IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Data Mining in analyzing Indian Liver Patient dataset for prediction of liver disease: a Naïve Bayes classifier model

[1]Asma Khatoon, [2]Sayed Azhar Sabri

[1]MTech (Computer Science and Engineering), [2]Full Stack Developer
[1]Department of Computer Science and Engineering,
[1]Rameshwaram Institute of Engineering and Technology, Affiliated to AKTU, Lucknow, India

Abstract : This study aims to detect potential liver patients with the ultimate goal of enabling early and effective treatment of these diseases through recommended liver function tests during health checkups. This research uses ILPD, a dataset of Indian liver patients consisting of 583 cases and 10 features. After data pre-processing, classification methods through the Naïve Bayes classifier technique were used to create a prediction system to obtain the diagnosis of liver disease. By achieving an accuracy of 65.57% concerning the training set and 64.70% for the testing set, it proves to be somewhat accurate in classifying liver patients and non-liver patients. These results indicate and support that the Naive Bayes algorithm is potentially a very reliable tool in diagnosis. The proposed approach provides a systematic machine-learning framework to assist physicians and healthcare professionals in early diagnosis and treatment planning, which will result in positive outcomes in the management of the liver disease.

Keywords **- Liver disease, Naïve Bayes, Indian Liver Patient Dataset (ILPD), Classification model.**

I. INTRODUCTION

This study aims at detecting probable liver patients, with the ultimate goal of enabling early and effective treatment of these namely diseases through liver function tests, as prescribed during health checkups. The research involved using the ILPD. Liver disease is a significant global health concern claiming approximately 2 million lives annually; Liver cirrhosis is responsible for 1.8% of deaths in Europe and, this figure has risen significantly over the last decades due to increased alcohol consumption rate, the proliferation of chronic hepatitis infections and the growing impact of obesity-related liver diseases [1].

Although liver disease is very serious, it is most often without identifiable symptoms that can delay early detection and effective treatment. Routine liver function tests (LFTs), which evaluates particular enzymes and proteins in the blood, are central in the early identification of liver dysfunction. However, doctors frequently underuse or misinterpret LFT results to the extent that investigations of the abnormalities are not pursued as outlined by national standards [2, 3].

Machine learning techniques in particular, data mining approaches offer some tentative guidance in improving the diagnostic process. Such methods provide an opportunity to extract meaningful patterns from large and complicated datasets where predictive models are built for liver disease detection. For this study, the focus exclusively rests on a Naïve Bayes algorithm-a very efficient and interpretable probabilistic classifier. In this research work, the Naïve Bayes Classifier is being constructed for the classification of liver patients, and it intends to build a classification model capable of effectively showing the classes "Liver Patient" and "Non-Liver Patient" that a dataset-encompassing 583 records and 10 attributes is used through the Indian Liver Patient Dataset. The attributes in the dataset, such as bilirubin levels, albumin, and enzyme counts, act as critical indicators for such a classification.

The main objectives of this research include:

DS Exploration and Pre-processing: Cleaned and analyzed to derive sensible patterns and ensure that the data used in the model training and testing is reliable.

Building a Naïve Bayes-Based Model for Prediction: An efficient classifier is constructed according to specific characteristics of the ILPD, which can predict the likelihood of liver disease.

Intending to contribute to the medical field, this research seeks to introduce a machine-led approach to liver disease predictive analysis. By employing the Naïve Bayes algorithm, this study brings to bear the importance of data diagnostic application in enhancing early detection and empowering healthcare practitioners in making informed decisions. The work done in this paper might inspire a future march toward ML applications in medical diagnostics.

II. LITERATURE REVIEW

Data mining is an immense time-honored asset for the healthcare industry, used for the determination of values from millions of records of data.

2.1 Data Mining in Healthcare

Data growth in the health-care industry is given an uptick due to changes along the information technology line that calls for research techniques to find out any relevant patterns. Data mining is a sub-space of artificial intelligence and machine learning has presented itself as an opportunity for arguments not yet concluded. In this context, data mining offers disease prediction and improved diagnostic assistance by using classification, cluster analysis, and rule-mining techniques [4, 5].

Although since the mid-1990s, data mining has been integral to healthcare data analysis together with predictive model development, this technique plays an integral role during the management of the early detection of liver diseases when an adequate interpretation of liver function tests helps with the diagnosis [4].

This section discusses the advantages of data mining concerning patient management, treatment optimization, and cost reduction in healthcare. This is achieved through classification algorithms that categorize patients based on similar health profiles, thereby allowing clinicians to provide customized care. The availability of voluminous historical data nurtures clinicians to diagnose complex cases. Studies stress that healthcare data mining enhances operational efficiencies and significantly improves the quality of life and survival rates of a patient [6].

There are major challenges of data-mining in the health-care industry. There are problems such as issues related with data integrity and a lack of decentralized data exchange among institutions, which may reduce predictive capability due to inconsistent input data. [16-19]. Research has stated that inaccurate predictions are often due to poor pre-processing of data and unsuitable choices of algorithm selected [4]. These must be addressed to ensure valid construction of effective systems.

Several studies validate the effectiveness of classification algorithms predicting diseases. Decision Trees and SVM have provided better results in accurately predicting cancer disease relapse over competing clustering techniques yielding an accuracy of 81% [7]. A recent hybrid task demonstrates that in heart disease prediction, hybrid models with Decision Trees using genetic algorithms came up with accuracies of 99.2%, excelling simpler models like KNN that weighed in at 61.39% [8].

Research ranked Random Forest models for chronic liver disease to an accuracy approx 87.48%, closely followed by Naïve Bayes' 82.65% accuracy [9], thus proving the relevance of the two in predicting liver-related conditions. Furthermore, various studies have examined the performance of classification algorithms relative to diseases, which note that Neural Networks, while effective in certain tasks, still lack the simplicity of explanation provided by Naïve Bayes models [8, 10].

2.2 Application of Naïve Bayes in Healthcare

Naïve Bayes is generally appreciated for its usefulness and simplicity in the analysis of medical datasets. Studies have shown Naïve Bayes to perform incredibly well in classification tasks, in particular, when handling large datasets with missing or noisy data. For instance, in the study regarding breast cancer survival, Naïve Bayes was found as the most accurate predictive model at 97.36%-an efficacy, surpassing both RBF Networks and J48 [11]. Likewise, it has found utility in diagnosing liver diseases, with the accuracy rate being slightly above 82.65% in predicting fatty liver disease [9].

Naïve Bayes is particularly useful for medical datasets, including liver function tests, since it can handle both categorical and continuous data directly. It helps to effectively screen for patients with liver disease when embedded in healthcare decision support systems. This, in turn, allows to focus on further investigations on patients identified through screening while streamlining the pathway through the healthcare system such that the optimal amount of specialist resources is used to treat the patients [12].

