

A  
*Project Report*  
on  
**University Recommendation System**

*submitted in partial fulfillment of the requirements  
for completion of DSci LAB*

*of*  
**TY COMP**  
*in*  
**Computer Engineering**  
*by*

**Avani Dhongde**  
**112003036**

**Shruti Jadhav**  
**112003050**

**Juhi Shekokar**  
**112003059**

*Under the guidance of*

**Archana Patil**  
*Professor*  
*Department of Computer Engineering*



Department of Computer Engineering,  
COEP Technological University (COEP Tech)  
(A Unitary Public University of Govt. of Maharashtra)  
Shivajinagar, Pune-411005, Maharashtra, INDIA

May, 2023

# 1 Problem Statement

For an aspiring student who wants to apply for higher studies in other countries, university selection process is a challenging task. Lot of different criteria need to be considered during application process based on individual's requirement. This problem can be addressed by modeling a recommender system based on various classification algorithms. In this project based on the Graduate and Undergraduate student dataset and user profile, a list of 5 best universities will be suggested based on which the applicant can apply to those universities and maximize his chances of getting admission into those universities.

## 2 Approach

As we have large set of data and User profile, we planned to use Knowledge based recommendation techniques using case based knowledge recommendation model. We used KNN (K Nearest Neighbors Model).

### 2.1 KNN Model

In KNN or K Nearest Neighbors Model, the trained data is compared with test data and distances are calculated using Euclidean distance. It then classifies an instance by finding its nearest neighbors and recommend the top n nearest neighbor universities. Algorithm is stated as below.

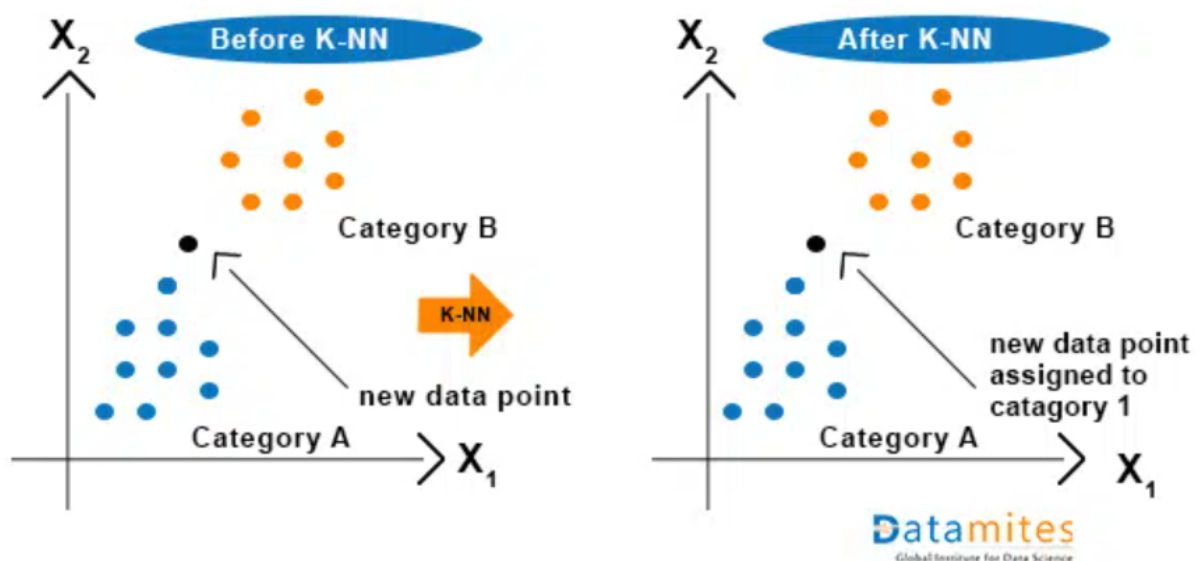


Figure 1: KNN

1. Initialize the value of k
2. For getting recommendation, iterate from 1 to number of trained data.
3. Calculate distance between test data and each row
4. Sort the distances in ascending order
5. Get top k rows and recommend to the user

