Performance Based Machine Learning Model to Enhance Performance of Students

Bhavesh Patel

Abstract*: Machine learning techniques are used by many organizations to analyze the data and finding some meaningful hidden pattern from the data, this process is useful by an organization to take the decision making process. Various organizations used like marketing, health care, software organization and education institute etc used it in decision making. We have used machine learning techniques to enhance the performance of students. It will be ultimately used by educational institute to improve the status of educational institute. This research paper includes Naïve Bayes (NB), Logistic Regression (LR), Artificial Neural Network(ANN) and Decision Tree machine learning techniques. Performance of these models have been compared using accuracy measures parameters and ROC index. This research paper has used various parameters like academic performance and demographic information to build the model. In addition to judge the performance also used some additional parameters to measure the performance like F-measure, precision, error rate and recall. The dataset is collected using survey methodology to build the model. As a conclusion found that the Artificial Neural Network model get the best performance among all the models.

***Keywords*: machine learning, precision, recall, Naïve Bayes, Artificial Neural Network, classification, performance**

I. INTRODUCTION AND OBJECTIVE OF THE RESEARCH

Great efforts done by many researchers to enhance the performance of students for various purposes like: detecting at-risk students, ensuring student retention, assigning courses and resources, and many more. The goal of this article is to forecast performance of students based on the different academic and demographics parameters. Student's performance can be predicted using machine learning techniques and recognize students at risk, so appropriate actions can be taken on this type of students to improve their performance. So, this research paper has various machine learning techniques to build the model and predict the students' performance. This research paper has been used the small student dataset to explore the result. This research paper has used Naive Bayes (NB), decision trees(DT), artificial neural network(ANN) and logistic regression(LR) machine learning classification techniques to build the model. The ROC index and other accuracy parameters are used to find out highly accurate model for prediction of performance of students. This research paper has used survey methodology to collect the dataset and build the model. The student's dataset is collected from the Faculty of Computer Applications department. This dataset contains information of 867 students. This research paper follows various activities to create, normalize and explore the results from the datasets like data gathering, pre-processing, evaluation pattern finding and result generation of four machine learning models. Ultimately main objective of the research is to finding the best model and analysis of the results. This research paper is divided into four sections: first section is Introduction and Objective of the research. Second section demonstrate the used methodology in research, third section shows the analysis and result and last fourth section discussed the concluded result.

Objective of the research:

the objective of this research article is to compare the model and find out the highest accurate model to predict the academic performance of the students.

II. LITERATURE REVIEW

Nowadays, many works related to this subject have been published. We found several literature reviews that looked at student academic performance patterns in different perspectives. Shahiria A.M. et. al. has used several mining techniques to evaluate the performance of students. They evaluated how these predictive algorithms could be used to find out the highly significant attributes in the student database [1]. P. Kavipriya used various data mining methods for predicting, analyzing, early warning, and evaluating student performance. He reviewed various classification methods such as decision trees, inexperienced Bayes algorithms, etc. He suggested that because it is difficult to predict student performance due to many challenges such as statistical imbalances, there is a need to install a support vector machine that offers the best accuracy in his study [2]. A. Mat et. al. has reviewed predictive modelling technique for student's academic performance. They have nursing various learning tasks for making the predictive models. Finally, they have used this models for course recommendation and career path planning [3]. Mishra T. et.al surveyed student performance and employment prediction using data mining. he focused primarily on the traditional educational establishment into his research work. They have demographic and socio-economic factors into the prediction model. He also said that some research work also has been done on employment forecasting [4]. Zaffar M. et. al used various prediction models for students they have taken admission in programming courses. Filter Feature selection algorithms have been used by them in the pre-processing stage for generating the result. Based on the result they found that student's attendance, mathematics result, physics result plays an important role in programming course [5].

Manuscript received on January 08, 2021.
Revised Manuscript received on January 15, 2021.
Manuscript published on February 28, 2021.

