

Project Progress Report

University Recommendation System



CSCI-5270-201 - Machine Learning

SUBMITTED BY: AMIT MAHARJAN

March 19, 2024

**Project Progress Report**

# **Project Progress**

## **Task Performed:**

Data Collection: I have found the dataset for my project “University Recommendation System” on Kaggle (<https://www.kaggle.com>) which contains information about Indian students. Kaggle is a popular platform for data enthusiasts, which hosts various resources like competitions and collaborative tools. This dataset was initially collected by Aditya Suresh Kumar from Eludix, which comprises 53,644 rows with 26 features each. This dataset appears to be well-balanced, with 27,955 entries for admitted universities and 25,689 for rejected ones. This dataset covers essential factors that are considered by US universities for graduate program admissions—like GRE scores (Verbal, Quantitative, and Analytical), TOEFL scores, undergraduate GPA, research and industry experience, journal publications, and more. It includes information about students admitted to 54 different US universities, including prestigious institutions like Massachusetts Institute of Technology, Princeton University, and George Mason University.

A portion of the dataset is showcased below:



Data Preprocessing: Following data preprocessing steps have been completed:

1. The 'userName' column has been 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.
2. The 'program' column contained 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.'
3. Similar to the 'userName' column, the 'userProfileLink' column is not significant for our classification task. Therefore, this column has also been removed.
4. The columns 'greV' and 'greQ' contain scores from both the old and new scales for the GRE test. We have converted the scores from the old scale to the new scale using the conversion scale dataset available on Kaggle. Here is a sample of the GRE conversion dataset:



1. The score scale of the GMAT test is as below:  
   a. Quantitative Section: Scored from 0 to 60

b. Verbal Section: Scored from 0 to 60

c. Analytical Writing Assessment (AWA): Scored from 0 to 6 in half-point increments

Based on these values, we have the following out of range score for GMAT in our dataset:  
a. gmatA = 102

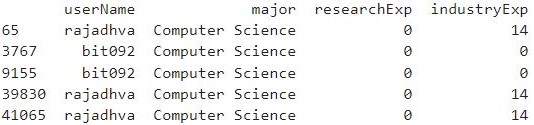
b. gmatQ = 102

c. gmatQ = 168

d. gmatV = 102

e. gmatV= 152

These values belong to the following 5 rows:



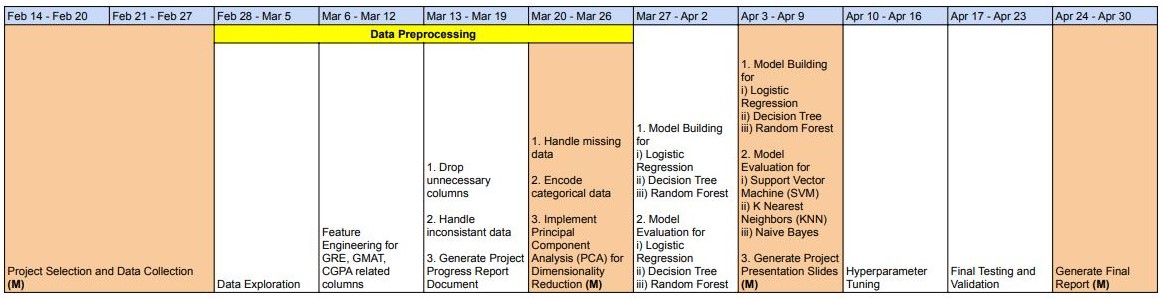
Therefore, we have removed these 5 data from our data set.

1. In real-world scenarios, individuals typically take either the GRE test or the GMAT test, but not both. However, we found 23 rows of data containing scores for both tests. To address this, we duplicated these rows, ensuring that each row corresponds to a single test, GRE or GMAT.
2. During the admission process for graduate programs in US universities, the committees typically consider either the GRE score or the GMAT score. Both of these tests have different scoring scales. To standardize these scores and facilitate comparison, we have transformed their values to range between 0 and 1. Additionally, we have merged both GRE and GMAT columns into a single column for streamlined analysis. As a result of this process, new columns have been created, namely 'GRE\_GMAT\_A', 'GRE\_GMAT\_V', and 'GRE\_GMAT\_Q', replacing the original columns 'greA', 'greV', 'greQ', 'gmatA', 'gmatV', and 'gmatQ'.
3. The column 'termAndYear' has been split into two separate columns named 'Year' and 'Semester'. For instance, the value 'Fall - 2015' in the 'termAndYear' column has been converted to Year = 2015 and Semester = Fall.
4. 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.

Challenges and Difficulties: Currently, the primary challenge I'm encountering lies in data preprocessing. Fortunately, having applied to various universities myself, I possess a good understanding of the admission process for graduate programs in US universities. Hence, I'm leveraging this domain knowledge to preprocess the data. Approximately 70% of the columns have undergone preprocessing. However, certain columns, such as 'Major', present a significant challenge due to the presence of inconsistent string values. I'm exploring potential solutions, considering either dropping this column entirely or assigning superclass values to its entries. For instance, replacing specific majors like computer engineering, computer science, information systems, and information technology with a broader category such as CS.

# **Project Timeline**

The chart below shows the project timeline for the "University Recommendation System." It includes tasks and milestones that are already done and those planned to be finished by the date mentioned in the chart.



**Notes:** Sections highlighted in brown with the word "(M)" at the end indicate milestones.

## **Completed Tasks and Milestones:**

* Project Selection and Data Collection (M)
* Data Exploration
* Feature Engineering for GRE, GMAT, CGPA related columns

## **Pending Tasks and Milestones:**

* Drop unnecessary columns
* Handle inconsistent data
* Generate Project Progress Report Document
* Handle missing data
* Encode categorical data
* Implement Principal Component Analysis (PCA) for Dimensionality Reduction (M)
* Model Building for
  + Logistic Regression
  + Decision Tree
  + Random Forest
* Model Evaluation for
  + Logistic Regression
  + Decision Tree
  + Random Forest
* Model Building for
  + Logistic Regression
  + Decision Tree
  + Random Forest
* Model Evaluation for
  + Support Vector Machine (SVM)
  + K Nearest Neighbors (KNN)
  + Naive Bayes
* Generate Project Presentation Slides (M)
* Hyperparameter Tuning
* Final Testing and Validation
* Generate Final Report (M)

## **Deliverables:**

* Final Project Report
* Final Presentation Slides
* Trained Machine Learning Model
* Code Repository

**Final project presentation date:**

* April 26