International Journal of Computer Science and Information Security (IJCSIS),
Vol. 19, No. 1, January 2021

A Comparative Study of Different Machine Learning Techniques to Predict the Result of an Individual Student using Previous Performances

Khalil Ahammad #1, Partha Chakraborty #2, Evana Akter #3, Umme Honey Fomey #4, Saifur Rahman #5

# Department of Computer Science & Engineering
Comilla University, Cumilla - 3506, Bangladesh
[1] khalil@cou.ac.bd
[2] partha.chak@cou.ac.bd
[3] evanaflora@gmail.com
[4] umme.fomey534@gmail.com
[5] saifurrahmany43@gmail.com

***Abstract*—Machine learning is a sub-field of computer science refers to a system's ability to automatically learn from experience and predict new things using the learned knowledge. Different machine learning techniques can be used to predict the result of the students in examination using previous data. Machine learning models can recognize vulnerable students who are at risk and take early action to prevent them from failure. Here, a model was developed based on the academic performance of the students and their result in the SSC exam. This paper also shows a comparative study of different machine learning techniques for predicting student results. Five different machine learning techniques were used to demonstrate the proposed work. They are Naive Bayes, K-nearest Neighbours, Support Vector Machine, XG-boost, Multi-layer Perceptron. Data were preprocessed before fitting into these classifiers. Among the five classifiers, MLP achieved the highest accuracy of 86.25%. Other classifiers also achieved a satisfactory result as all of them were above 80% accuracy. The results showed the effectiveness of machine learning techniques to predict the performance of the students.**

***Index Terms*—Machine learning, Result, Prediction**

I. Introduction

The concept of equal opportunity is a crucial factor that must be taken into consideration when talking about the development of a nation. In the educational factor, this idea is to guarantee every person has the same options for accessing and completing studies. There are some shortcomings in the education system in many developed countries like Bangladesh that it is still trying to overcome. Many students participate in the board exam every year without understanding their overall performance in earlier. As a result, student dropout rates create capital wastage for all actors in the education sector and also impact the institutions' assessment processes. Through evaluating student's previous results of all regular exams using machine learning techniques, teachers can anticipate the students resulting in the board exam. One of the greatest problems many education institutes face is enhancing the efficiency of educational processes in order to increase the success of students. To meet the expectations using machine learning models teachers can identify low-performance students and can update their teaching methods to offer additional guidance to eligible students. The early prediction may also allow students to gain a clear understanding of how well or bad they will do in a course and then take appropriate steps. Machine learning's basic concept is that it can automatically learn from practical experience. There are several supervised and unsupervised types of approaches to machine learning that are used to retrieve hidden information and correlations between data, which will ultimately assist decision-makers to take proper action in the future. Machine learning techniques such as Naive Bayes, SVM, KNN, etc. may be very helpful in predicting the performance of students based on the background and term exam performances of students.

In this work, A dataset was constructed consisting of academic results of different subjects and their respective GPA in the Secondary School Certificate Examination of 400 students from a renowned school[1]. Dataset was cleaned and preprocessed before fitting into the models. Five experiments with five different classifiers namely K-Nearest Neighbors, Support Vector Machine, Naïve Bayes, XG Boost, and Multi-layer Perceptron. All of the classifier's performance was evaluated using different evaluation matrices such as precision, recall, f1-score, and accuracy.