Summary

The review emphasized the importance of data mining, with special reference to classification algorithms, in the medical field. Despite the various methods, like Decision Trees and Random Forests, which have been applied for the same aim, Naïve Bayes stands out as a reliable approach to chronic liver disease prediction, as it is simple to understand and apply in practice. It is clear that through improvements in data pre-processing and techniques in model selection, Naïve Bayes stands to gain enormously in its contribution to diagnostic precision and in the streamlining of healthcare processes.

III. RESEARCH METHODOLOGY

The present research work applies the Naïve Bayes algorithm for chronic liver disease diagnosis and prognosis by use of the ILPD. The primary objective was to evaluate how efficiently Naïve Bayes acts as a probabilistic classifier for the analysis of healthcare data with respect to liver disease.

3.1 Naïve Bayes Algorithm:

Naïve Bayes is a widely used classification technique based on Bayes' Theorem. It provides for very efficient computational execution and can process quite large datasets. It can handle both categorical and continuous data, adding the flexibility from the medical analysis aspect. The term "naïve" indicates that the algorithm is based on the assumption that all features are conditionally independent given the target class, which simplifies the computations and, in practice, often produces good results [13].

Bayes Theorem:

The Naïve Bayes classifier is built upon Bayes' Theorem, and computes the posterior probability P (A|B), which is the probability of event A, given that event B has occurred. Mathematically, the theorem can be expressed as:
P(A|B) = (P(B|A)*P(A))/P(B)
Where, P (A|B) is the Posterior probability of the class (target) given the predictor (attribute). P (B|A) is likelihood, or the probability of the predictor given the class, P(A) is Prior probability of the class and P(B) is Prior probability of the predictor.
Using the framework developed through the mathematical representation of Bayes' theorem, the Naïve Bayes classifier can compute the probability that a given data point belongs to a specific class and choose the one where the probability is highest [14].
The Naïve Bayes algorithm is based on two assumptions:
The independence of features: Each of the predictor variables is conditionally independent given the target class, i.e. with respect to the class variable. Although this assumption reduces the computational complexity of the algorithm, in most real-life datasets, this assumption fails to hold [13].
Equal contribution of features: All input features contribute equally to prediction. The model does not assign different levels of importance to different attributes [13].
Despite these simplifying assumptions, Naïve Bayes has demonstrated competitive performance across various classification tasks, including medical diagnosis.

3.2 Implementation of Naïve Bayes for Liver Disease Prediction

The steps for implementing the Naïve Bayes classifier for liver disease prediction are detailed in the section below:

3.2.1 Dataset Acquisition

The Indian Liver Patient Dataset (ILPD) is used as a key dataset in this study. This includes clinical, demographic, and laboratory parameters, such as patient's age, sex, bilirubin amount, and the amount of different enzymes. These parameters serve as predictors to classify liver disease [14].

3.2.2 Data Preprocessing

One of the most crucial steps in developing a classification model is data preprocessing. The focus of our future work will be part-and-parcel of strong data detailing that drives things likely to be made ready for regression.
Filling of Missing Values: Cases with incomplete or missing data are dropped so as not to introduce inconsistencies during model training.
Feature Normalization: Continuous variable normalization is done to avoid uneven scale, which may lead to bias in predictions made by a model.
Encoding Categorical Variables: Categorical or non-numeric attributes such as gender are represented or converted into numeric form to allow the classifier to process them [14].

3.2.3 Dataset Partitioning

The dataset is segmented into two subsets: training and testing:
Training set includes 80% of the data, which will be used to train the model to learn probability distributions, and second is testing set for evaluating how effectively the model generalizes on predictions; this set is given a size of 20% of the data [14].

3.2.4 Model Training

The Naïve Bayes classifier is trained using the training datasets. During this process:
The algorithm calculated the chances of individual features occurring over various class labels based on their prior probabilities.
These calculated chances are then employed to compute posterior probabilities for classification [13].

Advantages for Naïve Bayes Algorithm in Liver Disease Prediction:

The Naïve Bayes classifier provides such a strong base for liver disease prediction because of the following advantages:
It requires little computation and has a very low running complexity, hence making it scalable for large datasets such as ILPD [13].
It can tolerate noisy data in case the data to be classified is inconsistent or has some missing values. This algorithm works with probability. It is based on probability rather than exact dependency of one feature on another [14].
Naïve Bayes algorithm can deal efficiently with any number of features, so this allows its application to complex medical diagnosis [14].
It is easy for its decisions to be interpreted. This is because it is probabilistically justifiable, allowing transparent reasoning around predictions for decision support in medicine [13].

IV. DATASET DESCRIPTION

The study takes recourse to the Indian Liver Patient Dataset (ILPD), which is a publicly available dataset commonly used for research into liver disease prediction and analysis. Clinical and laboratory data have been collected from patients suspected of suffering from liver disorders. This data is useful for training machine learning models that designed to predict chronic liver disease.
Properties of the Dataset: The ILPD contains information on 583 records of liver disease cases. The 10 attributes consist of demographic, biochemical, and clinical factors that are known to influence liver health; they include the following:
Age: The age of the patient, in years, is an important factor in assessing liver function.

Sex: The patient's gender or sex may also be considered for comparison (male vs. female), as liver disease proportion may vary according to sex.
Total bilirubin: A measure of total bilirubin in blood plasma to assess liver function.
Direct bilirubin: A bilirubin measure to what extent the liver processes bilirubin directly.
Alkaline phosphatase. Since this enzyme is associated with bile flow, abnormalities are often indicative of liver damage or bile duct obstruction.
Alanine aminotransferase: This laboratory test measures the serum enzyme ALT. The liver enzyme level rises whenever any liver cell is injured.
Aspartate transaminase: The occurrence of AST becomes elevated when damage occurs to the liver or muscle.
Total proteins: The total amount of proteins in the plasma used to assess the general health of the liver and liver function.
Albumin: This is made by the liver and is mainly decreased in cases of chronic liver disease.
Albumin/Globulin Ratio: An evaluation system based on the albumin-globulin ratio, providing further insights into liver function.
Class Label: The target variable in the data is the "liver patient status," which has, broadly, two classes:
1-Liver disease patient,
2-Non-liver disease patient.

4.1 Dataset Overview

The dataset is imbalanced since patients' data with liver disease dominate the data over non-diseased patients. This might require adopting strategies like oversampling, under sampling, or algorithm maladjustments during model training to address class imbalance. So, what is required as preprocessing must be done and will involve normalization and encoding before the data can be uploaded to the machine learning algorithm.
Missing and Noisy Data: The dataset is also accounted for its missing values and its noisy values, which become part of handling the model accuracy, while techniques like exclusion are applied since there is minimal amusingness and these values have been predefined.

4.2 Dataset Source

The ILPD dataset can be accessible via UCI Machine Learning Repository URL:https://archive.ics.uci.edu/dataset/225/ilpd+indian+liver+patient+dataset, and has become a common research dataset for analyzing organs in relation to liver diseases. Its structure and application in the real world make it suitable for machine learning work on healthcare.
On the basis of this dataset, a Naïve Bayes classifier is trained and evaluated for liver disease prediction so that reliable and easily interpretable results are produced to aid in early diagnosis and management.

