BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

Journal of Computer Science Research
https://ojs.bilpublishing.com/index.php/jcsr

ARTICLE

Student Performance Prediction Using A Cascaded Bi-level Feature Selection Approach

Wokili Abdullahi[1] **Mary Ogbuka Kenneth**[1*] **Morufu Olalere**[2]

1. Department of Computer Science, Federal University of Technology, Minna, Nigeria
2. Department of Cyber Security Science, Federal University of Technology, Minna, Nigeria

ARTICLE INFO

Article history
Received: 10 August 2021
Accepted: 19 August 2021
Published Online: 21 August 2021

Keywords:
Relief
Particle swarm optimization
Cascaded bi-level
Educational data mining
Binary-level grading
Five-level grading

ABSTRACT

Features in educational data are ambiguous which leads to noisy features and curse of dimensionality problems. These problems are solved via feature selection. There are existing models for features selection. These models were created using either a single-level embedded, wrapper-based or filter-based methods. However single-level filter-based methods ignore feature dependencies and ignore the interaction with the classifier. The embedded and wrapper based feature selection methods interact with the classifier, but they can only select the optimal subset for a particular classifier. So their selected features may be worse for other classifiers. Hence this research proposes a robust Cascade Bi-Level (CBL) feature selection technique for student performance prediction that will minimize the limitations of using a single-level technique. The proposed CBL feature selection technique consists of the Relief technique at first-level and the Particle Swarm Optimization (PSO) at the second-level. The proposed technique was evaluated using the UCI student performance dataset. In comparison with the performance of the single-level feature selection technique the proposed technique achieved an accuracy of 94.94% which was better than the values achieved by the single-level PSO with an accuracy of 93.67% for the binary classification task. These results show that CBL can effectively predict student performance.

1. Introduction

The role of education in the development of any country cannot be over emphasised. This is because of its' impacts on the social, economic and political developments in any society [1]. The quality of any nation is directly proportional to the quality of her education system, hence, the ongoing efforts to improve the quality of educational institutions. Academic performance of students in any educational institution is a measure of the institutions efficiency in knowledge delivery [2].

The interest of researchers and scholars on learning outcomes have grown exponentially and this account for the reason why scholars have been working hard to find out factors that affects good academic performance [1]. There are different factors that affects students' performance. They include: intelligence, state of health, motivation, anxiety, suitable learning environment, adequate education infrastructures, family and parental influences, societal influences, institutional influences [3].

In Computer Science, one of the active fields is data mining. Data mining deals with the process of extracting

**Corresponding Author:*
Mary Ogbuka Kenneth,
Department of Computer Science, Federal University of Technology, Minna, Nigeria;
Email: Kenneth.pg918157@st.futminna.edu.ng

DOI: https://doi.org/10.30564/jcsr.v3i3.3534

valuable information from raw data [4]. Data mining is crucial due to the rising amount of data and the immediate need to translate these data into practical information. Presently, data mining technique is being applied to different sectors of life. The educational sector is a significant area in which data mining is gaining increasing interest. In educational field data mining is referred to as Educational Data Mining (EDM). EDM emphasizes that useful knowledge is obtained from educational information systems such as the course management systems, registration systems, online learning management systems, and application systems. Predicting the academic success of students is a significant application of EDM. In the educational environment, the analysis and estimation of student performance is an integral aspect. This prediction task foresees the importance of an unknown variable that distinguishes students with outcomes such as pass or failure, grades and marks [5].

EDM emerges as a result of rapid growth in educational data and this has presented several challenges to researchers to develop more efficient data mining methods [6]. In EDM the features in educational data are ambiguous which leads to the curse of dimensionality problem. This issue of curse of dimensionality and noisy features can be solved using dimensionality reduction. Dimensionality reduction can be achieved via feature selection. The purpose of feature selection is to select an appropriate subset of features which can efficiently describe the input data, which reduces the dimensionality of feature space and removes irrelevant data. There are existing models for selection of student performance features. However these models were created using either a single-level embedded, wrapper-based or filter-based methods. Filter methods are fast and independent of the classifier but ignore the feature dependencies and also ignores the interaction with the classifier [7]. Since both embedded and wrapper based feature selection methods interact with the classifier, they can only select the optimal subset for a particular classifier. So the features selected by them may be worse for other classifiers [8,9]. Filter-based method is fit for dealing with data that has large amounts of features since it has a good generalization ability. Given the importance of features and the relevance between features, the filter-based feature selection algorithm can only rank the features and cannot optimally select the subset of the selected features. Therefore, particle swarm optimization was used to optimally select the subset of the selected features after performing filter-based selection. Hence, this research work proposes a cascaded bi-level feature selection approach to overcome the drawbacks of a single filter-based and wrapper-based selection techniques for student performance prediction. The contributions of this study are:

(1) Development of a cascaded bi-level feature selection technique for student performance prediction.

(2) Selection of optimal features using the cascaded bi-level feature selection technique.

(3) Evaluation of performance of the cascaded bi-level feature selection technique.

The following is how the rest of the paper is structured: A summary of related researches on student performance prediction is included in section two. The techniques used to accomplish the goal of student performance prediction is presented in section three. In section four results from the experimentation are presented and discussed. The study's conclusion is presented in section five and lastly future works are presented in section six.

2. Related Works

Students' viability of progress is essential to predict student performance. The significance of predicting student performance has led researchers to become more and more interested in this field. Therefore, various researches have been published to predict students' performance.

In the study by Lau [10] ANN was used to evaluate and predict the students' CGPA using the data about their socio-economic background and entrance examination results of the undergraduate students from a Chinese university. In order to evaluate the performance of ANN, computations of Mean Square Error (MSE), regression analysis, error histogram and confusion matrix are introduced to ensure the appropriateness of ANN's performance in mitigating the arising of over fitting issues. Overall, the ANN has achieved a prediction accuracy of 84.8%, and with AUC value of 0.86. However the proposed Artificial Neural Network (ANN) method performs poorly in classifications of students according to their gender, as high False Negative rates are obtained as results, which is likely due to high imbalance ratio of two different types of sample.

Olalekan [11] adapted Bayes' theorem and ANN to construct a predictive model for students' graduation probability at a tertiary institution. Four variables were used for prediction: Unified Tertiary Matriculation Test, Number of Sessions at the high school level, Grade Points at the high school level and Entry Mode. The data used was collected from the Computer Science School, Federal Polytechnic, Ile-Oluji, in Ondo State, Nigeria. The data were composed of 44 examples with five attributes. The study concludes that the ANN has a 79.31% higher performance accuracy than the 77.14% obtained by the Bayes classification model. The ANN precision improved as the hidden layers increased. As compared to other previous works, the overall accuracy in this study was low because of the small size

 DOI: https://doi.org/10.30564/jcsr.v3i3.3534

of data used. Expanding the data size would help enhance the accuracy of the classification of the model.