II. Related Work

A grammar-guided genetic programming algorithm, G3P-MI, was implemented by the University of Cordoba to predict whether the student will fail or pass a certain course [7]. The algorithm has a 74.29% accuracy. A platform that can forecast student performance using machine learning algorithms has been created by the Vishwakarma Engineering Research journal [8]. Two criteria were used, attendance of students and the related subject marks. A model for predicting student performance has been developed by Somiya College

[1] Feni Model High School, Feni, Bangladesh

TABLE I
FEATURE DESCRIPTION

Feature name	Description	Values	Data type
English	Marks of English in Exams	0-100	Integer
Bangla	Marks of Bangla in Exams	0-100	Integer
General Math	Marks of General Math in Exams	0-100	Integer
Physics	Marks of Physics in Exams	0-100	Integer
Chemistry	Marks of Chemistry in Exams	0-100	Integer
Biology	Marks of Biology in Exams	0-100	Integer
BDS	Marks of Bangladesh Studies in Exams	0-100	Integer

Mumbai [9]. The team correctly expressed the correlation with a student's past academic results. With data set growth, neural network output improvements have been documented. And their precision hit 70.48 percent. Artificial neural networks (ANNs) were used by De Albuquerque et al. [10] to forecast the success of the student in exams. This model used features such as grades, study periods, and school ratings for features, and high precision of 85% was obtained. Kotsiantis et al. [11] estimated the success of the pupil on final tests using techniques of machine learning. They used demographic features like sex, age from an e-learning method as inputs. They found that the strategy of Naïve Bayes obtained a higher average accuracy (73%) than the alternatives. The Eindhoven University of Technology performed an assessment of the efficacy of machine learning for dropout student outcome prediction [12]. The basic approach was to use various machine learning approaches such as CART, BayesNet, and Logit, to construct numerous prediction models. By using the J48 classifier, the most effective model was developed. Researchers from three separate universities in India undertook a related analysis [13]. A data set of university students was analyzed by different algorithms, after which the forecasts' accuracy and recall values have been compared. The architecture of the ADT decision tree provided the most correct outcomes. This [14] was done at the University of Minho, Portugal. Using decision trees, random forests, vector support machines, and neural networks it evaluates whether the student had passed the test in math and Portuguese language subjects were included in the data set. Such techniques were evaluated in terms of accuracy. Another paper [15], predicts student's success at the beginning of an academic cycle, based on their academic record. The research was performed on historical data stored within the information system of Masaryk University. The findings demonstrate that this technique is as successful as with machine learning techniques, such as support vector machines and it came to an accuracy of 85%.

III. Methodology

The full methodology of the proposed study for the student's performance prediction is shown in figure 1. The proposed study was designed on supervised machine learing techniques. The first step was to collect data of students regular exams marks. Then, the dataset was preprocessed by normalizing the data. After that, the preprocessed dataset was split into train and test parts. Five supervised classifiers were used namely, Naive Bayes, K-Nearest Neighbour, Support Vector Machine, XgBoost, and Multi-Layer Perceptron. These classifiers were trained with the training dataset. Finally, each classifier was evaluated with the test dataset by forming some evaluation matrix.

A. Data Description

The dataset of the student's performance was collected from Feni Model High School in Bangladesh. The performance data contained student's marks for the different subjects of class (9-10) school students of the academic year (2013-2014) and (2016-2017). After eliminating incomplete data, the dataset comprised of 400 students in the dataset.

The final dataset contained 31 columns. First two columns of the dataset are the Student name and his/her respective id. The latter 28 columns are the marks of 7 major and common subjects of science background students. For each student, the marks of four exams such as half-yearly examination, annual examination, pre-test, and test -examination were included. The final column is each student's GPA which they achieved in the Secondary School Certificate (SSC) Examination.

Student's GPA is the only Predictive variable of the dataset. Possible values of the Student's GPA is A+, A, A-, B, C, D, F follows the SSC Grading System from Intermediate and Secondary Education Boards, Bangladesh [17]. Here, A+ is the highest result that can be achieved by a student, and F is the lowest. The dataset consisted of 400 data items of students marks in the different subjects throughout the academic year and their performance in the SSC Exam. In the dataset, there were most data from A- category, precisely 93 data items. The lowest was from the D category. 58 students achieved the highest GPA A+ and 28 students failed in the exam.