4.3 Import Dataset

Here, we used the read.csv() function to load the dataset as a data-frame, as shown in Fig. 1, by providing the file path to the CSV. Fig. 2 illustrates that the dataset has been successfully inserted in RStudio. The str() function yields an overview of the dataset that includes details on total records, variable names, data types, and corresponding values.

```
#Read file
file <- "liver_file.txt"
data <- read.csv(file, sep="\t", header= TRUE)
str(data)
```

Fig.1: Import text file and read in csv format

```
> str(data)
'data.frame':   583 obs. of  11 variables:
 $ age    : int  65 62 62 58 72 46 26 29 17 55 ...
 $ gender : chr  "Female" "Male" "Male" "Male" ...
 $ TB     : num  0.7 10.9 7.3 1 3.9 1.8 0.9 0.9 0.9 0.7 ...
 $ DB     : num  0.1 5.5 4.1 0.4 2 0.7 0.2 0.3 0.3 0.2 ...
 $ alkphos: int  187 699 490 182 195 208 154 202 202 290 ...
 $ sgpt   : int  16 64 60 14 27 19 16 14 22 53 ...
 $ sgot   : int  18 100 68 20 59 14 12 11 19 58 ...
 $ TP     : num  6.8 7.5 7 6.8 7.3 7.6 7 6.7 7.4 6.8 ...
 $ ALB    : num  3.3 3.2 3.3 3.4 2.4 4.4 3.5 3.6 4.1 3.4 ...
 $ AGR    : num  0.9 0.74 0.89 1 0.4 1.3 1 1.1 1.2 1 ...
 $ Result : int  1 1 1 1 1 1 1 1 2 1 ...
```

Fig.2: View imported data

4.4 Data cleaning

Data cleaning can be used as a pre-requisite step for preparing a data-set to examine it ensure accuracy and consistency of raw data which often contain erroneous data, missing values, duplicity 'or' outliers are responsible for inaccurate conclusions. To remove duplicates prevents redundancy in handling missing values it ensures completeness either by using statistical methods or removing unusable record. Outliers if left unchecked it may create distortion in analysis so they are identified and managed using statistical methods standardization and data normalization improves its reliability so overall data cleanup enhances quality it gives correct output and better decision-making in different fields.

4.5 Remove duplicate rows

To eliminate redundancy, doublet entries are eliminated from the dataset using the distinct() function, as shown in Fig. 3. Alternatively the unique() function in R can achieve the same outcome. After this step, a total of 566 unique records remain in the dataset which will be utilized for the subsequent pre-processing steps.

```
#Unique rows from data frame
data %>% distinct()
#Remove duplicate
data1 <- data %>% distinct(.keep_all = TRUE)
```

Fig. 3: Remove duplicate data

4.6 Handling null values

Null values contained in dataset can be replaced by NA to analyze it effectively and support further study, as shown in Fig. 4. These NA values will be appropriately addressed in subsequent steps based on the project's requirements.

```
#Replace null values with NA
data1[data1 == ""] <- NA
```

Fig.4: Replace null values with NA

4.7 Dealing with Outliers

Outliers can show incorrect data or a situation where an assumption does not apply. Therefore, before using this data to build a predictive model, it must be handled appropriately (see Fig. 5). In this project, outliers below the minimum threshold are replaced with the first quantile value, while those above the maximum threshold are replaced with the third quantile value.
Therefore, first quantile, also known as the median of the bottom 25% of the dataset, while the third quantile represents the median of the top 75% of the dataset. The na.rm = TRUE option makes sure that any NA values will be removed before you calculate the quantiles. That is, NA values will be preserved in the data frame, but they will not be considered during the quantile calculation.

```
#1st Quantile for detecting min value
q1 = quantile(data1$age, probs = 0.25, na.rm = TRUE)
q1

benchmin = q1 - 1.5 * IQR(data1$age, na.rm = TRUE)
benchmin

#3rd Quantile for detecting max value
q3 = quantile(data1$age, probs = 0.75, na.rm = TRUE)
q3

benchmax = q3 + 1.5 * IQR(data1$age, na.rm = TRUE)
benchmax

data1$age = ifelse(data1$age > benchmax, q3, data1$age)
data1$age = ifelse(data1$age < benchmin, q1, data1$age)
```

Fig.5: Find and replace outliers

4.8 Missing Value Treatment

The data-set is analyzed for the presence of any missing values, as illustrated in Fig. 6. In this, the R sapply() function is employed, receiving a data frame as input and generating a vector as output. The command sum(is.na(x)) is executed to present the number of null entries in the data set.
Fig.7 presents the total number of missing values for every attribute that the dataset contains, and appropriate measures are taken to address them. Since the AGR field contains only four missing values, the corresponding rows are removed from the dataset, as illustrated in Fig. 8. After completing the data cleaning process, Fig. 9 confirms that no missing values remain in the dataset.

```
#Missing values
sapply(data1, function(x) sum(is.na(x)))
```

Fig.6: Finding missing values

```
> sapply(data1, function(x) sum(is.na(x)))
    age  gender      TB      DB alkphos    sgpt    sgot      TP     ALB     AGR  Result
      0       0       0       0       0       0       0       0       0       4       0
```

Fig.7: Missing values per feature

```
#Remove missing values
data1 <- data1[!is.na(data1$AGR),]
```

Fig.8: Remove rows with missing values

```
> sapply(data1, function(x) sum(is.na(x)))
    age  gender      TB      DB alkphos    sgpt    sgot      TP     ALB     AGR  Result
      0       0       0       0       0       0       0       0       0       0       0
```

Fig.9: View missing values after cleaning

4.9 Data Sampling

Samples are generated to derive insights about whole populations. Data sampling is a statistical approach which includes selecting, modifying, and examining a data subset to detect patterns and trends within a larger dataset. Before choosing an appropriate sampling method, the result field, which serves as the dataset's class label, is examined. The dataset is found to be imbalanced, as shown in Figs. 10 and 11, that is, there is high disparity in the distribution of different classes. The training of a model on imbalanced data can lead to biased predictions, often favoring the dominant class. Hence, this imbalance should be addressed to make sure that the model training is carried out impartially and the results are unbiased.

```
> counts <- table(data1$Result)
> prop.table(counts, margin=NULL)*100

       1        2
71.37809 28.62191
```

