

# Masters and Doctor of Philosophy admission prediction of Bangladeshi students into different classes of universities

Md Naimul Islam Suvon<sup>1</sup>, Sadman Chowdhury Siam<sup>2</sup>, Mehebuba Ferdous<sup>2</sup>, Mahabub Alam<sup>2</sup>,  
Riasat Khan<sup>2</sup>

<sup>1</sup>Department of Computer Science, University of Sheffield, Sheffield, United Kingdom

<sup>2</sup>Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh

## Article Info

### Article history:

Received Aug 19, 2021

Revised Jun 9, 2022

Accepted Jul 8, 2022

### Keywords:

Data normalization

Decision tree

Educational data mining

Oversampling

Random forest

## ABSTRACT

Many Bangladeshi students intend to pursue higher studies abroad after completing their undergraduate degrees every year. Choosing a university for higher education is a challenging task for students. Especially, the students with average and lower academic credentials (undergraduate grades, English proficiency test scores, job, and research experiences) can hardly choose the universities that could match their profile. In this paper, we have analyzed some real unique data of Bangladeshi students who had been accepted admissions at different universities worldwide for higher studies. Finally, we have produced prediction models based on random forest (RF) and decision tree (DT) techniques, which can predict appropriate universities of specific classes for students according to their past academic performances. Two separate models have been studied in this paper, one for Masters (MS) students and another for Doctor of Philosophy (PhD) students. According to the Quacquarelli Symonds (QS) World University Rankings, the universities where the students got admitted have been divided into 9 classes for MS students and 8 classes for PhD students. Accuracy, precision, recall and F1-Score have been studied for the two machine learning algorithms. Numerical results show that both the algorithm DT and RF have the same accuracy of 89% for PhD student data and 86% for MS student data.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## Corresponding Author:

Md Naimul Islam Suvon

Department of Computer Science, University of Sheffield

Sheffield S10 2TN, United Kingdom

Email: mnisuvon1@sheffield.ac.uk

## 1. INTRODUCTION

Nowadays, educational data has become more popular among researchers. Educational information mining is the process of acquiring necessary information from an extensive collection of educational datasets and finally making significant decisions from them [1]. Many students of Bangladesh apply for higher studies every year in different universities all over the world. The students spend a significant amount of money and time preparing for the application process. Unfortunately, most of them face difficulties deciding which universities they should apply to according to their different test scores. Many students tend to choose safe options, where there are high possibilities to get admitted. Conversely, some of them apply to an ambitious higher-class university, which does not conform with their academic profile, leading to an ultimate rejection. Many students face this kind of problem as they cannot evaluate their academic credentials according to the admissions criteria. There are plenty of consultancy centers in Bangladesh, where they evaluate the students' profiles and provide guidance for the application process in exchange for high consultation fees. But

eventually, they fail to find the perfect universities for students to apply; because they cannot evaluate their profile correctly. Sometimes they are misled by the senior graduates about the university's ranking and past admission decision patterns.

In the following paragraphs, related papers that are similar to this research have been reviewed. Most of the papers employed random forest (RF) and decision tree (DT) techniques to implement the prediction model. For instance, Acharya *et al.* [2] used machine learning (ML)-based methods to compare different regression algorithms to predict the applicants' chance of graduate admissions. The authors implemented a DT that breaks the dataset sequentially into a smaller subset, and in the meantime, the associated DT was developing accordingly. Finally, they applied RF regression. It is an additive type model, which helps to predict by combining decisions from a sequence of base models. They used multiple models to get excellent predictive work, known as model assembling. The authors found that the linear regression achieved the highest accuracy on their dataset (hypothetical open-source data of UCLA), which had low mean squared error (MSE) and a high R2 score compared to the other implemented regression techniques. Hmiedi *et al.* [3] made a regression model using the RF algorithm to predict the graduate admissions probabilities. This work used the same hypothetical open-source dataset from Kaggle of the University of California in Los Angeles as in [2]. The authors applied data augmentation to achieve a more diverse dataset and reduce overfitting and data preprocessing (data normalization and duplicate removal). They split the data into 70% and 30% for training and testing, respectively, and finally reported the proposed model's accuracy. Janani *et al.* predicted the chance of graduate admissions using the DT algorithm in [4]. The authors attained 93% accuracy by using the same open-source dataset in [5] and the DT classifier in output.