Dr. Bhavesh Patel, Assistant professor, Department of Computer Science Ganpat University, MCA,

Xu et al, have used various models of machine learning to forecast the performance of students. they have used kNN, RF, Logistic Regression, Linear Regression, and Proposed Progressive Prediction algorithm in their research work. As a result, they found that Proposed progressive prediction algorithm is the best algorithm among all the algorithms [6]. John Jacob et al. used different mining methods to predict performance of students. they have utilized techniques like Linear regression and decision tree to identify the poor students in academic. They also used clustering techniques to cluster the students as per their performance in the academic [7]. Hari Ganesh et al. learn several Data Mining applications in their research work. This research work surveyed different techniques of data mining and its algorithms in several regions of EDM. They discover that, EDM can also be used to find out the knowledge based process for the problems of primary students [8]. M. A. Al-Barrak et. al describe that ANN and Decision Tree both are well known classification methods for classifying the data and prediction. This research paper has used decision trees for predicting performance of students that can be helpful to students who need special attention. In this research paper, researchers have predicted students drop out rate using academic, socio-demographic and institutional data to forecast the final GPA of students [9].

III. RESEARCH MODEL

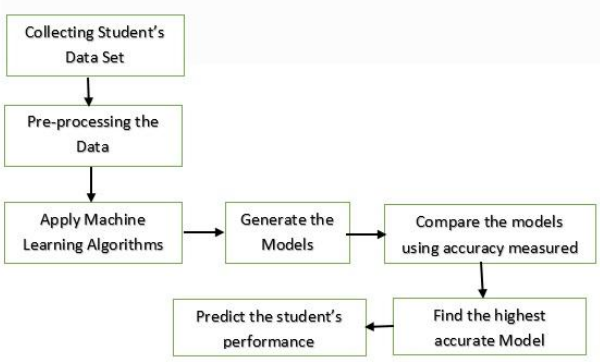

Fig. 1. Steps of proposed model

In this research paper followed the steps as per described in the figure. Initially collect the student's data set using the survey methodology. Further, preprocess the data to remove noisy and missing records from the data sets and normalize the data. After that applied various machine learning algorithms through the selected tools and generate the various models. Finally compare the models with accuracy measured parameters and find the highest accurate machine learning model and applied this model for predicting academic performance of students.

IV. DATA COLLECTION

This research paper has been collected total 867 student's data from Faculty of computer application. The data has been collected by preparing a questionaries and shared it among the students. After collecting the student's data we have transfored it .csv file to processed the data using WEKA tools. Further experiment is done using the WEKA tools. In this research we have collected the data based on the following student's academic and demographics parameters.

Table – I: Parameters used in research

ATTRIBUTES	ATTRIBUTES DEFINITION
Gender	Male =1 , Female =2
Department	BCA=1, B.Sc.(CS)=2, B.Sc. (IT) =3, M.Sc. (IT)=4, M.C.A =5
Family Income	Low=0, middle=1, high=2, very high=3
Family Education	No Education=0, Primary Education=1, Secondary Education=2, Graduation=3, Higher Graduation=4,
No. of siblings	No sibling=0, One=1, two=2, three=3, more than three=4
Attendance Ratio	Poor=0, average=1, good=2, very_good=3, excellent=4
Internal Results	Poor=0, average=1, good=2, very_good=3, excellent=4
Assignment Submission	Never submitted=0, irregular=1, Regular=2
Theory examination Result	Poor=0, average=1, good=2, very_good=3, excellent=4
Practical Examination Result	Poor=0, average=1, good=2, very_good=3, excellent=4
Using internet in study	Never=0, Always=1, Sometimes=2
Time spent is social media (hours)	Nominal value (1 to 24)
Final Outcome	Poor=0, Average=1, Good=2, very Good= 3, Excellent=4

V. METHODOLOGY USED IN RESEARCH

Artificial Neural Networks (ANN)

ANN is group of input and output units those are linked together using weighted connections. ANN absorbs by altering the weights to predict correct target. Backpropagation algorithm is highly used to train ANN. ANN has several benefits to use, like high resistance against noisy data and it gives good performance for classifying patterns using untrained dataset. ANN has many real world applications, such as speech recognition, handwriting and image identification etc. RNA can be recognize using their architecture. The architecture of fully connected multi-layer front feeder ANN is: an input layer, one or more hidden layers, and an output layer. Here, connections not going back to previous layer. In addition, every element in the L layer gives inputs to every element in the L + 1 layer [10].

