Int. J. Nonlinear Anal. Appl. 13 (2022) No. 1, 1239–1252
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2021.24724.2808

Analysis of performance of accuracy by adding new features individually using Relief-F and Budget Tree Random Forest (RFBTRF) method

K.Deepika[a,*], M. Sowjanya Reddy[b], N. Rajesh Pandian[c], R. Dinesh Kumar[b]

[a] *Department of Information Technology Kakatiya Institute of technology and science, Warangal- 506015, Telangana .*
[b] *Department of Computer Science and Engineering, Siddhartha Institute of Technology and Sciences, Narapally, Hyderabad, Telangana 500088*
[c] *Department of Computer Science and Engineering, PSNA College of Engineering and Technology Kothandaraman Nagar, Dindigul-624622 TamilNadu, India.*

(Communicated by Madjid Eshaghi Gordji)

Abstract

The education is very important for improving the values of students in the society. Different types of features like school related features, student related features, parent related features and teacher related features are influencing the success rate of students in their education. Identification of best features from the huge set of features for analyzing the success or failure of a student is one important challenge to the research community and academicians. The set of features information is collected for preparing the student dataset also one difficult task in the prediction of student academic performance. We collected a student dataset of different schools that contains 4965 student's information. The dataset contains information of 45 features of different categories such as school related features, student related features, parent related features and teacher related features. All features are not useful for predicting the academic performance of a student. The Data mining methods are applied in various research domains including education to extract hidden information from datasets. The feature selection algorithms are used to determine the best informative features by eliminating the irrelevant and redundant features. In this work, Relief-F Budget Tree Random Forest feature selection algorithm is used to identify the relevant features in the collected school dataset. Five different machine learning models are used to predict the efficiency of feature selection

*Corresponding author
Email addresses: deepika.srimanu@gmail.com (Dr.K.Deepika), souji.vj1@gmail.com (M.Sowjanya Reddy), nrajeshpandian@gmail.com (N. Rajesh Pandian), me.dineshkumar@gmail.com (Dr.R.Dinesh Kumar)

Received: June 2021 *Accepted:* October 2021

algorithm. The decision tree model shows best accuracy for student academic performance prediction compared with other models. The experimental results display that the RFBTRF algorithm identifies the best informative features for enhancing the accuracy of student academic performance prediction and also reduces the over-fitting issues. The experiment started with individual features and then continued with combination of different categories of features. It was observed that the accuracy of student academic performance prediction is decreased when some categories of features are added to other categories of features.

Keywords: Student Academic Performance Prediction, Machine Learning Models, Educational Data Mining, Feature Selection Algorithms.

1. Introduction

Education is very important for the society to improve knowledge about surroundings, ethical values and job skills in the humans. Different factors are influencing the success and failure rate of students in their education. The academicians are struggling a lot across the world to minimize the student failures. They follow different types of strategies like counselling, remedial classes, additional tests and assignments to improve the performance of a student. Even though, every year student failure rates, poor performances and college dropout numbers perturb academicians and parents. Most of the studies suggest that the performance of a student is affected by different categories of features such as social, economic, demographic and individual attributes. Specifically, some of the features like attendance percentage, health, social background, parents' occupation, parental support, alcohol consumption and internet usage hours are influencing more on student performance. Educational Data Mining (EDM) is one research domain that implements data mining and machine learning techniques in education to improve the performance of a student. The data mining techniques are helpful to identify hidden patterns in the student dataset and these patterns are extremely useful and interesting to academicians to develop methods for enhancing student academic performance. EDM is used for analysis and visualisation of student data, building recommendation systems for learners, constructing curriculum, detecting behaviours of learners, E-student modelling, student performance prediction and in several other areas. EDM consists of an amalgamation of techniques to analyse and improve the science of the teaching-learning processes. In these days, High dimensionality in case of more features is a bigger problem in machine learning approaches [15]. Several researchers do experiments on this problem to lessen the features count in the experiment. Researchers used statistical methods to avoid redundant data and to reduce noise data. Nevertheless, all features are not used to train the model. The feature selection techniques play major role to enhance the efficiency of a model by using with relevant and non-redundant features. Education is a necessary and complex process that affected by several factors such as policies of education, economic prosperity of nation etc. [16]. The accurate estimations of academic performance of a student at early times of degree programme help a lot to identify the weak students and enable the academicians to take the necessary actions to avoid the students from failure [13]. Many researchers developed different machine learning techniques for better analysis of knowledge of students. Most of these techniques shows good estimation in accurate prediction of Student Academic Performance (SAP) in future as well as estimate the effective student features [3]. Majorly, there are four important reasons such as avoid the curse of dimensionality, reduce the over-fitting problem by improving generalization, decrease the training time and reduce the number of features for the importance of feature selection techniques. In the field of data analysis and processing, the datasets are described with huge number of features or attributes that are used to estimate the usability and applicability

