International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Classification and Prediction of Student Academic Performance using Gray Wolf Optimization Based Relief-F Budget Random Forest

Kongara Deepika, Nallamothu Sathyanarayana

Abstract*: *The student academic prediction model helps to predict the student performance that helps the university to provide necessary care to the particular students. Efficient prediction model helps to encourage the student for better performance in the academic. In this research, the Relief-F Budget Tree Random Forest with Gray Wolf Optimization (RFBTRF-GWO) method is proposed for the feature selection. The Gray Wolf Optimization (GWO) helps to scale the relevant feature with ranking order from the features selected by the Relief-F Budget Tree Random Forest (RFBTRF). The selected features are given as input to the classifier for the effective prediction. The k-Nearest Neighbor (kNN) and Artificial Neural Network (ANN) are used for the classification. The proposed RFBTRF-GWO method is evaluated on the three datasets such as two UCI datasets and one collected dataset. The RFBTRF-GWO has a higher performance accuracy of 96.2 % while the existing method RFBTRF has an accuracy of 70.88 %.

Index Terms*: *Artificial Neural Network, Gray Wolf Optimization based Relief-F Budget Random Forest, k-Nearest Neighbor, and student academic prediction.

I. INTRODUCTION

The higher education institutions are now aware of the potential of learning educational data to increase the quality of their managerial decisions. The education institutions are involving in collecting the educational information and apply it with data mining techniques [1]. The major challenges involved in the process of predicting the student academic performance to develop the decision support system, thereby help in improving the teaching and learning practices. This is the complex process due to this involves in many aspects such as economic, cultural, social, academic background and demographic [2]. Early prediction of students who has the prone to drop their course, helps to prevent the student for such scenario. To reduce the risk of the problem, this is important to predict the risk as early as possible and provide some care to prevent the student to quit from their studies [3]. The structure data is needed to be provided to analyze various number of factors [4, 5].

A holistic method is needed to apply for the examination of the complex interactions between the various factors in learning space. The data collected from the student interaction is structured to enhance reliability [6, 7]. However, this is difficult to measure both the teaching quality and learning achievement. Several universities were involved in analyzing the teaching quality of students [8]. Examine with the student's past performance not only increases the complexity and also degrades the performance of the system [9, 10]. The student prediction model is developed based on the RFBTRF-GWO feature selection to increase efficiency. This method helps to select the relevant feature from the data to predict student academic performance. The proposed method is tested with three datasets namely math and Portuguese dataset from UCI and one collected university dataset. The RFBTRF-GWO is compared with other existing methods to investigate the performance of the prediction. This shows that the proposed method has a higher efficiency compared to the other existing method in student academic prediction.

The organization of the paper is that section II contains the related works, the proposed method explained in section III, section IV contains the experimental Design, section V discuss about the experimental result. The conclusion of this paper is made in section V.

II. RELATED WORKS

The tremendous growth in the university electronics data needs an effective method to handle and extract some meaningful information. The advancement in data mining techniques helps to provide examination of university data to increase the quality of educational process. Many research involves in the mining of information from the educational data and some recent methods are examined in this section.

RaheelaAsif, et al. [11] focused on the two aspects: predicting the student performance at the end of the four-year academic programme and typical progressions and combine them with prediction. The classification is made for two classes namely: low and high achieving students.

Manuscript published on 30 September 2019
* Correspondence Author
Kongara Deepika*, CSE, Talla Padmavathi College of Engineering, Telangana, India.
Dr. Nallamothu Sathyanarayana, CSE, Nagole Institute of Technology and Sciences, Telangana, India.

The decision tree and modified k-means algorithm were applied for the feature selection and classification. The developed method is tested on the collected dataset related to student performance. The experimental result shows that the developed method has higher performance. The method is needed to be tested on the standard dataset.

Concepción Burgos, et al. [12] applied the data mining techniques to the student history of data to predict the student drop out of a course. A logistic regression model was applied to filter out the feature and classify the data. An experiment is tested on over 100 students based on several distance learning course, shows that the performance of developed method. A tutorial action plan is developed based on the result of the predictive model. This predictive model helps to reduce the dropout rate of 14 % related to the previous academic years in which no dropout techniques are applied. The Madrid Open University data is used to test the performance of the developed method. The drawback of the logistic regression is that predict the data present in the linear nature of a continuous variable.

SumyeaHelal, et al. [13] developed the subgroup discovery to extract the significant aspects related to the student performance outcome. The method uses the student demographic and academic data, course and data retrieved from the institution learning management system to investigate the student performance. The result shows that the effectiveness of the subgroup discovery in identifying the factors. The Moodle data is used to analyze the performance of the developed method and the developed method shows the considerable performance. The efficiency of the prediction model is low.

Eduardo Fernandes, et al. [14] developed prediction method and applied in the public school of the Federal District of Brazil. The statistical model is applied to analyze the data and two datasets were used to test the performance. Gradient Boosting Machine (GBM) is applied for the classification in the student performance prediction. The method shows that student residence and school are the major factors in academic performance. The various factors are needed to be consider to increase the performance of the prediction.

