A machine learning approach to Predict the Engineering Students at risk of dropout and factors behind: Bangladesh Perspective

Sheikh Arif Ahmed
Computer Science and Engineering,
International Islamic University
Chittagong.
Chittagong, Bangladesh.
sheikharif1993@gmail.com

Shahidul Islam Khan
Computer Science and Engineering,
International Islamic University
Chittagong.
Chittagong, Bangladesh.
nayeemkh@gmail.com

Abstract— **Dropout rate in Bangladeshi universities getting high day by day. Especially in engineering subjects. Massive number of students taking various engineering subjects for their under graduation. However, the completion rate is low. There can be mainly two types of reasons behind dropout- Academic or personal reasons. The target of this study is to find the factors behind the high dropout rate in Bangladeshi universities engineering subjects and also to detect risky profiles for dropout so that their dropout can be prevented. Current and previous student's data is analyzed to find the factors and students at the risk of dropout, which can be useful for developing new strategies in the education system by universities or other educational authorities. SVM, random forest, neural network, etc. were used for creating the prediction model.**

Keywords—Machine learning, neural network, dropout detection

I. Introduction

Like other countries, student dropout is a threat to the Bangladeshi educational system. In the Bangladeshi system, minimum quality to getting a standard government or non-government job is graduation. Already millions of student completing graduation every year and competing for a job. So drop out student will be a national problem too.

Various study shows that there are some personal and some institutional or course-related reasons behind the dropout. Personal problems can be –

a. Lack of prerequisite knowledge base from previous education- Many of the students while in S.S.C. or H.S.C. do not get a proper education. Alternatively, sometimes, they do not get the motivation to study. [1]

b. Socio-Economic condition- Most of the students in Bangladesh have an economic problem. Sometimes their family has to bear the higher education cost. There are situations family cannot continue bearing the expenses. Students who have low economic condition do tuition to earn money with all their time and energy.[2]

c. Fit between major of study and personality- Most of the students do not get the major of under graduation what they want. If their major and personality do not match it will be a significant problem[3,4]

Alternatively, the problem can be institutional or academic course related-

a. Teaching quality- Sometimes, a teacher with an excellent academic background does not have the teaching quality to attract students.

b. Quality of management- Quality of university management is an excellent factor to student dropout

Many other reasons researcher are trying to find out so that they can predict before a student dropout and can take proper care to avoid the dropout.

Some models are statistical, and some are by data mining. As this mining is on educational purpose, these are known as educational data mining (EDM). We have also used EDM with machine learning to analyze and made a prediction model.

If a student completed graduation with an excellent academic background or bad, then we took his profile as a safe profile and whoever is already dropped out, their data is taken to predict future dropout risky profile.

II. Related Works

As dropout rate is increasing all over the world, many researchers trying to find the actual reason with their study.

Vinayak and Prageeth[5] used EDM to predict the risky profile for dropout. They have taken 54 attributes including personal or health data of a student with academic data. Firstly they extracted important features with various algorithms then made a prediction model using a Bayesian network. Their precision is 0.883. They have used data of only 50 students.

Otgontsetseg et al. [6] used 717 students online activity data to predict which profile might drop out in the middle of the semester using a decision tree based prediction system. Their accuracy is 79%

Prediction model designed to find the prospective dropout by tracking online activity, which can predict 40%-50% of students at risk. [7].

Costa et al. used four types of machine learning algorithms named decision tree, SVM, naive Bayes, and neural network for only a course "Introductory programming" [8].

Boris et al. used different kind of machine learning algorithm to predict dropout student using various academic data from an institution only. They collected data from the institution, not from the students. They also collected the data about financial aid taken by students.[9]

Marcell et al. advised a model to predict prospective student dropout using machine learning algorithm where he used data of 15825 students from Budapest University. [10]

Tio et al. worked on nonacademic data to predict students dropout at the university level. His main concern was psychological, financial and personality type. [11]

There are many other works related to this study in many countries i.e., However, most of them have some restrictions-

a. Few works are only for a course

b. Only for a university

c. Low accuracy

d. Few works only used online data; few used only personal or academic data.

e. Most importantly, dropout defined by non-graduate students who started their course, but what if they complete a four year under graduation program in more than six years?

