**Description**

HCV dataset is the dataset containing laboratory values and information such as demography, sex of several blood donors and Hepatitis C patients. The main objective of this dataset is to identify whether an observation represents blood donor, suspect blood donor, hepatitis, fibrosis or cirrhosis. There are 10 laboratory values affecting the result and demographic information like age, sex and id or no of the patient in the dataset.

The features in the dataset are discussed below:

* Patient ID/ No

Patient ID is just the id of the records mantained for the patients.

* Age

The age of the patient is giiven in this column.

* Sex

Sex of the patient is given in this column. Male patients are recorded as “m” and female patients are recorded as “f”.

* ALB

ALB column in the aAlbumin content in the patient. Albumin is the protein made by liver of our body. It helps to mantain fluid in our blood stream.

* ALP

ALK is the alkaline phosphatase in the patient body. It is the protein found in all body tissues. A blood test can be done to measure the ALP in human body.

* AST

ASP column represents the Asparatate transminase content in the patient. This is an enzyme that is made by liver.

* BIL

BIL represents the Bilirubin content in the patient. It is a yellowish pigment that is made during the normal breaksown of red blood cells.

* CHE

CHE represents the cholenisterase in the human body. It is an enzyme required for proper functioning in the nervous system.

* CHOL

CHOL is the choleestoral level in the pateint body. Cholestoral is a waxy substance found in our blood which helps to build healthy cells. High level of cholestoral increses risk of heart disease.

* CREA

CREA represent thr Creatinine in the patient bosy. It is a waste product produced by muscles from the breakdown of compund called creatine.

* GGT

GGT Gamma Glutamyl Transferase. It is an enzyme formed in liver which when elevated in blood causes damage to liver or bileducts.

* PROT

PROT is the prothromblin protien made by liver which clots the blood when one had cut injjury.

According to such features, the class category of the patient which should be classifies are listed out below:

1. Class 0 representing Blood Donor
2. Class 0’s representing suspect Blood Donor
3. Class 1 representing hepatitis
4. Class 2 representing Fibrosis
5. Class 3 representing Cirrhosis

The main objective in this dataset is to classify patient to above classes according to thier demographic data like age, sex and laboratry data such as ALB, ALP and os on.

**Data Preparation**

The dataset was downloaded from the UCI ML repository website and loaded using the read.table method in R. First ten observations of the data are shown below.

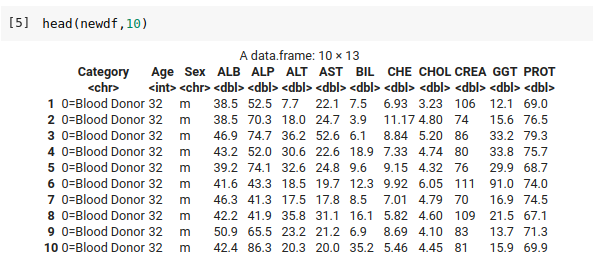


Fig: Screenshot showing first 10 observations of HCV dataset

**Data Preprocessing**

Following processes were involved in the data preprocessing step:

* Convert non numeric data to numeric

The category column values are represented as “0=Blood Donor”. However, for classification, we should convert it into unoredered factor. Thus, initially, such categorical values in the columns are converted in unordered factors. For this a new column, class is added which has the numeric representation of the category class and category class is removed. The column “Sex” is also converted to numeric form by replacing the “m” value with 0 and “f” value with 1.

* Balancing the dataset

Number of data representing each category is not equal and an imbalanced dataset leads to poor classification. Thus, to balance the dataset, we remove certain data from category 0 as a process of undersampling.

* Removing null values

While checking the null values in the dataset, there were 24 NA values.