Hien *et al.* [6], used the Bayesian network's technique to forecast the graduating student's cumulative grade point average (CGPA) based on the applicant's background (previously attended institutions, undergraduate CGPA, English test score, the field of study, age, gender, and marital status) during the admission process of the Asian Institute of Technology (AIT), Thailand. Finally, the study shows a mean absolute error of 0.22 and 0.20 grade points for the Master's (MS) and Doctoral of Philosophy (PhD) programs, respectively. A. Waters and R. Miikkulainen estimated the chance of admission of new applicants based on past admissions decisions at the Department of Computer Science of University of Texas at Austin, United States of America (USA) in [7]. They used a statistical ML technique (L1 regularized logistic regression) to evaluate this system from different numerical, categorical, and text features data. This system predicts a real-valued score for every student's file, similar to the traditional human reviewers. The proposed system graduate admissions evaluator (GRADE) attained an accuracy of 87.1% and reduced the total review time by 74%.

In this paper, we have worked to make the university selection procedure easier for the students according to their academic profiles. We work on a ML-based technique to predict the perfect university match for students pertaining to their past academic records, i.e., undergraduate university and CGPA, English proficiency test scores, job experience, and research papers. The students can evaluate their chance of getting admission to a higher rank or lower rank university. This paper's primary contribution is to work on an exclusive real dataset of Bangladeshi undergraduate students who have gone for higher studies abroad to USA, Canada, Germany, Australia, and different foreign countries for MS and PhD degrees for the last two years of 2018 and 2019 from more than 30 universities of Bangladesh. Next, we analyzed this data to find the essential features we need for our model. According to this year's Quacquarelli Symonds (QS) world university rankings [8], we have divided the accepted universities into four classes. Subsequently, we have developed three different approaches (each for the MS and PhD students' data) to make the model based on DT [5] and RF [9] approaches for assessing the possibility of a student's admission to a particular class of university. Finally, we have reported all the ML algorithms' performance for both the MS and PhD applicants' data in terms of the evaluation metrics, e.g., precision, recall, F1-score, accuracy, and confusion matrix [10]. To the best of our knowledge, this is the first time various multi-class classification models to select different universities worldwide have been done on the dataset of Bangladesh's students.

## 2. METHODOLOGY

The proposed approach for implementing the admission prediction is divided into different sections. Figure 1 represents the working sequences of the proposed model. In the subsequent paragraphs, the working methods of this paper have been described in detail.

### 2.1. Dataset

The primary contribution of this work is to create a unique dataset. We have collected our dataset from the Graduate Resources Enhancing Center (GREC), Bangladesh. GREC is one of the largest platforms for Bangladeshi students, where every year, many students take preparation for various standardized exams,

e.g., graduate record examination (GRE), international english language testing system (IELTS), test of english as a foreign language (TOEFL), standard assessment tests (SAT). We have collected the data for the last two years, 2018 and 2019, for students who have been accepted in various universities in the USA, Canada, Australia, Germany, and United Kingdom (UK). Initially, there were 230 students' data from more than 30 universities of Bangladesh, which contains master and doctoral applicants. Also, a single candidate got a chance in multiple universities simultaneously. Next, we separated the MS and PhD data and made two datasets to apply different prediction models. Finally, we obtained approximately 400 data for PhD students and 300 for the MS candidates. As this total of 700 data has been obtained from 230 candidates, an individual student got accepted in an average of 3 universities. Next, we have added a new feature to our dataset, the universities' QS world university rankings, where the students have been admitted. QS world university rankings, partnered with Elsevier, is the most accepted international rankings of universities worldwide. According to the university rankings, we have divided the candidates into nine classes for MS and eight classes for PhD students, where they have been admitted. Class A is the candidates who have been accepted in a university with a QS world university rankings between 1 to 50. Similarly, classes B, C, D, E, F, G, H, I are for university rankings between 50 to 100, 101 to 150, 151 to 200, 201 to 300, 301 to 400, 401 to 600, 601 to 750 and above 750, respectively.

