

Smart Bridge-Remote Summer Internship Program

Project Title : University Admission Prediction

Project ID : SPS_PRO_194

Developed by: Apeksha B, Chinmaya G R, Kruthika C S, Madhumati J B, Veena M R

1. ABSTRACT

This paper describes GRADE, a statistical machine learning system developed to support the work of the graduate admissions committee at the University of Texas at Austin Department of Computer Science (UTCS). In recent years, the number of applications to the UTCS PhD program has become too large to manage with a traditional review process. GRADE uses historical admissions data to predict how likely the committee is to admit each new applicant. It reports each prediction as a score similar to those used by human reviewers, and accompanies each by an explanation of what applicant features most influenced its prediction. GRADE makes the review process more efficient by enabling reviewers to spend most of their time on applicants near the decision boundary and by focusing their attention on parts of each applicant's file that matter the most. An evaluation over two seasons of PhD admissions indicates that the system leads to dramatic time savings, reducing the total time spent on reviews by at least 74%.

2. INTRODUCTION

Graduate programs in fields such as computer science have received increasing interest in recent years. While the number of applicants to such programs has grown two- to threefold (Figure 1), the number of faculty available to review applications has remained constant or grown very slowly over time. The result is that admissions committees face a prohibitively large workload, making it difficult to review applications thoroughly. This paper describes a system developed to support the work of the graduate admissions committees in the Department of Computer Science at the University of Texas at Austin (UTCS). The system, named GRADE (for Graduate Admissions Evaluator), uses statistical machine learning to estimate the quality of new applicants based on past admissions decisions. GRADE does not determine who is admitted or rejected from the graduate program. Rather, its purpose is to inform the admissions committee and make the process of reviewing files more efficient. The heart of GRADE is a probabilistic classifier that predicts how likely the committee is to admit each applicant based on the information provided in his or her application file. For each new applicant, the system estimates this probability, expresses it as a numerical score similar to those used by human reviewers, and generates human-readable information explaining what factors most influenced its prediction.

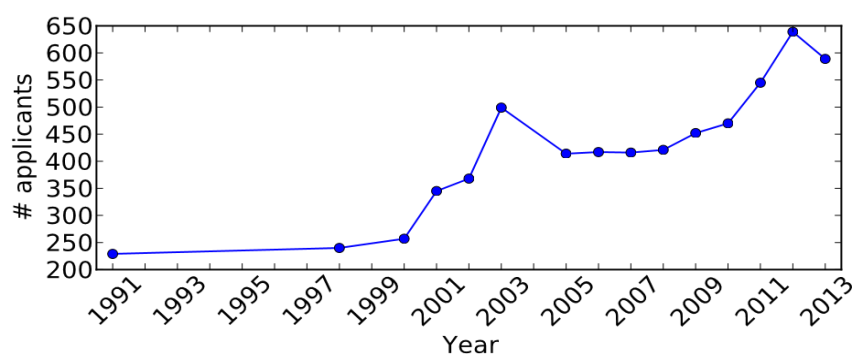


Figure 1: Number of applicants to the UTCS PhD program over time. Applicant pools have grown significantly in recent years, putting a strain on admissions committees, who have finite resources. (Data not available for some years.)

3. OVERVIEW

Students are often worried about their chances of admission in graduate school. The aim of this blog is to help students in shortlisting universities with their profiles. The predicted output gives them a fair idea about their admission chances in a particular university. This analysis should also help students who are currently preparing or will be preparing to get a better idea. 400 applicants have been surveyed as potential students for UCLA. The university weighs certain aspects of a student's education to determine their acceptance. The objective is to explore what kind of data is provided, determine the most important factors that contribute to a student's chance of admission, and select the most accurate model to predict the probability of admission.

To understand the impact of the prediction system, we first give a high-level overview of the graduate admissions process. UTCS, like many other departments, accepts applications for its graduate programs exclusively through an online system. Students fill out a series of forms with their educational history, test scores, research interests, and other information. Upon submitting the online application, a student's information is stored in a departmental database. References listed in the application are emailed and asked to submit letters of recommendation through the same online system. When the time window for accepting applications has closed, faculty members use an internal web-based system to review the pool of applicants. After reading each file, a reviewer submits a real-valued score in the range 0-5 to rate the quality of the applicant and enters a text comment explaining their score to other reviewers. The time required for each full review varies with the reviewer's style and experience, the quality and content of the application, and the stage of the review process, but a typical full review takes 10-30 minutes. The committee typically performs multiple review passes over the pool, and then admits or rejects each applicant based on the scores and comments of the reviewers who looked at his or her file. Although the primary criterion for this decision is quality, it is modulated to a significant degree by the current research opportunities in the department, i.e. how many new students the faculty request to be admitted in each research area.

In 2013, UTCS introduced a new, more efficient review process using GRADE to scale admissions to large applicant pools without sacrificing the quality of reviews. Instead of performing multiple full reviews on every file,

GRADE focuses the committee's attention to the files of borderline applicants, where it is needed most. The GRADE system and the new review process are described in detail in the following section.

3.1.Data Description

The dataset contains information about a student's:

- GRE Score
- TOEFL Score
- University Ratings
- Statement of Purpose Score
- Letter of Recommendation Score
- CGPA
- Whether the Student Has Done Any Research
- Chance of Admission (What We're Trying to Predict)

3.2.Software Design

- Jupyter Notebook Environment
- Spyder Ide
- Machine learning algorithms.
- Python (pandas, numpy, matplotlib,seaborn,sklearn)
- HTML
- Flask

