# College Admission

Project 5

**Analysis Tasks:** Analyze the historical data and determine the key drivers for admission.

**Predictive:**

- Find the missing values. (if any, perform missing value treatment)
  - There are no missing values in the dataset
- Find outliers (if any, then perform outlier treatment)
  - Outliers were there in the column “gre” and “gpa” and are removed by replacing the values with the mean values of the respective column.
- Find the structure of the data set and if required, transform the numeric data type to factor and vice-versa.
  - Some of the categorical column which are structured as numeric/integer due to encoded value of 0,1 or 0,1,2,3 for the ease of project analysis are converted into factors.
- Find whether the data is normally distributed or not. Use the plot to determine the same.
  - Data type was not normally distributed as the mean value of two important numeric column are not equal to mode and median value.in visualization also mean is not equal to median and mode.
- Normalize the data if not normally distributed.
  - The data is normalized by changing the values of numeric column into their under-root values. doing so, reduced the difference between mean mode and median values of respective columns. Now the data is almost equal to normally distributed.
- Use variable reduction techniques to identify significant variables.
  - First the multiple regression model is used to find the significant variable and then with the use of backward elimination method the insignificant variables are eliminated from consideration.
- Run logistic model to determine the factors that influence the admission process of a student (Drop insignificant variables)
  - The logistic regression model used to determine the factors that influences the admission process of student shows the probabilistic results of one to get admission or not on the basis of the significant variables.
- Calculate the accuracy of the model and run validation techniques.
  - Accuracy of the logistic regression model is calculated and validated by splitting the model int test and training set and then making logistic regression model on training set and making prediction over test set and cross checking with the original results.
- Try other modelling techniques like decision tree and SVM and select a champion model
  - Other models like decision tree, SVM and KNN are used to select the champion model.
- Determine the accuracy rates for each kind of model
  - Accuracy of all the models are checked by making confusion matrix.
- Select the most accurate model
  - By comparing all the accuracy’s, the most accurate model is selected.
- Identify other Machine learning or statistical techniques
  - Other machine learning operations like forest forecasting of the dataset is used.

All the codes used are attached in source code tab as in a pdf formats and the results of every particular segment is also attached in either screenshot section as shown by console in the screenshot section as pdf.

**Descriptive:**
Categorize the average of grade point into High, Medium, and Low (with admission probability percentages) and plot it on a point chart.
Cross grid for admission variables with GRE Categorization is shown below:

| GRE | Categorized |
|---|---|
| 0-440 | Low |
| 440-580 | Medium |
| 580+ | High |

From the visualization it is clear that
The students with gre score up to 440 has an average probability of 12.5% to get admission.
The students with a gre score of 440 to 580 has an average probability of 27.7% to get admission.
The students with gre score more than 580 has an average probability of 39.5% to get admission.

All the codes used are attached in source code tab as in a pdf formats and the results of every particular segment is also attached in either screenshot section as shown by console in the screenshot section as pdf.