B. Data Preprocessing

To ensure the maximum performance from the models the dataset was preprocessed before fitting into the models. Because of the numerical nature of our data, there was not much to preprocess. Preprocessing was done in two steps. Normalization and Label Encoding. RobustScaler was used for data normalization. Label Encoder was to transform categorical values like A+, A, B, etc into numerical values. Machine Learning algorithms work better with numerical values than categorical values.

C. Classification Model

After Preprocessing, the dataset was split into the training dataset and test dataset. The test data were 20% of the dataset.

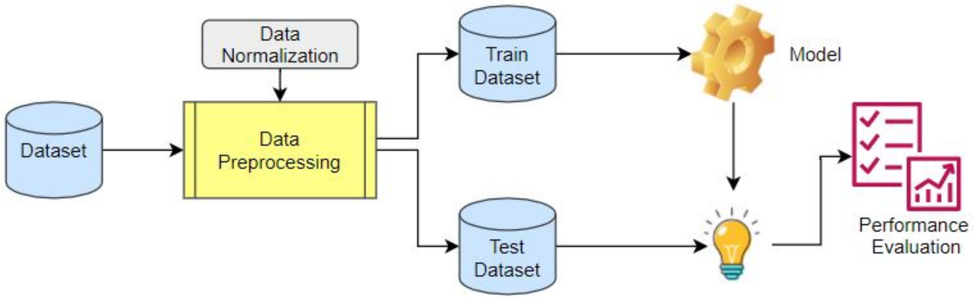

Fig. 1. Full methodology

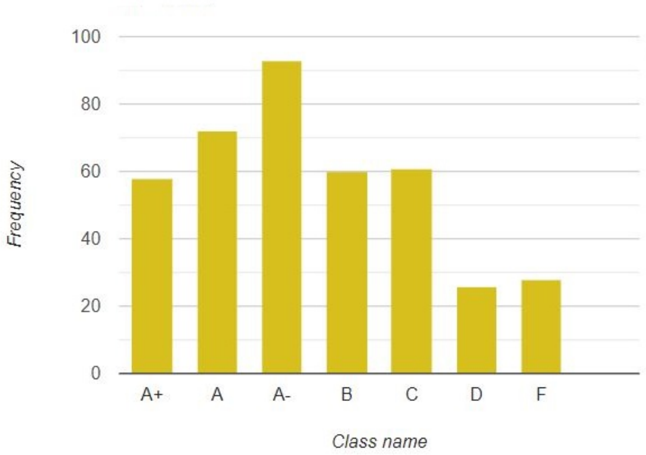

Fig. 2. Frequency of each GPA class in dataset

Five supervised classifiers were trained with 320 data items from the training dataset. These classifiers are Naive Bayes, K-Nearest Neighbour, Support Vector Machine, XgBoost, and Multi-Layer Perceptron.

Naive Bayes: The Naive Bayes Algorithm (NB) is a basic classification technique based on probability theory [1].

$$P(c|X) = \frac{P(X|c)P(c)}{P(X)} \qquad (1)$$

Here, c is performance class and X is the marks of individual subject as dependent feature vector (of size n) where: $X = (x_1, x_2, x_3, ..., x_n)$

K-Nearest Neighbors (KNN): In KNN, The nearest neighbor is determined according to the k-value that specifies the number of nearest neighbors to be considered and thus defines the sample data point class[3]. Minkowski distance formula was used in the proposed study for calculating distance of an observation from centroid.

$$D(X,Y) = (\sum_{i=1}^{n} |x_i - y_i|^{p})^{\frac{1}{p}} \qquad (2)$$

Support Vector Machine (SVM): SVM wokrs by forming a hyperplane (a decision boundary that maximizes margins between classes) in an N-dimensional space that classifies data points distinctly [16]. For linear kernel of SVM the equation for prediction for a new input (X) :

$$f(X) = sum(X * Xi) \qquad (3)$$

Here, X is the marks of individual subject as feature vector and Xi is support vectors that formed the hyperplane using training data.

