2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST) | 978-0-7381-3042-2/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICREST51555.2021.9331003

A Web Based Four-Tier Architecture using Reduced Feature Based Neural Network Approach for Prediction of Student Performance

Md. Anwar Hossen
Dept. of Software Engineering
Daffodil International University
Dhaka, Bangladesh
anwarcsejnu@gmail.com

Rakib Bin Alamgir
Dept. of Software Engineering
Daffodil International University
Dhaka, Bangladesh
alamgir35-903@diu.edu.bd

Arman Ul Alam
Dept. of Software Engineering
Daffodil International University
Dhaka, Bangladesh
armanulalam0@gmail.com

Fatema Siddika
Dept. of Computer Science and Engineering
Jagannath University
Dhaka, Bangladesh
shashi.csejnu@gmail.com

Shah Fahad Hossain
Dept. of Software Engineering
Daffodil International University
Dhaka, Bangladesh
shah35-1996@diu.edu.bd

Md. Shohel Arman
Dept. of Software Engineering
Daffodil International University
Dhaka, Bangladesh
arman.swe@diu.edu.bd

***Abstract*— Enhancing student's performance is a significant part of developing quality education in any educational institute. It is very difficult to get promising student performance without student categorization according to their academic performance as there are different standardized students. In this paper, our aim is to determine the performance of the students. For this purpose, a survey has been conducted on students in our university in order to collect data and to analyze and predict the student category based on their performance. Apart from this, another purpose of this study is to examine the effect of the reduced features on the classification model using state-of-art machine learning algorithms. Here, we propose a workflow of web-based four-tier architecture for the student performance prediction that will define the student's category in order to help them exactly pinpoint their learning capabilities. Hence, we used multiple supervised learning-based machine learning techniques for the prediction of student performance. Each of the student category categorized by considering on the top features. The analysis results indicate that we got the highest performance that is 88.00% by using the Artificial Neural Network (ANN) among the classifiers by showing its superiority to the existing model.**

Keywords*— *neural network, academic performance, chi2, web architecture, supervised learning.

I. Introduction

Student's performance defines how a student is accomplishing his or her studies & responsibilities. Among other aspects of human resource development education is one of the most important one. Nowadays, one of the serious challenges that the education field faces is student's underperformance or trend to drop out. The education institutes work hard to minimize this problem and prepare all the students for the future or their trend to drop out in the future. In 2017 in the United States the overall high school dropout rate was 5.4% according to the National Assessment of Education Progress (NAEP) [1]. In developed countries such as Bangladesh, the dropout rate was 19.89% in 2017 according to Bangladesh Bureau of Education Information and Statistics (BanBEIS) [2]. Authors in their paper [3], used machine learning approach to predict the dropout students or the students who might switch their schools. Also, they have ranked student on the possibility of dropout at high school graduation level. For this reason, they adapted the ML techniques such as Random Forest Search, Logistic regression. Early students' performance prediction is one of the effective solutions to minimize students' failure problem. Analyzing the students' performance can help in this matter. From human biological inspiration ANN has been developed so that it can perform tasks such as clustering, classifications etc. [4]. Additionally, accuracy of Machine Learning techniques may vary. So, studies applying different Machine Learning techniques focused on the accuracy [5] [6].

Consequently, a handful of data from Daffodil University students has been collected using survey questionnaires inspired by a Kaggle dataset. Features like nationality, questions asked in the classroom, meet with academic adviser, student Absence Days, parents Satisfaction, education Status of Parents, batch etc. are included in the dataset. After collecting, the data had been labeled. There were 18 attributes in the dataset. After numerical feature selection, we have come to the conclusion that the student's local Guardian Name has no effect on the outcome. So finally, we create a dataset with 17 attributes from which the top 10 attributes have been selected after encoding so that we could be able to compare the effectiveness of the attributes on the accuracy of the model. We have used Chi2 (Chi-Square) and Random Forest Importance (RFI) algorithm simultaneously to sort the most effective features from both implementation. Thereafter, data has been preprocessed and initial attributes were selected to create Training and Test sets so that the Artificial Neural Network algorithm can be applied to provide a prediction model so that level (high, medium, low) of each student can be identified.

