*USA University Recommendation System based on Student profile for UG/graduate studies*

**Rama Tejaswini Thotapalli (SJSU Id: 013785681)**

Under the guidance ofProfessor Shih Yu Chang, San Jose State University

**Background & Objectives of Project:**

For an aspiring student who wants to apply for higher studies in other countries, university selection process is a challenging task as lot of different criteria’s need to considered during application process based on individual’s requirement.

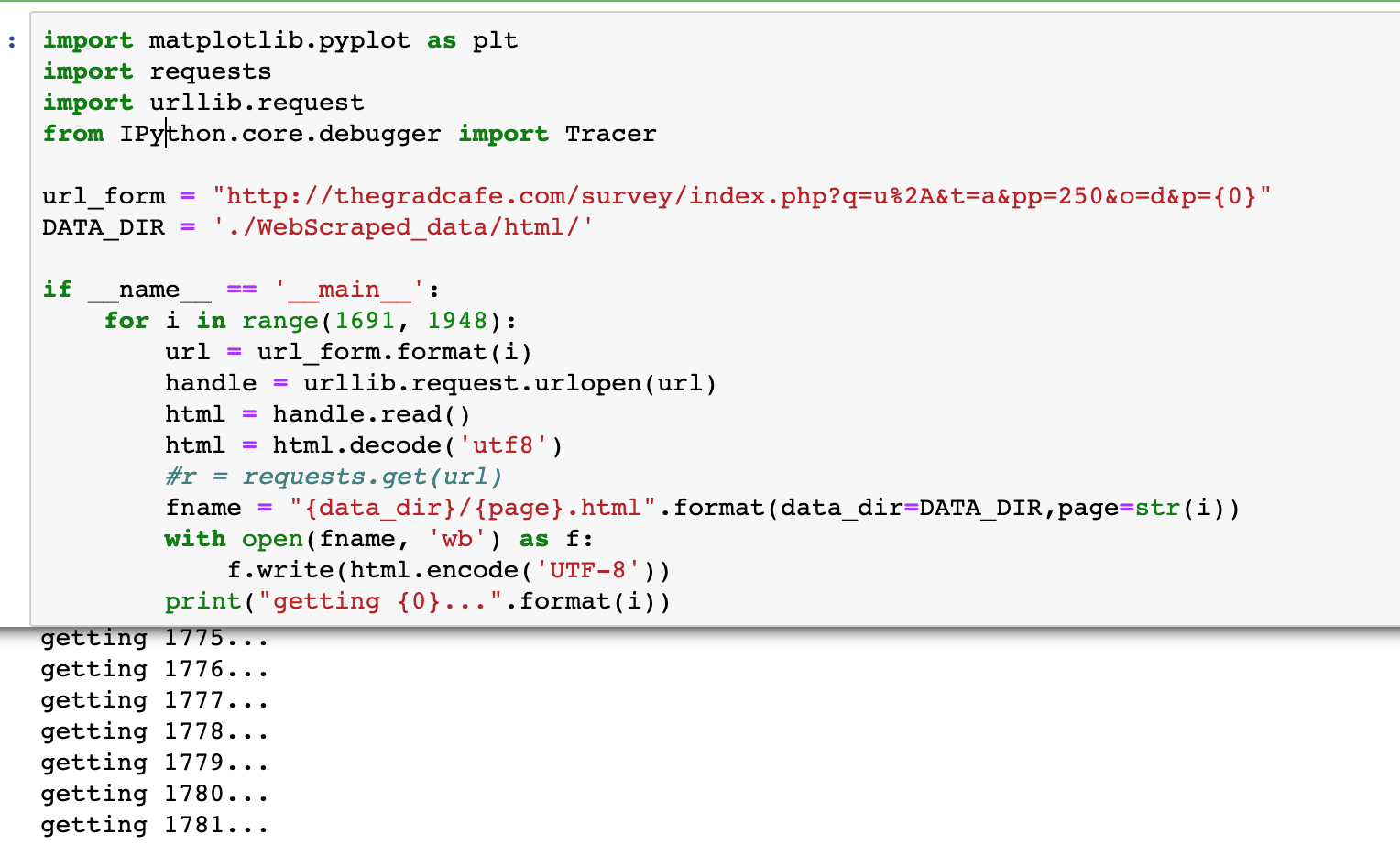
This problem can be addressed by modeling a recommender system based on various classification algorithms. In this project based on the student data set and the student profile who is looking for the admit, various models will be trained and a list of 10 best universities will be suggested such that it maximizes the chances of a student getting admit from that university list..