3.3.Flow Chart

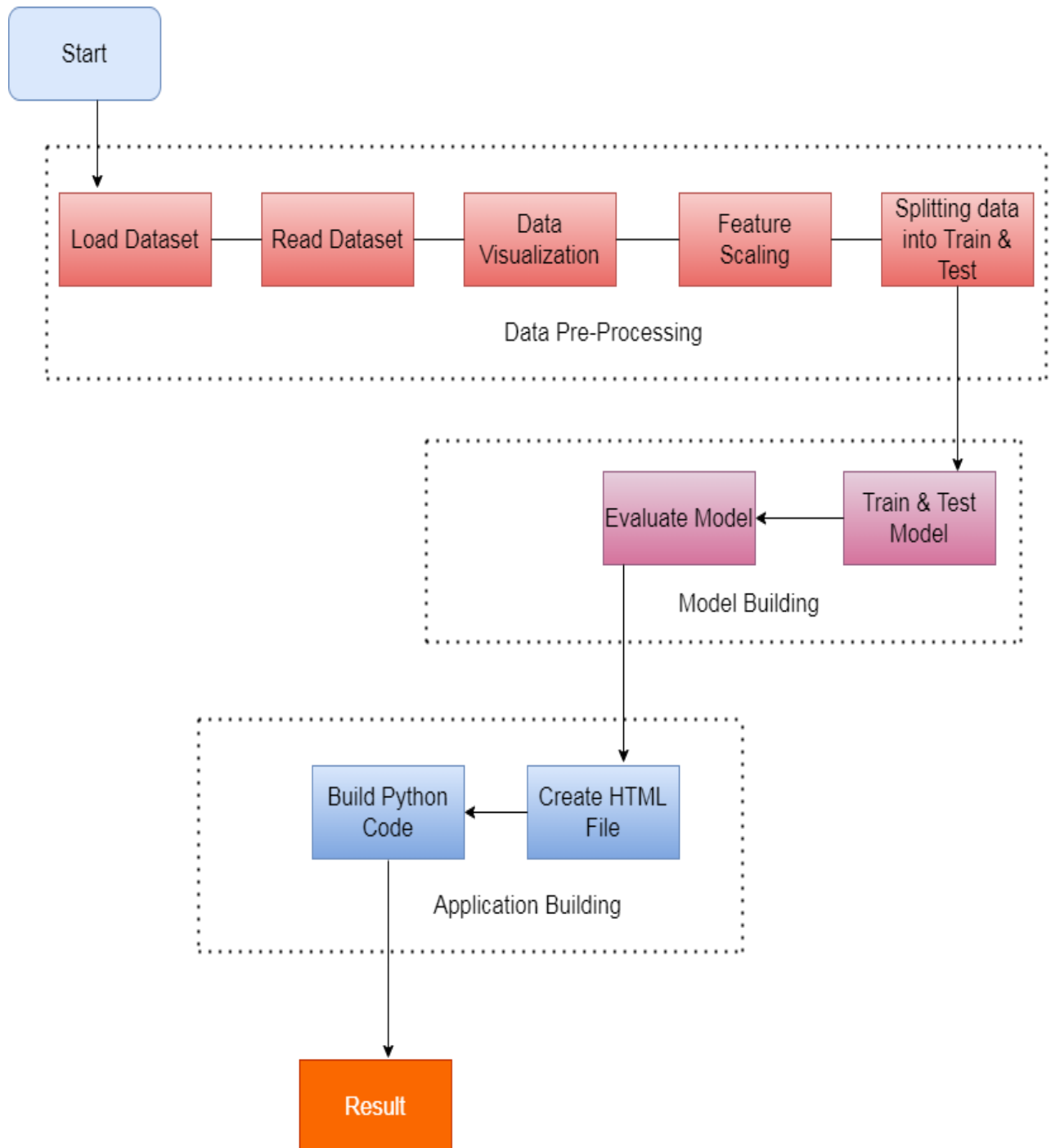


Fig 1.1.: Flow Chart

3.4.Importing Libraries and Dataset:

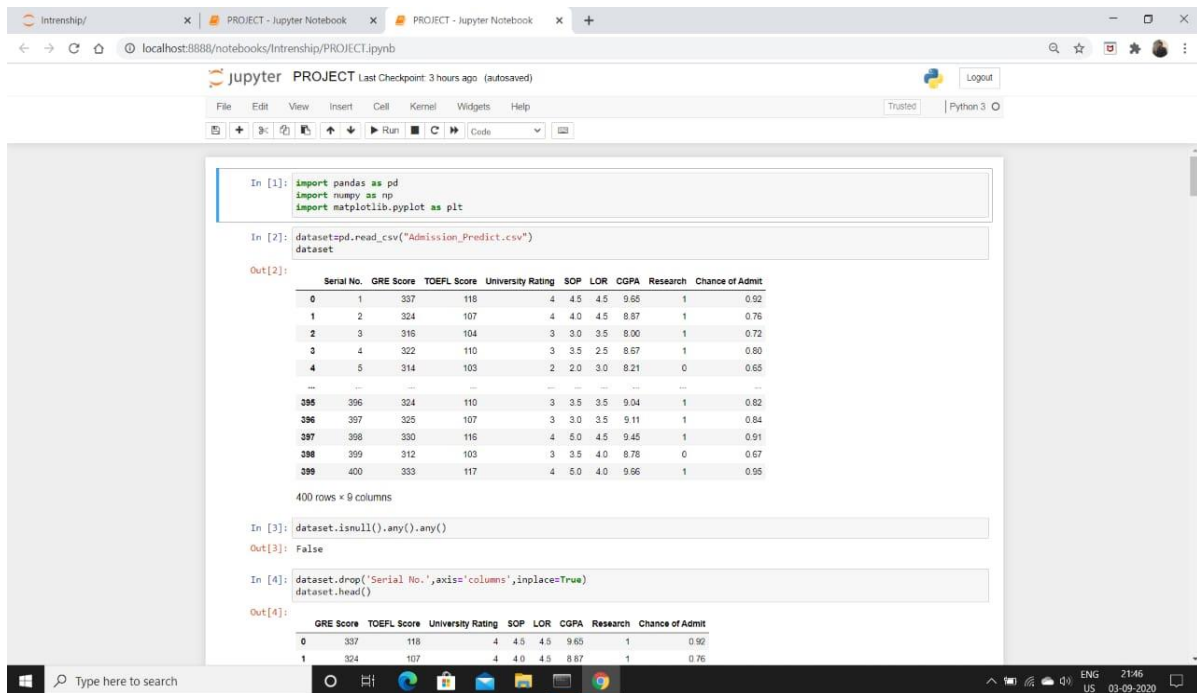


Fig 1.2 : Screenshot of Libraries and Dataset

3.5.Data Visualisation by Seaborn:

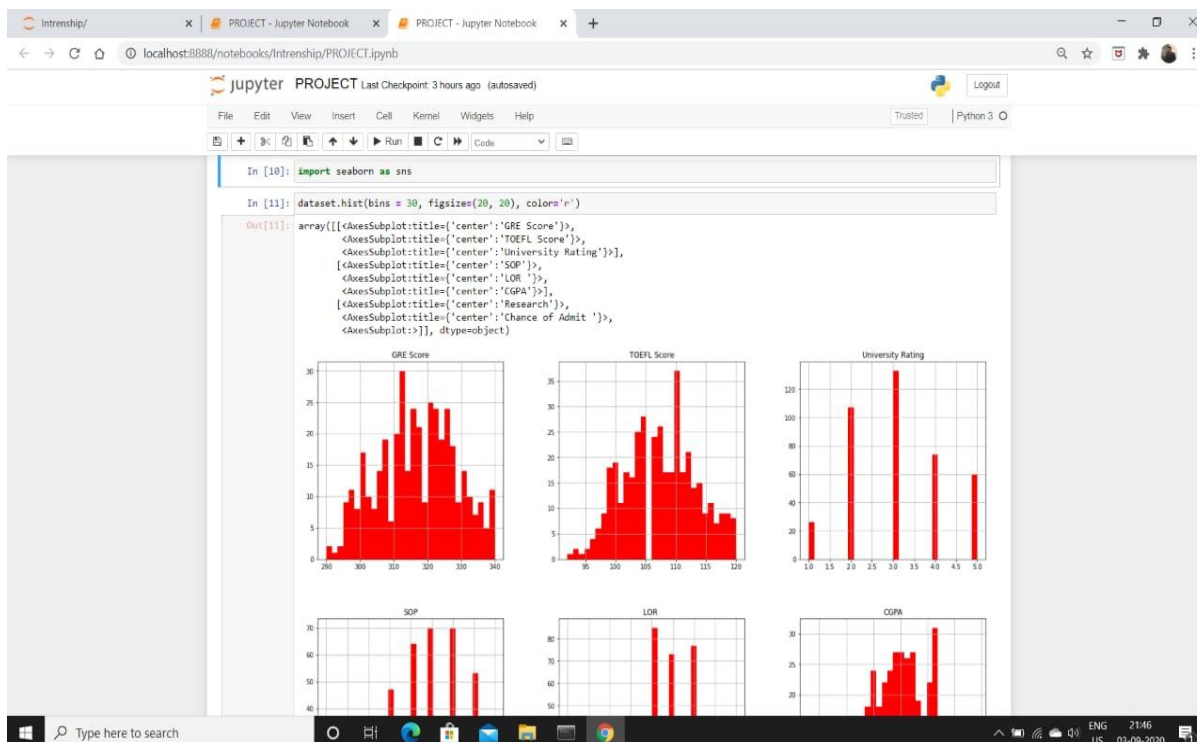


Fig 1.3.1. : Screenshot of Data Visualisation

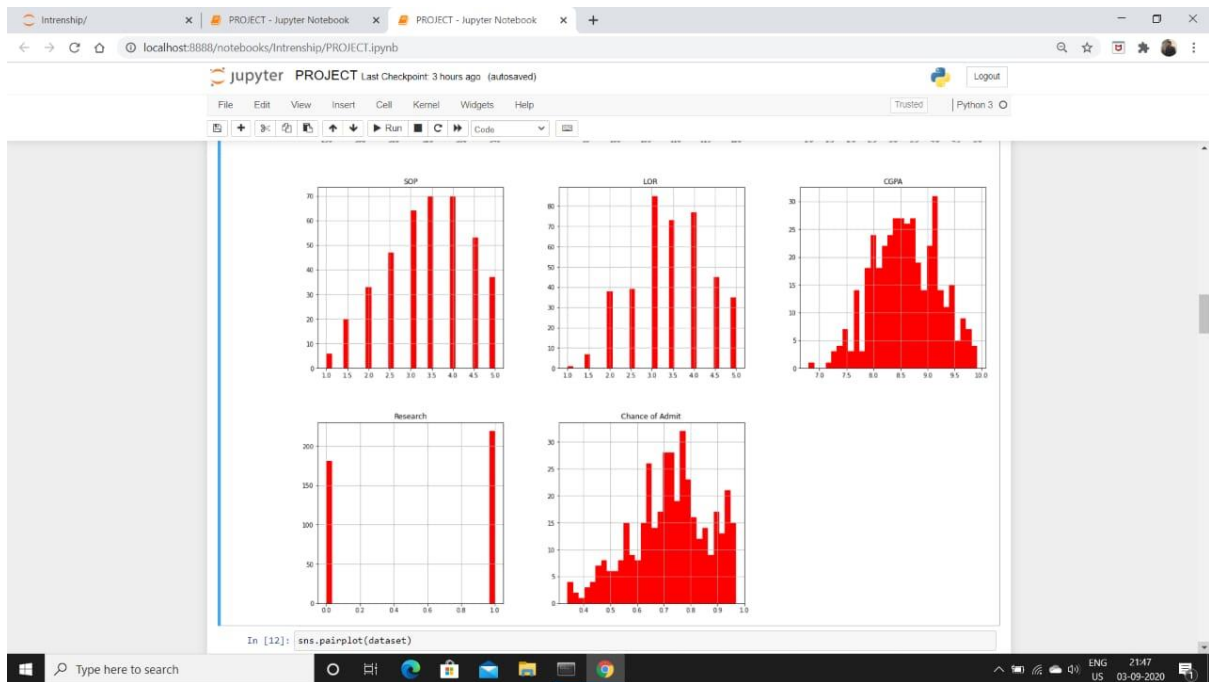


Fig 1.3.2. : Screenshot of Data Visualisation

3.6.Data visualisation by Pairplot

A **pairs plot** allows us to see both distribution of single variables and relationships between two variables. The pairs plot builds on two figures, the histogram and the scatter plot. The histogram allows us to see the distribution of a single variable while the scatter plots shows the relationship between two variables.