of the data [11]. The classification algorithms face a challenge of paying attention to data imbalance problems [5]. Several factors or features are influencing the student academic performance, but it is very difficult to identify which set of features impact is more in improving the academic performance of a student. In this aspect, the feature selection algorithms are used to identify which set of features are influencing more the performance of student. The feature selection algorithms play an important role in the recognition of relevant features for predicting the student academic performance. The feature selection algorithms are majorly divided into three classes such filter based, wrapper based and embedded based feature selection algorithms. The filter based feature selection algorithms identify the important features by computing the score of features. The score of a feature is high means the feature is more relevant for the class. The filter based methods are not using any machine learning models involvement to determine the importance of a feature. The wrapper based methods divides the features into different subsets. The machine learning models are used by the wrapper based methods to identify the important feature subset from a set of subsets. The embedded based methods used machine learning models directly to identify the important features from a set of features. The wrapper based and embedded based methods used machine learning models to identify the important features, whereas the filter based methods are not using any machine learning models to identify the importance of a feature. In this work, a filter based feature selection method named as Relief-F Budget Tree Random Forest Tree (RFBTRF) algorithm [7] is used to identify the informative features This paper is planned in six sections. The section 2 explains the analysis of existing works in student performance prediction. The section 3 presents the description about the collected student dataset used in this experiment. The section 4 explains the RFBTRF feature selection algorithm. The experimental results for student academic performance prediction are displayed in section 5. The conclusions to this work are described in section 6.

2. Literature Survey

The data mining techniques are used by several researchers in EDM to predict the academic performance of a student. Concepción Burgos et al., implemented [4] data mining methods to the dataset of student history for predicting the dropout students of a course. They applied logistic regression model for classification of data and filter irrelevant features. The proposed method performance was tested on various distance learning courses data of Madrid Open University that are registered by 100 students. The authors observed that the dropout reduced by 14% with the proposed method when compared with previous academic years. Eduardo Fernandes et al., proposed [8] a prediction method and applied this method on the dataset collected from public school of the Federal District of Brazil. The proposed method performance is tested by using two datasets and the method is applied for analysing the data. Gradient Boosting Machine (GBM) classifier is used for predicting the performance of a student. The experimental results show that the school and residence of a student are two factors influencing more the student academic performance. They observed that the requirement of different factors for improving the performance of student. Raheela Asif, et al., concentrated [2] on two tasks such as predicting the performance of a student at the end of academic programme of four-year and possible progressions of students. The dataset contains two classes such high and low achieving students. They used decision tree for feature selection and applied k-means algorithm for classification. The proposed method is tested on the dataset for predicting the performance of a student. They observed that the experimental results of proposed method show good performance. Sumyea Helal et al., established [10] a subgroup discovery for extracting the important aspects related to the outcome of student performance. The developed method used different types of information such as student academic data, course data, demographics data and data extracted

from learning management system of institution to investigate the performance of a system. The experimental results show that the subgroup discovery identifies the factors effectively. The developed method performance is analysed by using Moodle data. Feras Al-Obeidat et al., developed [1] a hybrid technique by combining fuzzy multi-criteria classification and decision tree to predict the academic performance of a student. The performance of a system is analysed by using various factors related to family size, age and school. The developed technique performance is tested by using two datasets such as UCI Math and UCI Portuguese datasets. The proposed method performance was analysed and compared with existing methods by using various standard classifiers. The feature selection algorithms were used to improve the system performance. Costa et al., developed [6] a fine-tuning method for predicting the performance of s student and applied data pre-processing methods. The combination of support vector machine, fine-tuning method and data pre-processing techniques attained higher efficiency than other standard methods. In data pre-processing method, SMOTE technique is used for balancing the data in the dataset and Information Gain algorithms is applied for dimensionality reduction. The parameters of SVM are fine-tuned by using Grid-Search method. The Grey Wolf Optimization technique was used as a feature selection method to enhance the efficiency of the proposed method. Li et al., solved [12] the class noise and imbalance problems in the dataset by proposing cost function based randomized learning algorithm. The randomized learning algorithm developed by using weight least square problem. The performance of proposed method is compared with existing standard methods for predicting the performance of student. The experimental results show that the proposed method solved the problems of label errors and data imbalance problems. Son and Fujita developed [9] an adaptive fuzzy technique to enhance the prediction accuracy of student performance. The proposed technique processes both local and global learning. In local learning, the settings of parameters in fuzzy technique were done by using hybrid method. In global learning, the random subsets were selected for training. In the parameter learning, the hybrid method is the combination of Particle Swarm Optimization and Gradient descent methods. The proposed method performance is validated by using different datasets of UCI and compared with various standard methods. The performance of learning rate and classification are needed to be enhanced for attaining best efficiency from the model. Radwan and Cataltepea solved [14] the problem of class imbalance in the dataset by applying two noise reduction methods. The combination of threshold method and over-sampling technique of SMOTE were used in the proposed method to choose the best boundary among classes and for balancing the training dataset. The proposed method is validated by using a dataset of UCI Portuguese dataset. The experimental results show that the developed method shows good performance in the process of noise reduction. They observed that the parameter selection methods enhanced the predictive performance of student.