Feras Al-Obeidat, et al. [15] developed hybrid technique of decision tree and fuzzy multi-criteria classification in the student academic prediction. There are several factors used to analyze the performance of the system. The UCI dataset were used to test the performance of the methods. The standard classifiers were used and compared with the existing method to analyze the effectiveness. The effectiveness of the method is low and feature selection techniques were used to increase the performance of the system.

A. Problem statement and solutions

The significant limitations of logistic regression and other existing method is cannot predict with the linear nature of continuous variables present in students' performance data like grade, age etc. In existing work, Relief-F algorithm is used for FS but it's not able to perform on incomplete and noisy data. To overcome these issues, FS method namely Relief-F and Budget Tree-Random Forest algorithm is used that reduce the irrelevant features and improve the prediction accuracy of Student Academic Performance.

- The Relief F algorithm calculates a feature score for each attribute based on the distance value and this provides number of features that affects the efficiency of the method. Hence, this requires scaling technique to reduce the number of features.
- Random forest algorithm constructs a collection of trees, where each tree is grown by random independent data sampling & feature splitting, which produce a collection of independent identically distributed trees.

Solution: To overcome this limitation the Grey Wolf Optimization (GWO) with Relief-F and Budget Tree-Random algorithm is used to optimize the imbalanced features. Here, optimization technique based GWO algorithm is applied for feature selection and it will select the more relevant features and improves the classification performance.

III. PROPOSED METHOD

The student academic prediction system helps to provide the information to the university about student performance. The major objective of this research is to provide efficient student academic prediction using RFBTRF-GWO. The RFBTRF-GWO methods is evaluated in the three datasets such as two UCI dataset and one collected dataset. The GWO is proposed in the RFBTRF to select the relevant features in the university data with the rank value. The RFBTRF-GWO select the relevant value and provides the features for binary classification. The two classifiers such as ANN, KNN are used to evaluate the performance of the proposed feature selection. The proposed method is evaluated and compared with the existing method to investigate efficiency.

A. Data acquisition and pre-processing

Data of the student performance is taken from the UCI dataset. The 10th standard student performance data are also collected for four schools in the year of 2013 to 2017. In UCI, there are two datasets such as Mathematics (Math) and Portuguese language. These datasets were used to evaluate the performance of the proposed method. The Math dataset consist of 395 instances and Portuguese dataset consist of 650 instances. The collected dataset has 4965 instances. The factors such as school name, age, gender, travel time, hobbies, health details etc.., were collected based on school reports and questionnaires [16]. The student performance is classified as low and high performance. The input data were converted as string format of integer or float. The pre-processed data were applied for the FS process, which is explained in the next sub section.

B. Relief-F and Budget Tree Random forest

The main process of FS is to control the size of the features in the dataset. First, the original data were used without involves in the dropping information. The irrelevant features are removed to reduce the dimension of the data.

This helps to mining accuracy, decrease computation time and increase the performance of the classification. The Relief-F algorithms were used for the feature selection and this doesn't able to process the incomplete and noisy data [17]. Hence, RFBT-RF is applied to solve these problems. Relief-F algorithm measures the feature score for each attributes and rank them in order to eliminate the irrelevant features [18]. These scores can be applied as feature weights to guide downstream model. The features can be selected by initiating the weight in the system. The weight initialization is provided in the Eq. (1).

$$Weight\, C[A] = 0.0 \tag{1}$$

At each iteration, Relief-F algorithm involves in the feature vector (x) related to one random instance, and instance close to x are measured from Euclidean distance of each class. The shorter distance from same class is called 'near-hit', and closer instance from another class is 'near-miss'. Algorithm analyzes the hits and miss of each class in random instances, which is mathematically measured in Eq. (2),

$$C_i = C_i - (x_i - nearHit_i)^2 + (x_i - nearmiss)^2 \tag{2}$$

Once the hit and miss are measured, then the value is applied to BT-RF algorithm. The feature cost is processed using the Eq. (3),

$$C[A] = C[A] - \frac{\sum_{j=1}^{k} diff\,(A, r, h_{ji})}{(m,k)} \tag{3}$$

Where $C[A]$ consist of weight value of all attributes A that is depend on r_i value, which is random instance. The k is search as nearest neighbors from same class, called nearest hit h_j. The random forest Function F aims to minimize the expected loss subject to a budget constraint is shown in Eq. (4). The label pairs (x, y) are distributed as $(x, y)d \sim H$. The aim is to learn the classifier f from the function F, which reduce the expected loss related to budget constraint:

$$f_{\in}^{\min} FE_{xy}\left[L(y, f(x))\right], E_x\left[C(f, x)\right] \le B \tag{4}$$

