_i - y1_i)^2}{(x1_i + y_{1i})} \tag{5}$$

3.5.2 Linear Discriminant Analysis (LDA)

The LDA is a reference algorithm in supervised

classification. It can be understood in two complementary ways:

- A geometric approach which amounts to looking for hyperplanes that best separate the groups;
- A model approach assumes that the distributions of the covariates are Gaussian vectors with different parameter values for each group.

The proposed work considers $(x_1, y_1), \ldots, (x_n, y_n)$ a sample where x_i has values in $\mathbb{R}^d$ and y_i in $\{0,1\}$. the geometric approach amounts to looking for a straight line of $\mathbb{R}^d$ with equation $a_1 x_1 +, \ldots + a_d x_d = 0$ such that [35]:

- the centres of gravity of each group project on this line are at best separated $\Longrightarrow$, maximizing the inter-class distance.
- Projected observations are close to their projected centre of gravity $\Longrightarrow$ and minimize intra-class distance.

The compromise between these two distances is obtained by maximizing the Rayleigh coefficient, which is the quotient between these two distances:

$$J(a) = \frac{B(a)}{W(a)} = \frac{a^t B_a}{a^t W_a} \tag{6}$$

Where B and W are the inter and intra class matrices, respectively defined by:

$$B = \frac{1}{n}\sum_{k=1}^{K} n_k (g_k - g)(g_k - g)^t \tag{7}$$

and

$$W = \frac{1}{n}\sum_{k=1}^{K} n_k V_k \tag{8}$$

Where,

$$V_k = \frac{1}{n_k}\sum_{i:Y_i=k} (X_i - g_k)(X_i - g_k)^t \tag{9}$$

Here g designates the centre of gravity of the cloud x_i, $i = 1, \ldots, n$ and $g_k, k = 0,1$ the centres of gravity of the two groups. the solution is given by an eigenvector associated with the largest eigenvalue of $W^{-1}B$.

The model approach assumes that the vectors $X|Y = k$, $k = 0,1$ are Gaussian vectors with expectation $\mu_k \in \mathbb{R}^d$ and variance-covariance matrix Σ. These parameters are estimated by maximum likelihood, and the a posteriori probabilities are deduced by the Bayes formula [35]:

$$P(Y = k|X = x) = \frac{\pi_k f_{X|Y=k}(x)}{f(x)} \tag{10}$$

3.5.3 Support Vector Machines (SVM)

Decision support systems, which are generally called machine learning methods, are the process of estimating what the output corresponding to the given inputs is, with the help of previous knowledge and experience, in cases where the actions to be taken cannot be clearly defined. SVM is a machine learning method built on strong statistical theories. Vapnik first proposed it in 1995 for classification and regression type problem-solving. Traditional machine learning methods address the need for large amounts of training data, slow convergence rates, local minima, and over/underfitting problems. SVM overcomes this problem by operating based on internal risk minimization. the variant of SVM used in classification applications is SVS (Support Vector Classification), and the variant used in regression applications is SVR (Support Vector Regression). SVM is also successful in applications with high dimensions but few data. Because of these features, SVM has been used in many application areas such as data mining, customer fraud detection and image classification [36].

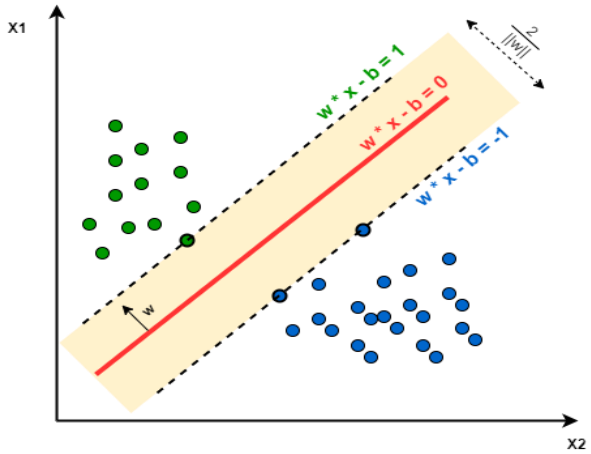


Fig. 5 SVM classifier structure [36]