3. Description about Dataset

The dataset is collected from different schools that contain 4965 student's information. Each student information is described with four categories of features such as student features, parent features, school features and teacher features. The table 1 shows the student features. The set of Features considered in this experiment are

- Student features: Consist of Nine Attributes like Student id, age, Distance from home to school, health problem, year, Attendance, Hostel, Reason to choose the school, goout time.
- Parent Features: Consist of Nine Attributes like Family size, Pstatus, fedu, medu, mjob, fjob, guardian, famsup, pAlchoholic.

Table 1: Student Features

Feature Category	Name of a Feature	Description
Student Features	School Name	Five School Names (GB, SA, WP, SP, LL)
	Student_Id	Numeric
	Stud_Name	Student Name Characters
	Section	Section Name
	Gender	F – Female M Male
	Age	Numeric
	Address	U – Urban R - Rural
	Distance from Home to School Health Problem	How Many Kilometers Health Problem for Student (Yes / No)
	Year	6 Years Data (2012, 2013, 2014, 2015, 2016, 2017)
	Attendance	For 200 Classes
	Hostel	Yes / No
	Reason	Reason to choose this School : Close to Home School to Home Course Preferences Other
	Go out	Going out with friends (1 – very low to 5 – very high)

- School Features: Consist of six Attributes like Traveltime, study time, schoolsup, feerange, Activities, Have Qualified teacher.
- Teacher Features: Consist of 18 features like TelExp, Tfeedback, Tmarks , HinExp, Hfeedback, Hinmarks, EngExp, Engfeedback, Engmarks, mathExp, Mfeedback, Mathmarks, SciExp, Scifeedback, Scimarks, SocExp, Socfeedback, Socmarks.
- Considered SA-1, SA-2, SA-3 Marks and Grade as Class label for predicting the Performance of Students.

4. RFBTRF Feature Selection Method

The main aim of this work is predicting student academic performance using only a few features for each student. The prediction was done by using classification algorithms. The feature selection algorithms are helped to find the required set of features for analysis. Student academic performance was predicted by using each set of selected features. Based on these results, a small student feature set was passed as input for the classification algorithms to give the best prediction results. In this work, a feature selection algorithm named as RFBTRF method is used to identify best informative features for predicting student academic performance. The Fig. 1 shows the RFBTRF method.

The steps in the RFBTRF method are depicted in figure 2.

Table 2: Parent Features

Feature Category	Name of a Feature	Description
Parent Features	Family SIze	Family Size LT3 – Less than three GT3 – Greater than three
	Pstatus	Parent Cohabitation Status T – Living Together A - Apart
	Fedu	Father Education 0 – None 1 – Primary Education (upto 4th Class) 2 – 5th to 9th Class 3 – Secondary Education (upto Inter) 4 – Higher Education
	Medu	Mother Education Same as Fedu
	Mjob	Mother Job Teacher Healthcare Related Govt Services At Home Other
	Fjob	Father Job same as Mjob
	Guardian	Students Guardian (Mother, Father, Other)
	Famsup	Family Educational Support (Yes or No)
	Palcoholic	Parent Alcoholic Consumption Rating (1 – low, 2 – Medium, 3 – High)

Table 3: shows the school related features

Feature Category	Name of a Feature	Description
School Features	Travel Time	Number of Minutes or Hours
	Study Time	Weekly Study Time 1 – Less than 2 Hours 2 – 2 to 5 Hours 3 – 5 to 10 Hours 4 – 10 Hours
	Schoolsup	School Support - whether school gives extra educational support (Yes or No)
	Fee Range	Below 20000 – 1 Above 20000 - 2
	Activities	Conduct curricular activities
	Have Qualified Teachers	Yes / No

