Department of Electrical and Computer Engineering
North South University

Senior Project Design

Towards the Analysis and Detection of MS and PhD Admission of Bangladeshi Students into different Ranking University

Team Members:

Name:	ID:
1. Md.Fahad Arafin	1520319042
2. Md. Faysal Ahmed	1521094642
3. Porinita Haque	1711204042

Faculty Advisor:
Dr. Mahdy Rahman Chowdhury
Associate Professor
Department of Electrical and Computer Engineering
North South University

Summer 2021

Declaration

This is to declare that no part of this report or the project has been previously submitted elsewhere for the fulfillment of any other degree or program. Proper acknowledgement has been provided for any material that has been taken from previously published sources in the bibliography section of this report.

.....

Masudur Rahim
Department of Electrical and Computer Engineering
North South University, Bangladesh

.....

Md. Mohaimanul Masud Sunny
Department of Electrical and Computer Engineering
North South University, Bangladesh

.....

Fahim Al Ifran Rahim
Department of Electrical and Computer Engineering
North South University, Bangladesh

Approval

The Senior Design Project entitle “All object on-chip tractor beam using SPP” by Masudur Rahim, Md. Mohaimanul Masud Sunny and Fahim Al Ifran Rahim has been accepted as satisfactory and approved for partial fulfillment of the requirement of BS in EEE degree program.

……

Supervisor's Signature

Dr. Mahdy Rahman Chowdhury

Associate Professor

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

……

Dr. Mohammad Rezaul Bari

Associate Professor & Chair

Department of Electrical and Computer Engineering

North South University.

Dhaka, Bangladesh.

Acknowledgement

First of all, we would like to express our profound gratitude to our honorable course instructor, Dr. Mahdy Rahman Chowdhury, for his constant and meticulous supervision, valuable suggestions, his patience and encouragement to complete this research.

We would like to thank everybody who supported us and provided with guidance for the completion of this research.

Abstract

Towards the Analysis and Detection of MS and PhD Admission of Bangladeshi Students into different Ranking University

Many Bangladeshi students intend to pursue higher studies abroad after completing their undergraduate degrees every year. Choosing a university for higher education is an ambiguous task for students. Usually, they face various problems in selecting the perfect university for them according to their profile. 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, 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 MS students and another for PhD students. According to the QS World University Rankings, the universities where the students got admitted have been divided into nine classes for Masters (MS) students and eight classes for PhD students. Random Forest and Decision tree algorithms are used for making the multi-class classification models. F1-score, accuracy, weighted accuracy, and the receiver operating characteristic curves have been studied for the two machine learning algorithms. Numerical results show that for MS data using random forest and decision tree we got same accuracy which is 86%. Again for PhD data using random forest and decision tree we got same accuracy which is 89%.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Chapter 1
Overview

In this paper, we have tried to make the university selection procedure easier for the students according to their academic profile. We work on a machine learning-based approach to predict the perfect university match for students pertaining to their past academic records, i.e., undergraduate university and CGPA, English proficiency score, job experience, and research paper. 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 Canada, US, Uk and Australia for MS and PhD degrees for the last two years of 2018 and 2019 from more than 30 universities of Bangladesh.

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 failed to find the perfect match of universities for students to apply; because they could not evaluate their profile correctly. Sometimes they are misled by the senior graduates about the university's ranking and past admission decision patterns.

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 machine learning-based approach to predict the perfect university match for students pertaining to their past academic records, i.e., undergraduate university and CGPA, English proficiency test's score, job experience, and research paper. 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. The initial dataset contains a lot of features, including the undergraduate university, subject, and grades, admitted university, subject and research area, funding sources, GRE and English aptitude test scores, research papers, job or research experiences, etc. Next, we have analyzed this data to find the essential features that we need for our model. According to this year of 2021 QS World University Rankings, we have divided the accepted universities into four classes [2]. Subsequently, we have developed three different approaches (each for the MS and PhD students' data) to make the model for assessing the possibility of a student's admission to a particular class of university. We have implemented the decision tree algorithm [3] and two ensemble learning methods, random forest [3] and adaptive boosting [4], in this work. Finally, we have reported all the machine learning algorithms' performance for both the MS and PhD applicants' data in terms of the evaluation metrics, e.g., precision, recall, F1-score, accuracy, weighted accuracy, ROC curve, AUC, [5], etc. 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.

Project Description

In this work, we have tried to make the university selection procedure easier for the students according to their academic profile. We work on a machine learning-based approach to predict the perfect university match for students pertaining to their past academic records, i.e., undergraduate university and CGPA, English proficiency score, job experience, and research paper. 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 Canada, US and UK for MS and PhD degrees for the last two years of 2018 and 2019 from more than 30 universities of Bangladesh.

Summary

In this chapter, we discussed the importance and the goals of our project. Students of our country suffer a lot during the time of application. They could not decide which university is suitable according to their study profile. Our goal is to analysis all of the data very well and make a classification model.

Chapter 2

Related Work

Introduction

In this section, related papers that are similar to our work have been discussed. The authors used different machine learning approaches for classifying the chances of a particular university based on students' academic profile.

Similar Existing Works

In this section, related papers that are similar to our work have been discussed. In the paper [6], N. T. N. Hien and P. Haddawy used the Bayesian network's approach to predict the graduating student's cumulative grade point average based on the applicant's background (previously attended institutions, undergraduate CGPA, English test score, the field of study, age, gender, marital status, etc.) at the time of admission. They evaluated the stratified ten-fold cross-validation technique of three years' admissions data of the Asian Institute of Technology (AIT), Thailand. The study shows a mean absolute error of 0.22-grade points for a master's program and 0.20 for the doctoral program. The study used the Bayesian network's approach; it gives departmental faculty members valuable information in making admission decisions. The Bayesian network prediction model represented a case-based retrieval mechanism that the same similarity measure used by the case-based system. The case-based system shows the past student most similar to the evaluating students.

In this paper [7], 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, USA. They used a statistical machine learning technique (L_1 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 GRADE (graduate admissions evaluator) attained an accuracy of 87.1% and reduced the total review time by 74%.

In this paper [8], M. S. Acharya, A. Armaan, and A. S. Antony used Machine Learning based methods, where they compared different regression algorithms to predict the applicants' chance of graduate admissions. They used linear regression to predict results and support vector regression to use kernel tricks to predict data. Next, they implemented a decision tree that breaks the dataset sequentially into a smaller subset, and in the meantime, the associated decision tree was developing accordingly. Finally, they applied random forest regression. It is an addictive type model, and this model helps to predict by combining decisions from a sequence of base models. Each base model is a decision tree, and the random forest model's result is the decision trees' cumulative output. They used multiple models to get excellent predictive work, which is 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 MSE and a high R2 score compared to the other implemented regression techniques.

In this paper [9], I. Hmiedi *et al.* made a regression model using the Random Forest 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 [8]. 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 75% for training and 25% for testing and finally reported the proposed model's accuracy.

P. Janani *et al.* predicted the chance of graduate admissions using the decision tree Algorithm in [10]. They used a classification algorithm with the decision tree classifier to predict the output due to its simple logic, effectiveness, and interpretability. This model works by creating a tree-

like structure by dividing the dataset into several smaller subsets based on different conditional logic. The authors attained 93% accuracy by using the same open-source dataset in [8] and the decision tree classifier in output.

Summary

In this chapter, we discussed about the similar work that are being done before by researchers. There are many works on this filed but there are not many who have tried multi-classification. So, in our project we have tried something different by showing multi-classification detection by giving level.