**Dataset:**

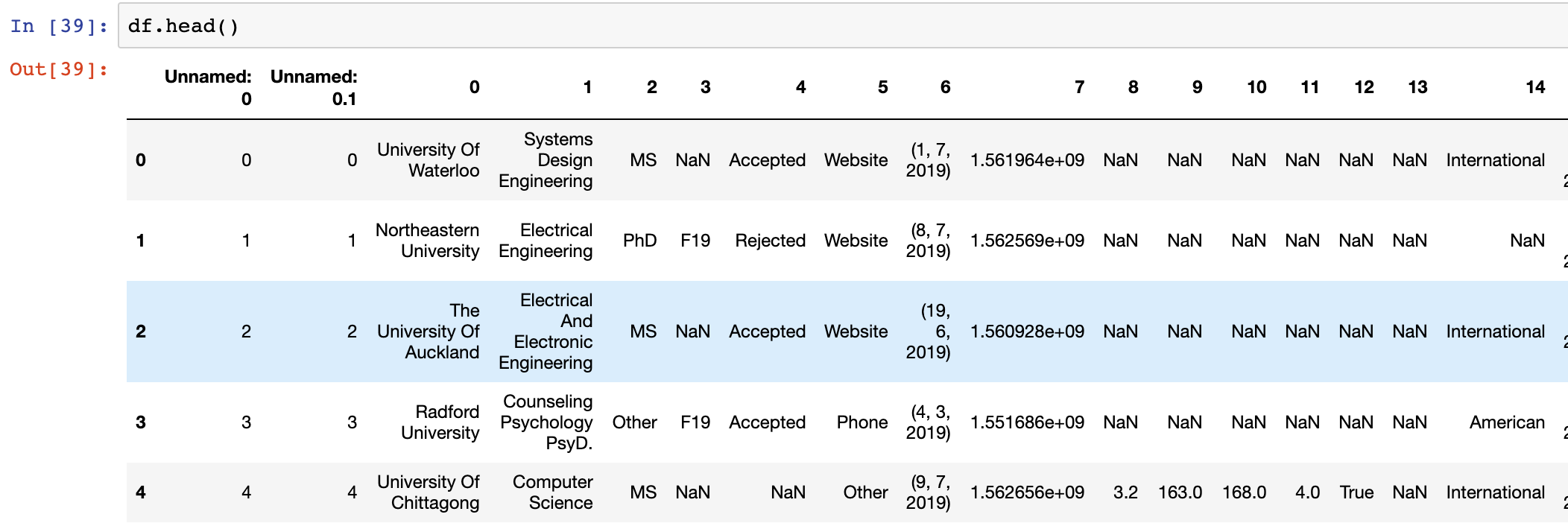
The first step in building any recommendation system is the identification of the data set. In order to build the classification model for the recommender system, this data has to be organized with appropriate labels. This core data for the application process is not readily available on the internet for direct consumption. However, this whole approach is based on making maximum use of the available information. The graduate student data was scraped from the following websites [www.thegradcafe.com](http://www.thegradcafe.com), and the Undergraduate university student data was scraped from <https://collegescorecard.ed.gov/data/>.

Graduate Student Dataset:

For Graduate Student data, we scraped [www.thegradcafe.com](http://www.thegradcafe.com) website. About 271807 rows of raw student data was obtained as a result of web scraping. Each sample corresponds to the profile of a student. We have got 1949 html pages of the data and need to change it into CSV files.



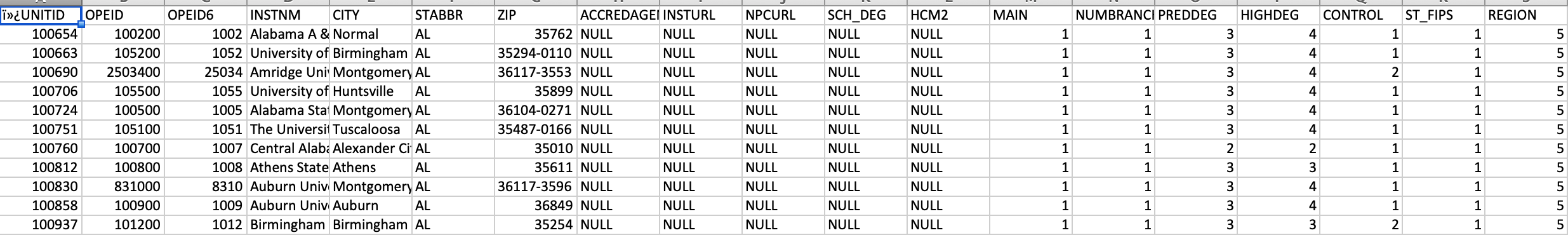
After scraping the final data will look like



The list of attributes are made as dataset for pre-process cleansing. For graduate students the dataset consists of University Name, Major, Degree, Season, Decision, Decision Method, Decision Date, Undergraduate GPA, Is New GRE Verbal, GRE Quant,   GRE Writing, Status,   Postdate Comments, Research Experience, Recommendations, Undergraduate GPA. For Under graduate students dataset consists of Student profile and SAT scores.

