ISSN 2278-3091

Volume 9, No.3, May - June 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse221932020.pdf

https://doi.org/10.30534/ijatcse/2020/221932020

Machine Learning Algorithms and Ensemble Technique to Improve Prediction of Students Performance

Randhir Singh[1], Saurabh Pal[2]
[1]Research Scholar, Department of Computer Applications,
VBS Purvanchal University, Jaunpur, India,
randhirs76@gmail.com
[2]Head, Dept. of Computer Applications,
VBS Purvanchal University, Jaunpur, India
drsaurabhpal@yahoo.co.in

ABSTRACT

Measuring Student's performance is necessary for the mordent society. Applications of Machine learning algorithms increased the growth in various fields like disease prediction, student's performance prediction, and crop productions prediction and in various other fields. The main aim of this study is to improve the prediction of students' performance using various machine learning algorithms and ensemble technique to get better accuracy over individual machine learning algorithms. We have used the student's dataset, which consists of 1000 instances and 22 attributes for evaluating the performance of students. In this paper we have applied four machine learning algorithms Decision Tree (DT), Naïve Bayesian (NB), K-Nearest Neighbors (KNN) and Extra Tree (ET) and then we have developed a model to combine the results of each individual base learner using Bagging and Boosting ensemble methods. The results obtained using bagging and boosting ensemble techniques were compared to select the best model.

The results of all machine learning algorithms and ensemble techniques are tested with various factors like accuracy, sensitivity, specificity and f1-score. After comparison of results we find that bagging is the best method which gives the better result as compared to bagging ensemble techniques. The developed model can be applied on the admission seeking students to identify the perdition of their performance in the selected course, which can be beneficial for both the students and Institution.

Key words: Educational Data Mining, Machine Learning; K-Nearest Neighbor Classifier, Extra Tree Classifier, Ensemble Technique.

1. INTRODUCTION

Academic performance evaluation is a type of clustering problem where clusters are their grades like first division, second division, third division and fail, which measures the intelligence level of students. The Intelligence level bases groupings are essential for selecting the deserving students and to provide them good education so that they achieve their aim of life. As cost of electronics equipments prices are reduced the application of IT equipments are increased in the educational institutions for collecting huge amounts of data about their students. At present days generally affiliating Universities collect registration data using registration for and examination data using examination form and also Universities provide e-mark sheet, certificate, migration to students, but most time these data remains unused because Universities does not further analyzed these data. Data mining techniques can help to develop machine learning model with the help of these data. The application of data mining techniques on these data is Educational Data Mining (EDM). EDM includes study of new tools and development of new models and applying new algorithms on a huge amount of data to find useful unknown patterns and, thus, help to understand students' behavior and learning capabilities. Therefore, educational data mining provides discovering of new applications for solving problems which are related to educational fields.

Now a day's higher education institutions are utilizing ITC (Information Technology and Communication) resources, as Management Information Systems, which are providing high volumes of data [1]. Perfect machine learning techniques can be used in analyzing educational data. In this paper, we use machine learning classifiers to develop a new model for prediction of academic performance of students.

The objective of the study is to find the factors taken admission in courses which affect low academic performance in Bachelor of Computer Applications (BCA) for students at the United College of Management, Prayagraj, UP. The developed model can be effectively work for BCA department to take right decisions and for monitor and support students which need special attention to enhance the performance of students and quality of Institute.

The rest paper is prepared as follows. The next Section 2, describe the previous work done on EDM field using machine learning algorithms. In Section 3 we discus about the dataset used, the machine learning algorithms applied, ensemble methods applied to enhance the results of single classifiers. Section 4 presents the results and discussion. In Last section 5 we have discussed the conclusion of the research with findings achieved by this study.

2. RELATED WORK

A review of relevant literatures was carried out. Ahamad and Elaraby [2] evaluated the 6 years (2005–2010) enrollment students' data in a specific course program across, with many features collected from the university database. The work predicted the final grades in the particular program.

Pandey and Pal [3] presented an educational data mining approach to classify students' according to performers or underperformers class using Naïve Bayes algorithm.

Bhardwaj and Pal [4] did a relative study for investigation of various decision tree algorithms on an academic dataset in order to predict the student's academic performance. The work primarily concentrates on choosing the best decision tree algorithm, and then provides a standard for them individually. It was discovered that the CART decision tree technique performed reasonably better on the dataset used for testing, that was obtained based on the accuracy and precision produced at the validation stage.