The missing data pattern was visualized using the md.pattern() function in the mice package available in R. The code used for this analysis ig given below:

install.packages('VIM')

library(VIM)

aggr\_plot <- aggr(trimmedDf, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(data), cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Pattern"))

Following results were obtained:

Variables sorted by number of missings:

Variable Count

ALP 0.15517241

CHOL 0.02586207

ALB 0.00862069

ALT 0.00862069

PROT 0.00862069

Category 0.00000000

Age 0.00000000

Sex 0.00000000

AST 0.00000000

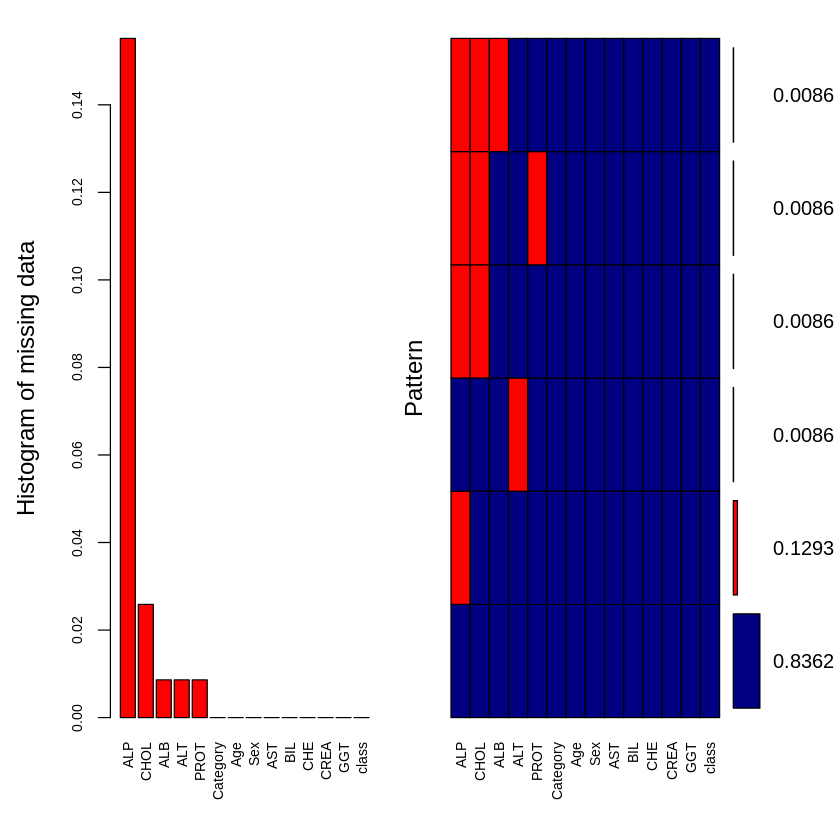
BIL 0.00000000

CHE 0.00000000

CREA 0.00000000

GGT 0.00000000

class 0.00000000

Fig: Pattern observed for missing values in hcv dataset

From the above analysis, we can deduce that most of the values missing are in the columns like “ALP” and “CHOL”.

Such missing data imputation was done using the mice() function from the mice package. We used predective mean matching imputation methos using the mice() function. The code for missing data imputation is given below:

naremovedDF <- mice(trimmedDf,m=5,maxit=50,meth='pmm',seed=500)

summary(naremovedDF)

After this imputation we checked the NA values in the dataset as follows.



After performing above steps in the dataset in R, the first ten observations of the dataset are shown below:

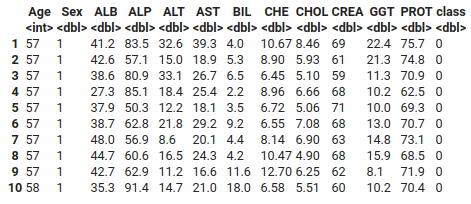


Fig: Screenshot representing first 10 instances of data after preprocessing

**Data Analysis**

* **Correlation Plot**

Initially, we evaluate the correlation between the columns in the dataset. Following correlation plot is obtained.