II. Related Works

Nowadays, the classification problem is the most discussed topic to data mining and machine learning researchers. Machine learning acquires significantly high accuracy in the classification-based problem that shows in a different studies. It is a process of predicting the value of a feature or attribute (categorical) based on the values of other attributes. Classes are often called labels or targets. Its sub-processes are: train classifier on a specific portion of dataset to generate its mode.

After that apply the generated model on the rest of the testing dataset. This technique has been used into many fields such as Medical, Economics, Stock Exchange, Fraud Detection and even in Education. Using EDM (Educational Data Mining) techniques, authors has worked with student's performance prediction on student's academic attributes, social attributes and on all features [7]. For this task they applied Decision Tree, Nave Bayes and K-NN classification techniques and determine which one is the most appropriate for early prediction of student performance. Three algorithms were tested for each subsequent steps and then they were compared and finally found that Decision Tree classifier gives the best results when used with students social and academic attributes. Another paper [8], for the students of postgraduate degree, the authors have developed a system of selecting the suitable subject with machine learning approach. In this case, their aim was to detect best subject for a student for post-graduation degree depending on previous educational record to reduce the possibility of dropout of student and guide them to successfully complete post-graduation degree by suggesting them suitable subject. In another research, for distant learners, authors come up with a personal module [9].

III. Methodology

A. Dataset Collecting

In this study, we used the collected dataset from Daffodil International University. This dataset contains 17 attributes and 518 observations.

TABLE I LIST OF HIGHLY CORRELATED FEATURE

Attributes	Type	Description
gender	Categorical	male, female
batch	Numerical	16 - 52
nationality	Categorical	Bangladesh, Somalia
placeOfBirth	Categorical	District Name
department	Categorical	SWE,CSE,MCT, ESDM
semester	Categorical	Summer, fall, spring
section	Categorical	A – I, N, P, R, T
lastSemesterGradePoint	Numerical	2.00 - 4.00
questionAskInTheClassroom	Categorical	high, medium, low
questionAskedInTheClassroom	Numerical	0 - 60
goThroughCourseMaterials	Categorical	high, medium, low
goThroughCourseMaterial	Numerical	high, medium, low
meetWithAcademicAdvisor	Numerical	0 - 9
groupStudyHours	Numerical	0 - 98
studentAbsentDays	Numerical	0 - 21
parentsSatisfaction	Categorical	Yes, No
educationStatusOfParents	Categorical	SSC, HSC, Above HSC

The dataset used in this study has been collected specifically from Daffodil International University Students as shown in Table I. This dataset consists of 518 student's record with 18 features.

B. Handling Categorical Features

The data we have collected, needs some preprocessing before moving further. Such as, both the features and labels contained categorical data and we have handled that by encoding both. Features like gender, place of birth and nationality, department, semester, section, parent's satisfaction, parent's education status are categorical data. Each of this features were taken and made encoded by one hot encoding. To achieve this purpose Panda's get dummies function has been used. The label or outcome is also taken into account as categorical as it contained three different categories: High, Medium, and Low. So, each of this different categories also encoded. To achieve such purpose we have used label encoding and labeled the categories manually. Then we checked for highly correlated features and removed them.

C. Feature Selection

1) Random Forest Importance: It is another popular method for feature ranking. The positive results from each of the decision tree are taken and with the final average result the random forest can be calculated for determining the feature importance. For gathering all the values of feature in one, Scikit-Learn Random Forest library has been utilized [10] [11].