Under Graduate student dataset:

Under Graduate student data is taken from the College rank score card website <https://collegescorecard.ed.gov/data/>. The data before cleaning looked like below.



So to train and model this data, the data preprocessing is done.

Data Preprocessing:

In order to use the obtained data for our analysis, we need to do the preprocessing and cleansing, as there are lots of anomalies in the dataset. For this we use pandas and numpy frameworks.

Cleansing the data was done by

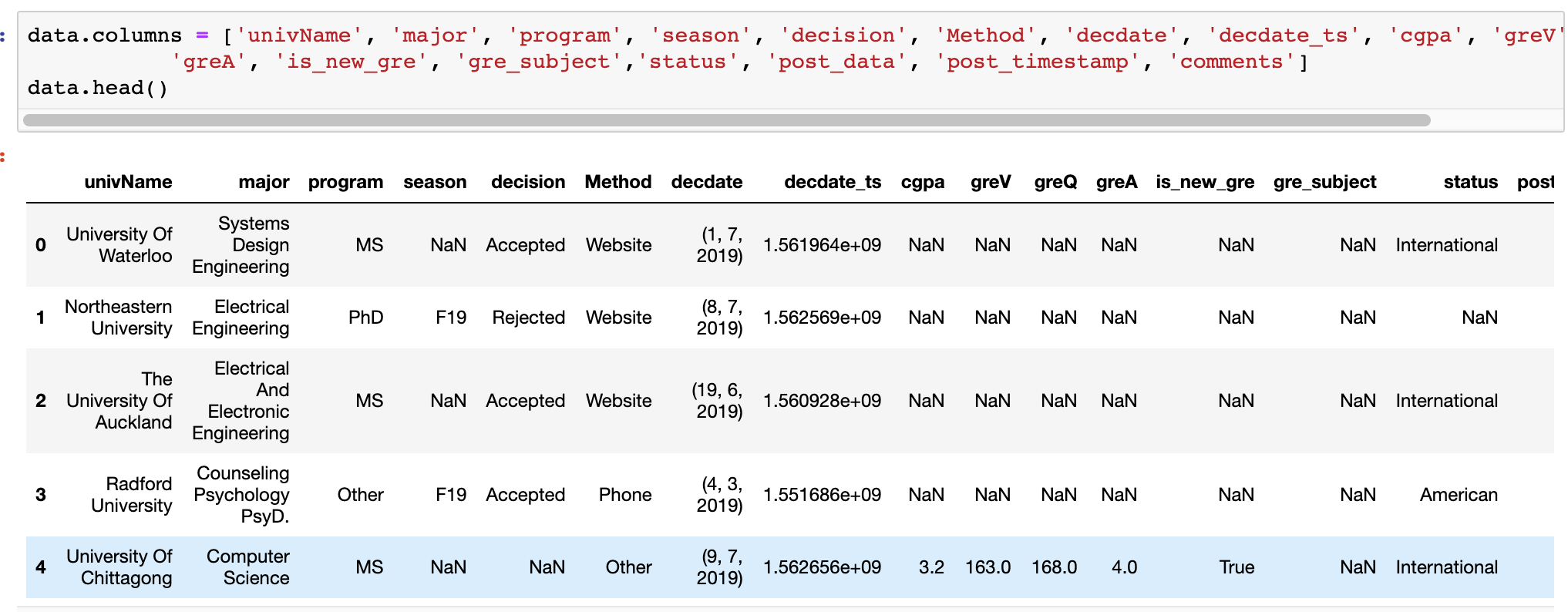
* Removing the irrelevant columns by using the drop column feature
* Filling the null values with the appropriate value or deleting the row containing null values.
* Removing the spaces in the data and reducing the size of the dataset.

In our graduate dataset, The GRE scores were also cleansed since they contained scores of both old and new versions of the examination. Similarly the 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.

*x* normalized = (*x* – *x* minimum) / (*x* maximum – *x* minimum)

Where x is the value of the GPA

The Processed data will look like below.



**Undergraduate Data:**

In Undergraduate data, we have taken below few rows of data like Institution name, city , Tution Fees, Sat Score, Admission rate, Debt and Men Ratio.

INSTNM', 'CITY', 'STABBR' ,'TUITIONFEE\_OUT', 'SAT\_AVG\_ALL', 'ADM\_RATE\_ALL', 'DEBT\_MDN\_SUPP', 'UGDS\_MEN'])

This data will be used for training the model and test data as SAT score and Maximum tution fees.

**Analysis and Methodology of project:**

Here I have used Knowledge based recommendation System where User inputs are taken into account and compare with the training data.