Fig.10: Each target class percentage

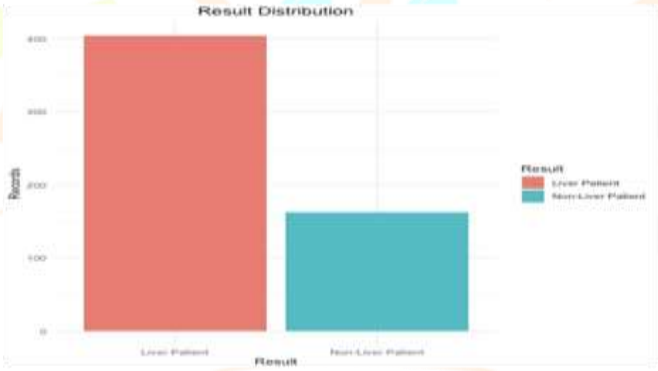

Fig.11: Bar graph of target class

4.10 Stratified Sampling

We adopted stratified sampling to balance the representation of various categories within the data. The population is divided into homogeneous strata (subgroups) to obtain equal representation across different categories[15]. There are three methods to determine sample size in stratified sampling. These methods include equal allocation, proportional allocation, and optimal allocation. This analysis takes equal allocation method: same number of samples from each stratum. A total of 566 records are drawn for this approach, where each subgroup contributes 162 records this maintains the dataset diverse enough while providing each class enough data to efficiently train the model. The stratified() function from the splitstackshape package is used, where inputs like the dataset, the name of the stratifying column, and the number of records required from each subgroup should be given (see Fig 12). Otherwise, dplyr package can be applied for stratified sampling in which group_by () is used to create grouped data and sample_n() is specified to set the sample size (see Fig 13), after sampling the target class labels were equal in their distribution as depicted in Figs 14 and 15, this balance ensured that the model trained on an equal number of liver disease cases and non-liver disease cases eliminating bias and further enhancing predictive performance.

```
data2 <-stratified(data1, c("Result"), 162)
```

Fig.12: Stratified sampling with Splitstackshape package

```
data3 <- data2 %>% group_by(Result) %>% sample_n(162)
```

Fig. 13: Stratified sampling with dplyr package

```
> counts <- table(data3$Result)
> prop.table(counts, margin=NULL)*100

 1  2
50 50
```

Fig.14: Percentage of each target class after sampling

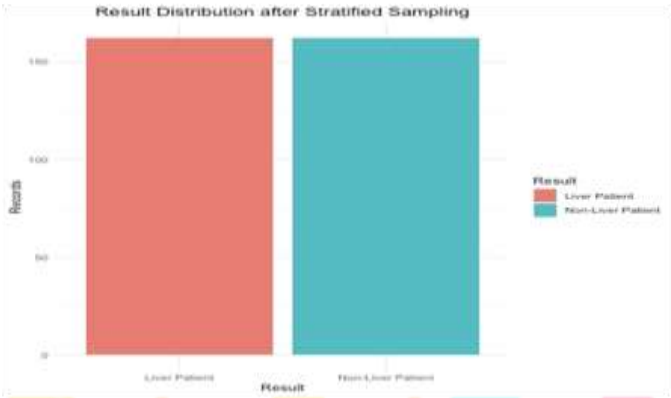

Fig.15: Bar graph of target class after sampling

V. RESULTS AND DISCUSSION

5.1 Correlation Analysis

Correlation analysis is an assessment of variable correlations in the dataset. This study provides a capability of making deductions based on the interdependencies between variables and affecting feature selection and model optimization.

In this research, correlation analysis was conducted on the features of the Indian Liver Patient Database System (ILPD) to show relations that are likely to have an implication on the prediction of liver disease risk, and attribute correlations were determined using the Pearson correlation coefficient and the results were summarized in correlation matrices.

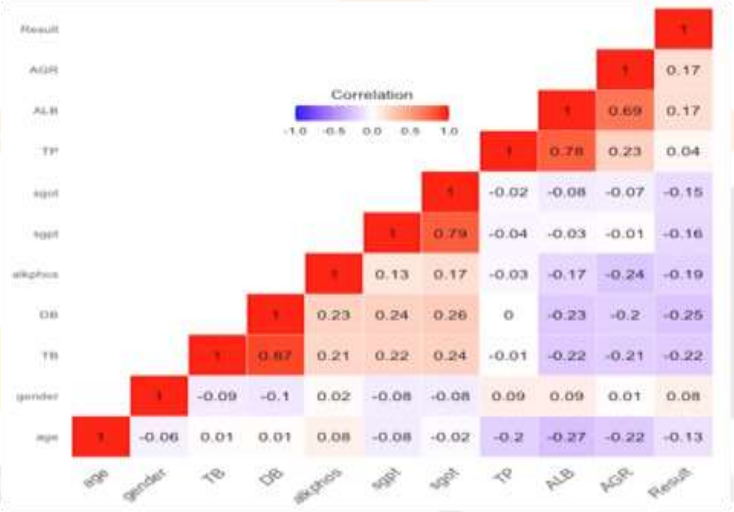

Fig.16: Correlation among attributes

5.2 Observations from Correlation Analysis

Strong Correlations:

Total Bilirubin and Direct Bilirubin: The direct bilirubin is a fraction of the total bilirubin and will therefore be expected to show a fairly strong correlation: any change in one is amplified by a varying intensity in the other.

Alamine Aminotransferase (SGPT) and Aspartate Aminotransferase (SGOT): Strong positive correlation is observed. Both enzymes are liver function markers, and usually, go up together in cases of liver damage.

Total Bilirubin and Direct Bilirubin: These two attributes show a strong positive correlation, suggesting that variations in one attribute are closely paralleled in the other. This strong correlation is understandable, as direct bilirubin is a fraction of total bilirubin.

Total Protein and Albumin: These feature also present strong correlation, as they are biologically related, for albumin is one of the major constituents of total protein.
Albumin, and Albumin and Globulin Ratio: There is a strong correlation between albumin and the albumin-to-globulin ratio in that this ratio is derived from the level of albumin.

Moderate Correlations:

Age and Albumin: The negative correlation suggests that as age increases, Albumin (ALB) tends to decrease, age could influence Albumin indirectly because older individuals might have lower liver function, reduced protein synthesis, or worse nutritional intake, which could cause lower albumin levels.
Total Bilirubin and Aspartate Aminotransferase (SGOT): Both Bilirubin and Aspartate Aminotransferase show moderate positive correlation because their levels can rise due to hepatocellular damage (damage to liver cells), so there is an expected relationship.
Direct Bilirubin and Aspartate Aminotransferase(SGOT): The moderate positive correlation means an increase in direct bilirubin also tends to increase Aspartate aminotransferase The increase in direct bilirubin is often due to liver dysfunction (such as in hepatitis or cholestasis, which is a blockage of bile flow). Under conditions that affect the liver, both DB and SGOT levels may rise because of liver damage.
Total Protein and Albumin and Globulin Ratio (AGR): Both are moderately positively correlated. The Albumin/Globulin Ratio is the ratio of albumin and globulin, two main blood plasma proteins. Total Protein is the summation of albumin and globulin. Most conditions that increase total protein will also increase albumin and globulin; therefore, they will increase the AGR.

Weak or No Correlations:

Attributes with correlation values below 0.3 show weak or negligible associations, indicating that these variables are largely independent of one another. Many of the correlations are weak but still significant, such as between gender and TP (0.092) or age and SGPT (-0.084).