Table 4: Teacher Features

Feature Category	Name of a Feature	Description
Teacher Features	Tel_Teacher	Telugu Teacher Name
	Tel_Exp	Telugu Teacher Experience
	T_Quali	Telugu Teacher Qualification
	T_feedback	Telugu Teacher Interaction with student / Feedback Rating (1 – Very bad, 2 – Bad, 3 – Good, 4 – Excellent)
	Hindi, English, Maths, Science, Social	Hindi, English, Maths, Science, Social Teachers Name, Experience, Qualification, Feedback
	Subject Marks	Telugu, Hindi, English, Maths, Science, Social Studies marks
	SA - 1	First Prefinal Exam Grade Marks
	SA – 2	Second Prefinal Exam Grade Marks
	SA - 3	Final Grade Marks
	Grade Range (0 – 10)	9.6 – 10 | 8.6 – 9.5 | 7.6 - 8.5 | 6.6 – 7.5 | 5.6 - 6.5 | 4.6 – 5.5 | 3.6 - 4.5 A1 | A2 | B1 | B2 | C1 | C2 | D

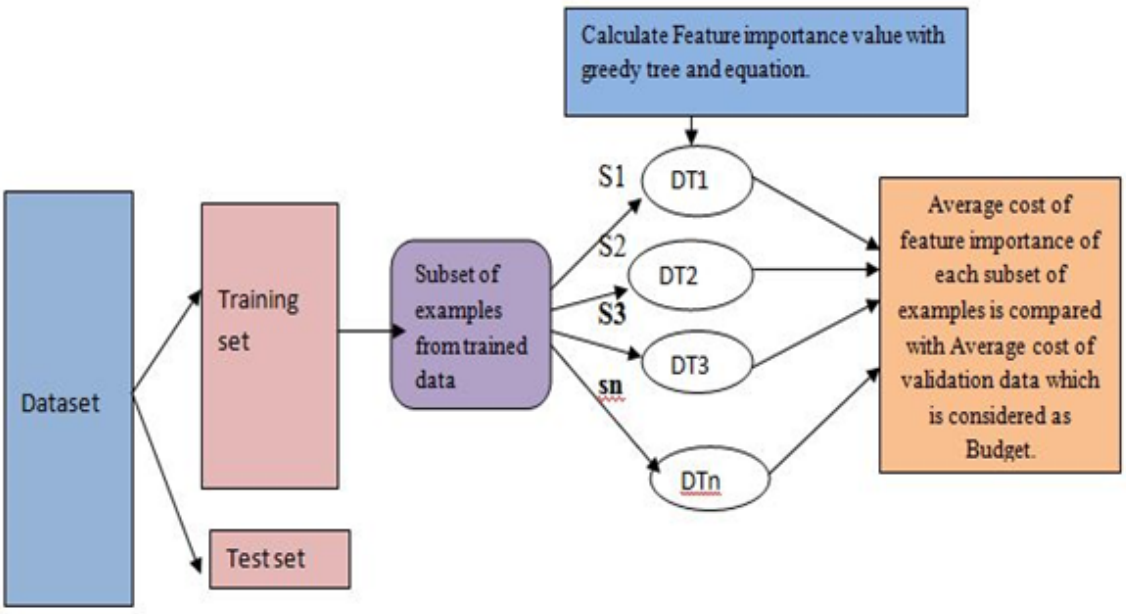

Figure 1: RFBTRF Feature selection method

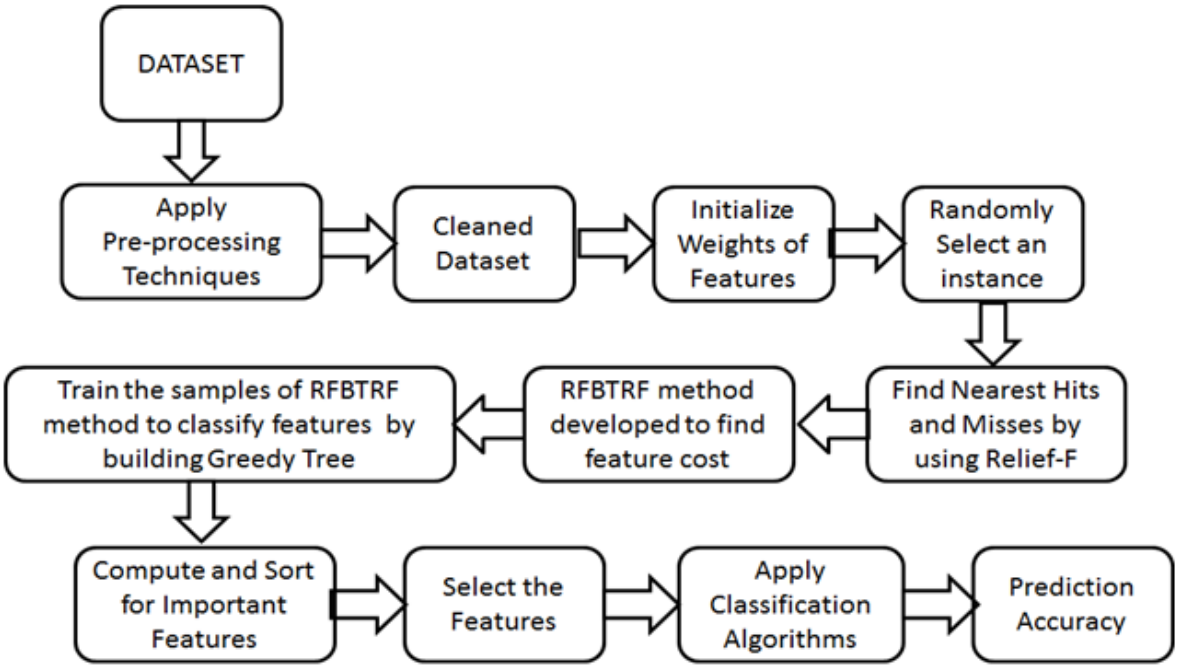

Figure 2: Basic structure of the Relief-F Budget Random Forest Feature Selection Method

RFBTRF Algorithm steps

1. Initially Apply Relief-f method.
2. Randomly select an instance and Calculate hits and misses

$$C_i = C_i - (x_i - nearHit_i)^2 + (x_i - nearMiss_i)^2 \tag{4.1}$$

3. Based on this hits and misses calculate cost of feature.

$$C\,[A] = C\,[A] - \frac{\sum_{j=1}^{k} diff(A, r_i, h_j)}{(m, k)} \tag{4.2}$$

Whereas C[A], is indicated as weight value of all attributes A and depend on value ri. The ri is the randomly selected instance. The random instance searches for k of its nearest neighbours from the same class, called nearest hits hj.