Salal [12] presented a model for student performance classification based on the Eurostat Portuguese data set consisting of 33 attributes and 649 instances. Nine classifiers namely: ZeroR, Naïve Bayes, Random Forest, Random Tree, Decision Tree (J48), REPTree, Simple Logistic, JRip, and OneR were utilized in this study. In this study feature selection was performance using filter-based technique. All the nine classifiers had performance improvement when trained with the selected features. For instance the decision tree classifier with an accuracy of 67.79% when trained with all the feature attained 76.27% when trained with the selected features. This shows that student's attributes affect the student performance. The drawback of the proposed system is that filter-based feature selection techniques ignore the feature dependencies and also ignore the interaction with the classifier.

Magbag and Raga [13] focused on building a model to predict first-year students' academic success in tertiary education. This research aimed to allow early intervention to help students stay on course and reduce non-continuance. The data utilized in this paper were obtained from three higher education institutions in Central Luzon, primarily in the cities of Angeles, San Fernando and Olongapo. The study subjects included first-year students from 8 academic departments from 2018-2019; Arts and Sciences, Engineering and Architecture, Computer Studies, Criminology, Education, Hospitality and Tourism, Business and Accountancy, Nursing and Allied Medical Sciences. The dataset was composed of 4,762 examples. The dataset was pre-processed, and missing values were deleted, leaving 3,466 available samples. Using Correlation-based Feature Selection, Gain Ratio and Information Gain for feature rating, feature selection was carried out. Using these selected features, the NN and logistic regression models were trained and evaluated. In comparison with similar works, the scale of the dataset used rendered the scheme more robust. However, the accuracy of 76% achieved in this analysis is low.

Ünal [14] performed student performance prediction using feature selection. Decision tree, random forest and Naïve Bayes were employed on the educational datasets to predict the final grades of students. In this study two experiments were conducted. The first experiment deals with training the classifiers without feature selection. And the second experiment deals with training the classifiers after feature selection. In the second experiment wrapper feature selection technique was used to select the most relevant feature set, while the irrelevant features were removed. The second experimentation produced an improved accuracy due to the applied feature selection than the first experiment without feature selection. For instance the accuracy of Naïve Bayes improved from 67.80% in the first experiment to 74.88% in the second experiment. The EuroStat dataset from secondary education of two Portuguese schools were used. This issue with the feature selection technique used in this study is that they are classifier dependent. That is a set of features selected by a particular classifier and works well for that classifiers. Those not mean those set of features will also perform well for other classifiers/models.

3. Methodology

This section presents the research methods that were followed in the study. It provides information on the dataset, data encoding method, feature selection and data classification. The diagram presenting each of the method used for student performance prediction is presented in Figure 1.

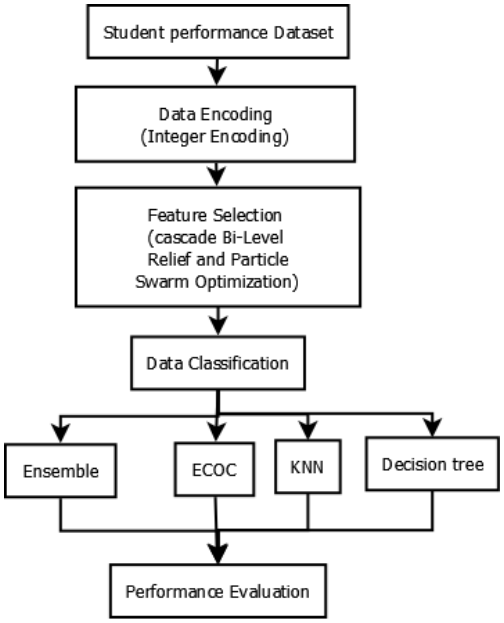

Figure 1. Proposed System

3.1 Dataset

This study's dataset is known as the student performance dataset. The data were obtained from the University of California at Irvine's repository. The student performance data were collected from two public schools in the Alentejo region of Portugal during the 2005-2006 academic year. The dataset consists of secondary school accomplishment statistics from two Portuguese schools. The data were acquired through school reports and surveys and include student grades, demographic, social, and school-related characteristics. Two datasets are provided, one for mathematics and the other for Portuguese lan-

DOI: https://doi.org/10.30564/jcsr.v3i3.3534

guage performance. The mathematics data set consist of 395 instances and 33 attributes where 32 attributes are the predictors while one attribute (attribute 33) is the target. The Portuguese data set consist of 649 instances with 33 attributes. In Cortez and Silva [15], the two datasets were modeled under binary and five-level classification tasks. These two classification task is explained in section 3.2.

3.2 Data Preprocessing

The overall assessment in the original data, as in several other countries, is on a scale of 0-20, with 0 being the worst and 20 being the best. Because the students' final score is in the form of integers, and the predicted class should be in the form of categorical values, the data had to be transformed to categories according to a scoring policy. Two separate grading systems were employed in this study: binary grading and five-level grading. First, the final grade was divided into five categories. These ranges are described using the Erasmus framework. The scale 0-9 equates to grade F, which is the lowest grade and corresponds to the mark "fail". The remaining class labels (10-11, 12-13, 14-15, and 16-20) correspond to D (sufficient), C (satisfactory), B (good), and A (excellent) respectively. Secondly, the final grade was categorized into two (binary) categories: fail and pass. Table 1 shows the five-level grading categories. The binary-level grading categories are shown in Table 2. In Table 2, the range of 0-9 corresponds to F, and it means "fail"; the range of 10-20 refers to A, B, C, and D, and it means "pass."

Table 1. Five-level grading categories

1	2	3	4	5
Excellent	Good	Satisfactory	Sufficient	Fail
16-20	14-15	12-13	10-11	0-9
A	B	C	D	F

Table 2. Binary-level grading categories

0	1
Fail	Pass
0-9	10-20
F	A, B, C, D

3.3 Data Encoding

There are both numeric variables and categorical variables in the dataset used. Categorical variables are usually represented as 'strings' or 'categories' and are finite in number. In this phase, the categorical data types of attributes were converted to numeric attributes. Data encoding was done because specific machine learning algorithms such as Naïve Bayes, support vector machine and Ensemble need numeric attribute types to work. The integer encoding technique was employed in this research. Integer encoding involves mapping each string attribute to an integer value. Integer encoding was used for the categorical (string) class because the integer values have a natural ordered relationship between each other and machine learning algorithms may be able to understand and harness this relationship. And the categorical attributes has an order relationship [16]. Table 3 indicate gender representation after integer encoding.