5.1.2 Interpretation of Correlation Values

Strong correlation: Values of correlation that do not exceed ±1, which is usually taken to refer to a correlation value above ±0.7, may indicate a strong relationship.
Moderate correlation: Values within the range of ±0.3 to ±0.7 suggest a moderate level of correlation.
Weak or No Correlation: Values below ±0.3 indicate low or no correlation.
This analysis provides some basis for the selection of features most relevant to the prediction task while eliminating redundancy produced by highly correlated attributes. Having more useful variables in place allows for the Naive Bayes classifier to become more efficient and precise in its predictions.

5.2 Density Plot

A density plot is the commonly adopted visual illustration depicting the spread of a variable in the dataset. It fits a smooth curve, determining the probability distribution of the variable, which aids in illustrating its spread clearly. In contrast to histogram plots, density plots remain unaffected by changes in bin number or size, thereby producing consistent and interpretable shapes of the distributions.
Key takeaways can be extracted from the density plots generated for the ILPD dataset.
Age: Density distributions are slightly left-skewed. The median age, therefore, is slightly lower than the mean age. The peak of the density curve tends toward the right side with most patients clustered around the age of 40, and the density estimate is around 0.024.
Total bilirubin: It is noted that the distribution exhibits positive skewness as the mean is greater than the median. Most patients have total bilirubin values less than 1, with the density estimate peaking at something above 0.7.
Direct Bilirubin: The distribution is yet positively skewed, just like its precedent, while most patients exhibit a direct bilirubin value of under 0.5. The density estimate, in turn, reaches beyond 1.5, as the mean is greater than the median.
Alkaline Phosphatase: The ALP attribute follows a right-skewed distribution with the median smaller than the mean. Nearly all patients recorded ALP levels around 200, the density estimate for which is 0.007.
Alanine Aminotransferase: The ALT distribution is also right-skewed, with most patients' ALT variable afflicted with values below 75. Its density maximum estimate lies above 0.022.
Aspartate Aminotransferase: ASP values are observed to be right-skewed, with most patients' coming in around 40. Highest density estimate= 0.019.
Total Protein: Distribution is left-skewed, which means the median exceeds the mean. Most patients' total protein figures hover around 7, with a density estimate greater than 0.4.
Albumin: Albumin distribution shows a normal curve with no skewness, which indicates that the mean equals the median. Majority of patients have albumin values a little over 3, with a density estimate that goes beyond 0.5.
A/G Ratio: The distribution of A/G ratio is normal with no significant skewness, thus balancing the mean and median.

5.2.1 Interpretation of Peaks in Density Plots

A density plots peak signifies the highest density of values in a given range. The variables range of values is depicted on the horizontal axis whereas the vertical axis represents kernel density estimates. The area under the curve indicates the probability of a value occurring within a certain range.
The density plots illustrate various distributions of important liver function characteristics and their effects on patient health. The right skewness of the bilirubin levels indicates an abnormal level in some patients, which is vital for liver disease prediction. However, normally distributed features such as albumin suggest a stable baseline in the entire dataset. The information obtained

from density plots helps one take a more in-depth study about the dataset, modeling a more reliable preprocessing and feature selected Naïve Bayes classification model.

5.3 Visualizing the Results with sjPlot

The sjPlot package provides strong visualization tools for data analysis, leading to intuitive charts and tabular output. This section focuses on stacked bar charts as a way to show the predictive regions of the dataset with respect to target class attributes (liver patients or non-liver patients). Such visualizations bring out patterns and relationships within the data that are of relevance in predictive modeling.

5.3.1 Key Observations from sjPlots visualizations

Age Distribution (Fig.17): Instances of the target class subgroups are available for the respective age values (given target group classes, liver patients, and non-liver patients). To no one's surprise, a non-liver patient aged 85 years is actually presented in the dataset, which brings to mind that the dataset also consists of patients represented across different age brackets.

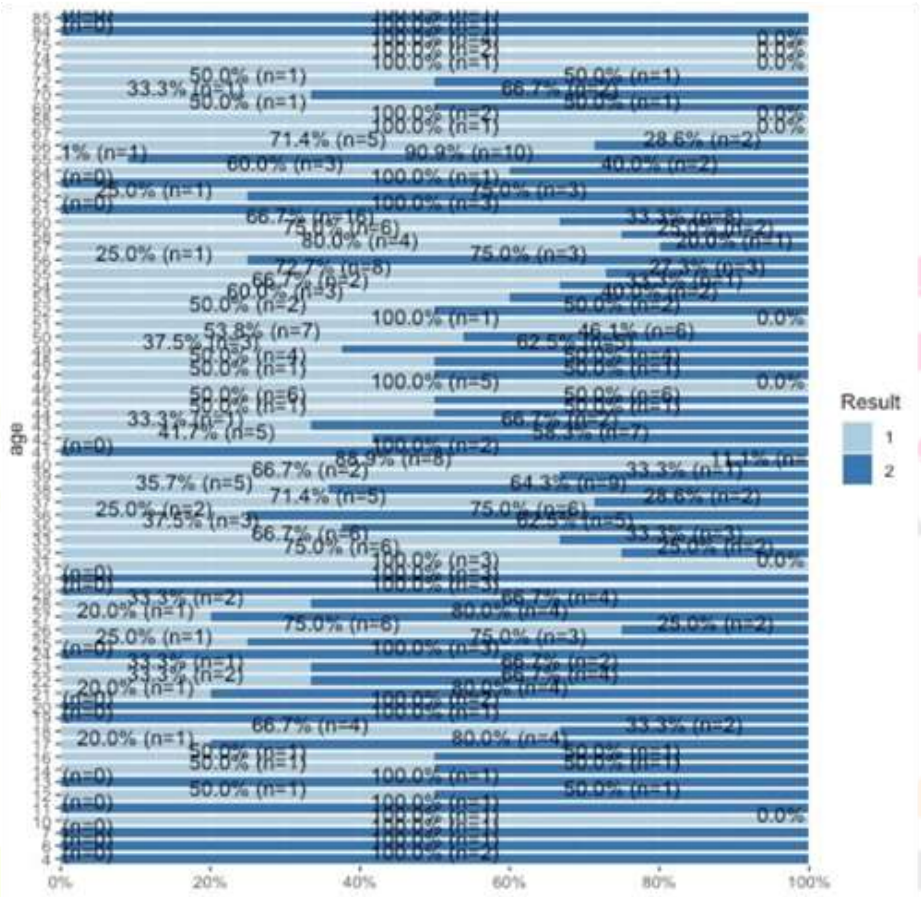

Fig.17: Stacked bar chart for Patients Age and Result

(i) Gender Distribution (Fig. 18):

The data of liver patients and non-liver patients are not equal across gender groups, depicting a higher number of female non-liver patients as compared to male non-liver patient with suspecting potential biases.