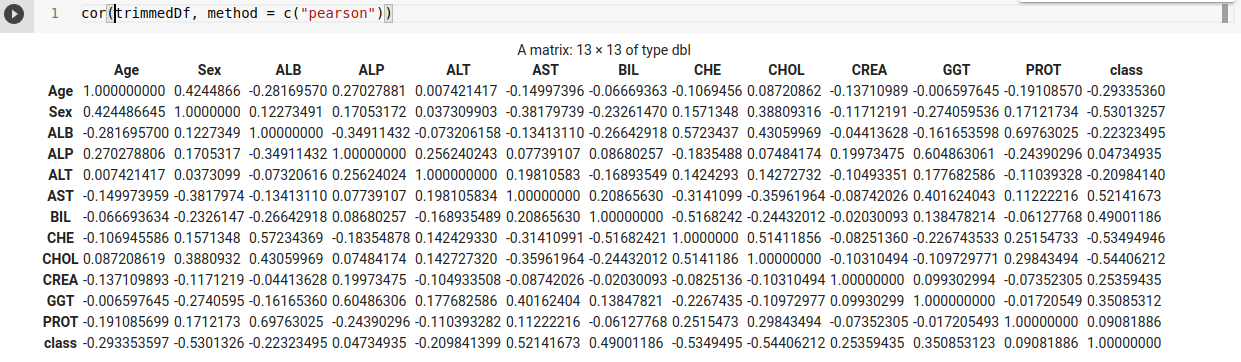


Fig: Correlation plot of the different categorical variables with the response variable

The correlation plot is obtained using the “corrr” package of R. This plot is the Pearson correlation plot of each and every categorical variables in the dataset. The correlation coefficient indicates the level of linear relationship between the two variables. The correlation coefficient close to -1 indicates strong negative linear relationship whereas close to +1 indicates strong positive linear relationship. From above correlation plot, we can deduce that the features “AST” and “BIL” have moderate positive linear relationship with the class column. Similarly, features like “Sex” and “CHE” have moderate negative relationship with the class column. We can also deduce the similar features from the correlation plot. We can deduce that features like “PROT” and “ALB” are highly correlated and columns like “ALP” and “GGT” are also highly correlated. These similar features can be discarded in the feature selection step. Thus from correlation plot evaluation, we discard features like “ALB” and “GGT”.

**Visualization of correlation plot**

The correlation values obtained above from the corrr package was used to visualize the correlation plot. The “corrplot” package in R was used to visualize the correlation plot of the data.

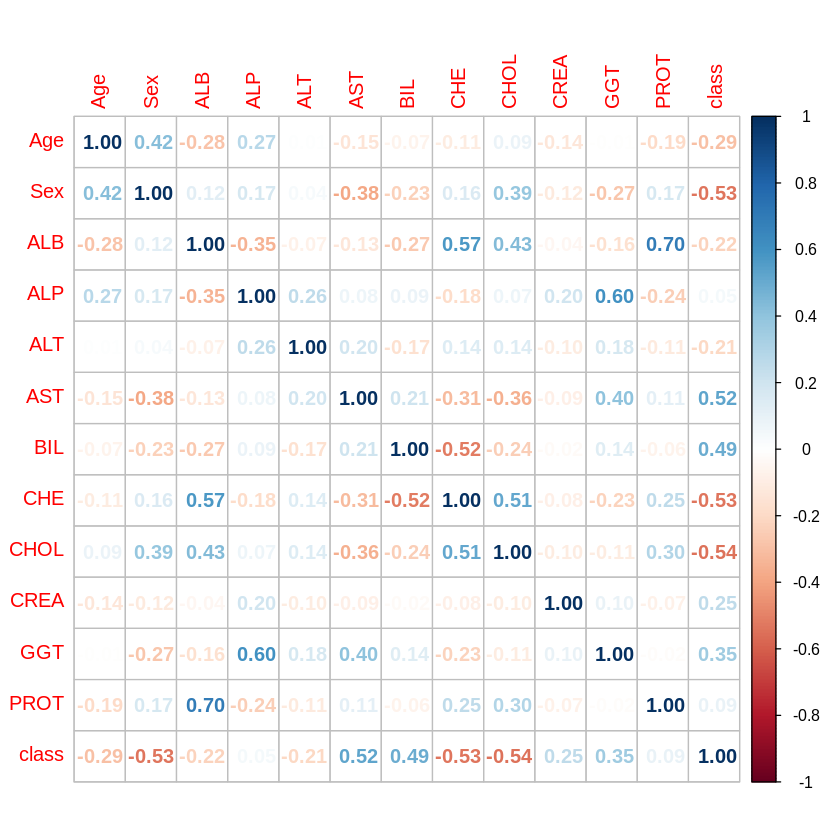
Following code was used to visualize the corrrelation values for the datset:

library("corrplot")

