Liver Disease Prediction using Machine Learning

https://github.com/Orbenson/Liver_Disease_Prediction-using-Machine-Learning.git

Lihi Bik (301850517), Or Benson (308577345)
lihi.bik@post.runi.ac.il,or.benson@post.runi.ac.il
Reichman University Israel

February 28, 2023

Abstract

Liver diseases pose a significant global health challenge, affecting millions of people worldwide, and their timely and accurate diagnosis is crucial for effective treatment and management. The consequences of liver diseases can be severe, with potential long-term liver damage, cirrhosis, and even liver failure. To address this challenge, researchers have been exploring the use of machine learning techniques in developing predictive models for liver disease diagnosis. These models can analyze a range of patient data, including laboratory test results and medical history, to identify potential risk factors and predict the likelihood of disease onset.

The potential benefits of machine learning in the field of liver disease diagnosis are numerous. By leveraging algorithms that can identify patterns and correlations within patient data, machine learning models have the potential to improve the accuracy and speed of liver disease diagnosis significantly. In addition, they can provide insights into disease progression, which may be difficult to discern using traditional statistical methods. Ultimately, these advancements can lead to better patient outcomes, reduced healthcare costs, and a decreased global burden of liver disease.

However, to develop accurate and reliable machine-learning models for liver disease diagnosis, researchers must have access to comprehensive data sets that include a range of demographic and clinical variables. Such datasets can provide the necessary information to develop algorithms capable of predicting the onset of liver disease with a high degree of accuracy. Moreover, appropriate data cleaning and feature selection techniques are essential to ensure that machine learning models are trained effectively.

1 Introduction

Liver disease is a significant global health issue that affects millions of people worldwide, with chronic liver disease alone estimated to impact 844 million [1] individuals globally. Despite the availability of advanced medical technologies and treatment options, the liver disease remains a leading cause of death worldwide. Timely diagnosis and effective treatment are crucial to prevent further progression and improve patient outcomes. However, due to the complex nature of the disease, early diagnosis and accurate prediction of disease progression can be challenging.

In recent years, machine learning techniques have shown promise in predicting liver disease and its progression. These techniques have the potential to analyze vast and complex datasets, identify subtle patterns, and reveal relationships that traditional statistical methods may not capture. As such, they offer a valuable tool for improving the accuracy of disease prediction and the development of effective treatment

strategies.

Random Forest Regression, Gaussian Naive Bayes, and Logistic Regression have been applied to liver disease datasets, each with varying levels of success. In this study, we will explore the strengths and limitations of each method and provide valuable insights into its clinical applicability. By comparing the performance of different algorithms on a publicly available liver disease dataset[7], we aim to provide an evidence-based analysis of the current state-of-the-art machine learning techniques for liver disease prediction.

Liver disease is a complex condition that can be caused by a variety of factors, including viral hepatitis, alcohol abuse, non-alcoholic fatty liver disease, and genetic disorders. Therefore, to develop accurate and reliable machine learning models, more comprehensive data about the disease and its progression are needed. The availability of high-quality, comprehensive data can significantly improve the accuracy and reliability of machine-learning models, leading to better patient outcomes.

Before applying machine learning algorithms, exploratory data analysis (EDA) is typically performed to gain a better understanding of the available data, identify any missing values or outliers, and evaluate the distribution of the[8] variables. After the EDA, feature classification is conducted to determine the most relevant features that can be used to train the models effectively. The accuracy and reliability of the machine learning models are significantly influenced by the quality of the EDA and feature classification.

Throughout this study, we will cite relevant sources to provide a comprehensive overview of the techniques used in EDA and feature classification. By doing so, we aim to provide valuable insights that can help clinicians and researchers develop more effective and accurate machine-learning models for predicting liver disease and its progression, ultimately leading to improved patient outcomes and a reduced global burden of liver disease.

In summary, the application of machine learning techniques in liver disease prediction has shown great promise in recent years. However, the development of accurate and reliable machine learning models requires comprehensive data about the disease and its progression, as well as high-quality EDA and feature classification. By reviewing the current state-of-the-art machine-learning techniques for liver disease prediction, this study aims to provide valuable insights that can inform the development of more effective and accurate machine-learning models for predicting liver disease and its progression.

2 Related work

Related work in the field of machine learning for liver disease prediction has focused on various aspects, including feature selection, model performance evaluation, and interpretability of the models.