Logistic Regression (LR):

LR is a one type of mathematical modeling technique that defines association among different independent variables $X_1 ... X_K$ and a dependent variable. LR uses logistic function as a mathematical equation and it has value between 0 and 1 for specified input. This logistic model shows the probability of any event that is always a value between 0 and 1. The following equation shows the logistic model.

$$P(D = 1|X_1, X_2, \dots, X_k) = 1/1 + e^{-(\alpha + \Sigma_1^k \beta_i X_i)} \quad (1)$$

Here α and β shows the parameters of model. In the training phase, to find the best values of model, it uses Gradient Descent Algorithm [11].

Naïve Bayes (NB)

It is one type of classification model. It is simplest disparity of Bayesian network. It describes each instance is independent than other instance. The following formula is used in Naïve Bayes model.

$$V_{max} = \max_{v_j \in V} P(v_j) \prod_i P(a_i|v_j) \quad (2)$$

Here v shows target of the model, $P(ai|vj)$ and $P(vj)$ both can be find out by counting the frequencies in training dataset [12].

Decision Tree (DT)

DT model shows tree structure that resembles flowchart. In this structure, internal node shows attributes of test, branch shows the result of the test. Further, leaf node shows the target object label, and the first node shows the root node. The decision tree is either binary or non-binary tree. Decision tree has not required previous knowledge of any problem. so it is commonly used classification technique. It can also be easily converted into classification rules and generated rules are easy to understand. They are used in several real-world applications like molecular biology, manufacturing, medicine and financial analysis etc. The most common decision tree algorithms include CART, C4.5 and ID3 [10].

Validation Methods and Performance Measure Parameters:

this research paper used 3-fold cross validation technique. Here, the database will be distributed in three equal sets. Due to three fold's cross validation the testing and learning sets are executed three times. In this method, the machine learning algorithm picks one set for testing purpose and other two sets for training purpose. Finally aggregate all folds or execution to count performance and accuracy of model. Classification models performance can also be evaluated using the ROC index. It is also a one type of most useful performance measure under the curve. ROC index is calculated using predicted score. Following equation three is used to count ROC index [21]. Furthermore, except ROC index, other important measures also used to find the accuracy of model like F- Measure and Error rate of classification error. Here, the following formula four is used for calculating the F-Measure. Generally, F-Measure use to find out the misclassification rate. [13]

$$ROC\ index = \sum_{i=2}^{|T|} \left(FPR(T[i]) - FPR(T[i-1])\right) \times \left(TPR(T[i]) + TPR(T[i-1])\right)/2 \quad (3)$$

Here $|T|$ shows thresholds those used in research. $FPR(T[i])$ shows the false positive rate. $TPR(T[i])$ shows the true positive rate. Model with greater ROC-index describes good classification model. A model is strong model if that have ROC-index value above 0.7 and a model is week if that have value below 0.6 [13].

$$F\ Measure = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (4)$$

$$Precision = \frac{TP}{(TP + FP)} \quad (5)$$

$$Recall = \frac{TP}{(TP + FN)} \quad (6)$$

Here TP is a True Positives. It shows the data rows in test sets having positive target and also those are predicted as positive target. Here TN is a True Negatives. It shows the data rows in test sets having negative target and also those are predicted as negative target. Here FP is False Positives. It shows the data rows in test set having negative target but those are predicted as positive target. Here FN is False Negatives. It describes the number of data rows in the test set positive target but those are predicted as negative target [13].

VI. EXPERIMENT AND RESULT ANALYSIS

In this research paper, Naïve Bayes(NB), Decision Tree(DT), Artificial Neural Network(ANN) and Logistic Regression (LR) models are used. Each model's accuracy and performance measures are described into the following table.