M<-cor(trimmedDf)

corrplot(M, method="number")

Following correlation plot was obtained:

Fig: Correlation plot of the data

From above correlation plot, we can observe that the class column is highky correlated to attributes like “AST” and “BIL”. They have good positive correlation and this indicated that they are important attribute fot the class variable. Similarly, the attributes like “CHE” and “CHOL” have strong negative correlationship with class variable.

* **Scatter Plot**

Scatter plot is generated for the dataset using the pairs.panels function from the “psych” package in R. This feature is used to produce a matrix scatter plot with bivariate scatter plots below the diagonal, histograms on the diagonal, and the correlation of Pearson above the diagonal. Following scatter plot is obtained using this package.

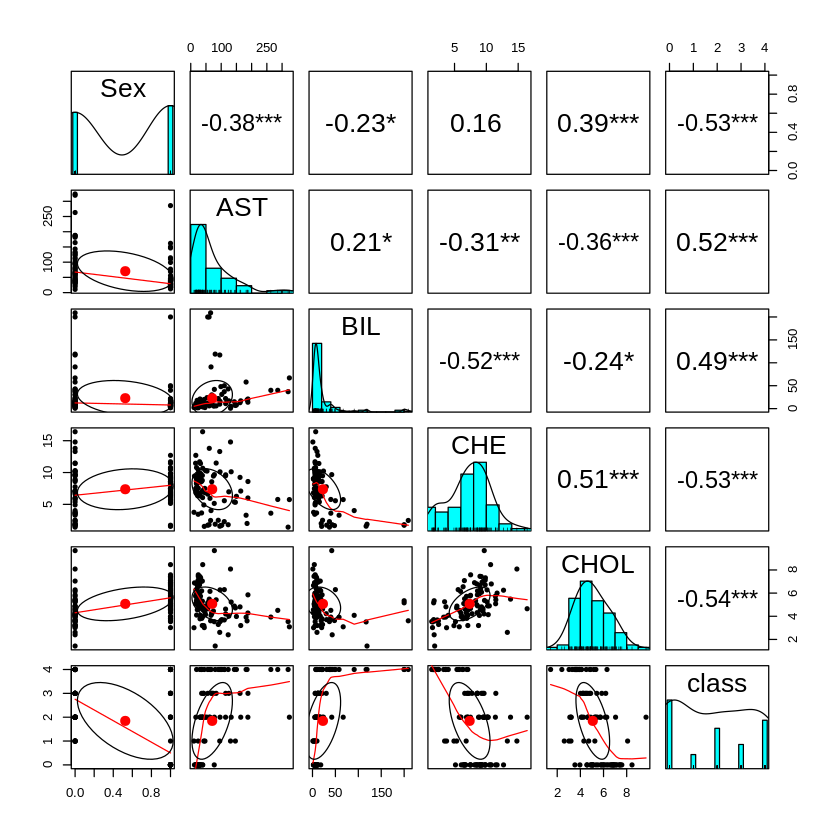


Fig: Scatter plot of the feature selected HVC dataset

Summary of the trimmed dataframe is also generated using summary() method in R. Following is the summary generated.

From the above scatter plot, we can evaluate that features such as “Sex” follows binomial distribution whereas features like “PROT”, “CHE”, “CHOL”, “Age” follows normal distribution.