Table 3. Integer encoding for Gender Attribute

Male	1
Female	2

3.4 Feature Selection

Feature selection is a process that selects pertinent features as a subset of original features [17]. In real-world situations, relevant features are often unknown a priori. Hence feature selection is a must to identify and remove the irrelevant/redundant features for student performance prediction [18]. This paper proposed a novel cascade bi-level feature selection approach for the classification of student performance data, which used filtering technique such as Relief (RF) and optimization technique such as Particle swarm Optimization (PSO).

The proposed method is divided into two levels. In the first level (level 1) the Relief technique was used to select 20 sets of features based their shared information. The selected 20 sets of features were gathered and a new feature subset is generated. In the second level (level 2) the new 20 feature subset is used as input to the PSO and an optimized feature subset is selected. The proposed feature selection scheme is presented in Figure 2.

3.4.1 Relief Feature Selection Technique

Relief is a feature selection algorithm that uses a filter-method approach that is particularly sensitive to feature interactions.

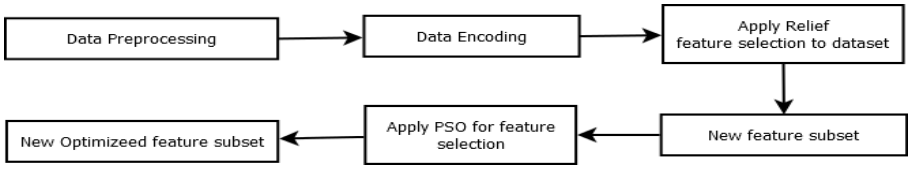

Figure 2. Proposed feature selection technique

 DOI: https://doi.org/10.30564/jcsr.v3i3.3534

Relief calculates a proxy statistic for each feature that can be used to measure feature 'quality' or 'relevance' to the target definition [19]. These feature statistics are known as feature weights (W [A] = weight of feature 'A') or feature 'scores,' and they can vary from -1 (worst) to +1 (best).

The advantages of using the Relief method is that it is computational fast even when there is a big amount of data. Time complexity is not a problem because a consistent number of trials is completed. As a result, the relief technique may complete faster than other filter-based approaches that require all of the data to be considered [20].

Given the importance of features and the relevance between features, Relief filter-based selection feature algorithm selects relevant feature based on their relationship with dependent variable however in Relief method the interaction with the classifier and each feature is considered independently thus ignoring feature dependencies. To enable feature dependencies and classifier interaction, the particle swarm optimization is used to optimally select a subset from the selected features.

3.4.2 Particle Swarm Optimization (PSO)

Particle swarm optimization (PSO) is a computational technique for solving problems by iteratively trying to enhance a candidate solution in terms of a quality metric [21]. It solves a problem by generating a population of possible solutions, which are referred to as particles, and moving them around in the search space using a simple mathematical formula based on their position and velocity [22]. Consider the global optimum of m-dimensional function defined in Equation 1.

$$G(x_1, x_2, x_3, x_4, ..., x_m) = G(X) \quad (1)$$

Where x_i is the search variable, which represents the set of free variables of the given function. The aim is to find a value x^* such that the function $G(x^*)$ is either a maximum or a minimum in the search space. The PSO algorithm is a multi-agent concurrent search method in which each particle represents a potential solution in the swarm. All particles go through a multidimensional search space, where each particle adjusts its position based on its own and neighbouring experiences (Poli et al., 2007 Suppose x_i^s denotes the position vector of particle I in the multidimensional search space at time step s, then Equation 2 is used to update the position of each particle in the search space.

$$x_i^{s+1} = x_i^s + v_i^{s+1} \ \ with\ x_i^0 \sim U(x_{min}, x_{max}) \quad (2)$$

Where v_i^s is the velocity vector of particle *I* that drives the optimization process and reflects both the own experience knowledge and the social experience knowledge from the all particles. $U(x_{min}, x_{max})$ is the uniform distribution where x_{min} and x_{max} are its minimum and maximum values respectively.

The velocity of the particle *i* updated using Equation 3.

$$v_i^{s+1} = w * v_i^s + c_1 * r_{1i} * (p_i - x_i^s) + c_2 * r_{2i} * \left(p_g - x_i^s\right) \quad (3)$$

Where s denotes the sth iteration in the process, w is inertia weight and c_1 and c_2 are acceleration constants. r_{1i} and r_{2i} are random values uniformly distributed in [0,1]. p_i and p_g represent the elements of $pbest$ and $gbest$ respectively. The flowchart for selection of features using PSO is shown in Figure 3.

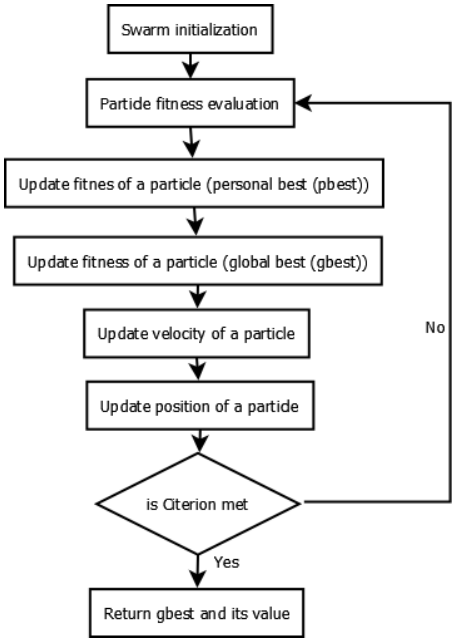

Figure 3. Flow chat for PSO feature selection

Therefore, in a PSO method, all particles are initiated randomly and evaluated to compute fitness together with finding the personal best (best value of each particle) and global best (best value of particle in the entire swarm). After that a loop starts to find an optimum solution. In the loop, first the particles' velocity is updated by the personal and global bests, and then each particle's position is updated by the current velocity. The loop is ended with a stopping criterion predetermined in advance [23].

3.5 Data Classification

In this research four machine learning classification models were used for training and classification: Error-Correcting Output Codes (ECOC), Decision Tree (DT), Ensemble, and K-Nearest Neighbor (KNN).

