Received: October 1, 2018

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Relief-F and Budget Tree Random Forest Based Feature Selection for Student Academic Performance Prediction

Kongara Deepika [1]* **Nallamothu Sathyanarayana [2]**

[1]*Talla Padmavathi College of Engineering, India*
[2]*Nagole Institute of Technology and Sciences, India*
* Corresponding author's Email: deepika.srimanu@gmail.com

Abstract: Now-a-days, research in educational mining focuses on modelling student's performance. Many universities include large volumes of data related to student's details, performance, management details, educational process, and etc. Moreover, most of the data remains unused because inability of the university administration to handle it, also huge volumes of data are difficult to perform. In this paper, hybrid Feature Selection (FS) method namely Relief-F and Budget Tree-Random Forest (RFBT-RF) is proposed for selecting active features to reduce high dimensionality and handle uncertainty of data. The proposed feature selection method selects only relevant features instead of selecting redundant and irrelevant features for the classifiers. Also, RFBT-RF method is applied on multiple classifiers like Decision Tree (DT), Naive Bayes (NB) and Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) for predicting the Student Academic Performance (SAP). RFBT-RF method was applied on three databases such as UCI (maths), UCI (Portuguese) and Collected dataset. Results showed that, RFBT-RF algorithm achieved 6.85% of improved SAP accuracy compare to the existing Logistic Regression (LR) model.

Keywords: Artificial neural networks, Budget tree-random forest, Decision tree, Educational data mining, K-nearest neighbour, Support vector machine.

1. Introduction

Education is a complex process, and its effectiveness is affected by many factors like nation's economic prosperity, educational policies, etc. [1]. Except the teacher, students can also get useful information through cooperation with other students. Cooperation with other students enables the transfer of knowledge acquired by individual members of the group to all other members. Furthermore, questions of a group member can point out the deficiencies in the knowledge of other members, or encourage the entire group to consider a certain relevant topic [2]. Social Networking Sites (SNSs) are very popular today, especially among the younger population. SNSs enables the students to share their knowledge in a simple way. Social relationships are a significant part of the university experience of undergraduate students. The social media has become common among teenagers, most of the relationships maintained in real life are also translated online [3].

Educational Data Mining (EDM) is the significant application which employs data mining techniques to analyse the educational data. The EDM applications improves the pedagogical support, forecasting of student performance, clustering educational data and etc. [4, 5]. Additionally, some students are not able to learn quickly on that condition defining the learner models like demographics, characteristics, preferences and cognitive traits for improving the learning skills [6]. The social media activities highly influence and divert the student's concentration towards the non-educational topics [7]. In educational institutes, student's academic details are stored in large datasets and extracts the useful information's like academic policies on how to improve student retention rates, allotment of teaching details and support resources, create intervention strategies or student performance, and etc. [8]. This research work enhances the prediction rate of SAP by an effective FS method namely RFBT-RF method. The RFBT-RF method reduces the redundancy and irrelevant features from the dataset and forwards the relevant features to the multiple classifiers. The proposed RFBT-RF method improves prediction performance of multiple classifiers, those are, DT, SVM, KNN, and NBC. The major contribution of the hybrid RFBT-RF method in SAP prediction is described below.

- Design an efficient FS method to avoid the miss prediction rate and analyse the SAP effectively.
- The proposed FS method helps to improve the multiple machine learning classifiers SAP forecasting performance.
- The Relief-F algorithm construct the large tree hence, prediction time complexity increases. To rectify this problem BT-RF algorithm based hybrid algorithm is used.

This paper is composed as follows. Section II presents a broad survey of several recent papers on SAP prediction analysis. In section III, an effective FS method namely RFBT-RF is presented for improve the SAP prediction. The section IV shows the comparative experimental result for proposed and existing SAP prediction approaches. The conclusion is made in Section V.

2. Literature review

Researchers suggested several techniques on the prediction of SAP based on machine learning techniques. In this scenario, a brief evaluation of some important contributions to the existing literatures are presented below.

A. Pardo, Feifei Han, and Robert A. Ellis, [9] highlighted self-regulated learning features performed efficiently with respect to SAP analysis. This result provided robust evidence of the advantages of combining self-reported and observed data sources to gain more precise insight of the learning experience leading to more effective overall improvements. The paper is based on online learning approach and the model didn't discuss the relation between teachers and their students for the purposes of learning.