Livieris, et al. [5] developed an Artificial Neural Network (ANN) classifier to predict the performance of students in Mathematics. In their work, they discovered that the modified spectral Perry trained artificial neural network performs better classification compared to other classifiers.

Kotsiantis, et al. [6] explored machine learning techniques for dropout prediction of students in distance learning. This study contributed in that it carved the path for educational data mining and one of the first works to implement machine learning methods in an academic environment. Their algorithm was fed on demographic data and several project assignment rather than class performance data to make prediction of students.

Moucary, et al. [7] applied a hybrid technique on K-Means Clustering and Artificial Neural Network for students who are pursuing education while adopting a new foreign language as a means of instruction and communication. Firstly, Neural Network was used to predict the student's performance and then fitting them in a particular cluster which was form using the K-Means algorithm. This clustering helped in serving a powerful tool to the instructors to identify student's capabilities during their early stages of academics.

Hongsuk, et al. [8] develop a Deep Neural Network model to estimate flow of traffic conditions. A Traffic Performance Logistic regression was used for congested traffic condition and a non-congested traffic condition. The 3 layer mold was able to estimate the congestion with a 99% of accuracy.

Zoric and Alisha [9] study the performance perdition on 76 students from the University of Applied Sciences Baltazar, Zaprešić. Using Neural Network author achieved prediction rate 93,42 % due to insufficient data of lower-grade students (only 8%).

Hassan et al. [10] used 1170 students' data set to predict student's performance. They used Linear Discriminant Analysis, Gradient Boosting Classifier, Random Forest Classifier, SVC, KNN and Decision Tree Classifier and achieved highest accuracy of 89.74% in case of Decision Tree.

Gamao and Gerardo [11] developed a predictive model for the analysis of dropouts using the students' cumulative record to achieve through the integration of the MMFA with the relevant classification algorithms such as NB and DT. The MMFA was able to explore the optimized solution or model through the search space with the use of firefly behavior using the mutation process. Thus, the accuracy achieved through the study can be used by academic administrators of the Davao del Norte State College in crafting academic policies that enhance students' performance for both curricular and extra-curricular activities to minimize dropouts. Future researchers may consider integrating the MMFA with other relevant classification algorithm to explore its potential use for whatever it serves best.

Gil et. Al [12] predicted the causes of students' dropout are an important and challenging task for schools and educators. Accordingly, they checked whether the use of data mining processes could be beneficial in addressing this issue in every school. Wherein the student dropout indicators successfully predicted using the data mining classification technique. To identify the dropout student indicators, the most commonly used data mining approaches were used based on C4.5 and Naive Bayes. These two different classification algorithms trained and tested using a ten-fold cross-validation approach. It alerts the educator to take appropriate action to improve student performance through specialized coaching and counseling.

Singh and Pal [13] described different machine learning techniques for evaluating the performance of students. Five machine learning techniques PCA, SVM, LDA,RNC and ET are used to classify the prediction of students. The best accuracy find among these different techniques is 94.86% from SVM. The second highest accuracy obtained is 93.21% in the case of LDA. They got the highest accuracy in the literature available on student's performance prediction. The machine learning-based method reduces generation errors and obtains more information by using the first-stage prediction as a feature rather than a separate training.

3. METHODS

In this paper we have used three machine learning classifier algorithms, namely: Decision Tree (DT), Naïve Bayesian (NB), K-Nearest Neighbors (KNN) and Extra Tree (ET) for the purpose of testing our dataset which was taken from United Institute of Management, Prayagraj. A brief description of the classifiers used in this study is described below.

- **Decision Tree (DT):**

  Decision tree is a type of supervised learning based predictive modeling tool. Decision tree is based on graphical representation of all possible solution on different conditions. A decision tree is generated from root following top-down approach that involves partitioning of data; entropy is used to calculate homogeneity of data. Category based data as well as numerical data both work with this model.

- **Naïve Bayesian (NB):**

  Naive Bayes is a popular data classification technique. It is based on the probability theory concept and based on assumption that there is no dependency among predictors. In other words it assumed that the presence of a particular feature in a class is not related to each other.

- **K-Nearest Neighbors (KNN):**

  K nearest neighbor algorithm is a classifier which creates different types of cases that is based on similarity measure It is a supervised machine learning algorithm applied for classification and regression problem. It is non-parametric because it does not have any assumptions about the distribution of data. In classification algorithm, learning depends on ‘how similar’ a data is from the other.