3.5.1 Error-Correcting Output Codes (ECOC)

DOI: https://doi.org/10.30564/jcsr.v3i3.3534

The ECOC technique is a tool that allows the issue of multiclass classification to be interpreted as multiple problems of binary type, enabling the direct use of native binary classification models [24]. ECOC designs are independent of the classifier depending on the implementation. ECOC has error-correcting properties and has shown that the learning algorithm's bias and variance can be decreased [25].

3.5.2 Decision Tree (DT)

A decision tree is a supervised learning model in which data is continually separated based on a specific parameter. The decision tree employs a tree-like structure to progress from observations about an item (represented by the branches) to inferences about the item's target value (defined in the leaves) (Kolo *et al.*, 2015). Entropy is a popular technique used in determining which attribute to position at the root or the different levels of the tree [26]. Entropy is a measure of randomness in processed information [26]. The larger the entropy, the more challenging it is to draw any conclusions from that data. A branch with an entropy of zero, for example, is chosen as the root node, and further division is required for a branch with an entropy greater than zero [27]a novel concept of a non-probabilistic novelty detection measure, based on a multi-scale quantification of unusually large learning efforts of machine learning systems, was introduced as learning entropy (LE. In Equation 4, entropy for a single attribute is expressed.

$$E(S) = \sum_{i=1}^{n} - p_i log_2 p_i \quad (4)$$

Where S represents the present state, $p_i$is the probability of an event i of state S.

3.5.3 K-Nearest Neighour (KNN)

This is among the simplest machine learning model [28]. An item is classified based on its "distance" from its neighbours, and it is allocated to the most common class of its k closest neighbours [29,30]. The Euclidean distance is a linear distance between two points in Euclidean space [31]. If two vectors x_i and x_j are given where $x_i = (x_{i1}, x_{i2}, ..., x_{in})$ and$x_j = (x_{j1}, x_{j2}, ..., x_{jn})$, Then the Euclidean distance between x_i andx_j is given in Equation 5:

$$ED(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \quad (5)$$

3.5.4 Ensemble Classifier

An ensemble learning model combines predictions from multiple models with a two-fold goal: the first objective is to maximize prediction accuracy compared to a single classifier [32]. The second gain is more critical generalizability due to multiple advanced classifiers. As a result, solutions, where a single prediction model would have problems, can be discovered by an ensemble. A key rationale is that an ensemble can select a set of hypotheses out of a much larger hypothesis space and combine their predictions into one [33]. Via voting or weighted voting of their forecast for the final estimates, classifiers in the ensemble learning model are merged into meta-classifiers [34].

3.6 Performance Metrics

In this study, the accuracy, precision, recall, and f-score performance measures were used to evaluate the proposed method. This measure is explained below.

3.6.1 Accuracy

The rate of correct classifications is used to define accuracy. This is the number of correct guesses divided by the total number of right forecasts. The exact formula is given in Equation 6:

Accuracy=

$$\frac{\text{True Positive} + \text{True negative}}{\text{True Positive} + \text{True negative} + \text{False Positive} + \text{False negative}} \quad (6)$$

3.6.2 Precision

Precision is a metric used to calculate how many positive predictions are accurately made. The number of true positive elements is derived by dividing the total number of true positives by the total number of false positives. The formula in equation is used to define precision 7:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (7)$$

3.6.3 Recall

Sensitivity is another term for recall. The amount of correct positive predictions that could have been made from all positive predictions is calculated by recall. The recall is calculated using the formula in Equation 8.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (8)$$

3.6.4 F-Score

The f-score of a model is defined as the harmonic average of recall and precision. F-Score is represented in Equation 9.

$$\text{F} - \text{Score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (9)$$

 DOI: https://doi.org/10.30564/jcsr.v3i3.3534

4. Results and Discussion

In this work two classification tasks were carried out. These classification tasks are: binary-level grading classification and the five-level grading classification task. The Mathematics and the Portuguese dataset was used. The mathematics data set consist of 395 instances and 33 attributes while the Portuguese data set consist of 649 instances with 33 attributes. The datasets were divided in the ratio of 4:1 for training and testing (80% for training and 20% for testing).

4.1 Binary-level Grading Classification

The binary classification deals with classification using the two classes which are pass and fail. The dataset original label consist of 0-20 labels or grades. Where 0 is the worst grade and 20 is the best score. In the binary classification the integer labels were categorized into two classes where Fail (0) represents grade 0-9 and Pass (1) represents grade 10-20. Using the binary labels the four classifiers (ECOC, Ensemble, KNN and Decision Tree) were trained and tested using the original features (no feature selection done), sub-features using relief feature selector, sub-features using PSO feature selector and sub-features using the cascade bi-level feature selector. Table 4 presents the accuracy, precision, recall and f-score of all the four classifiers when trained with the original 32 features.

From Table 4 it can be seen that ensemble classifier performed best for Mathematics dataset with accuracies of 91.14%, f-score of 85.52%, precision of 77.38% and recall of 86.10% when compared with ECOC, KNN and DT. While KNN performed least with an accuracy of 70.89%, f-score of 54.90%, precision of 43.75% and recall of 73.68%. For Portuguese dataset ECOC and Ensemble achieved same accuracy of 81.25% which presents them as the best performer. Table 5 presents the accuracy, precision, recall and f-score of all the four classifiers when trained with the Relief selected sub-features.

To properly evaluate the performance of ECOC, Ensemble, DT and KNN classifiers when trained with Relief selected feature subsets for both Mathematics and Portuguese data set, their precision, recall, f-score and accuracy are presented in Table 5. The ensemble classifier performed best with an accuracy of 92.41% and f-score of 86.35% for Mathematics dataset. While ECOC performed best with an accuracy of 93.08% and f-score of 82.35% for Portuguese dataset. Table 6 presents the accuracy, precision, recall and f-score of all the four classifiers when trained with the PSO selected sub-features.

In Table 6 Ensemble classifier obtained the best performance for both Mathematics and Portuguese data sets with an accuracy of 93.67% and an f-score of 87.90% for Mathematics data set and an accuracy of 94.62% and an f-score of 82.05% for Portuguese data set. ECOC, KNN and DT performance equally when trained with PSO selected Mathematics sub-features. However for the Portuguese data set KNN performed least with an accuracy of 91.42% and f-score of 70.59%.

A comparison of the classification results of the four classifiers for the original feature sets, relief feature set, PSO selected features and the cascade bi-level feature sets are shown Table 7.