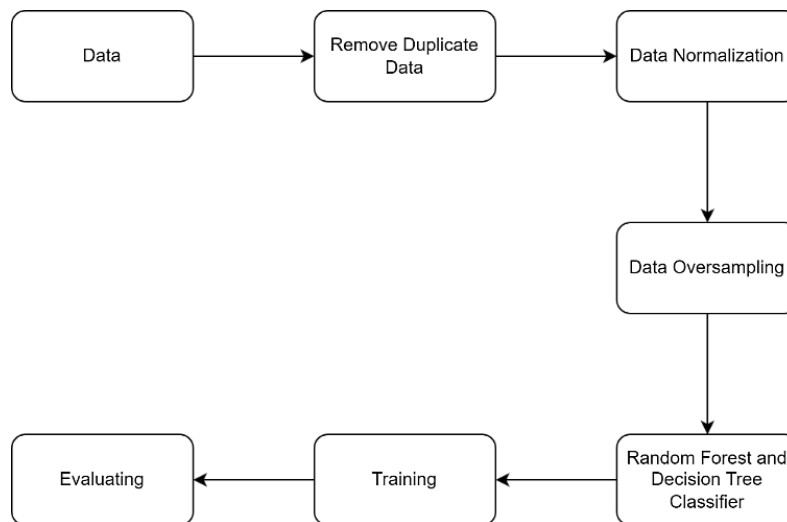


Figure 1. Working sequences of the proposed system

Figures 2(a) and 2(b) show the number of MS and PhD students in each university class. The initial dataset contains many features in three forms, numerical, categorical, and text features. Those are student's name, admitted university name with its state and country, admitted department, intended research area, types of funding (fellowship, assistantship, external scholarship), intended semester, undergraduate university name with CGPA and department, IELTS/TOEFL score, GRE score, publications (conference or journal), job experience, research experience, application method, and funding source.

## 2.2. Data analysis

It is crucial to analyze data before building a model because there might be many missing, inconsistent, and duplicate data [11]. We have performed some analysis on both the MS and PhD datasets so that the applied algorithm can quickly analyze them. Most of the students' data are structured in a uniform format so that the algorithm can easily interpret it, except the undergraduate and admitted universities' names. Some students may describe the same undergraduate institution as BUET, Bangladesh University of Engineering and Technology, Bangladesh U of Engineering and Technology, etc. The previously attended undergraduate universities' names are rephrased in uniform abbreviated string formats at the first data preprocessing step, e.g., BUET, DU, RUET, SUST, etc. The admitted graduate universities' names are not required to be preprocessed as they are divided into different classes and consequently have been removed from the feature vectors. Figures 3(a) and 3(b) demonstrate the correlation matrix between various features for PhD and MS students, respectively

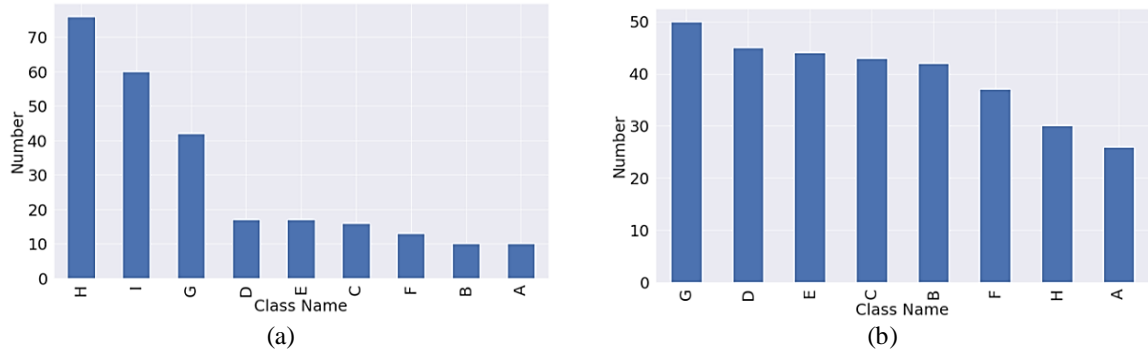


Figure 2. Number of students in each class of university, (a) MS and (b) PhD

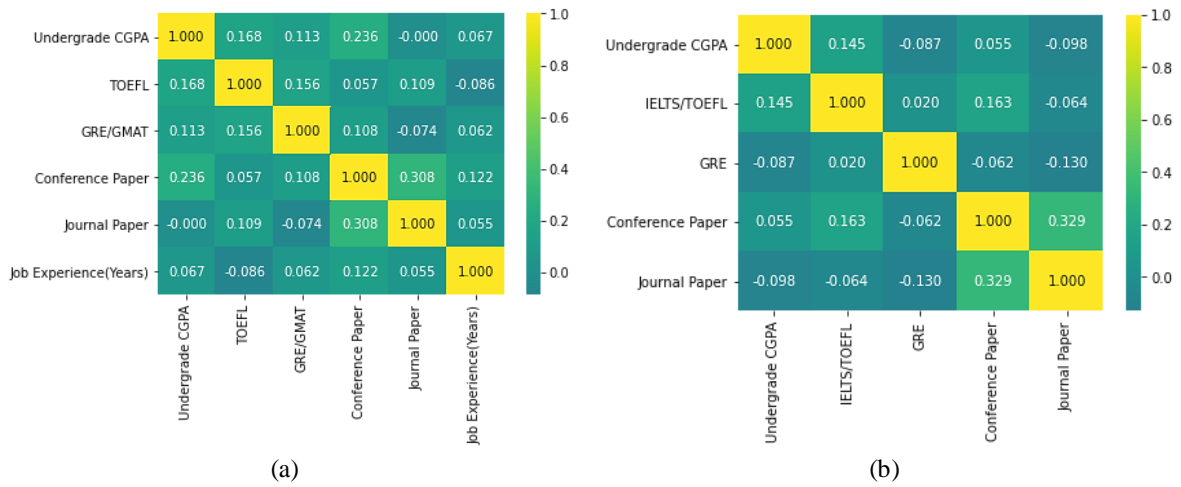


Figure 3. Correlation matrix between features for (a) PhD and (b) MS students

Finally, we have found out the important features [12] which mostly influenced the model. Feature importance is calculated as [13]:

$$RFfi_i = \frac{\sum_{j \in \text{all trees}} normfi_{ij}}{T} \quad (1)$$

In (1),  $normfi_{ij}$  represents the normalized feature importance for  $i$  in tree  $j$  and  $T$  denotes the total number of trees.