As shown in Figure 5, SVM classifies by finding the separator plane with the maximum spacing between classes. the equation of the plane he finds is expressed by equation (11).

$$f(x) = \langle w, x \rangle + b \tag{11}$$

Here, $w \in \mathbb{R}^n$ is the weight vector, and b is the scalar constant. N-dimensional x vectors represent the training data. the dot product of w and x is added to the b scalar, and the result of the function is obtained. an N-dimensional vector represents each training data. the m numbers of data that make up the dataset are labelled with one of the elements in the set $y \in \{+1, -1\}$, and they have to satisfy the condition in equation (12). Here, the $\xi_i \geq 0$ condition is met.

$$y_i[\langle w, x_i\rangle + b] \geq 1 - \xi_i, i = 1, \ldots, m \quad (12)$$

To find the optimum separator plane, the minimum value of the objective function in equation (13) must be found, depending on the condition in equation (12).

$$C\sum_{i=1}^{n}\xi_i + \frac{1}{2}\|w\|^2 \quad (13)$$

Equation (13) should be made minimum using conditions of the form (12). Here, C is the user-defined positive parameter that allows for striking a tradeoff between the complexity of the separator plane and the classification accuracy. the following set of equations is obtained if the primitive problem is given in equations (12) and (13) is written in a binary optimization problem structure with Lagrangian multipliers. Here, the target function in the first part is tried to be maximized according to the conditions of inequality and inequality in the second part of the equation (14). the target function value also depends on the inner product of the input dataset.

Maximum

$$Q(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,k=1}^{n}\alpha_i y_i \alpha_k y_k \langle x_i, x_k\rangle \quad (14)$$

Condition

$$\sum_{i=1}^{n}\alpha_i y_i = 0, \qquad \alpha_i \geq 0, \forall i \quad (15)$$

The following decision function is created by using positive value Lagrange multipliers α_i. Data with such multiplier values are called support vectors. These data are the data that best represent the separator plane in the dataset.

$$f(x) = sign\left[\sum_{i=1}^{\#sv}\alpha_i y_i K(x, x_i) + b\right] \quad (16)$$

The best Lagrange multiplier values are found from the optimization problem established using equations (11-16). Then the separator plane is created with these values. This plane is much more sparse than the general dataset, created based only on support vector points.

The SVM classifier described so far can be used successfully in linear applications. However, in non-linear applications, kernel functions are needed for SVM to classify. the data in the non-linear input space is transformed into a linear high-dimensional feature space with the kernel function. the inner product kernel function in equation (17) also provides this transformation.

$$K(x, x_i) = K(x_i, x) = \varphi(x)^T\varphi(x_i) \quad (17)$$

in the proposed study, the RBF kernel function was used. This function is expressed by equation (18).

$$K(x, x') = \exp\left[\frac{-\|x - x'\|^2}{\sigma^2}\right] \quad (18)$$

σ is a user-defined positive real number and represents the width.

4. Simulation Results and Discussion

4.1 Performance Evaluation Parameters

A confusion matrix consist of TP, TN, FP, FN. This matrix is useful for two main reasons: first, because the data represents the classification results of each data set, and second because the matrix is responsible for other metrics.

Table 1. Evaluation parameters

TPTP	Indicated the number of records that were classified as correctly classified.
TNTN	Indicated the number of records classified as not classified correctly.
FPFP	Indicated the number of records that were classified as incorrectly classified.
FNFN	Indicated the number of records classified as not Classified incorrectly.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (19)$$

$$Precision = \frac{TP}{TP + FP} \quad (20)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (21)$$

$$F - Score = \frac{2TP}{2TP + FP + FN} \quad (22)$$

Table 2. Genetic Algorithm Parameters

Population Size	40
No of Iteration	100
No of variables	1
Lower bound	Min(index)
Upper bound	Max(index)

Min(index) and Max(index) depend on the dataset used for training.

