Machine Learning Project Mid-term Report

1. Who is your teammate or you are working alone?

I'm working alone. LIN JIA u1091732

2. A brief introduction to the problem you want to solve using machine learning techniques.

This dataset is created by Mohan S Acharya [1] to estimate chances of graduate admission from an Indian perspective. This dataset is inspired by the UCLA Graduate Dataset.

The dataset contains several parameters which are considered important during the application for masters programs. The parameters included are : GRE Scores (out of 340); TOEFL Scores (out of 120); University Rating (out of 5); Statement of Purpose and Letter of Recommendation Strength (out of 5); Undergraduate GPA (out of 10); Research Experience (either 0 or 1); Chance of Admit (ranging from 0 to 1) [2]. In a such fair size of dataset, about several different important categories/parameters are included. I will solve the what's the key categories/parameters and what's the least, as well, how these categories/parameters interrelated among themselves. By controlling the variables, to help predictions more efficient and meaningful.

3. The motivation - why do you want to use machine learning techniques? Why not the traditional or existing methods?

Machine learning techniques I will use: decision tree, random forest, linear regression and etc. to predict the chance of admit with high accuracy by applying above algorithms and then comparing their scores.

A traditional algorithm takes some input and some logic in the form of code and drums up the output. As opposed to this, a Machine Learning Algorithm takes an input and an output and gives the some logic which can then be used to work with new input to give one an output. The logic generated is what makes it machine learning [3].

4. What you have done to reach your goal. Note that just "We collected data" will NOT be enough.

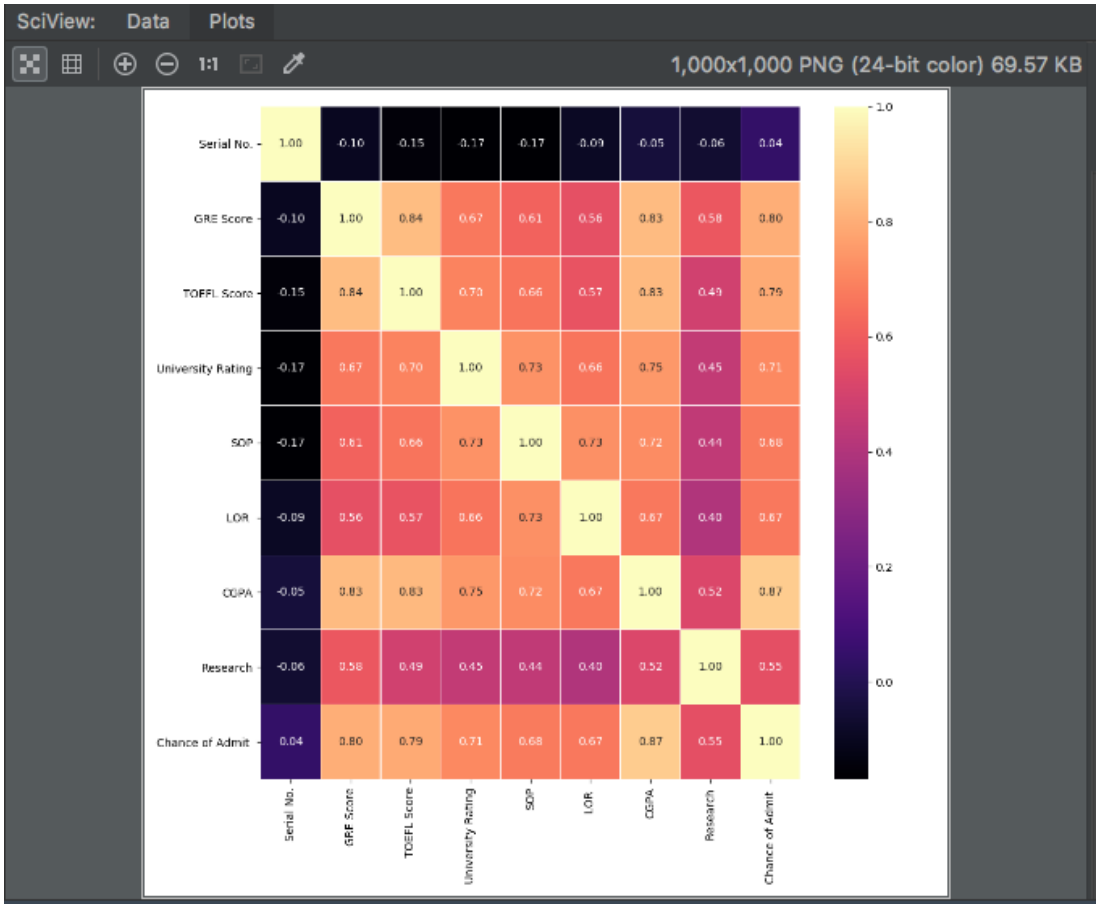
A. Did the background research on the algorithms what I will use in theoretically, read recent decades books and papers to see if innovations had been made.

B. Read the dataset, check the basic information about this dataset.

```
There are 9 columns:
Serial No., GRE Score, TOEFL Score, University Rating, SOP, LOR , CGPA, Research, Chance of Admit , <class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 9 columns):
Serial No.           400 non-null int64
GRE Score            400 non-null int64
TOEFL Score          400 non-null int64
University Rating    400 non-null int64
SOP                  400 non-null float64
LOR                  400 non-null float64
CGPA                 400 non-null float64
Research             400 non-null int64
Chance of Admit      400 non-null float64
dtypes: float64(4), int64(5)
memory usage: 28.2 KB
```

C. Check is there any missing value in the dataset.

D. According to the heat map of the data analysis, we can see that The 3 most important features for admission to the Master: CGPA, GRE SCORE, and TOEFL SCORE. And The 3 least important features for admission to the Master: Research, LOR, and SOP.



5. What is your detailed plan for the rest of the project.

Week 10 (spring break) - finalize the introduction section, and start to write methodology section of the report as well as citations.

Week 11 - continue to prepare the dataset, analyze the correlation between each categories/parameters, try to visualize the relationship maybe by subplot or some other method to show.

Week 12 - write the decision tree python code and do the analysis, meanwhile, supplement methodology and references.

Week 13 - write the random forest python code and do the analysis, meanwhile, supplement methodology and references.

Week 14 - write the linear regression python code and do the analysis, meanwhile, supplement methodology and references.

Week 15 - do the comparison for above analysis, write result and analysis section of the report. I wish to compare different models to check for best model depending on r_squared score and accuracy score value.

Week 16 - Write the summary section of the report and finalize the everything. The application of the goal: with the high accuracy predictions, graduate applicant/student can self-define and local their capacities to help them to make the decision if or not apply for a graduate school, and if so how much chance they will have by competitor information at the same period.

Refences

[1] Mohan S Acharya, Asfia Armaan, Aneeta S Antony : A Comparison of Regression Models for Prediction of Graduate Admissions, IEEE International Conference on Computational Intelligence in Data Science 2019.

[2] Graduate Admissions. Predicting admission from important parameters. 2019. https://www.kaggle.com/mohansacharya/graduate-admissions

[3] Richa Bhatia. How Do Machine Learning Algorithms Differ From Traditional Algorithms? 2018. https://www.analyticsindiamag.com