TABLE II RANDOM FOREST IMPORTANCE

Features	Importance
goThroughCourseMaterial	0.101468
studentAbsentDays	0.055337
batch	0.058384
questionAskedInTheClassroom	0.075167
groupStudyHours	0.091830
department_ESDM	0.041830
placeOfbirth_Feni	0.040781
nationality_Nigeria	0.002035
placeOfbirth_Bosasosomalia	0.000170
questionAskeInTheClassroom_H	0.013035

2) Chi Square: It is mainly used to determine feature selection whether input features are relevant to the outcome to be predicted [12]. It determines the highest valued features from the Chi2 statistic test which consist of training and test set. The mathematical equation of Chi2 algorithm has given in the reference [13].

TABLE III CHI SQUARE WEIGHT RESULT

Variable	Chi Square Weight
goThroughCourseMaterial	72.046270
studentAbsentDays	46.797132
batch	39.071508
questionAskedInTheClassroom	36.148678
groupStudyHours	9.818743
department_ESDM	6.441762
placeOfbirth_Feni	5.978561
nationality_Nigeria	5.574314
placeOfbirth_Bosasosomalia	4.771242
questionAskeInTheClassroom_H	4.739952

D. Proposed Architecture

Our System model consists of several components. Firstly, we pre-processed the dataset, after that we weighted the features by applying Random Forest Importance and Chi2 technique. Each of the variables had been sorted and top ten common variables according to their weights were taken for training purpose. After that, we trained the model with different combination of variables and find out that top 3 features (go through course material, student absent days, question asked in the classroom) have the best impact on the training model and provide best accuracy. So, we have train our ANN model with top 3 variables. Fig. 1. shows the proposed System Model:

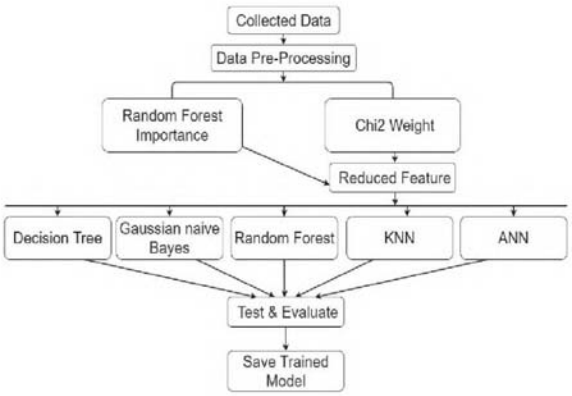

Fig. 1. System Model

The prediction accuracy of the machine learning techniques may vary on different conditions. Hence, most of the studies, applying ML classification techniques have been focused on the prediction and their accuracy for our proposed web-based four-tier architecture. It consists four-tier architecture including machine technique can be used to get, store and test new user data to predict the performance of students. The proposed architecture workflow and contributions of the paper are summarized as follows,

• Tier one focuses on collecting data from students who have completed their secondary, higher secondary school certificate.

• Tier 2 stores those new data into a drive so that these data can be used in future.

•Tier 3 uses saved and trained machine learning classification ANN model to test new data.

•Finally, tier 4 represents the result of the whole system for the user.

The proposed four-tier architecture has been shown in Fig. 2.

In Fig. 2, we designed our model in four tier. In first tier we are focused in collecting data by taking a survey with various questions among the different level of students of Daffodil International University. From this survey we have collected 518 records with 17 attributes such as: (1) Demographic features such as gender, place of birth and nationality. (2) Academic background such as batch, grade, section, department, semester. (3) Behavioral factors such as questions asked in classroom, going through educational resources, participating in group study, students absence days, meeting with advisor and parents satisfaction.