Where $L(y, y)$ represent the loss function, $C(f, x)$ is the cost value and B is used specified budget constraint. The feature acquisition cost $C(f, x)$ is a modular support features based on function f on example x, that is acquired each feature with fixed constant costs. Without cost constraint, the problem is equivalent to a supervised learning problem and adding cost creates a combinatorial problem. In practice, the data is trained as $(x_i, y_i), \ldots, (x_n, y_n)$ drawn IID with $(x_i, y_i)d \sim H$. The empirical loss subjection is minimizing a budget constraint in Eq. (5).

$$f_{\in}^{\min} F\frac{1}{n} L(y_i, f(x_i)), \frac{1}{n}\sum_{i=1}^{n} C(f, x_i) \le B \tag{5}$$

The classifier function f is a random forest with consist K random trees, $D_1, D_2, \ldots, D_K$ that are learn on training data. The expected cost of the instance x is measured as in Eq. (6).

$$E_f\left[E_x\left[C(f, x)\right]\right] \le \sum_{j=1}^{K} E_{Dj}\left[E_x\left[C(D_j, x)\right]\right] \tag{6}$$

Where, RHS are average related to the random trees. The trees of random forest are uniformly distributed in the RHS scales. The upper bound analyze the typical features of random forest due to the low feature correlation among trees. The Greedy tree is mathematically shown in Eq. (7).

$$(t) := \min_{g_t \in G_t} \max_{i \in outcomes} \frac{c(t)}{F(S) - F(S_{gt}^{i})} \tag{7}$$

Whereas, S_{gt}^{i} is the example set in S that has outcome i based on the classifier $g(t)$ with feature importance t. Greedy tree helps to process all the features simultaneously in RFBT-RF. The decision tree is construct by budget RF by using the Greedy tree as sampled features from the training data unless the budget B is decreased as evaluated based on validation data. The tree ensemble was returned as output. The random forest selects less number of feature than Relief-F. The more relevant and less number of feature increases the performance and reduce the computation time. The GWO method with the Relief-F and random forest, examine the feature influences and reduces the number of feature to perform effectively.

C. Grey Wolf Optimization

Grey Wolf Optimization (GWO) method is inspired by the hunting behavior of grey wolves. The GWO mimic the hunting process of grey wolves including searching prey, encircling, tracking and attacking [19, 20]. The method follows the manner of social leadership hierarchy. The grey wolf has the strict role of following the social dominant hierarchy. These are involving in four levels. The top-level wolf are said to be α that makes decision and the second level wolf is β helps in support for decision and other pack activity. The β analyze the α command and provide feedback to alpha. The next level rank is omega ω, which plays the role of scape goat. The other wolves such as scouts and sentinels are called delta δ. The delta has to submit to alpha and beta, but it dominates the omega. The process of GWO is explained as below

1. The search agent is a member of grey wolf pack
2. The solution denotes the grey wolf position
3. The quality of solution related to the position of the corresponding wolf
4. The global solution related to prey location
5. The three best solutions denote the alpha, beta and delta in the hierarchy
6. The iteration is the hunting process of the wolf pack

To simulate the hunting process of the grey wolf and measure the three best solution $(x_\alpha, x_\beta, x_\delta)$, and update the position of wolf by a group evolving is in the Eq. (8 - 11)

$$y1 = x_\alpha^j - a1 * \left| c1 * x_\alpha^j - x_i^t(t) \right| \quad (8)$$

$$y2 = x_\beta^j - a2 * \left| c1 * x_\beta^j - x_i^j(t) \right| \quad (9)$$

$$y3 = x_\delta^j - a1 * \left| c1 * x_\delta^j - x_i^t(t) \right| \quad (10)$$

$$x_i^j(t+1) = \frac{y1 + y2 + y3}{3} \quad (11)$$

Where $x_i^j(t)$ is the i^{th} solution in the j^{th} dimension at the t^{th} iteration. The random numbers $c1, c2\,and\,c3$ were uniformly distributed in [0, 2]. The random numbers $a1, a2, and\,a3$ are uniformly distributed in $\left[-2*\left(1-\frac{t}{G}\right), 2*\left(1-\frac{t}{G}\right)\right]$; $i = 1, 2, ..m$ indexes each solution in population with size m; $j = 1, 2, ...n$; $j = 1, 2, ..., n$ indexes each dimension of the n -dimensional problem to be solved; $t = 1, 2, ...G$ is the number of iteration.

The random number fluctuation range $a1, a2, \text{and}\, a3$ linearly decreases as the iteration number t gradually increases the pre-determined maximal number G. Minimizing the fluctuation characteristics analyze the process of the prey tactic in the hunting process. If the amplitude is less than 1, the wolf attacks towards the prey and related search agent will analyze a local optimal solution. If amplitude is more than 1, searching process analyze more distinct solutions.

The GWO process is considerably simple. This is start from the random wolf pack and continuously update the position of each wolf unless iterator is beyond an iteration.

Initialization
repeat
Update wolf position based on Eq. (8)–(11)
Provide the new position to the wolves
Update of $\alpha, \beta, \text{and}\, \delta$
Until the stopping criterion is reached.

