

Project Proposal

University Recommendation System



CSCI-5270-201 - Machine Learning

SUBMITTED BY: AMIT MAHARJAN

February 27, 2024

# **Project Overview**

## **Project Type**

Predictive analytics using supervised learning algorithms

## **Problem that will be addressed**

The United States is widely regarded as a premier destination for graduate studies among international students. However, the university application process can prove challenging for individual applicants. Students invest significant time and financial resources in applying to specific universities, with the outcome often falling short of their expectations. Consequently, their efforts may feel wasted if they do not receive the desired admission results. To address this issue, I propose the development of a University Recommendation System. This system will leverage predictive analytics to identify USA universities with a higher likelihood of admission approval based on each student's unique profile.

## **Motivation for the problem**

As an international student studying at ETSU in the USA, I've experienced the challenges of applying to universities firsthand. With thousands of universities across the country [1], it's simply not feasible to apply to all of them. Due to constraints like time and finances, most individuals apply to only 10-15 universities at most [2].

The application fees alone can add up significantly. Each university charges anywhere from $30 to $135, with an average of $74 per application [3]. For example, applying to just 10 universities could cost $740. But that's not all; many universities also require official GRE and English Proficiency Test scores.

If, say, 5 out of the 10 universities an individual applies to require official scorecards, the additional costs can pile up. Sending GRE scores costs $35 per university [4], while TOEFL scores cost $25 each [5]. This adds another $300 to the total expenses, bringing it to $1040 in total.

For someone from Nepal, that's equivalent to 137,886.11 Nepali Rupees—a substantial amount, considering it surpasses the average monthly salary in Nepal.

## **Previous Work**

Swaminathan et al. from the University of California San Diego have published a study comparing different machine learning models for their University Recommendation System [6]. 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 [7]. 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 Overview**

I have discovered a dataset on Kaggle (<https://www.kaggle.com>) containing information about Indian students. Kaggle, a popular platform for data enthusiasts, hosts various resources like competitions and collaborative tools. The dataset, initially collected by Aditya Suresh Kumar from Eludix, comprises 53,644 rows with 26 features each. It appears to be well-balanced, with 27,955 entries for admitted universities and 25,689 for rejected ones. Covering essential aspects 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—the dataset seems comprehensive. However, it requires preprocessing before building machine learning models. 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:



# **Problem Solving Approach**

Tentative steps to be carried out for building the University Recommendation System include:

1. Data Preprocessing: The dataset exhibits several issues, such as numerous missing values and inconsistencies. For instance, the "major" column contains both acronyms and full forms, while GRE scores are provided in both old and new formats. Additionally, undergraduate aggregate scores vary in scale, with some in GPA, others in a 10-point scale, and some in percentages. Moreover, there are unnecessary columns that do not contribute to predicting university lists. Therefore, the initial step involves preprocessing the data to ensure consistency and readiness for model training, aiming for high accuracy.
2. Building and Training ML Models: The preprocessed datasets will serve as the foundation for constructing and training the ML models using various popular supervised algorithms. For this project, we will employ several algorithms that were extensively covered during our lectures and lab sessions. These models include:
   * 1. Logistic Regression
     2. Decision Tree
     3. Random Forest

In addition to the aforementioned models, we will also explore other supervised algorithms that were not extensively discussed during this course, such as:

* + 1. Support Vector Machine (SVM)
    2. K Nearest Neighbors (KNN)
    3. Naive Bayes

By utilizing a diverse set of algorithms, we aim to comprehensively evaluate and compare their performance in predicting university recommendations.

1. Testing the ML Models: To assess the performance of our ML models, we will adhere to the 75:25 rule principle for splitting the dataset into training and testing subsets. Accordingly, after training each model using 75% of the dataset, we will evaluate its effectiveness using the remaining 25% of the data. By comparing the prediction results from each of the six models mentioned earlier, we aim to identify the most suitable model for our project—the University Recommendation System.
2. Predicting Outcomes: Upon completing the training and testing phases, we will proceed to predict outcomes for a sample profile. This involves applying the trained model to the profile data and generating a list of universities where the individual may have received admission. Through this process, we aim to provide personalized recommendations tailored to the individual's qualifications and preferences.

# **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?

# **Expected Learning Outcomes**

This project encompasses all the techniques covered in our course, employing a list of six models to predict admission outcomes at US universities and classify them as either acceptance or rejection. Through this practical application, I aim to showcase my understanding of these concepts in a real-world context. Additionally, I will explore three additional models not covered in the course, allowing me to visualize their behavior and gain insights into their strengths and limitations.

Moreover, my concurrent enrollment in the Data Analytics and Visualization course this semester provides me with valuable skills in preprocessing and analyzing datasets.

As highlighted in a survey conducted among Crowdflower employees in 2015, cleaning and organizing data can be a time-consuming task, with 66.7% of respondents citing it as a major challenge before data analysis. Hence, this project offers an opportunity to learn techniques for preprocessing data and ensuring dataset cleanliness, thereby enhancing my proficiency in data management and analysis.

# **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] *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>

[3] *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>

[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] 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)

[7] 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