Case Study Assignment

Data Mining

DataSet:-Chronic Kidney Disease

Challenges in Dataset:

- Dataset has many missing values.
- Datatype of the columns also incorrect
- All missing values are assigned with “?” and “ ?” which is in string format.
- Some data are stored with spacing in front of the starting letter. ex: “ yes” & “yes”.
- Outliers in the dataset

Data Preprocessing And EDA

- All the missing values are in “?” string format so we replace all the “?” values with nan.
- Datatype of all the columns is in object/string format so we change the datatype of numeric value columns into the float for calculation.
- Map the categorical feature into 1 and 0 format for model input.
- Fill the missing values with mode/median and mode according to the box plot and distribution graph plot.
- Check the correlation of columns with the target column and found that our features have both positive and negative correlation with our target value which is a good for training.
- In fig 1 it shows that some of the features contain a high number of outliers. So we filled those columns missing values with the median.
- In fig 2 it shows that a few columns are creating bell curves and equally distributed but most of them are following left skewed and right skewed patterns.
- Fig 3 contains the correlation values of the features. It shows how the features are correlated with each other and we found that features are correlated with the target columns but not so much correlated with each other which is good for the model.

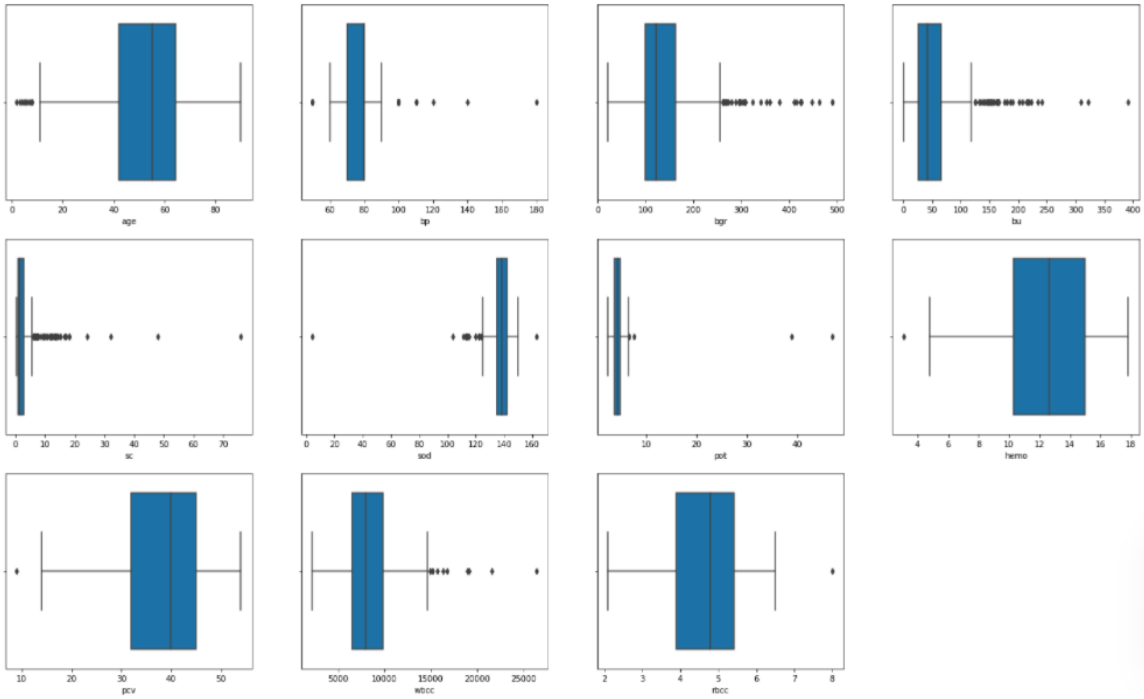

fig 1: Box Plot

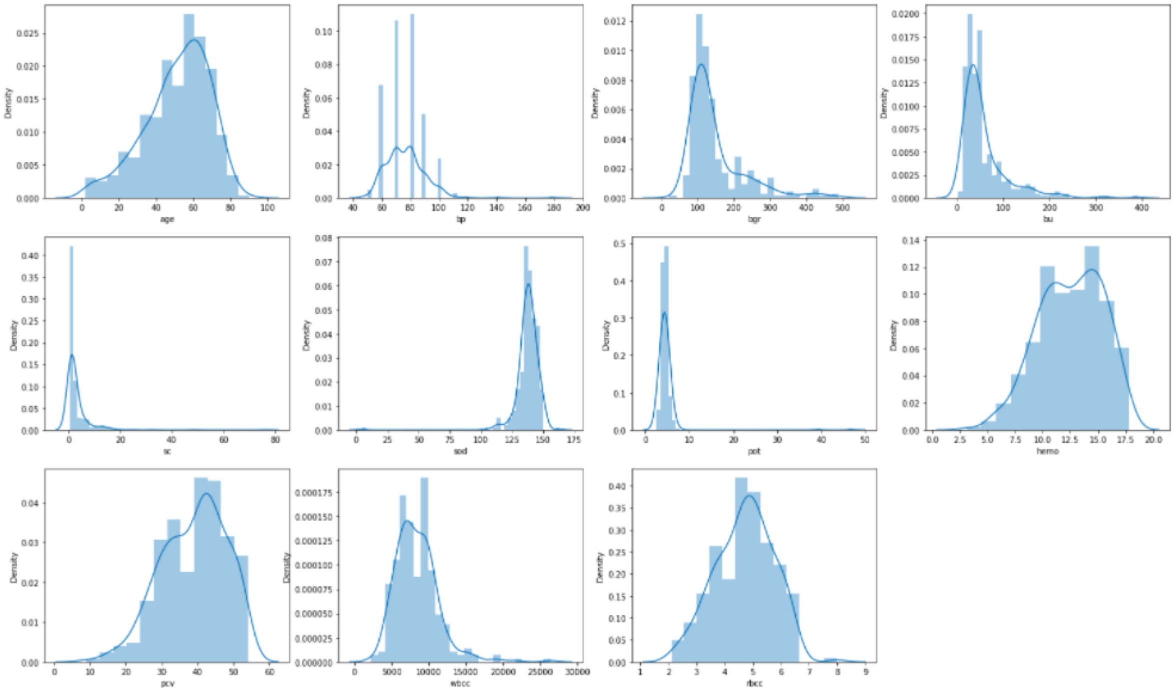

Fig2: Distribution plot

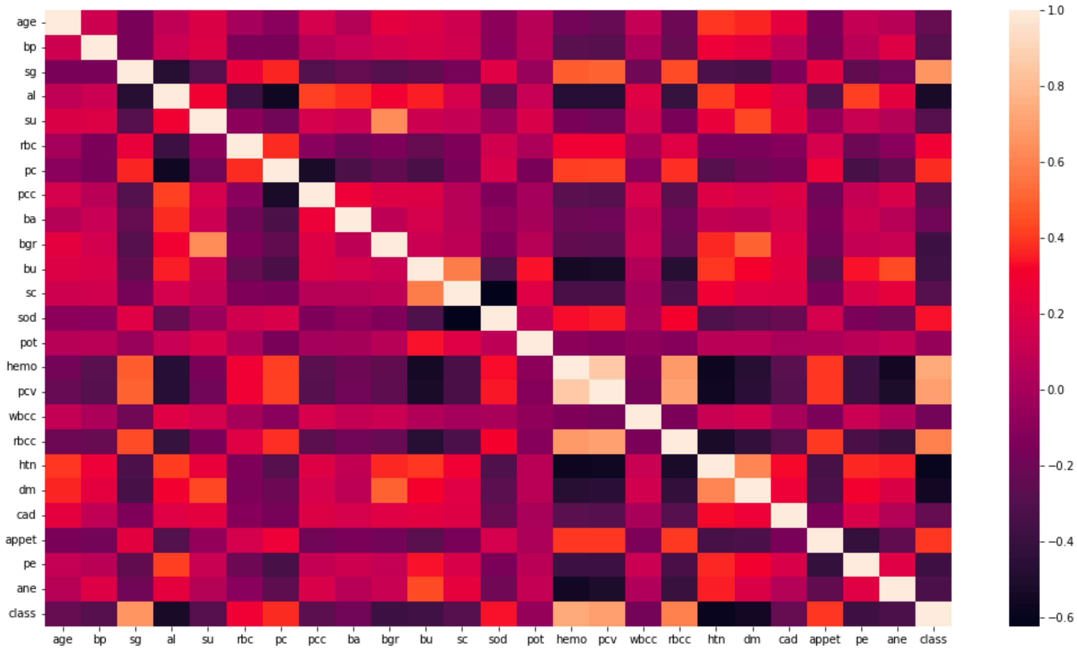

Fig3: Heatmap of correlation of data

Model Building

- We use different algos for building the model like:- random forest, extra tree, LogisticRegression etc.
- We got 99% accuracy with Random forest and Extra tree algo.
- We divide the dataset into 75% training and 25% test.

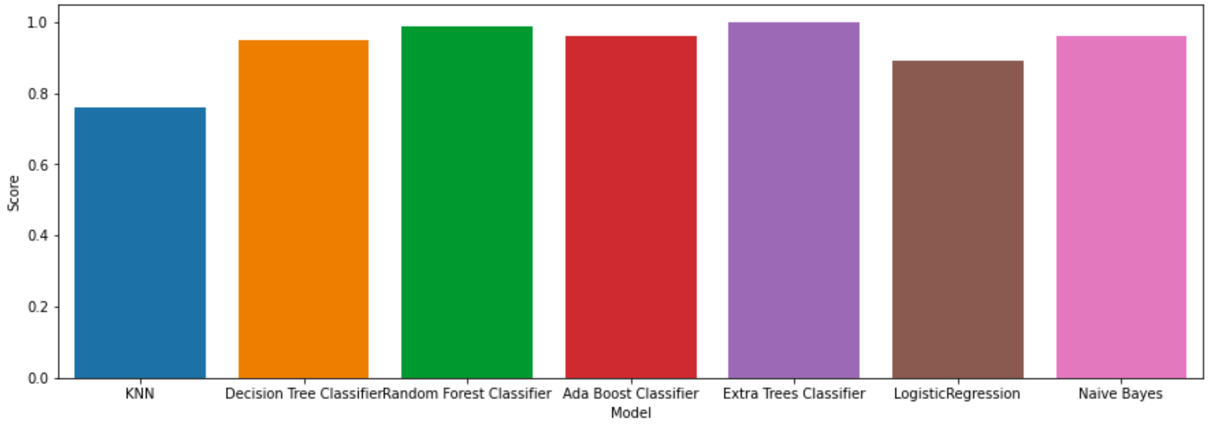

Fig 4: Model Accuracy