$$normfi_{ij} = \frac{fi_i}{\sum_{j \in \text{all features}} fi_j} \quad (2)$$

In (2), each feature's importance on a DT is normalized to a value between 0 and 1 by dividing by the sum of all the feature importance values. Here  $fi_i$  is the importance of feature  $i$ , which can be calculated by:

$$fi_i = \frac{\sum_{j: \text{node } j \text{ splits on feature } i} ni_j}{\sum_{k \in \text{all nodes}} ni_k} \quad (3)$$

According to Figure 4(a), for MS data, the essential feature is undergraduate CGPA, which feature importance score is almost 0.24, and the second and third one is GRE and IELTS/TOEFL, with a score of 0.23 for both. Similarly, in Figure 4(b), for Ph.D. data, Bachelor's CGPA, GRE and IELTS/TOEFL are the most significant features for our model. Consequently, it means that these three features will perform a crucial role in our prediction.

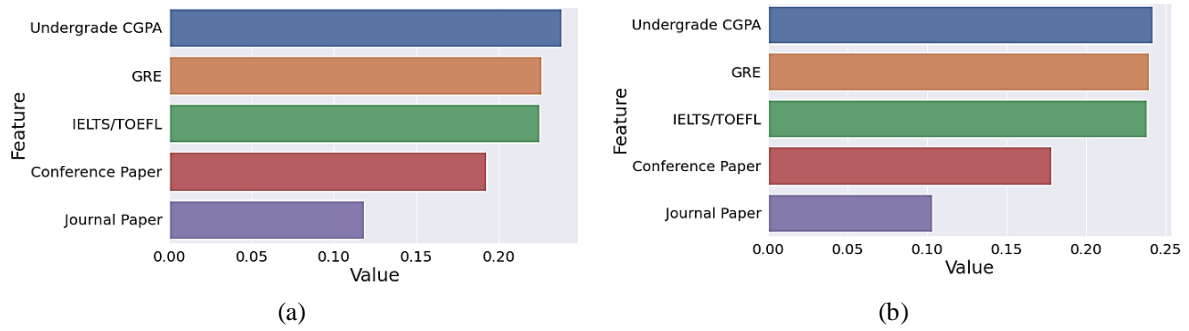


Figure 4. Feature importance scores of various features for (a) MS and (b) PhD students' data

### 2.3. Data preprocessing

In this step, we have gone through a few steps to perform data cleaning and data transformation. Firstly, we have removed missing data from the table. Next, we normalized the test scores and merged them into a single feature using (4) [14]. We also one-hot encoded [15] the categorical data by creating binary or dummy variables as input in the ML model. We then applied a transformation to numerical data by performing normalization on them. We have used MinMaxScaler transformation [16], all of our data transformed between 0 and 1.

$$Z = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (4)$$

We have classified the universities based on the ranking into various classes for the proposed prediction model as explained in section 3.1. According to Figure 2, our dataset is biased towards the H class for MS and the G class for PhD data. If the dataset is not balanced, the proposed admission prediction model will get biased towards the majority class and perform poorly towards minority classes, so we need to balance our class data. For balancing, we have used oversampling method using synthetic minority oversampling technique (SMOTE). In SMOTE, the minority class is oversampled and generates new data to increase it to the same number as the majority class [17]. In this work, for MS, the majority class is H which has 76 samples, and for PhD the majority class is G, with 50 samples. After oversampling, all the classes of MS and PhD consist of 86 and 50 samples, respectively. So now, our dataset is equally distributed towards all predicted classes, which will help us get an unbiased prediction. After oversampling applying SMOTE technique, the size of the dataset has increased. For PhD we have now  $(50 \times 8) = 400$  samples and for MS  $(76 \times 9) = 684$  samples in total.

### 2.4. Modeling

A ML model is a representation of an algorithm that sifts through massive amounts of data in search of patterns or predictions. In this paper, two ML algorithms, DT and RF, have been used to predict the admission possibilities of MS and PhD applicants. In the following subsection, these algorithms are explained briefly.