D. Classifiers

Gray Wolf Optimization based Relief-F Random forest (RFBTRF-GWO) is proposed in this research for the relevant feature selection and this helps to improve the prediction of student academic performance. The RFBTRF-GWO method select the features based on the order of importance and this given to the input of the classifier such as kNN and ANN. The kNN classifiers was used in the student academic predictions and ANN is applied to investigate its performance.

kNN: Many researchers have tried to use kNN classifier for pattern recognition and classification with respect to the training data [21]. The result of comparing the fuzzy version with the Crisp version shows that the fuzzy algorithm has a lower error rate. The kNN algorithm is simple method for solving the classification problem that has competitive results and has higher performance than other data mining techniques [22]. The classifier is tested and proved to be have the capacity to solve the problem of the other algorithms.

ANN: ANNs is the mathematical models that mimic the complex neuron pattern interconnect in the human brain. From the examination of many applications, ANN has higher efficiency in pattern recognition [23]. The structure of the ANN is similar to the human brain. The neural network system has two computational processes [24, 25]. The number of features selected by the Relief-F, random forest and RFBTRF-GWO method in Table 1.

Table I. Feature selected by various methods

Dataset	Relief F	RFBT	GWO
UCI (Mathematics) dataset	'G2', 'G1', 'failures', 'higher', 'schoolsup', 'school', 'Pstatus', 'internet', 'address', 'romantic', 'nursery', 'Dalc', 'famsize', 'guardian', 'paid', 'sex', 'traveltime', 'activities', 'famsup', 'Fjob', 'famrel', 'goout', 'Medu', 'studytime', 'absences'	'internet', 'higher', 'Fjob', 'nursery', 'address', 'famsize', 'activities', 'Pstatus', 'sex', 'famsup', 'traveltime', 'reason', 'schoolsup', 'paid', 'Mjob', 'Medu', 'guardian', 'Fedu', 'age', 'failures', 'school', 'studytime'	'internet', 'higher', 'Fjob', 'nursery', 'address', 'famsize', 'activities', 'Pstatus', 'sex', 'famsup'
Support values	5370, 778, -3166, 3542, 1462, 5294, -2212, -3242, -3442, -1296, -3178, 1402, 646, -2522, 9154, 6276, 288, 888, 530, 2808, 7596, 4536, 2972, -1876, -3358.	0.009, 0.024, 0.012, 0.03, 0.028, 0.024, 0.017, 0.012, 0.017, 0.035, 0.019, 0.012, 0.019, 0.007, 0.011, 0.018, 0.021, 0.018, 0.026, 0.031, 0.194, 0.405.	0.399, 0.394, 0.349, 0.325, 0.189, 0.188, 0.161, 0.087, 0.063, 0.031.
UCI Portugues	'G2', 'G1', 'failures', 'higher', 'paid', 'schoolsup', 'Pstatus', 'internet', 'school', 'nursery', 'Dalc', 'address', 'famsize', 'romantic', 'guardian', 'sex', 'famsup', 'activities', 'traveltime', 'absences', 'Fjob', 'famrel', 'studytime', 'Walc', 'Medu', 'reason', 'Mjob', 'freetime', 'Fedu', 'health'	'freetime', 'famrel', 'guardian', 'studytime', 'Medu', 'school', 'romantic', 'Pstatus', 'famsup', 'higher', 'internet', 'Mjob', 'nursery', 'activities', 'famsize', 'age', 'failures', 'schoolsup', 'sex', 'Fedu', 'Fjob', 'address', 'reason', 'paid', 'traveltime'	'freetime', 'famrel', 'guardian', 'studytime', 'Medu', 'school', 'romantic', 'Pstatus', 'famsup', 'higher'