For Graduate University Recommendation I have used Case based knowledge recommendation as it will take the User inputs and compare with trained data.

For Undergraduate Recommendation System, I have used Constraint based Knowledge recommendation system where user inputs taken into account as constraints and based on the constraints we will compare with trained data.

I used two different models like K-Nearest Neighbors for Graduate data and Feature weighted algorithms for Undergraduate data.

# K Nearest Neighbor:

# In KNN, 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.

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

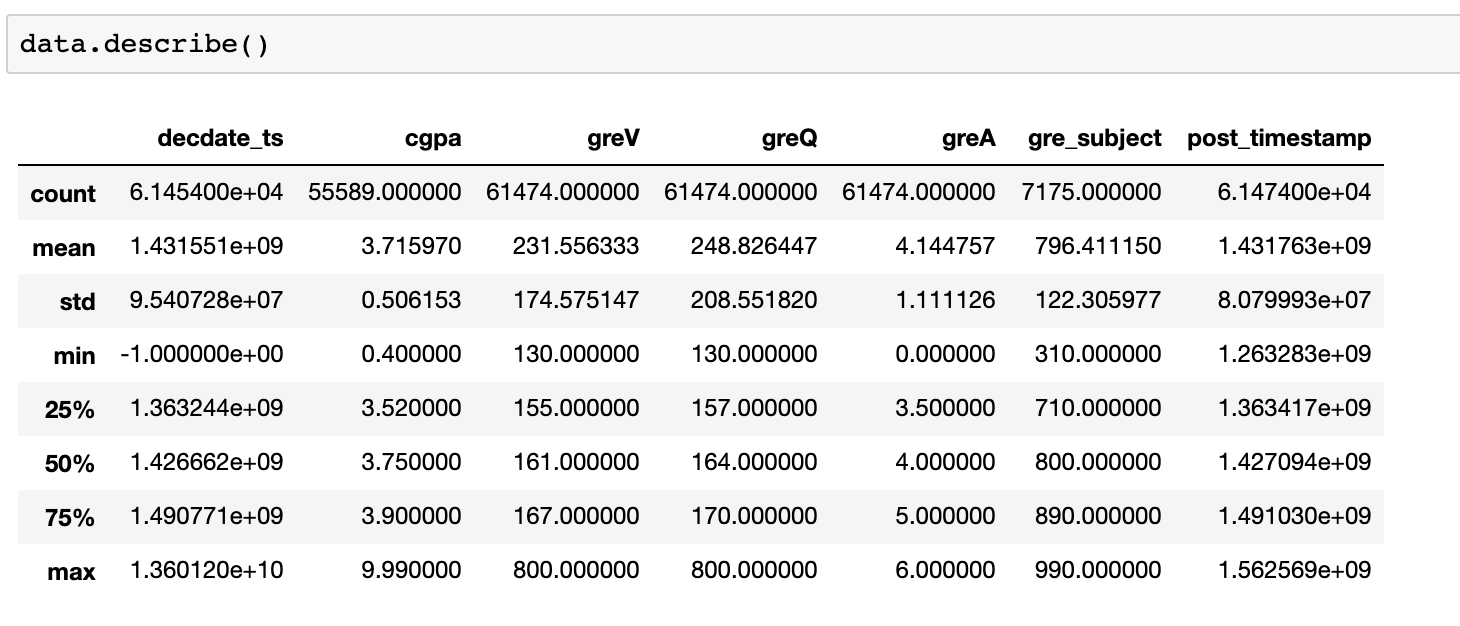
# Feature weighted algorithm:

# The weightage of all the features are taken and find the similarity score. Based on the similarity score, the universities with highest similarities will be recommended to student. Suppose w1, w2 are weights and f1 and f2 are features the similarity is calculated by formula Similarity