R. C. Zhang, H. M. Lai, P. W. Cheng, and C. P. Chen, [10] presented Computer-based Graduated Prompting Assessment System (CGPAS) that is designed through feedback to support 2D graphing. This study includes three different contributions such as first, the TML-based assessment was developed and tracked because the assessment system has become the dominant mode for communicating with learners. Second, the study was student data which were examined across eight time periods to understand the influence of graduated prompting assessment on students' academic performance. Third, quasi-experiments were used to test the derived hypotheses. Due to the limitation of the CGPA, the system student's usage did not specify whether they were interrupted by other factors outside the system.

F. Al-Obeidat, A. Tubaishat, A. Dillon, and B. Shah, [11] used the data analytic technique for predicting student's performance by considering their past experience. They have implemented a hybrid classification technique which was a combination of fuzzy multi-criteria and DT classification. This approach used to identify the key factor of student's success/failure but, not adaptable for new student data.

S. Ikbal, A. Tamhane, B. Sengupta, M. Chetlur, S. Ghosh, and J. Appleton, [12] developed a model to make an early prediction of academic performance risks for students at various granularity levels. The main contribution of the paper was to develop a generic framework to predict academic performance risks for K-12 students at various granularity levels of the curriculum. The method was unable to handle the missing data and class imbalance problems which lead to poor accuracy.

C. Grunschel, M. Schwinger, R. Steinmayr, and S. Fries, [13] presented Motivational Regulation Strategies (MRS) that efficiently analyse the SAP. This approach observes the relationship between the MRS and SAP. The major benefit of MRS was positive indirect effects on SAP and academic performance. But, this approach not able to analyse the new SAP data because it analyse only historical data.

So, an appropriate FS methodology namely (RFBT-RF) is implemented that enhances the performance of SAP prediction based on different classifier and to overcome the above mentioned drawbacks.

3. Proposed methodology

The real time applications produced numerous data for analysing the desired model by considering SAP related data. In this research, the prediction of SAP by using a RFBT-RF algorithm has been made from a UCI and collected dataset. The FS method namely RFBT-RF, selects the relevant data and remove the least relevant or irrelevant features from the dataset, after that ranked data is filtered. Hence,

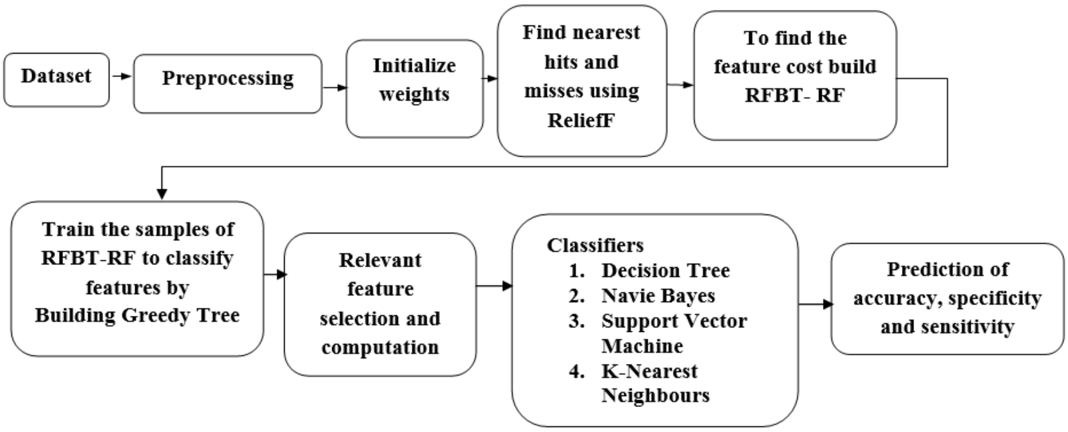
Dataset
Preprocessing
Initialize weights
Find nearest hits and misses using ReliefF
To find the feature cost build RFBT- RF
Train the samples of RFBT-RF to classify features by Building Greedy Tree
Relevant feature selection and computation
Classifiers
1. Decision Tree
2. Navie Bayes
3. Support Vector Machine
4. K-Nearest Neighbours
Prediction of accuracy, specificity and sensitivity

Figure.1 Basic architecture of the proposed method

proposed FS method improves the classification performance of different classifiers such as SVM, KNN, NBC, and DT. The proposed FS method architecture is shown in Fig. 1.

3.1 Data collection and pre-processing

An academic student's performance related data is taken from the UCI machine Learning Repository database. Also, researchers collected 10th standard student academic performance related data from four schools in the year of 2013-2017. The UCI (maths) database includes 395 instances, UCI (Portuguese) database includes 650 instances and collected database includes 4965 instances. The data attributes include school name, age, gender, travel time, distance from school to home, hobbies, health details and etc. were collected by using school reports and questionnaires. The student's performance categorized into two groups such as low performance and high performance. Here, input data files are converted string format to integer or float. Furthermore, these pre-processed values are forwarded to the FS process.