### 3 Methodology

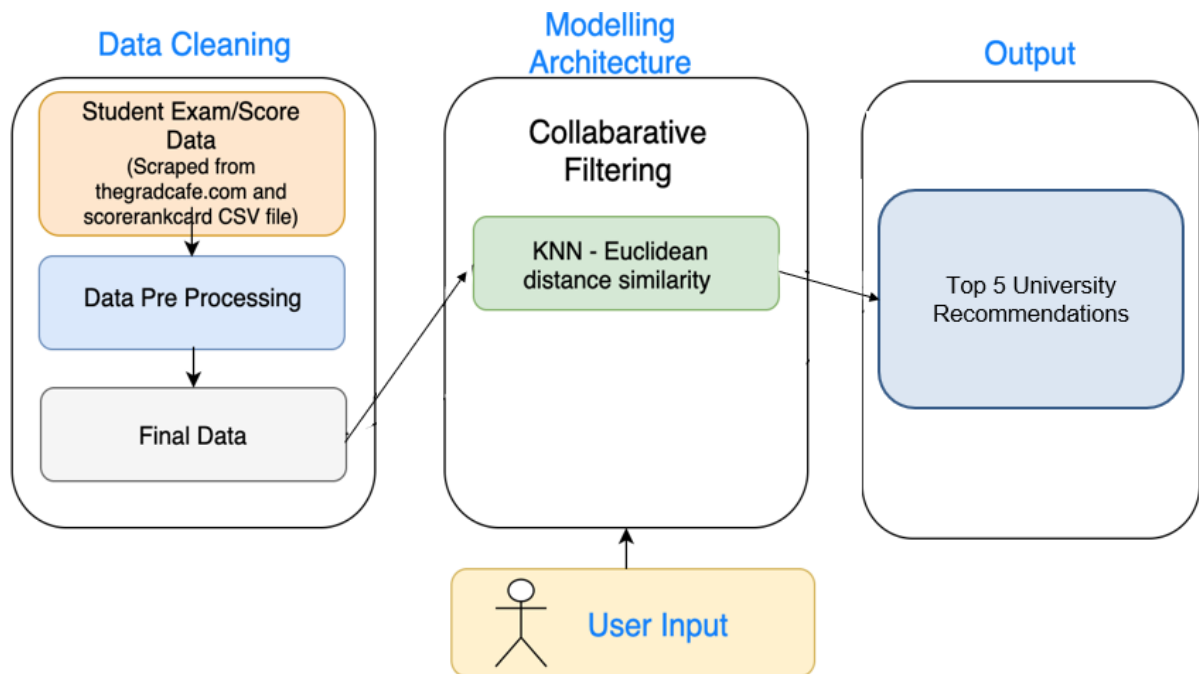


Figure 2: Methodology

#### 3.1 Dataset Generation

The first step in building any recommendation system is the identification of the data set.

Graduate student data was scraped from [www.thegradcafe.com](http://www.thegradcafe.com). About 25000 rows of raw student data was obtained as a result of web scraping, which is being processed to use as final dataset.

The Graduate Student data has the columns related to their GRE and TOEFL scores and University name they got admitted/rejected into.

In order to use the obtained data for our analysis, preprocessing and cleansing of dataset is required. The following are the data processes we did in our project

1. Data Scraped from website are loaded into different files. Merged all data.
2. Data from the admitted student rows are taken and rejected student rows are deleted.
3. Column names are set to the dataset.
4. The new/old GRE scores were also cleansed
5. Null values are deleted or filled with appropriate values
6. GPA scores available were based on different point systems, so all the GPA scores were uniformly scaled to 4 point scale by using normalize functions.
7. All the unnecessary columns are dropped.
8. Changed the order of columns making train and test dataset for algorithm

```
data.columns = ['univName', 'major', 'program', 'season', 'decision', 'Method', 'decdate', 'decdate_ts', 'cgpa', 'greV', 'greA', 'is_new_gre', 'gre_subject', 'status', 'post_data', 'post_timestamp', 'comments']
data.head()
```

	univName	major	program	season	decision	Method	decdate	decdate_ts	cgpa	greV	greQ	greA	is_new_gre	gre_subject	status	post
0	University Of Waterloo	Systems Design Engineering	MS	NaN	Accepted	Website	(1, 7, 2019)	1.561964e+09	NaN	NaN	NaN	NaN	NaN	NaN	International	
1	Northeastern University	Electrical Engineering	PhD	F19	Rejected	Website	(8, 7, 2019)	1.562569e+09	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
2	The University Of Auckland	Electrical And Electronic Engineering	MS	NaN	Accepted	Website	(19, 6, 2019)	1.560928e+09	NaN	NaN	NaN	NaN	NaN	NaN	International	
3	Radford University	Counseling Psychology PsyD.	Other	F19	Accepted	Phone	(4, 3, 2019)	1.551686e+09	NaN	NaN	NaN	NaN	NaN	NaN	American	
4	University Of Chittagong	Computer Science	MS	NaN	NaN	Other	(9, 7, 2019)	1.562656e+09	3.2	163.0	168.0	4.0	True	NaN	International	