Motivation:
This study solves most of these issues while predicting with a good accuracy rate. A good model was needed to predict dropout the engineering sdtudetns in risk. Also it is necessary to see which kind of featues could be cruicial for predicting more accurately.

a. We have taken personal, academic, and also institutional data to check if there any co-relation between the dropout and those data.

b. This data was collected from 12university of different cites in Bangladesh. Among them, few are private, and few are public.

c. Unlike other class out is clearly defined by three categories. This work included those students in a separate category in term of drop out whoever taken more than six years to complete their graduation.

d. Comparison of the prediction with personal, academic and also using the both kind of data is done.

e. Important featues are found which are co related mostly to the dropout.

III. Methodology & Implementation

Previously it was told that there are few reasons behind student dropout including personal and academic reasons. We have divided these caused in three groups. After a deep study of previous works, discussing with experts and depending on the previous results, we have made a questionnaire of 28 questions. Then preprocessed the data for analyzing with machine learning and statistical methods. Figure 1 shows the method we followed.

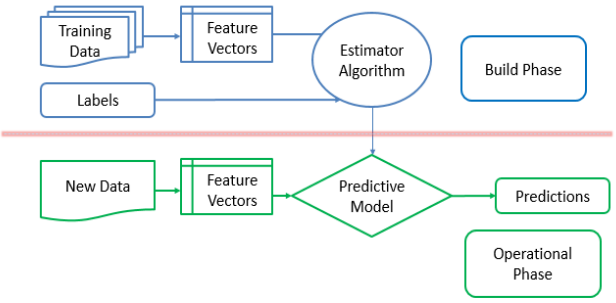


Fig. 1 Methodology

This methodology follows a few steps-

1. Collection of Data
2. Preprocessing
3. Prediction model and knowledge development
4. Using the model for risky profile prediction

A. Data collection

We made a list of 54 attributes then selected 28 most important, which can be important factor for dropout. These are very confidential data, so we collected those data anonymously without any identity. We have taken these survey by hardcopy and by google form. Our data was collected from 480 students of various Bangladeshi university, where the class label was if a student is graduated or not?

B. Preprocessing

We had some missing data which we imputed by mean[12]. We have collected our data in a normalized way so that it can easily be analyzed, i.e., most of the data is taken on a scale of 3. Here in Table 1 variable list and possible answer set is shown described briefly then-

TABLE I. VARIABLE TABLE

	Description	**Variables**
Personal Data	Name	
	Gender {M,F}	
	Family income range {1,2,3}	FIR
	Hour of Tuition service to other-{1,2,3}	HOT
	Daily time spent at social media {1,2,3)	TSSM
	Time spent with friends{1,2,3}	TSF
	Lives in/with {1,2,3}	LWF
	Position in siblings{1,2,3}	PS
	Distance of University{1,2,3}	DU
	Any law breading during Student life	LB
	Connected with student politics	SP
	Currently, in a profession with related subject to graduation?	PRG
	Married or unmarried {1,2}	MU
Academic Data	S.S.C result {1,2,3}	SSC
	H.S.C. result {1,2,3}	HSC
	Study gap/fail in S.S.C. or H.S.C.{1,2}	SG
	Number of Retake course{1,2,3}	RC
	Semester gap in university {1,2,3}	SGU
	Math score in S.S.C{1,2,3}	MSSC
	Math score in H.S.C{1,2,3}	MHSC
	Number of discollegiate subjects for poor attendance {1,2,3}	DFA
Institutional Data	University Type? {1,2}	UT
	Number of P.H.D. holder in university? {1,2,3}	PHDU

	Industry collaboration in university?{1,2}	IC
	Ragging in campus? {1,2}	RIC
	Club based extra curriculum activities {1,2,3}	ECA
	Special encouragement program for good results? {1,2}	EP
	Treatment program for dropout prevention? {1,2}	TP
Prediction	Completed Graduation? {1,2,3}	CG

Here is a short description of some important variables:

FIR- As we have seen, the economic status of a family contribute a lot for dropout we have kept this in our questionnaire. There is three possible answers: 1. Bellow 10k 2. Bellow 30k 3. Bellow 50k 4. Others

HOT- Hour of Tuition. This question is related to the previous one. Maximum students give tuition for extra income. However, if a student gives much tuition to support their educational expenses or family, it is a problem. They will not get much time for their study. There are three possible answers: 1. Less than 3 hours 2. Less than 6 hours 3. More than 6 hours.