Table-II: Result of machine learning models using accuracy measure parameters

MODEL	ANN	DT	LR	NB
TRUE POSITIVE	78	78	62	58
FALSE POSITIVE	19	18	17	25
TRUE NEGATIVE	62	60	63	55
FALSE NEGATIVE	20	20	25	33
PRECISION	82.18	80.15	82.16	77.67
RECALL	80.78	79.82	77.62	74.52
F-MEASURE	81.22	80.2	79.23	76.45
ACCURACY	81.52	79.15	78.32	75.32
ERROR RATE	18.48	20.85	21.68	24.68
ROC INDEX	0.831	0.792	0.762	0.752

Abbreviations:

DT- Decision Tree
NB – Naïve Bayes
ANN – Artificial Neural Network
LR – Logistic Regression

As per described into the table, lowest ROC index in Naïve Bayes model. It's value is 0.762. Even accuracy rate is also lowest that is 75.32 and error rate is highest that is 24.68. So based on above result prove that Naiive Bayes is a week model. At contrast, Artificial Neural Network (ANN) model having highest accuracy that is 81.52. Further, higher ROC index value that is 0.831 and the lowest error rate that is 18.48. This result itself proves that ANN model is highly accurate machine learning model to predict academic performance of students. The following figures show the result analysis into the form of charts.