Figure 3: Data before Pre-Processing

### 3.2 Pre-Processing

Pre-processing is the stage where the data is changed after being obtained. It's also to help the classifier to perform better by reducing some bad data that may cause an inaccuracy. Pre-Processing cleans our webscrapped dataset and makes it ready for use in the model.

	univName	cgpa	greV	greQ	greA
14	Ohio State University	4.00	150.0	166.0	3.0
17	Texas A&M University	3.57	157.0	151.0	5.5
46	University Of California, Irvine	3.66	155.0	167.0	4.0
64	Boston University	3.10	161.0	157.0	4.0
203	Oregon State University	3.38	154.0	170.0	4.0

Figure 4: Final Dataset

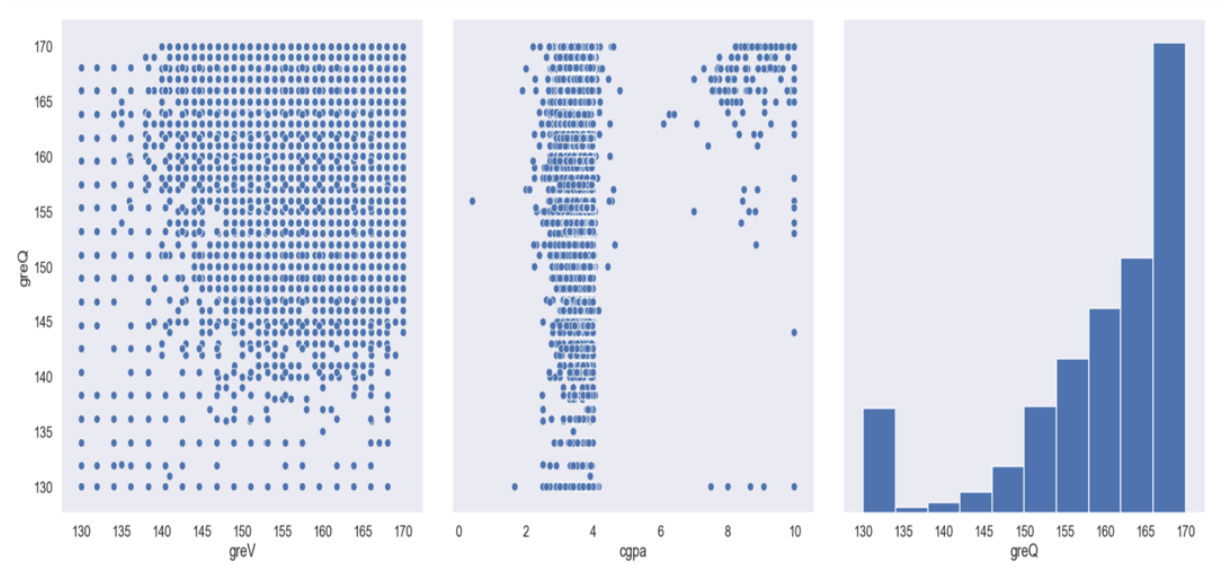


Figure 5: Data Visualization

### 3.3 Implementation

In KNN, trained and Test data will be sent to algorithm to find the euclidean distance and the top 5 nearest neighbors are taken into consideration.

The features like GREA, GREV, GREQ, CGPA of the user as test data are taken as features and provide weightage to them to find the similarity score. and provide top 5 recommendations to the user.

Developed a User interface where user can easily provide his/her score details to the application and get recommendations.