4.2 Dataset

Portuguese Secondary education consists of three 3 years of study followed by 9 years of primary education and higher education. Most students enter the public school

system for free. the database was created from two sources: paper references of different quality (for example, a three-hour forecast and screenings); and the questionnaire used to complete the above information. There are two datasets: Mathematics subjects (where 395 samples have been taken) and Portuguese subjects (with 649 samples). Some properties are ignored due to a lack of unique values. Various parameters have been considered to create attributes, including school name, student age, student gender, travel time from home, the total distance from school to home, student's hobbies, student health information, etc. the dataset is collected through school reports and surveys. Details about attributes can be found in the UCI repository [30].

in this article, Portuguese math and linguistic scores (e.g. G3 in Table 3) are modelled using three supervised approaches:

- the threshold is set for the pass or fail. if G3≥10, then students are pass below that threshold are considered to fail otherwise;
- 5 class classification Erasmus qualifications transformation system (Table 3);

Table 3. Multiclass classification system

	1	2	3	4	5
origin	(excellent /very good)	(good)	(satisfactory)	(sufficient)	(fail)
Portugal	16-20	14-15	12-13	10-11	0-9
Ireland	**A**	**B**	**C**	**D**	**F**

KNN, LDA and SVM models require pre-processing before installing the models. the dummy variable is changed from encoding 1 to encoding **0.** Then the DMDM model will be loaded. Other methods accept default settings for KNN (e.g. T = 500), LDA (e.g. E = period 100) and SVM. in addition, KNN and SVM training will be improved by tuning hyperparameters with an internal grid search.; where H and γ values are selected from a set of predefined structures.

For each DMDM model, three input configurations were tested as the G1, and G2 classes were expected to have a significant impact:

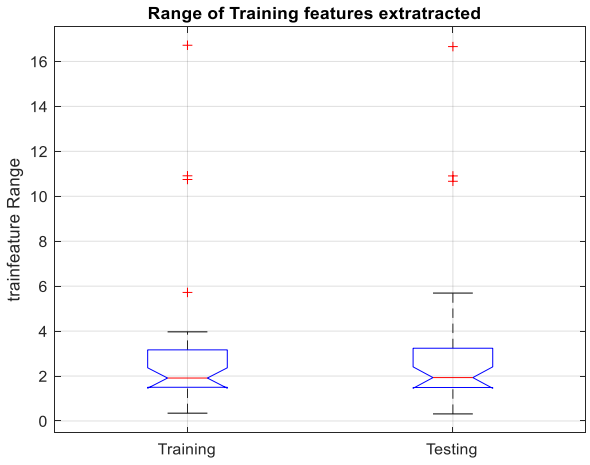
- **A** – it incorporates all the carriable of table 1; it does not include G3 (exit);
- **B** – same as A, it does not include G2 (second semester);
- **C** - same as B; it does not include G1 (first-semester class).

for configuration **A**, this pattern corresponds to the notes of the second period (G2 or double version/5 steps). If the second class is not available (configuration **B**), the value of the first period (or the binary/5-level variant) is used. If the value is missing (configuration **C**), the most common class (for a classification task) or the average output value is returned. the test series results are shown in Tables 4 through 9 for accuracy, precision, F-score, and precision. Setting **A** gives the best results, as expected. Assessed performance declined when the second-hour grade (**B**) was unknown, and the worst results were obtained when student grades (**C**) were not used. Using only the latest estimates available (SVM method for the first two input configurations), mathematical classification (binary and 5-level), and Portuguese input option **A** is the best choice for regression purposes.

This indicates that non-evaluative data is not useful in this case. But the scenario changed for another experiment. in 8 cases, SVM was the best choice, LDA with the 4 best results, and KNN with lower accuracy.

4.2 Simulation Results

(a) Mathematics data



(b) Portuguese data

Fig. 6 Notch plot of training features for Mathematics and Portuguese datasets

Pre-test data for the control group showed that the lower quartile Q1 was 1.5 points and the Q3 or upper quartile was 2.8 points out of 4 points, providing an interquartile RIRI

range of 1.8 points means 50. the percentage of data on students ranged from 1.8 to 2.8, with an average of 2. the features of the proposed system are less biased. Skewness indicates that the data may not be normally distributed. Therefore, the filtered features have a stable distribution of the classifier data as a training set.

The table below (Tables 4 to 9) shows the simulation results for various classifier configurations with and without feature selection using the Portuguese and Mathematics datasets with and without feature selection.