#### 2.4.1. Decision tree technique

DT are statistical models that generate trees based on each feature's information gain and return the predicted output by making decisions based on the tree nodes [5]. We chose the Iterative Dichotomiser-3 (ID3) algorithm [18], which is mostly used to generate DT. It uses a greedy top-down approach [19] to select the tree nodes based on the information gain of a particular feature, and (6) is used to measure the information gain. We calculated the gain by the difference in the parent node's entropy and its corresponding child nodes' average entropy.

$$IG = E(P) - WA \times E(C) \quad (5)$$

Here  $IG$  represents information gain,  $WA$  denotes weight average,  $E(P)$  and  $E(C)$  represent parent and child entropy, respectively. Next, entropy is calculated, which is a measure of disorder obtained by selecting a particular feature.

$$E(S) = \sum_{i=1}^c -p_i \log_2 p_i \quad (6)$$

The ID3 algorithm constructs the trees by choosing a variable with maximum gain and splits the data based on the attributes (if the attribute is available or not in the data) [20]. It keeps constructing child nodes by selecting different variables until all the features are explored and conclusive. As our dataset has distinct features like CGPA, standardized test scores, number of publications, etc., the DT model is a perfect choice as it can make step by step decisions from each of the features.

### 2.4.2. Random forest approach

RF model is an approach based on the ensemble method. It is one of the popular approaches which offers an optimized predictive algorithm by combining several models [9]. Several weak learners are combined to build a more robust model in the ensemble method [21], [22]. We chose DT as weak models for our proposed model and then used these models to construct a powerful learner [23]. Each weaker model creates a new dataset by considering a random subset of data from every tree's original dataset. Then the tree is trained on the unique subset of data. In this way, we trained a large number of individual DT models. The subset data can be chosen with no replacement, i.e., pasting or relief, also known as bagging or bootstrap aggregating. The RF model then follows the prediction that matches the most trees' output like a voting system. We have used RF in our model as DT tend to overfit a lot. It is worth mentioning that, we can eliminate the disadvantage of overfitting with RF by averaging the result.

## 3. RESULTS AND DISCUSSION

This work's primary objective is to make a reliable prediction of a student's admission into a specific university class. We analyzed our unique dataset of 400 PhD students and 300 MS candidates using two ML methods. At first, we have split the dataset into training and test subset by the ratio of 7:3. To show each method's performance, we have measured different evaluation matrices, i.e., precision, recall, F1-score, accuracy, and confusion matrices. Precision indicates the percentage of relevant cases among the retrieved ones ( $TP$  divided by  $TP + FP$ ), and recall specifies the percentage of relevant data that have been retrieved ( $TP$  divided by  $TP + FN$ ).  $TP$  and  $FP$  represent true positive and false positive, respectively, and  $FN$  means false negative [24].

### 3.1. Performance of decision tree technique

According to Table 1, for the DT MS model, F1-score and recall value is 86% for both and 87% for precision. The accuracy for the MS model is not outstanding because of the university classes of lower-ranking, i.e., G, H and I. On the other hand, the accuracy of the DT PhD model is not impressive because of the higher-ranking university predicted classes A, B, C and D. Precision and recall for the PhD model is 89% and F1-Score is 88%. Finally, the overall accuracy of the MS model is 86% which is lower than the PhD model.

Table 1. F1-Score, precision, recall, and accuracy of different classes for decision tree

Class	MS			PhD		
	F1-Score	Precision	Recall	F1-Score	Precision	Recall
<b>A</b>	96%	100%	93%	85%	79%	91%
<b>B</b>	94%	95%	93%	82%	78%	86%
<b>C</b>	94%	96%	91%	78%	75%	81%
<b>D</b>	93%	98%	89%	80%	82%	77%
<b>E</b>	86%	83%	90%	85%	91%	80%
<b>F</b>	94%	95%	93%	87%	96%	79%
<b>G</b>	79%	70%	90%	90%	87%	94%
<b>H</b>	67%	68%	60%	99%	100%	98%
<b>I</b>	71%	74%	68%	X	X	X
<b>Overall</b>	86%	87%	86%	88%	89%	89%
<b>Overall Accuracy</b>		86%			89%	

Figures 5(a) and 5(b) depict the confusion matrix of the proposed DT MS and PhD models. We got the lowest accuracy for MS class H with 61% accuracy. Class B gives the highest accuracy, which is 97% for MS dataset. For PhD, class B offers the lowest accuracy of 82% and highest for Class H with 100% accuracy.

