Introduction:

This a brief summary of the analysis of a drug dataset. The outputs as well as the essential findings have been mentioned here.

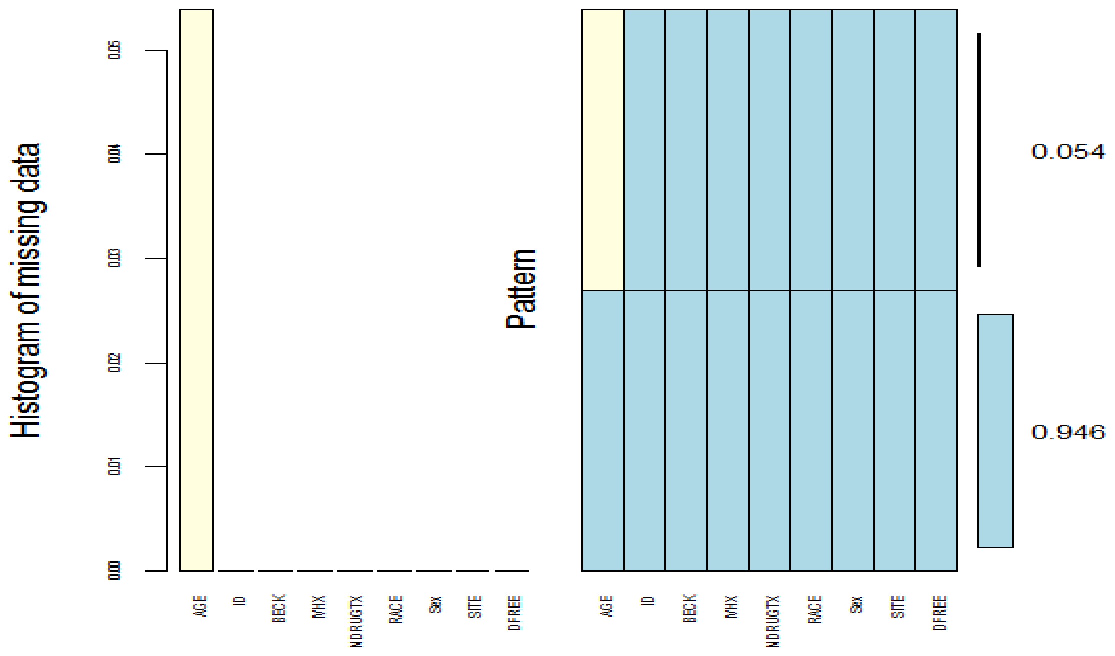
The analysis has been carried in four steps:

1. Loading & cleaning the dataset
2. Exploratory Data Analysis
3. Implementation of the ML Algorithms
4. Comparative study of ROC curves

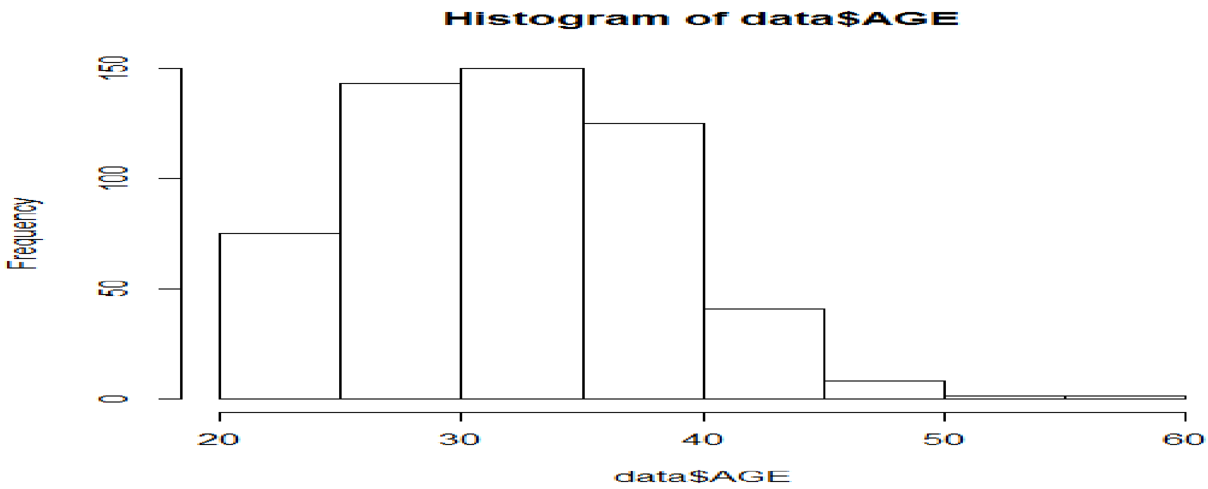
Loading & Cleaning the dataset:

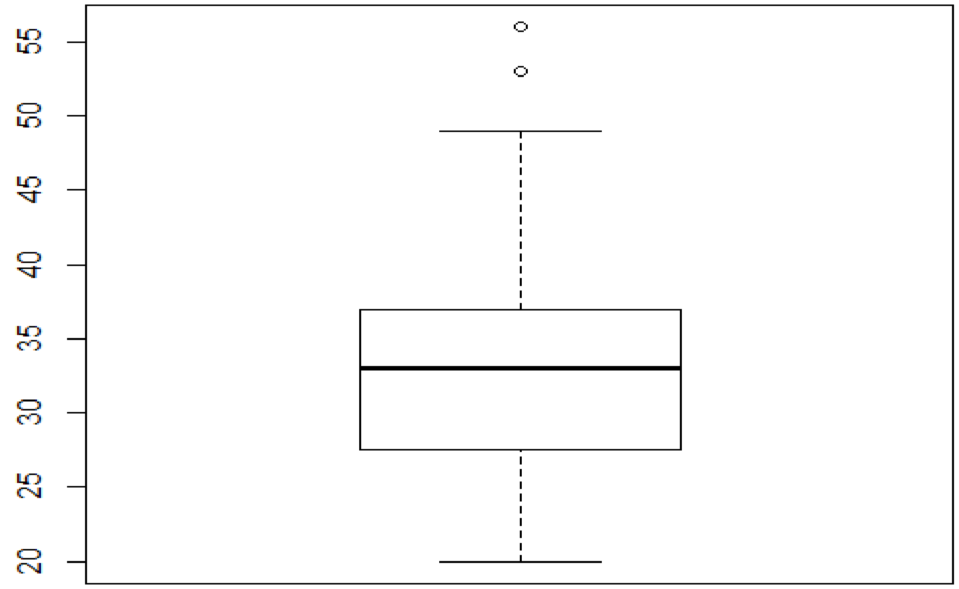
In order to proceed with this step, the dataset & required packages are first of all loaded into the R environment. The dataset contains NAs coded as empty strings, hence we convert the empty strings to NA while reading the data.

The summary is then checked which shows that except for the Age attribute, NAs are not present. The NAs are further investigated using the aggregate function which gives the following plot:



It can be seen from the plot that 94.6 % of the data is free from missing values. All of the NAs are present in the Age attribute. To observe the AGE, a histogram and boxplot is plotted:

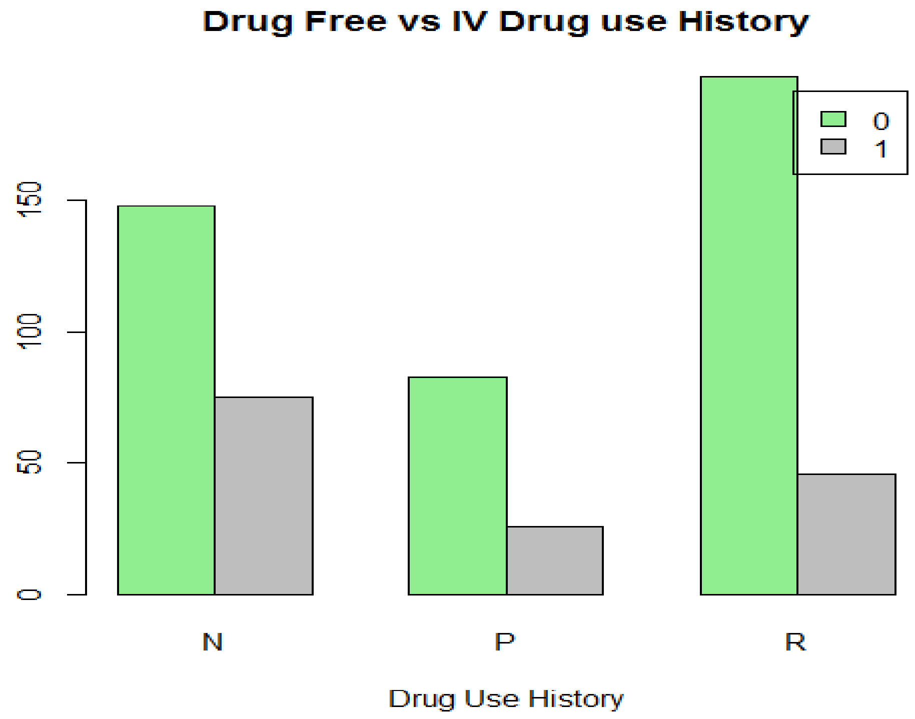


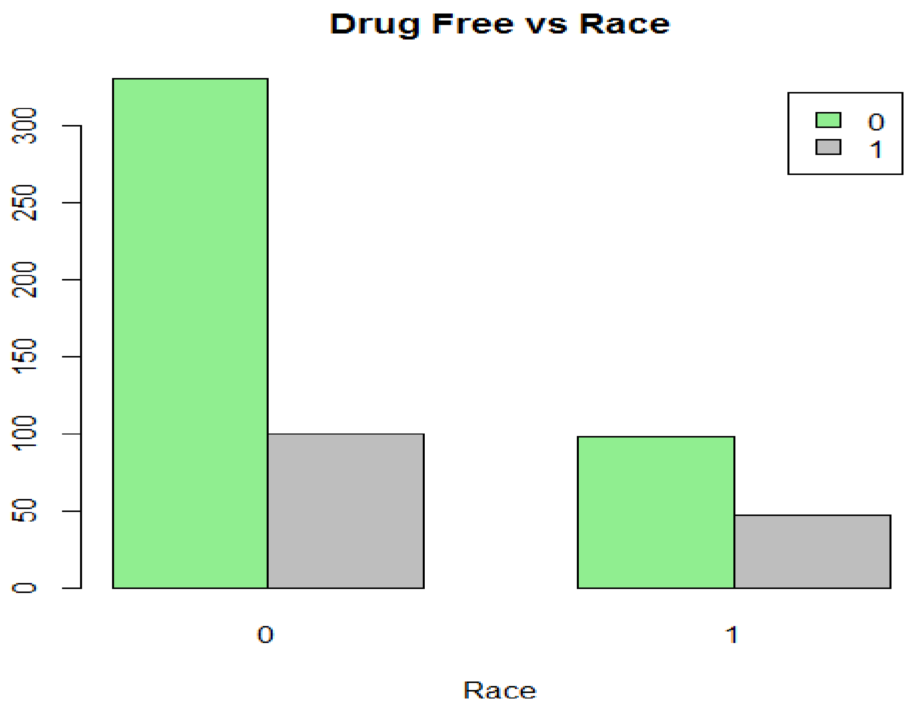


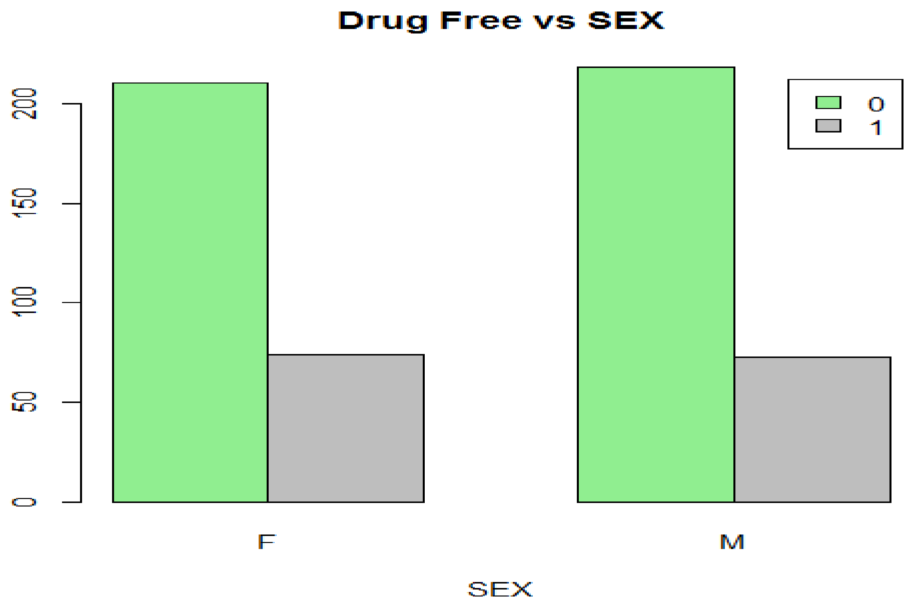
It can be observed that the Age ranges between 20 -56, the first & third quartile is around 27 and 37 and the median is 33. Median seems to be a better choice to replace NAs as there are outliers in the data.

