**CSCI-5270-201**

**Machine Learning**

****

**Final Report**

**University Recommendation System**

Submitted By: Amit Maharjan

4/29/2024

# Introduction

University Recommendation System is a Machine Learning application that helps international students predict their chances of admission to graduate programs in 54 different universities located in the USA. The system is designed to assist international students in navigating the complex process of applying to graduate programs in the United States. With thousands of universities to choose from [1] and significant financial costs involved, the system aims to streamline the application process by providing personalized recommendations based on students' profiles and preferences. By leveraging predictive analytics and supervised learning algorithms, the system aims to improve the efficiency and effectiveness of the application process, ultimately assisting students in making informed decisions about their academic futures.

# Problem Statement

One of the most challenging factors international students face while pursuing the dream of studying abroad is choosing the right university. There are thousands of universities in the US [1], and it is not feasible to apply to all of these universities due to time and financial constraints. Most of these universities require an application fee ranging from $30 to $135, with an average of $74 per application [2]. Even if an individual chooses to apply to 10 different universities [3], the application fee cost is already $740. Additionally, these universities may also ask for official GRE and English Proficiency Test scores. The Educational Testing Service (ETS), which conducts GRE and TOEFL, charges $35 to send the official GRE test score [4] and $25 to send the official TOEFL test score [5]. Even if only 5 out of 10 universities ask for official test scores, this adds $300 to the cost, which in total becomes $1040. This amount when converted to Nepalese local currency becomes NRP 139,067.76, which is more than the average salary of a family in Nepal [6]. Therefore, the need to apply to the right university is important.

# Motivation

The motivation for this project originated from the challenges international students face when applying for graduate studies in the United States. The US is a premier destination for higher education, attracting numerous international applicants seeking to pursue their academic aspirations. However, the complex university application process creates barriers for individual applicants, including time constraints and financial limitations. With the ability to apply to only 10-15 universities at most [3], international students often find themselves struggling with the daunting task of selecting the right university. Having experienced these challenges firsthand, and recognizing that many others have encountered similar obstacles, there exists an importance of developing a solution to streamline the university selection process. A misstep in this decision-making process can have far-reaching consequences, affecting both time and financial resources, potentially resulting in delays in academic pursuits. Therefore, the development of a University Recommendation System using predictive analytics and supervised machine learning algorithms aims to assist international students in making informed decisions about their university choices, mitigating the risks associated with selecting the wrong institution.

# Questions that the project will answer

This project aims to alleviate the confusion and uncertainty faced by international students aspiring to apply for graduate studies at US universities. By providing insights into their likelihood of admission to specific institutions, students can make more informed decisions and better plan their academic pursuits. This proactive approach ensures that their time and financial resources are invested wisely, ultimately leading to a more efficient and effective application process.

Some questions that University Recommendation System will address via this project are:

1. Which universities are most suitable for an individual to apply based on their profile and preferences?
2. How can international students make informed decisions about where to apply and invest their time and resources?
3. What are the key factors international students need to consider for admission to their preferred universities?
4. Which ML model is most suitable for the data we have?

# Literature Review

Swaminathan et al. from the University of California San Diego have published a study comparing different machine learning models for their University Recommendation System [7]. They tested Baseline, K Nearest Neighbor, Random Forest, and SVM models to predict which universities international students could potentially get admitted to. Their findings showed that SVM performed the best, while Baseline had the lowest accuracy among the four models.

Elahi et al. conducted research on a University Recommender System that considered user preferences [8]. The study examined the factors that students prioritize when choosing a university and evaluated the usability of the recommendation system to assist students in making informed decisions about their study destination. The research aimed to assess how well the system could suggest universities to students. They discovered that methods like SVD and KNN were adept at predicting student preferences. Students felt that SVD understood their preferences well, while KNN provided diverse options. According to this study, the most crucial factors for students when selecting a university were teaching quality, cost, and research opportunities.