3.2 Feature selection using Relief-F and budget tree-random forest

The main responsibility of FS process is to control the size of the feature subset. At first, choose the original features subset without dropping the information. In next step, avoids the unrelated and redundant features for decreasing the dimensionality of the data. Accordingly, FS increases the mining accuracy, decreases the computation time and improves the result comprehensibility. In existing work, Relief-F algorithm is used for FS but it's not able to perform on incomplete and noisy data [14]. To overcome these issues, FS method namely RFBT-RF is used that reduce the irrelevant features and improve the prediction accuracy of SAP.

Consider, Relief F algorithm calculates a feature score for each attribute then it applies ranking order in that, but only high ranking features are selected for FS [15]. Alternatively, these scores may be applied as feature weights to guide downstream modelling. The features can be selected by first initializing their weights. The Eq. (1) represents the weight initialization,

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

Whereas, $C[A]$ is indicated as weight value of all attributes A. At each iteration, Relief-F algorithm consider the feature vector (x) belongs to one random instance, and the feature vectors of the instance closest to x (by Euclidean distance) from each class. The closest same-class instance is called 'near-hit', and the closest different-class instance is called 'near-miss'. After that, that algorithm find the hits and misses for each class in random instances that is mathematically calculated in Eq. (2),

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

After detection of hit and miss rates, the weight value is forwarded to the BT-RF algorithm. The major advantage of budget random forest algorithm is to minimize the prediction error. Random forest algorithm constructs a collection of trees, where each tree is grown by random independent data sampling & feature splitting, produce a collection of independent identically distributed trees. The feature cost is computed by the following Eq. (3),

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

Whereas, weight value of attributes A is $C[A]$ and depend on r_i value. The r_i is the randomly selected instance. But, the search of random instance then searches for k of its nearest neighbours from the same class, called nearest hits h_j. The goal of the random forest functions F that minimizes expected loss subject to a budget constraint is shown in Eq. (4).

$$f \underset{\in}{\min} F E_{xy}[L(y, f(x))], E_x[C(f,x)] \leq B \quad (4)$$

Whereas $L(y, \hat{y})$ is a loss function and $\hat{y} = f(x)$, $C(f,x)$ is the cost of evaluating the function of f on example x and B is a user specified budget constraint. In this paper, the feature acquisition cost $C(f,x)$ is a modular function applied to the $f(x)$ is acquiring each feature has a fixed constant cost. Then, minimize the empirical loss subject to a budget constraint is shown in Eq. (5),

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

In our context the classifier f is a random forest, T consisting of K random trees, $D1, D2, \ldots, DK$ are learnt on training data. Consequently, the expected cost for an instance x during prediction-time is written in Eq. (6),

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

The random forest tree algorithm equally distributes the RHS scale value with number of trees. The upper bound of the trees show the typical behaviour of the random forest because of low features correlation between the trees. The Greedy-Tree is mathematically shown in Eq. (7),

Compute

$$(t) := \underset{g_t \in G_t}{min} \underset{i \in outcomes}{max} \frac{c(t)}{F(S) - F\left(S_{g_t}^{i}\right)}, \quad (7)$$

Whereas, $S_{g_t}^{i}$is the set of examples in S that has outcome i using classifier g_t with feature importance t. GREEDYTREE helps to compute all the features simultaneously in RFBT-RF. The BUDGETRF iteratively builds decision trees by calling GREEDYTREE as a subroutine on a sampled subset of examples from the training data until the budget B is exceeded as evaluated using the validation data. The ensemble of trees then returned as output. As shown in subroutine GREEDYTREE, the tree building process is greedy and recursive. The pseudocode for proposed FS method is shown below.

3.2.1. Algorithm BUDGET RF

```
procedure $BUDGETRF(F, B, C, ytr, Xtr, ytv, Xtv)$
$T \leftarrow \Phi$
while Average cost using validation set on $T \leq B$ do
    Randomly sample $n$ training data with replacement to form $X^{(i)}$and $y^{(i)}$
    Train $T \leftarrow GREEDYTREE(F, C, y^{(i)}, X^{(i)})$
    $T \leftarrow T \cup T$
return $T/T$
rf.fit$(X^{(i)}, Y^{(i)})$
sorted(zip(map(lambda x: round(x, 4), rf.feature_importances_), names), reverse=True)
```