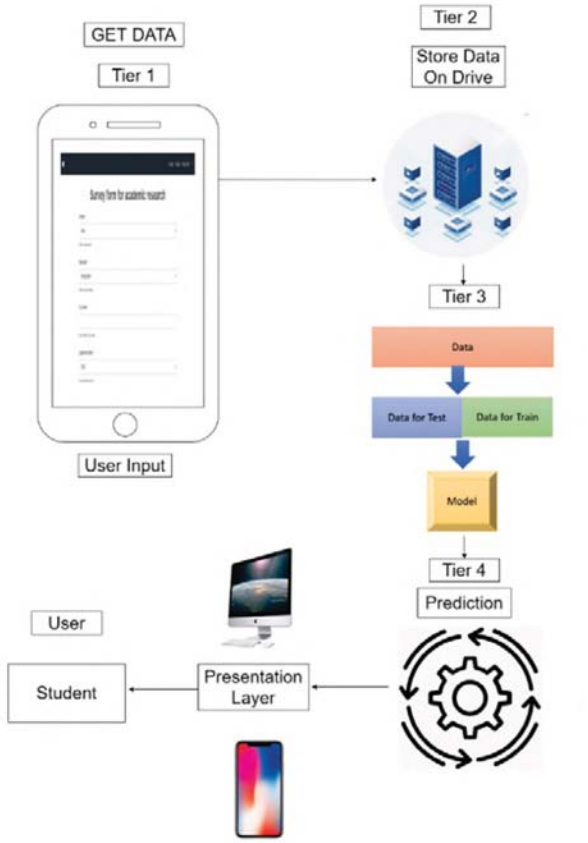

Fig. 2. Proposed architecture for student performance prediction

In tier 2, we preprocessed our collected dataset and initial attributes were selected to create Training and Test set and stored into a drive. We have encoded the categorical variables including the label. After that we weighted the features by applying Random Forest Importance and Chi2 technique. Each of the variables had been sorted and top ten common variables according to their weights were taken for training purpose.

In tire 3, we have separated the training and test set and the test set contained 30% of the whole dataset. we trained the model with different combination of variables and find out that top 4 weighted features have the best impact on the training model and provide best accuracy. Our proposed Artificial Neural Network model are massively parallel systems with large numbers of interconnected simple processors [14]. It has been inspired from biological background that can be utilized to perform certain tasks such as classification, pattern recognition, clustering etc. [15]. So, we have train our ANN model with top 4 variables and save the model for further processing. We have integrated our trained model in a web server.

In tire 4, we represent the whole system result to the user as a web or mobile application.

IV. RESULT AND DISCUSSION

A. Analysis of the Result

In this experiment, we considered different analysis to examine the five machine learning classification techniques for the classification for student prediction. Table IV shows the prediction results of Decision Tree (DT), Gaussian Naive Bayes (NB), Random Forest (RF) and K-nearest neighbors (KNN) and Artificial Neural Network algorithms.

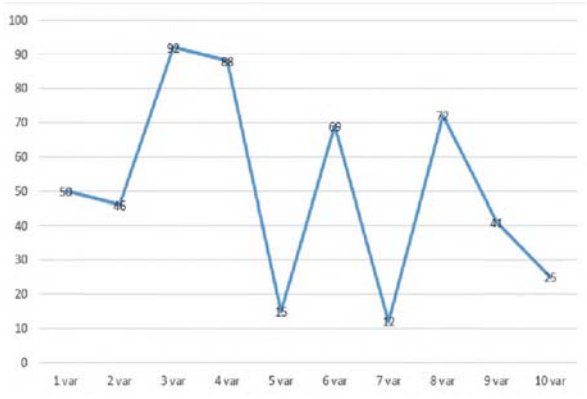

Fig. 3. Neural Network Accuracy Considering Different Variables

TABLE IV Accuracy Comparison of Reduced and Ten Feature

Models	Accuracy (Ten Feature)	Accuracy (Reduced Feature)
ANN	25%	92%
KNN	41%	37%
GNB	53 %	53%
RF	44 %	44%
DT	42 %	41%

Algorithm 1: Student Category Classification

```
Xtrain, Xtest, Ytrain, Ytest =
TrainTestSplit(dataset.features,tSize = 0.30)
j = 0
while j <= dataset:featureColumns:length do
        featureLabels=data
        set.ColumnNames[top 10];
        inputLayerDimension = dataset.featureColumn.lemgth;
        hiddenLayerAct = relu;
        outputLayerAct = sigmoid;
        createModel   =   createANNModel(inputLayerDimension,
        hiddenLayerAct, outputLayerAct);

        createModel. t(Xtrain, Ytrain, epochs = , batchSize = 40);
        Y_pred = classifier.predict(Xtest);
        for prediction in Y pred do
            return scores;
        end
        accuracy:= (TP+TN) / (TP+FP+TN+FN);
        return model accuracy;
        pop one feature from dataset;
        X = create new dataframeColumn.Values;
        Xtrain, Xtest, Ytrain, Ytest =
        TrainTestSplit(dataset.features,tSize = 0.30)
        j += 1
end
```

