PREDICTION OF PRESENCE OF LIVER DISEASE IN INDIA USING INDIAN LIVER PATIENT DATASET

AKSHAY KUMAR VT

CW-Statistical Modelling and Machine Learning

CONTENT

INTRODUCTION

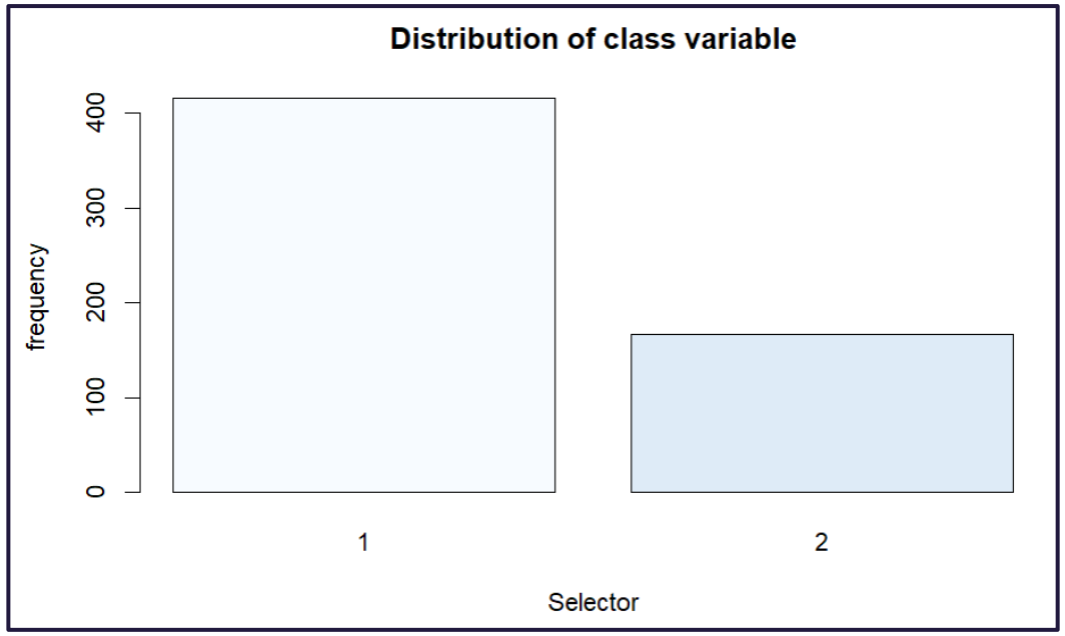
WHAT & WHY

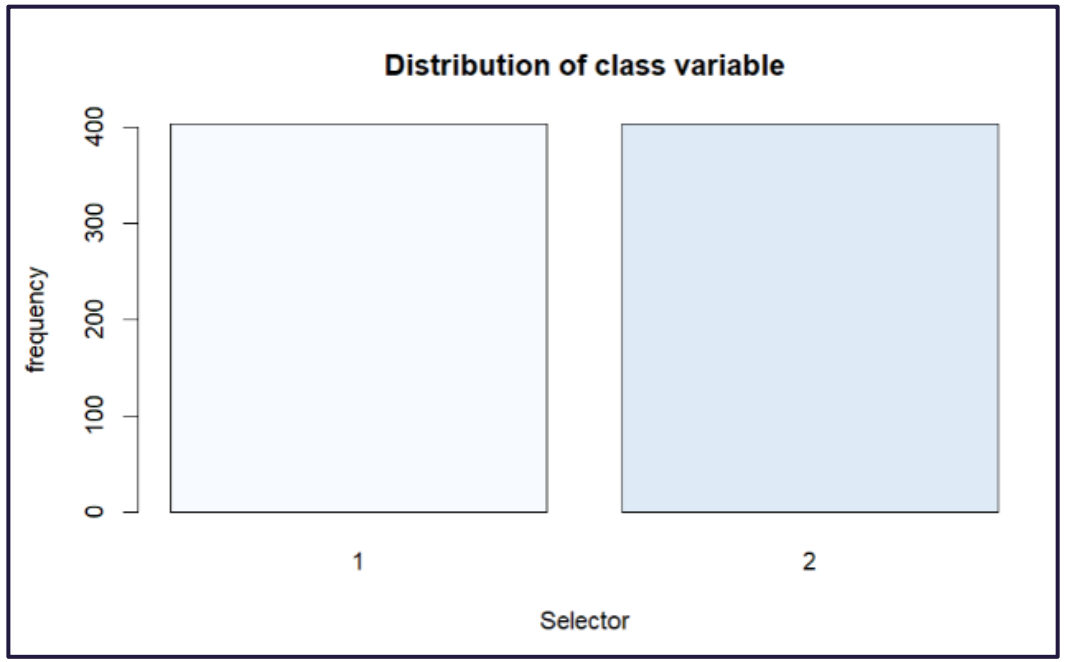
- Prediction model for liver cirrhosis
- Indian Liver Patient Dataset
- 3.17% of total deaths in India happens due to liver diseases
- 8th largest reason for deaths in India
- 83rd position in the world
- Lack of early symptoms and diagnosis

