PREDICTIVE ANALYTICS FOR LIVER DISEASE

A BUSINESS ANALYTICS REPORT

BY HARSINI SATHYAMURTHI

EXECUTIVE SUMMARY

Report Objective

In order to investigate the connections between liver function tests and patient health outcomes, this research analyzes the Liver Patient Dataset. In order to effectively detect the existence of liver disease and enhance healthcare decision-making through predictive analytics, machine learning models such as Decision Trees, Bagging, and Random Forests are being developed.

Methodology

Descriptive statistics were used analyse the distributions of key features and data quality. To create machine learning models, decision trees, bagging, and random forests were used. The model's performance was evaluated using ROC curves and AUC values, and Random Forest provided information on the importance of every feature.

Business Insights

Important indicators of liver function, including ALT, AST, and total bilirubin, have been found to be reliable indicators of liver illness. The most reliable and accurate classification results were obtained with Random Forest. By emphasizing important diagnostic indicators and encouraging early liver disease diagnosis, these findings help guide therapeutic procedures.

MODEL PERFORMANCE SUMMARY

Model Accuracy

Random Forest outperformed Decision Tree and Bagging models, achieving the best AUC and classification accuracy. This demonstrates how well the model predicts the existence of liver disease.

Model Stability

10-fold cross-validation confirmed that the Random Forest and bagging models showed improved model stability with less overfitting. Throughout the training rounds, Random Forest consistently delivered strong performance.

Key Predictors & Insights

Three important indicators of liver illness were found to be ALT, AST, and total bilirubin. These biomarkers improve liver disease risk assessment and offer useful information to promote early therapeutic treatments.

Business Recommendations

Adopt Random Forest models for reliable liver disease classification in healthcare analytics. Use insights from feature importance to inform clinical decision-making and prioritize critical liver function indicators.

Table of Contents

Introduction

Machine Learning (ML) helps to understand the patterns and present predictions from the datasets for businesses where the machines are trained by using different models. It has become essential to use **Machine Learning (ML) techniques in healthcare departments** in this modern world. It is useful when analyzing complex clinical data, as it supports and provides early **diagnosis of liver conditions** and related diseases. Timely intervention, made possible by accurate liver disease prediction, can greatly enhance patient outcomes and lower death rates.

Predictive models may be built using the liver health dataset provided in this experiment, which contains biochemical indicators related to liver function and patient demographic data. With differing mortality rates across locations and etiologies, liver cirrhosis continues to be a major worldwide health concern despite advancements in therapy (*Ye et al., 2023*).

To categorize individuals according to whether they have liver illness, this project will use **R programming** to develop several machine learning methods. To assess the efficacy of the models, important performance indicators such as accuracy, precision, true positive rate (sensitivity), and area under the **ROC curve** (**AUC**) will be used. Recent studies demonstrate the effectiveness of ensemble learning techniques, which on comparable liver datasets produced excellent prediction accuracy (*Ganie et al., 2024*). By utilizing these methods, the study seeks to determine which model is most suited to assist medical professionals in making data-driven diagnostic choices.

Data Cleaning

Before modelling, data cleaning will be an essential first step to guarantee the dataset's dependability and quality. After loading and exploring the liver dataset, names **(Liver_data) <- new_column_names** were used to **rename the columns** for clarity. The naniar package's **vis_miss()** function was used to visually and quantitatively identify missing values, and **Liver_data <- na.omit (Liver_data)** was used to eliminate rows that had missing values.

To facilitate classification, **factor()** was used to transform the target variable into a factor with relevant labels. Likewise, numeric columns were expressly designated as numeric, while categorical variables such as gender were **transformed into factors**. A clean, consistent dataset prepared for machine learning modelling was the result of this method.

Descriptive Statistics

Descriptive Statistics is an analytics technique that summarizes a complex dataset that provides the trends, understanding the outliers and ultimately gives the measure values such as the measures explaining the central limits, ranges and other measures. This procedure ensures data quality, directs feature selection, and serves as the foundation for comprehending the data before applying predictive modelling.