Table 4. Simulation outcome for Portuguese data using KNN classifier with and without feature selection

	KNN				Feature Selection with KNN			
Class	**Accuracy**	**Precision**	**Sensitivity**	**F-Score**	**Accuracy**	**Precision**	**Sensitivity**	**F-Score**
F	0.745	0.638	0.811	0.357	0.782	0.639	0.894	0.373
A	0.859	0.778	0.928	0.423	0.823	0.686	0.945	0.397
B	0.749	0.545	0.92	0.342	0.683	0.469	0.819	0.298
C	0.693	0.5	0.815	0.31	0.781	0.679	0.853	0.378
D	0.569	0.366	0.615	0.229	0.793	0.706	0.855	0.387

Table 5. Simulation outcome for Portuguese data using LDA classifier with and without feature selection

Class	LDA				Feature Selection with LDA			
	Accuracy	**Precision**	**Sensitivity**	**F-Score**	**Accuracy**	**Precision**	**Sensitivity**	**F-Score**
F	0.85	0.837	0.859	0.424	0.914	0.893	0.933	0.456
A	0.98	1	0.961	0.49	0.912	0.875	0.945	0.454
B	0.777	0.6	0.928	0.364	0.731	0.558	0.853	0.337
C	0.659	0.423	0.801	0.277	0.772	0.646	0.864	0.37
D	0.657	0.48	0.744	0.292	0.796	0.673	0.892	0.384

Table 6. Simulation outcome for Portuguese data using SVM classifier with and without feature selection

Class	SVM				Feature Selection with SVM			
	Accuracy	**Precision**	**Sensitivity**	**F-Score**	**Accuracy**	**Precision**	**Sensitivity**	**F-Score**
F	0.848	0.782	0.9	0.418	0.878	0.806	0.943	0.434
A	0.908	0.875	0.936	0.452	0.933	0.897	0.967	0.465
B	0.741	0.529	0.917	0.336	0.785	0.644	0.897	0.375
C	0.763	0.6	0.889	0.358	0.819	0.719	0.898	0.399
D	0.751	0.649	0.815	0.361	0.853	0.782	0.913	0.421

Table 7. Simulation outcome for Mathematics data using KNN classifier with and without feature selection

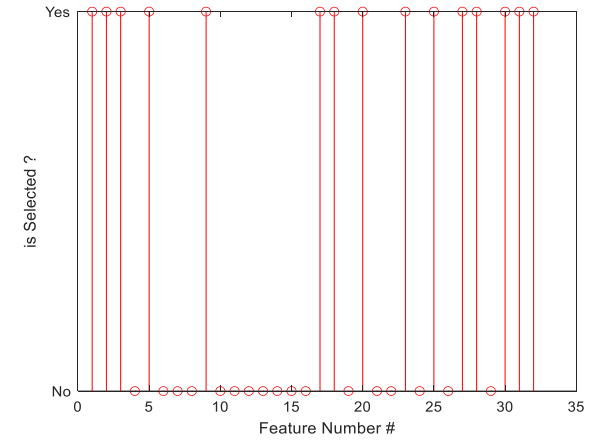
Class	KNN				Feature Selection with KNN			
	Accuracy	**Precision**	**Sensitivity**	**F-Score**	**Accuracy**	**Precision**	**Sensitivity**	**F-Score**
F	0.763	0.595	0.897	0.358	0.819	0.729	0.89	0.401
A	0.808	0.643	0.959	0.385	0.896	0.857	0.929	0.446
B	0.61	0.357	0.723	0.239	0.701	0.459	0.889	0.303
C	0.682	0.507	0.779	0.307	0.739	0.579	0.853	0.345
D	0.758	0.691	0.797	0.37	0.715	0.595	0.784	0.338

Table 8. Simulation outcome for Mathematics data using LDA classifier with and without feature selection