# Dataset

The dataset utilized for this project was extracted from Kaggle, which is a renowned platform for data enthusiasts, accessible via the link: <https://www.kaggle.com>. This dataset was originally compiled by Aditya Suresh Kumar from Eludix. It consists of 53,644 entries each containing 26 features. Impressively, the dataset exhibits a balanced distribution, with 27,955 entries for admitted universities and 25,689 for rejected ones. The dataset covers crucial aspects which are integral part of US university graduate program admissions, such as GRE scores (Verbal, Quantitative, and Analytical), TOEFL scores, undergraduate GPA, research and industry experience, as well as journal publications. It consists of data about students admitted to 54 diverse US universities, including prestigious institutions like Massachusetts Institute of Technology, Princeton University, and George Mason University. However, to effectively utilize this dataset for machine learning model development, preprocessing steps are required to ensure data accuracy and consistency.

A portion of the dataset is showcased below:



Below are all 26 columns of the dataset:

* userName
* major
* researchExp
* industryExp
* specialization
* toeflScore
* program
* department
* toeflEssay
* internExp
* greV
* greQ
* userProfileLink
* journalPubs
* greA
* topperCgpa
* termAndYear
* confPubs
* ugCollege
* gmatA
* cgpa
* gmatQ
* cgpaScale
* gmatV
* univName
* admit

# Implementation

## Data Collection

The dataset which is utilized for constructing the ML models required for the "University Recommendation System" was sourced from Kaggle and originally compiled by Aditya Suresh Kumar from Eludix. It comprises of 53,644 number of rows and 26 number of columns. The dataset contains information about Indian students who had applied to US universities for their graduate studies. Among these data, there exists columns "greV" and "greQ," which represent the scores obtained by students in GRE Verbal and Quantitative exams, respectively. These columns contain both the old and new GRE scoring systems which shows importance of converting them to a consistent format. To address this, a conversion scale dataset was extracted from Kaggle, containing 61 rows and 3 columns. Below is a sample of the GRE conversion dataset:



## Data Cleaning

Below are the data cleaning steps implemented in this project:

* The 'userName' column was dropped because it does not contribute any meaningful information to our project's classification of student admissions. Admission decisions for graduate programs in US universities are influenced by various other factors such as GRE scores, TOEFL scores, undergraduate GPA and many more. Hence, the 'userName' column has been removed.
* Similar to the 'userName' column, the 'userProfileLink' column is not significant for our classification task. Therefore, this column has also been removed.
* The 'program' column contains inconsistent data, such as 'Both MS and PhD' and 'MS/PhD,' which represent the same condition. To ensure consistency, the values of 'Both MS and PhD' were replaced with 'MS/PhD.' Additionally, for programs listed as 'MS/PhD,' two rows were created with identical information for all columns except the 'program' column - one labeled as 'MS' and the other as 'PhD.’
* The rows containing missing values in the 'program' column were removed due to the presence of numerous missing values in other columns for these rows.
* The columns 'greV' and 'greQ' contained scores from both the old and new scales for the GRE test. These values were converted from the old scale to the new scale using the conversion scale dataset available on Kaggle.
* The newly created columns 'newGreV' and 'newGreQ', along with the existing column 'greA', contained out-of-range values. The score range for the GRE Verbal and Quantitative sections is from 130 to 170, while the score range for the GRE Analytical Writing section is from 0 to 6. Consequently, rows with out-of-range values were deleted.
* There were numerous missing values in the columns 'gmatA', 'gmatV,' and 'gmatQ', representing the GMAT test scores obtained by the students. Only 23 entries contained valid data, while 51200 rows had incorrect or missing values. The 'major' column for these 23 rows included entries such as Computer Science, Engineering Management, Information Technology Management, MIS, and Information Technology. Typically, in real-world scenarios, these majors require the GRE test instead of the GMAT test. Consequently, these three columns were dropped entirely.
* The 'cgpa' column exhibits various scaling systems, with some values in percentages, others in a 10-point scale, and some in a 4 or 5 scale. Fortunately, our dataset includes a column named 'cgpaScale.' Therefore, we have divided the values in the 'cgpa' column by those in the 'cgpaScale' column to ensure consistency across the dataset.
* The resulting new column 'newCgpa' had 1260 missing values. Therefore, the rows with missing values in 'newCgpa' column was deleted.
* The column 'topperCgpa' has been dropped because it does not contribute any meaningful information in classifying admission or rejection to universities.
* The column 'toeflEssay' had 37659 missing values, which is high compared to the total remaining number of rows in the dataset. Therefore, instead of imputing the missing values, the column itself has been deleted.
* The column 'toeflScore' contained both missing values and out-of-range values. The score range for the TOEFL test is from 0 to 120. Therefore, rows with missing values in the 'toeflScore' column or values less than 0 or greater than 120 have been deleted.
* The column ‘specialization’ was also dropped as it had too many missing values i.e., 16675.
* The rows having missing values for ‘ugCollege’ column were also deleted.
* The values in 'major' column had too much text involved, so we dropped the column as a whole for simplicity.
* Similar to the 'major' column, the 'department' and 'ugCollege' columns were also dropped for simplicity.
* One Hot Encoding was applied to the columns: 'program' and 'Semester' due to their low number of distinct values.
* Label Encoding was applied to the columns 'Year' and 'univName' due to their numerous distinct values. The mapping value of the 'univName' column with its original value was stored to extract the real name of the university after the model predicts the decision.
* The outliers were identified from all these columns by plotting the boxplot diagram for each individual column and were removed using Isolation Forest.