- **Extra Tree (ET):**

  This method is an ensemble method which stands for Extremely Randomized Trees. This algorithm develops randomizing of tree for numeric input features. It often leads to increased accuracy when compared to the ordinary random forest.

  Figure 1 shows the structure of methodology used in this research paper.

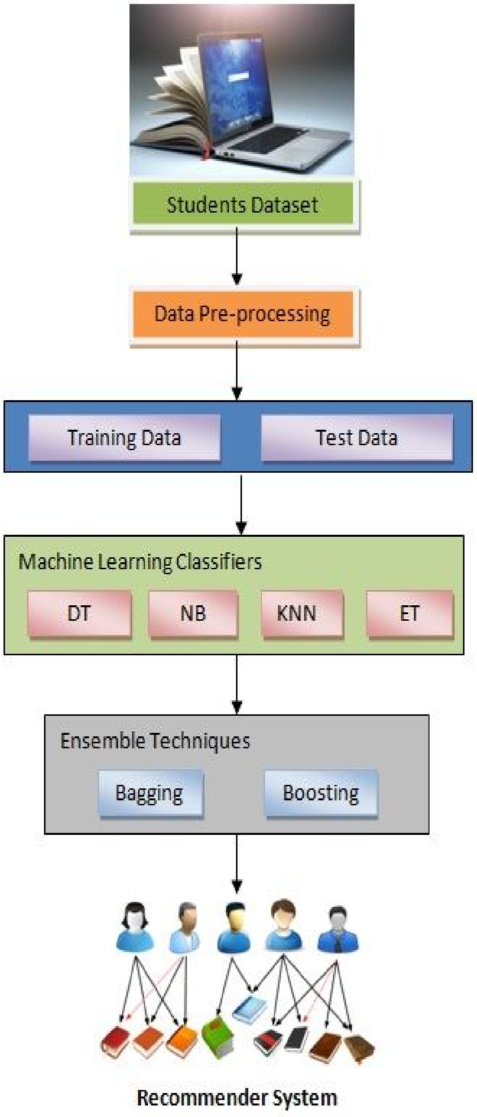


Figure 1: Methodological approach for Performance Prediction

3.1 Dataset Analysis

The data used in this study is of Bachelor of Computer Applications program, which has been collected from United Institute of Management, Prayagraj. The BCA course is divided in 3 years which consist of two semesters per year; therefore total six semester examination completes the whole BCA course. In this research paper we have taken count of only final semester results. The data is collected with the permission of examination and admission departments from the year 2014 to 2019 and total number of students passed from the institution is 1000, therefore total 1000 instances are available with 22 attributes; these attributes are collected from the registration as well as examination form. The target and other variables discussed in this study are listed in table 1.

Table 1: Student Dataset

Feature	Attribute	Domain
S1	Sex of Students	1= Female, 2=Male
S2	Students category	1= General, 2=OBC, 3=SC, 4=ST, 5=Minority
S3	Discussion at home	1=Always, 2=Almost Always, 3=Sometimes, 4= Never
S4	Own Computer /Laptop	1=Yes, 2=No
S5	Laptop shared with family	1=Yes, 2=No
S6	Study desk at home	1=Yes, 2=No
S7	Own mobile phone	1=Yes, 2=No
S8	Own Gaming system	1=Yes, 2=No
S9	Heating/Cooling systems at	1=Yes, 2=No
S10	Absent from school	1=Once a week or more, 2=Once every two weeks, 3=Once a month, 4=Never or almost never
S11	How often use computer/Laptop at home	1=Every day or almost every day, 2=Once or twice a week, 3=Once or twice a month, 4=Never or almost never
S12	How often use computer at School	1=Every day or almost every day, 2=Once or twice a week, 3=Once or twice a fifteen days, 4= Once or twice in a month, 5=Never or almost never
S13	Access textbooks	1=Yes, 2=No
S14	Completed assignments	1=Yes, 2=No
S15	Collaborate with classmates	1=Yes, 2=No
S16	Communicate with teacher	1=Yes, 2=No
S17	Students grade in Senior Secondary education	1 =90% -100%, 2= 80% - 89%, 3= 70% - 79%, 4= 60% - 69%, 5= 50% - 59%, 6= 40% - 49%, 7= < 40%
S18	Fathers qualification	1=primary, 2=middle, 3=graduate/Post-graduate, 4=doctorate
S19	Mother's Qualification	1=primary, 2=middle, 3=graduate/ost-graduate, 4=doctorate
S20	Father's Occupation	1=Service, 2=business, 3=NA
S21	Mother's Occupation	1=House-wife , 2=employed, 3=business, 4=NA
S22	Performance in B.C.A	1= > 60%, 2= >45 & <60%, 3= >36 & <45%, 4= < 36%}