Chapter 3

Methodology

Introduction

Our approach for making the model is divided into different sections. Fig. 1 represents a flowchart of our proposed model. These are two classification algorithms that we will use in our model which are Random Forest and Decision Tree. In the subsequent sections, the working methods of this paper have been described in detail.

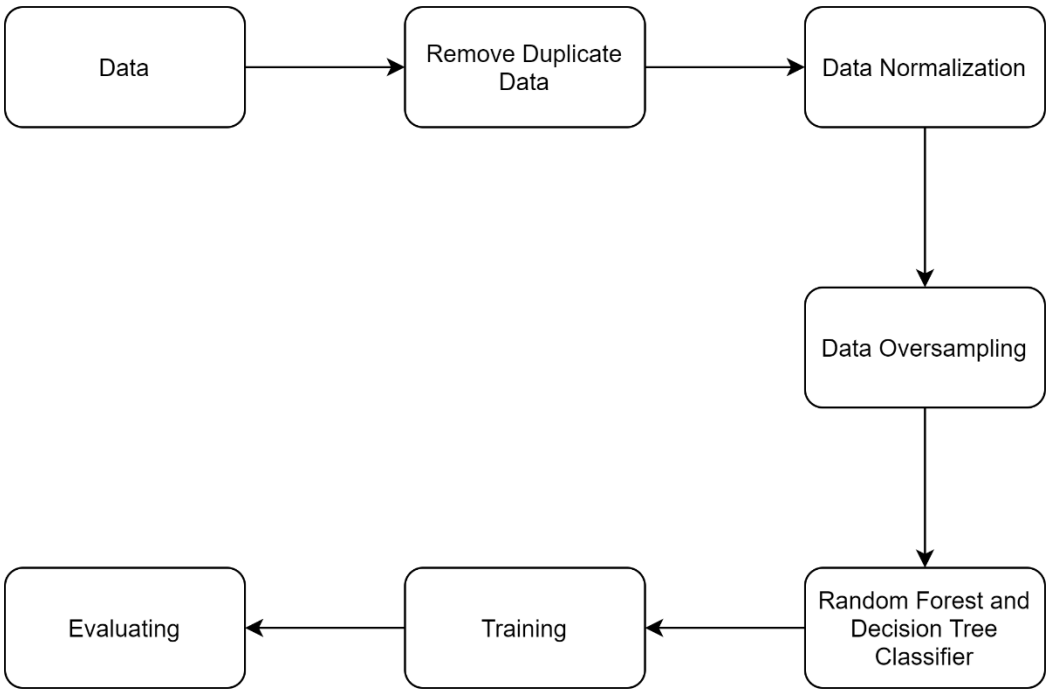


Fig. 1. Proposed System Model.

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., GRE, IELTS, TOEFL, SAT. They collect information from the students who are admitted to different universities worldwide with scholarships. We have collected the data for the last two years, 2018 and 2019, for students who have been accepted in various universities in mostly the USA, Canada, Australia, Germany, and the 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. As expected, the academic credentials are scattered quite differently between the master and doctoral students, with doctoral students were tending to achieve better academic performance than the master's students. Next, we separated the MS and PhD data and made two datasets to apply different prediction models for them. Then we have expanded our dataset by processing it into the multiple universities accepted candidate's data. 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(as we know that one candidate can get chances in many universities so in total for 230 candidates there are total 700 data split according to accepted university). An individual student got accepted in average of 3 universities. Next, we have added a new feature to our dataset, the universities' QS World University Ranking of 2021. Where the students have been admitted. our main target is analysis top universities so our prediction can give us the accurate result As we know that ranking varies and change every year but the top universities do not varies so much. 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, 10q to 150, 151 to 200, 201 to 300, 301 to 400, 401 to 600, 601 to 750 and above 750, respectively. Our work's final objective is to find in which class of university a candidate should apply with his academic profile. As our data is custom data so it can gives us truly prediction. In future if we work more and more for this custom dataset it can gives more accuracy.

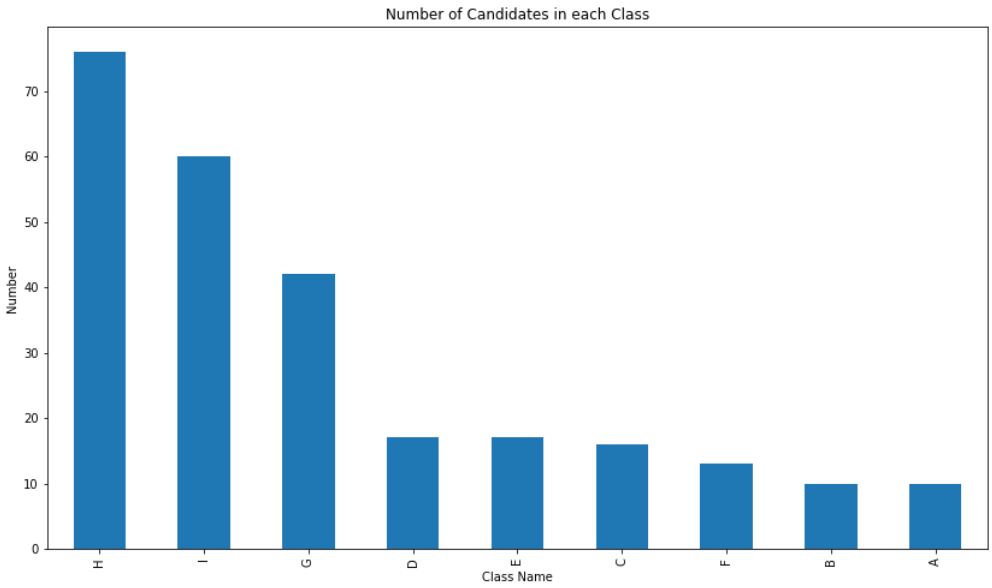


Fig. 2. Number of MS students in each class of university.

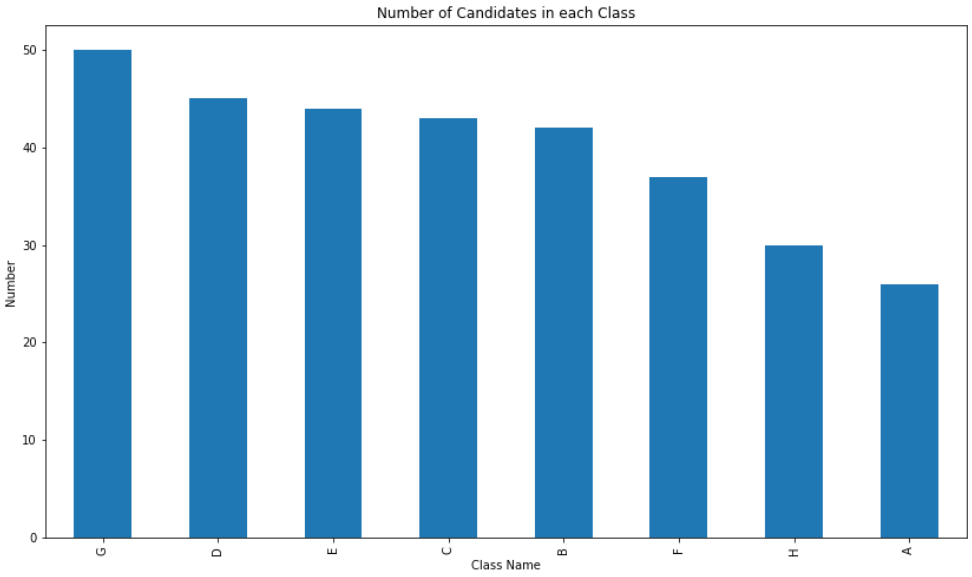


Fig. 3. Number of PhD students in each class of university.