Summary Statistics

Eleven variables make up the liver disease dataset, including biochemical indicators like total bilirubin, ALT, AST, and albumin levels as well as demographic characteristics like age and gender. Figure 1.1 gives the summary statistics for the dataset.

```
> summary(Liver_data)
      Age            Gender     Total_Bilirubin  Direct_Bilirubin Alkaline_Phosphatase      ALT               AST
 Min.   : 4.00   Female:140   Min.   : 0.400   Min.   : 0.100   Min.   :  63.0       Min.   :  10.00   Min.   :  10.0
 1st Qu.:33.00   Male  :439   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   :44.78                Mean   : 3.315   Mean   : 1.494   Mean   : 291.4       Mean   :  81.13   Mean   : 110.4
 3rd Qu.:58.00                3rd Qu.: 2.600   3rd Qu.: 1.300   3rd Qu.: 298.0       3rd Qu.:  61.00   3rd Qu.:  87.0
 Max.   :90.00                Max.   :75.000   Max.   :19.700   Max.   :2110.0       Max.   :2000.00   Max.   :4929.0
 Total_Protein      Albumin      Albumin_Globulin_Ratio          Target
 Min.   :2.700   Min.   :0.900   Min.   :0.3000         Liver_Disease   :414
 1st Qu.:5.800   1st Qu.:2.600   1st Qu.:0.7000         No_Liver_Disease:165
 Median :6.600   Median :3.100   Median :0.9300
 Mean   :6.482   Mean   :3.139   Mean   :0.9471
 3rd Qu.:7.200   3rd Qu.:3.800   3rd Qu.:1.1000
 Max.   :9.600   Max.   :5.500   Max.   :2.8000
```

Figure 1.1 Overall Summary Statistics

To characterize the distribution and spread of these variables and to assist detect any outliers and skewed distributions, summary statistics including minimum, maximum, median, mean, and quartiles were calculated.

Data Distribution and Visualizations

Frequency tables and bar plots were used to analyze the distribution of the target variable, which shows whether liver disease is present or not. A class imbalance that must be considered in modelling is indicated by the bar plot, which showed that patients with liver illness much outnumber those without. Figure 1.1 below presents the Bar chart visualizations.

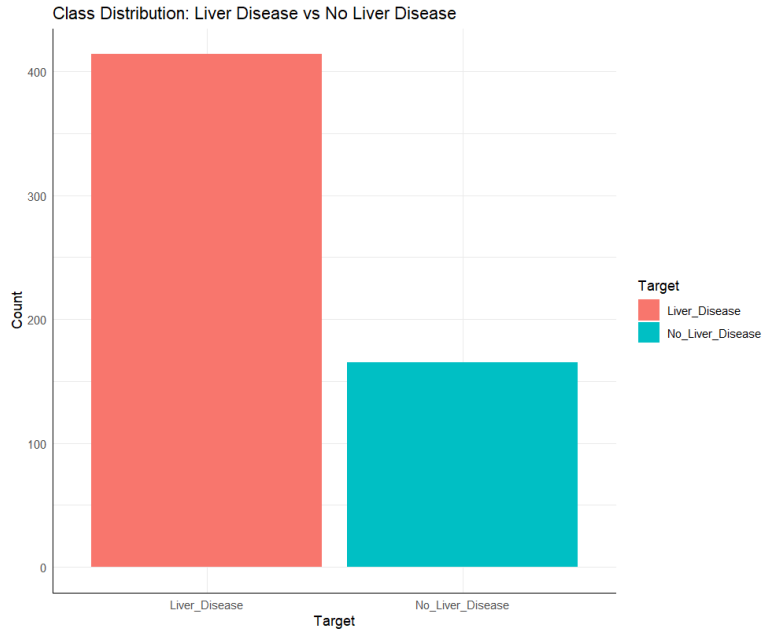


Figure 1.2 Bar chart visualizations

Boxplots were used to further investigate continuous variables. Figure 1.3 below presents the Boxplot visualizations with outliers.