Support values	15742.0, 5520.0, -15248.0, 12762.0, 7828.0, 22590.0, -10746.0, -12340.0, -11834.0, -3150.0, -10896.0, 5884.0, 1074.0, -8006.0, 36576.0, 23406.0, 5168.0, 30928.0, 2190.0, 15674.0, 36006.0, 15830.0, 6646.0, -5272.0, -12138.0, -13652.0, 13528.0, -9562.0, -12666.0, -1726.0	0.032, 0.018, 0.021, 0.017, 0.021, 0.025, 0.033, 0.018, 0.023, 0.017, 0.017, 0.036, 0.005, 0.035, 0.018, 0.018, 0.024, 0.016, 0.023, 0.023, 0.024, 0.024, 0.026, 0.223, 0.251.	0.194, 0.177, 0.143, 0.135, 0.1, 0.083, 0.073, 0.026, 0.02, 0.016.
School dataset	'studytime', 'Alchoholic', 'activities', 'Soc_feedback', 'Mjob', 'E_feedback', 'Qualified_teachers', 'family_support', 'schoolsup', 'hostel', 'guardian', 'fee_range', 'Pstatus', 'T_feedback', 'ENG', 'M_feedback', 'year', 'Medu', 'SOC', 'Sci_feedback', 'famsize', 'H_feedback', 'healthproblem', 'SCI', 'TEL', 'MAT', 'Fedu', 'HIN', 'Fjob', 'M_Exp', 'Soc_Exp', 'Reason_to_choose_school', 'goout', 'H_Exp', 'E_Exp'	'SA-1', 'Alchoholic', 'studytime', 'Soc_feedback', 'MAT', 'M_feedback', 'activities', 'HIN', 'TEL', 'SCI', 'ENG', 'Sci_feedback', 'Student_ID', 'T_feedback', 'attendence', 'E_feedback', 'Fedu', 'SOC', 'year', 'age', 'Fjob', 'Distance_from_home_to_school', 'Medu', 'H_feedback', 'goout', 'fee_range', 'E_Exp', 'Reason_to_choose_school', 'Sci_Exp', 'Mjob', 'traveltime', 'Tel_Exp', 'M_Exp'	'SA-1', 'Alchoholic', 'studytime', 'Soc_feedback', 'MAT', 'M_feedback', 'activities', 'HIN', 'TEL', 'SCI'
Support values	-71718.0, -4828.0, -40988.0, 42324.0, 49780.0, -51012.0, 69410.0, 8112.0, 6104.0, 43806.0, 59132.0, 27242.0, 46976.0, 102388.0, 15252.0, 69392.0, 71722.0, 242402.0, -12846.0, 248300.0, 71722.0, 62380.0, 204194.0, 71722.0, 2342.0, 58412.0, 4056.0, 42428.0, 3848.0, 99956.0, 11680.0, 54642.0, 604.0, 43878.0, 9832.0.	0.018, 0.008, 0.007, 0.003, 0.01, 0.015, 0, 0.005, 0.006, 0.004, 0, 0.012, 0.007, 0.005, 0.007, 0, 0, 0.12, 0.005, 0.109, 0, 0.006, 0.05, 0, 0.005, 0.017, 0.004, 0.006, 0.005, 0.013, 0.004, 0.053, 0.005.	0.12, 0.119, 0.113, 0.096, 0.084, 0.08, 0.05, 0.049, 0.027, 0.012.

The selected features and the respective values are provided in the Table. (1). The feature values and support values are provided, based on that student prediction is made. The support values measure the influences of the features in the student performance. For instance, The GWO in the UCI math dataset has measured the support value of the internet has 0.3993 in the student performance.

IV. EXPERIMENTAL DESIGN

For experimental simulation, PyCharm software was employed on PC with 3.2 GHz with i5 processor. In order to estimate the efficiency of proposed RFBT-RF algorithm, the performance of the proposed method was compared with the existing method [20]. In experimental evaluation, three databases were used. Those are, UCI (maths), UCI (Portuguese) database and Collected school database. The performance of the RFBT-RF methodology was compared by means of accuracy, precision, recall and F-score.

Performance measure is defined as the relationship between the input and output variables of a system understand by employing the suitable performance metrics like precision and recall. The general formula for calculating the precision and recall of the SAP prediction is given in the Eq. (12) and (13).

$$\Pr ecision = \frac{TP}{TP + FP} \tag{12}$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{13}$$

Accuracy is the measure of statistical variability and a description of random errors. The general formula of accuracy for determining student performance prediction using different classifier efficiency is given in the Eq. (14).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{14}$$

Where, TP is represented as true positive, FP is denoted as false negative, TN is represented as true negative and FN is stated as a false negative. F-score is the measure of accuracy test and it considers the both precision P and recall R of the test in order to calculate the score. The general formula for F-score is given in the Eq. (15).

$$F - measure = 2.\frac{\Pr ecision.\mathrm{Re}call}{\Pr ecision + \mathrm{Re}call} \times 100 \tag{15}$$

V. EXPERIMENTAL RESULT

The student academic prediction model is tested on the two UCI datasets such as math, and Portuguese. The RFBTRF-GWO is also tested on the collected student datasets to effectively examination the performance. The common metrics such as accuracy, precision, recall and F-measure are measured from the RFBTRF-GWO and compared with existing method. The two classifiers are used to analyze the performance of the RFBTRF-GWO such as kNN and ANN. The kNN classifier is used in the existing method to analyze the performance of the feature selection and in addition to that ANN is used in this method to investigate the performance. This section provides the detailed description about the performance of the proposed feature selection method.

Table II. RFBTRF-GWO evaluated in UCI (Math) Dataset

UCI (Math) Dataset						
Classifier	Feature Selection Method	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)	Feature Selected
KNN	RFBTRF	70.88	71	71	71	'internet', 'higher', 'Fjob', 'nursery', 'address', 'famsize', 'activities', 'Pstatus', 'sex', 'famsup', 'traveltime', 'reason', 'schoolsup', 'paid', 'Mjob', 'Medu', 'guardian', 'Fedu', 'age', 'failures', 'school', 'studytime'
	RFBTRF-GWO	96.2	96	96	96	'internet', 'higher', 'Fjob', 'nursery', 'address', 'famsize', 'activities', 'Pstatus', 'sex', 'famsup'
ANN	RFBTRF	81.01	81	81	81	'internet', 'higher', 'Fjob', 'nursery', 'address', 'famsize', 'activities', 'Pstatus', 'sex', 'famsup', 'traveltime', 'reason', 'schoolsup', 'paid', 'Mjob', 'Medu', 'guardian', 'Fedu', 'age', 'failures', 'school', 'studytime'
	RFBTRF-GWO	96.2	96	96	96	'internet', 'higher', 'Fjob', 'nursery', 'address', 'famsize', 'activities', 'Pstatus', 'sex', 'famsup'