Fig. 2 and Fig. 3 show the number of MS and PhD students in each university class. Our 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/TOFEL score, GRE score, publications (conference or journal), job experience, research experience, application method, funding source, etc.

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 such as finding prediction, regression plot, co relation matrix, one-hot encoding etc. 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.

In Fig. 4 and Fig. 5, we have analyzed the Bangladeshi institutions from where most students got accepted for higher studies for the last two years, 2018 and 2019. For the MS program, the University of Dhaka (DU), and Bangladesh University of Engineering and Technology (BUET), have the highest number of students accepted into different universities worldwide. On the other hand, BUET is at the top, and DU is the second position for the PhD program. Not surprisingly, both of these universities' students are mostly dominating in both programs. They are the best undergraduate institutions in Bangladesh regarding academic and employer reputation, acceptance rate, tuition fees, faculty-to-student ratio, etc.

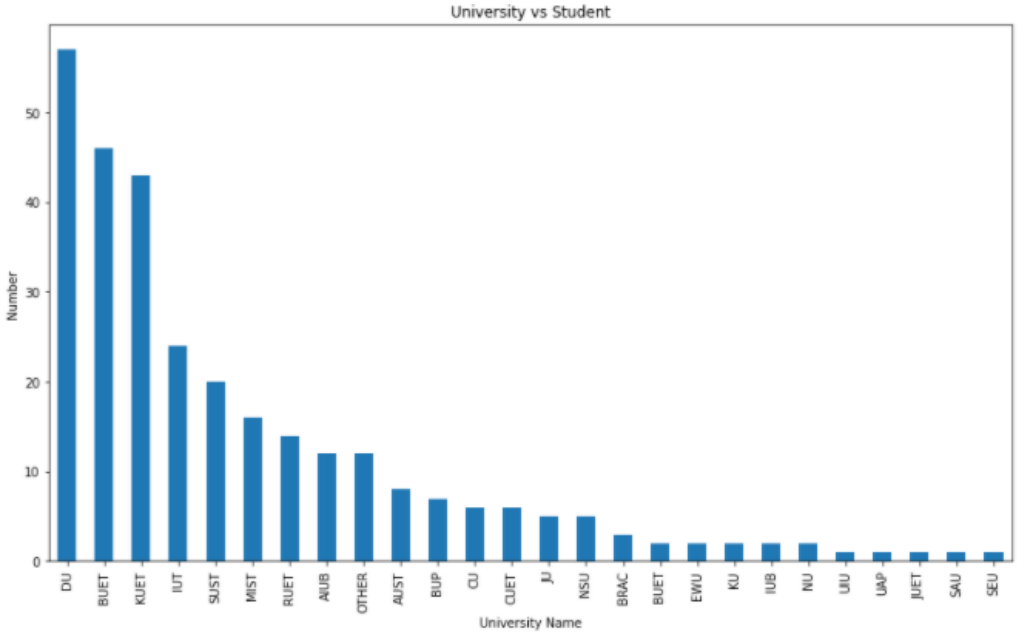


Fig. 4. The number of students who got accepted for MS according to the undergraduate university.

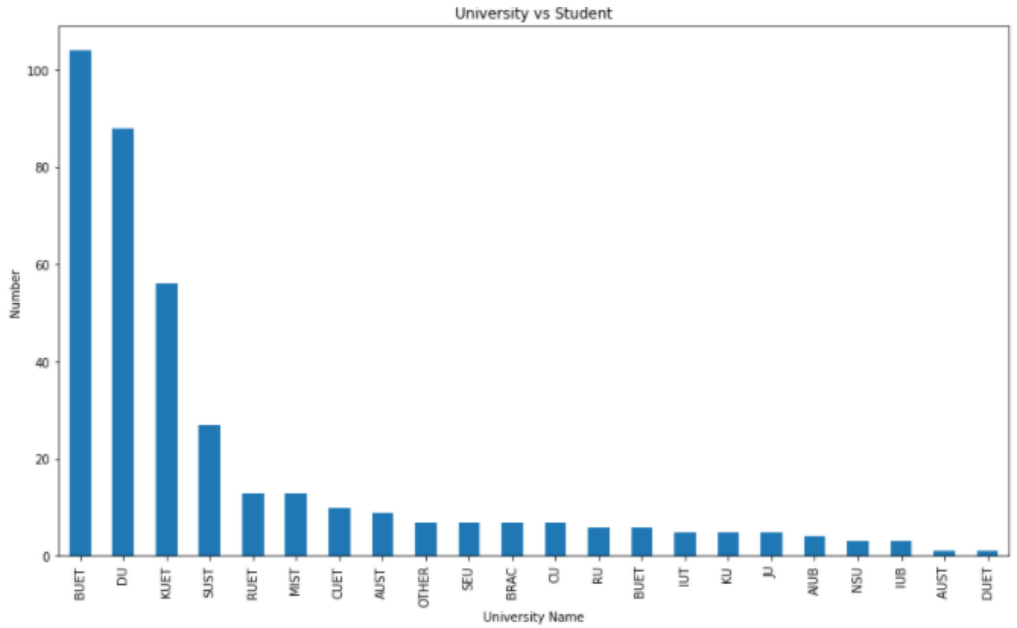
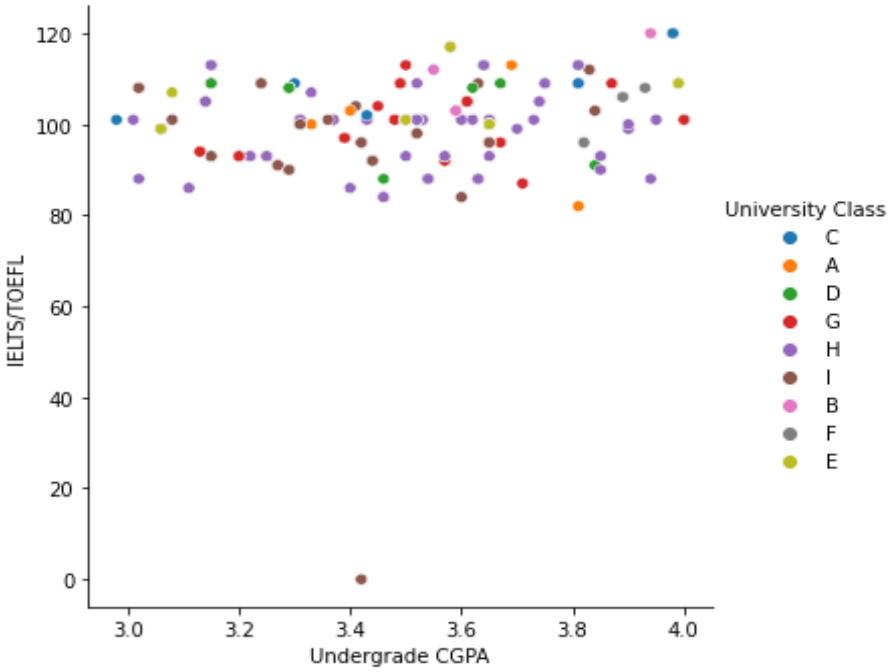
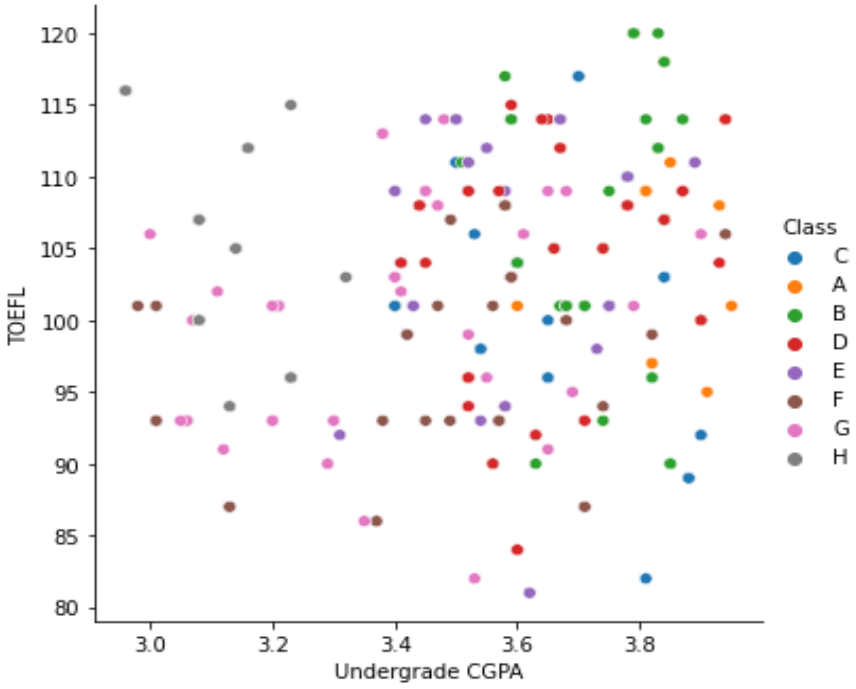