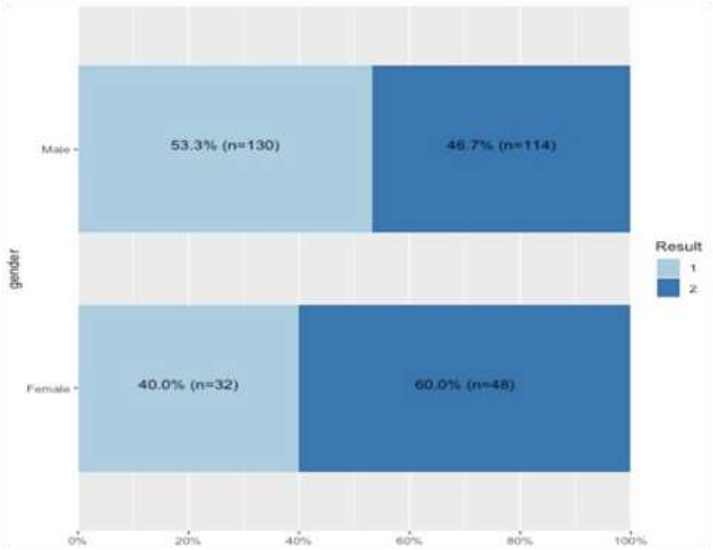

Fig.18: Stacked bar chart for Patients Gender and Result

(ii) Total Bilirubin (Fig. 19):

Key Patterns: Patients with total bilirubin values of 2.7 and above or 0.4 are 100% liver patients. Conversely, patients with a total bilirubin value of 5.3 are exclusively non-liver patients. These patterns indicate the strong diagnostic potential of bilirubin levels for liver disease prediction.

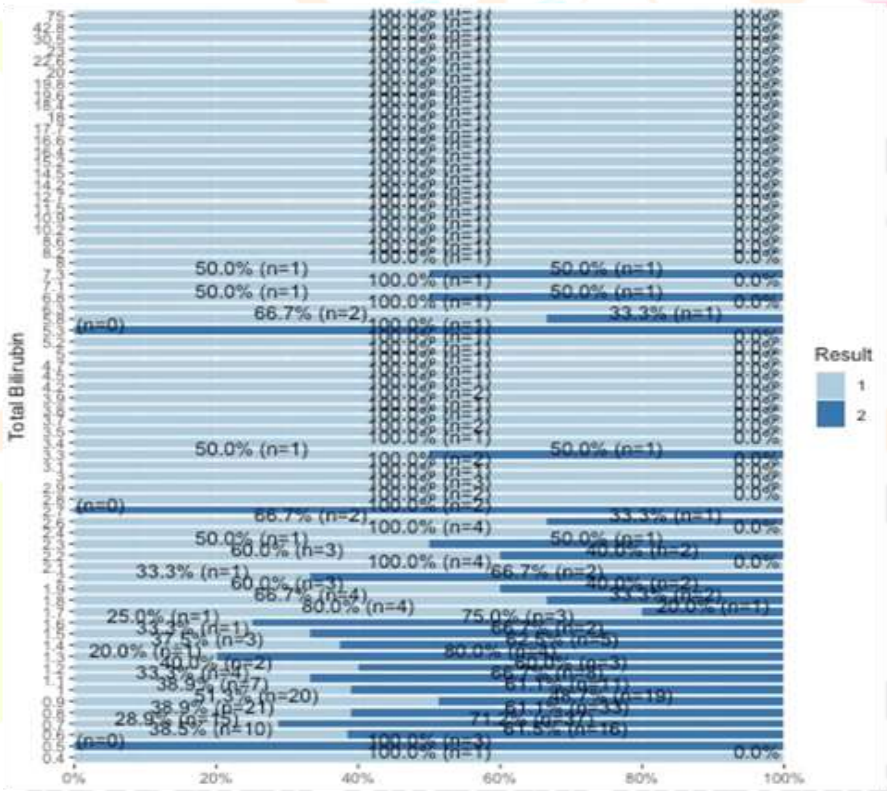

Fig.19: Stacked bar chart for Patients Total Bilirubin and Result

(iii) Direct Bilirubin (Fig. 20):

Patients having direct bilirubin levels of 1.5 or higher would be classified as pachycysts or hepatitic patients. However, an anomaly presented itself in that a direct bilirubin equal to 3.6, with 50% liver patients and 50% non-liver patients, suggesting some possible overlaps.

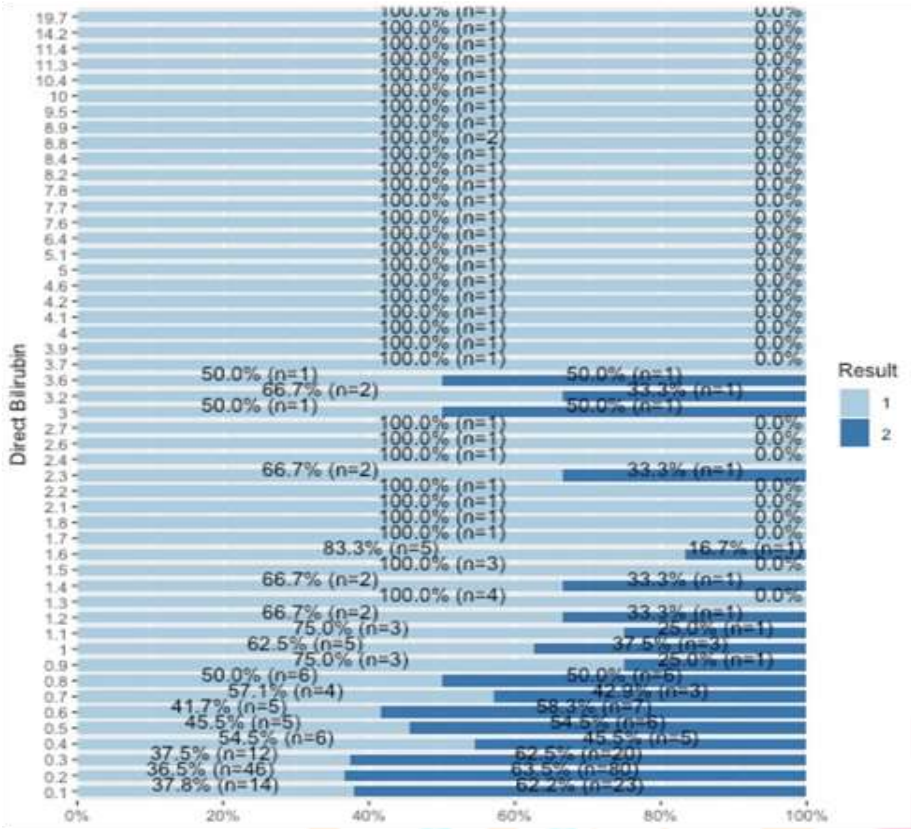

Fig.20: Stacked bar chart for Patients Direct Bilirubin and Result

(iv) Total Protein (Fig. 21):

Specific values like 2.7, 2.8, 3, 3.6 and 4 to 4.4 show 100% liver patients. On the other hand, values like 3.7 and 3.9 are entirely associated with non-liver patients. Intermediate values show records from both class subgroups, indicating variability in their diagnostic significance.