Subroutine – GREEDYTREE

```
procedure GREEDYTREE $(F, C, y, X)$
$S \leftarrow (y, X)$
if $F(S) = 0$then return
for each feature $t = 1$to $m$ do
    Compute $(t) := \underset{g_t \in G_t}{min} \underset{i \in outcomes}{max} \frac{c(t)}{F(S)-F(S_{g_t}^{i})}$, importance for feature $t$
    Where $S_{g_t}^{i}$is the set of examples in $S$that has outcome $i$ using classifier $g_t$ with feature $t$.
$\hat{t} \leftarrow argmin_t R(t)$
Compute $\hat{g} \leftarrow \underset{g_{\hat{t}} \in G_{\hat{t}}}{argmin} \underset{i \in outcomes}{max} \frac{c(\hat{t})}{F}(S) - F(S_{g_{\hat{t}}}^{i})$ // Feature importance from Eq. (7)
Make a node using feature $\hat{t}$and classifier $\hat{g}$
for each outcome $i$ of $\hat{g}$ do
    GREEDYTREE $(F, C, y_{\hat{g}}^{i}, X_{\hat{g}}^{i})$to append as child nodes.
```

3.3 Classification

After FS, the efficiency of the selected feature subset of the proposed method is evaluated by different classifiers: SVM, NB, DT, and KNN algorithms. All the other classification methods are taking much time to process the data but, the proposed algorithm significantly reduce the time to build the classification model. All different classifiers used in SAP process is described below.

- **Support Vector machine:** This classifier is a supervised learning method that use in both regression and classification work. Its search the nearest Support Vectors (SV) to determine the confinement in the training set and separate the new test vectors by using SV vectors [16].

- **Naïve Bayes Classifier:** The NBC predicts conditionally independent class of given class labels that are stated as problem of instances and then modifies the feature vector into feature values, where the class labels are drawn from some finite set. Moreover, NBC method identify the hidden information between subjects that affected the student performance [17].
- **K-Nearest Neighbour:** This algorithm initially calculates the distance between the data points and these points are closest to the training sets. When the number of training samples are less, the KNN classifier is no longer optimal. However, if training set includes more number of samples, then time complexity is high for similarity calculation [18].
- **Decision tree:** This algorithm classifies the instances by sorting from root node to leaf node, which gives the classification of a specific instance. Each node in the tree denotes a test of some instances and every branch descending from node corresponds to one of the possible values for attribute. This algorithm is simple, fast and easy to comprehend [19].

In this study, four machine learning classifiers are used for SAP. The RFBT-RF method is proposed for improve the prediction performance of classifiers. The RFBT-RF method is applied in the above mentioned 4 classifiers but, DT algorithm shows better SAP prediction.

3.3.1. Decision tree

The DT algorithm is the tree like structure, in that every internal node indicates the “test” on an attribute. Each branch of the tree indicates the test outcomes, leaf node indicates the class label and path from root to leaf indicates the classification rules. An Iterative Dichotomiser 3 (ID3) is the common DT algorithm that employs greedy algorithm for selecting the best attribute to split the dataset on each iteration. An ID3 algorithm uses different attributes in decision trees construction. It effectively generates the trees in four steps.

- Compute entropy function to dataset.
- For each attributes/features three estimation process is required such as (i) entropy calculation, (ii) consider average entropy value for current attribute, and (iii) estimation of gain for current attribute.
- Consider highest gain attribute
- Iteration repeated until desired tree construction.

The following Eq. (12) effectively exhibits the entropy function.

$$Entropy\ (E) = \sum_{i=1}^{k} -p_i \log_2 p_i \qquad (12)$$

Here, E represents the decision or rule and K corresponds to the number of output variable classes, and p_i the possibility of the class i. In this algorithm the quality of the split is characterized by the information gain is shown in Eq. (13),

$$Gain(E, A) = Entropy(E) - \sum_{v \in values(A)} \frac{|E_v|}{E} Entropy(E_v) \qquad (13)$$

Values (A) characterize the possible values of attribute A, E_v indicates the subdivision of rule, E which consists of value v in E. The Entropy (E) effectively evaluates the entropy of an input attribute A which contains k categories, Entropy $E_v$represents the entropy of an attributes category with regard to the output attribute, and $\frac{|E_v|}{E}$denotes the probability of the j^{th} category in the attribute. The difference between the entropy of the node and an attribute represents the information gain of an attribute. Hence, FS based RFBT-RF method reduces the feature space dimension, remove irrelevant features and improves the classifier’s performance. The experimental outcomes of RFBT-RF with different classifiers are shown in the following section.