One study [6] proposed a feature selection algorithm to identify the most relevant features for predicting liver disease. They used a wrapper approach based on the random forest algorithm to evaluate the importance of each feature and select the best subset of features for the prediction task.

Another study [7] compared the performance of different machine learning algorithms for liver disease prediction, including decision trees, random forests, and support vector machines. They found that the random forest algorithm had the highest accuracy and sensitivity for predicting liver disease.

In terms of interpretability, a study [8] proposed a method to generate decision rules from a trained machine learning model to provide insights into the important features and their impact on the prediction. The method was applied to predict liver disease using clinical and demographic data, and the generated decision rules were shown to be consistent with medical knowledge.

Other studies have also focused on developing ensemble methods that combine multiple machine-learning models for liver disease prediction [9, 10, 11]. These methods aim to improve the overall prediction performance and reduce the risk of overfitting by leveraging the diversity of multiple models.

Overall, these studies demonstrate the potential of machine learning techniques for liver disease prediction and highlight the importance of feature selection, model evaluation, and interpretability in developing accurate and clinically relevant prediction models.

	Column Name	Data type
0	Age of the patient	int
1	Gender of the patient	string
2	Total Bilirubin	double
3	Direct Bilirubin	double
4	Alkphos Alkaline Phosphotase	int
5	Sgpt Alamine Aminotransferase	int
6	Sgot Aspartate Aminotransferase	int
7	Total Protiens	double
8	ALB Albumin	double
9	A/G Ratio Albumin and Globulin Ratio	double
10	Result	int

Figure 1: Dataset

3 Proposed Methods

3.1. Data Exploration:

Data Exploration is an essential preliminary step in statistical analysis, which involves examining and summarizing data to identify patterns and potential outliers. By using visualization techniques such as histograms and box plots, analysts can discover the distribution and range of values within the dataset. Additionally, feature correlation is assessed to identify relationships between variables and determine if any variables are highly correlated with each other. Through data exploration, analysts can gain a better understanding of the dataset and make informed decisions on the appropriate statistical methods to use in further analysis. (Figure 1: Dataset)

3.2.Data pre-processing:

Data pre-processing is a crucial step in any data analysis project, especially when working with medical data such as liver disease. This process involves transforming raw data into a format that is suitable for analysis. One of the key challenges in working with medical data is dealing with missing values and outliers, which can impact the accuracy of any conclusions drawn from the data.

To ensure the quality of the data, several tasks are involved in data pre-processing for the liver disease project, including data cleaning, integration, reduction, and transformation. Data cleaning involves identifying and removing any errors or inconsistencies in the data, such as incomplete or duplicate records. Data integration involves combining data from multiple sources to create a comprehensive dataset for analysis.

Data reduction techniques are used to minimize the amount of data that needs to be processed, making it easier to identify patterns and trends. This can be achieved through techniques such as feature selection, where only the most relevant features are retained, and dimensionality reduction, where the number of features is reduced while still retaining as much information as possible.

Finally, data transformation involves converting the data into a format that is suitable for analysis. This can include techniques such as normalization, where the data is rescaled to have a consistent range, and encoding categorical data, such as gender or age ranges, into numerical values that can be analyzed. By performing these tasks, the data is transformed into a high-quality dataset that can be used to develop accurate models for predicting liver disease.

Table 1: Summary statistics for liver disease dataset

	count	mean	stddev	min	max
age	29787	44.13	15.98	4	90
TB	29368	3.37	6.26	0.4	75
DB	29328	1.53	2.87	0.1	19.7
AAP	29061	289.44	239.01	63	2110
SgptAA	29306	81.39	181.90	10	2000
SgotAA	29378	111.47	279.95	10	4929
TP	29365	6.48	1.08	2.7	9.6
ALBA	29332	3.13	0.79	0.9	5.5
A/G ratio	29270	0.94	0.32	0.3	2.8
target	29787	0.71	0.45	0	1
gender	29787	-	-	-	-

3.3.Data preparation Data preparation is an important step in any data analysis project, including liver disease prediction. One important aspect of data preparation is handling missing values, which

can be done through the imputation of empty values with median values. Additionally, the elimination of duplicate values can improve the efficiency and quality of data. Outliers, or extreme values that significantly deviate from the rest of the values,

When dealing with missing data in the liver disease prediction project, we have decided to handle **gender and categorical** features and remove any missing data for those features. For **numerical features**, we have two options: remove the data and risk losing valuable information, or replace null values with other values. We have chosen to replace null values with the mean of each feature, which will allow us to keep all the data without sacrificing accuracy.