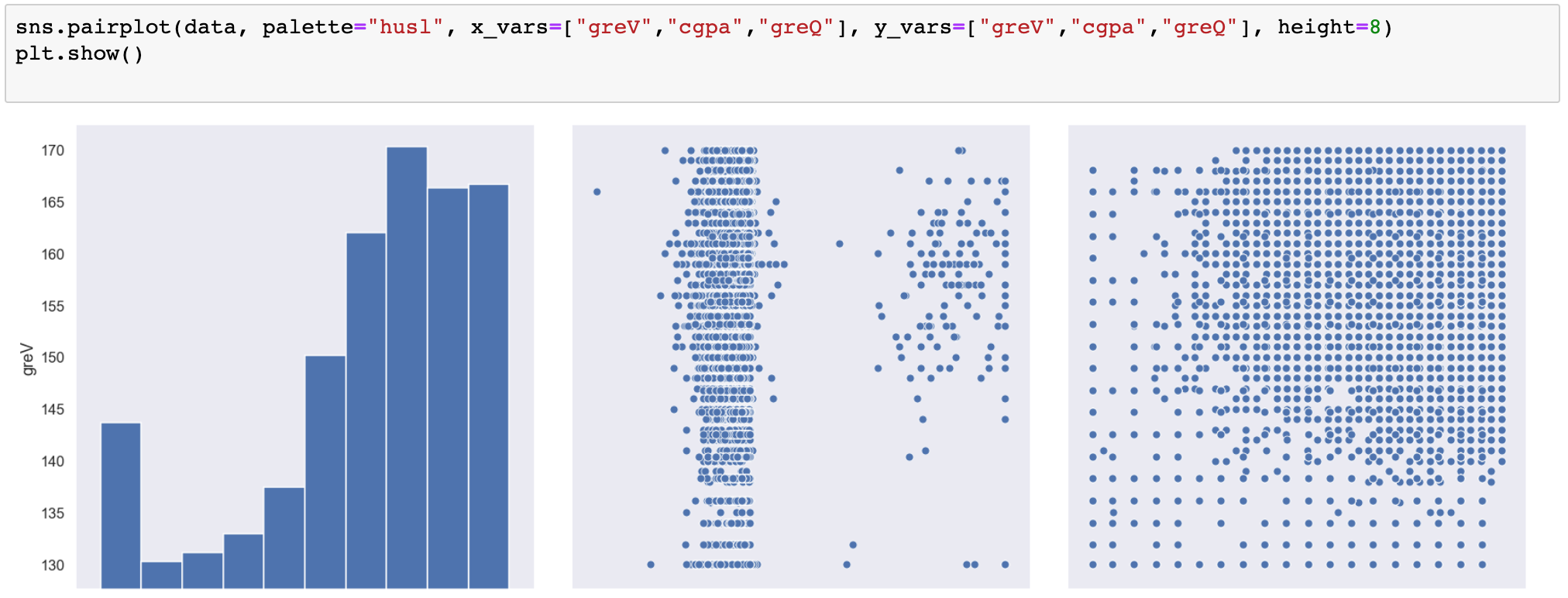
# score = w1\* f1+w2\*(1-f2)

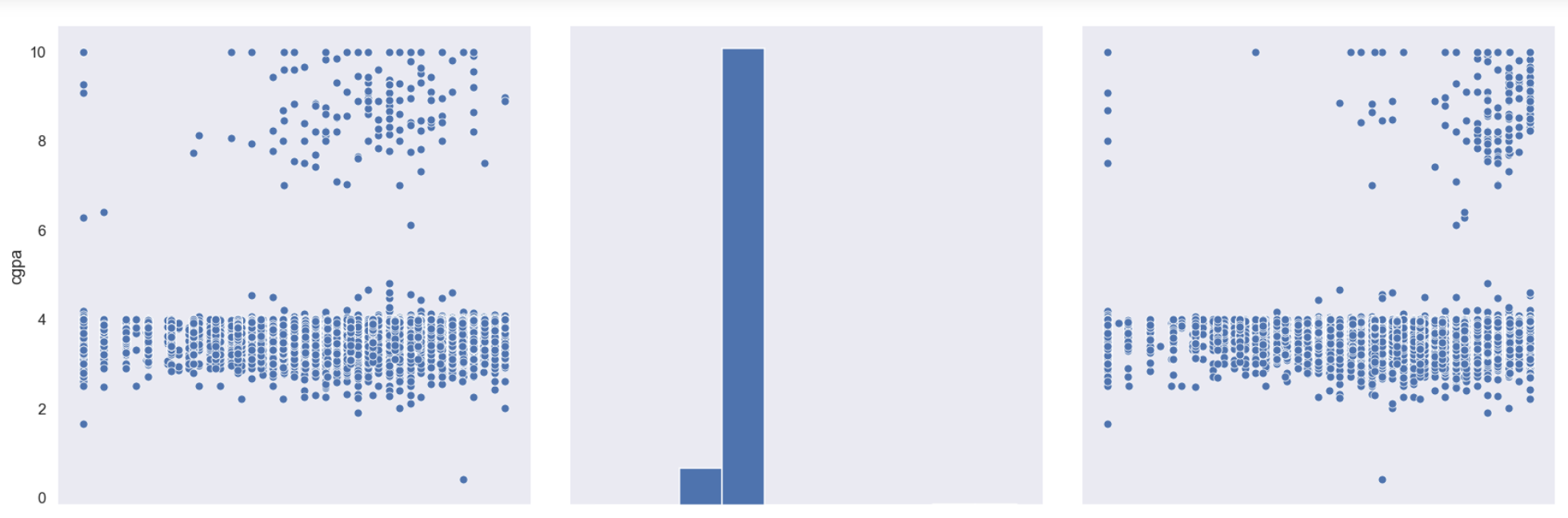
Results of the models:

Below are the results for the Data Preprocessing and modeling. Below is the description of data for the



Exploratory data analysis: pair plots of GRE verbal, GRE Quantitative and GPA.