4. After calculating the cost of features, the weighted values are forwarded to Relief-F Budget tree Random Forest Algorithm, Which iteratively builds the decision trees with low acquisition cost by greedy tree, where each tree is grown by random independent data sampling and feature splitting, by information gain that produce a collection of independent identically distributed tree.
5. Calculate the Budget based on average cost of validation set.
6. The samples of RFBTRF are trained by greedy tree function for classifying the features.
7. The greedy tree calculates the feature importance value, based on the cost value and feature set that lie within the budgeted feature dimension.

$$computer(t) = \overset{min}{g_t \epsilon G_t} i \epsilon \overset{max}{outcomes} \frac{c(t)}{F(s) - F(\mathrm{S}^i_{gt})} \tag{4.3}$$

$$computer(t) = \frac{Cost\ of\ features}{feature\ selected - set\ of\ examples\ for\ previously\ trained\ samples\ average} \tag{4.4}$$

Table 5: The accuracies of SAP prediction when student features are used

Classifiers	Accuracy
KNN	64.45
ANN	64.35
NB	85.90
DT	92.34
SVM	64.35

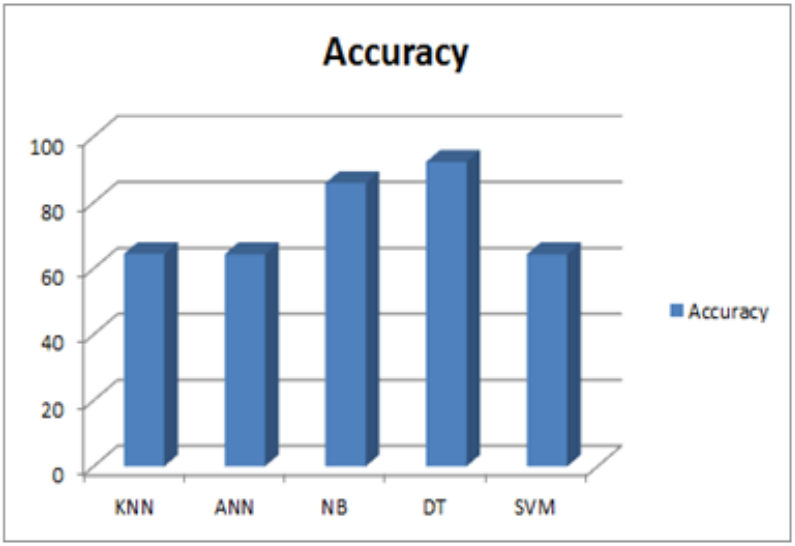

Figure 3: The accuracies of different classifiers for SAP prediction when student features are used

Whereas S^i_{gt}, is the set of examples in S that has outcome i using classifier with feature importance t. Features that are less than the budget are eliminated.
8. Select the features.
9. Apply classifiers for prediction of accuracy.