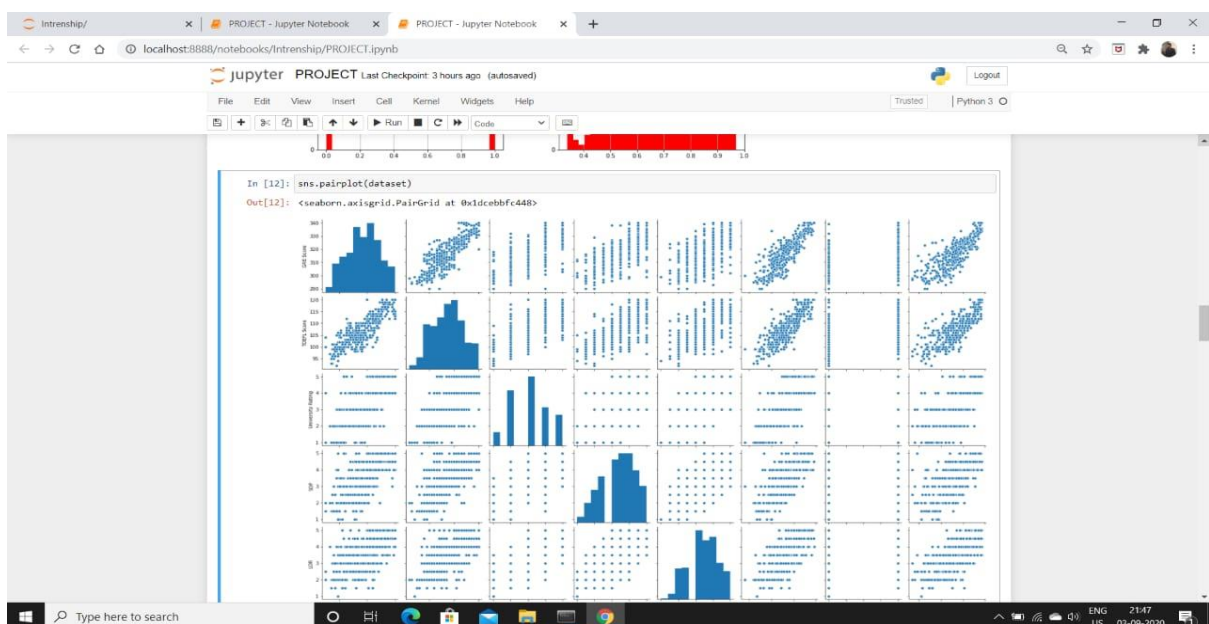


Fig 1.4.1. : Screenshot of Data Visualisationby Pairplot

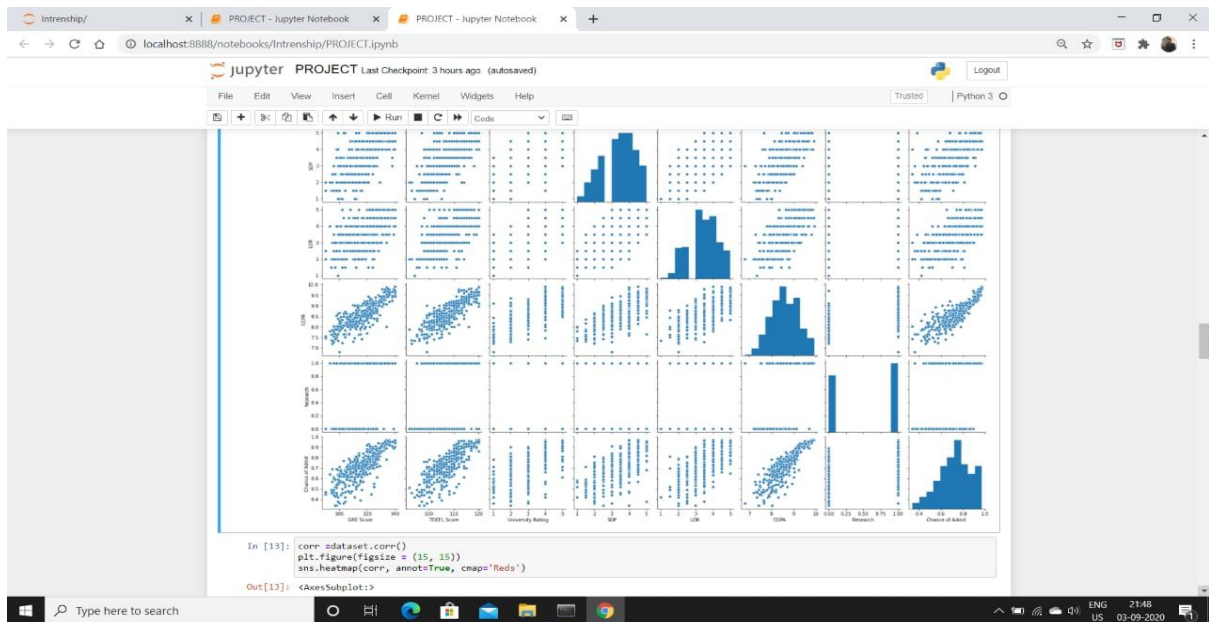


Fig 1.4.2. : Screenshot of Data Visualisationby Pairplot

3.7.Heatmap:

The **heatmap** is a way of representing the data in a 2-dimensional form. The data values are represented as colors in the graph. The goal of the heatmap is to provide a colored visual summary of information.

