**Introduction**

I have chosen a data about various factors which were noted during admission to the university. There are 8 features in total i.e., GRE score, TOEFL score, university rating, SOP (Statement of Purpose), LOR (Letter of Recommendation), CGPA, research and chance of admission (chance of Admit). Total 400 entries/samples are included in this dataset. Chance of Admit is the binary/categorical variable (has a continuous value between 0 and 1) which is the response variable and helps in prediction of the admission. I have implemented following parts in this project

1. Data loading and importing libraries
2. Data visualization and correlation using Seaborn
3. Data splitting into test and train using sciklearn packages
4. Running Multiple Linear Regression model from pandas
5. **Importing necessary libraries**

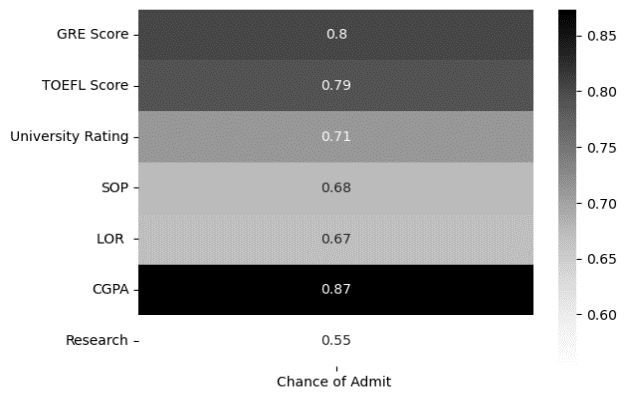
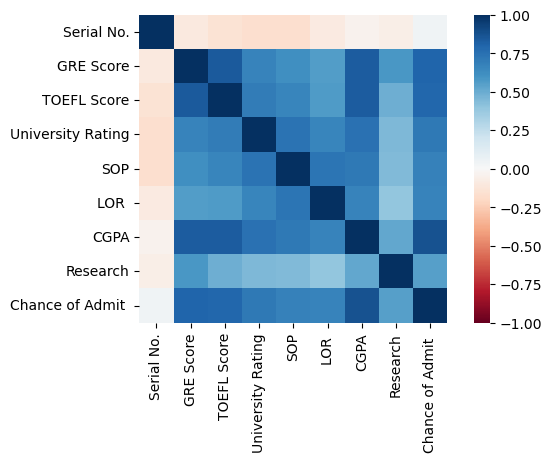
Different data manipulation and analysis libraries can be used in python, Pandas is a powerful library which I used for my data analysis, additionally matplotlib and seaborn were imported to make graphs and plots.

1. **Loading Dataset**

I loaded my excel dataset from my PC with following commands. Although data can also be found online with this link <https://www.kaggle.com/code/suneelpatel/graduate-admission-analysis-and-prediction/data>.

1. **Data Exploration/Data Visualization**

To better understand my data and to observe any patterns in it, I carried out exploratory data analysis with codes presented in the separate file named “Codes”. This data consists of 400 rows and 9 columns. Data was explored and checked for any null values in the dataset, but it seems very clear as no null or duplicate value was observed in the data. To visualize the correlation of all variables first a heatmap is plotted. Scatter matrix was created using following codes, individual correlation between various variables can also be plotted so see the picture more clearly. I checked the correlation of Chances of Admit with all the variables, the column ‘Serial No.’ does not seem to be necessary for admission prediction so this column is dropped out of the data. A pairplot image was also generated to see all the correlations. I was clear from the correlation map that CGPA has a strong correlation with their chance of admit, followed by GRE Score and TOFEL Score. Chance of Admit is weakly correlated with Research.



A)

B)

**Figure 1: Correlation Visualization.** A)Heatmap image showing correlation between all X variables in the data. B) the image shows only the correlation of Chances of Admit with other variables after dropping serial number column from the data.

1. **Data Splitting**

Before fitting the data into any model, I splitted the data into Train and Test, using train\_test\_split module of scikitlearn library (it will split the data randomly), where I chose the splitting ratio as 30% for test data. The train data will train the model and verify the test set. X (x1, x2…) is all the independent variables (CGPA, GRE Score, TOFEL Score, SOP, Research etc) while Y is the dependent variable i.e., Chance of Admit (need to predict).

1. **Building Regression Model**

I used and compared two different regression models i.e., multiple Linear regression and RandomForestRegressor models to predict the chances of admission in the university based on independent (predictor) variables. I am using the chance of admission as target variable and other variables as predictors. My results showed that linear regression model performed well with r2 value of 0.8 as compared to 0.71 from RandomForestRegressor. Although the r-square values from both models show variance but the mean absolute error was near around 0.05 for both models (linear regression MSE= 0.044 and RandomForestRegressor MSE=0.049).

**Conclusions**

This dataset was consisting of 8 independent variables and chance of Admit was dependent variable, I had to predict the factors which can affect the chances of admission to the university. Exploratory data analysis could show that CGPA, GRE Score and TOFEL Score are the top three features that have strong correlation with chances of Admit. After analysing the dataset with two different models, the linear regression remain the best model with high r2 and low MSE value.