Prediction of presence of Liver Disease in India Using Indian Liver Patient Dataset

Akshay Kumar Valappil Thodi,

School of Computing, Engineering & Intelligent Systems,

Ulster University, Derry/Londonderry Campus,

Abstract: Indian liver disease patient dataset, obtained from UCI ML Repository, has been used to built classification models to predict a subject's chances of having liver disease. Under simple functions in R Studio, a classification tree, a SVM ad a Naïve-bayes model were constructed out of which SVM showed the best accuracy because of its non-parametric nature and tolerance towards outliers. It is concluded that 'Recall' is the best measure of performance in this context and tree model has the highest number.

I. Introduction

Liver health is at an all time decline in India. According to WHO research, 3.17% of total deaths in India happens due to liver diseases. [1] This puts India in the number 83 position in the world, also, this makes liver disease 8th largest reason for death in India as well which is way ahead than dementia or most of the cancer variants. Out of all the liver diseases, cirrhosis tops the list.

Cirrhosis is a scarring stage of the liver which is a result of long-time liver damages. The liver can still function even with cirrhosis, but eventually it leads to liver failure. There are no earlier symptoms, but as it gets worse, the patient starts to have jaundice, they vomit blood, they get itchy skin and swollen legs and much more. There are three major methods of diagnosis for liver cirrhosis. Blood tests, scans and liver biopsy. And there are three major classes of causes for the disease as well, namely, excessive alcohol consumption, long exposure to hepatitis (B or C) and severe form of "Non-Alcoholic fatty liver disease (NAFLD)" [2].

This study focuses on building a model to predict whether a subject has liver disease or not based on blood tests. The intention comes from the fact that detecting the condition at early stage becomes a matter of arch importance since that makes it easy to arrest the condition. But, since there are no major symptoms in the early stages, subjects never really go for the specific tests. This model aims to solve that problem by getting the required data from general full body medical check-ups and blood tests and predict the prospects at earlier stages itself. The model doesn't replace the diagnosis itself, but instead it predicts the possibility of a subject having the disease and the positive class can be subjected to further medical tests. The data set used for this study is the Indian Liver Patient dataset (ILPD). 'R studio' is being used here in the study for both the analysis of the data and construction of the model.

II. Dataset Description

Indian Liver Patient dataset (ILPD), was collected with the intention of building a model to predict the possibility of liver disease in subjects. It was obtained from UCI Machine Learning repository [3].

The data contains 11 attributes including the target variable and 583 patient records collected from the state of Andhra Pradesh, one of the southern states in India. Out of these 584 records, 416 are diagnosed with liver disease and the rest are without the disease. Attribute "Selector" includes this class information. Out of the 10 independent attributes, "gender" is the only categorical

variables and the rest are numerical and are results of blood tests. More information on each variable is given in the appendix section (A).

III. Methodology

ILPD has been analysed using R Studio and excel (very few tables). R Studio has a very large inbuilt packages for data exploration and visualizations which makes it easy to use and understand the data set. At later stages when it comes to building the model, *rpart* was used for classification tree, *e1071* for SVM model and Naïve-Bayes model.

First step into any data analysis process is to load and understand the data using the tool (here R Studio). As already mentioned, there are 11 variables including the class variable "Selector", out of which "Gender" is categorical and the rest are numerical. There are 583 data points and 4 missing values under the variable "AG ratio". As far as the class variable is concerned, it is highly imbalanced as shown below in the graph. 71% with liver disease and the rest without.

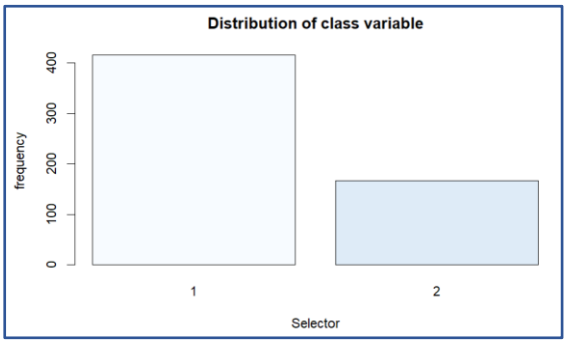


Figure 1: Bar plot of "Selector" variable

Also, from the summary of the dataset it is clear that attributes are all at different scales. This is observed as the medians are highly different form each other. Foor example, the variable "DB" has a median of 0.3 while the variable "Alkphos" has a median of 208. For more details on the summary, please refer the appendix (B).