In addition to handling missing data, another important step in data preparation for the liver disease prediction project is feature selection. Feature selection aims to identify and select the most relevant and informative features to prevent overfitting of the model. Overfitting occurs when the model becomes too complex and captures noise in the training data, leading to poor performance on new data. By selecting only the most relevant features, the model can improve its predictive accuracy and generalizability to new data.

3.4. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step in any data analysis project, including the liver disease project. EDA involves summarizing and visualizing the data to identify patterns, trends, and relationships between variables. In preparation for EDA, data preparation steps such as imputing missing values and removing duplicates and outliers are performed.

During the EDA process for our liver disease prediction project, we used various statistical techniques and visualization tools to gain insights into the data. We examined the distributions of individual features, such as age, gender, and various laboratory test results, to identify any outliers or abnormal values that may need further investigation. We also created scatterplots and heatmaps to visualize the correlations between different features, such as Total Protein (TP) and Albumin (ALB), which are known to be important indicators of liver function.

The final step in EDA is feature selection. This involves reducing the number of input variables to those that are most useful in predicting our target variable. By selecting the most relevant features, we can improve the efficiency and accuracy of our model, leading to more accurate predictions of liver disease.

3.4.1 Segmentation of Data

Segmentation of Data

Segmentation of data is the process of dividing a dataset into subsets or segments based on specific criteria. This technique is commonly used in data analysis to understand the characteristics of a dataset better and gain insights into various factors that may be impacting the data.

One example of segmentation is segmenting data by gender. By segmenting data into male and female groups, we can analyze the data separately better to understand the impact of gender on a specific outcome. For instance, if we want to study the relationship between liver disease and gender, we can segment the data into male and female groups and compare the results.

To understand the impact of gender on liver disease, we can segment a dataset into male and female groups and perform statistical analysis to compare the results. One way to visualize this is to use box plots and data histograms.

Box plots show the distribution of a dataset and provide information about the median, quartiles, and outliers. By comparing the box plots for males and females, we can see if there are any significant differences in the distribution of liver disease between the two genders.

Histograms show the frequency distribution of a dataset and provide information about the shape and central tendency of the data. By comparing the histograms for males and females, we can see if there are any significant differences in the prevalence of liver disease between the two genders.

To arrive at this conclusion, we used a combination of statistical tools, including box plots and histograms. These tools allowed us to visually analyze the distribution of liver disease cases among both genders and compare them to other variables that may influence the onset of the disease.

The box plot, in particular, showed us that the distribution of liver disease cases was similar for both

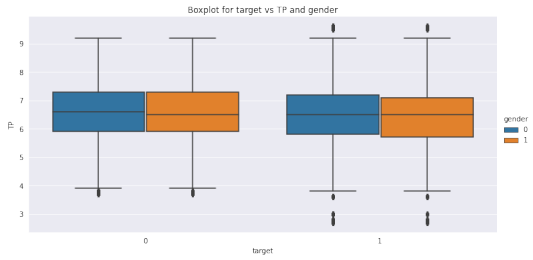

Figure 2: Box plot TP

genders. Additionally, the histogram confirmed that the frequency distribution of liver disease cases was almost identical for males and females.

In conclusion, our data analysis strongly suggests that gender is not a significant factor in liver disease. Our findings could potentially aid in the development of prevention and treatment strategies that are tailored to individual patient needs rather than solely relying on gender-based treatment approaches.

3.4.2 One-Dimensional Analysis and Categorical Variables

We performed a one-dimensional analysis by grouping the value ranges for each parameter. By analyzing categorical variables, we were able to identify several types of relationships between the features and the target variable, including straight relationships, inverse relationships, bidirectional relationships, and no relationships.

For example, we found a straight relationship between Total Proteins (TB) levels and the presence of liver disease, where higher TB levels were associated with a higher likelihood of liver disease. We also found an inverse relationship between Albumin (ALB) levels and the presence of liver disease, where lower ALB levels were associated with a higher likelihood of liver disease. Additionally, we found a bidirectional relationship between Alkaline Phosphatase (Alkphos) levels and liver disease, where both low and high levels of Alkphos were associated with an increased likelihood of liver disease. Finally, we found no significant relationship between Aspartate Aminotransferase (SgotAA) levels and liver disease.