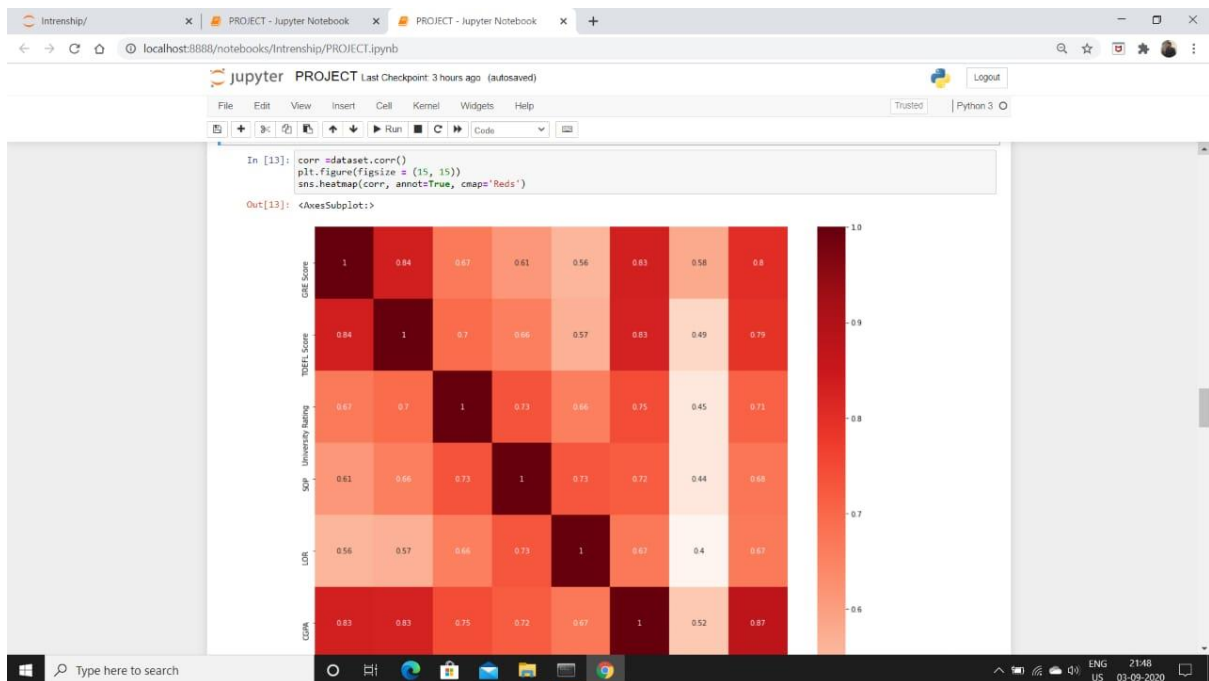


Fig 1.5.1. : Screenshot of Heatmap

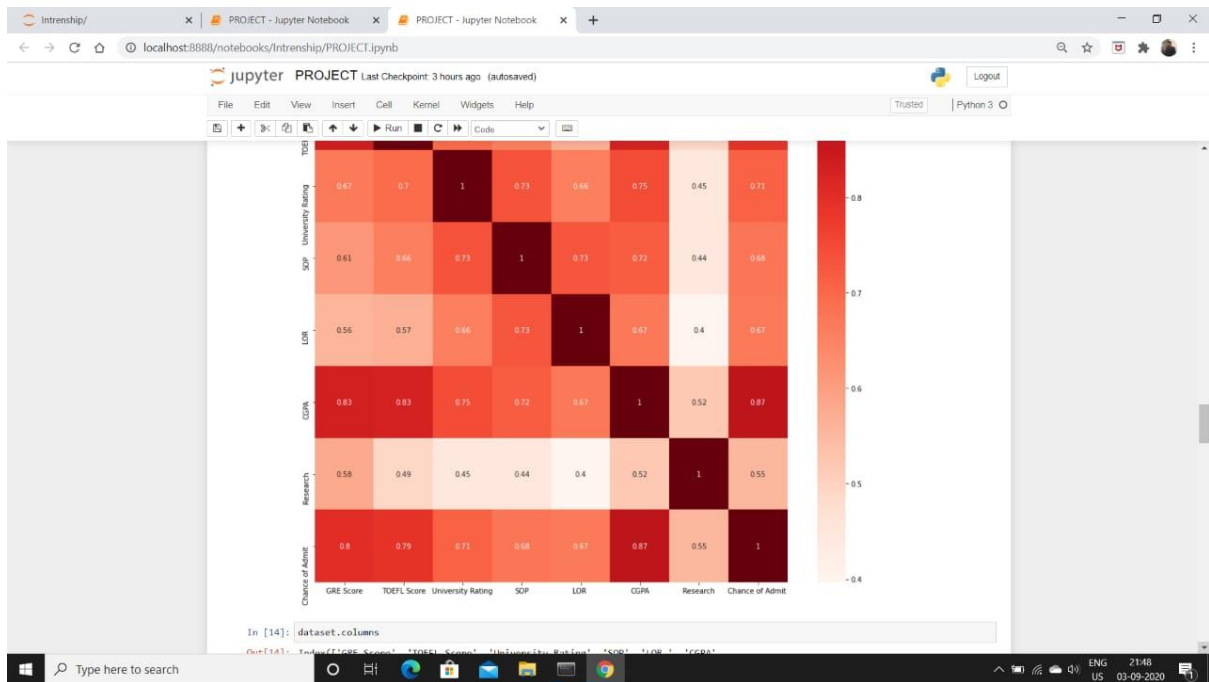


Fig 1.5.2. : Screenshot of Heatmap

The top three features that affect the Chance to Admit are:

1. CGPA
2. GRE Score
3. TOEFL Score

CGPA

The Cumulative Grade Point Average is a 10 point grading system. From the data shown below, it appears the submissions are normally distributed. With a mean of 8.6 and standard deviation of 0.6.

GRE Score

The Graduate Record Examination is a standardized exam, often required for admission to graduate and MBA programs globally. It's made up of three components:

1. Analytical Writing (Scored on a 0-6 scale in half-point increments)
2. Verbal Reasoning (Scored on a 130-170 scale)
3. Quantitative Reasoning (Scored on a 130-170 scale)

TOEFL Score

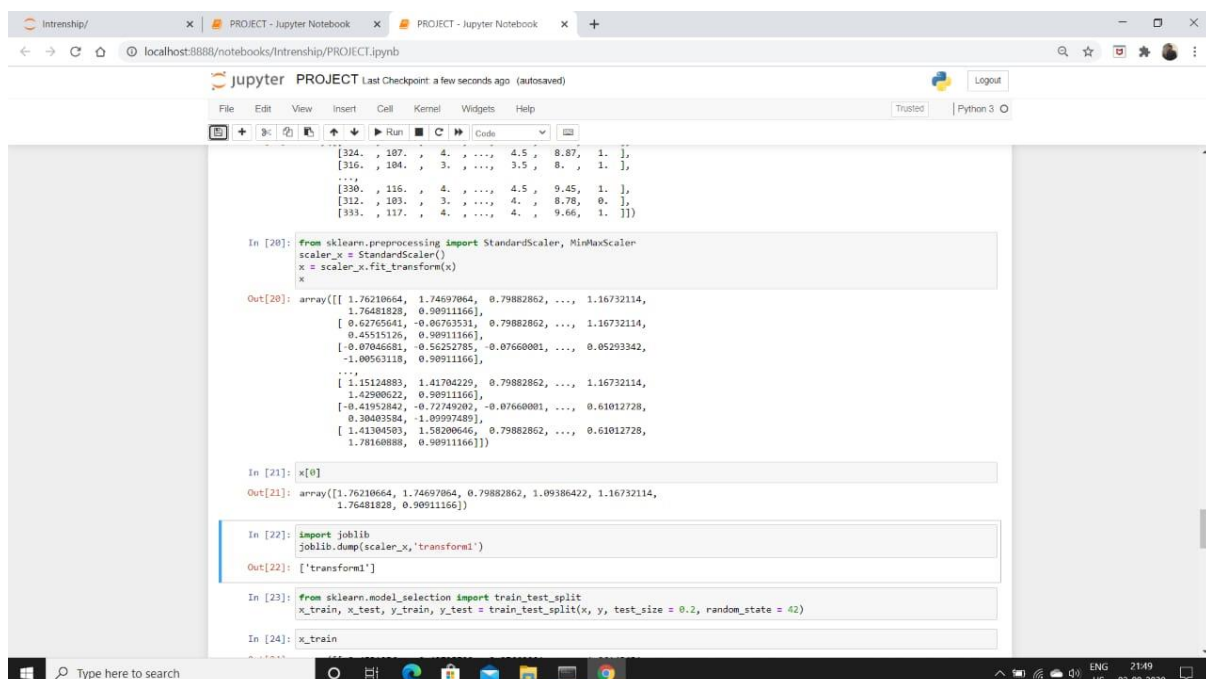
The Test of English as a Foreign Language is a standardized test for non-native English speakers that are choosing to enroll in English-speaking universities.