(a) Missing values

As observed earlier, there were 4 missing values in the variable "AG ratio". Since, the count is too small, it can be ignored and removed. This is being done using the r function *na.omit(Data Frame)*.

(b) Duplicate entries

Duplicate entries can be easily removed from a data frame in r using the function *distinct (Data Frame)*. From the dimensions of the data before and after removing the duplicates, it is understood that there were 13 duplicate entries in the data.

(c) Balancing the Data

Highly imbalanced data results in a biased model [4]. There are multiple methods to handle biased data. SMOTE or Synthetic Minority Oversampling Technique is one such method where the algorithm generates synthetic samples of the minority class [4]. In R, SMOTE-NC function under the package RSBID easily does the job. This specific function deals with datasets with both nominal and continuous variables as the name suggests (NC). The distribution of the target variable after SMOTE is given below in the figure 2.

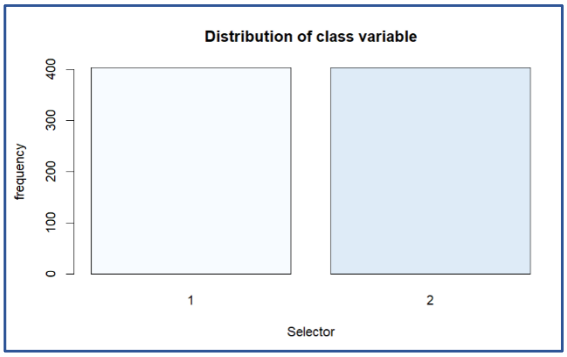


Figure 2 Bar plot of the "Selector" variable after SMOTE

(d) Scaling variables

When the attributes are at different scales, it affects different calculations within the models' algorithms. This can be easily solved using the *scale* function in r. Scaling results can be checked by taking the summary of the data frame before and after the scaling process. Check appendix(C) for the summary after scaling.

(e) Multicollinearity

Multicollinearity is when multiple independent variables are correlated with each other. When this happens, it affects the built model in different ways, such as, it makes it difficult to tell which are the most important variables in the model, the variable co-efficient may not be accurate and over-fitting. One way to identify this is to plot a correlation plot. Given below is the correlation plot for the ILPD dataset.

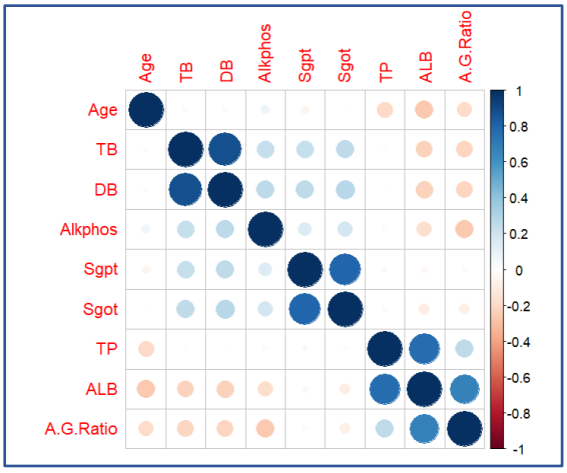


Figure 3: Correlation plot

From the image it is clear that (TB, DB), (Sgpt, Sgot) and (ALB, TP) are relatively correlated. To get a closer look, Pearson correlation coefficient for each of these pairs were calculated and were found to be 0.88, 0.79 and 0.76 respectively. Since they had a correlation more than 80%, one of DB and TB could be removed to solve the multicollinearity issue. There are other methods such as combining the variables to form a new variable or using dimension reduction methods such as PCA. Here, the variable DB has been removed.

(f) Outlier detection

Detecting and solving outlier problem is one of the major tasks in data cleaning. One of the ways to identify outliers is to observe the boxplot. So, box plots were constructed for each of the variables. It turns out, there are outliers in all of them except for "Age" variable. Check appendix (D) for all the boxplots.

(g) Modelling

Since the classes are already known, the models needed in this process are supervised classification models. The models chosen here are Classification Tree, Support Vector machine (SVMs) and naïve Bayes model. These models have been chosen for specific reasons. Classification tree is a non-parametric model which doesn't assume any distribution on the data set and also is tolerant towards outliers. SVM is also highly tolerant towards outliers and these two models accounts for the outlier problem of the dataset that has not been solved yet. The third model Naïve- bayes is not tolerant towards outliers and that will give a fairly good idea on how far the outliers affects a machine learning model.