In the below, Fig. 4. shows the prediction accuracy of four machine learning classifiers for student performance prediction using top ten variables. Now, for Artificial Neural Network(ANN) firstly we have applied the Chi2 algorithm without proposed variables. After that, Random Forest Importance has been applied. The main reason behind doing this was to find out the top n important variables to be selected for the training set. The result of Chi2 and Random Forest Importance on top 24 variables.

B. Performance Evaluation

At first with the top 10 variables we have done a multilayered Neural Network approach. After that, gradually we have reduced one feature at a time to inspect the accuracy difference basis on features. The results obtained for the number of used variables where we can find out that the top four variables (questions asked in the classroom, student absence days, go through course material, batch) provide the best accuracy compared to more than four variables. Also, after analyzing the dataset, our observation has been provided into Fig.4, Fig. 5, Fig. 6, Fig. 7 respectively:

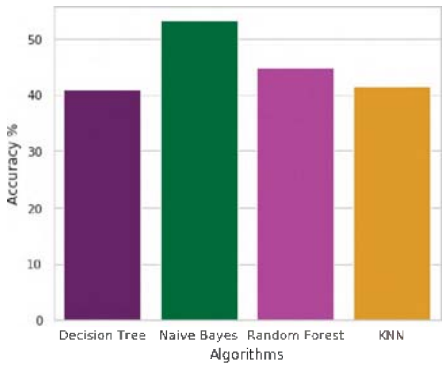

Fig. 4. Prediction accuracy of classification techniques

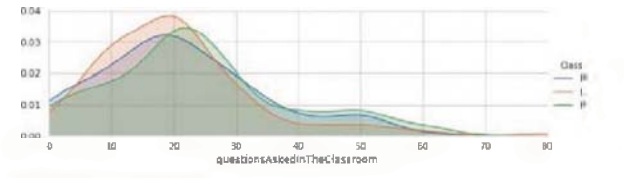

Fig. 5. Ratio of questions asked in the classroom

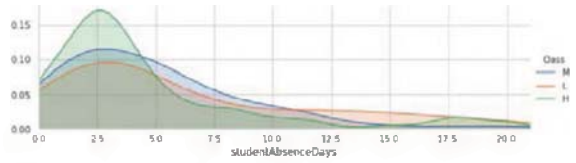

Fig. 6. Ratio of student absence days

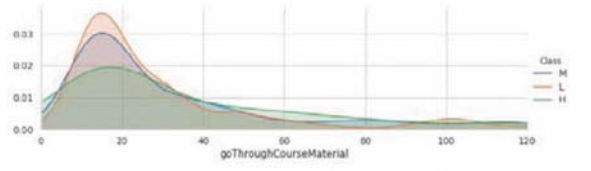

Fig. 7. Ratio of go through course material

C. Result Discussion

From Fig. 7 we can see that, students who went through course materials less than 20 hours have the low performance. On the other hand students who went through course materials more than 50 hours have the high performance. Again, from Figure 5 we can clearly view that, students who were absent more than 8 days have the low performance. On the other hand students who were absent less than 5 days have the high performance. And from Figure 5 we got that the students who asked questions in the classroom less than 25 times has the low performance and students who asked questions more than 25 times have the high performance.

Also, following results clearly shows that the ANN achieved the highest accuracy and KNN the lowest than the remaining algorithms. ANN outperformed the other classifiers and had best performance after reducing variables which is 92%.