TSSM- Social media is nowadays a massive source of time wastage for students. They waste much time doing nothing but watching things in social media. There are three possible answer-

TSF-Few students spent too much time with friends. Even they do not attend class due to gossiping with friends. There we have kept three values- No, 1-2 hour, More Than 2 hour.

LWF-Lives in/with is a crucial factor for student dropout. Many students living outside the home can not study due to the environment. Here we kept three value- With family, in the dormitory or in Outside.

PS- Position in siblings. This is also critical. Values are- first, middle, last

DU- Distance of university. Some universities are out of the city. Students are living in the city. This is a very critical factor. Values are- within 3 kilometers, within 10 kilometers, more than 10 kilometers.

LB- Law breaking students are violent and less concentration on their study. Values are yes and no.

SP- Connected with student politics or not. Values are yes or no.

PRG- Profession related to the subject of graduation. Values are yes and no.

MU- Married or unmarried.

SSC- S.S.C. (Secondary School Certificate) GPA

HSC-H.S.C(Higher Secondary School Certificate) GPA

SG- Study gap in S.S.C or H.S.C. Yes or No.

RC- Number of Retake course in under graduation.

SGU- Semester gap in University.

MSSC- Math score in S.S.C depending on grade range we kept three range.

MHSC-Math score in H.S.C depending on grade point we have kept three range.

DFA- Number of subjects students could not sit the exam because of poor attendance in those subjects.

UT- Type of university. Public or Private.

PHDU- Number of P.H.D. holder in university.

IC- Industry collaboration in university. Yes or no.

RIC- Ragging in Campus. Yes or No.

ECA- Extra curricular activiites in university.

EP- Encouragement program for good results. Yes or No.

TP- Treatment program for students to prevent dropout. Yes or No.

CG- Completion of Graduation. As we have three types of students. 1. Graduated 2. Graduated but taken more than six years (Graduated lately) 3. Not graduated. We can say students whoever graduated safely are not in risk of dropout, which is “Safe” profile, whoever taken more than six years to complete a four-year graduation program is “Medium” profile. Moreover, finally, those who could not graduate are Risky profile.

Figure-2 shows the heat map depending on these three types of category

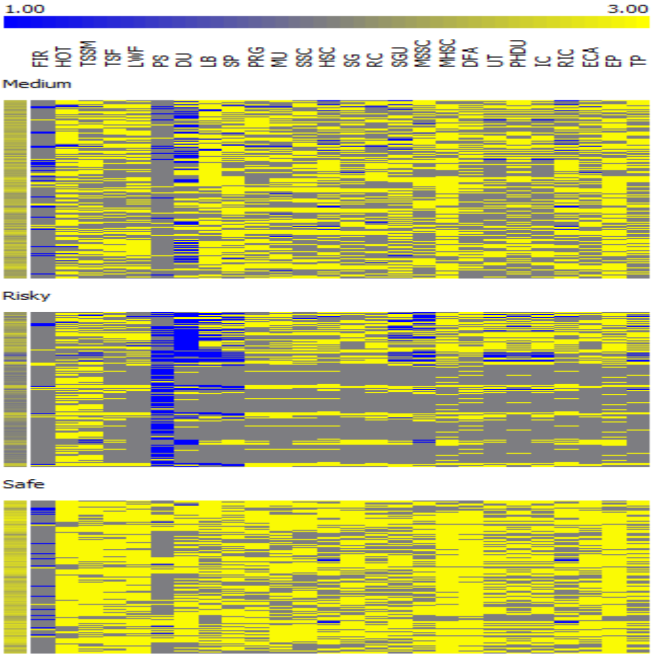


Fig.2 Heat map

C. Prediction model and knowledge development

We have used different algorithms to predict these profiles. As we have three types of data, namely personal, academic, and institutional, we have made our model firstly only with personal data then with academic data and finally compared those with all three category type. Here is an overview of our model in figure 3-