- Classification Tree

This is a non-parametric model which has a tree like flow-chart structure to classify the data. It has a root node, internal nodes and leaf nodes in its structure where the leaf nodes are the classes in the target variable [5]. The visual representation of the model built is given in appendix (E)

- Support Vector Machine

An SVM classifies the data by choosing the best hyperplane which creates the best margin between the boundary and the closest data point. An SVM works best if the classification problem is binary and is rarely affected by outliers [6]. Appendix (F) has the summary output of the SVM model.

- Naïve Bayes Model

This model uses bayes theorem to predict the probability of any data point to be in a particular class and based on the cut off classify the data [7]. Since it choses the class based on the existing data points, it is highly affected by the outliers. Summary output of the model can be found in appendix (G).

(h) Fitting of the model

All the model has been fit on a training data which is 80% of the original data split maintaining the class balance and the performance has been measured on the remaining 20% i.e., test data. As mentioned earlier, *rpart* function was used to build the classification tree model, *svm* function under e1071 package to build the SVM model and *naiveBayes* function to build the NB Model.

IV. Results

The image in appendix (H) summarises the performance of each model built. As can be seen from the figure, in terms of accuracy i.e., percentage of rightly classified data points, SVM performs the best. For better understanding, look at the table in appendix (I).

V. Discussion

As mentioned in the previous section, there are 4 major performance measures that can be calculated from the models' confusion metrices (Appendix (E, F & G)). Given below are the equations for each of those metrices.

Accuracy = (TP + TN)/ (TP+TN+FP+FN)

Precision = (TP)/(TP+FP)

Recall = (TP)/(TP+FN)

F1Score= (2*Precision*Recall) / (Precision+Recall)

Confusion Matrix		Actual	
		+	-
Prediction	+	True Positive (TP)	False Positive (FP)
	-	False negative (FN)	True Negatives (TN)

Figure 4: Typical CM

Which of these 4 needs to be considered to find the best model would be decided based on the intention of the study and the intention is to predict whether a subject has a possibility of having liver disease so that they can be subjected to further tests and if positive treated. So, it is important to select performance measure which predicts maximum true positives out of total actual positives. This is, as per the formulae given, "Recall".

VI. Conclusion & Recommendations

As per this logic, Classification tree shows the best result with 76.25% recall. This could be because the model is non-parametric and is highly tolerant towards the outliers.

It is shown by the studies that alcohol consumption has a high positive correlation with liver disease so, finding more data on the alcohol consumption of the subject could improve the model.

VII. References

[1] WORLDHEALTHRANKINGS, "INDIA: LIVER DISEASE," 2020. [Online]. Available: https://www.worldlifeexpectancy.com/india-liver-disease. [Accessed November 2023].

[2] NHS, "Overview Cirrhosis," NHS, 29 June 2020. [Online]. Available: https://www.nhs.uk/conditions/cirrhosis/. [Accessed November 2023].

[3] UCI, "ILPD (Indian Liver Patient Dataset)," UCI, 20 may 2012. [Online]. Available: https://archive.ics.uci.edu/dataset/225/ilpd+indian+liver+patient+dataset. [Accessed November 2023].

[4] D. K. P. P. Sotiris Kotsiantis, "Handling imbalanced datasets: A review," GESTS International Transactions on Computer Science and Engineering, vol. 30, pp. 25-36, 2005.

[5] O. Z. M. Lior Rokach, Data Mining with Decision Trees, World Scientific, 2008.

[6] L. Wang, Support Vector Machines: Theory and Applications, Springer, 2005.

[7] F. Sabry, Naive Bayes Classifier, One Billion Knowledgeable, 2023.

VIII. Appendix

(A)

Variables Table

Variable Name	Role	Type	Demographic	Description
Age	Feature	Integer	Age	Age of the patient. Any patient whose age exceeded 89 is listed as being of age "90".
Gender	Feature	Binary	Gender	Gender of the patient
TB	Feature	Continuous		Total Bilirubin
DB	Feature	Continuous		Direct Bilirubin
Alkphos	Feature	Integer		Alkaline Phosphotase
Sgpt	Feature	Integer		Alamine Aminotransferase
Sgot	Feature	Integer		Aspartate Aminotransferase
TP	Feature	Continuous		Total Proteins
ALB	Feature	Continuous		Albumin
A/G Ratio	Feature	Continuous		Albumin and Globulin Ratio
Selector	Target	Binary		Selector field used to split the data into two sets (labeled by the experts)