The test is split up into 4 sections:

1. Reading
2. Listening
3. Speaking
4. Writing

All sections are scored out of 30, giving the exam a total score of 120 marks. In this dataset, the TOEFL scores have a mean of 107 and a standard deviation of 6.

3.8.Feature scaling:



```
Intership/ PROJECT - Jupyter Notebook PROJECT - Jupyter Notebook +
localhost:8888/notebooks/Intership/PROJECT.ipynb
jupyter PROJECT Last Checkpoint: a few seconds ago (autosaved)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3
In [20]: from sklearn.preprocessing import StandardScaler, MinMaxScaler
scaler_x = StandardScaler()
x = scaler_x.fit_transform(x)
Out[20]: array([[ 1.76210664,  1.74697064,  0.79882862, ...,  1.16732114,
 1.76481828,  0.90911166],
 [ 0.62765641, -0.06763531,  0.79882862, ...,  1.16732114,
 0.45515126,  0.90911166],
 [-0.07046681, -0.56252785, -0.07660001, ...,  0.05293342,
-1.00563118,  0.90911166],
 ...,
 [ 1.15124883,  1.41704229,  0.79882862, ...,  1.16732114,
 1.42900622,  0.90911166],
 [-0.41952042, -0.7749208, -0.07660001, ...,  0.61012728,
 0.30403584, -1.0997489],
 [ 1.41304503,  1.58200646,  0.79882862, ...,  0.61012728,
 1.78160888,  0.90911166]])

In [21]: x[0]
Out[21]: array([ 1.76210664,  1.74697064,  0.79882862,  1.09386422,  1.16732114,
 1.76481828,  0.90911166])

In [22]: import joblib
joblib.dump(scaler_x, 'transform1')
Out[22]: ['transform1']

In [23]: from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2, random_state = 42)

In [24]: x_train
```