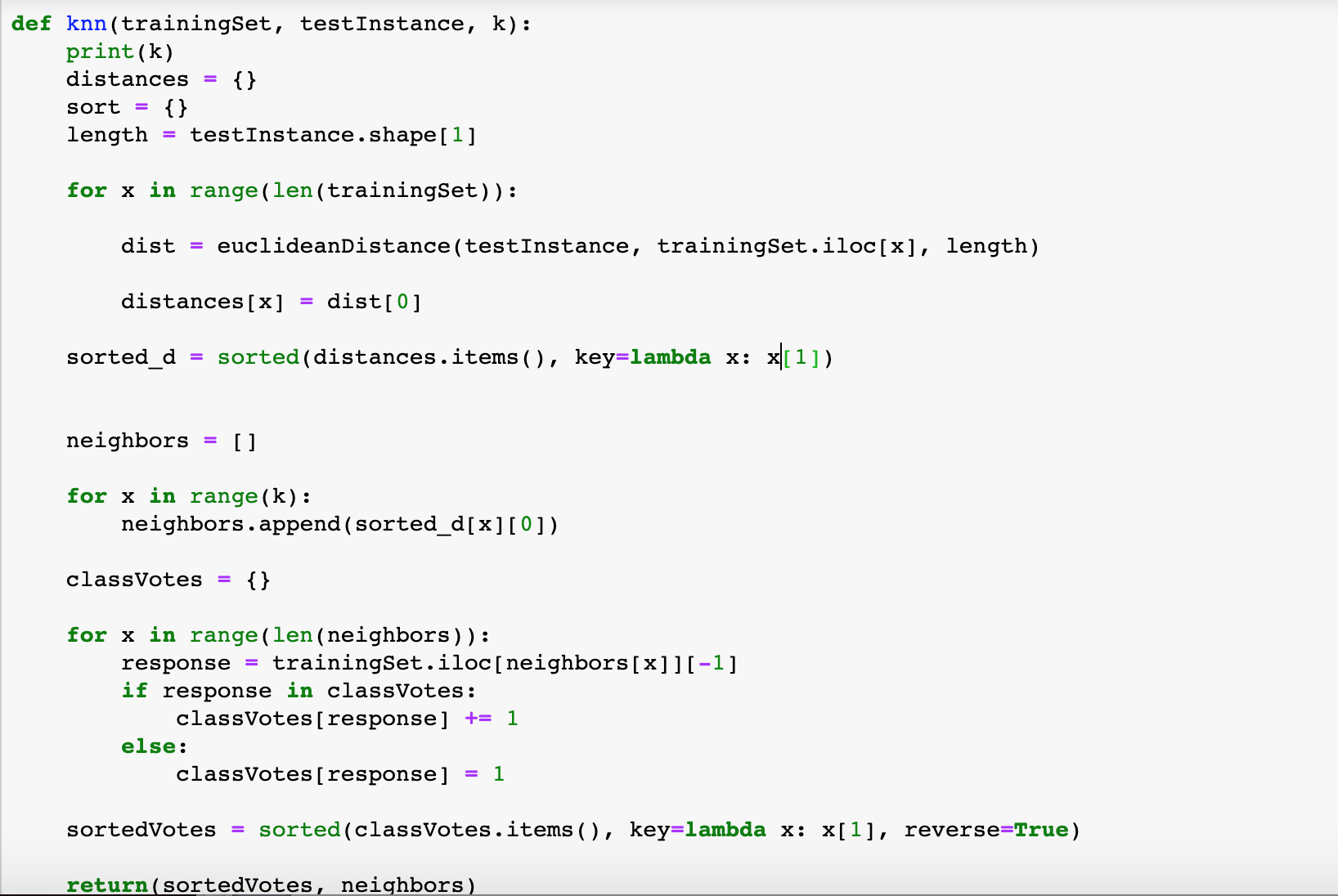


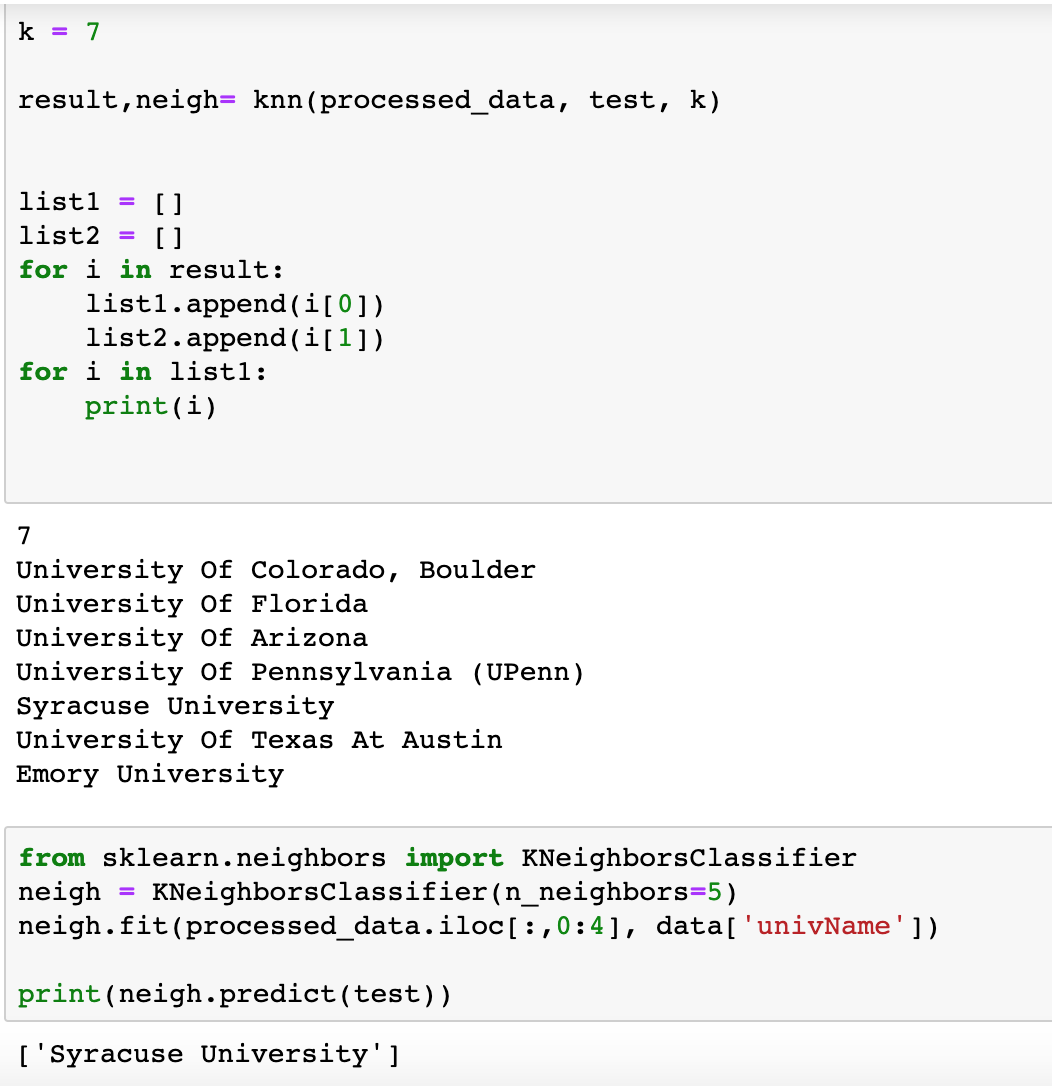