Fig. 5. The number of students who got accepted for PhD according to undergraduate university.



(a)

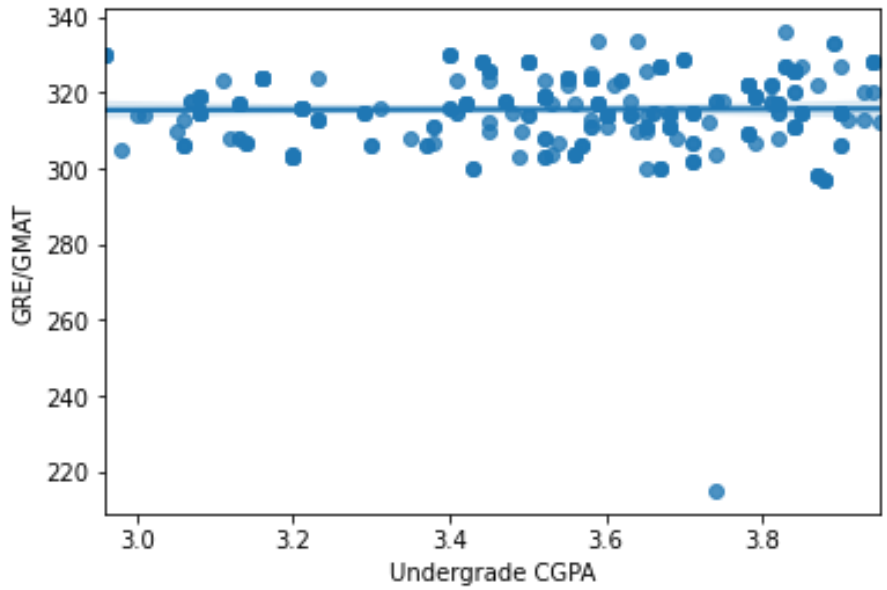


(b)

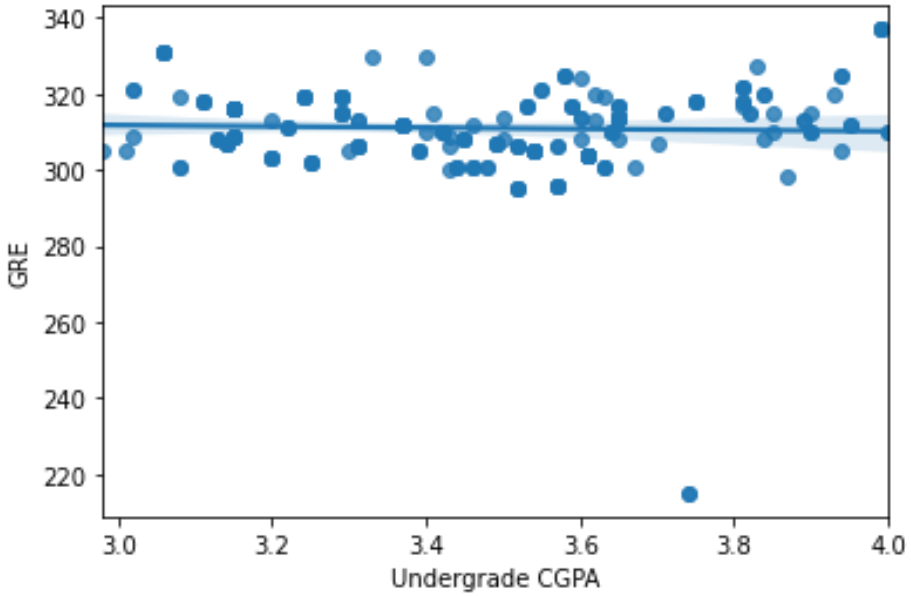
Fig.6. (a) PhD Students CGPA vs IELTS/TOEFEL data plotting

(b) MS Students CGPA vs TOEFL data plotting

Fig.6(a,b) shows the data plotting of our university classes between CGPA and English test (IELTS or TOEFL). For (a) we have anomaly data that has a value zero and its CGPA is just above 3.4 but for MS students, all the scores are from 80 to 120.



(a)



(b)

Fig.7. (a) PhD Students CGPA vs GRE Regression line (b) MS Students CGPA vs GRE Regression line

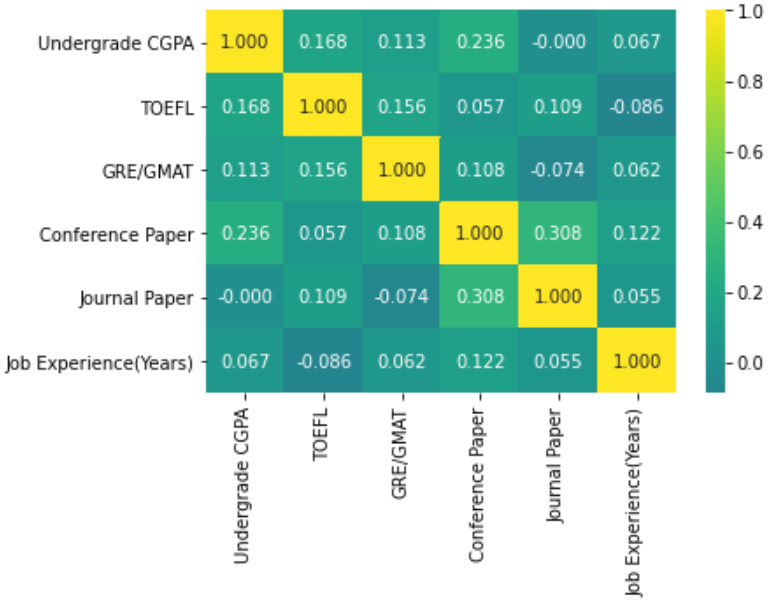
Fig.7(a, b) shows the regression line of CGPA vs GRE of PhD and MS students. Here in both PhD and MS students, we have data from way below 290 GRE score, which is 215.