In Table 7 an accuracy of 91.14% was obtained for Mathematics dataset classification using the original

Table 4. Binary-level classification results before feature selection

Before Feature Selection Classification Results								
	Mathematics				Portuguese			
Classifiers	**Accuracy (%)**	**F-Score (%)**	**Precision (%)**	**Recall (%)**	**Accuracy (%)**	**F-Score (%)**	**Precision (%)**	**Recall (%)**
ECOC	89.87	83.24	88.75	78.33	92.31	81.25	79.67	83.10
Ensemble	91.14	85.52	77.38	86.10	92.31	81.25	79.67	83.10
KNN	70.89	54.90	43.75	73.68	89.23	77.55	73.08	82.61
DT	87.34	77.76	70.00	87.31	91.54	77.06	80.51	84.35

Table 5. Binary-level Classification Results for Relief Selected Features

Relief Selected Features Classification Results								
	Mathematics				Portuguese			
Classifiers	**Accuracy (%)**	**F-Score (%)**	**Precision (%)**	**Recall (%)**	**Accuracy (%)**	**F-Score (%)**	**Precision (%)**	**Recall (%)**
ECOC	91.14	85.71	87.50	84.00	93.08	82.35	80.77	84.00
Ensemble	92.41	86.35	79.71	95.00	93.08	81.63	76.92	86.96
KNN	79.75	66.67	66.67	66.67	91.54	77.55	73.08	82.61
DT	89.87	82.61	79.17	86.36	92.31	79.17	82.61	86.36

DOI: https://doi.org/10.30564/jcsr.v3i3.3534

Table 6. Binary-level Classification for PSO Selected Features

PSO Selected Features Classification Results								
	Mathematics				Portuguese			
Classifiers	**Accuracy (%)**	**F-Score (%)**	**Precision (%)**	**Recall (%)**	**Accuracy (%)**	**F-Score (%)**	**Precision (%)**	**Recall (%)**
ECOC	92.41	84.21	76.19	94.12	93.85	78.95	75.00	83.33
Ensemble	93.67	87.90	85.71	90.00	94.62	82.05	80.00	84.21
KNN	92.41	84.21	76.19	94.12	91.42	70.59	60.00	85.71
DT	92.41	84.21	76.19	94.12	93.85	78.95	75.00	83.33

Table 7. Comparison of Feature Selection Techniques for Binary-level Classification Task

Accuracy (%)								
	Mathematics				Portuguese			
Feature Selection	**ECOC**	**Ensemble**	**KNN**	**DT**	**ECOC**	**Ensemble**	**KNN**	**DT**
Before Feature Selection	89.87	91.14	70.89	87.34	92.31	92.31	89.23	91.54
Relief	91.14	92.41	79.75	89.87	93.08	93.08	91.54	92.31
PSO	92.41	93.67	92.41	92.41	93.85	94.62	91.42	93.85
Cascaded Bi-level	93.67	94.94	92.89	92.89	95.38	96.15	93.85	93.85

feature sets. While the original Portuguese dataset obtained the highest accuracy of 92.31% when trained with ECOC and Ensemble classifier. The highest classification accuracy obtained using the Relief selected features for the Mathematics data set is 92.41%. The PSO selected sub-features obtained a classification accuracy of up to 93.67% for the Mathematics data set. The proposed cascaded bi-level obtained a classification accuracy of 94.94% for the Mathematics data set. For the Portuguese data set the highest classification accuracy obtained for classification using the Relief selected features is 93.08% by Ensemble and ECOC classifiers. The PSO selected sub-features obtained a classification accuracy of up to 94.62% for the Portuguese data set. The proposed cascaded bi-level obtained a classification accuracy of 96.15% for the Portuguese data set. In conclusion the proposed technique selected the best sub-features that achieved a higher classification accuracy than the sub-features selected by a single-level relief or PSO selector.

The selected features by Relief, PSO and Cascaded Bi-level feature selectors used for training and testing of the four models are presented in Table 8.

From Table 8, for the Mathematics dataset Relief selector selected 20 feature sets from the 32 original feature sets, PSO selected 16 features and cascaded bi-level selector selected 11 features from the original feature sets. For the Portuguese dataset Relief selector selected 20 feature sets from the 32 original feature sets, PSO selected 14 features and cascaded bi-level selector selected 8 features from the original feature sets. From the selected features in Table 5 it can be seen that the G1 and G2 features were selected by all the feature selectors. This shows that first period grade (G1) and the second period grade (G2) are relevant for the final grade prediction.

Table 8. Selected feature sets by Relief, PSO and Cascaded bi-level feature selectors for binary-level grading

Selected Feature sets		
	MATHEMATICS	**PORTUGUESE**
Relief	G2, G1, Sex, Paid, Failures, Activities, Romantic, Famsup, Stidytime, Higher, Mjob, Pstatus, Dalc, Medu, Guardian, Goout, Walc, Absences, Age, School	School, G2, G1, Activities, Sex, Address, Famsup, Failures, Nursery, Reason, Romantic, Higher, Medu, Famrel, Schoolsup, Fedu, Internet, Goout, Studytime, Health
PSO	School, Age , Famsize, Medu, Fjob, Guardian, Failures, Famsup, Paid, Activities, Nursery, Internet, Romantic, Freetime, G1, G2	Age, Address, Famsize, Fjob, Reason, Traveltime, Studytime, Failures, Famsup, Paid, Freetime, Goout, G1, G2
Cascaded Bi-Level	G2, G1, Sex, Activities, Famsup, Studytime, Mjob, Medu, Guardian, Goout, Walc	G2, G1,Nursery, Reason, Romantic, Higher, Schoolsup, Goout

Table 9 presents a comparison of the performance of the proposed technique with related work that used the student performance dataset from UCI repository with respect to binary classification. The results obtained showed

 DOI: https://doi.org/10.30564/jcsr.v3i3.3534

that the proposed technique achieved a higher student prediction accuracy than related work.

Table 9. Comparison of Binary Classification Performance with Related Work

Techniques	Mathematics	Portuguese
	Highest Obtained Accuracy (%)	Highest Obtained Accuracy (%)
Ünal [14]	93.67	93.22
Shah [35]	93.80	
Cascaded Bi-level	**94.94**	**96.15**

4.2 Five-Level Grading Classification

The five-level grading classification deals with classification using the five classes which are excellent (5), good (4), satisfactory (3), sufficient (2) and fail (1). The original label of 0-20 labels or grades were categorized into the aforementioned five classes. Using the five-level grading the four classifiers (ECOC, Ensemble, KNN and Decision Tree) were trained and tested using the original features (no feature selection done), sub-features using relief feature selector, sub-features using PSO feature selector and sub-features using a cascade bi-level feature selector. The five-level grading classification result is shown in Table 6.