Trained data for Modelling:



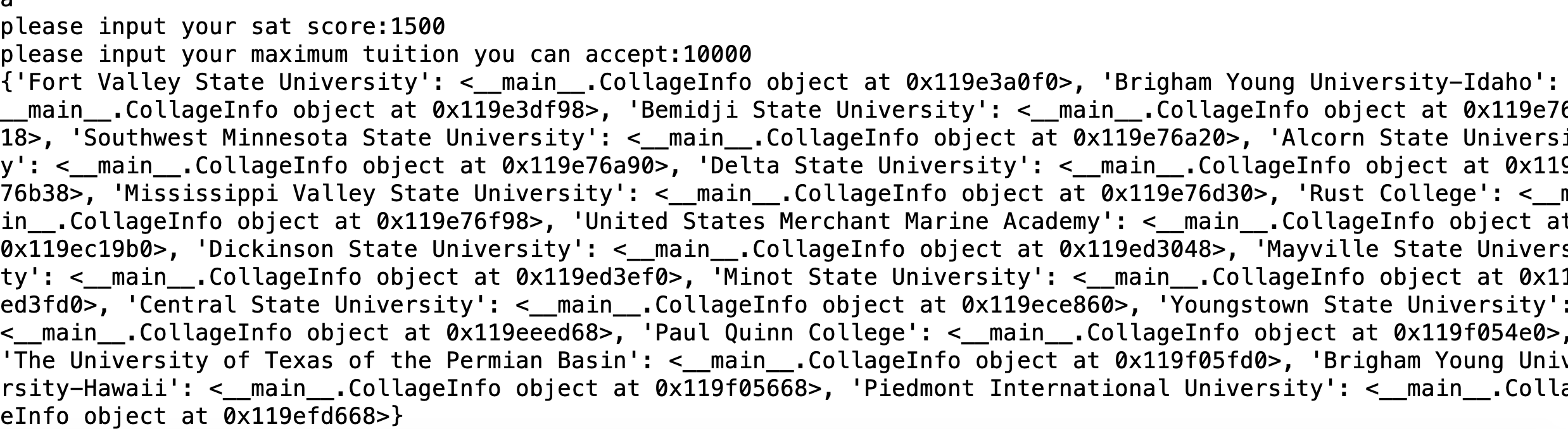
K Nearest Neighbour model:



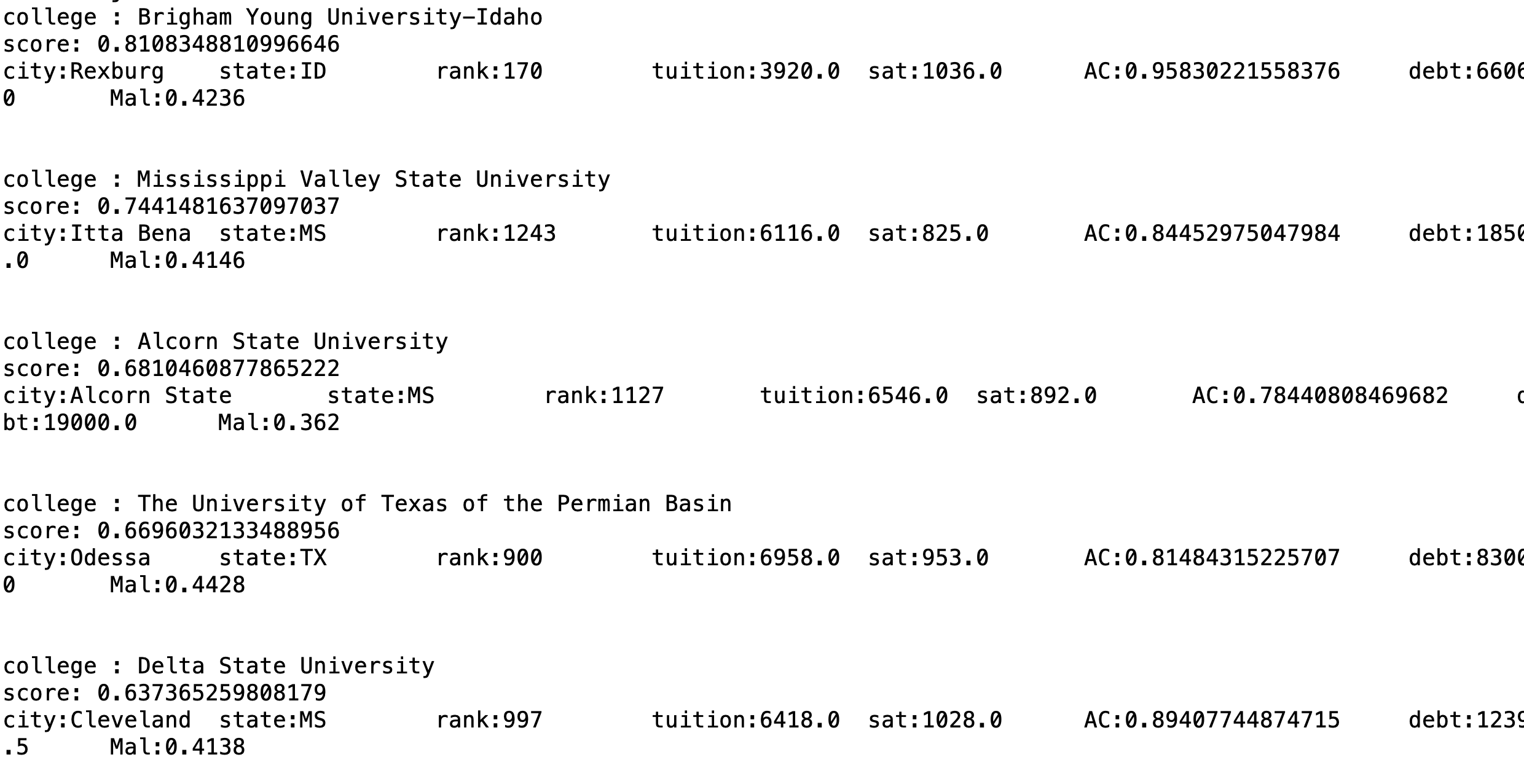


For Undergraduate Universities recommendation, weighted algorithm outputs:

Input of the data for SAT Score and Maximum Tution Fees:



Output:



Web Application:

I have developed a web application for University Recommendation System. Below are the Screenshots for this.