D. Comparison with Reference Work

In Table V, we made a comparison that shows our predictive model performance which is better than the reference model [5]. We have attained the highest accuracy using the Random Forest Importance and Chi Square feature selection method from the dataset and accomplished 92% accuracy with reduced feature by applied the Artificial Neural Network (ANN) whereas the best model of reference work accomplished accuracy 85% with reduced feature based model.

TABLE V

COMPARISION WITH REFFERENCE WORK

Model name	Accuracy (Reduced Feature)
Artificial Neural Network (ANN)	92%
Reference work [5]	85%

V. CONCLUSION

In this paper we have presented five supervised learning based machine learning technique on a real dataset of students, collected from Daffodil International University in 2019 academic year to predict performance of students. Afterwards we investigated the variables effect on model accuracy and later we compared the five classifiers performance and evaluated their performance for getting a better insight from the prediction model, we have reduced the features. After reducing the features, we have achieved highest accuracy up to 92% from ANN and achieved lowest accuracy 37% from KNN. So, this ANN model can be effectively applied to predict student performance. However, lots of work remain undone, the performance of the models not up to the mark.

REFERENCES

[1] "Dropout rates | National Assessment of Education Progress (NAEP). " [Online]. Available: https://nces.ed.gov/fastfacts/display.asp?id=16. (accessed: Oct. 10, 2020)

[2] "Higher Secondary dropout rates | Bangladesh Bureau of Education Information and Statistics (BanBEIS). ". http://lib.banbeis.gov.bd/BANBEIS_PDF/Pocket%20Book%20on%20Bangladesh%20Education%20Statistics%202017.pdf. (accessed: Oct. 10, 2020)

[3] Aguiar, E., Lakkaraju, H., Bhanpuri, N., Miller, D., Yuhas, B., & Addision, K. L. (2015, March). Who, when, why: A machine learning approach to prioritizing students at risk of not graduating high school on time. In Proceeding of the *Fifth International Conference on Learning Analytics and Knowledge* (pp. 93-102). ACM.

[4] https://medium.com/technology-invention-and-more/everything-you-need-to-know-about-artificial-neural-networks-57fac18245a1. (accessed: Oct. 10, 2020)

[5] M. M., Alam, K., Mohiuddin, A. K., Das, Md. K., Islam, Md. S., Kaonain, Md. H., Ali (2018). A Reduced Feature Based Neural Network Approach to Classify the Category of Students.

[6] H., Agrawal, H., Mavani (2015). Students' Performance Prediction using Machine Learning. International Journal of Engineering Research & Technology (IJERT).

[7] Hafez Mousa, Ashraf Maghari (2017 August). School Students' Performance Prediction Using Data Mining Classification. *International Journal of Advanced Research in Computer and Communication Engineering.*

[8] A., Usiobaifo, & 2.0., Osaseri. (2016). A machine learning approach for predicting postgraduate students' performance. *Proceeding of INCEDI 2016 Conference. 779-784.*

[9] Li. M., Cohen, W., Koedinger, K. R., & Matsuda, N. (2010, June). A machine learning approach for automatic student model discovery. In Education Data Mining 2011.

[10] Random forest feature importance. (n.d.). blog.datadive.net/selecting-good-features-part-iii-random-forests/. (accessed: Oct. 10, 2020)

[11] RASCHKA, S. M. (2017). PYTHON MACHINE LEARNING -.S.l.: PACKT PUBLISHING LIMITED.

[12] Sklearn.feature selection.chi2. (n.d). https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html. (accessed: Oct. 10, 2020)

[13] Liu, H., &Setiono, R. (1995, November). Chi2: Feature selection and discretization of numeric attributes. In Tools with artificial intelligence, 1995. proceedings. seventh international conference on(pp. 388-391). IEEE.

[14] K Anil, Jain, Mao Jianchang and K. M Mohiuddin, "Artificial Neural Networks: A Tutorial", *Michigan State University*, 1996.

[15] Overview of Artificial Neural Networks and its Applications. https://hackernoon.com/overview-of-artificial-neuralnetworks-and-its-applications-2525c1addff7. (accessed: Oct. 10, 2020)