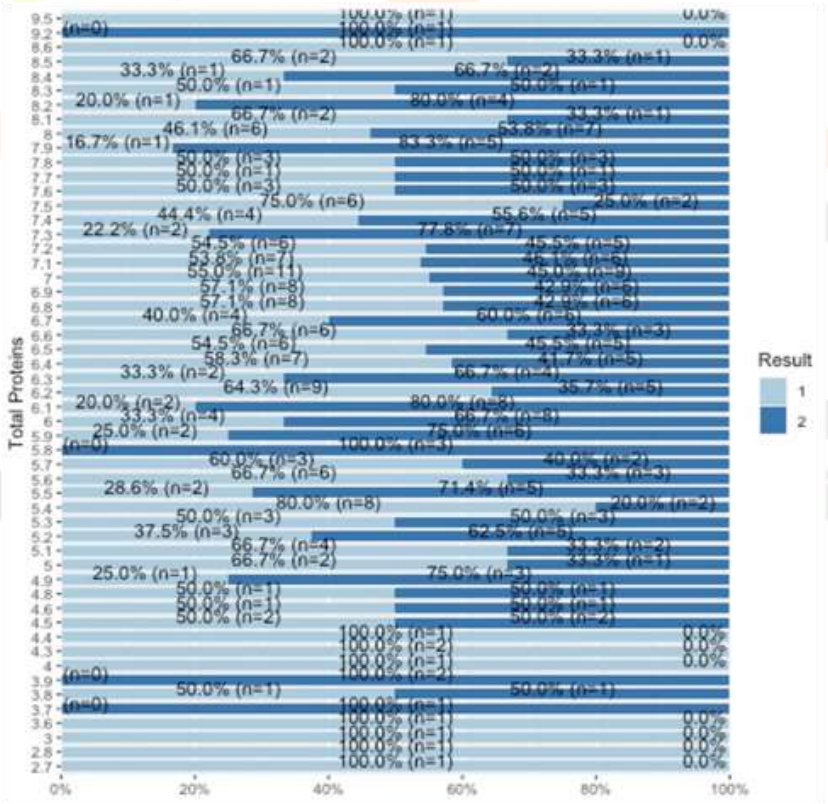

Fig.21: Stacked bar chart for Patients Total Proteins and Result

(v) Albumin (Fig. 22):

Most of the albumin values are instances by both subgroups of the target class.
Notable exceptions: Albumin values of 0.9, 1.5, and 5.5 are exclusively from liver patients-100%.

Albumin values are 4.4, 4.6, 4.7, and 5.0 for non-liver patients-100%.

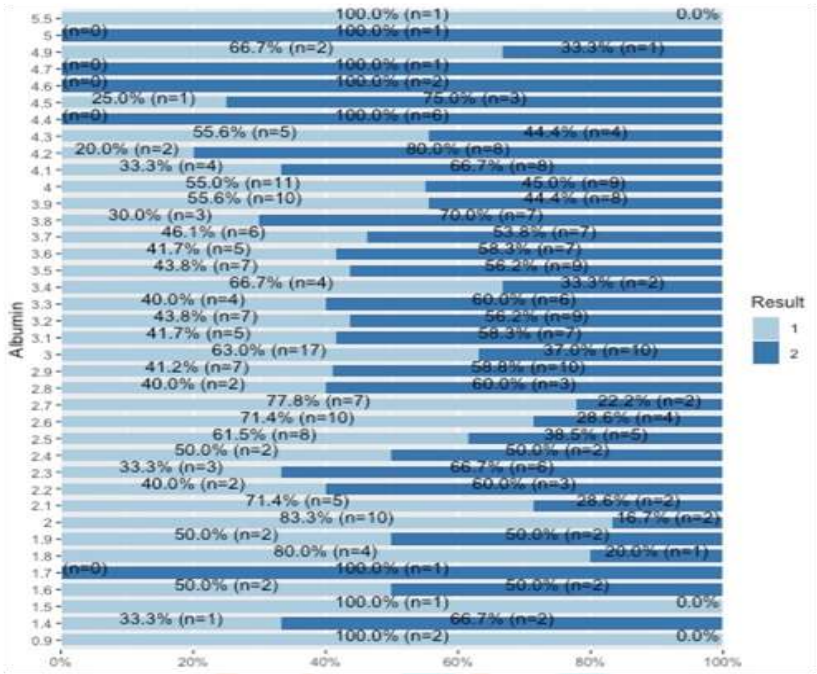

Fig.22: Stacked bar chart for Patients Albumin and Result

Insights and Implications

Stacked bar charts of sjPlot provide a good picture of attribute distributions for the dataset. Observations such as the skewed representation of certain attribute values and the patterns in the target class subgroups assist in
Identifying significant predictors like bilirubin, albumin, and total protein for liver disease classification.
Identifying potential abnormalities and overlaps in the data that will eventually direct towards more feature engineering and model improvement.
With such visualizations, effective use of the dataset can easily be understood which further substantiates the point for the use of naive bayes classification for liver disease prediction.
Naïve Bayes is Bayes theorem and an assumption-based classifier which considers predictor independence. Thus is easy to build and is particularly useful for a large database. Despite its simplicity, Naïve Bayes outshines many complex classifiers.

5.4 Application of Naïve Bayes

Model training: The naïve_bayes() function is used to fit the model on the training dataset. This function identifies the class for each feature and assumes that each feature has its own unique conditional distribution (see Fig 23).
The summary produced by the model consists of:
Number of Classes: Two (liver patients and non-liver patients).
Number of Records and Features: It denotes the structure of the dataset.
Conditional Distributions: Probability of each feature.
Prior Probabilities: Both classes are considered to have identical prior probabilities, with no bias being introduced at the start before any new data is supplied (see Fig. 24).

```
model1 <- naive_bayes(Result ~. , data = train)
```

Fig.23: Model training using Naïve Bayes Classifier

```
================================================== Naive Bayes ==========================

Call:
naive_bayes.formula(formula = Result ~ ., data = train)

---------------------------------------------------------------------------------------------

Laplace smoothing: 0

---------------------------------------------------------------------------------------------

A priori probabilities:

        1         2
0.5171103 0.4828897

---------------------------------------------------------------------------------------------

Tables:

---------------------------------------------------------------------------------------------
:: age (Gaussian)
---------------------------------------------------------------------------------------------

age           1        2
  mean 46.11765 41.81102
  sd   15.29660 17.11016

---------------------------------------------------------------------------------------------
:: gender (Bernoulli)
---------------------------------------------------------------------------------------------

gender           1         2
  Male   0.8235294 0.7244094
  Female 0.1764706 0.2755906

---------------------------------------------------------------------------------------------
:: TB (Gaussian)
---------------------------------------------------------------------------------------------

TB            1        2
  mean 3.853676 1.167717
  sd   8.516954 1.052124

---------------------------------------------------------------------------------------------
:: DB (Gaussian)
---------------------------------------------------------------------------------------------

DB            1         2
  mean 1.5191176 0.4047244
  sd   2.9001536 0.5475571

---------------------------------------------------------------------------------------------
:: alkphos (Gaussian)
---------------------------------------------------------------------------------------------

alkphos        1        2
   mean 325.0662 222.0709
   sd   295.6707 154.7849

---------------------------------------------------------------------------------------------

# ... and 5 more tables

---------------------------------------------------------------------------------------------
```