Fig.8(a, b) shows the correlation matrix between features we are working with, and for both PhD (a) and MS (b), we can see that conference paper and journal paper has similarity over 30% as we know the characteristics of papers are same that's why these two features have high similarity. This similarities can be seen from co relation matrix graph.

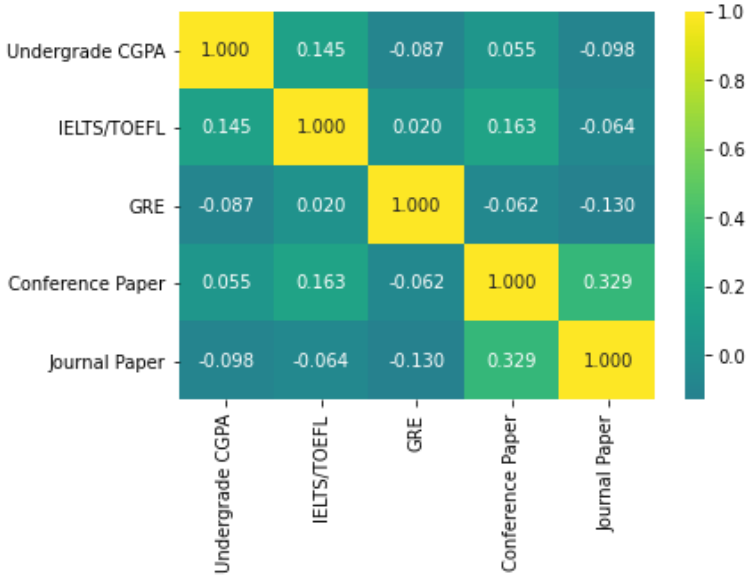
Finally, we have found out the important features [12] which mostly influenced our model. As we do work for the first time in machine learning algorithm so in the future work we will consider calculating new features from existing ones.

Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher node probability value refers to the more important feature. At the random forest level which we have used, the final feature importance is its average over all the trees. The sum of the feature's importance value on each tree is calculated and divided by the total number of trees, as shown in Eq. 1.

$$RFfi_i = \frac{\sum_{j \in all\ trees} normfi_{ij}}{T} \tag{1}$$



(a)



(b)

Fig.8. (a) Correlation matrix of PhD students between features

(b) Correlation matrix of MS students between features

In Eq. 1, $RFfi_i$ is the importance of feature i calculated from all the tress in the random forest model, $normfi_{ij}$ denotes the normalized feature importance for i in tree j and T is the total number of trees.

$$normfi_{ij} = \frac{fi_i}{\sum_{j \in all\ features} fi_j} \tag{2}$$

In Eq. 2, each feature's importance on a decision tree 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 Eq. 3. In Eq. 3, ni_j is the importance of node j, which can be obtained from Eq. 4 using Gini importance, assuming only two child nodes (binary tree). Here w_j is the weighted number of samples reaching node j, C_j. The impurity value of node j and $left(j)$ and $right(j)$ is the child node from left and right split on node j, respectively.

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

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \tag{4}$$

In Fig. 9, for MS data, it can be seen that the essential feature is undergraduate CGPA, which score is almost 0.7, and the second one is GRE, with a score of 0.15. Similarly, in Fig.10, for Ph.D. data, Bachelor's CGPA and GRE are the top two needed features for our model. So, it means that for our prediction, these two features will perform a crucial role.

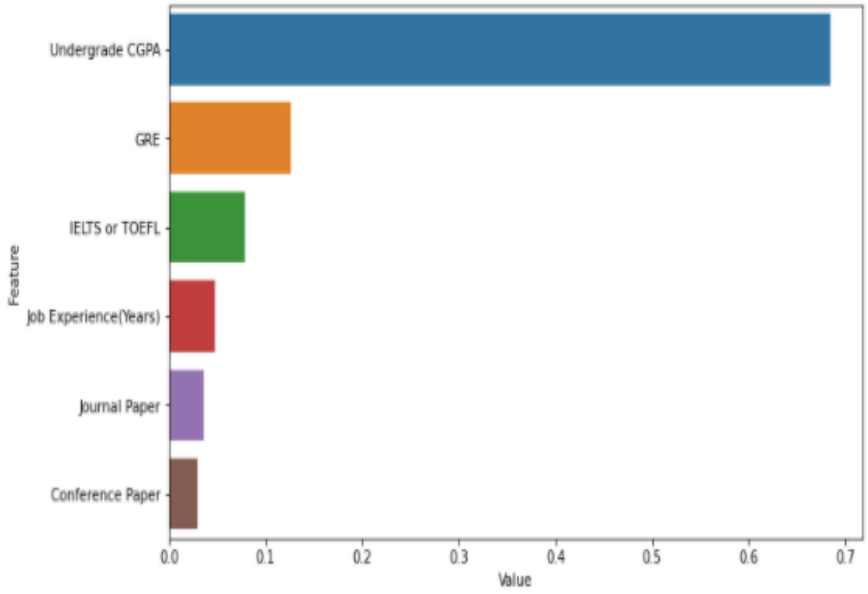


Fig. 9. Feature importance scores of MS students' data.

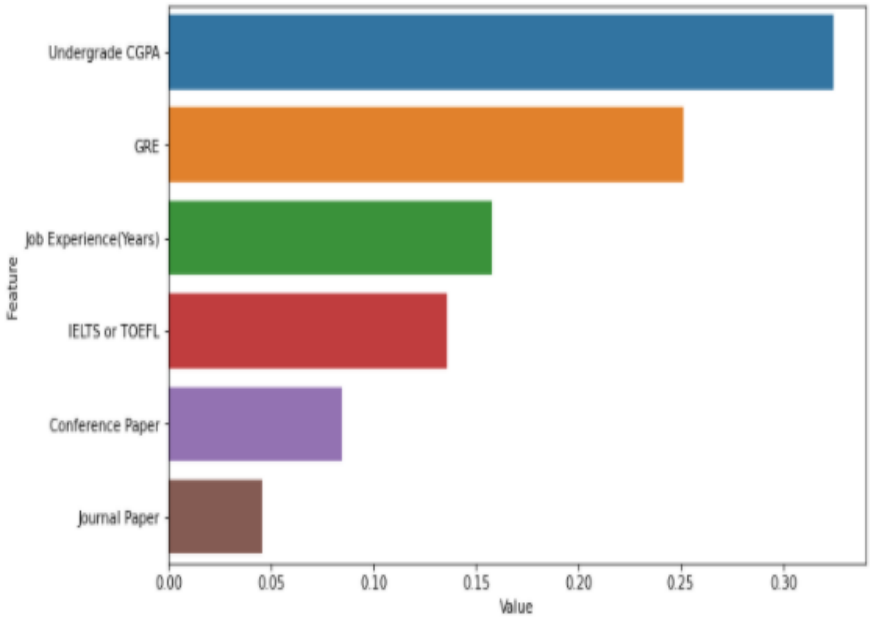


Fig. 10. Feature importance scores of PhD students' data

In our raw data, there was a lot of information about an individual student. Our model returned significantly biased results with all the provided data with a low accuracy score on the training step. We performed feature engineering and dimensionality reduction of the raw data and selected the following eight features, which consequently improved the predictive model's performance and effectiveness.