4. Experimental result and discussion

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 LR based FS methods [20]. In experimental analysis, 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.

4.1 Performance measure

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. (17) and (18).

$$Precision = \frac{TP}{TP+FP} \quad (17)$$

$$Recall = \frac{TP}{TP+FN} \quad (18)$$

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. (19)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (19)$$

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. (20).

$$F - Score = 2.\frac{Precision.Recall}{Precision+Recall} \times 100 \quad (20)$$

4.2 Analysis of existing and proposed feature selection method performance

In this study, four classification models were generated such as SVM, KNN, NBC, and DT. The performance of these classifiers were evaluated on the test data in terms of accuracy, precision, recall and f-score. According to the result, the proposed RFBT-RF method helps to improve the classification result and efficiently classifies the student's performance. The RFBT-RF algorithm was then applied to the different classifiers such as SVM, KNN, NBC and DT. Also, the input data was taken from the collected school database. Table 1 shows performance of different classifiers.

4.3 Analysis of existing UCI and proposed database performance

In this experimental research, efficiency of RFBT-RF based SAP prediction is compared with the existing LR model with respect to three databases like UCI (maths, Portuguese) and School database. In Table 2, precision, recall, accuracy, TP, TN, and F-score value of proposed methodologies is compared with the two class's namely low and high performance. Moreover, the different classifiers are used to predict the SAP such as SVM, NBC, KNN, and DT. In UCI (maths) database, the maximum precision value of LR and RFBT-RF method achieved approximately 69% and 89% respectively. The LR and RFBT-RF method achieved approximately 62% and 89% of maximum recall. Additionally, both the method achieved 62% and 88.60% of maximum accuracy with respect to SVM and NBC classifier respectively. The maximum accuracy performance of LR and RFBT-RF method achieved approximately 62% and 88.60%. In the Table 2, performance of selected features is addressed. In existing work, LR method was used for FS but several features were repeated. The proposed RFBT-RF algorithm reduces the redundancy of the feature and irrelevant features.

In UCI (Portuguese) database, SVM, NBC, and KNN classifiers achieved approximately 66.69% of accuracy with respect to LR method. The DT classifier achieved approximately 69.23% of accuracy. In Collected school database, LR and RFBT-RF method both achieved 92% and 98% of precision and recall respectively with respect to DT algorithm. The accuracy of LR and RFBT-RF method is achieved 91.03% and 97.88% respectively. The F-score performance is 91% and 98% respectively. The graphical representation of UCI (maths) and UCI (Portuguese) database performance is shown in Fig. 2 and Fig. 3.

The Fig. 2 and Fig. 3 represents the proposed RFBT-RF and existing LR method's performance in UCI (maths) and UCI (Portuguese) with different classifiers. Compared to the LR method, the proposed RFBT-RF method shows better results in all classifiers. The UCI (maths) database employs minimum instances approximately 395 instances and UCI (Portuguese) database employs maximum instances approximately 650 instances.

Table 1. The RFBT-RF feature selection method using different classifiers Performance analysis

Evaluation Parameter	Classifiers			
	SVM	KNN	NBC	DT
TP	639	639	576	633
FN	0	3	338	349
Precision	41	77	93	98
Recall	64	65	92	98
Accuracy	64.35	64.65	64.65	97.88
F-Score	50	51	92	98

Table 2. Performance analysis of existing and proposed feature selection method with different classifiers