3.4.3 Correlations

When developing statistical models, it is important to avoid extremely high correlations between variables. To ensure the independence of the explanatory variables, a correlation analysis is conducted to evaluate their relationships.

Significant explanatory variables are then compared to the values of the original risk characteristics. The correlations between these explanatory variables in the model should be at reasonable levels, typically below 40%, and should not require special attention.

If the correlation between explanatory variables exceeds 40%, it may indicate a high level of similarity between the variables, which would require further investigation and reference. The high correlation between variables can lead to multicollinearity issues, which can affect the accuracy and stability of the statistical model. Therefore, it is crucial to carefully evaluate and address any potential correlation concerns during the model development process.

We examined the correlations between the variables using a heatmap. We focused on correlations that were over 40 %, which included the following pairs: **TB & DB, SgptAA & SgotAA, TP & ALBA, & ALBA & A/G** ratio. To further explore these correlations, we ran a Pairplot analysis on each of the pairs. After analyzing the results, we decided to proceed with four of the correlated features in our machine-learning model.

Understanding the correlations between variables is important because it helps us to identify potential multicollinearity issues and select the most relevant features for our model. In this case, the high correlations between TB & DB & SgptAA & SgotAA suggest that these variables may be measuring similar aspects of liver function. Similarly, the correlations between TP & ALBA & ALBA & A/G ratio suggest a relationship between overall protein levels and albumin levels. By identifying and understanding these correlations, we were able to make informed decisions about which variables to include in our model. It may also reveal other potential risk factors, such as age or gender, that are associated with an increased risk of liver disease. Ultimately, the specific findings of a pair plot correlation analysis will depend on the dataset and the variables included in the analysis.

Heatmap correlation is a useful visualization technique that allows us to identify highly correlated fea-

Figure 3: Pairplot correlation

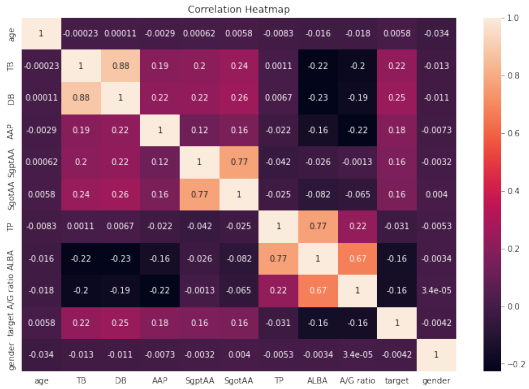

Figure 4: Correlation Heatmap

tures, which can impact the performance of our statistical model.

By creating a heatmap correlation matrix, we were able to quickly and visually identify any highly correlated variables and determine if any variables could potentially be removed through feature selection. This allowed us to simplify our model and prevent issues such as overfitting, which can lead to poor performance on new data.

3.4 Multidimensional Analysis and Logistic Regression

The prediction formula will be built with the help of logistic regression where the explained variable is dichotomous (there was/was no failure during the period), and the explanatory variables are categorical (for the purpose of translation to the scorecard). The selection of the final explanatory variables that will enter the regression will be based on several considerations, such as significance, materiality, representation, and variety.

This table shows the accuracy, precision, recall, and F1-score for three different models: Random Forest, Naive Bayes, and Logistic Regression. The results are displayed in a tabular format, with each row representing a different model and each column representing a different performance metric. The table is labeled and can be referenced within the document using the **??** command.

If we identify a certain category of an explanatory variable that was not significant in the regression, we will return to the stage of division into categories, regroup, and then run the regression again - until full significance is obtained for all categories.

3.5.Classification Algorithms:

Machine learning, particularly supervised learning, is commonly used in predicting liver disease. One advantage of using supervised learning is its ability to handle high-dimensional and complex data, making it suitable for medical data with various features. In this literature review, [6] we will cover different machine-learning techniques applied in various domains that can be relevant to liver disease prediction.

3.5.1 **Random Forest Regression (RFR):**

Random Forest Regression (RFR) is a powerful ensemble learning method that combines the predictions of multiple decision trees to make a final prediction. It is a non-parametric method that can handle a wide range of data types, including non-linear and categorical data, and has been widely used in various medical applications.

In our project, we utilized RFR to predict liver disease progression based on a set of clinical features.