DATASET DESCRIPTION

WHICH & WHERE

- Indian Liver Patient Dataset
- Source : UCI machine learning Repository
- Collected from Andhra Pradesh, southern state in India
- 11 attributes and 583 records
- Target variable : “Selector”
- Categorical variable : “Gender”
- Rest are numerical blood test results
- “Selector” : 416 with liver disease
- Highly imbalanced data, missing values (4), Duplicates (13), Multicollinearity, Outliers, varying scales



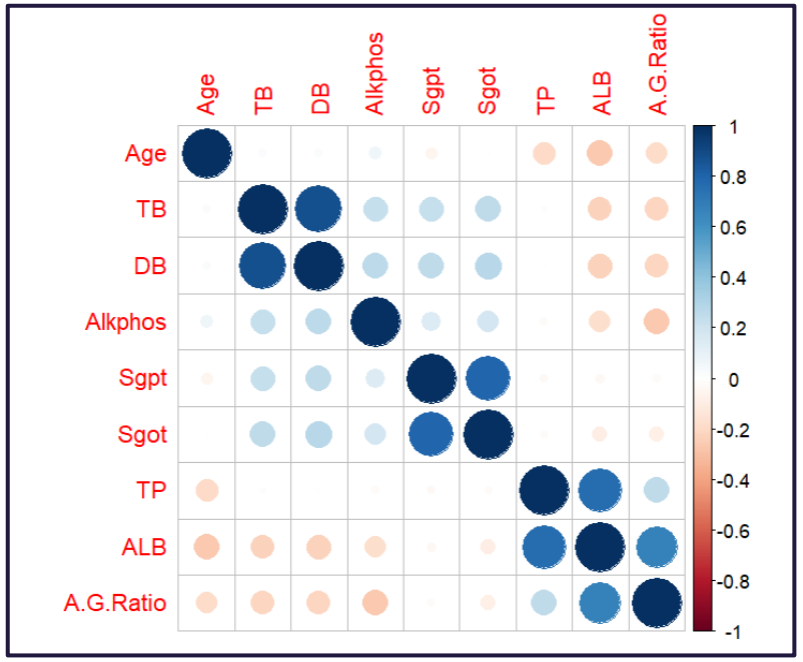


PRE-PROCESSING

1. Missing values : only 4, removed using "*na.omit(df)*" function
2. Duplicate entries : 13, removed using "*distinct(df)*" function
3. Class balancing : Using "SMOTE-NC"
   - Before (1:2 = 71%:29%)
   - After (1:2 = 1:1)

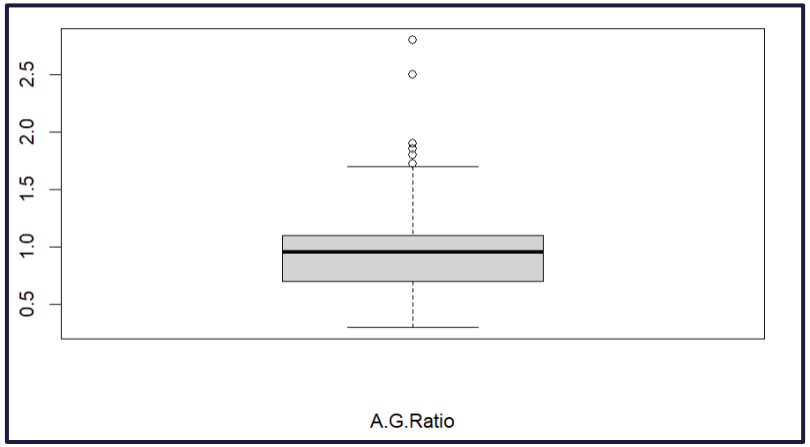
```
      Age          Gender          TB              DB              Alkphos          Sgpt            Sgot
 Min.   : 4.00   Female:142   Min.   : 0.400   Min.   : 0.100   Min.   :  63.0   Min.   :  10.00   Min.   :  10.0
 1st Qu.:33.00   Male  :441   1st Qu.: 0.800   1st Qu.: 0.200   1st Qu.: 175.5   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
```