Fig 1.6. : Screenshot of Feature scaling

3.9. Training and Testing variables:

Fig 1.7. : Screenshot of Training and Testing variables

3.10. Building model:

Fig 1.8. : Screenshot of Building model

3.11. Evaluation of model:

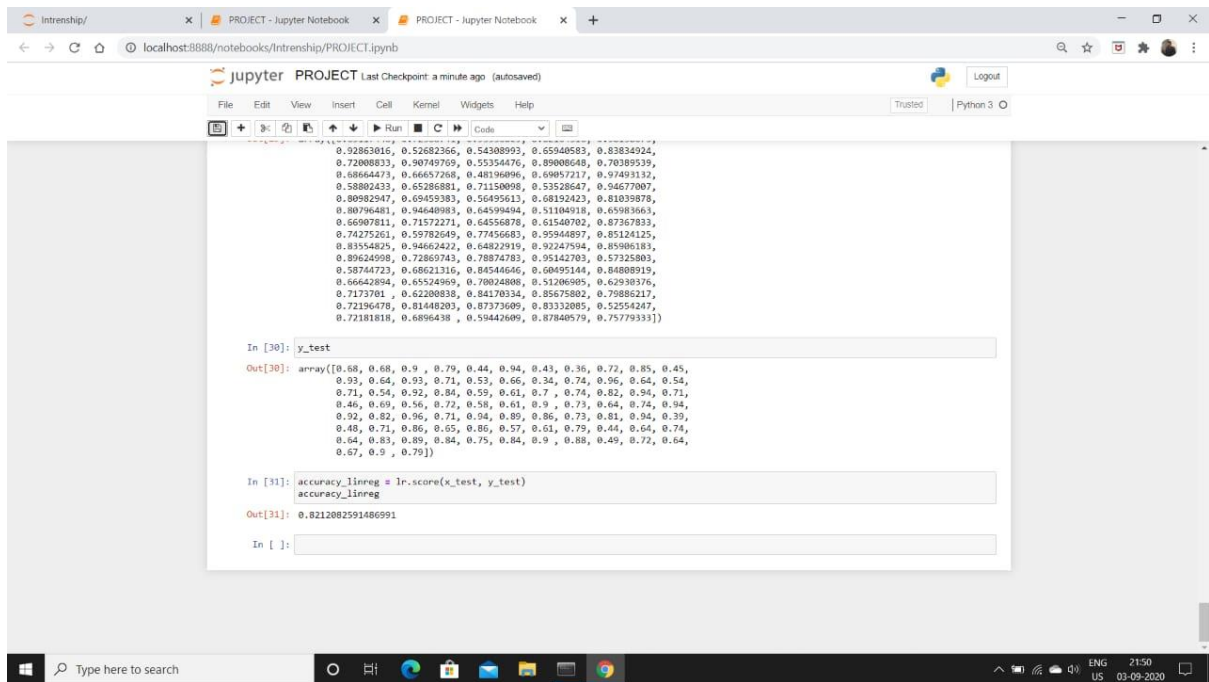


Fig 1.9. : Screenshot of Evaluation of model

3.12. HTML File in spyder:

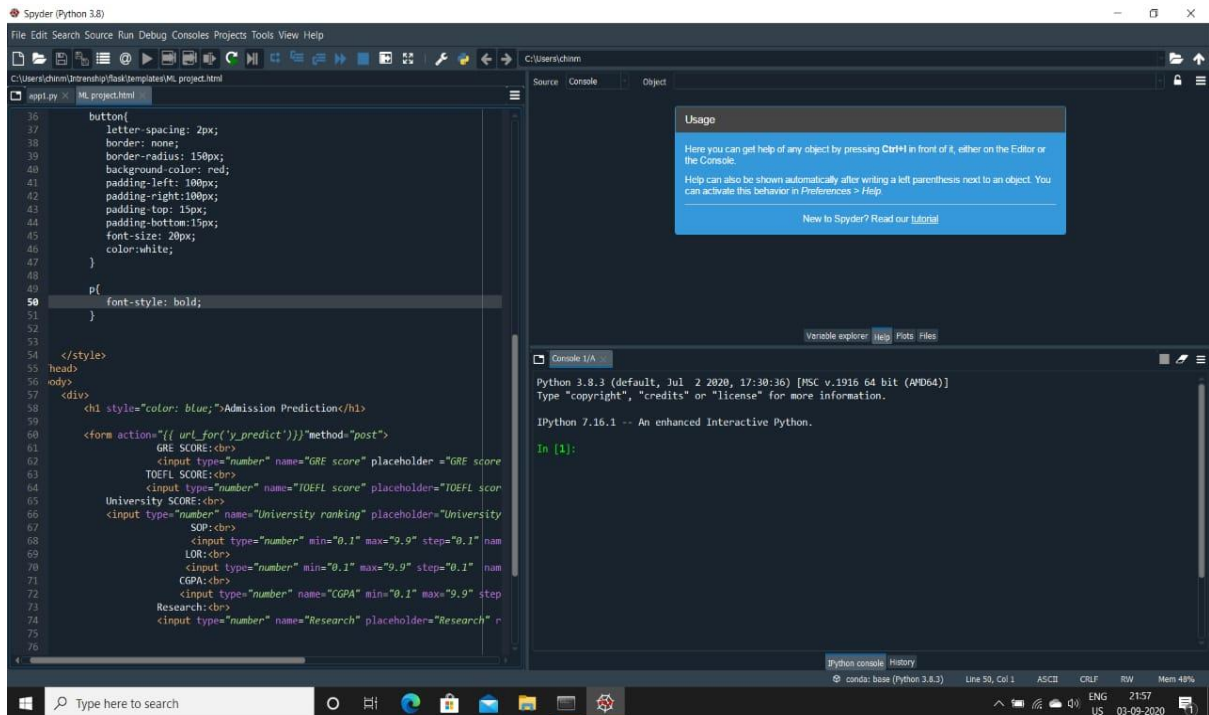
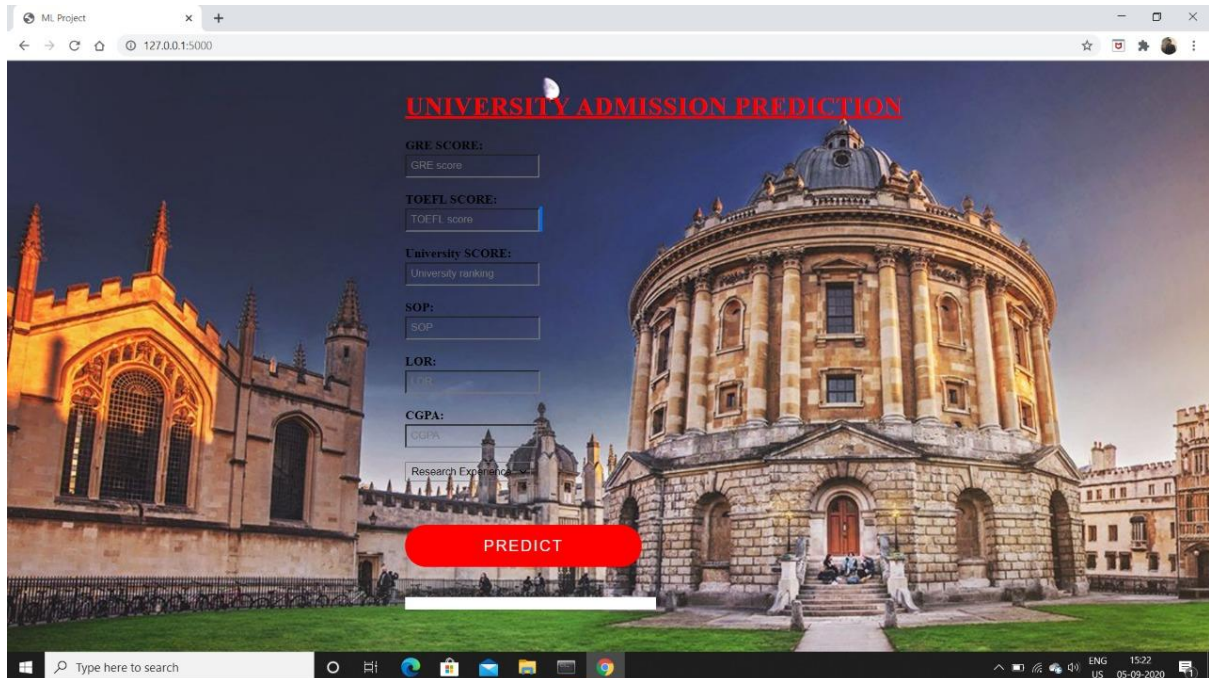


Fig 1.10 : Screenshot of HTML File

4. OUTPUT:

4.1.Web page before entering input:



The screenshot shows a web browser window with the title "ML Project" and the address bar displaying "127.0.0.1:5000". The page features a background image of a large, historic building with a prominent dome. The title "UNIVERSITY ADMISSION PREDICTION" is displayed in red text at the top. Below the title, there are several input fields for user data: GRE SCORE, TOEFL SCORE, University SCORE, SOP, LOR, CGPA, and Research Experience. A red "PREDICT" button is located at the bottom of the form.