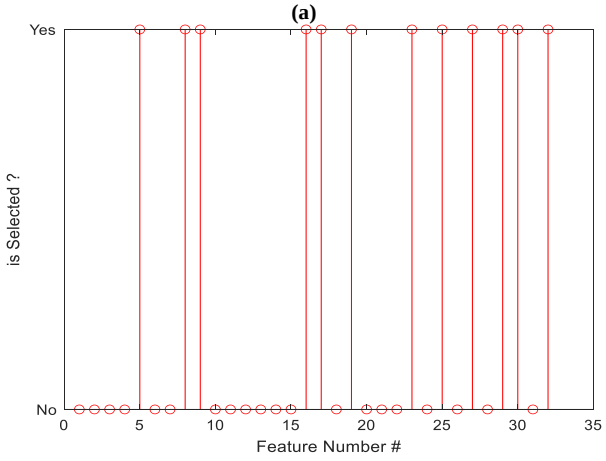
Class	LDA				Feature Selection with LDA			
	Accuracy	**Precision**	**Sensitivity**	**F-Score**	**Accuracy**	**Precision**	**Sensitivity**	**F-Score**
F	0.857	0.763	0.939	0.421	0.883	0.9	0.87	0.442
A	0.922	0.871	0.971	0.459	0.967	1	0.939	0.484
B	0.774	0.634	0.88	0.369	0.749	0.545	0.92	0.342
C	0.767	0.641	0.857	0.367	0.65	0.421	0.776	0.273
D	0.81	0.724	0.875	0.396	0.698	0.508	0.819	0.314

Table 9. Simulation outcome for Mathematics data using SVM classifier with and without feature selection

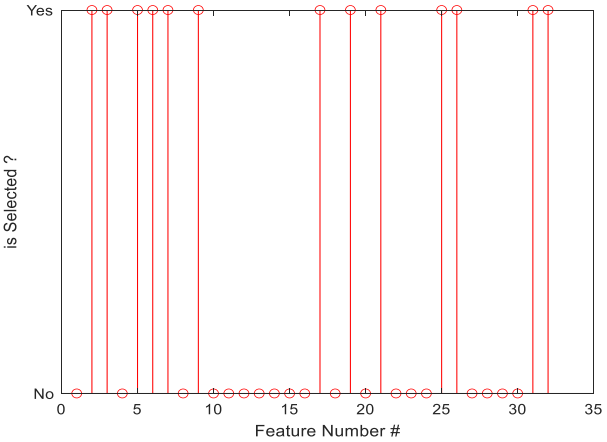
Class	SVM				Feature Selection with SVM			
	Accuracy	Precision	Sensitivity	F-Score	Accuracy	Precision	Sensitivity	F-Score
F	0.854	0.744	0.953	0.418	0.906	0.868	0.939	0.451
A	0.872	0.771	0.967	0.429	0.859	0.778	0.928	0.423
B	0.778	0.63	0.894	0.37	0.751	0.543	0.931	0.343
C	0.768	0.65	0.85	0.368	0.766	0.619	0.877	0.363
D	0.823	0.766	0.865	0.406	0.815	0.732	0.878	0.399



(a)

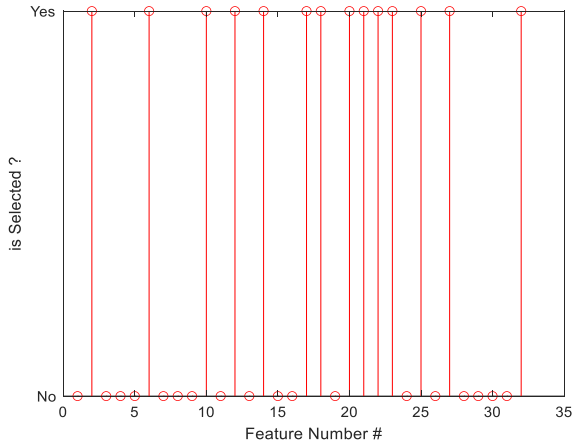


(b)

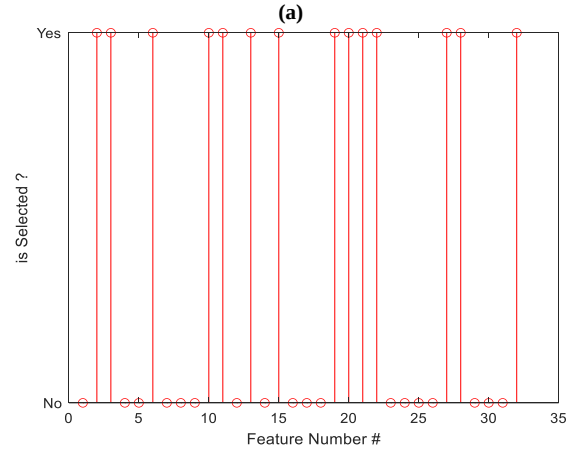


(c)