From Table 10 it can be seen that ensemble classifiers performed best for both Mathematics and Portuguese dataset with an accuracies of 72.68% and 80.05% respectively. DT also performed least for both Mathematics and Portuguese dataset with accuracies of 64.56%, and 76.92% respectively. Table 11 presents the accuracy, precision, recall and f-score of all the four classifiers when trained with the Relief selected sub-features.

To properly evaluate the performance of ECOC, Ensemble, DT and KNN classifiers when trained with Relief selected feature subsets for both Mathematics and Portuguese data set for the five-level grading version, their precision, recall, f-score and accuracy are presented in Table 11. The ensemble classifier performed best with an accuracy of 79.75% and f-score of 91.80% for Mathematics dataset. While ECOC performed best with an accuracy of 93.08% and f-score of 82.35% for Portuguese dataset.

Table 12 is classification results of ECOC, Ensemble, KNN and DT when trained with PSO selected feature sets. In Table 12 Ensemble classifier obtained the best performance for both Mathematics and Portuguese data

Table 10. Five-level classification results before feature selection

Before Feature Selection Classification Results								
	Mathematics				Portuguese			
Classifiers	Accuracy (%)	F-Score (%)	Precision (%)	Recall (%)	Accuracy (%)	F-Score (%)	Precision (%)	Recall (%)
ECOC	72.42	77.73	85.00	71.67	72.31	80.00	72.73	88.89
Ensemble	72.68	78.33	86.01	72.98	74.62	80.05	70.73	92.12
KNN	69.62	73.33	73.33	73.33	70.77	76.92	68.18	88.24
DT	64.56	68.45	68.45	68.45	66.15	78.82	78.82	78.82

Table 11. Five-level Classification for Relief Selected Features

Relief Selected Features Classification Results								
	Mathematics				Portuguese			
Classifiers	Accuracy (%)	F-Score (%)	Precision (%)	Recall (%)	Accuracy (%)	F-Score (%)	Precision (%)	Recall (%)
ECOC	75.95	88.52	84.38	93.10	93.08	82.35	80.77	84.00
Ensemble	79.75	91.80	87.50	95.25	93.08	81.63	76.92	86.96
KNN	75.95	87.50	87.50	87.50	91.54	77.55	73.08	82.61
DT	70.89	85.25	81.25	89.66	92.31	79.17	82.61	86.36

Table 12. Five-level Classification for PSO Selected Features

PSO Selected Features Classification Results								
	Mathematics				Portuguese			
Classifiers	Accuracy (%)	F-Score (%)	Precision (%)	Recall (%)	Accuracy (%)	F-Score (%)	Precision (%)	Recall (%)
ECOC	75.95	78.26	75.00	81.82	77.69	77.78	77.78	77.78
Ensemble	78.48	80.85	79.17	82.26	78.64	78.28	69.23	90.00
KNN	74.68	80.00	75.00	85.71	76.92	76.92	76.92	76.92
DT	72.15	78.26	75.00	81.82	71.77	62.52	76.92	52.63

DOI: https://doi.org/10.30564/jcsr.v3i3.3534

sets with an accuracy of 78.26% and an f-score of 80.85% for Mathematics data set and an accuracy of 78.64% and an f-score of 78.28% for Portuguese data set. For both Mathematics and Portuguese data set DT performed least with an accuracy of 72.15% and f-score of 78.26% for Mathematics data set and accuracy of 71.77% and f-score of 62.52% for Portuguese data set.

In Table 13 the Ensemble classifier produced the best performance for both the Mathematics and Portuguese data sets. Ensemble classifier got an accuracy of 84.81%, f-score of 92.31%, precision of 93.75% and recall of 90.91% for Mathematics dataset. For the Portuguese data set the Ensemble classifier obtained an accuracy of 83.85%, f-score of 87.50%, precision of 77.78% and recall of 100%. Table 14 is a comparison of the performance based on accuracy of the Relief, PSO and Cascaded bi-level feature selection techniques.

In Table 14 the highest classification accuracy which was obtained by Ensemble classifier using the Relief selected features for the Mathematics data set is 79.75%. The PSO selected sub-features obtained a classification accuracy of up to 78.48% from Ensemble classifier for the Mathematics data set. The proposed cascaded bi-level obtained the highest accuracy of 84.81% when compared with Relief and PSO performance for the Mathematics data set. For the Portuguese data set the highest classification accuracy obtained for classification using the Relief selected features is 76.92% by Ensemble classifier. The PSO selected sub-features obtained a classification accuracy of up to 78.64% for the Portuguese data set using the Ensemble classifier. The proposed cascaded bi-level obtained highest accuracy of 83.85% when compared with Relief and PSO performance for the Portuguese data set. Training with the original complete 32 feature sets obtained the least accuracy as compared with training with the selected Relief, PSO and Cascaded bi-level feature sets. In conclusion the proposed technique selected the best sub-features that achieved a higher classification accuracy than the sub-features selected by a single-level relief or PSO selector.

From Table 15, for the Mathematics dataset Relief feature selector selected 20 feature sets from the 32 original feature sets, PSO selected 16 features and cascaded bi-level selector selected 10 features from the original feature sets. For the Portuguese dataset Relief selector selected 20 feature sets from the 32 original feature sets, PSO selected 13 features and cascaded bi-level selector selected 6 features from the original feature sets. From the selected features in Table 8 it can be seen that the G1 and G2 features were selected by all the feature selectors. This shows that first period grade (G1) and the second period grade (G2) is relevant for the final grade prediction for the five-level grading as it is important in the binary-level grading classification task.