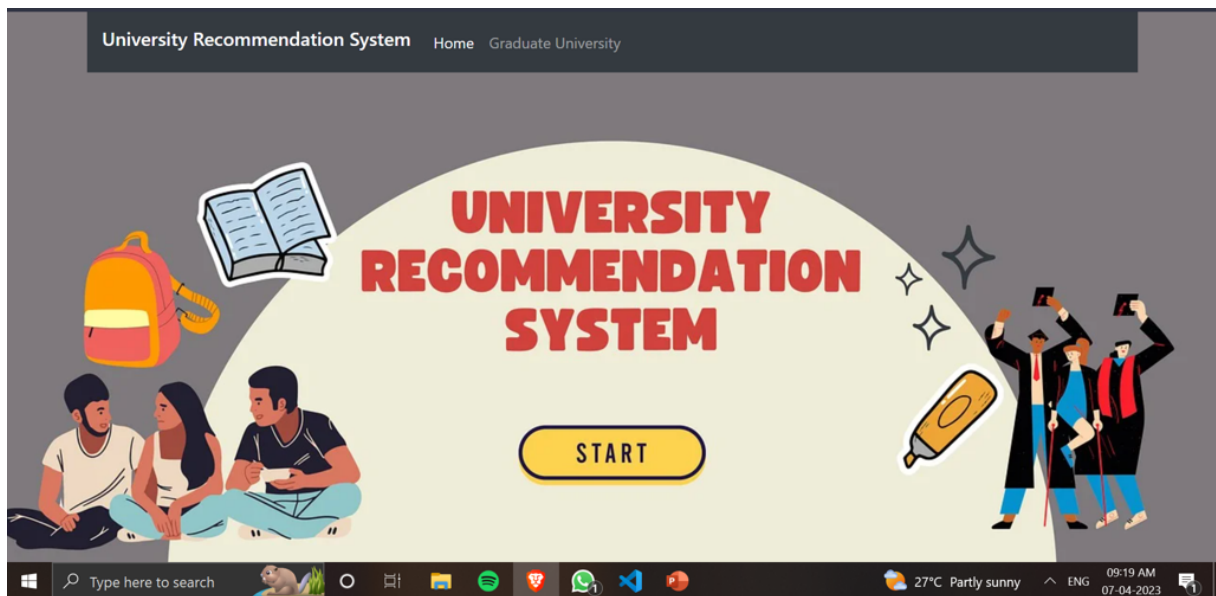


Figure 6: Home Page

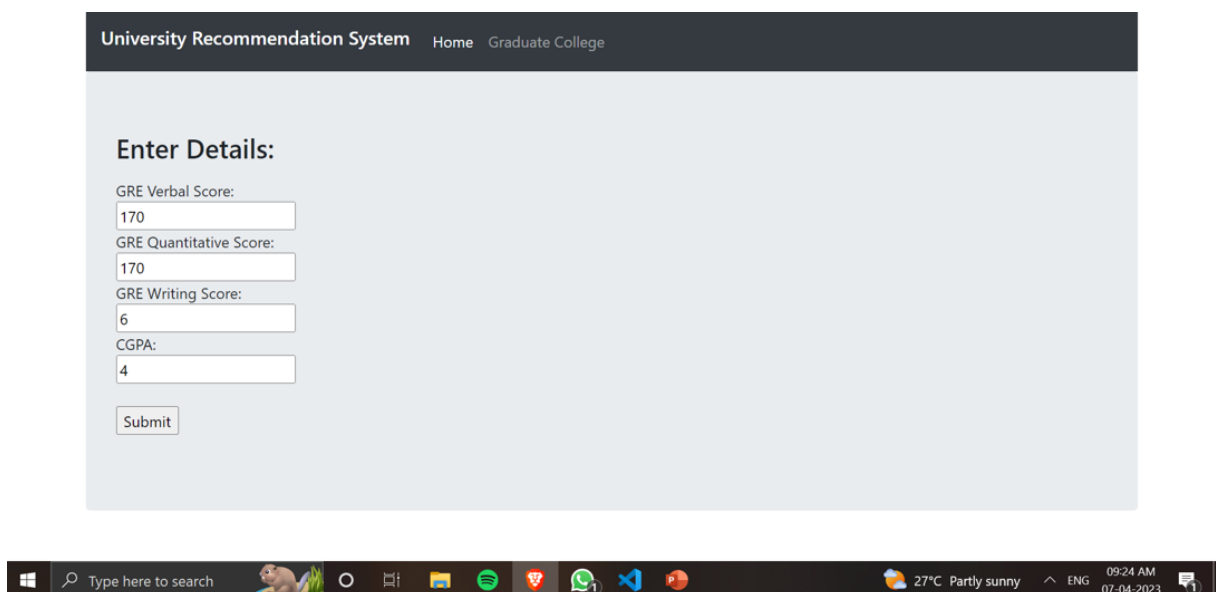


Figure 7: Enter Details

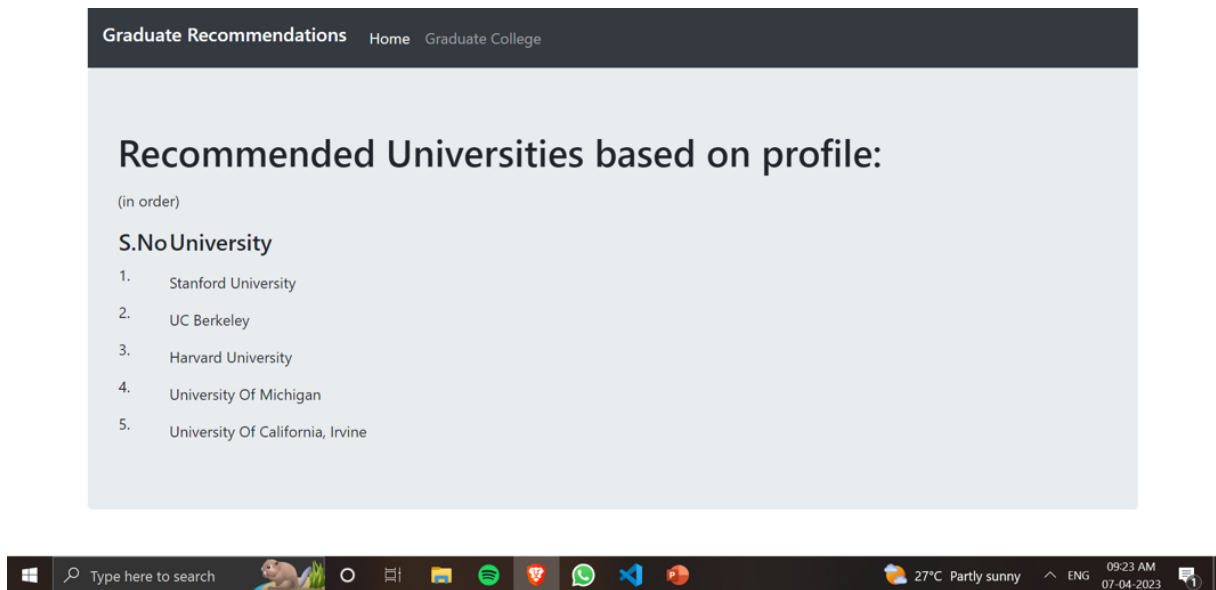


Figure 8: Results

## 4 Evaluation and Accuracy

As there is User interface to the application, any user can test the website any time. Using sklearn accuracy score function, we find the accuracy of the model as 92.45%. Planning to develop the feedback page where users can come back and provide feedback to the website and provide data in which University he/she got the admit. (if time permits).

```

# Defining X and y
X = data.drop('univName',axis=1)
y = data.univName
# Splitting data into train and test
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.25,random_state=42)
# Importing and fitting KNN classifier for k=3
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=3)
knn.fit(X_train,y_train)
# Predicting results using Test data set
pred = knn.predict(X_test)
from sklearn.metrics import accuracy_score
accuracy_score(pred,y_test)

```

[30] ✓ 0.4s

... 0.92455009

Figure 9: Accuracy

By capturing admission information we can train the algorithm to increase accuracy. We have also implemented selective analysis, and get an accuracy of 59.21% for the ANN model.

## **5 Conclusion**

In short, This project will help students in decision making of which University to choose for their higher education in other countries like USA. With the User friendly web Interface which takes in test scores, helps students apply to the Universities in which there is a high chance of getting the admit. Data Mining techniques like KNN and ANN are used.

## **6 Bibliography**

We would like to express our sincere gratitude to Archana Patil Ma'am, Assistant Professor at COEP Technological University, for giving us the opportunity to work on the mini project "University Recommendation System." Without her encouragement and guidance, this project would not have been possible.