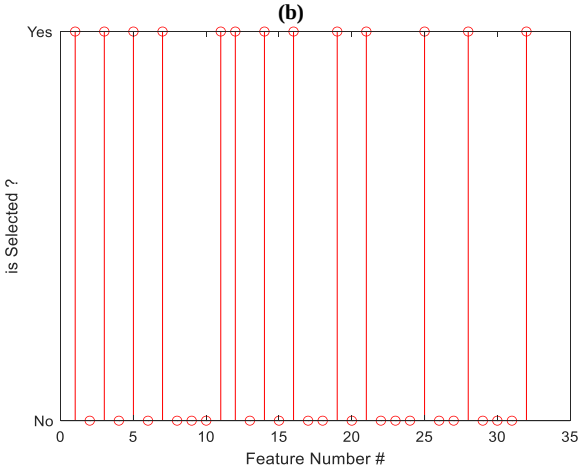
Fig. 6 Feature selection graphs of outcome for Portuguese data using (A) KNN, (B) LDA, and (C) SVM classifiers, respectively



(a)



(b)



(c)

Fig. 7 Feature selection graphs of outcome for Mathematics data using (A) KNN, (B) LDA, and (C) SVM classifiers, respectively

Table 10. Comparison of accuracy

Method	Accuracy of datasets used	
	Portuguese	Maths
KNN	0.541	0.567
LDA	0.62	0.713
SVM	0.673	0.713
Feature Selection with KNN	0.644	0.623
Feature Selection with LDA	0.69	0.71.3
Feature Selection with SVM	0.759	0.723

Table 11. Comparative analysis of maths data

Methods	Accuracy	Selected Features
Logistic Regression (SVM) [37]	62.05%	sex, Fedu, sex, Pstatus, sex, Mjob, sex, study-time, age, reason, Medu, sex, guardian, sex, Fjob, famsize, address, travel-time, sex, sex
Proposed KNN	62.3%	internet, higher, Fjob, P status, nursery, activities, sex, Mjob, famsize, address, schools-up, Medu, Fedu, age, travel-time, paid, reason, failures, study-time.
Proposed LDA	71.3%	-----
Proposed SVM	72.3%	------

Table 12. Performance analysis of existing and proposed GAGA optimized ensemble classifiers with different classifiers for UCI Portuguese data

Methods	Accuracy	Selected Features
Logistic Regression (SVM)[37]	67.69%	sex, sex, travel-time, sex,sex, address, age, studytime, sex, sex, sex, famsize, Mjob, sex, guardian
RFBT-RF (SVM)[38]	66.92%	free time, family, school, romantic, guardian, higher, study time, famsup, internet, age, nursery, Medu, Fedu, paid, activities, Mjob, Fjob, address, P status, schoolsup, famsize, reason, failures, sex, travel-time.
Proposed KNN	64.4%	free time, family, school, romantic, guardian, higher, studytimeage, nursery, Medu, Fedu, paid, Fjob, address, P status, schoolsup, famsize, reason, failures, sex, travel-time.
Proposed LDA	69 %	------
Proposed SVM	75.9 %	------

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 Relief-F and Budget Tree-Random Forest with an improvement of 9 % accuracy. At the same time, the GA-SVM-based method gives 72.3% and 75.9 % in mathematics and Portuguese datasets, respectively, which is very low compared to the proposed method.

5. Conclusion

Numerous studies have been carried out to estimate students' academic performance with machine learning algorithms. With the appropriate data pre-processing process and the selection of the right algorithm, it is possible to improve the prediction results. in this study, various classifier algorithms are proposed to predict secondary school students' success in Mathematics and Portuguese lessons. in the data pre-processing step with the proposed algorithm, the problem of unbalanced class distribution is handled with data normalization methods. in the feature selection phase, genetic algorithm methods are used. After normalizing the data to the [0, 1] interval, hyper-parameter tuning is performed for the proposed classifier algorithms for five-level classification in the training phase. Experiments show that the SVM method is the most appropriate method for the unbalanced class distribution problem.