Table 13. Five-level Classification for Cascaded Bi-level Selected Features

Cascaded Bi-level Selected Features Classification Results								
Classifiers	**Mathematics**				**Portuguese**			
	Accuracy (%)	**F-Score (%)**	**Precision (%)**	**Recall (%)**	**Accuracy (%)**	**F-Score (%)**	**Precision (%)**	**Recall (%)**
ECOC	83.54	90.00	84.38	96.43	83.08	88.24	83.33	93.75
Ensemble	84.81	92.31	93.75	90.91	83.85	87.50	77.78	100
KNN	81.01	88.14	81.25	96.30	77.38	74.29	72.22	76.47
DT	73.42	81.97	78.13	86.21	72.67	76.19	88.89	66.77

Table 14. Comparison of Feature Selection Techniques for Five-level Classification Task

Accuracy (%)								
	Mathematics				**Portuguese**			
Feature Selection	**ECOC**	**Ensemble**	**KNN**	**DT**	**ECOC**	**Ensemble**	**KNN**	**DT**
Before Feature Selection	72.42	72.68	69.62	64.56	72.31	74.62	70.77	66.15
Relief	75.95	79.75	75.95	70.89	76.15	76.92	74.62	72.31
PSO	75.95	78.48	74.68	72.15	77.69	78.64	76.92	71.77
Cascaded Bi-level	83.54	84.81	81.01	73.42	83.08	83.85	77.38	72.67

 DOI: https://doi.org/10.30564/jcsr.v3i3.3534

Table 15. Selected feature sets by Relief, PSO and Cascaded bi-level feature selectors for Five-level grading

	Selected Feature sets	
	MATHEMATICS	**PORTUGUESE**
Relief	G2, G1, Sex, Medu, Walc, Studytime, Address, Paid, Schoolsup, Mjob, Failures, Higher, Pstatus, Dalc, school, Freetime, Age, Famsup, Internet, Absences	G2, G1, School, Activities, sex, Studytime, Higher, Medu, Failures, Schoolsup, Nursery, Health, Famsup, Goout, Pstatus, Address, Fedu, Internet, Reason, Walc
PSO	Sex, Age, Famsize, Medu, Failures, Schoolsup, Famsup, Paid, Activities, Nursery, Internet, Romantic, Famrel, Freetime, G1, G2	Sex, Medu, Failures, Schoolsup, Paid, Activities, Internet, Famrel, Freetime, Goout, Health, G2, G1
Cascaded Bi-level	G2, G1, Walc, Studytime, Address, Paid, Schoolsup, Failures, Dalc, Internet	G2, G1, sex, Famsup, Pstatus, Address

Table 16 presents a comparison of the performance of the proposed technique with related works that used the Student performance dataset from UCI repository with respect to five-level grading classification. The results obtained showed that the proposed technique achieved a higher student prediction accuracy than related works based on Portuguese and Mathematics data set.

Table 16. Comparison of Five-Level Grading Performance with Related Work

	Mathematics	**Portuguese**
Techniques	**Highest Obtained Accuracy (%)**	**Highest Obtained Accuracy (%)**
Salal [12]		76.73
Ünal [14]	79.49	77.20
Proposed Technique	**84.81**	**83.85**

5. Conclusions - Future Works

This study developed a cascade bi-level feature selection technique for predicting students' academic performance. The Cascade bi-level feature selection technique achieved using Relief filter-based algorithm and Particle Swarm Optimization (PSO) algorithm. First the relief algorithm was used to select features based on their relevance to the target class. This selected features were fed as input to the PSO. The PSO then optimally selects the subset of the selected features based on the particle fitness. The Relief, PSO, and the Cascade bi-level selected features were analyzed using Error-Correcting Output Code (ECOC), ensemble, Decision Tree and K-Nearest Neighbour (KNN) machine learning models. The cascaded bi-level feature selection technique was evaluated against single-level feature selection techniques and against related works. The accuracy performance metric was used to perform this assessment. The proposed cascaded bi-level feature selection technique obtained an accuracy of 94.94% for Mathematics data set and 96.15% for Portuguese data set using the binary-level grading version of the data set. The cascaded bi-level feature selection technique also obtained an accuracy 84.81% for Mathematics data set and 83.85% for Portuguese data set using the five-level grading version of the data set. The results indicate the effectiveness of the cascaded bi-level feature selection technique in achieving an improved student performance prediction as it selects the best sub-features.

This study utilized Relief a filter-based technique and Particle swarm optimization a wrapper technique for feature selection. For future work other filter and wrapper-based feature selection techniques can be utilized, which can provide an insight on which filter and/or wrapper-based selection techniques produces better results when combined. In this study, the bi-level selection approach was considered. It is recommended that further research should explore multiple-level techniques for feature selection.

References

[1] J. M. Adán-Coello and C. M. Tobar, 'Using Collaborative Filtering Algorithms for Predicting Student Performance', in *Electronic Government and the Information Systems Perspective*, vol. 9831, A. Kő and E. Francesconi, Eds. Cham: Springer International Publishing, 2016, pp. 206-218. DOI: 10.1007/978-3-319-44159-7_15.

[2] E. Jembere, R. Rawatlal, and A. W. Pillay, 'Matrix Factorisation for Predicting Student Performance', in *2017 7th World Engineering Education Forum (WEEF)*, Kuala Lumpur, Nov. 2017, pp. 513-518. DOI: 10.1109/WEEF.2017.8467150.

[3] K. David Kolo, S. A. Adepoju, and J. Kolo Alhassan, 'A Decision Tree Approach for Predicting Students Academic Performance', *Int. J. Educ. Manag. Eng.*, vol. 5, no. 5, pp. 12-19, Oct. 2015. DOI: 10.5815/ijeme.2015.05.02.

[4] S. Hussain, N. Abdulaziz Dahan, F. M. Ba-Alwi, and N. Ribata, 'Educational Data Mining and Analysis

DOI: https://doi.org/10.30564/jcsr.v3i3.3534

of Students' Academic Performance Using WEKA', *Indones. J. Electr. Eng. Comput. Sci.*, vol. 9, no. 2, p. 447, Feb. 2018.
DOI: 10.11591/ijeecs.v9.i2.pp447-459.

[5] M. Imran, S. Latif, D. Mehmood, and M. S. Shah, 'Student Academic Performance Prediction using Supervised Learning Techniques', *Int. J. Emerg. Technol. Learn. IJET*, vol. 14, no. 14, p. 92, Jul. 2019.
DOI: 10.3991/ijet.v14i14.10310.

[6] M. A. Amoo, O. B. Alaba, and O. L. Usman, 'Predictive modelling and analysis of academic performance of secondary school students: Artificial Neural Network approach', *Int. J. Sci. Technol. Educ. Res.*, vol. 9, no. 1, pp. 1-8, May 2018.
DOI: 10.5897/IJSTER2017.0415.

[7] Z. M. Hira and D. F. Gillies, 'A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data', *Adv. Bioinforma.*, vol. 2015, pp. 1-13, Jun. 2015.
DOI: 10.1155/2015/198363.

[8] A. Daud, N. R. Aljohani, R. A. Abbasi, M. D. Lytras, F. Abbas, and J. S. Alowibdi, 'Predicting Student Performance using Advanced Learning Analytics', in *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, Perth, Australia, 2017, pp. 415-421.
DOI: 10.1145/3041021.3054164.