XGBoost: Another machine learning strategy that belongs to the category of ensemble methods is Gradient Boosting [6]. It works by constructing a strong prediction model by combining multiple weak prediction models that use multiple variants of the same training data [5]. In the proposed work, various parameters were adjusted to minimize the prediction error, such as maximum tree depth, minimum loss reduction needed to make a further partition on a leaf node of the tree, or phase size shrinkage used in the update to avoid over-fitting.

Multi-layer Perceptron (MLP): MLP is especially appropriate for making a classifier for the data that sets the example of vector attribute values in one or more classes. The Multi-layer perceptron is commonly used when there is very little knowledge about the structure of the problem.

A one-hidden-layer MLP can be written in matrix notation as,

$$y\prime = \sigma(W_2\sigma(W_1x + b_1) + b_2) \qquad (4)$$

Here, $y\prime$ is the output, x is input vector, b_1 and b_2 is bias, W_1, W_2 is weights, and σ is the activation function. A loss function is defined to evaluate the performance of the classifier.

$$SSE = (y - y\prime)^2 \qquad (5)$$

Here, SSE is Sum of Squares Error for the predicted output $y\prime$ as y was the actual output.

To get the best performance some of the parameters of the classifiers were tuned. The best set of parameters used in the proposed model given below:

- KNN: optimal set of parameters is ’n neighbors’: 15, ’n_jobs’: -1, ’leaf_size’: ’100’

- SVM: optimal set of parameters is 'C': 10, 'gamma': 0.1, 'kernel': 'linear'
- XgBoost: optimal set of parameters is 'objective': 'reg:linear', 'max depth': 5, 'learing_rate': 0.1 , 'alpha': 10
- MLP: optimal set of parameters is 'hidden_layer_sizes': (100,100,100), 'activation': 'relu', 'solver': 'sgd', 'alpha': 0.0001, 'tol': 0.000000001

IV. Experimental Result Analysis

The performance was evaluated by testing those classification models on 80 testing data items. In test dataset, there were 13 data items from the A+ category, 12 from A, 21 from A-, 11 from B, 15 from C, 2 from D, and 6 from the F category. Four evaluation matrix namely, Precision, Recall, F1-score, and Accuracy were used to evaluate the performance of each classifier. F1-score for each classifier in each performance category are shown in table II. Here, MLP and NB achieved highest f1-score 0.93 in predicting category C and F. Average highest f1-score was attained in category F and the lowest average was in category D.

TABLE II
F1-scores for each result class prediction

Class Name	F1 Score				
	NB	KNN	SVM	XGBoost	MLP
A+	0.81	0.87	0.82	0.92	0.83
A	0.78	0.85	0.83	0.86	0.80
A-	0.88	0.84	0.88	0.82	0.89
B	0.76	0.50	0.73	0.76	0.82
C	0.84	0.90	0.83	0.77	0.93
D	0.40	0.67	0.40	0.40	0.67
F	0.93	0.92	0.82	0.89	0.92

In five classifiers, the MLP performed highest from others with an accuracy of 86.25% and an average weighted F1-score of 0.86. Table III shows the value of four evaluation matrices for each classifier. The weighted average value was used for Precision, Recall, and F1-score. Apart from MLP, other classifiers also showed satisfactory results. KNN was in the second with an accuracy of 82.5% and an F1-score of 0.81. In F1-score, Naive Bayes scored 0.82 though it has slightly less accuracy of 82% than KNN. XGBoost Performed the lowest accuracy of 81% and an F1-score of 0.81.

TABLE III
Performance evaluation for each classifier

Classifier	Precision	Recall	F1-Score	Accuracy
NB	0.83	0.82	0.82	82%
KNN	0.84	0.82	0.81	82.5%
SVM	0.82	0.81	0.81	81.25%
XGBoost	0.82	0.81	0.81	81%
MLP	0.87	0.86	0.86	86.25%

ROC curve identifies how well a classifier can distinguish between classes. ROC curve for NB is shown in figure 3, for KNN in figure 4, for SVM in figure 5, for XGBoost in figure 6, and for MLP in figure 7. ROC curve shows the comparison between the true-positive rate and false-positive rate. Best ROC AUC value of 0.81 was attained by the NB and XGBoost classifier.