### 3.2. Performance of random forest approach

Evaluation metrics of the implemented RF model are depicted in Table 2. For the MS model, F1-score and precision values are 86% and 85% for recall. The accuracy of this model is moderate because of

the lower-ranking university predicted classes of G, H and I. On the other hand, for the PhD model, the accuracy is modest because of the higher-ranking university predicted classes, i.e., C and D. F1-score and recall for PhD model is 89% and precision is 88%. Although the overall accuracy for both DT and RF algorithms is the same for MS and PhD, the individual precision, recall and F1-score of different classes varies significantly. Interestingly, the confusion matrix for the RF technique has not been reported in this paper as it follows the same trend of the DT model.

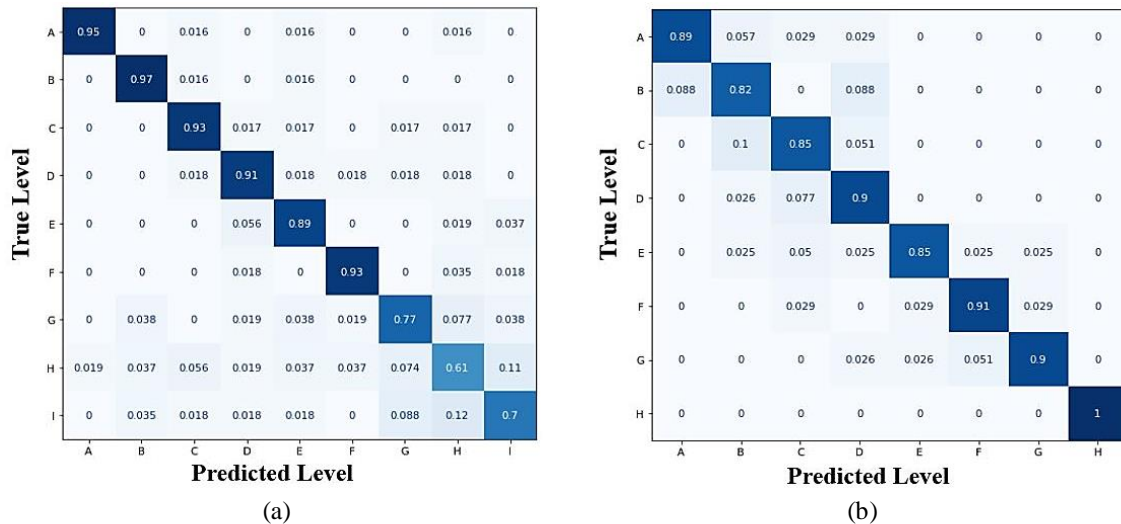


Figure 5. Confusion matrix of decision tree for (a) MS model and (b) PhD model

Table 2. F1-Score, precision, recall, and accuracy for random forest

Class	MS			PhD		
	F1-Score	Precision	Recall	F1-Score	Precision	Recall
<b>A</b>	95%	95%	96%	87%	93%	82%
<b>B</b>	89%	89%	98%	86%	85%	88%
<b>C</b>	87%	87%	96%	79%	77%	82%
<b>D</b>	89%	89%	94%	84%	79%	85%
<b>E</b>	88%	88%	85%	89%	92%	86%
<b>F</b>	92%	92%	95%	90%	92%	87%
<b>G</b>	73%	73%	88%	91%	93%	91%
<b>H</b>	80%	80%	56%	99%	98%	100%
<b>I</b>	77%	77%	65%	X	X	X
<b>Overall</b>	86%	86%	85%	89%	88%	89%
<b>Overall Accuracy</b>		86%			89%	

### 3.3. Prediction with the most importance features

We can observe that for MS data, the highest accuracy of 86% has been achieved by the RF method. For PhD dataset, a similar accuracy has been attained by RF and DT algorithms. We also tried to implement the k-fold cross-validation technique [25], but unfortunately, the classification performance improved insignificantly. The desired result has not been obtained because of too much noisy data. We have analyzed the most important eight features, i.e., undergraduate CGPA and university, English test scores, research papers, and job experiences, described in the data analysis section. We are now finding out how many classifications are true and false according to the most significant features, undergraduate CGPA and GRE. First, we have found out the average undergraduate CGPA and GRE scores. We find the average CGPA is 3.53, and the mean average GRE score is 300.1. Finally, we set four conditions to the undergraduate CGPA and GRE score intuitively to find out the true and false class labeling. The conditions are:

- 1) If  $CGPA > \text{Average CGPA}$  and  $GRE > \text{Average GRE}$ , it is in class A and B
- 2) Else if  $CGPA > \text{Average CGPA}$  and  $GRE < \text{Average GRE}$ , it is in class C and D
- 3) Else if  $CGPA < \text{Average CGPA}$  and  $GRE > \text{Average GRE}$ , it is in class E and F
- 4) Else if  $CGPA < \text{Average CGPA}$  and  $GRE < \text{Average GRE}$ , it is in class G, H, and I (for MS data only)

Undergraduate CGPA is a higher important feature than the GRE score, as shown in Figures 4(a) and Figure 4(b). Out of 400 PhD data, we have found that only 140 data match these conditions, and for 300 MS data, it only matches 100 data, unfortunately. As the scope of admission and funding opportunities is

uncertain, many Bangladeshi students with excellent academic profiles apply to only individual universities, where the chances of admissions are almost definite for them. Our obtained dataset comprises many cases where the students have an outstanding academic background to get admission at class A or B universities, but actually, they only applied to lower-class universities. This section confirms this fact, and hence we can estimate the reasons behind not obtaining higher accuracy.