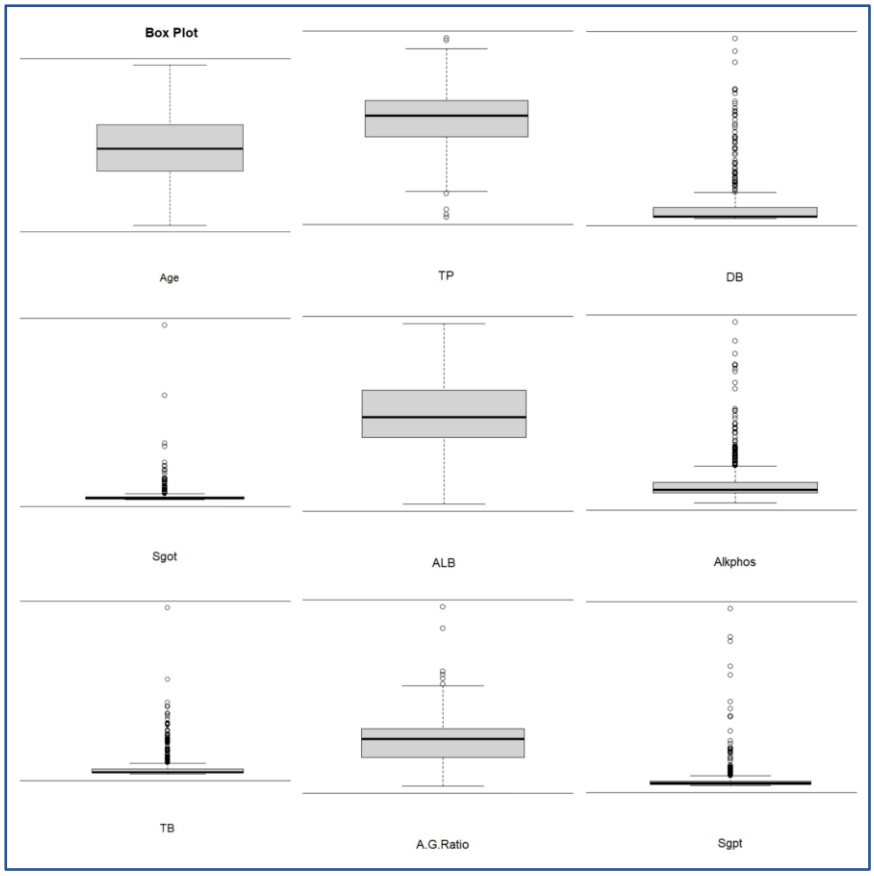
(B) Summary of the data set

```
      Age           Gender          TB               DB            Alkphos          Sgpt             Sgot
 Min.   : 4.00   Female:142   Min.   : 0.400   Min.   : 0.100   Min.   :  63.0   Min.   :  10.00   Min.   :  10.0
 1st Qu.:33.00   Male  :441   1st Qu.: 0.800   1st Qu.: 0.200   1st Qu.: 175.5   1st Qu.:  23.00   1st Qu.:  25.0
 Median :45.00                Median : 1.000   Median : 0.300   Median : 208.0   Median :  35.00   Median :  42.0
 Mean   :44.75                Mean   : 3.299   Mean   : 1.486   Mean   : 290.6   Mean   :  80.71   Mean   : 109.9
 3rd Qu.:58.00                3rd Qu.: 2.600   3rd Qu.: 1.300   3rd Qu.: 298.0   3rd Qu.:  60.50   3rd Qu.:  87.0
 Max.   :90.00                Max.   :75.000   Max.   :19.700   Max.   :2110.0   Max.   :2000.00   Max.   :4929.0

       TP              ALB          A.G.Ratio        Selector
 Min.   :2.700   Min.   :0.900   Min.   :0.3000   Min.   :1.000
 1st Qu.:5.800   1st Qu.:2.600   1st Qu.:0.7000   1st Qu.:1.000
 Median :6.600   Median :3.100   Median :0.9300   Median :1.000
 Mean   :6.483   Mean   :3.142   Mean   :0.9471   Mean   :1.286
 3rd Qu.:7.200   3rd Qu.:3.800   3rd Qu.:1.1000   3rd Qu.:2.000
 Max.   :9.600   Max.   :5.500   Max.   :2.8000   Max.   :2.000
                                 NA's   :4
```