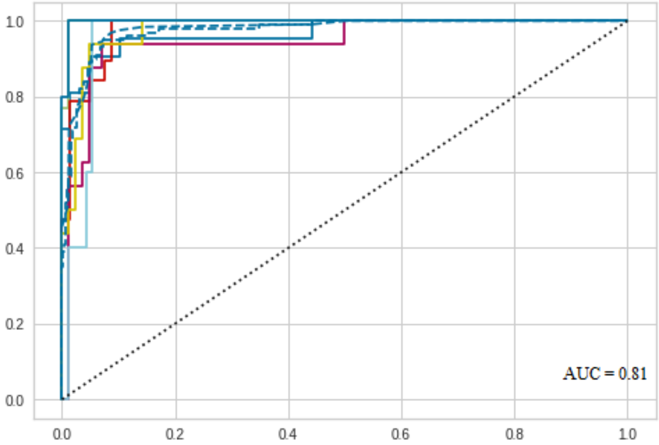

Fig. 3. ROC curve for NB

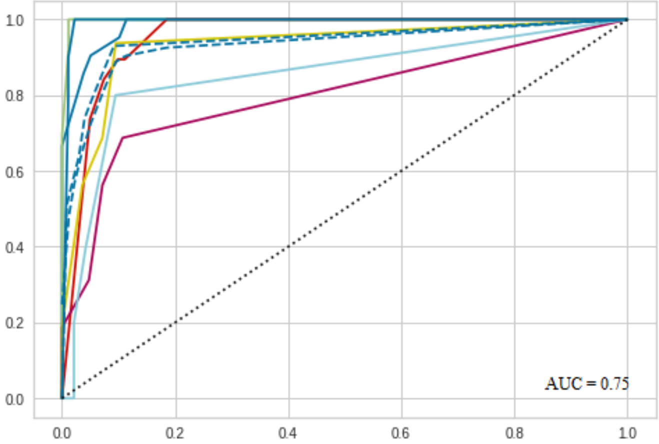

Fig. 4. ROC curve for KNN

Result Comparison: A comparative study with previous works on student results prediction was done here. In the proposed study, for predicting student results in the SSC exam, only regular term exam marks were used as features. In the comparison table IV, the proposed work was compared with previous works that used exam marks or grades as features. This model outperformed with 96.25% accuracy in all the previous work shown here. The closet accuracy of 96.19% was achieved by Dinh Thi Ha et al. [20] using Naive Bayes and MLP classifiers. An artificial neural network model with grades and environmental data of students attained 77.04% accuracy in [19]. An accuracy of 79% was achieved using the LDA approach [21] which is also lower than the proposed model. An accuracy of 81.73% achieved while predicting the students performance of the public universities of Bangladesh in [23].

TABLE IV
COMPARISON STUDY WITH PREVIOUS WORK

Algorithms	Attributes	Accuracy	Paper Information
MLP	Marks	**86.25%**	**Proposed Study**
NB, MLP	Marks	86.19%	Dinh Thi Ha et al. (2020)
ANN	Grades Demographics Environment	77.04%	Hussein Altabrawee et al. (2019)
LDA	Grades	79%	Abu Zohair (2019)
KNN, SVM	Internal assessment Grades	80%	Mayilvaganan el al. (2014)
XgBoost	Grades Students demographic	73%	Osmanbegovic et al. (2008)