1. Undergrad University
2. Undergrad CGPA
3. TOEFL/IELTS
4. GRE
5. Conference Paper
6. Journal Paper
7. Job Experience (Years)
8. Undergrad University Class

Then we applied the required preprocessing to these data features, which are described in the following section.

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. Since we worked with GRE, IELTS, and TOEFL test results, we found some of the test results were unavailable as GRE is only required at the United States universities, and all the universities require either IELTS or TOEFL. We normalized the test scores and merged them into a single feature using Eq. 5. We also one-hot encoded the categorical data by creating binary or dummy variables for using them as input in the machine learning model. We then applied a transformation to numerical data by performing normalization on them. As we have used MinMaxScaler transformation, all of our data transformed between 0 to 1.

$$z = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (5)$$

For our prediction we have classified the universities based on the ranking into 9 classes (A, B, C, D, E, F, G, H, I) for MS data and 8 classes(A, B, C, D, E, F, G, H) for PhD data as explained in Section 3.1.

In Fig.3 and Fig.4, our dataset is biased towards the diction, we have classified the universities based on the ranking into 9 classes

H class for MS and G for PhD data. If we don't balance this data, our 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 SMOTE. In SMOTE, the minority class oversampled and made new data to increase it to the same number as the majority class. Here for MS, the majority class is H which has 76 samples, and for PhD the majority class is G which has 50 samples. After oversampling, all the classes of MS and PhD have 86 and 50 samples, respectively. So now, our dataset is equally distributed towards all of our predicted classes, which will help us to get an unbiased prediction. After oversampling, our dataset has increased. For PhD we have now (50*8) = 400 samples and for PhD (76*9) = 684 samples in total.

Modeling

1) Decision Tree

Decision trees are built by recursively splitting our training samples using the features from the data that work best for the specific task. This is done by evaluating certain metrics, like the Gini index or the Entropy for categorical decision trees, or the Residual or Mean Squared Error for regression trees. The process is also different if the feature that we are evaluating at the node is discrete or continuous. For discrete features all of its possible values are evaluated, resulting in N calculated metrics for each of the variables, being N the number of possible value for each categorical value. For continuous features the mean of each two consecutive values (ordered from lowest to highest) of the training data are used as possible thresholds. The result of this process is, for a certain node, a list of variables, each with different thresholds, and a calculated metric (Gini or MSE) for each variable/threshold tandem. Then, we pick the variable/threshold combination that gives us the highest/lowest value for the specific metric that we are using for the resulting children nodes (the highest reduction or increase in the metric). We won't go into how these metrics are calculated, as it is off the topic of this introductory post, however I will leave some resources at the end for you to dive deeper if you are interested. At the moment just think of these metrics (Gini for categorical trees and Mean Squared Error for regression trees) as some sort of error which we want to reduce.

We decided to use Decision trees as they are statistical models that can generate trees based on each feature's information gain and return the predicted output by making decisions based on the tree nodes [3]. We chose the ID3 algorithm [13], which is mostly used to generate decision trees. It uses a greedy top-down approach to choose the tree nodes base on the information gain of a particular feature, and Eq. 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. Here IG is information gain, WA denotes weight average, $E(P)$and $E(C)$ represent parent and child entropy, respectively.

$$IG = E(P) - WA \times E(C) \quad (6)$$

Eq. 7 is used for measuring entropy, which is a measure of disorder obtained by selecting a particular feature.

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

(7)

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). 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 decision tree model is a perfect choice as it can make step by step decisions from each of the features.

2) Random Forest

We also tried the random forest model, which 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 [3]. In the ensemble method, several weak learners are combined to build a more robust model. We chose decision trees as weak models for our model and then used the weak models to construct a powerful learner. 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 decision tree models. The data subset can be chosen with no replacement, i.e., pasting or relief, also known as Bagging or Bootstrap aggregating. The random forest model then follows the prediction that matches the most trees' output like a voting system. We have used random forest in our model as decision trees tend to overfit a lot. So, with random forest, we can eliminate the disadvantage of overfitting by averaging the result.

Summary

In this chapter, we have discussed the methodology of our project and how we achieved it with proper implementation. We have shown multiple implementations for our problem with proper parameters and process.

Chapter 4

Performance Evaluation

Introduction

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 three machine learning methods. At first, we have split the dataset into training and test subset by the ratio of 70:30. For showing the performance of each method, we have found out different evaluation matrices. They are precision, recall, F1-score, accuracy, weighted 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

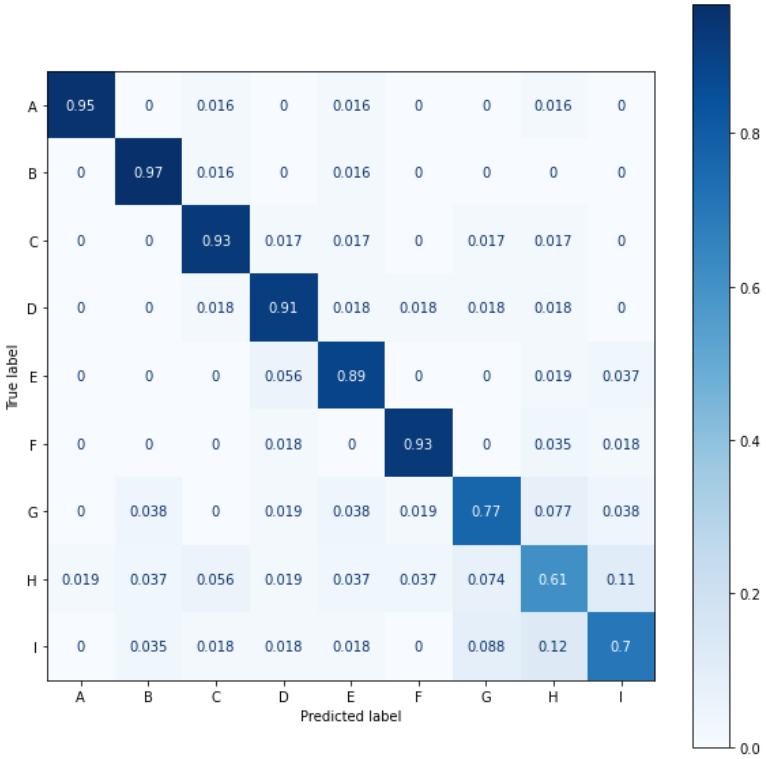
have been retrieved (TP divided by $TP + FN$). TP and FP mean True positive and False Positive, respectively, and FN means False Negative. We have also shown the ROC curve for the best-performed model.