Among the future studies, estimating student academic performance with different classification algorithms and learning these algorithms with automatic-machine learning methods of hyper-parameters can be shown. Again, analyzing trainer performance with data mining algorithms and techniques is among the future studies.

References

[1] Baker, R.S, Big Data and Education. New York: Teachers College, Columbia University, (2015).
[2] Romero, C. and Ventura, S., Educational Data Mining and Learning Analytics: an Updated Survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3) (2020) E1355.
[3] Aldowah, H., Al-Samarraie, H. and Fauzy, W.M., Educational Data Mining and Learning Analytics for 21st Century Higher Education: A Review and Synthesis. Telematics and Informatics, 37 (2019)13-49.
[4] Sutha, K., & Tamilselvi, J. J. A Review of Feature Selection Algorithms for Data Mining Techniques. International Journal on Computer Science and Engineering, 7(6) (2015) 63.

[5] Patel, H., & Prajapati, P. International Journal of Computer Sciences and Engineering Open Access. Int. J. Comput. Sci. Eng, 6(10) (2018).

[6] Tsiakmaki, M., Kostopoulos, G., Kotsiantis, S. and Ragos, O., Transfer Learning From Deep Neural Networks for Predicting Student Performance. Applied Sciences, 10(6) (2020) 2145.

[7] Andrade, T.L.D., Rigo, S.J. and Barbosa, J.L.V., Active Methodology, Educational Data Mining and Learning Analytics: A Systematic Mapping Study. Informatics in Education, 20(2) (2021).

[8] Mangina, E. and Psyrra, G., Review of Learning Analytics and Educational Data Mining Applications. in Proceedings of EDULEARN21 Conference5 (2021) 6.

[9] Aggarwal, C.C., Data Mining: the Textbook New York: Springer. 1 (2015).

[10] Schmidhuber, J., Deep Learning in Neural Networks: an Overview. Neural Networks, 61 (2015) 85-117.

[11] Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P. and Gao, R.X., Deep Learning and Its Applications to Machine Health Monitoring. Mechanical Systems and Signal Processing, 115 (2019) 213-237.

[12] Yang, J., Zhang, X.L. and Su, P., Deep-Learning-Based Agile Teaching Framework of Software Development Courses in Computer Science Education. Procedia Computer Science, 154 (2019) 137-145.

[13] Kastrati, Z., Dalipi, F., Imran, A.S., Pireva Nuci, K. and Wani, MAMA, Sentiment Analysis of Students' Feedback With NLP and Deep Learning: A Systematic Mapping Study. Applied Sciences, 11(9) (2021) 3986.

[14] Mamoun, A. and Alshanqiti, A., Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review. Applied Sciences, 11(1) (2021) 237.

[15] Ajibade, S.S.M., Ahmad, N.B. and Shamsuddin, S.M., December. A Data Mining Approach to Predict Students' Academic Performance Using Ensemble Techniques. in International Conference on Intelligent Systems Design and Applications. (2018) 749-760. Springer, Cham.

[16] Chaudhury, P., Mishra, S., Tripathy, HKHK and Kishore, B., March. Enhancing the Capabilities of the Student Result Prediction System. in Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies. (2020) 1-6.

[17] Salal, Y.K., Abdullaev, S.M. and Kumar, M., Educational Data Mining: Student Performance Prediction in Academic. International Journal of Engineering and Advanced Technology, 8(4C) (2019) 54-59.

[18] Hamoud, A., Selection of the Best Decision Tree Algorithm for Predicting and Classifying Students' Actions. American International Journal of Research in Science, Technology, Engineering & Mathematics, 16(1) (2020) 26-32.

[19] John M., Using Machine Learning to Predict Student Performance. Msc. Thesis, University of Tampere, Tampere, Finland, (2017).