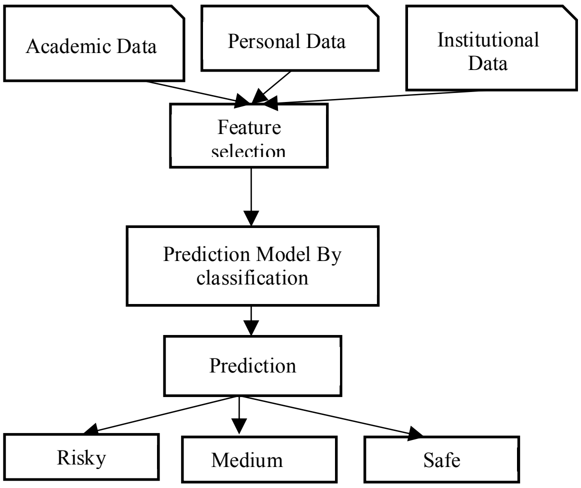


Fig.3 Model Overview

Feature selection is an essential part of this work. For this, we have used the gain ratio.

Gain ratio: Amount of information gained by knowing the value of the attribute[13]

Gain Ratio for an attribute P =$\frac{Gain(A)}{Split\ (A)}$

Where ,

- Gain=(Entropy of distribution before the split)–(entropy of distribution after it)
- Entropy(P_1,P_2,..,P_n)=−P_1log(P_1)−P2log(P_2)−⋯−P_n log(P_n) = $-\sum_1^n P(n)\ logP(n)$
- Split Info(A) = $f(x) = \sum_1^n \left(\frac{|An|}{|A|} \times log2 \frac{|An|}{|A|}\right)$, where A = Attribute

Not only that feature selection will accelerate the algorithm implementation but also it will give the essential factors which are related to student dropout. These attributes can be used in the future to overcome the dropout problem. Alternatively, students can judge themselves if they are at risk or not by these factors. Our selected feature by "Gain ratio" value shown in Table 2. Here we have selected the within the threshold 0.1

TABLE II. FEATURES SELECTION

Selected Important Features	Gain Ratio
Lives in/with	0.1625
Connected with student politics	0.155
Distance of University	0.1433
Special encouragement program for good results?	0.1391
Math Score in HSC	0.1295
Time spent with friends	0.1295
Math Score in SSC	0.1134
Club based extra curriculum activities	0.0971
Study Gap in SSC/HSC	0.0944
Number of discollegiate subjects for poor attendance	0.0944

Classification:

Now we have our class labels. We can build a classification model to predict whether a user is safe or unsafe using those labels. We have used several methods to do that. Figure 4 below clearly shows that we created a classification model with three types of data. Firstly taken a full data set, in the second one taken only personal data and finally only with academic data. In the following section, we will see the comparison of results for each type of model.