Table III. RFBTRF-GWO evaluation in UCI Portuguese dataset

UCI (Portuguese)						
Classifier	**Feature Selection Method**	**Accuracy**	**Precision**	**Recall**	**F-Measure**	**Feature Selected**
KNN	RFBTRF	67.69	68	68	68	'freetime', 'famrel', 'guardian', 'studytime', 'Medu', 'school', 'romantic', 'Pstatus', 'famsup', 'higher', 'internet', 'Mjob', 'nursery', 'activities', 'famsize', 'age', 'failures', 'schoolsup', 'sex', 'Fedu', 'Fjob', 'address', 'reason', 'paid', 'traveltime'
	RFBTRF-GWO	93.07	93	93	93	'freetime', 'famrel', 'guardian', 'studytime', 'Medu', 'school', 'romantic', 'Pstatus', 'famsup', 'higher'
ANN	RFBTRF	79.23	79	79	79	'freetime', 'famrel', 'guardian', 'studytime', 'Medu', 'school', 'romantic', 'Pstatus', 'famsup', 'higher', 'internet', 'Mjob', 'nursery', 'activities', 'famsize', 'age', 'failures', 'schoolsup', 'sex', 'Fedu', 'Fjob', 'address', 'reason', 'paid', 'traveltime'
	RFBTRF-GWO	95.38	95	95	95	'freetime', 'famrel', 'guardian', 'studytime', 'Medu', 'school', 'romantic', 'Pstatus', 'famsup', 'higher'

The RFBTRF-GWO method is evaluated in the UCI math dataset and compared with the RFBTRF method to understand the effectiveness of the proposed method. Two classifiers such as ANN and kNN are used to evaluate the performance of the RFBTRF-GWO method. The different metrics are measured from the RFBTRF-GWO and RFBTRF, and shown in the Table 2. As the table shows that the GWO method has higher performance in both classifiers than the existing RFBTRF method. The RFBTRF-GWO method has the F-measure of 96 % in kNN while existing method RFBTRF has the F-measure of 71 %. The proposed method has the higher performance in other parameter as well. From the examined two classifier, the kNN has the higher performance than the ANN when combining with the proposed feature selection algorithm in UCI math database.

The RFBTRF-GWO method is evaluated in the UCI Portuguese dataset with two classifiers and compared with the RFBTRF method, as shown in Table 3. This shows that the proposed method has the higher performance compared to the RFBTRF method. The features selected by the method are also compared with existing method. The RFBTRF-GWO method has the higher performance in the four parameters. The accuracy of the proposed RFBTRF-GWO method in ANN method is 95.38 % while compared with the RFBTRF method has 79.23 %.

The RFBTRF-GWO is processed in the collected database to analyze the performance of the method in the different data. The existing method of RFBTRF is evaluated in the same collected dataset and compared with RFBTRF. The RFBTRF-GWO and RFBTRF are analyzed with the different metrics in the collected student dataset, as shown in Table 4. The table shows that the proposed feature selection method has the higher performance compared to the existing method. The feature selected by the method is also shown in the table along with the performance metrics. The F-measure of the RFBTRF-GWO is 94 % compared with the RFBTRF method of 91 %. This algorithm performed well in student collected dataset but performance lagging in standard datasets because of its imbalance ratio between different classes of training and testing. The collected data have lower variation in the data causes it to have the imbalance data.

Table IV. RFBTRF-GWO evaluated in collected student dataset

Collected Student Database						
Classifier	Feature Selection Method	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)	Feature Selected
KNN	RFBTRF	64.65	65	65	65	'SA-1', 'Alchoholic', 'studytime', 'Soc_feedback', 'MAT', 'M_feedback', 'activities', 'HIN', 'TEL', 'SCI', 'ENG', 'Sci_feedback', 'Student_ID', 'T_feedback', 'attendence', 'E_feedback', 'Fedu', 'SOC', 'year', 'age', 'Fjob', 'Distance_from_home_to_school', 'Medu', 'H_feedback', 'goout', 'fee_range', 'E_Exp', 'Reason_to_choose_school', 'Sci_Exp', 'Mjob', 'traveltime', 'Tel_Exp', 'M_Exp'
	RFBTRF-GWO	96.87	97	97	97	'SA-1', 'Alchoholic', 'studytime', 'Soc_feedback', 'MAT', 'M_feedback', 'activities', 'HIN', 'TEL', 'SCI'
ANN	RFBTRF	91.33	90	92	91	'SA-1', 'Alchoholic', 'studytime', 'Soc_feedback', 'MAT', 'M_feedback', 'activities', 'HIN', 'TEL', 'SCI', 'ENG', 'Sci_feedback', 'Student_ID', 'T_feedback', 'attendence', 'E_feedback', 'Fedu', 'SOC', 'year', 'age', 'Fjob', 'Distance_from_home_to_school', 'Medu', 'H_feedback', 'goout', 'fee_range', 'E_Exp', 'Reason_to_choose_school', 'Sci_Exp', 'Mjob', 'traveltime', 'Tel_Exp', 'M_Exp'
	RFBTRF-GWO	94.36	94	94	94	'SA-1', 'Alchoholic', 'studytime', 'Soc_feedback', 'MAT', 'M_feedback', 'activities', 'HIN', 'TEL', 'SCI'