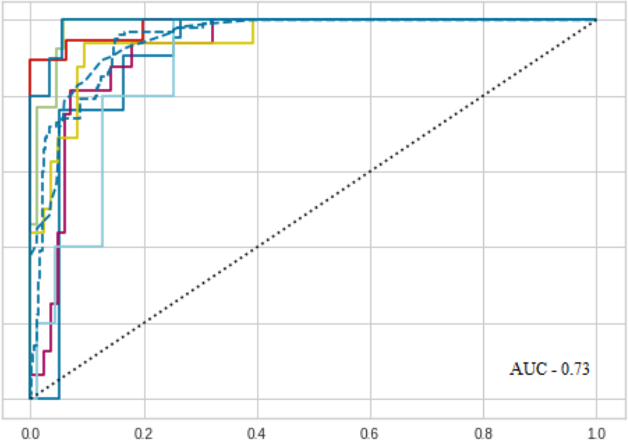

Fig. 5. ROC curve for SVM

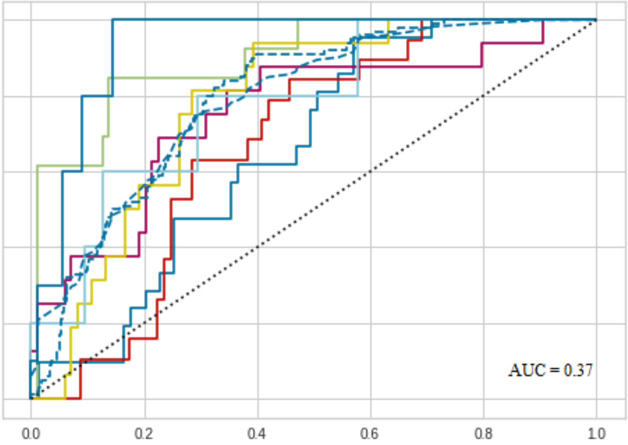

Fig. 7. ROC curve for MLP

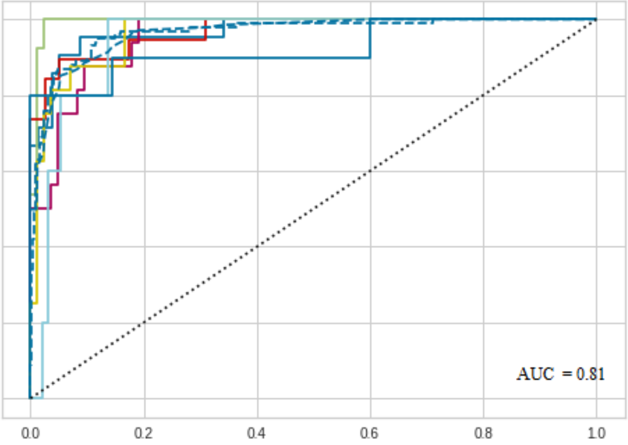

Fig. 6. ROC curve for XGBoost

V. CONCLUSION

This paper is very versatile in predicting student performance. The prediction of student performance gets tough day by day as factors affecting student performance does not always remain limited only to scores in the previous regular exams. Proposed models can be used to early recognize vulnerable students who are at risk and take early action to prevent them from failure and that can help us take necessary measures to enhance the quality of education in the institution. With a relatively small dataset, the proposed models performed good results as all the classifier's accuracy was more than 80% where MLP achieved the highest 86% prediction accuracy. In the proposed work, the research work was conducted on a limited dataset. The collected data were from only one school and two academic years. A large dataset from different schools that contain student results from more academic years can give a better understanding of student's academic success prediction. Factors used here for predicting student's GPA are only the marks of regular exams throughout the academic year. Other environmental factors such as the school environment, residence environment, teaching quality of teaches, etc also play a great role in student's performance in exams. These factors were not included in the proposed study. In future work, these environmental factors as well as regular exam marks will be considered. Different neural network structures such as CNN, RNN, etc will be used with a large dataset in future work.

REFERENCES

[1] I.H. and Frank E. (2000), Data Mining – Practical Machine Learning Tools and Techniques, Second edition, Morgan Kaufmann, San Francisco.