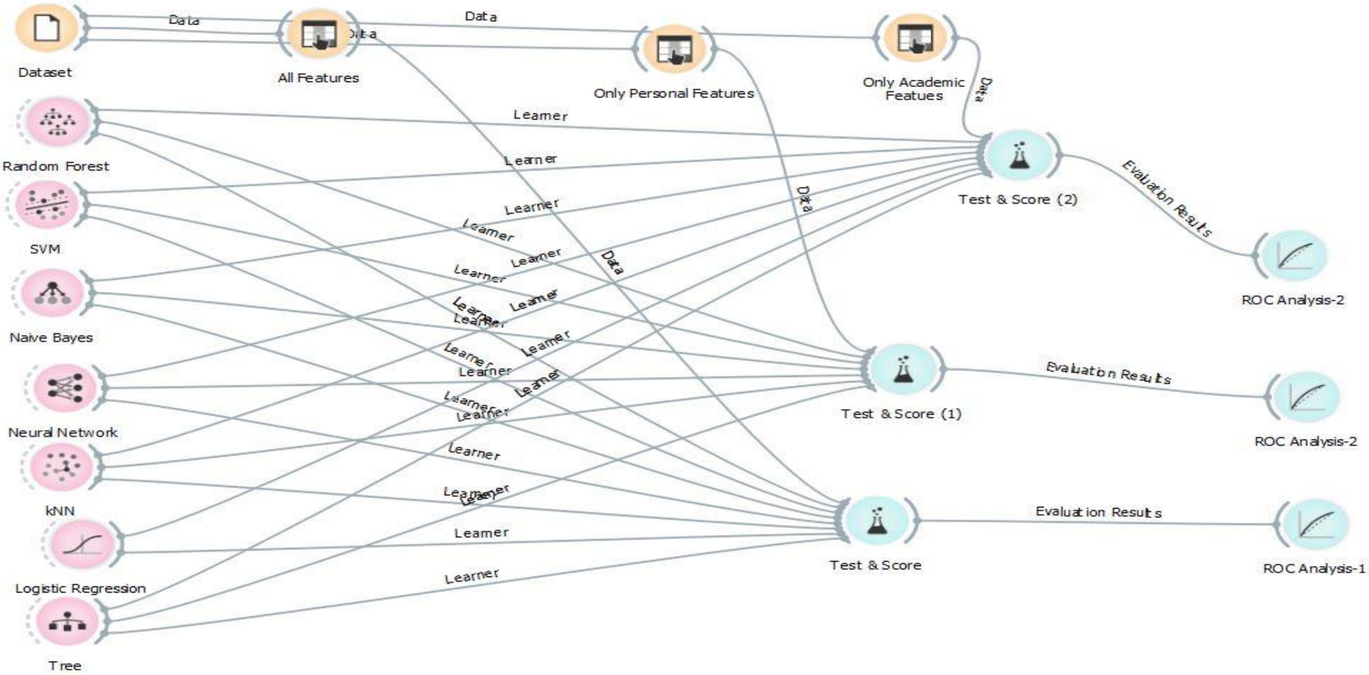


Fig. 4 Model design in Orange software

IV. Results and Discussion

Here evaluation of performance for the classification done by few error metrics. **Table 3** shows the context of confusion matrix-

TABLE III. CONFUSION MATRIX

		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

Here,

TP = True Positive = Predicted as positive and original member of positive class

FP = False Positive = Predicted as positive but original member of negative class

FN = False Negative = Predicted as negative but original member of positive class

TN = True Negative = Predicted as negative and originally a member of negative class

Several standard terms for evaluating by confusion matrix for two class-

Accuracy	ACC = (TP + TN) / (P + N)
Sensitivity or Recall	TPR = TP / (TP + FN)
Precision	PPV = TP / (TP + FP)
F1 Score	F1 = 2TP / (2TP + FP + FN)

.

Fig 5 shows a measurement system of classifier algorithms.

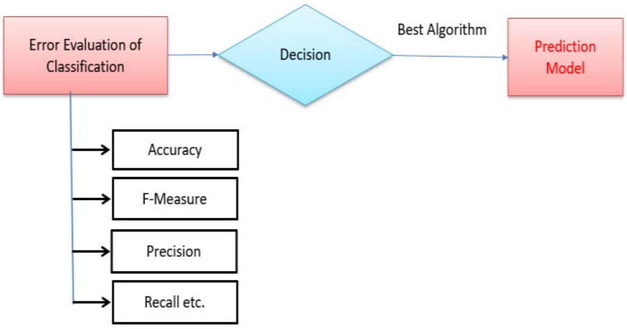
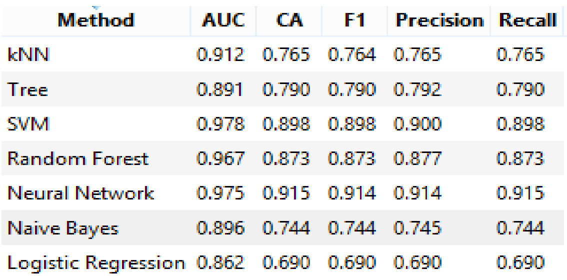


Fig 5. Measurement of Classification algorithms

Figure 6 shows the result of the various classifier for the full dataset



Method	AUC	CA	F1	Precision	Recall
kNN	0.912	0.765	0.764	0.765	0.765
Tree	0.891	0.790	0.790	0.792	0.790
SVM	0.978	0.898	0.898	0.900	0.898
Random Forest	0.967	0.873	0.873	0.877	0.873
Neural Network	0.975	0.915	0.914	0.914	0.915
Naive Bayes	0.896	0.744	0.744	0.745	0.744
Logistic Regression	0.862	0.690	0.690	0.690	0.690

Fig. 6 Result of the various classifier for the full dataset

For better comparison, we can see the ROC curve in figure 7.