UNIVERSITY ADMISSION PREDICTION

GRE SCORE:
GRE score

TOEFL SCORE:
TOEFL score

University SCORE:
University ranking

SOP:
SOP

LOR:
LOR

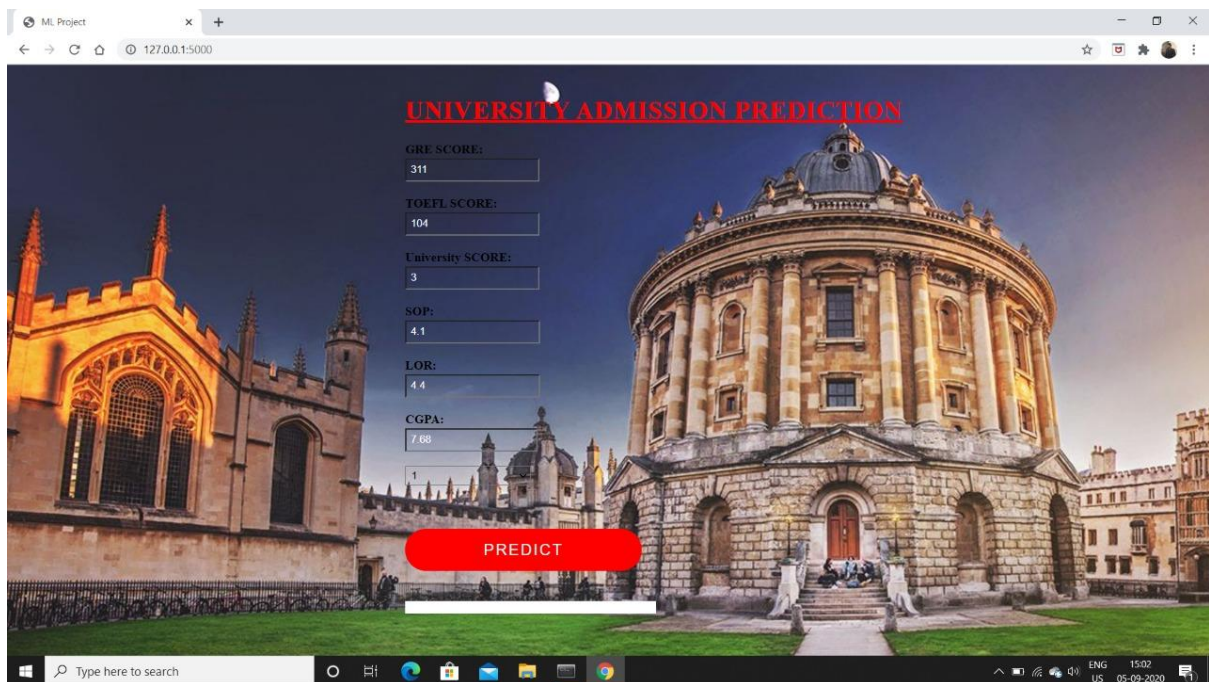
CGPA:
CGPA

Research Experience:
Research Experience

PREDICT

Fig 2.1. : Screenshot of output

4.2.Web page after entering input:



The screenshot shows the same web browser window as in Fig 2.1, but with the input fields filled with numerical values. The "PREDICT" button remains red and is still visible at the bottom of the form.

UNIVERSITY ADMISSION PREDICTION

GRE SCORE:
311

TOEFL SCORE:
104

University SCORE:
3

SOP:
4.1

LOR:
4.4

CGPA:
7.68

Research Experience:
1

PREDICT

Fig 2.2. : Screenshot of output

4.3.Output:

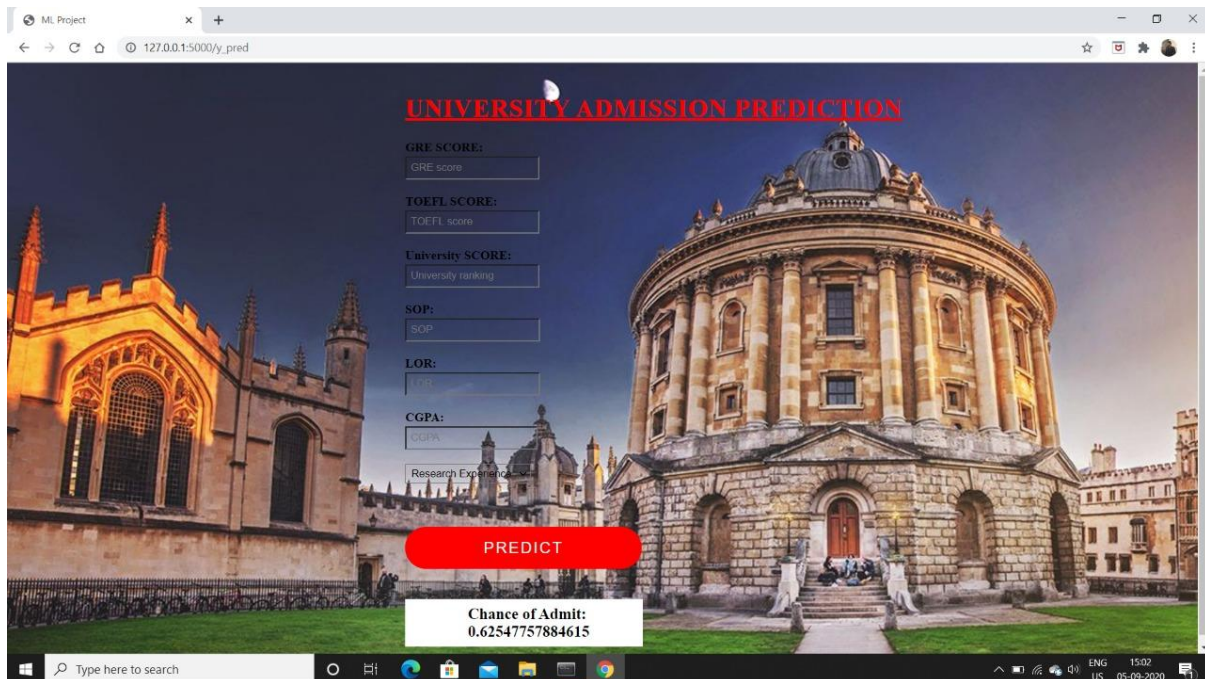


Fig 2.3 : Screenshot of output

5. ADVANTAGES AND DISADVANTAGES OF PROPOSED SYSTEM

5.1.Advantages

- It helps student for making decision for choosing a right college. Here the chance of occurrence of error is less when compared with the existing system.
- It is fast, efficient and reliable.
- Avoids data redundancy and inconsistency.
- Very user-friendly.
- Easy accessibility of data.

5.2. Disadvantages

- Required active internet connection.
- System will provide inaccurate results if data entered incorrectly.

6. CONCLUSION

This paper describes GRADE, a system that uses statistical machine learning to scale graduate admissions to large applicant pools where a traditional review process would be infeasible. GRADE allows reviewers to identify very high and low-scoring applicants quickly and reduces the amount of time required to review each application. While all admission decisions are ultimately made by human reviewers, GRADE reduces the total time spent reviewing files by at least 74% compared to a traditional review process, and makes it possible to complete reviews with limited resources without sacrificing quality.

7. BIBLIOGRAPHY

1. <https://towardsdatascience.com/why-random-forest-is-my-favorite-machine-learning-model-b97651fa3706>
2. <https://towardsdatascience.com/predicting-ms-admission-afb9c5c599>
3. <https://www.mdpi.com/2306-5729/4/2/65/html>