[20] Başer S H, Hökelekli O, Kemal A., Estimation of Student Performance in Secondary Education With Data Mining Methods, Journal of Computer Science and Technologies, 1(1) (2020) 22-27.

[21] Ünal, F., Data Mining for Student Performance Prediction in Education. Data Mining-Methods, Applications and Systems.(2020).

[22] Athani, S.S., Kodi, S.A., Banavasi, M.N. and Hiremath, P.S., May. Student Academic Performance and Social Behaviour Predictor Using Data Mining Techniques. in 2017 International Conference on Computing, Communication and Automation (ICCCA). (2017) 170-174. IEEE.

[23] Ma, X. and Zhou, Z., Student Pass Rates Prediction Using Optimized Support Vector Machine and Decision Tree. in 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC). (2018) 209-215. IEEE.

[24] Troussas, C., Virvou, M. and Mesaretzidis, S., Comparative Analysis of Algorithms for Student Characteristics Classification Using A Methodological Framework. in 2020 6th International Conference on Information, Intelligence, Systems and Applications (IISA). (2020) 1-5. IEEE.

[25] Singh, M., Verma, C., Kumar, R. and Juneja, P., Towards Enthusiasm Prediction of Portuguese School's Students Towards Higher Education in Real-Time. in 2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM). (2020) 421-425. IEEE.

[26] Walia, N., Kumar, M., Nayar, N. and Mehta, G., Student's Academic Performance Prediction in Academic Using Data Mining Techniques. in Proceedings of the International Conference on Innovative Computing & Communications (ICICC).(2020).

[27] Srivastava, A.K., Chaudhary, A., Gautam, A., Singh, D.P. and Khan, R., Prediction of Students Performance Using KNN and Decision Tree-A Machine Learning Approach. Strad Research, 7(9) (2020) 119-125.

[28] Zaffar, M., Hashmani, M.A., Savita, K.S., Rizvi, S.S.H. and Rehman, M., Role of FCBF Feature Selection in Educational Data Mining. Mehran University Research Journal of Engineering & Technology, 39(4) (2020) 772-778.

[29] Xu, X., Wang, J., Peng, H. and Wu, R., Prediction of Academic Performance Associated With Internet Usage Behaviours Using Machine Learning Algorithms. Computers in Human Behavior, 98 (2019) 166-173.

[30] Student Performance Data Set, UCI Machine Learning Repository, Online Available At: Https://Archive.Ics.Uci.Edu/Ml/Datasets/Student+Performance

[31] Cortez, P., & Silva, A. M. G Using Data Mining to Predict Secondary School Student Performance, (2008).

[32] Farissi, A. and Dahlan, H.M., 2019, September. Genetic Algorithm-Based Feature Selection for Predicting Student's Academic Performance. in International Conference of Reliable Information and Communication Technology,. Springer, Cham. (2019) 110-117.

[33] Farissi, A. and Dahlan, H.M., Genetic Algorithm-Based Feature Selection With Ensemble Methods for Student Academic Performance Prediction. in Journal of Physics: Conference Series 1500(1) (2020) 012110. IOP Publishing.

[34] Shrestha, S. and Pokharel, M. Educational Data Mining in Moodle Data. International Journal of Informatics and Communication Technology (IJ-ICT), 10(1) (2021) 9-18.

[35] Injadat, M., Moubayed, A., Nassif, A.B. and Shami, A., 2020. Systematic Ensemble Model Selection Approach for Educational Data Mining. Knowledge-Based Systems, 200 (2020) 105992.

[36] Burman, I. and Som, S., February. Predicting Student's Academic Performance Using A Support Vector Machine. in 2019 Amity International Conference on Artificial Intelligence (AICAI) (2019) 756-759. IEEE.

[37] Mason, C., Twomey, J., Wright, D. and Whitman, L., Predicting Engineering Student Attrition Risk Using A Probabilistic Neural Network and Comparing Results With A Back Propagation Neural Network and Logistic Regression. Research in Higher Education, 59(3) (2018) 382-400.

[38] Deepika, K. and Sathyanarayana, N., Relief-F and Budget Tree Random Forest-Based Feature Selection for Student Academic Performance Prediction. International Journal of Intelligent Engineering and Systems, 12(1) (2019) 30-39.