Feature Selection Method	Classifiers	Parameters						Selected Features
		TP	TN	Precision	Recall	Accuracy	F-Score	
UCI Database (Maths)								
Logistic regression [20]	SVM	1	48	52	62	62.05	57	sex,sex,sex,Fedu,sex,Pstatus,sex,Mjob,sex,studytime,age,reason,Medu,sex,guardian,sex,Fjob,famsize,address,traveltime,sex,sex
	NBC	24	22	69	58	58.27	58	
	KNN	6	39	53	57	56.96	54	
	DT	8	32	49	49	48.10	49	
UCI Database (Portuguese)								
Logistic Regression [20]	SVM	0	8	46	68	67.69	55	sex, sex, traveltime, sex,sex, address, age, studytime, sex, sex, sex, famsize, Mjob, sex, guardian.
	NBC	24	64	69	68	67.69	68	
	KNN	5	83	63	68	67.69	60	
	DT	13	77	66	68	69.23	65	
UCI Database (Maths)								
Proposed RFBT-RF	SVM	15	48	81	80	79.74	78	internet, higher, Fjob, Pstatus, nursery, activities, famsup, sex,Mjob, famsize,address, schoolsup, Medu,Fedu, age, traveltime, paid,reason, Guardian,failures, studytime
	NBC	25	45	89	89	88.60	89	
	KNN	12	44	70	71	70.88	69	
	DT	20	47	85	64	86.70	83	
UCI Database (Portuguese)								
Proposed RFBT-RF	SVM	0	87	46	67	66.92	54	freetime, famrel, school, romantic, guardian, higher, studytime, famsup, internet, age, nursery, Medu, Fedu, paid, activities, Mjob, Fjob, address, Pstatus, schoolsup, famsize, reason, failures, sex, traveltime.
	NBC	37	69	84	82	81.53	82	
	KNN	3	35	66	68	67.69	59	
	DT	24	30	79	80	80.0	79	
Collected School Database								
Logistic Regression [20]	SVM	581	279	87	87	86.06	87	healthproblem, Reason_to_choose_school, family_support, Medu, age, famsize, Mjob, studytime, age, age, age, age, hostel, age, Pstatus.
	NBC	569	342	93	92	91.74	92	
	KNN	569	247	82	82	82.17	82	
	DT	608	296	92	92	91.03	91	
Proposed RFBT-RF	SVM	639	0	41	64	64.35	50	Alchoholic, activities, SOC, HIN, SA-1,SCI,T_feedback, studytime, TEL,Soc_feedback, ENG,M_feedback,MAT, E_feedback,Student_ID,attendence,Sci_feedback,year,Mjob,Medu,Fedu,Distance_from_home_to_school,age,goout,H_feedback, E_Exp,Fjob,Sci_Exp, Reason_to_choose_school, Soc_Exp,traveltime,Tel_Exp, fee_range.
	NBC	576	338	93	92	92.04	92	
	KNN	639	3	77	65	64.65	51	
	DT	633	349	98	98	97.88	98	

According to the Fig. 4, the performance of the collected database in terms of efficient parameters. In Collected database, the RFBT-RF method achieved approximately 97.88% of prediction accuracy in DT classifier. Additionally, compare to the LR method, proposed RFBT-RF method improved the prediction accuracy of all classifier's results efficiently. The table 3 shows the comparative study of the existing and proposed SAP research works.