3.2 Data Preprocessing

The first step shown in methodology is data preprocessing. Data preprocessing includes (i) a method to select students' records and choosing important attributes and (ii) the students records are not clean and include inconsistent data. So apply different methods of data cleaning to clean such anomalies. The dataset for the study was collected from the United Institute of Management, PrayagRaj. The dataset had 1000 instances and 22 attributes. The student's dataset is pre-processed using equation

$$x_{norm} = (x - x\_mean)/\rho$$

Where

x_{norm}= Normalized value of x,

x_{mean} =Mean value of x and

ρ=Standard deviation of the given population.

3.3 Ensemble Techniques

Ensembles techniques improve the accuracy of predication as compared to single classifier on the dataset. In this paper, we have applied two ensemble methods to improve the performance of classification algorithms. The two most popular ensemble techniques: Bagging classifier and Boosting (Adaboost classifier) are used to combine the results obtained by the four machine learning classifiers.

- **Bagging Classifier:** Bagging method is applied to decrease the variance calculated by decision tree classifier. The objective of bagging ensemble method is dividing the dataset into various subsets for training selected at random with substitution. Now, these data subsets are trained using decision trees. Now, the average of the results obtained by each data subset is taken which gives better results as compared to single classifier.
- **Boosting Classifier:** Boosting classifier is another important ensemble technique. It is applied to create a group of classifiers. In boosting method, classifiers are trained serially by classifiers fitting data and then analyzing errors. Decision trees are tainted successive to fit from the data and with the objective to get improved accuracy at each stage. Bagging method is used to converts weak classifiers to a good model.

4. RESULTS

Before applying machine learning classifiers the dataset is visualized using histogram and density map. The histogram and density maps of all attributes of student dataset represent the bar of frequency of different values and density map is a smooth continuous curve, which is formed by estimating the density from the data individually. In a density map, continuous curve is drawn at each individual data point and then all these curves are summed up to get a single smooth curve accordingly. We measured the density of each attributes on the basis of target variables classification and represent in Figure 2.

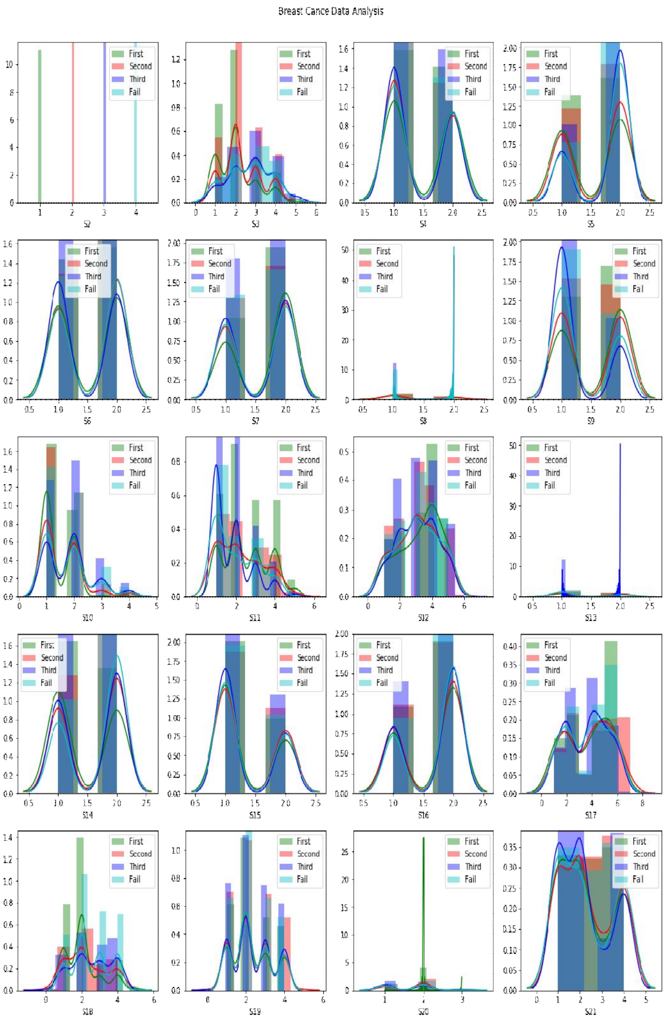


Figure 2: Dataset visualization using histogram

The analysis of dataset and implementation of classification in this paper has been done using Python code. The student dataset is divided into 80% as training set and 20% as test dataset using 10-fold cross validation.