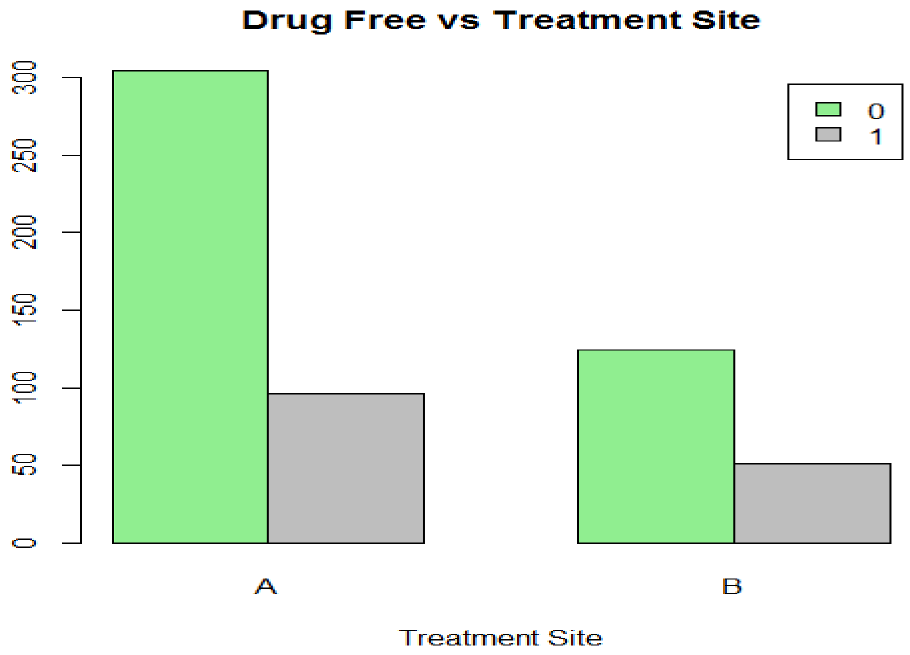
Exploratory Data Analysis:

Barplots are plotted to do a comparative study of the factor variables with the target. These are the plots obtained:




Drug Free vs Race
0
50
100
150
200
250
300
0
1
Race
🟩 0
▨ 1


Drug Free vs SEX
0
50
100
150
200
F
M
SEX
🟩 0
▨ 1

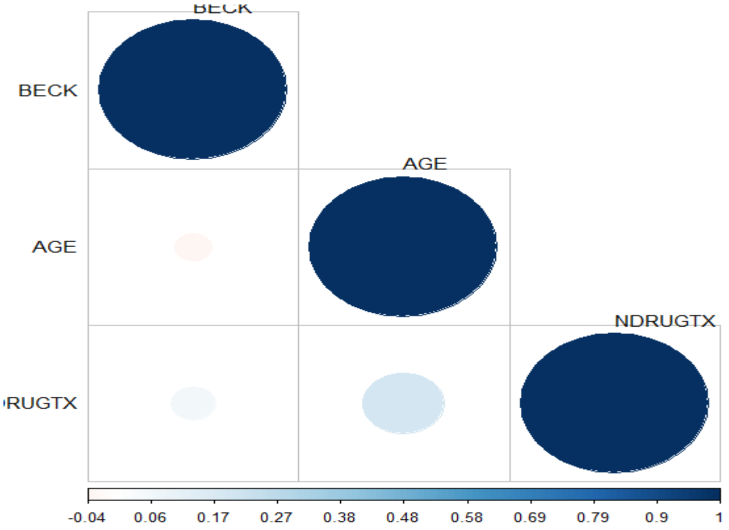

Drug Free vs Treatment Site
0
50
100
150
200
250
300
A
B
Treatment Site
🟩 0
▨ 1

Following inferences can be drawn from the above plots:

- People with recent IV Drug use history form the highest ratio of non-drug free to drug- free individuals.
- People who never had a drug history are in the ratio of 1:2 for drug free individuals to non-drug free individuals.
- White people constitute of good ratio of non-drug free individuals to drug free individuals.
- It is also noticeable that Male and Female individuals are in the same ratio of drug free to non-drug free people, where non-drug free individuals are dominant.
- Treatment site B has a better recovery ratio than treatment site A

A correlation analysis is done among the numeric features which results in the following table and plot:

	AGE	BECK	NDRUGTX
AGE	1.00000000	-0.04108021	0.18957761
BECK	-0.04108021	1.00000000	0.05925075
NDRUGTX	0.18957761	0.05925075	1.00000000



It can be clearly observed that there is no correlation among the attributes.