[9] B. K. Francis and S. S. Babu, 'Predicting Academic Performance of Students Using a Hybrid Data Mining Approach', *J. Med. Syst.*, vol. 43, no. 6, p. 162, Jun. 2019.
DOI: 10.1007/s10916-019-1295-4.

[10] E. T. Lau, L. Sun, and Q. Yang, 'Modelling, prediction and classification of student academic performance using artificial neural networks', *SN Appl. Sci.*, vol. 1, no. 9, p. 982, Sep. 2019.
DOI: 10.1007/s42452-019-0884-7.

[11] A. M. Olalekan, O. S. Egwuche, and S. O. Olatunji, 'Performance Evaluation Of Machine Learning Techniques For Prediction Of Graduating Students In Tertiary Institution', in *2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS)*, Ayobo, Ipaja, Lagos, Nigeria, Mar. 2020, pp. 1-7.
DOI: 10.1109/ICMCECS47690.2020.240888.

[12] Y. K. Salal, S. M. Abdullaev, and M. Kumar, 'Educational Data Mining: Student Performance Prediction in Academic', vol. 8, no. 4, p. 6, 2019.

[13] A. Magbag and R. R. Jr, 'Prediction Of College Academic Performance Of Senior High School Graduates Using Classification Techniques', vol. 9, no. 04, p. 6, 2020.

[14] F. Ünal, 'Data Mining for Student Performance Prediction in Education', *IntechOpen*, p. 12, 2020.
DOI: http://dx.doi.org/10.5772/intechopen.91449.

[15] P. Cortez and A. Silva, 'Using data mining to Predict Secondary School Student Performance', p. 9, 2008.

[16] C. Seger, 'An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing', Bachelor Degree, KTH ROYAL INSTITUTE OF TECHNOLOGY SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE, Sweden, 2018.

[17] J. Wang, S. Zhou, Y. Yi, and J. Kong, 'An Improved Feature Selection Based on Effective Range for Classification', *Sci. World J.*, vol. 2014, pp. 1-8, 2014.
DOI: 10.1155/2014/972125.

[18] B. Kumari and T. Swarnkar, 'Filter versus Wrapper Feature Subset Selection in Large Dimensionality Micro array: A Review', vol. 2, p. 6, 2011.

[19] R. P. L. Durgabai and Y. Ravi Bhushan, 'Feature Selection using ReliefF Algorithm', *IJARCCE*, pp. 8215-8218, Oct. 2014.
DOI: 10.17148/IJARCCE.2014.31031.

[20] R. J. Urbanowicz, R. S. Olson, P. Schmitt, M. Meeker, and J. H. Moore, 'Benchmarking Relief-Based Feature Selection Methods for Bioinformatics Data Mining', *ArXiv171108477 Cs*, Apr. 2018, Accessed: Jul. 13, 2021. [Online]. Available: http://arxiv.org/abs/1711.08477.

[21] S. Talukder, 'Mathematical Modelling and Applications of Particle Swarm Optimization', Master's Thesis, Blekinge Institute of Technology, 2011.

[22] S. Sengupta, S. Basak, and R. A. P. Ii, 'Particle Swarm Optimization: A survey of historical and recent developments with hybridization perspectives', p. 34, 2019.

[23] B. Sahu and D. Mishra, 'A Novel Feature Selection Algorithm using Particle Swarm Optimization for Cancer Microarray Data', *Procedia Eng.*, vol. 38, pp. 27-31, 2012.
DOI: 10.1016/j.proeng.2012.06.005.

[24] G. Armano, C. Chira, and N. Hatami, 'Error-Correcting Output Codes for Multi-Label Text Categorization', p. 12, 2013.

[25] S. Escalera, O. Pujol, P. Radeva, and P. Ivanova, 'Error-Correcting Ouput Codes Library', *J. Mach. Learn. Res.*, vol. 11, p. 4, 2010.

[26] 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', *Int. J. Comput. Eng. Res.*,

vol. 08, no. 9, p. 8, Sep. 2018.

[27] I. Bukovsky, W. Kinsner, and N. Homma, 'Learning Entropy as a Learning-Based Information Concept', *Entropy*, vol. 21, no. 166, pp. 1-14, 2019. DOI: 10.3390/e21020166.

[28] Z. Zhang, 'Introduction to machine learning: k-nearest neighbors', *Ann. Transl. Med.*, vol. 4, no. 11, pp. 218-218, Jun. 2016. DOI: 10.21037/atm.2016.03.37.

[29] S. P. Arade and J. K. Patil, 'COMPARATIVE STUDY OF DIABETIC RETINOPATHY USING K-NN AND BAYESIAN CLASSIFIER', *Int. J. Innov. Eng. Res. Technol.*, vol. 4, no. 5, pp. 55-61, 2017.

[30] A. Kataria and M. D. Singh, 'A Review of Data Classification Using K-Nearest Neighbour Algorithm', *Int. J. Emerg. Technol. Adv. Eng.*, vol. 3, no. 6, pp. 354-360, 2013.

[31] X. Gu, L. Akoglu, and A. Rinaldo, 'Statistical Analysis of Nearest Neighbor Methods for Anomaly Detection', in *33rd Conference on Neural Information Processing Systems*, canada, 2019, p. 11.

[32] E. A. Amrieh, T. Hamtini, and I. Aljarah, 'Mining Educational Data to Predict Student's academic Performance using Ensemble Methods', *Int. J. Database Theory Appl.*, vol. 9, no. 8, pp. 119-136, Aug. 2016. DOI: 10.14257/ijdta.2016.9.8.13.

[33] O. W. Adejo and T. Connolly, 'Predicting student academic performance using multi-model heterogeneous ensemble approach', *J. Appl. Res. High. Educ.*, vol. 10, no. 1, pp. 61-75, Feb. 2018. DOI: 10.1108/JARHE-09-2017-0113.

[34] A. Almasri, E. Celebi, and R. S. Alkhawaldeh, 'EMT: Ensemble Meta-Based Tree Model for Predicting Student Performance', *Sci. Program.*, vol. 2019, pp. 1-13, Feb. 2019. DOI: 10.1155/2019/3610248.

[35] M. B. Shah, M. Kaistha, and Y. Gupta, 'Student Performance Assessment and Prediction System using Machine Learning', in *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*, Mathura, India, Nov. 2019, pp. 386-390. DOI: 10.1109/ISCON47742.2019.9036250.

DOI: https://doi.org/10.30564/jcsr.v3i3.3534