5. Experimental Results

The experiment is conducted for student performance prediction. Five different classifiers are used in the experiment to predict the accuracy of student performance prediction when different sets of features are used. The Table 1 shows the experimental results of student academic performance prediction when student features are used to represent the vectors.

In Fig. 3, the DT classifier attained best accuracy of 92.34% for student academic performance prediction. The DT classifier shows best performance for SAP prediction when compared with other classifiers. The Table 2 shows the experimental results of student academic performance prediction when school features are used to represent the vectors.

In Fig. 4, the DT classifier attained best accuracy of 96.57% for student academic performance prediction. The DT classifier shows best performance for SAP prediction when compared with other classifiers. The Table 3 shows the experimental results of student academic performance prediction when Teacher features are used to represent the vectors.

Table 6: The accuracies of SAP prediction when school features are used

Classifiers	Accuracy
KNN	94.86
ANN	92.34
NB	92.04
DT	96.57
SVM	93.85

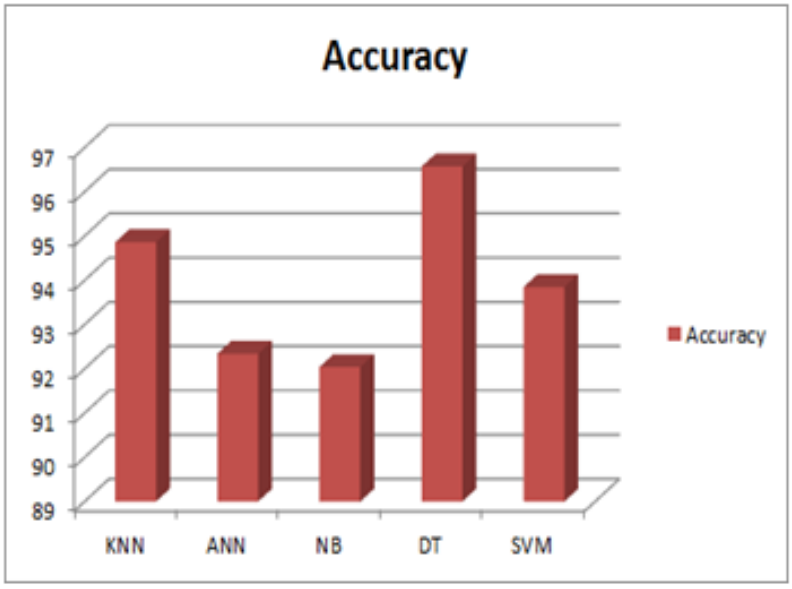

Figure 4: The accuracies of different classifiers for SAP prediction when school features are used

Table 7: The accuracies of SAP prediction when Teacher features are used

Classifiers	Accuracy
KNN	92.95
ANN	80.26
NB	93.55
DT	96.77
SVM	70.09

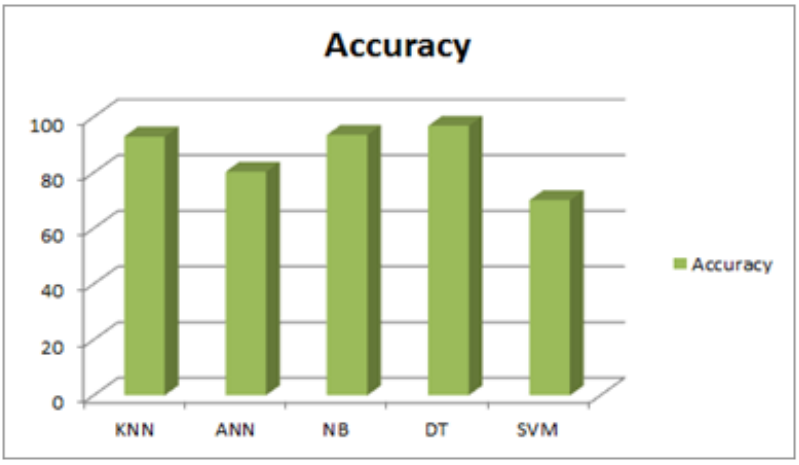

Figure 5: The accuracies of different classifiers for SAP prediction when teacher features are used

Table 8: The accuracies of SAP prediction when combination of student features and parent features are used

Classifiers	Accuracy
KNN	64.45
ANN	55.58
NB	90.23
DT	98.89
SVM	64.35

In Fig. 5, the DT classifier attained best accuracy of 96.77% for student academic performance prediction. The DT classifier shows best performance for SAP prediction when compared with other classifiers. The Table 4 shows the experimental results of student academic performance prediction when combination of student features and parent features are used to represent the vectors. In Fig. 6, the DT classifier attained best accuracy of 98.89% for student academic performance prediction. The DT classifier shows best performance for SAP prediction when compared with other classifiers. The Table 5 shows the experimental results of student academic performance prediction when combination of student features, parent features and school features are used to represent the vectors.