To measure the performance of the predicted model accuracy, Recall, specificity, precision and F-1 scores are evaluated using the formula shown in table 2.

Table 2: Formulas

Sr. No.	Performance Metrics	Formula
1	Accuracy	$\frac{(TP+TN)}{(TP+FP+TN+FN)}$
2	Recall	$\frac{TP}{(TP+FN)}$
3	Specificity	$\frac{FP}{(FP+TN)}$
4	Precision	$\frac{TP}{(TP+FP)}$
5	F-1 score	$2\frac{(Precision*Recall)}{(Precision+Recall)}$

The value calculated for four classifiers using equations shown in table 2 are described in Table 3.

Table 3: Output of Evaluating Algorithms

Classifier	Accuracy	Recall	Specificity	Precision	F-1 Score
Decision Tree	81.27	74.264	89.58	72.47	70.63
Naive Bayesian	86.83	81.55	98.99	82.89	72.78
KNN	84.92	80.36	94.97	75.66	71.78
ET	82.72	72.78	89.15	72.76	72.78

On comparing the results, we finds that the best accuracy achieved by Naïve Bayesian classifier as 86.83%, from the table it is clear that nearly all classifiers predict the accuracy in between 81% to 87%, which proves the selection of these classifiers are best for predicting the performance of students. K-nearest Neighbor classifiers also performs well as its accuracy achieved is 84.72%. The comparison of the four classifiers on the basis of accuracy is presented in figure 3 using box and whisker plot to illustrate the mean value of prediction.

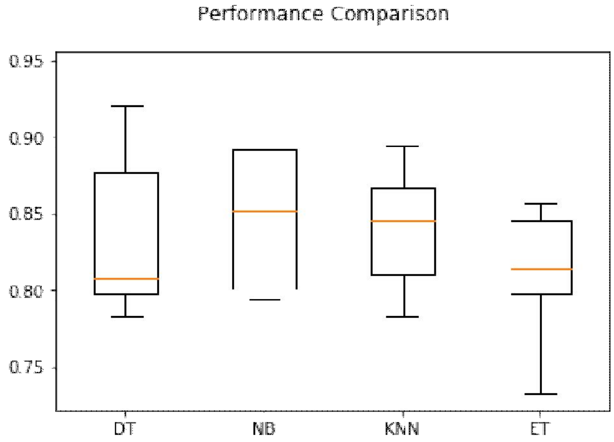


Figure 3: Accuracy of different algorithms

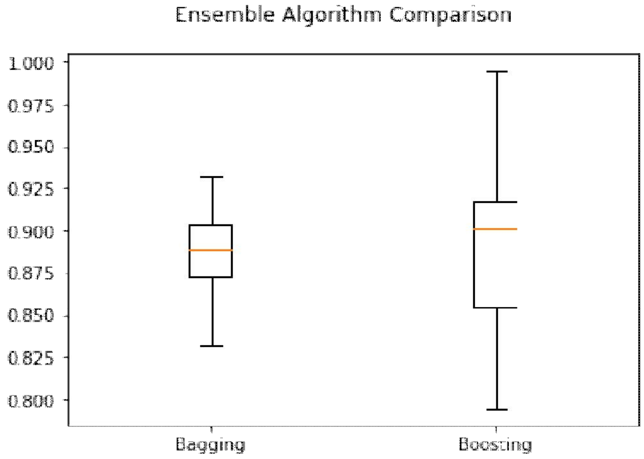
To improve the results of the machine learning classifiers, we have used Bagging and Boosting ensemble techniques. Ensemble techniques are applied to improve the accuracy of machine learning classifiers. After applying the Bagging and Boosting the results obtained by the two methods are shown in table 4.