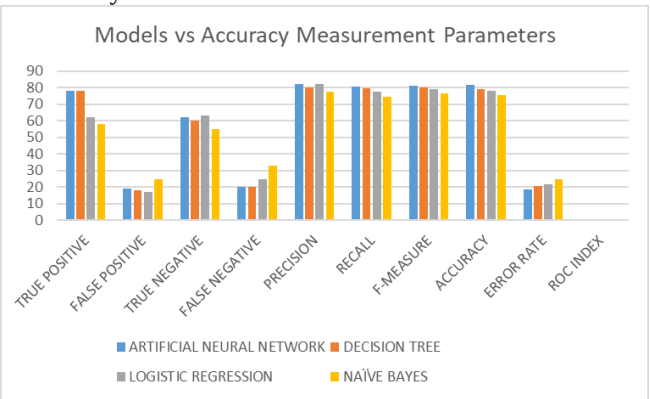

Fig.2. Experiment Results of Models Vs Accuracy Measurement Parameters

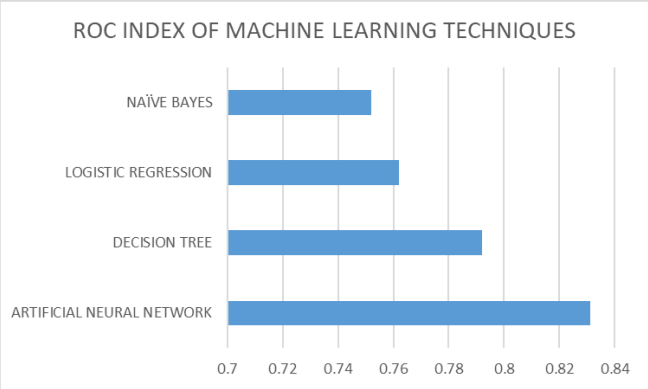

Fig. 3. Experiment Result of ROC Index of different models of research

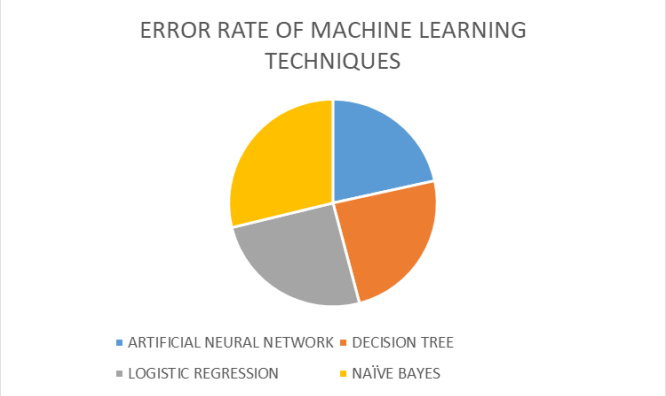

Fig. 4. Experiment Result of Error rate of machine learning techniques

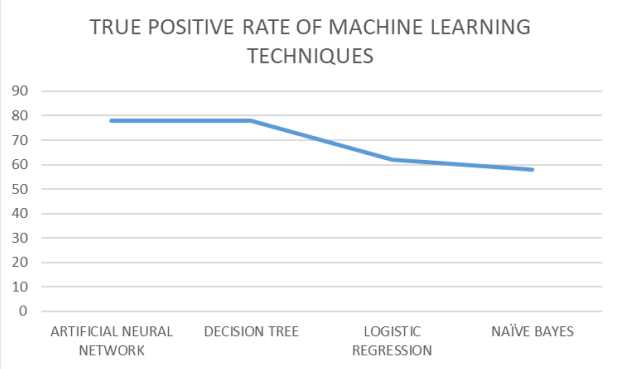

Fig. 5. Experiment result of True positive rate of models

VII. THE CONCLUSION

This research paper has address the problem of identifying students with poor academic achievements. this research has been used total 867 student's data from Faculty of computer application to address the problem. This dataset has used academic achievement and demographics parameters of students in this research paper. The objective paper is to enhance performance of students. To achieve this goal, this research paper has used four machine learning techniques Decision tree (DT), Naïve Bayes(NB), Artificial Neural Network(ANN) and Logistic Regression (LR). These used models are compared using accuracy measured parameters like true positive (TP), true negative(TN), false positive(FP), false negative(FN), precision, recall, error rate and ROC Index. ROC Index and error rates are highly acceptable measurements to prove the accuracy of model. So, by considering these two accuracy measurement parameters if checked the model then find that Artificial Neural Network (ANN) is highly accurate machine learning model in this research. This research paper proved that, Artificial Neural Network (ANN) has highest ROC index (0.831) and lowest error rate (18.48) among all the compared model. At contrast, Naïve Bayes model has Lowest ROC index (0.752) and highest error rate (24.68) model. So based on this experiment it proves that, Artificial Neural Network is highest accurate model and Naïve bayes is week model for used data set to predict the academic performance of students.

REFERENCES

1. A.M. Shahiri, W. Husain, and N.A. Rashid, "A Review on Predicting Student's Performance using Data Mining Techniques", Proceeding Computer Science, vol. 72, (2015), pp. 414-422.
2. P. Kavipriya, "A Review on Predicting Students' Academic Performance Earlier, Using Data Mining Techniques," International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 12, December 2016 ISSN: 2277 128X.
3. Asiah, M., et al.: "A review on predictive modeling technique for student academic performance monitoring." In: EDP Sciences (eds.) MATEC Web of Conferences 2019, EAAIC, vol. 255, p. 03004. EDP Sciences (2019)
4. Mishra, T., Kumar, D. and Gupta, S., "Students' Performance and Employability Prediction through Data Mining: A Survey", Indian Journal of Science and Technology, Vol.10(24), (2017)
5. Maryam Zaffar, Manzoor Ahmed Hashmani, and KS Savita. "A Study of Prediction Models for Students Enrolled in Programming Subjects". In 4th International Conference on Computer and Information Sciences (ICCOINS), 2018.
6. J. Xu, K. H. Moon, and M. Van Der Schaar, "A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs," IEEE J. Sel. Top. Signal Process., vol. 11, no. 5, pp. 742–753, 2017.
7. John Jacob, Kavya Jha,Paarth Kotak, Shubha Puthran "Educational Data Mining Techniques And Their Applications", IEEE Interna-tional Conference On Green Computing and Internet Of Things (ICGCIoT),2015.
8. S.Hari Ganesh A.Joy Christy "Applications of Educational Data Mining: A Survey ", IEEE Sponsored 2nd International Conference ICIIECS-2015.
9. M. A. Al-Barrak, and M. Al-Razgan, "Predicting Students Final GPA Using Decision Trees: A Case Study", Int. J. Inf. Edu.Tech., vol. 6, no. 7, pp. 528–533, 2016,
10. J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques", 3rd ed., 2012.
11. D. G. Kleinbaum and M. Klein, "Logistic Regression A Self-Learning Text", 3rd ed., New York, 2010.
12. S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Predicting Students ' Performance in Distance Learning Using Machine,", 2004.
13. J. D. Kelleher, B. Mac Namee, and A. D'Arcy, "Fundamentals of Machine Learning for Predictive Data Analytics. Algorithms, Worked Examples, and Case Studies". 2015.

AUTHORS PROFILE

Dr. Bhavesh Patel, Assistant professor in Ganpat University, MCA, Department. I have completed Ph.D in Computer Science. As a Researchers, written 8 International research papers and 4 National Research papers and published in reputed journals. My research area is data mining and artificial intelligence.