#### 4. CONCLUSION

The undergraduate students from developing countries like Bangladesh invest a lot of money, time, and energy while applying for graduate studies. In this work, mathematical models have been developed to predict universities' admissions possibilities from the students' perspective. An individual student with his past academic records will be informed about which range of universities he should apply, and his chance of admission. We have used two algorithms, DT and RF, to make a separate MS and PhD applicants model. We found that the RF and DT of both the classifier model for both the MS and PhD data performed the same in terms of F1-score and accuracy. This validates the reason behind other research papers used the RF algorithm to anticipate graduate admissions. The DT algorithm offers similar prediction accuracy to the RF for the PhD and MS data. Many students in our acquired dataset applied and eventually got accepted in lower-ranking universities, although they have an outstanding profile, which leads to some noises and outliers. In the future, more data of the candidates of MS and PhD degrees can be collected and adopted a more proper way to divide them into different classes. We can increase the system's reliability by adding text features to the data, e.g., statements of purpose, research proposals, and letters of recommendation.

#### REFERENCES





- [1] C. Romero and S. V. Ventura, *Data mining in E-learning*. WIT Press, 2006.
- [2] M. S. Acharya, A. Armaan, and A. S. Antony, "A comparison of regression models for prediction of graduate admissions," Feb. 2019, doi: 10.1109/iccids.2019.8862140.
- [3] I. Hmiedi, H. Najadat, Z. Halloush, and I. Jalabneh, "Semi supervised prediction model in educational data mining," Dec. 2019, doi: 10.1109/acit47987.2019.8991048.
- [4] J. P. H. P. V. and M. P. S., "Prediction of MS graduate admissions using decision tree algorithm," *International Journal of Science and Research (IJSR)*, vol. 9, no. 3, pp. 492–495, 2020.
- [5] H. Ahmed and A. Nandi, *Condition monitoring with vibration signals: Compressive sampling and learning algorithms for rotating machines*. Wiley, 2019.
- [6] N. T. N. Hiên and P. Haddawy, "A decision support system for evaluating international student applications," Nov. 2007, doi: 10.1109/fie.2007.4417958.
- [7] A. Waters and R. Miikkulainen, "Grade: Machine-learning support for graduate admissions," *AI Magazine*, vol. 35, no. 1, p. 64, Mar. 2014, doi: 10.1609/aimag.v35i1.2504.
- [8] "QS World University Rankings 2020," *Quacquarelli Symonds Limited*. 2020, Accessed: Dec. 26, 2021. [Online]. Available: <https://www.qs.com/portfolio-items/world-university-rankings-2020>.
- [9] N. Nurhachita and E. S. Negara, "A comparison between deep learning, naïve bayes and random forest for the application of data mining on the admission of new students," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 2, p. 324, Jun. 2021, doi: 10.11591/ijai.v10.i2.pp324-331.
- [10] M. N. I. Suvon, R. Khan, and M. Ferdous, "Real time bangla number plate recognition using computer vision and convolutional neural network," in *2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAET)*, Sep. 2020, pp. 1–6, doi: 10.1109/iicaet49801.2020.9257843.
- [11] J. Huang, Y.-F. Li, and M. Xie, "An empirical analysis of data preprocessing for machine learning-based software cost estimation," *Information and Software Technology*, vol. 67, pp. 108–127, Nov. 2015, doi: 10.1016/j.infsof.2015.07.004.
- [12] A. P. Cassidy and F. A. Deviney, "Calculating feature importance in data streams with concept drift using Online Random Forest," in *2014 IEEE International Conference on Big Data (Big Data)*, Oct. 2014, pp. 23–28, doi: 10.1109/bigdata.2014.7004352.
- [13] S. M. Shaharudin, N. Ahmad, and S. M. C. M. Nor, "A modified correlation in principal component analysis for torrential rainfall patterns identification," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 9, no. 4, pp. 655–661, Dec. 2020, doi: 10.11591/ijai.v9.i4.pp655-661.
- [14] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Applied Soft Computing*, vol. 97, p. 105524, Dec. 2020, doi: 10.1016/j.asoc.2019.105524.
- [15] R. Karthiga, G. Usha, N. Raju, and K. Narasimhan, "Transfer learning based breast cancer classification using one-hot encoding technique," in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, Mar. 2021, pp. 115–120, doi: 10.1109/icaiss50930.2021.9395930.
- [16] M. M. Ahsan, M. A. P. Mahmud, P. K. Saha, K. D. Gupta, and Z. Siddique, "Effect of data scaling methods on machine learning algorithms and model performance," *Technologies*, vol. 9, no. 3, p. 52, Jul. 2021, doi: 10.3390/technologies9030052.
- [17] Y. Yan, R. Liu, Z. Ding, X. Du, J. Chen, and Y. Zhang, "A parameter-free cleaning method for SMOTE in imbalanced classification," *IEEE Access*, vol. 7, pp. 23537–23548, 2019, doi: 10.1109/access.2019.2899467.
- [18] C. Jin, L. De-lin, and M. Fen-xiang, "An improved ID3 decision tree algorithm," in *2009 4th International Conference on Computer Science & Education*, Jul. 2009, pp. 127–130, doi: 10.1109/iccse.2009.5228509.
- [19] A. Ahmad and G. Brown, "Random projection random discretization ensembles - ensembles of linear multivariate decision trees," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 5, pp. 1225–1239, May 2014, doi: 10.1109/tkde.2013.134.
- [20] N. A. Mashudi, N. Ahmad, and N. M. Noor, "Classification of adult autistic spectrum disorder using machine learning approach," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 3, pp. 743–751, Sep. 2021, doi:







- 10.11591/ijai.v10.i3.pp743-751.
- [21] N. Quadrianto and Z. Ghahramani, "A very simple safe-Bayesian random forest," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1297–1303, Jun. 2015, doi: 10.1109/tpami.2014.2362751.
  - [22] Z. Wang, C. Cao, and Y. Zhu, "Entropy and confidence-based undersampling boosting random forests for imbalanced problems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 12, pp. 5178–5191, Dec. 2020, doi: 10.1109/tnnls.2020.2964585.
  - [23] L. Zhao, S. Lee, and S.-P. Jeong, "Decision tree application to classification problems with boosting algorithm," *Electronics*, vol. 10, no. 16, p. 1903, Aug. 2021, doi: 10.3390/electronics10161903.
  - [24] M. M. Alam, M. N. I. Suvon, and R. Khan, "Social distance measurement and face mask detection using deep learning models," in *International Conference on Intelligent Emerging Methods of Artificial Intelligence & Cloud Computing*, Springer International Publishing, 2022, pp. 540–549.
  - [25] K. Pal and B. V Patel, "Data classification with k-fold cross validation and holdout accuracy estimation methods with 5 different machine learning techniques," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, Mar. 2020, pp. 83–87, doi: 10.1109/iccmc48092.2020.iccmc-00016.

## BIOGRAPHIES OF AUTHORS







**Md Naimul Islam Suvon**     completed his bachelor's degree in Computer Science and Engineering from North South University, Dhaka, Bangladesh. He is currently doing his Master's in Computer Science with speech and language processing in University of Sheffield, UK. He is currently working on deep learning and NLP methods on several domain. He has recently published multiple IEEE conference and journal papers on different domain of deep learning. His current research interests include deep learning, machine learning, NLP speech processing, and computer vision. He can be contacted at email: naimul.suvon@northsouth.edu or mnisuvon1@sheffield.ac.uk.







**Sadman Chowdhury Siam**     obtained his B.Sc. degree in Computer Science and Engineering from North South University, Dhaka, Bangladesh. He is currently working on artificial intelligence and face recognition-related problems. His research interests involve deep learning, computer security, and computational intelligence. He can be contacted at email: sadman.siam@northsouth.edu.







**Mehebuba Ferdous**     recently completed her graduation in Computer Science and Engineering from North South University, Dhaka. Her research interest includes database and networking. She can be contacted at email: mehebuba.ferdous@northsouth.edu.



**Md Mahabub Alam**     is currently studying MSc. in Information and Communication Engineering at TU Darmstadt, Germany. He completed his bachelor's in Computer Science and Engineering at North South University, Dhaka. His research interests are in deep learning, networking, and security. He can be contacted at email: mahabub.alam01@northsouth.edu.



**Riasat Khan**     pursued his bachelor's degree in Electrical and Electronic Engineering from Islamic University of Technology, Bangladesh. He obtained his master's and Ph.D. degrees in Electrical Engineering from New Mexico State University, Las Cruces, USA. At present, he is with the Electrical and Computer Engineering Department of North South University, Dhaka, Bangladesh. His main research interests include cardiac electrophysiology, computational bioelectromagnetics, model order reduction, and machine learning. He can be contacted at email: riasat.khan@northsouth.edu.