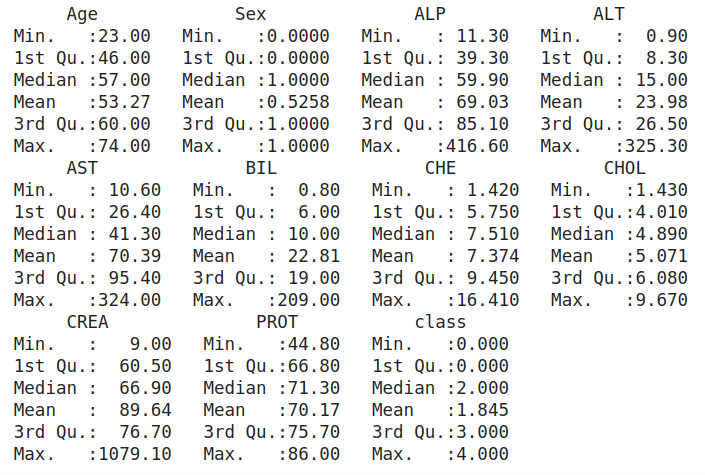


Fig: Summary of the trimmed dataset

**Classification**

The major task in this dataset is to classify the patient whether they belong to “Blood Donor” or “Suspect Blood Donor” or “Hepatitis” or “Fibrosis” or “Cirrhosis”. To develop a classification model, extreme gradient boosting is used. The extreme gradient boosting is done through the use of “xgboost” package in R. Initially the cleaned and preprocessed to a train test split of ratio 0.75: train and 0.25: test. After the train test split, the xgb.DMatrix is prepared for the entire dataset separately for training and testing data. The xgb model developed using this package is subjected to fitting the training dataset and also a k fold cross validation with k =5. After the training of the model and validating model with 5 fold cross validation, out-of-fold `prediction errors were assessed. After assessing such errors, the out-of-fold prediction were obtained as follows:

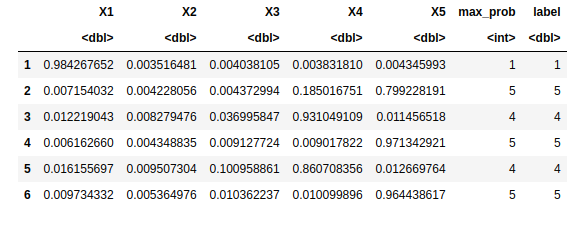


Fig: Out-of-fold prediction errors obtained from xgboost 5 fold cross validation

Here the columns X1, X, X3, X4, X5 represents the cross validation score at every fold.

For the evaluation of the multi-class classification model, confusion matrix and other evaluation metrics were generated from the model subjecting the model with actual test data and the predictions generated from the model from xgboost classification model. Following classification results were obtained from the model:

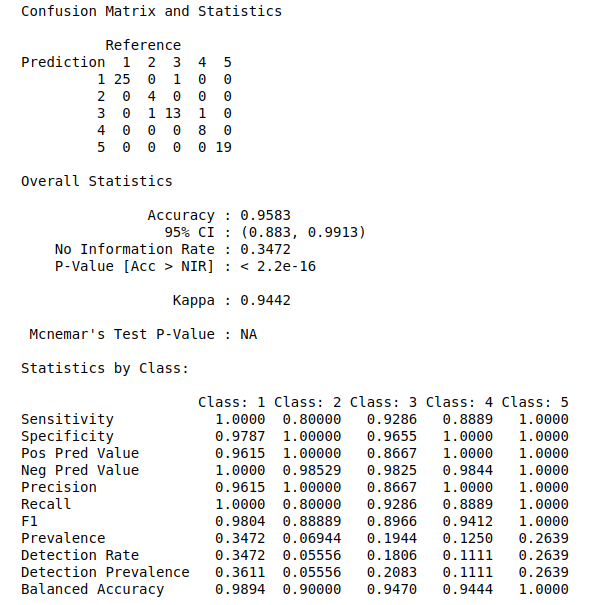


Fig: Classification report for the test data in HCV dataset

Thus, using the xgboost classification technique for this dataset, we achieved an overall accuracy of 95.8% in this dataset. The F1 score of the classification model is very high for three classes whereas around 0.88 F\_score was obtained for the other two classes.