In Fig. 7, the DT classifier attained best accuracy of 98.89% for student academic performance prediction. The DT classifier shows best performance for SAP prediction when compared with other classifiers. The Table 6 shows the experimental results of student academic performance prediction

Table 9: The accuracies of SAP prediction when combination of student features, parent features and school features are used

Classifiers	Accuracy
KNN	64.45
ANN	64.55
NB	91.13
DT	98.89
SVM	64.35

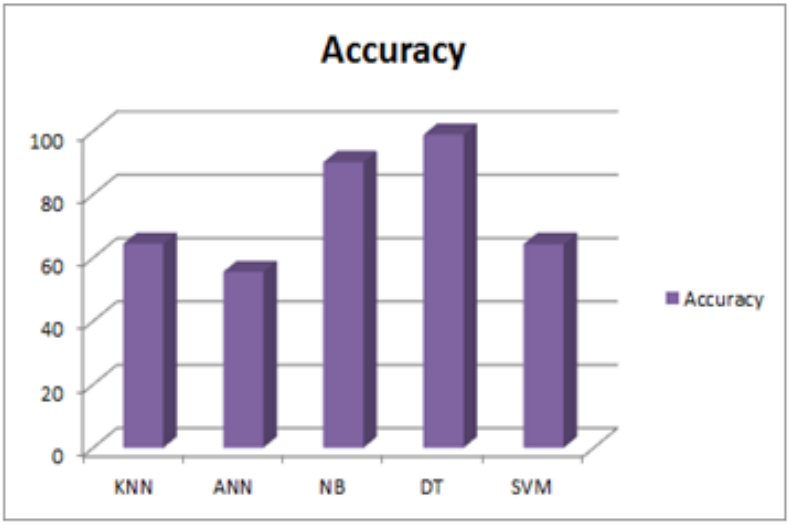

Figure 6: The accuracies of different classifiers for SAP prediction when student features and parent features are used

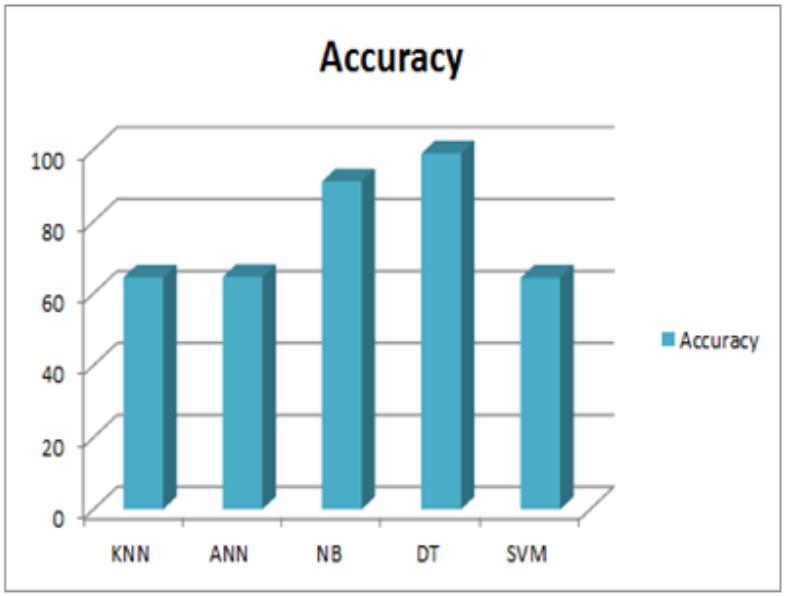

Figure 7: The accuracies of different classifiers for SAP prediction when student features, parent features and school features are used

Table 10: The accuracies of SAP prediction when combination of student features, parent features, school features and teacher features are used

Classifiers	Accuracy
KNN	64.65
ANN	91.33
NB	92.04
DT	97.88
SVM	64.35

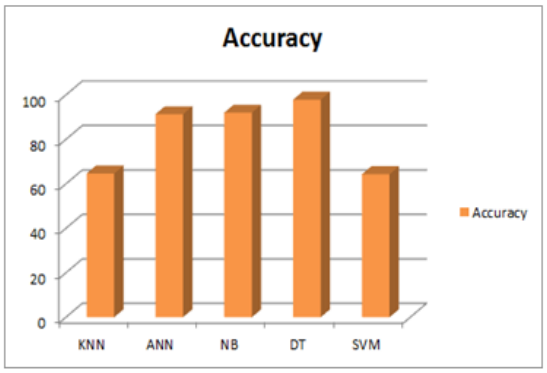

Figure 8: The accuracies of different classifiers for SAP prediction when student features, parent features, school features and teacher features are used

when combination of student features, parent features, school features and teacher features are used to represent the vectors. Table 6. The accuracies of SAP prediction when combination of student features, parent features, school features

In Fig. 8, the DT classifier attained best accuracy of 97.88% for student academic performance prediction. The DT classifier shows best performance for SAP prediction when compared with other classifiers.