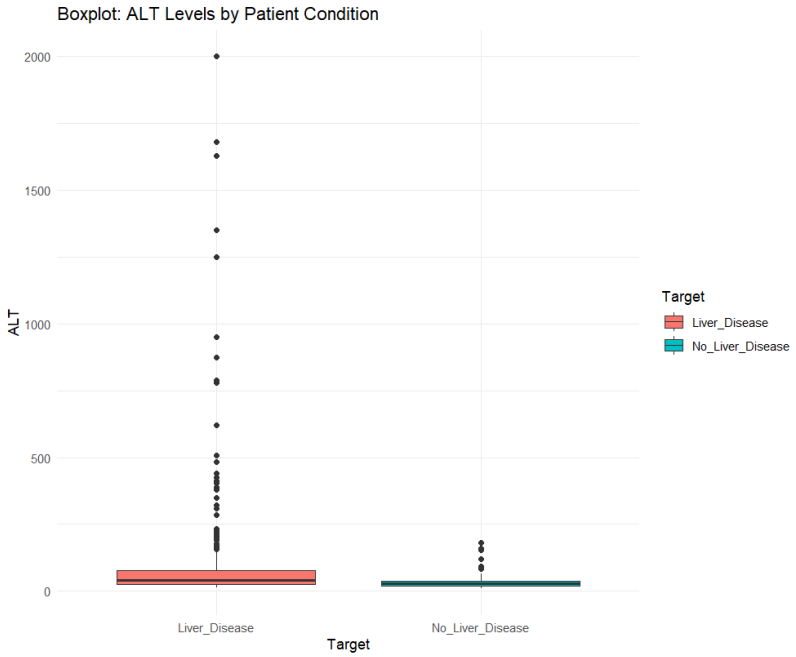


Figure 1.3 Boxplot visualizations

For instance, the ALT level boxplot demonstrated significantly higher median values and a wider dispersion among patients with liver illness than those without, underscoring the significance of this biomarker as a prediction tool.

Correlation Insights

To examine the correlations between numerical variables, a correlation matrix was computed and displayed using a heatmap. Strong positive correlations between albumin and albumin-globulin ratio, as well as between total bilirubin and direct bilirubin, were seen in the heatmap, indicating considerable redundancy between these variables. The Figure 1.4 gives the Correlation matrix for all the variables.

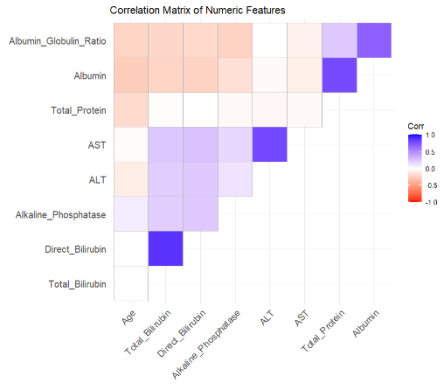


Figure 1.4 Correlation matrix

To prevent multicollinearity and enhance model interpretability and predictive performance, feature selection requires an understanding of these connections.

Methodology

The Methodology section gives the outline of **machine learning techniques** used to classify the liver disease in patients based on the biological and chemical variables and the demographics from the liver dataset. Modelling is carried into a **structured, stepwise approach** keeping both **model accuracy** and **interpretability** into consideration.

The first activity performed in the Liver dataset was **cleaning** where missing values were removed, data types were checked and factorized followed by renaming the column names for more interpretability. The **outliers** were monitored through the **visualization** plot (Boxplot) but not altered due to limitations within the calculations and difficulties. Hence assuming the outliers does not affect the results much.

The dataset was split into two where 70% of it is set into a training set and the remaining 30% into a testing set. This stratified sample is created by using **createDataPartition()** function from the **caret** package. By this way, the consistency of the proportions of the target are kept avoiding kurtosis or skewness in the samples. In assistance with the evaluation metrics given, the models are evaluated and the interpretations for the datasets were made.

Modelling Approach

The following gives the machine learning models that were developed to classify the liver disease status.

- **Decision Tree:** A baseline classifier in Decision Tree was trained using the **rpart()** algorithm. The tree was later pruned by using the **complexity parameter (CP)** plot to avoid generalizing and overfitting of the data.
- **Bagging:** Then for further improvement in the predictive performance, an ensemble bagging model was created using the **bagging()** function from the **ipred** package. This combines the predictions of the different decision trees using bootstrapped decision trees. This reduces the model variance.
- **Random Forest:** A random forest model was implemented with the help of **ranger** method via the **caret::train()** interface, with a **10-fold cross-validation**. The model was optimized to improve and maximize the Area Under the Curve (**AUC**). The variable importance (**VI**) was also inspected to give better interpretations of the model.

Model Evaluation Metrics

Model performance was evaluated on the split test set using the following metrics.

- **Confusion Matrix:** This step is used to assess the classification of accuracy, sensitivity and the specificity of the models.
- **ROC Curve and AUC:** These help in evaluating each model's ability to differentiate the two classes in the data. AUC is used for the primary comparison metric due to the effectiveness in the imbalanced classification works.

The models used were selected based on the type and size of the datasets featuring important analysis in the model generation.

Modelling and Evaluation

This Model Evaluation section presents the development and the evaluation results for each of the models developed. Visual aids by including the model plots and performance curves are included in the relevant subsections to support the interpretation and the comparison.