On UCI math dataset "internet" has 39%, "higher" has 39%, "Fjob" has 34%, "nursery" has 32%, "address" has 18%, "famsize" has 18%, "activities" has 16%, "Pstatus" has 8%, "sex" has 6%, and "famsup" has 3% of impact on predicting G3. On UCI Portuguese dataset "freetime" has 19%, "famrel" has 17%, "guardian" has 14%, "studytime" has 13%, "Medu" has 10%, "school" has 8%, "romantic" has 7%, "Pstatus" has 2%, "famsup" has 2% and "higher" has 1% of impact on predicting G3. On School dataset parent "SA-1" has 12%, "Alchoholic" has 11%, "studytime" has 11%, "Soc_feedback" has 9%, "MAT" has 8%, "M_feedback" has 8%, "activities" has 5%, "HIN" has 4%, "TEL" has 2%, and "SCI" has 1% of impact on predicting Grade of student.

Based on these features students are having impact on their academic performance and inferences to parents and Teachers need to be given to improve the student academic performance on school dataset.

Table V. Evaluation of different method in standard database

Dataset	Methods	Accuracy (%)
Math	SVM	86.3
	DT	90.7
	Random Forest	91.2
	DTFMC [15]	82.28
	RFBTRF + kNN	70.88
	RFBTRF-GWO + kNN	96.2
Portuguese	SVM	91.4
	DT	93
	Random Forest	92.6
	DTFMC [15]	85.82
	RFBTRF + kNN	67.69
	RFBTRF-GWO + kNN	93.07

The different techniques used in the student performance predictions are compared with RFBTRF-GWO method. The classifiers such as Support Vector Machine, Decision Tree (DT), Random Forest, and DTFMC [15] are compared with the proposed method in Table 5. The RFBTRF-GWO method has the highest performance compared to the other classifier. Therefore, the proposed method has a higher efficiency compared to the other classifiers in student performance prediction and this method can be applied to practical use.

VI. CONCLUSION

Early prediction of student performance helps the universities to encourage the student to improve the academic performance. The data mining techniques help to investigate the number of factors involves in the student performance. In this research, RFBTRF-GWO is proposed to select the features from the data to forecast student performance. The RFBT-RF selects the features from dataset and given as input to the GWO to order the features based on its support value. The best rank order of the selected features is considering for the classification. The classifiers predict student performance based on the selected features. The outcome of the method is compared with existing method, which shows the proposed method has the higher performance than existing method in performance prediction. The proposed RFBTRF-GWO has an accuracy of 93.07 % in UCI Portuguese datasets while the existing method has the accuracy of 67.69 %.

REFERENCES

1. V. L. Miguéis, A. Freitas, P. J. Garcia, and A. Silva. "Early segmentation of students according to their academic performance: a predictive modelling approach," *Decision Support Systems, 115*, pp. 36-51.
2. S. Helal, J. Li, L. Liu, E. Ebrahimie, S. Dawson, D. J. Murray, and Q. Long. (2018). Predicting academic performance by considering student heterogeneity. *Knowledge-Based Systems, 161*, pp. 134-146.
3. C. Márquez-Vera, A. Cano, C. Romero, A.Y.M. Noaman, H. MousaFardoun, and S. Ventura. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems, 33(1)*, pp. 107-124.