(C) Summary after scaling

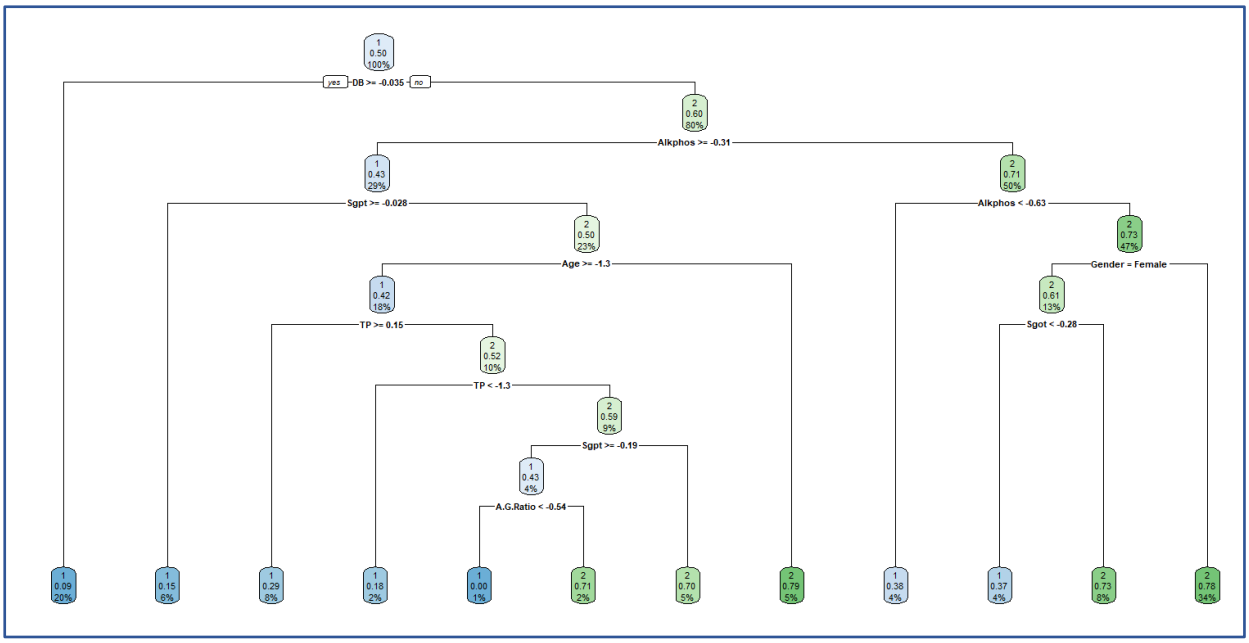
```
       Age.V1           Gender          TB.V1              DB.V1           Alkphos.V1          Sgpt.V1
Min.   :-2.4514073   Female:167   Min.   :-0.419419   Min.   :-0.431838   Min.   :-0.942109   Min.   :-0.361331
1st Qu.:-0.7248726   Male  :641   1st Qu.:-0.350707   1st Qu.:-0.390930   1st Qu.:-0.453906   1st Qu.:-0.283567
Median : 0.0583184                Median :-0.326469   Median :-0.357937   Median :-0.325294   Median :-0.225244
Mean   : 0.0000000                Mean   : 0.000000   Mean   : 0.000000   Mean   : 0.000000   Mean   : 0.000000
3rd Qu.: 0.7550143                3rd Qu.:-0.159159   3rd Qu.:-0.145483   3rd Qu.: 0.058502   3rd Qu.:-0.089157
Max.   : 2.8515207                Max.   :13.448684   Max.   : 7.586100   Max.   : 8.410631   Max.   :12.534523
      Sgot.V1             TP.V1              ALB.V1          A.G.Ratio.V1     Selector
Min.   :-0.317862   Min.   :-3.710611   Min.   :-3.0507498   Min.   :-2.269295   1:404
1st Qu.:-0.264150   1st Qu.:-0.602304   1st Qu.:-0.6719122   1st Qu.:-0.596583   2:404
Median :-0.214214   Median : 0.082383   Median :-0.0111240   Median : 0.072502
Mean   : 0.000000   Mean   : 0.000000   Mean   : 0.0000000   Mean   : 0.000000
3rd Qu.:-0.082958   3rd Qu.: 0.665920   3rd Qu.: 0.7818218   3rd Qu.: 0.526126
Max.   :19.604407   Max.   : 3.000070   Max.   : 3.0285017   Max.   : 6.094267
```