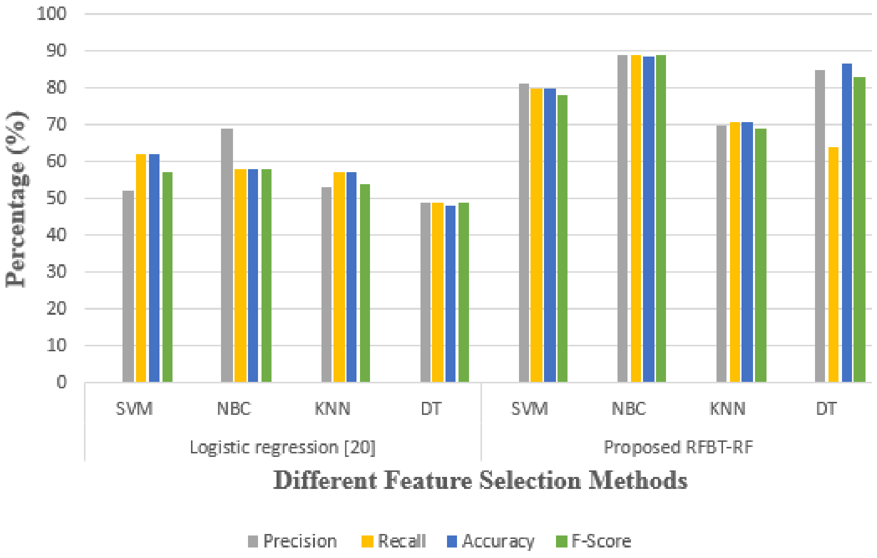

Fig.2 UCI (Maths) database performance

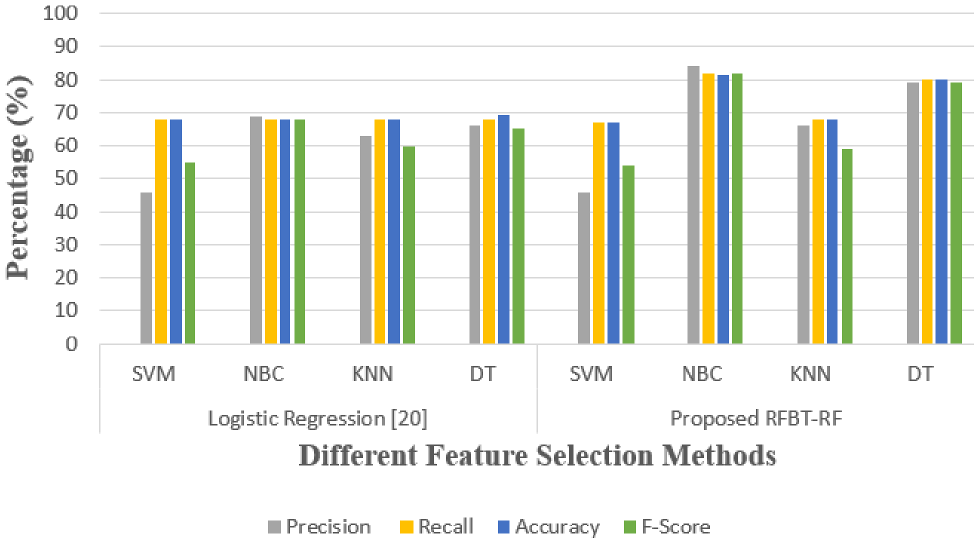

Fig.3 UCI (Portuguese) database performance

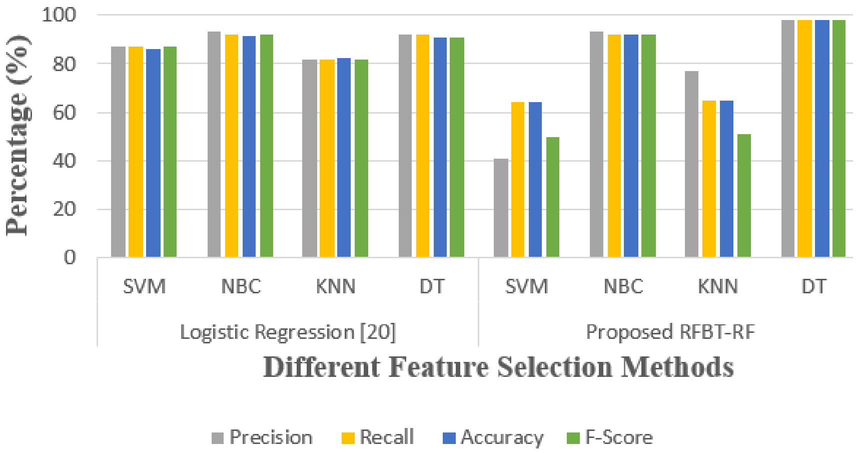

Fig.4 Collected database performance

Table 3. Comparative study of proposed and existing work of student performance prediction in academic area

Methodologies	Database	Precision (%)	Recall (%)	Accuracy (%)
Probabilistic neural network [20]	Collected for 682 first-year freshman students at a case study public urban university from 2010 to 2011	96.7	87.0	94.7
Naïve Bayesian [21]	Ministry of education in Gaza Strip for 2015	94.65	93.6	93.17
KNN [21]		63.4	62.9	63.45
Decision Tree [22]	Dataset of 240 samples collected randomly through survey at university located at India	78.9	96.4	92.5
Proposed (RFBT-RF)	UCI (maths)	89	89	88.60
	UCI (Portuguese)	84	82	81.53
	School Database	98	98	97.88

5. Conclusion

Data mining techniques are applied to higher education more and more to give insights about educational and administrative problems in order to increase the managerial effectiveness. In this paper, an effective RFBT-RF feature selection method is proposed to reduce the irrelevant features and improve the prediction rate of student's performance. The RFBT-RF method improves the classifiers performance in SAP prediction. In experimental analysis, three databases are used for SAP such as UCI (maths), UCI (Portuguese) and collected database. These databases results are compared with the different FS based LR and RFBT-RF methods with different classifiers such as SVM, KNN, NBC, and DT in terms of precision, recall, accuracy and f-score. The proposed RFBT-RF method achieved 81.53% of accuracy in UCI (Portuguese), 88.60 % of accuracy in UCI (maths) and 97.88% of accuracy in collected database. In future, this work can be extended by improving the optimal selection of attribute based on the correlation among them to identify student's academic performance with an efficient classification technique.

References

[1] J. Xu, K. H. Moon, and M. V. D. Schaar, "A machine learning approach for tracking and predicting student performance in degree programs", *IEEE Journal of Selected Topics in Signal Processing*, Vol.11, No.5, pp.742-753, 2017.

[2] D. Lambić, "Correlation between Facebook use for educational purposes and academic performance of students", *Computers in Human Behavior*, Vol. 61, pp.313-320, 2016.