In this experiment, the combination of student features and parent features attained best accuracy of 98.89% for predicting the student academic performance. It was observed that the accuracy was not changed when added the school features to the combination of student features and parent features. It was also observed that the addition of teacher features to the combination of student features, parent features and school features, there was a diminishing in the accuracy of student academic performance prediction. The school features alone attained good accuracy for predicting the academic performance of a student.

6. Conclusions

Feature selection plays an important role in reducing irrelevant features and improves the performance of machine learning Algorithms. RFBTRF feature selection method is used to reduce the irrelevant features and improve the student performance prediction rate. Experimental results demonstrated the addition of new features on the collected school dataset. RFBTRF feature selection method is used on the dataset and different classifiers such as SVM, KNN, NBC, ANN and DT performance is compared in terms of accuracy. By addition of new features on the dataset most important features are selected using RFBTRF Method and classifier performance is measured. From

the experimental results in addition of new features some classifier performance is stable and some classifier performance is either increasing or decreasing, this is due to the instances used in the dataset and basing on the nearest neighbour instances in the dataset.

References

[1] F. Al-Obeidat, A. Tubaishat, A. Dillon and B. Shah, *Analyzing students' performance using multi-criteria classification*, Cluster Comput.,21(1)(2018) 623-632.

[2] R. Asif, A. Merceron, S. A. Ali and N. G. Haider, *Analyzing undergraduate students' performance using educational data mining*, Comput. Educ., 113 (2017) 177-194.

[3] A. Asselman, M. Khaldi and S. Aammou, *Evaluating the impact of prior required scaffolding items on the improvement of student performance prediction* , Educ. Inf. Technol., (2020) 1-23.

[4] C. Burgos, M. L. Campanario, D. de la Peña, J. A. Lara, D. Lizcano and M. A. Martínez, *Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout*, Comput. Electr. Eng. , 66. (2018)., pp. 541-556.

[5] R. C. Chen, *Using deep learning to predict user rating on imbalance classification data*, IAENG Int. J. Comput. Sci., 46(2019) 109–17.

[6] E. B. Costa, B. Fonseca, M. A.Santana, F. F. de Araújo and J. Rego, *Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses*, Comput. Hum. Behav., 73(2017) 247-256.

[7] K. Deepika and N. Sathyanarayana, *Relief-F and Budget Tree Random Forest based feature selection for student academic performance prediction* , Int. J. Intell. Eng. Syst., 12 (1) (2019) 30-39.

[8] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. V. Erven, *Eduational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil*, J. Bus. Res., 94 (2019) 335-343.

[9] H. Fujita, *Neural-fuzzy with representative sets for prediction of student performance*, Appl. Intell., 49(1) (2019) 172-187.

[10] S. Helal , J. Li, L. Liu, E. Ebrahimie, S. Dawson and D.J. Murray, *Identifying key factors of student academic performance by subgroup discovery*, Int. J. Data Sci. Anal. ,7(3)(2018) 227-245.

[11] J. K. Jaiswal and R. Samikannu, *Application of random forest algorithm on feature subset selection and classification and regression*, World Congr. Comput. Commun. Technol.(WCCCT), IEEE, 2017, p. 65–68.

[12] M. Li, C. Huang , D. Wang, Q. Hu, J. Zhu and Y. Tang, *Improved randomized learning algorithms for imbalanced and noisy educational data classification*, Computing, 101(6)(2019) 571-585.

[13] M. Pandey and S. Taruna, *Towards the integration of multiple classifier pertaining to the Student's performance prediction*, Perspect. Sci., 8 (2016) 364-366.

[14] A. M. Radwan and Z. Cataltepe, *Improving performance prediction on education data with noise and class imbalance*, Intell. Autom. Soft Comput.,(2017) 1-8.

[15] X. Wang, R. Chen, F. Yan, Z. Zeng and C. Hong, *Fast adaptive K-means subspace clustering for high-dimensional data*, IEEE Access, 7 (2019) 42639–42651.

[16] J. Xu, K. H. Moon and M. V. D. Schaar, *A machine learning approach for tracking and predicting student performance in degree programs*, IEEE J. Sel. Top. Signal Process. , 11 (5)(2017) 742-753 .