## Feature Extraction

Below are the feature extraction steps implemented in this project:

* The column 'termAndYear' was splitted into two separate columns named 'Year' and 'Semester'. For instance, the value 'Fall - 2015' in the 'termAndYear' column was converted to Year = 2015 and Semester = Fall. Resultant sample of data after the operation is shown below:



* The resulting column 'Year' contained some out-of-range values such as 2, 6, 20131, 20133, etc. Therefore, the rows containing these out-of-range values in the 'Year' column were deleted. Additionally, rows with missing values in the 'Year' column were also eliminated.
* The values in the 'industryExp' and 'internExp' columns were added to create a new column named 'profExp', as both columns refer to professional working experience.
* There were three columns named 'researchExp', 'journalPubs', and 'confPubs', which somehow describe the research experience of the individual. Assuming that one journal publication or conference publication takes a 12-month time period, we created a new column named 'resExp' by multiplying the values in 'journalPubs' and 'confPubs' by 12 and adding the resultant value to the value in the 'researchExp' column.

## Feature Selection

Below are the feature selection steps implemented in this project

* Old GRE columns, 'greV', and 'greQ', were dropped as they were of no use after we had created new columns with consistent scaling.
* Similarly, the old column 'termAndYear' was also dropped as we had splitted the value of this column into two new columns.
* Old CGPA columns, 'cgpa', and 'cgpaScale', were also dropped as a new column had been created with consistent scaling.
* The column 'topperCgpa' does not provide any information in either predicting the university decision for the student or helping in calculating CGPA-related values for individual students. Therefore, this column has also been dropped.
* The old work experience columns, 'industryExp', and 'internExp', and the old research experience columns, 'researchExp', 'journalPubs', and 'confPubs', were also dropped as we had created new columns which contain the compiled values of all these columns.

## Model Building

The number of rows was reduced to 41,836 after all these preprocessing steps. The dataset resulted in a perfectly balanced dataset, with 20,920 records for the 'Rejected' class and 20,916 records for the 'Accepted' class. The dataset was then split into training and testing datasets in a 3:1 ratio, with 25% of the data segregated for testing and the remaining 75% used for training the models. Additionally, the dataset was standardized using the 'Standard Scaler' to facilitate faster convergence. Six different supervised machine learning models were employed to predict the classification of admission or rejection to universities. These machine learning models include:

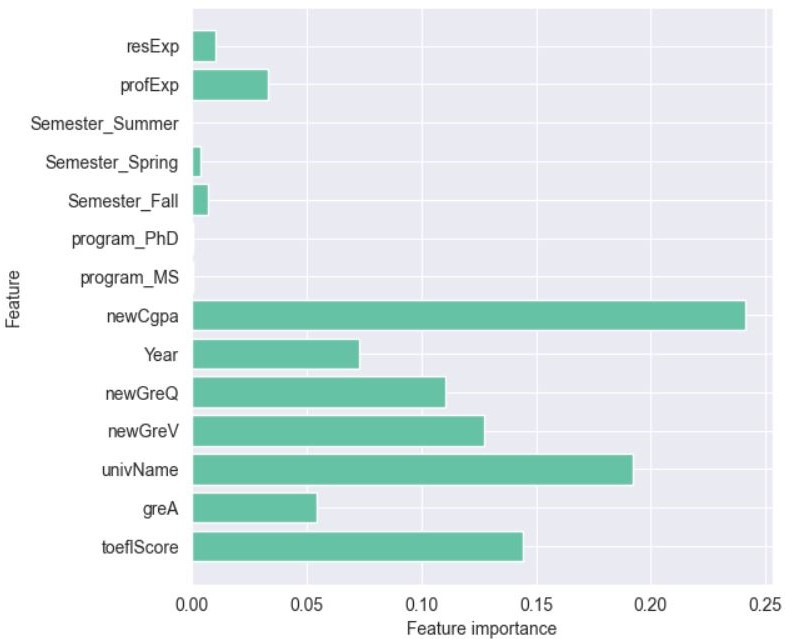
1. Logistic Regression
2. Decision Tree
3. Random Forest
4. Support Vector Machine (SVM)
5. K Nearest Neighbors (KNN)
6. Naive Bayes

## Results and Insights

Using the 25% of the dataset segregated as a test dataset during the data splitting process, the obtained accuracies are as follows:

* Logistic Regression - 53.53%
* Decision Tree – 63.88%
* Random Forest - 63.75%
* Support Vector Machine (SVM) - 56.64%
* K Nearest Neighbors (KNN) - 55.53%
* Naive Bayes - 52.72%

Among these models, the Decision Tree model achieved the highest accuracy at 63.88%. The feature importance graph derived from this decision tree indicates that columns such as 'newCgpa', 'univName', 'toeflScore', and GRE columns including 'newGreQ', 'newGreV', and 'greA' are significant factors contributing to the determination of admission or rejection to US universities for graduate studies.



During the application process, students typically apply to a mix of high-ranked universities where admission probability may be lower and some less competitive universities where admission is more assured. This approach ensures that students secure admission to at least some universities to avoid wasting time or potentially delaying their academic pursuits. Therefore, we assess the precision of the model for selecting high-ranked universities and the recall for identifying universities with high admission certainty. In our case, the ‘Decision Tree’ model can be used for selecting high-ranked universities due to its highest precision of 64%, while 'Naive Bayes' can be utilized for choosing universities with a high certainty of admission, given its high recall of 89.53%.

# Answers to Research Questions

Below are the responses to the questions that the University Recommendation System will tackle through this project:

1. Which universities are most suitable for an individual to apply based on their profile and preferences?

* By inputting the data of the individual, the University Recommendation System will predict the chances of getting admission or rejection to 54 different universities. Therefore, those universities which have been classified as having a chance of admission are the most suitable universities for individuals to apply to based on their profile.

1. How can international students make informed decisions about where to apply and invest their time and resources?

* By opting for the universities recommended by the University Recommendation System, students can wisely invest their time and resources. Conversely, avoiding universities not suggested by our models can save their time and resources, including money.

1. What are the key factors international students need to consider for admission to their preferred universities?

* The marks obtained in undergraduate studies, the rank or prestige of the undergraduate institute, English proficiency score, and GRE score are the key factors international students need to consider for admission to their preferred universities. This information has been derived from the feature importance graph obtained from the Decision Tree model mentioned above.

1. Which ML model is most suitable for the data we have?

* The ‘Decision Tree’ model is most suitable for selecting high-ranked universities due to its highest precision of 64%, while 'Naive Bayes' is most suitable for choosing universities with a high certainty of admission, given its high recall of 89.53%.