Decision Tree

The Decision Tree model was trained and developed using the training dataset and pruned to optimize generalization. The pruned model obtained has an **AUC** of **0.7253,** providing a clear and interpretable baseline setup.

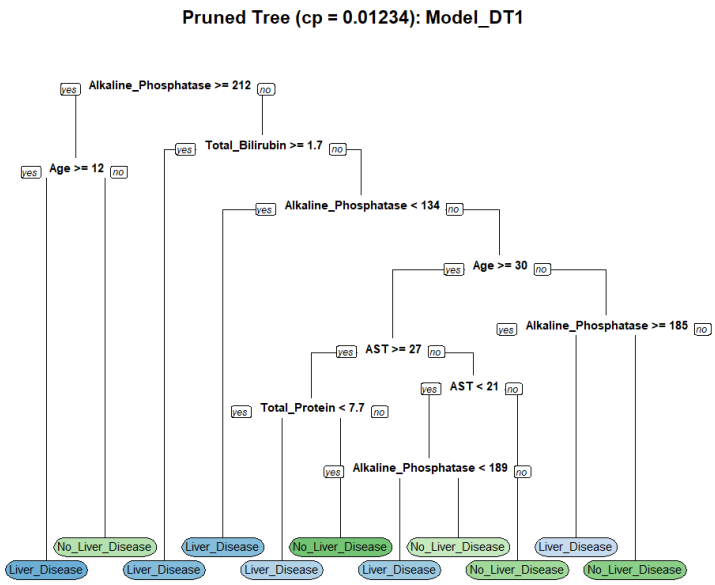


Figure 1.5 Decision Tree

Though the decision tree gives good result with excellent transparency and interpretability, its predictive power was moderate. This single tree structure increases overfitting possibilities limiting generalizing of unseen data.

Bagging

The ensemble Bagging model improves the performance of the model. This technique achieved the **AUC** of **0.08018** on the testing dataset. The figure gives the bagging curve.