Table 4: Output of Ensemble Classifiers

Ensemble Classifier	Accuracy	Recall	Specificity	Precision	F-1 Score
Bagging	89.56	70.94	93.41	73.41	68.47
Boosting	91.76	76.33	99.39	80.55	65.77

From the table it is clear that Boosting is the better ensemble technique as compared with bagging because the accuracy of boosting classifier is 91.76% which is also approximately 5% more than the single machine learning classifier Naïve Bayesian. The comparisons of two ensemble techniques are shown in Figure 4.

Figure 4: Accuracy of ensemble methods



5. CONCLUSION

The main objective of this paper is to build an efficient model which can improve the prediction of performance accuracy of students. Machine learning techniques and ensemble methods are widely used in student's performance prediction these days. Ensemble methods are used to improve the results of single machine learning classifiers. In this paper four machine learning classifiers Decision Tree, Naïve Bayesian, K-nearest Neighbor and Extra Tree are used as base learning algorithms and then two ensemble techniques Bagging and Boosting are used to enhance the results of single base learners. The best accuracy among these different machine learning classifiers is 86.83% from Naïve Bayesian and 91.76% in boosting ensemble technique.

The results obtained by this paper can be used to find the non-performing students and to pay more attention on these students to improve the performance. This can improve the quality of higher educational and may be beneficial for the higher education institutes.

REFERENCES

1. Ferguson, R**. The State of Learning Analytics in 2012: A Review and Future Challenges**. Tech. Rep.; Knowledge Media Institute; http://kmi. open. ac. uk/publications/techreport/kmi-12-01, 2012. https://doi.org/10.1145/2330601.2330616
2. Ahmed A, Elaraby I. **Data mining: A prediction for student's performance using classification method**. World Journal of Computer Application and Technology. Vol. 2, Issue 2, pp:43-47, 2014.
3. Pandey U, Pal S. **Data mining: A prediction of performer or underperformer using classification.** (IJCSIT) International Journal of Computer Science and Information Technology. Vol. 2, Issue 2, pp. 686-690, 2011.
4. Bhardwaj B, Pal S. **Data mining: A prediction for performance improvement using classification**. (IJCSIS) International Journal of Computer Science and Information Security. Vol. 9, Issue. 4, pp. 136-140, 2012.
5. Livieris I. & Pintela P. **An improved spectral conjugate gradient neural network training algorithm**. International Journal on Artificial Intelligence Tools. Vol. 21, Issue. 1, pp. 1250009-1-21, 2012. https://doi.org/10.1142/S0218213011004757
6. Kotsiantis S, Pierrakeas C, Pintelas P. **Predicting students' performance in distance learning using machine learning techniques**. Journal of Applied Artificial Intelligence. Vol. 18, Issue. 5,pp 411-426, 2004.
7. Moucary C, Khair M, Zakhem W. **Improving student's performance using data clustering and neural networks in foreign-language based higher education**. The Research Bulletin of Jordan ACM. Vol. 2, Issue. 3, pp. 27-34, 2011.
8. Hongsuk Y, Jung H, Bae S. **Deep Neural Networks for traffic flow prediction**. In Big Data and Smart a Computing (BigComp), IEEE International Conference. IEEE. pp. 328-331, 2017. https://doi.org/10.1109/BIGCOMP.2017.7881687
9. Bilal Zorić, A. (2019). **Predicting Students' Success Using Neural Network.** RENET-Society for Advancing Innovation and Research in Economy, Zagreb, pp. 58-66, 2019.
10. Hasan, H. R., Rabby, A. S. A., Islam, M. T., & Hossain, S. A. (2019, July). **Machine Learning Algorithm for Student's Performance Prediction**. In 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) pp. 1-7. IEEE, 2019.
11. Gamao1, A.O. & Gerardo, B.B. **Prediction-Based Model for Student Dropouts using Modified Mutated Firefly Algorithm.** International Journal of Advanced Trends in Computer Science and Engineering, Vol.8, No.6, pp 3461-3469, 2019. https://doi.org/10.30534/ijatcse/2019/122862019

12. Gil, J.S., Delima, A. J. P. & Vilchez, R. N. **Predicting Students' Dropout Indicators in Public School using Data Mining Approaches**. International Journal of Advanced Trends in Computer Science and Engineering, Vol.9, No.1, pp 774-778 2020. https://doi.org/10.30534/ijatcse/2020/110912020
13. Singh, R. & Pal, S. **Application of Machine Learning Algorithms to Predict Students Performance**. International Journal of Advanced Science and Technology, Vol 29, No. 5, pp 7249-7261, 2020.