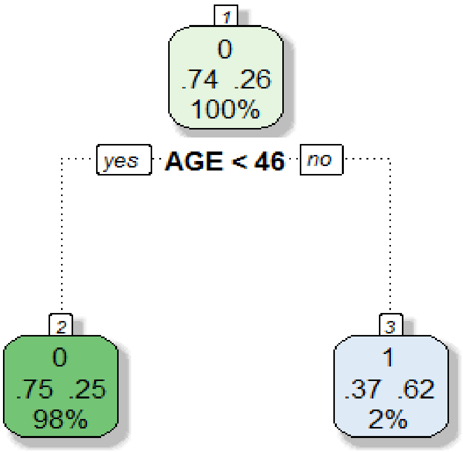
Implementing the ML Algorithms:

The dataset is first divided into training and testing set. The first step is to identify the important attributes. Random forest is very effective to find a set of predictors that best explains the variance in the response variable. Variable importance based on the mean decrease in accuracy is tabulated below:

AGE	BECK	IVHX	NDRUGTX	RACE
0.0117567568	-0.0006756757	0.0059459459	0.0077027027	0.0005405405
Sex	SITE			
0.0037837838	0.0009459459			

Using the above, we eliminate one attribute at a time to build the model to obtain a good accuracy as well as AUC.

- GLM is first used to fit the model. The baseline model gives an accuracy of 73.83%.
- Decision tree gives an accuracy of 74.42 %.
- Random forest results in an accuracy of 75%.



Comparative study of ROC curves:

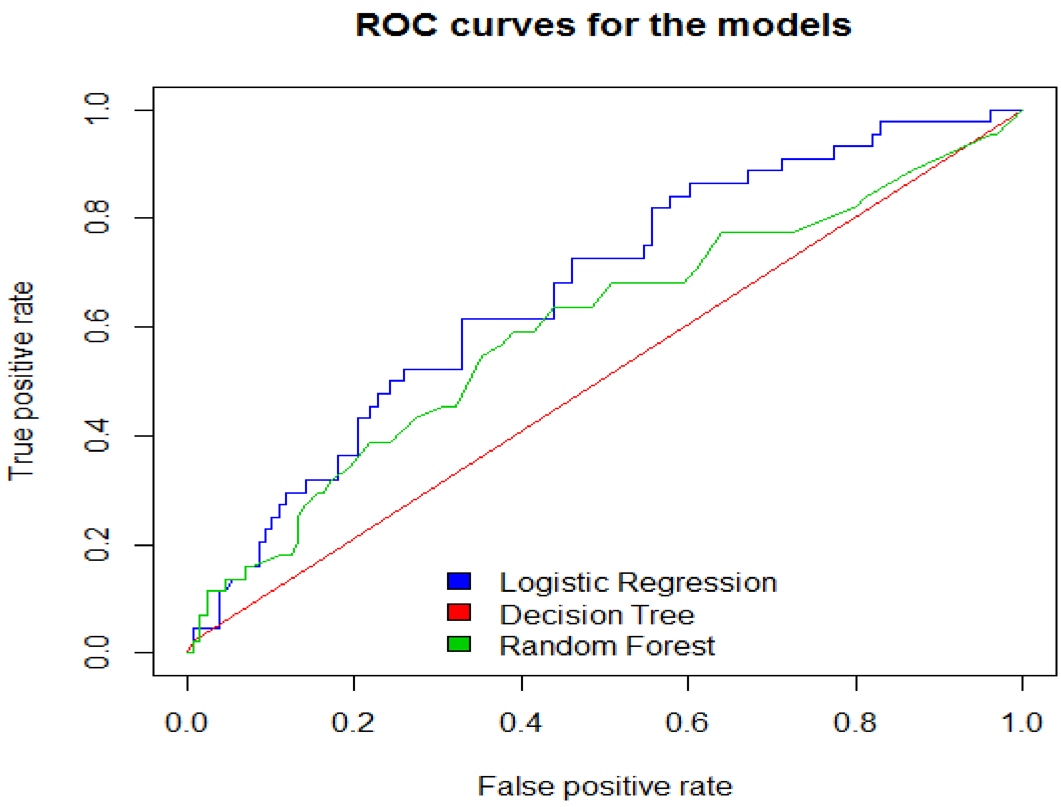
Area under the curve is calculated for the three models. The output obtained is as follows:

AUC of Logistic Regression: 0.668

AUC of Decision Tree: 0.507

AUC of Random Forest: 0.598

The ROC curves obtained is plotted :


ROC curves for the models
True positive rate
False positive rate
🟦 Logistic Regression
🟥 Decision Tree
🟩 Random Forest