# Conclusion

After constructing the dataset based on my results, including undergraduate scores, GRE scores, English proficiency scores, industrial experience, research experience, etc., the system suggested the following universities where I could potentially gain admission or face rejection:

**Accept**

* Arizona State University
* Clemson University
* George Mason University
* New Jersey Institute of Technology

**Reject**

* Columbia University
* New York University
* Princeton University
* University of Texas Austin

Overall, I believe the model performed quite well. During my application process, I applied to New Jersey Institute of Technology and Clemson University, where I received admission, while the universities listed as rejected were top-ranked Ivy League institutions, where I anticipated facing difficulty in gaining admission.

In conclusion, the developed recommendation system has proven to be effective in improving decision-making for international students seeking admission to universities. The utilization of machine learning algorithms, particularly the Decision Tree model, has significantly enhanced accuracy in predicting admission outcomes. Furthermore, the Naive Bayes model demonstrated superior recall, emphasizing its efficacy in identifying universities where admission is more certain. These findings underscore the importance of employing advanced analytics tools to aid students in navigating the complex process of university selection, ultimately facilitating informed decisions and optimizing their chances of admission.

# Future Research Directions

Future research directions could explore leveraging Natural Language Processing techniques to enhance data preprocessing, especially in addressing challenges encountered with the 'Major' and 'Department' columns. This could involve developing algorithms to automatically extract relevant information from textual data, thereby improving the efficiency and accuracy of data preprocessing tasks. Furthermore, there is potential for investigating the integration of Letter of Recommendation (LOR) and Statement of Purpose (SOP) data into the recommendation system, which could enrich the predictive capabilities and provide a more comprehensive assessment of students' suitability for admission to specific universities.

# Learning Outcome

The project provided valuable insights into the application of various machine learning algorithms, encompassing six different supervised techniques. It underscored the significance of employing diverse data preprocessing techniques to ensure the quality and reliability of the input data. Additionally, the importance of scaler and normalization methods in standardizing features for improved model performance was highlighted. Furthermore, the project demonstrated the necessity of effectively handling additional data post-model development to adapt to evolving requirements. Finally, the utilization of encoding techniques showcased the importance of transforming categorical data into a format suitable for machine learning algorithms, thereby enhancing predictive accuracy and robustness.

# Reference List

[1] *How Many Universities Are in the U.S. and Why That Number Is Changing*. (n.d.). Retrieved February 25, 2024, from <https://www.usnews.com/education/best-colleges/articles/how-many-universities-are-in-the-us-and-why-that-number-is-changing>

[2] *College Application Fees: What It Costs to Apply | BestColleges*. (n.d.). Retrieved February 25, 2024, from <https://www.bestcolleges.com/research/college-application-fees-how-much-does-it-cost/#graduate-application-fees>

[3] *How Many Colleges Should You Apply To? | Coursera*. (n.d.). Retrieved February 25, 2024, from <https://www.coursera.org/articles/how-many-colleges-should-i-apply-to>

[4] *Sending Your GRE General Test Scores*. (n.d.). Retrieved February 25, 2024, from <https://www.ets.org/gre/test-takers/general-test/scores/send-scores.html#accordion-d35ebe9746-item-75f4d8a94c>

[5] *Sending Your TOEFL iBT Scores*. (n.d.). Retrieved February 25, 2024, from <https://www.ets.org/toefl/test-takers/ibt/scores/send-scores.html>

[6] *What Is The Average Salary In Nepal? And Its Minimum Wage*. (n.d.). Retrieved April 27, 2024, from <https://www.timechamp.io/blogs/what-is-the-average-salary-in-nepal-and-its-minimum-wage/>

[7] Swaminathan, R., Manley, J., A53096254, G., Krishnakumar, S., & Suresh, A. (n.d.). *University Recommender System for Graduate Studies in USA*. Retrieved February 25, 2024, from [www.edulix.com](http://www.edulix.com)

[8] Elahi M, Starke A, El Ioini N, Lambrix AA and Trattner C (2022) Developing and Evaluating a University Recommender System. Front. Artif. Intell. 4:796268. doi: 10.3389/frai.2021.796268