[2] P. M. Arsad, N. Buniyamin, J.-l. A. Manan,2013, A neural network students' performance prediction model (nnsppm), in: Smart Instrumentation, Measurement and Applications (ICSIMA), 2013 IEEE International Conference on, IEEE, 2013, pp. 1–5.

[3] T. N. Phyu, “Survey of Classification Techniques in Data Mining,” Int. MUlticonference Eng. Comput. Sci., vol. I, pp. 18–20, 2009.

[4] V. Vapnik.1995, The Nature of Statistical Learning Theory. NY: Springer-Verlag.

[5] Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16, pages 785–794, New York, NY, USA. ACM.

[6] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5):1189–1232.

[7] Amelia Zafra, Cristobalal Romero, Sebastian, ”Predictic Academic Achievement Using Multiple Instance Genetic Programing” in 9th International Conference on Inteligent Systems Design and Apllications.January 2009.

[8] 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.

[9] Havan Agrawal, Harshil Mavani”Student Performance Prediction using Machine Learning ” (March 2015)

[10] Arnold KE, Pistilli (2012) Course signals at purdue: using learning analytics to increase student success. In:2nd International conference on learning analytics and knowledge (LAK’12), pp 267–270

[11] Wu Zhang, Syed Muhammad Raza Abidi and Sadaqat Ali ”Using machine learning to predict student difficulties from learning session data”,10 February 2018

[12] Thai Nge, N., Janecek, P., Haddawy, P. A comparative analysis of techniques for predicting academic performance. In Proc. of 37th Conf. on ASEE/IEEE Frontiers in Education, 2007.

[13] Yadav et al. Using Machine Learning to Predict Student Performance (2012)

[14] Cortez, Paulo Silva, Alice Maria Gonçalves”Using data mining to predict secondary school student performance,BRITO, A. ; TEIXEIRA, J., eds. lit. – “Proceedings of 5th Annual Future Business Technology Conference, Porto, 2008”.

[15] Al-Barrak, M.A., Al-Razgan, M.: Predicting students final gpa using decision trees: a case study. Int. J. Inf. Educ. Technol. 6(7), 528 (2016)

[16] Zhang Y. (2012) Support Vector Machine Classification Algorithm and Its Application. In: Liu C., Wang L., Yang A. (eds) Information Computing and Applications. ICICA 2012. Communications in Computer and Information Science, vol 308. Springer, Berlin, Heidelberg.

[17] SSC Grading System - Education Board Bangladesh. http://www.educationboard.gov.bd/grads.htm

[18] M. Mayilvaganan, D. Kalpanadevi, Comparison of classification techniques for predicting the performance of students academic environment, in: Communication and Network Technologies (ICCNT), 2014 International Conference on, IEEE, 2014, pp. 113–118.

[19] Altabrawee, Hussein and Ali, Osama and Qaisar, Samir. (2019). Predicting Students’ Performance Using Machine Learning Techniques. JOURNAL OF UNIVERSITY OF BABYLON for pure and applied sciences. 27. 194-205. 10.29196/jubpas.v27i1.2108.

[20] Dinh, Thi and Dinh, Ha and Pham, Thi and Loan, and Cu Nguyen, Giap and Thi, Nguyen and Thi Lien Huong, Nguyen. (2020). An Empirical Study for Student Academic Performance Prediction Using Machine Learning Techniques. International Journal of Computer Science and Information Security.

[21] Abu Zohair, L.M. Prediction of Student’s performance by modelling small dataset size. Int J Educ Technol High Educ 16, 27 (2019). https://doi.org/10.1186/s41239-019-0160-3

[22] E. Osmanbegovi, M. Sulji. Data mining approach for predicting student performance, Economic Review 10 (1).

[23] Zulfiker, M.S., Kabir, N., Biswas, A.A., Chakraborty, P., Rahman, M.M.: Predicting students’ performance of the private universities of Bangladesh using machine learning approaches. Int. J. Adv. Comput. Sci. Appl. 11(3), 672–679 (2020).