4. A. Pardo, F. Han, and R. A. Ellis. (2016). Exploring the relation between self-regulation, online activities, and academic performance: A case study. *In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pp. 422-429.
5. H. Fujita, "Neural-fuzzy with representative sets for prediction of student performance," *Applied Intelligence, 49(1),* pp. 172-187.
6. G.D. Saenz, L. Geraci, T. M. Miller, and R. Tirso, "Metacognition in the classroom: The association between students' exam predictions and their desired grades", *Consciousness and cognition, 51*, pp. 125-139.
7. M. R. M. Veeramanickam, M. Mohanapriya, B.K. Pandey, S. Akhade, S.A. Kale, R. Patil, and M. Vigneshwar. Map-reduce framework based cluster architecture for academic student's performance prediction using cumulative dragonfly based neural network. *Cluster Computing,* pp.1-17, 2018.
8. A. Khan, and S. K. Ghosh. (2018). Data mining based analysis to explore the effect of teaching on student performance. *Education and Information Technologies, 23(4),* pp. 1677-1697.
9. W. Xing, R. Guo, E. Petakovic, and S. Goggins. (2015). Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. *Computers in Human Behavior, 47,* pp. 168-181.
10. J. Xu, K. H. Moon, and V. M. DerSchaar (2017). A machine learning approach for tracking and predicting student performance in degree programs. *IEEE Journal of Selected Topics in Signal Processing, 11(5),* pp.742-753.
11. R. Asif, A. Merceron, S. A. Ali, and N. G. Haider. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education, 113*, pp. 177-194.
12. C. Burgos, M. L. Campanario, D. de la Peña, J. A. Lara, D. Lizcano, and M. A. Martínez. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering, 66,* pp. 541-556.
13. S. Helal, J. Li, L. Liu, E. Ebrahimie, S. Dawson, and D.J. Murray. (2018). Identifying key factors of student academic performance by subgroup discovery. *International Journal of Data Science and Analytics,* pp.1-19.
14. E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Van Erven. (2019). Eduational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research, 94,* pp. 335-343.
15. F. Al-Obeidat, A. Tubaishat, A. Dillon, and B. Shah, "Analyzing students' performance using multi-criteria classification," *Cluster Computing,* pp. 1-10.
16. P. Cortez, and A. M. G. Silva. (2008). Using data mining to predict secondary school student performance.
17. R. P. L. Durgabai, and Y. R. Bhushan. (2014). "Feature selection using ReliefF algorithm", *International Journal of Advanced Research in Computer and Communication Engineering, 3(10),* pp. 8215-8218.
18. T. R. Reddy, B.V. Vardhan, M. GopiChand, and K. Karunakar. (2018). Gender Prediction in Author Profiling Using ReliefF Feature Selection Algorithm. *In: Proc. of International Conf. on Intelligent Engineering Informatics, Springer,* Singapore, pp. 169-176, 2018.
19. D. A. Adeniyi, Z. Wei, and Y. Yongquan. (2016). Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. *Applied Computing and Informatics, 12(1),* pp. 90-108.
20. E. Emary, H. M. Zawbaa, and A. E. Hassanien. (2016). Binary grey wolf optimization approaches for feature selection. *Neurocomputing, 172,* pp. 371-381.
21. K. Luo. (2019). Enhanced grey wolf optimizer with a model for dynamically estimating the location of the prey. *Applied Soft Computing.*
22. N. Liu, X. Xu, Y. Li, and A. Zhu. (2019). Sparse representation based image super-resolution on the KNN based dictionaries. *Optics & Laser Technology, 110,* pp. 135-144.
23. S. Raith, E. P. Vogel, N. Anees, C. Keul, J. F. Güth, D. Edelhoff, and H. Fischer. (2017). Artificial Neural Networks as a powerful numerical tool to classify specific features of a tooth based on 3D scan data. *Computers in biology and medicine, 80,* pp. 65-76.
24. A. Jafarian, S. M. Nia, A. K. Golmankhaneh, and D. Baleanu. (2018). On artificial neural networks approach with new cost functions. *Applied Mathematics and Computation, 339,* pp. 546-555.
25. U. Anitha, S. Malarkkan, G. A. Jebaselvi, and R. Narmadha. (2019). Sonar image segmentation and quality assessment using prominent image processing techniques. *Applied Acoustics, 148*, pp. 300-307.

AUTHORS PROFILE

Mrs. K. Deepika Research Scholar of JNTUH, Hyderabad, in Data mining specialization. Presently working as Associate professor in Talla Padmavathi College of Engineering with 12 years of teaching experience in computer science and Engineering, and has a 6 years of research experience in this area. She has published many journals and International conferences on Data mining area. She has interest in this area and attended many workshops for enhancing the knowledge in this area. She had worked on Data mining project at her M.tech and published national and international conference. She has a good qualification in Engineering with B.tech and M.Tech in computer science and Engineering. She has worked on many techniques in Data mining area and guided students in their B.Tech and M.tech projects. She has handled subjects related to Data mining and other subjects in the computer science and engineering department and enhanced her knowledge in teaching and research area.

Dr. N. Satyanarayana who possessed the highest qualification in engineering. He is presently working as the Principal at Nagole Institute of Technology and Science. His qualifications are M.Sc, AMIE (ET), M.Tech (CS), Ph.D(CSE), MISTE, MCSI. He obtained his Ph.D in Computer Science in area of Advanced Computer Architecture and guiding many scholars in area of Networking and Data warehousing and Data Mining. He had a vast experience in Research area of Networking and Data Mining and has published many papers in Journals and National and International Conferences. He had a remarkable record of merit in pursuit his engineering studies. He has a disposition to inspire both students and staff to achieve their right goals. He has 15 years' experience in teaching and research into the learning systems and the collective efforts of the faculty. With his extensive and rich research experience he is able to run the institution and guide the scholars.