[3] A. Krasilnikov, and A. Smirnova, "Online social adaptation of first-year students and their academic performance", *Computers & Education*, Vol. 113, pp.327-338, 2017.

[4] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Van Erven, "Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil", *Journal of Business Research*, In Press, 2018.

[5] H. Muhonen, E. Pakarinen, A. H. Poikkeus, M. K. Lerkkanen, and H. Rasku-Puttonen, "Quality of educational dialogue and association with students' academic performance", *Learning and Instruction*, Vol.55, pp.67-79, 2017.

[6] C. Mejia, B. Florian, R. Vatrapu, S. Bull, S. Gomez, and R. Fabregat, "A novel web-based approach for visualization and inspection of reading difficulties on university students", *IEEE Transactions on Learning Technologies*, Vol.10, No.1, pp.53-67, 2017.

[7] E. Alwagait, B. Shahzad, and S. Alim, "Impact of social media usage on students' academic performance in Saudi Arabia", *Computers in Human Behavior*, Vol.51, pp.1092-1097, 2015.

[8] M. Goga, S. Kuyoro, and N. Goga, "A recommender for improving the student academic performance", *Procedia-Social and Behavioral Sciences*, Vol.180, pp.1481-1488, 2015.

[9] A. Pardo, F. Han, and R.A. Ellis, "Combining university student self-regulated learning indicators and engagement with online learning events to predict academic performance", *IEEE Transactions on Learning Technologies*, Vol.10, No.1, pp.82-92, 2017.

[10] R. C. Zhang, H. M. Lai, P. W. Cheng, and C. P. Chen, "Longitudinal effect of a computer-based graduated prompting assessment on students' academic performance", *Computers & Education*, Vol.110, pp.181-194, 2017.

[11] F. Al-Obeidat, A. Tubaishat, A. Dillon, and B. Shah, "Analyzing students' performance using

multi-criteria classification", *Cluster Computing*, pp.1-10, 2017.

[12] S. Ikbal, A. Tamhane, B. Sengupta, M. Chetlur, S. Ghosh, and J. Appleton, "On early prediction of risks in academic performance for students", *IBM Journal of Research and Development*, Vol. 59, No. 6, pp.5-1, 2015.

[13] C. Grunschel, M. Schwinger, R. Steinmayr, and S. Fries, "Effects of using motivational regulation strategies on students' academic procrastination, academic performance, and well-being", *Learning and Individual Differences*, Vol. 49, pp.162-170, 2016.

[14] R.P.L. Durgabai, and Y.R. Bhushan, "Feature selection using ReliefF algorithm", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 3, No.10, pp.8215-8218, 2014.

[15] T.R. Reddy, B.V. Vardhan, M. GopiChand, and K. Karunakar, "Gender Prediction in Author Profiling Using ReliefF Feature Selection Algorithm", In: *Proc. of International Conf. on Intelligent Engineering Informatics*, pp.169-176, 2018.

[16] G. Pratiyush and S. Manu, "Classifying Educational Data Using Support Vector Machines: A Supervised Data Mining Technique", *Indian Journal of Science and Technology*, Vol.9, No.34, pp.1-5, 2016.

[17] M. Makhtar, H. Nawang, W. Shamsuddin, and S. Nor, "Analysis on Students Performance Using Naïve Bayes Classifier", *Journal of Theoretical & Applied Information Technology*, Vol.95, No.16, pp.3993-4000, 2017.

[18] T. Anderson and R. Anderson, "Applications of Machine Learning to Student Grade Prediction in Quantitative Business Courses", *Global Journal of Business Pedagogy*, Vol.1, No.3, p.13, 2017.

[19] A. Nichat and A. Raut, "Predicting and Analysis of Student Performance Using Decision Tree Technique", *International Journal of Innovative Research in Computer and Communication Engineering*, Vol.5, No.4, pp.7319-7327, 2017.

[20] C. Mason, J. Twomey, D. Wright, and L. Whitman, "Predicting engineering student attrition risk using a probabilistic neural network and comparing results with a backpropagation neural network and logistic regression", *Research in Higher Education*, Vol.59, No.3, pp.382-400, 2018.

[21] I.A.A. Amra and A.Y. Maghari, "Students performance prediction using KNN and Naïve Bayesian", In: *Proc. of the 8th International Conference on Information Technology,* pp.909-913, 2017.

[22] S. Sivakumar, S. Venkataraman, and R. Selvaraj, "Predictive modeling of student dropout indicators in educational data mining using improved decision tree", *Indian Journal of Science and Technology*, Vol.9, No.4, pp.1-5, 2016.