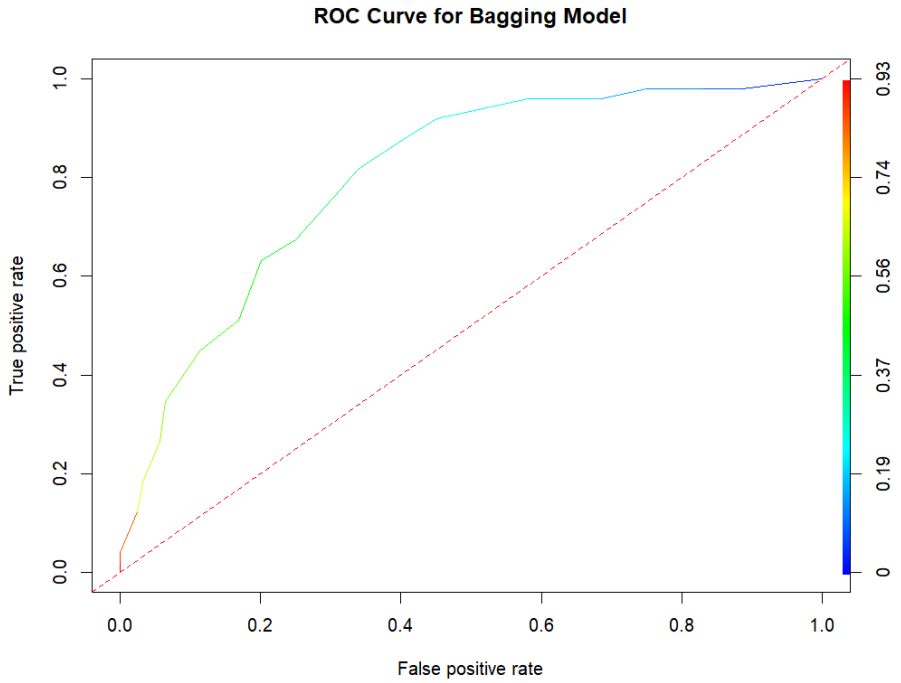


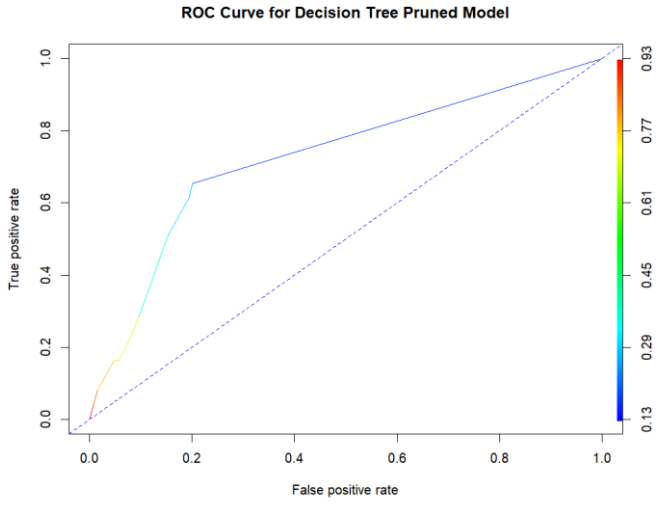
Figure 1.6 ROC curve for bagging

The Key performance metrics are these.

- **Accuracy**: 0.763
- **Sensitivity**: 0.8871
- **Specificity**: 0.4490
- **Positive Predictive Value** (**PPV**): 0.8029

Random Forest

The random forest model was implemented having 10 fold cross validation. The figure 1.7 gives the ROC curve for random forest.



The Random Forest model achieved the highest **AUC** of **0.8453** on the test set. The Key performance metrics:

- Accuracy: 0.7746
- Sensitivity: 0.9839 → Excellent ability to detect Liver Disease
- Specificity: 0.2449
- Positive Predictive Value (PPV): 0.7673

Variable importance analysis further revealed key predictors of liver disease:

- ALT (Alanine Aminotransferase)
- Alkaline Phosphatase
- Direct Bilirubin

The Random Forest model balances strong predictive accuracy with robustness and interpretability, making it the optimal classifier for this task.

Results and Interpreations

The Liver Patient Dataset was used to create and assess **three machine learning models**: **Random Forest, Bagging, and Decision Tree**. **AUC scores** for the test set and **ROC curves** were used to evaluate performance. The Decision Tree model had a lower specificity and showed indications of overfitting, but it provided a straightforward baseline. Accuracy was more evenly distributed thanks to the bagging's increased stability and decreased volatility. The **best AUC** and classification accuracy were attained by **Random Forest**, which was trained using **10-fold cross-validation**, demonstrating its exceptional generalization capacity.

The Random Forest model's feature importance analysis revealed that the most significant predictors of liver illness were ALT, AST, and total bilirubin. Since high levels of these markers are known to be indicative of liver dysfunction, these findings have clinical significance.

All things considered, the Random Forest model is suggested for implementation because of its predictability, robustness, and capacity to offer comprehensible insights via feature significance rankings - thereby facilitating data-driven medical decision-making for the treatment of liver illness.

Recommendations and Conclusions

ince the Random Forest model provides the most accurate and trustworthy findings for liver disease diagnosis, it should be used in real-world situations. To maintain the model functioning properly, it's critical to update it often with fresh patient data. When testing patients, pay particular attention to vital liver health indicators such as bilirubin, ALT, and AST. When physicians or regulators want precise explanations, use more straightforward models, such as decision trees. Additionally, monitor the model's accuracy over time by tracking its performance.

Using machine learning approaches, our research was able to predict liver illness and provide useful clinical insights from the Liver Patient Dataset. Compared to Decision Tree and Bagging methods, Random Forest produced the greatest AUC and accuracy among the models assessed, exhibiting the best predictive performance and stability. Across cross-validation

folds, Random Forest's ensemble nature consistently produced findings while successfully reducing overfitting.

Key liver biomarkers that can guide clinical screening and early intervention methods, ALT, AST, and total bilirubin, were highlighted for their predictive power by feature significance analysis. The results confirm the usefulness of machine learning models, especially Random Forest, in improving the evaluation of liver disease risk and assisting in medical decision-making.

A more precise, data-driven categorization of liver diseases may be made possible by integrating Random Forest models into clinical processes. Additional patient demographic and lifestyle data integration may be investigated in future research to improve predicted accuracy and enable individualized healthcare treatments.

References

Ganie, S. M., Dutta Pramanik, P. K., & Zhao, Z. (2024). Improved liver disease prediction from clinical data through an evaluation of ensemble learning approaches. *BMC medical informatics and decision making*, *24*(1), 160-124. https://doi.org/10.1186/s12911-024-02550-y

Ye, F., Zhai, M., Long, J., Gong, Y., Ren, C., Zhang, D., Lin, X., & Liu, S. (2022). The burden of liver cirrhosis in mortality: Results from the global burden of disease study. *Frontiers in public health*, *10*, 909455-909455. https://doi.org/10.3389/fpubh.2022.909455