Performance of Decision Tree Model

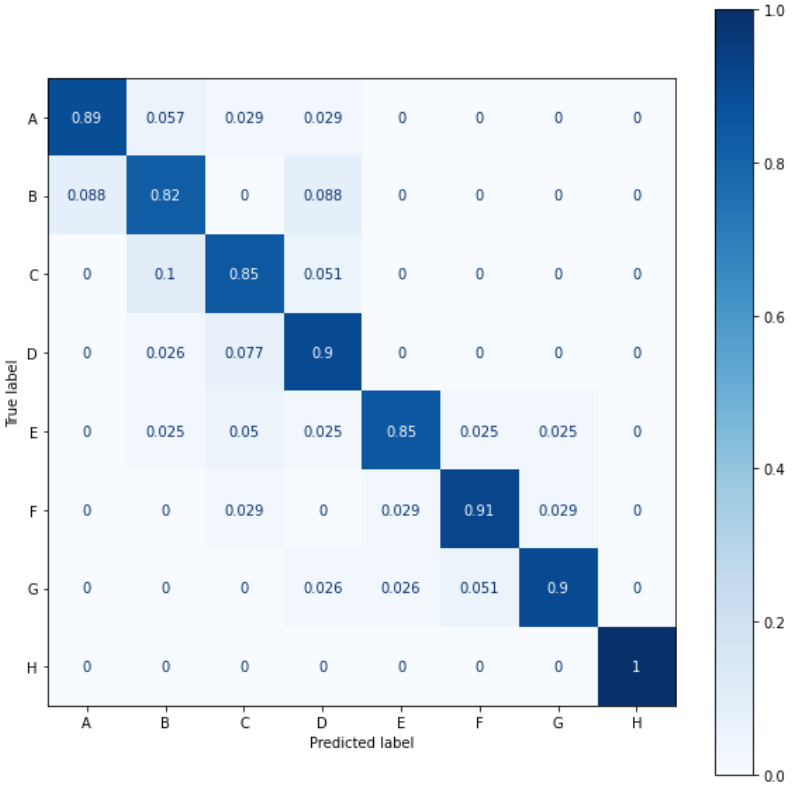
The decision tree is one of the most powerful machines learning algorithms that split a node by searching for the most important feature. We have made two models, one for the MS students and another for the PhD students. From Table 1, it can be found that the MS model's accuracy is lower than the PhD model. The result of MS metrics is 85% for F1-Score and precision and 86% for recall, and for PhD it has 89% accuracy for all of them. The overall accuracy for MS is 86% and 89% for PhD model.

Table 1: F1-Score, Precision, Recall, Accuracy and Weighted Accuracy for Decision Tree

	F1-Score	Precision	Recall	Accuracy	Support
MS	85%	85%%	86%	86%	171
PhD	89%	89%	89%	89%	100



(a)



(b)

Fig. 11 (a) Confusion Matrix of Decision Tree MS model
(b) Confusion Matrix of Decision Tree PhD model

Fig.11 (a, b) shows the confusion matrix of our MS and PhD models. Where we can see that both of our models are not biased. We got the lowest accuracy for MS(a) class H, so if we merge H and G classes, we can get a better result. Class B gives the highest accuracy, which is 98%. For PhD(b), class D gives the lowest accuracy, 81%, and got highest for Class H, 100%.If we do merge this is for the future work.

Performance of Random Forest Model

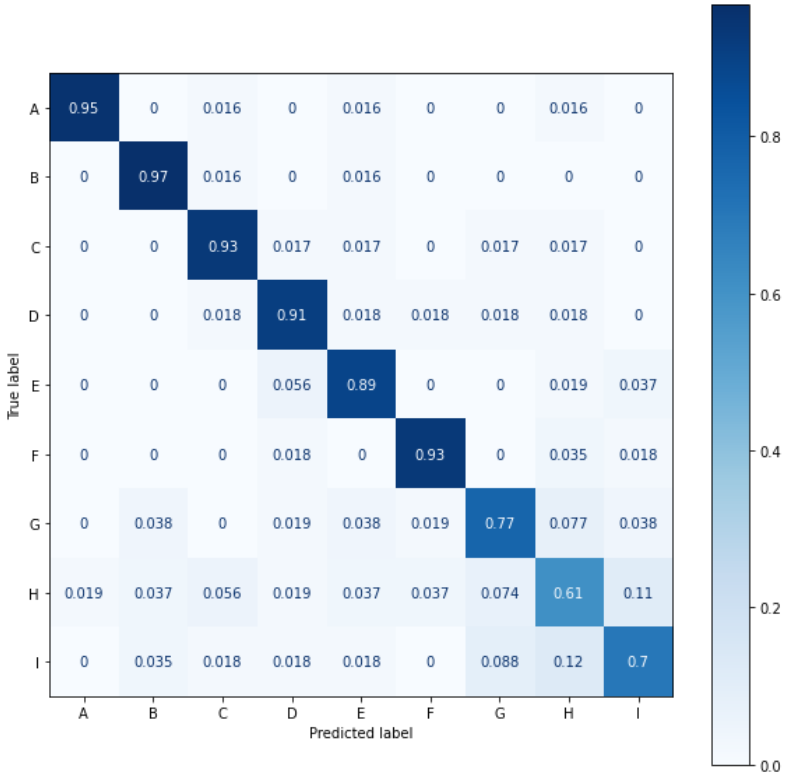
We have shown different result metrics of our random forest model in Table 2. From Table 1, it can be found that the MS model's accuracy is lower than the PhD model. The result of MS metrics is 85% for F1-Score and 86% for precision, recall and for PhD it has 86% for recall and 86% for F1-score and precision. The overall accuracy for both MS and PhD models is 86%.

Table 2: F1-Score, Precision, Recall, Accuracy and Weighted Accuracy for Random Forest

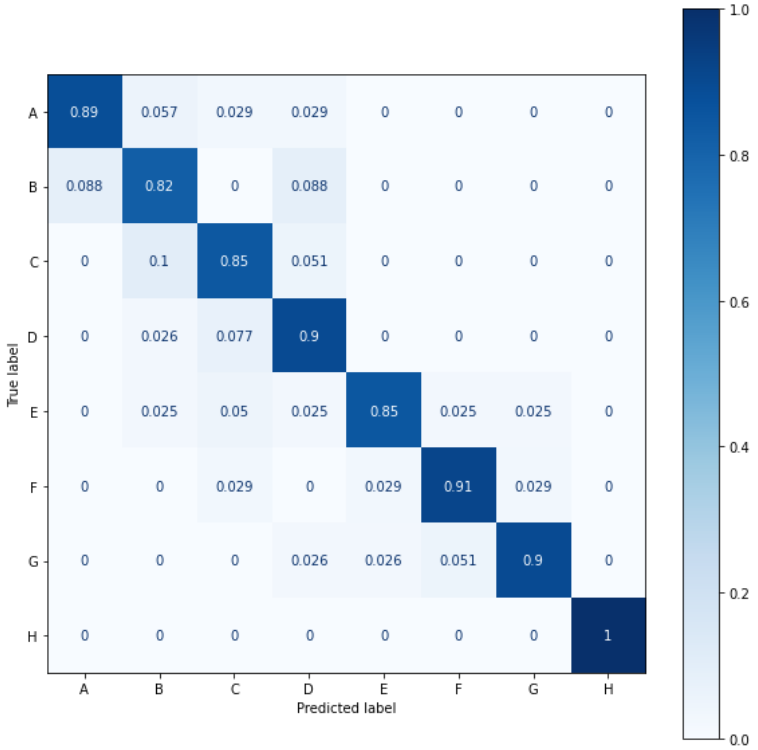
	F1-Score	Precision	Recall	Accuracy	Support
MS	85%	85%	85%	86%	171
PhD	89%	89%	89%	89%	100

Fig.12 (a, b) shows the confusion matrix of our MS and PhD models. Where we can see that both of our models are not biased. For MS(a) class H, we got the lowest accuracy so if we merge

these two classes, we can get a better result. Class B gives the highest accuracy, which is 98%. For PhD(b), class D gives the lowest accuracy, 81%, and got highest for Class H, which is 100%. For both of method decision tree and random forest, we got exactly the same accuracy.