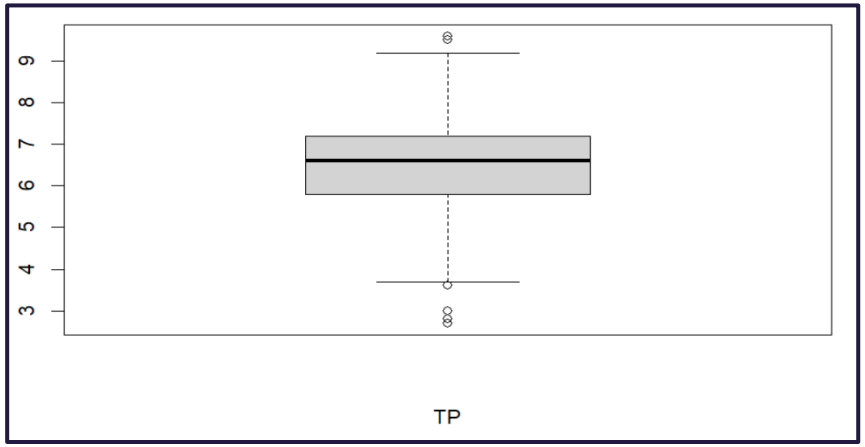
```
      Age.V1          Gender          TB.V1              DB.V1             Alkphos.V1          Sgpt.V1
 Min.   :-2.4514073   Female:167   Min.   :-0.419419   Min.   :-0.431838   Min.   :-0.942109   Min.   :-0.361331
 1st Qu.:-0.7248726   Male  :641   1st Qu.:-0.350707   1st Qu.:-0.390930   1st Qu.:-0.453906   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
```

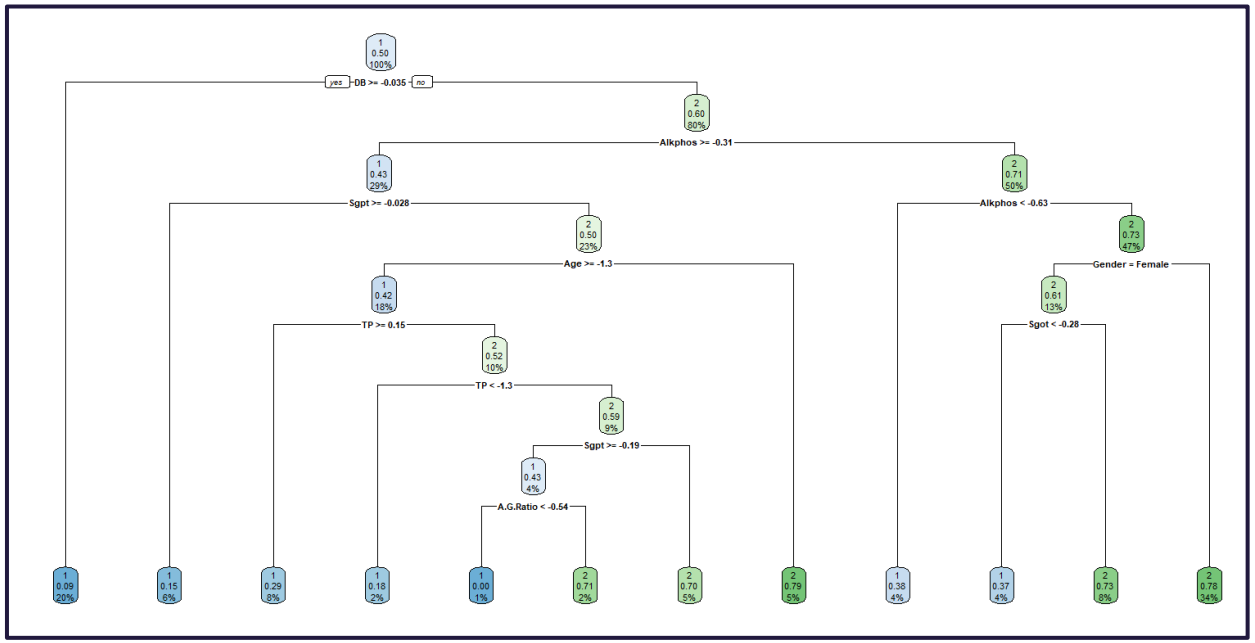


PRE-PROCESSING

1. Scaling : done using the function "*scale(df)*"
2. Multicollinearity : one pair of variables (DB & TB) showed a correlation of 88%, One of them can be removed
3. Outlier detection : Done using boxplots (example given below, for more, check appendix (D) of report)