(D) Box Plot



(E) Classification Tree summary and plot

```
          1 51 16
          2 29 64

                Accuracy : 0.7188
                  95% CI : (0.6423, 0.7869)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 1.489e-08

                   Kappa : 0.4375

 Mcnemar's Test P-Value : 0.07364

              Sensitivity : 0.6375
              Specificity : 0.8000
           Pos Pred Value : 0.7612
           Neg Pred Value : 0.6882
               Prevalence : 0.5000
           Detection Rate : 0.3187
     Detection Prevalence : 0.4188
        Balanced Accuracy : 0.7188

         'Positive' Class : 1
```



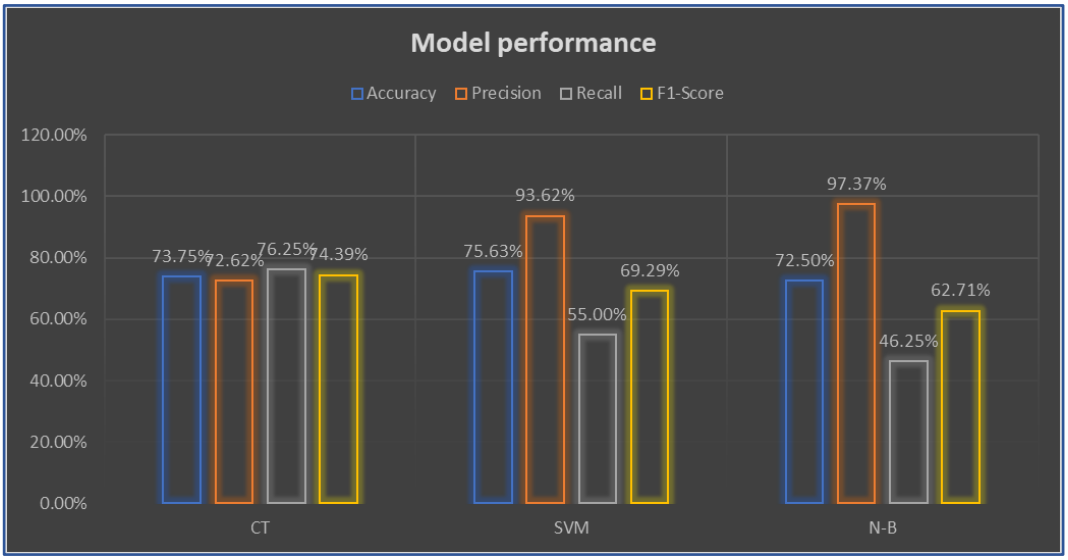
(F) SVM Model Summary

```
Confusion Matrix and Statistics

          Reference
Prediction  1  2
         1 46  6
         2 34 74

               Accuracy : 0.75
                 95% CI : (0.6755, 0.815)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 8.762e-11

                  Kappa : 0.5

 Mcnemar's Test P-Value : 1.963e-05

            Sensitivity : 0.5750
            Specificity : 0.9250
         Pos Pred Value : 0.8846
         Neg Pred Value : 0.6852
             Prevalence : 0.5000
         Detection Rate : 0.2875
   Detection Prevalence : 0.3250
      Balanced Accuracy : 0.7500

       'Positive' Class : 1
```

(G) NB Model summary

```
Confusion Matrix and Statistics

          Reference
Prediction  1  2
         1 37  2
         2 43 78

               Accuracy : 0.7188
                 95% CI : (0.6423, 0.7869)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 1.489e-08

                  Kappa : 0.4375

 Mcnemar's Test P-Value : 2.479e-09

            Sensitivity : 0.4625
            Specificity : 0.9750
         Pos Pred Value : 0.9487
         Neg Pred Value : 0.6446
             Prevalence : 0.5000
         Detection Rate : 0.2313
   Detection Prevalence : 0.2437
      Balanced Accuracy : 0.7188

       'Positive' Class : 1
```

(H) Performance table



(I) Performance Summary table

raw data	CT	SVM	N-B
Accuracy	73.75%	75.63%	72.50%
Precision	72.62%	93.62%	97.37%
Recall	76.25%	55.00%	46.25%
F1-Score	74.39%	69.29%	62.71%