(a)



(b)

Fig. 12 (a) Confusion Matrix of Random Forest MS model
(b) Confusion Matrix of Random Forest PhD model

Prediction With the Most Importance Features

We can observe that for MS data, the highest accuracy of 69% has been achieved by the random forest method. For PhD dataset, a similar accuracy has been attained by random forest and decision tree algorithms. We tried to implement the k-fold cross-validation technique [15] also, but unfortunately, the performance of the classification 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 of all, 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> Average CGPA and GRE > Average GRE, it is in class A and B
2. Else if CGPA > Average CGPA and GRE < Average GRE, it is in class C and D
3. Else if CGPA < Average CGPA and GRE> Average GRE, it is in class E and F
4. Else if CGPA < Average CGPA and GRE < 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 Fig. 9 and Fig. 10. 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.

Chapter 6
Limitations

Most of the works of the researches have limitations and future works. Our model has many more scopes to improve. We are the first to work with this data, so there are many more data analysis and processing that we cannot don't can be done to improve our model. The main work that we can do in the future is to delete data anomalies and add more data, for example, the data of 2021 which will help our model learn more. Secondly, we can use some boosting algorithms and neural networks or ensemble methods to increase our accuracy. But our main focus should be on data as it is custom-made data. There are a lot of noises in the data which we need to remove some features which we can merge and make one feature out of it. For example, we can merge conference paper and journal paper features into one feature named Papers.

Chapter 7
Conclusion & Future Work

The undergraduate students from developing countries like Bangladesh invest a lot of money, time, and energy while applying for graduate studies. In this work, machine learning algorithm 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, decision tree and random forest, to make a separate MS and PhD applicants model. We found that the random forest and decision tree 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 random forest algorithm to anticipate graduate admissions. The decision tree algorithm offers similar prediction accuracy to the random forest 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 adopt 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, letters of recommendation, etc.

Question

Need explanation on why this dataset needs machine learning models? What characteristics of the dataset requires you to adopt ML algorithms. Provide an explanation from a technical point of view.

Answer

At first, our dataset was a word file containing the name, age, CGPA, and all educational information of an MS or PHD accepted student to a foreign university. So we manually chose our features as the dataset was not that big. This takes a lot of time. So, in short, we make this dataset in a way so that we can use machine learning or any deep learning algorithm. We, first of all, choose all the popular features that have been used in some previous works also. We make two tabular data with a Predicting label University class. The features are CGPA, GRE score, IELTS, Paper, etc. As we are working with features and want to know by a student profile that in which university classes they might have high probabilities to get chances. A highly accurate model was not our acceptance in this work as from customized noisy data, you don't accept to get a highly accurate model right away. The dataset was unbalanced. We have also used the oversampling(data augmentation) method to get good accuracy. Also, our work is not on exceptional students. That's why we removed some outliers. Because there will always be some exception students who have low academic scores but got chances in highly ranked universities because of experience, our analysis is on the common students because they are the ones who always want to know in which ranges university they should try not the exceptional students. Regarding the research paper feature, you will also see that the students with high academic scores are ahead, and exceptional students are minimal, only one or two. So after making our dataset with proper features, we have tried different analyses such as correlation matrix to see the relation between our features and many more. Then we finally preprocessed data and trained the model. Preprocessing was crucial for our dataset as some features were discrete and some were continuous. So as we were trying to predict a class from multiple classes, what is the best option rather than a machine learning model? There are some dynamic algorithms for classification problems, and from them, we have used Decision Tree and Random Forest. If you want to apply some algorithms like machine learning or even rule based on a dataset, there are so many in the present. So there is always some further work on a dataset, especially new datasets like us. There are two types of researches that are very popular right now. One is to optimize an algorithm such as using regularization parameter, new loss function implementation, and another is to work with a noisy custom dataset, analyze it and increase the accuracy. Without any preprocessing, our accuracy on the first version of our dataset was below 50%. So for prediction choosing machine learning models was based on the previous paper's performance and popularity.

Bibliography

1. C. Romero and S. Ventura, Data Mining in E-Learning, vol. 2, WIT Press, 2006, pp. 1-19.
2. "University Rankings," QS World University Rankings. [online]. Available: https://www.topuniversities.com/university-rankings/world-university-rankings/2021. [Accessed: 09-August-2020].
3. A. K. Nandi, H. Ahmed, "Decision Trees and Random Forests," *Condition Monitoring with Vibration Signals: Compressive Sampling and Learning Algorithms for Rotating Machines*, IEEE, 2019, pp.199-224.
4. B. Schölkopf, J. Platt, and T. Hofmann, "AdaBoost is Consistent," *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, MIT Press, 2007, pp.105-112.
5. J. M. Rosa, "Introduction to ΣΔ Modulators: Fundamentals, Basic Architecture and Performance Metrics," *Sigma-Delta Converters: Practical Design Guide*, IEEE, 2018, pp.1-27.
6. N. T. N. Hien and P. Haddawy, "A decision support system for evaluating international student applications," *2007 37th Annual Frontiers in Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports, Milwaukee*, WI, pp. F2A-1-F2A-6, 2007.
7. A. Waters and R. Miikkulainen, "GRADE: Machine Learning Support for Graduate Admissions," AI Magazine, vol. 35, pp. 64-75, 2014.
8. M. S. Acharya, A. Armaan and A. S. Antony, "A Comparison of Regression Models for Prediction of Graduate Admissions," *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, Chennai, India, pp. 1-5, 2019.
9. I. Hmiedi, H. Najadat, Z. Halloush and I. Jalabneh, "Semi Supervised Prediction Model in Educational Data Mining," *2019 International Arab Conference on Information Technology (ACIT)*, Al Ain, United Arab Emirates, pp. 27-31, 2019.
10. P. Janani, V. P. Hema, S. P. Monisha, "Prediction of MS Graduate Admissions using Decision Tree Algorithm," *International Journal of Science and Research (IJSR)*, vol. 9, pp. 492 – 495, March 2020.
11. J. Huang, Y. F. Li, and M. Xie, "A Systematic Analysis of Data Preprocessing for Machine Learning- based Software Cost Estimation. Information and Software Technology", Elsevier, 2015, vol. 67, pp. 108-127.
12. A. P. Cassidy and F. A. Deviney, "Calculating feature importance in data streams with concept drift using Online Random Forest," *2014 IEEE International Conference on Big Data (Big Data)*, Washington, DC, 2014, pp. 23-28.
13. C. Jin, L. De-lin and M. Fen-xiang, "An improved ID3 decision tree algorithm," *2009 4th International Conference on Computer Science & Education*, Nanning, 2009, pp. 127-130.
14. Y. Freund and R. E. Schapire, "A Short Introduction to Boosting," *Journal of Japanese Society for Artificial Intelligence*, vol. 14, September 1999, pp. 771-780.
15. K. Pal and B. V. Patel, "Data Classification with k-fold Cross Validation and Holdout Accuracy Estimation Methods with 5 Different Machine Learning Techniques," *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2020, pp. 83-87.

16. S. Mitrofanov and E. Semenkin, "An Approach to Training Decision Trees with the Relearning of Nodes," 2021 International Conference on Information Technologies (InfoTech), 2021, pp. 1-5, doi: 10.1109/InfoTech52438.2021.9548520.
17. Y. Guo, Y. Zhou, X. Hu and W. Cheng, "Research on Recommendation of Insurance Products Based on Random Forest," 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), 2019, pp. 308-311, doi: 10.1109/MLBDBI48998.2019.000