CT		Actual	
		1	2
Prediction	1	51	16
	2	29	64

MODELLING

Classification Tree

- Supervised classification algorithm
- Tree like structure with root node, internal nodes and leaf nodes (target classes)
- Highly tolerant towards outliers
- Non-Parametric
- Visualization
- Confusion matrix

MODELLING

SVM		Actual	
		1	2
Prediction	1	46	6
	2	34	74

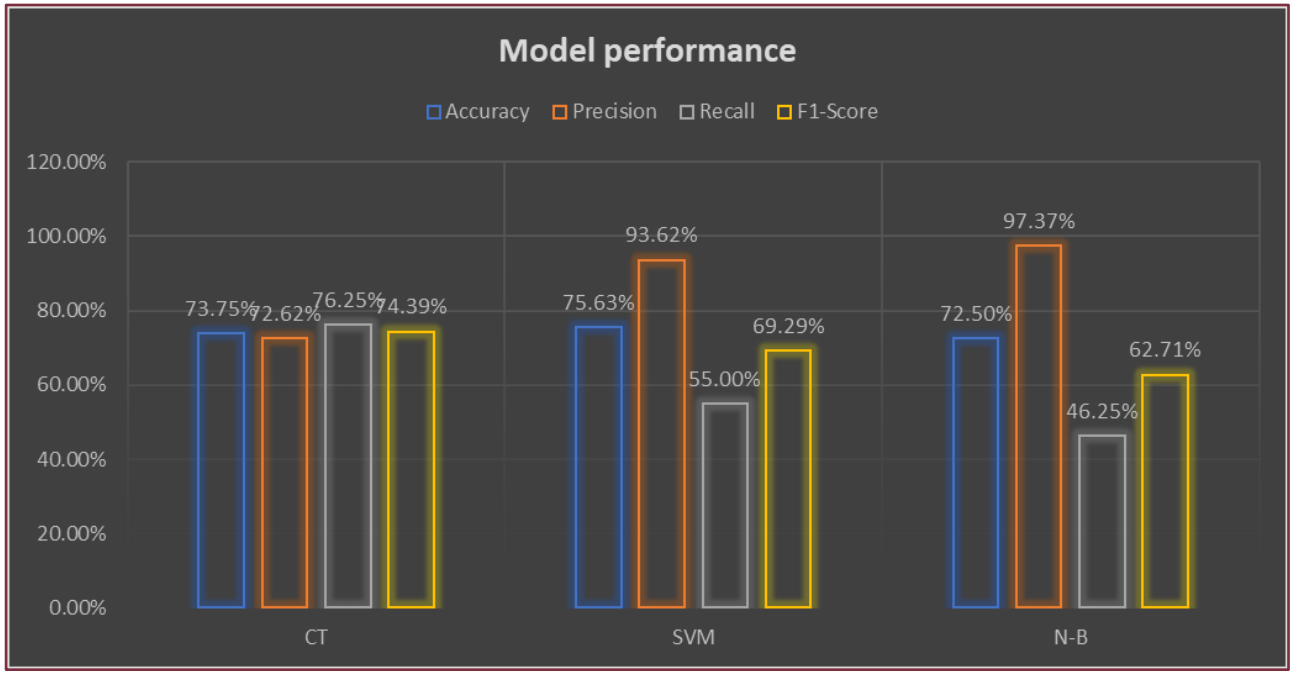
Support Vector Machine

- Supervised classification algorithm
- Chooses hyperplane with the best margin between the boundaries and closest data points (Support vectors)
- Highly tolerant towards outliers
- Confusion matrix

N-B Model		Actual	
		1	2
Prediction	1	37	2
	2	43	78

Naïve-bayes Classification

- Supervised classification algorithm
- Classifies data points based on bayes theorem.
- Highly affected by outliers
- Chosen to compare the performance
- Confusion Matrix



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

RESULTS & DISCUSSION

Performance summary

- SVM performs the best in terms of accuracy
- Is 'accuracy' the best measure of performance in this context?
- Accuracy = (TP + TN)/ (TP+TN+FP+FN)
- Precision = (TP)/(TP+FP)
- Recall = (TP)/(TP+FN)
- F1Score= (2*Precision*Recall) / (Precision + Recall)
- Needs to reduce FN as much as possible
- Recall is the most suitable, CT performs the best

- Try other methods like ANN
- Trye SVMs with different Kernal functions
- Gather data for alcohol consumption

FUTURE RECOMMENDATIONS

THANK YOU

Akshay Kumar VT

[redacted]

[redacted]