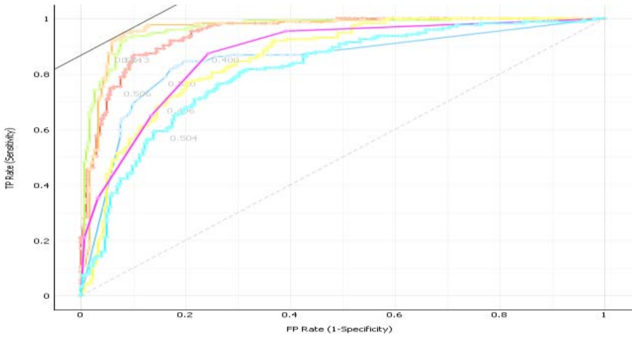
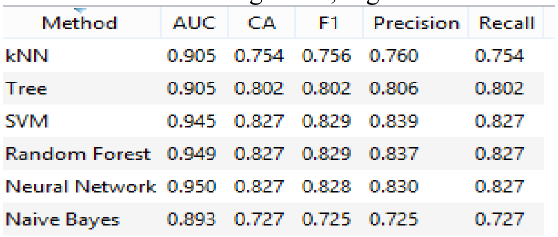


Fig. 7 ROC curve analysis

From this, we can see that the neural network is giving the best results in every aspect.

Now we will see the classification result and ROC curve analysis only by "Personal Data" in figure 8. Figure 9 and For "Academic Data" in figure 10, Figure 11



Method	AUC	CA	F1	Precision	Recall
kNN	0.905	0.754	0.756	0.760	0.754
Tree	0.905	0.802	0.802	0.806	0.802
SVM	0.945	0.827	0.829	0.839	0.827
Random Forest	0.949	0.827	0.829	0.837	0.827
Neural Network	0.950	0.827	0.828	0.830	0.827
Naive Bayes	0.893	0.727	0.725	0.725	0.727

Fig. 8 classification results for Personal Data

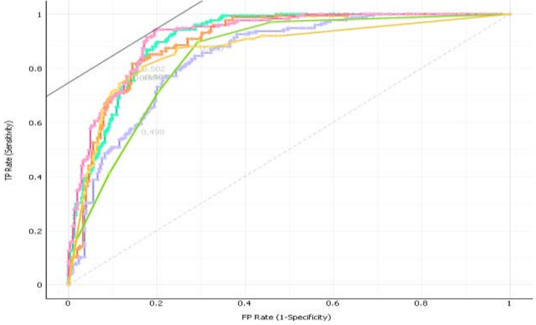
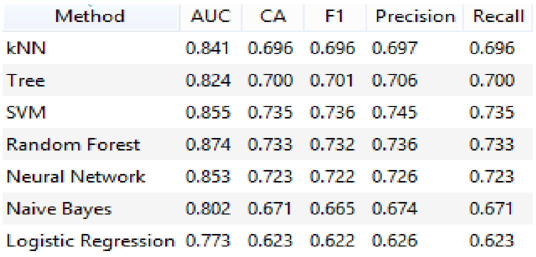


Fig. 9 ROC curve analysis for only Personal Data



Method	AUC	CA	F1	Precision	Recall
kNN	0.841	0.696	0.696	0.697	0.696
Tree	0.824	0.700	0.701	0.706	0.700
SVM	0.855	0.735	0.736	0.745	0.735
Random Forest	0.874	0.733	0.732	0.736	0.733
Neural Network	0.853	0.723	0.722	0.726	0.723
Naive Bayes	0.802	0.671	0.665	0.674	0.671
Logistic Regression	0.773	0.623	0.622	0.626	0.623

Fig. 10 Classification results for Academic data

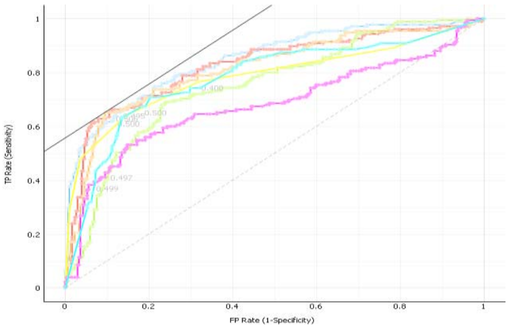


Fig. 11. ROC curve for classifiers with Academic Data

So, we can see if we take only personal or Academic, it will be less effective. We need to take personal, academic, and institutional data together for better performance. From the results, we can see the "Neural network" for all data performs the best. Accuracy of which is 0.915. With this, we made our model to predict prospective dropout student. Also, we have shown the crucial factors in feature selection by gain ratio.