Fig.24: Summary of Naïve Bayes Model

Prediction: The model's prediction for the training dataset with respect to the target variable (Result) was achieved with the help of the predict() function using the designated features.
Actual labels were used to perform a cross-tabulation analysis comparing predictions with true classes, hence showing whether or not they were correctly classified (see Fig. 25).

```
prediction1 <- predict(model1, train)
(crstab1 <- table(prediction1, train$Result))
train1 = sum(diag(crstab1))/sum(crstab1) * 100

prediction2 <- predict(model1, test)
(crstab2 <- table(prediction2, test$Result))
test1 = sum(diag(crstab2))/sum(crstab2) * 100
```

Fig.25: Accuracy check of Naïve Bayes Model

Model Evaluation: The Naïve Bayes model was evaluated for its accuracy by testing both the training and testing data on it:
Training Dataset Accuracy: 65.57% (Fig. 26).
Testing Dataset Accuracy: 64.70%.
These accuracies reflect reasonably good predictive information since the model learned the structure in the data and remained constant on both datasets.

```
> cbind(train1, test1)
      train1   test1
[1,] 65.57971 64.70588
```

Fig.26: Prediction accuracy on train set and test set

Insights and Implications

Strengths of Naïve Bayes: Ease and effective handling of larger datasets. The capability to accurately classify data with minimal assumptions about feature independence.
Performance considerations: Results are about 65% accuracy which is moderate, the model's strength and simplicity make it a great choice for initial analysis and quick predictions.
Some improvements in that aspect can be made where some feature engineering strategies or further preprocessing steps may lead to enhanced model efficiency.
The Naïve Bayes algorithm is highly reliable and computationally efficient, when predicting liver disease with the ILPD dataset. Its capacity to produce something tell-tale, even when working with simplistic assumptions, showcases its worth in medical data analysis along with those in other domains.

VI. CONCLUSIONS

This study explores a machine learning-driven approach to predict liver disease. The Naïve Bayes-based algorithm is used to create a model which achieved an accuracy of 65.57% during training and 64.70% during testing. The experimental results confirm that as a robust and effective tool the proposed model offers a reliable solution for the early detection of liver disease in healthcare environment.

VII. AUTHOR STATEMENTS

We would like to express my heartiest gratitude to Samar Fatma ma'am for their invaluable help and guidance during this study. This study utilizes publicly accessible data for research purpose. Authors declare equal rights in this paper. Asma Khatoon conceived the study, performed the computational analysis, interpreted the results and wrote the paper. Sayed Azhar Sabri supervised the analysis, contributed to the result analysis and reviewed the paper. This study is not related to either human or animal use.There is no conflict of interest in this study.

REFERENCES

[1] Blachier, M., Leleu, H., Peck-Radosavljevic, M., Valla, D. C., & Roudot-Thoraval, F. (2013). A review on the epidemiological burden of liver disease in Europe. Journal of Hepatology, 58(3), 593-608.
[2] Macpherson, I., Nobes, J. H., Dow, E., Furrie, E., Miller, M. H., Robinson, E. M., & Dillon, J. F. (2020). Enhancing patient outcomes in liver disease through intelligent liver function testing. The Journal of Applied Laboratory Medicine, 5(5), 1090-1100.
[3] Standing, H. C., Jarvis, H., Orr, J., Exley, C., Hudson, M., Kaner, E., & Hanratty, B. (2018). Primary care perspectives on early detection of liver disease: A qualitative study. British Journal of General Practice, 68(676), e743-e749.
[4] Househ, M., & Aldosari, B. (2017). Potential risks associated with data mining in healthcare. Informatics Empowers Healthcare Transformation, 80-83.
[5] Ahmed, K. P. (2017). Comparative analysis of data mining tools for disease prediction. Journal of Pharmaceutical Sciences and Research, 9(10), 1886-1888.
[6] Kumar, S. R., Gayathri, N., Muthuramalingam, S., Balamurugan, B., Ramesh, C., & Nallakaruppan, M. K. (2019). Processing and mining medical big data for e-healthcare. In Internet of Things in Biomedical Engineering (pp. 323-339). Academic Press.
[7] Ojha, U., & Goel, S. (2017, January). Predicting breast cancer recurrence using data mining techniques. In 2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence (pp. 527-530). IEEE.
[8] Almarabeh, H., & Amer, E. (2017). Evaluating accuracy in healthcare predictions using data mining techniques. International Journal of Computer Applications, 168(3), 12-17.
[9] Wu, C. C., Yeh, W. C., Hsu, W. D., Islam, M. M., Nguyen, P. A. A., Poly, T. N., & Li, Y. C. J. (2019). Using machine learning algorithms to predict fatty liver disease. Computer Methods and Programs in Biomedicine, 170, 23-29.
[10] Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identifying significant features and applying data mining techniques for heart disease prediction. Telematics and Informatics, 36, 82-93.
[11] Chaurasia, V., Pal, S., & Tiwari, B. B. (2018). Classifying benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology, 12(2), 119-126.
[12] Hashi, E. K., Zaman, M. S. U., & Hasan, M. R. (2017, February). Designing an expert clinical decision support system for disease prediction using classification techniques. In 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 396-400). IEEE.
[13] Prabhakaran, S. (2018). Understanding the working of the Naïve Bayes algorithm: A comprehensive guide with examples and code.
[14] Chauhan, N. S. (2022). A detailed overview of the Naïve Bayes algorithm. KDNuggets. Retrieved from https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html.
[15] Kim, Y. J., Oh, Y., Park, S., Cho, S., & Park, H. (2013). Utilizing data mining for stratified sampling design. Healthcare Informatics Research, 19(3), 186-195.
[16] Kandati, D. R., & Gadekallu, T. R. (2022). Federated learning for COVID-19 detection using genetic clustering. Electronics, 11(17), 2714.
[17] Rehman, M. U., Shafique, A., Ghadi, Y. Y., Boulila, W., Jan, S. U., Gadekallu, T. R., & Ahmad, J. (2022). A novel chaos-based privacy-preserving deep learning approach for cancer diagnosis. IEEE Transactions on Network Science and Engineering, 9(6), 4322-4337.
[18] Yang, Y., Wang, W., Yin, Z., Xu, R., Zhou, X., Kumar, N., & Gadekallu, T. R. (2022). AI bots for COVID-19 combat: Mixed game-based AoI optimization. IEEE Journal on Selected Areas in Communications, 40(11), 3122-3138.
[19] Pandya, S., Gadekallu, T. R., Reddy, P. K., Wang, W., & Alazab, M. (2022). InfusedHeart: An innovative knowledge-infused framework for cardiovascular event diagnosis. IEEE Transactions on Computational Social Systems.