Similar kind of work in India got 72% accuracy[5], in Japan [6] they have got 89% accuracy.

V. Conclusions & Future Work

Dropout prediction is a hot topic because of the high dropout rate over the world. However, most of the case dropout factors are different in different socio-economic and geological condition. So we have given a model for Bangladeshi students to predict dropout where we used a different kind of data including academic, personal, and institutional. As many researchers used only academic or only personal data, we have shown that using both kinds of data gives good accuracy. Unlike others, we have included late graduated students in a separate profile, namely medium. For our model, neural networks work best. We have also shown some essential factors which we have found by feature selection using gain ratio. Living criteria, student politics, Distance of University, Special encouragement program for good results, Math Score in HSC and SSC, attendance, etc. are an essential factor for dropout from engineering subjects in Bangladeshi universities. Students should be aware of these things so that they can be safe from the dropout. Universities should take necessary steps to prevent dropout following these critical factors.

Limitation of this works is, it can only be used for Bangladeshi students whoever completed their S.S.C., H.S.C. from a science background, and doing engineering. There are many kinds of educational medium, like Polytechnique, Madrasa, etc.

The future researcher can work on classifying dropped out students by their problem type so that different kind of students can be treated differently for the dropout prevention program.

References

[1] A. M. Graffigna, L. Hidalgo, A. Jofré, M. del C. Berenguer, A.Moyano, and I. Esteybar, "Tutorial Practice as a Strategy of retention at the School of Engineering," Procedia - Soc. Behav. Sci.,vol. 116, no. 2007, pp. 2489–2493, 2014.

[2] P. Kaur, M. Singh, and G. S. Josan, "Classification and PredictionBased Data Mining Algorithms to Predict Slow Learners inEducation Sector," in Procedia Computer science, vol. 57, pp.500–508, 2015.

[3] D. Ktoridou, "Measuring the Compatibility between Engineering Students ' Personality Types and Major of Study : A first step towards preventing Engineering Education Dropouts," no. April, pp. 192–195, 2014.

[4] J.L. Holland, "Exploring careers with a typology: What we have learned and some new directions", American psychologist,vol. 51(4), pp-397-406, 1996.

[5] Hegde, Vinayak, and P. P. Prageeth. "Higher education student dropout prediction and analysis through educational data mining." In 2018 2nd International Conference on Inventive Systems and Control (ICISC), pp. 694-699. IEEE, 2018.

[6] Sukhbaatar, Otgontsetseg, Kohichi Ogata, and Tsuyoshi Usagawa. "Mining Educational Data to Predict Academic Dropouts: a Case Study in Blended Learning Course." In TENCON 2018-2018 IEEE Region 10 Conference, pp. 2205-2208. IEEE, 2018.

[7] Halawa, Sherif, Daniel Greene, and John Mitchell. "Dropout prediction in MOOCs using learner activity features." Proceedings of the Second European MOOC Stakeholder Summit 37, no. 1,pp. 58-65, 2014.

[8] Evando B. Costa, Baldoino Fonseca, Marcelo Almeida Santana, Fabrisia Ferreira de Araujo, Joilson Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses", Computer in Human Behavior, vol. 73,, pp. 247-256, 2017.

[9] Perez, Boris, Camilo Castellanos, and Dario Correal. "Applying Data Mining Techniques to Predict Student Dropout: A Case Study." In 2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI), pp. 1-6. IEEE, 2018.

[10] Nagy, Marcell, and Roland Molontay. "Predicting Dropout in Higher Education Based on Secondary School Performance." In 2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES), pp. 000389-000394. IEEE, 2018.

[11] Dharmawan, Tio, Hari Ginardi, and Abdul Munif. "Dropout Detection Using Non-Academic Data." In 2018 4th International Conference on Science and Technology (ICST), vol. 1, pp. 1-4. IEEE, 2018.

[12] Vink, Gerko, Laurence E. Frank, Jeroen Pannekoek, and Stef Van Buuren. "Predictive mean matching imputation of semicontinuous variables." Statistica Neerlandica 68, no. 1, pp. 61-90, 2014..

[13] Karegowda, Asha Gowda, A. S. Manjunath, and M. A. Jayaram. "Comparative study of attribute selection using gain ratio and correlation based feature selection." International Journal of Information Technology and Knowledge Management 2, no. 2, pp. 271-277, 2010.