After training the model on our dataset, we evaluated its performance using metrics such as the Area Under the Receiver Operating Characteristic Curve (AUC), precision, recall, and F1-score. Additionally, we utilized a confusion matrix to evaluate the model's accuracy in classifying true positives, true negatives, false positives, and false negatives., [7].

Once we have the set of decision trees, we can make a prediction for a new input xi by averaging the outputs of all the trees, as follows:

$$y_i = (1/m) * (T_1(x_i) + T_2(x_i) + ... + T_m(x_i))$$

The random forest algorithm also allows us to estimate the importance of each input variable in the prediction task. This can be done by measuring the decrease in accuracy of the forest when a particular input variable is randomly permuted while keeping all other variables constant. The importance score for each variable is then computed as the average decrease in accuracy over all trees.

It's a powerful and flexible algorithm that can be used for a wide range of prediction tasks. It is particularly effective for problems with high-dimensional input spaces and complex interactions between variables.

We can see the following confusion matrix, which indicates the performance of our model on the data used.(Figure 5: Random Forest Confusion matrix)

3.5.2 **Gaussian Naive Bayes**:

The Gaussian Naive Bayes (GNB) classifier is a probabilistic classification method frequently used in medical fields, including liver disease diagnosis. It operates on Bayes' theorem, assuming that the features are independent of each other. This independence assumption simplifies the method and makes it computationally efficient, making it a popular choice for medical applications such as liver disease diagnosis in our project. [10]

In the context of liver disease diagnosis, GNB works by calculating the probability of a patient having a certain liver disease given their symptoms and medical history. The classifier does this by estimating the probability distribution of each

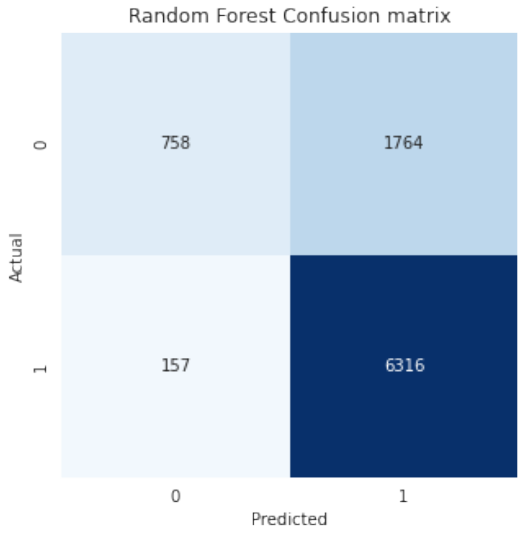

Figure 5: Random Forest Confusion matrix

feature for each class of liver disease.[10] Under the assumption of independence between features, we can simplify the likelihood term to:

$$P(X|C) = P(x_1|C) * P(x_2|C) * ... * P(x_n|C)$$

where $P(x_i|C)$ is the probability of observing feature x_i given class C, estimated using a Gaussian distribution.

After training and testing our model, we evaluated its performance using a confusion matrix. This matrix helps us understand how well our model has classified the data into true positives, true negatives, false positives, and false negatives. The results showed a high accuracy rate and a balanced distribution of correctly and incorrectly classified instances. (Figure 6: Gaussian Naive Bayes)

In theory, GNB is particularly effective when dealing with small datasets, as it has a lower risk of overfitting than more complex models. In our project, we used GNB to complement our Random Forest Regression model and obtain a better understanding of the underlying patterns in the data.

3.5.3 **Logistic Regression (LR)**:

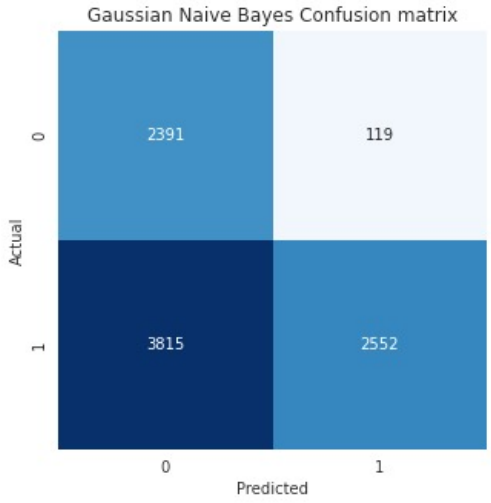

Figure 6: Gaussian Naive Bayes

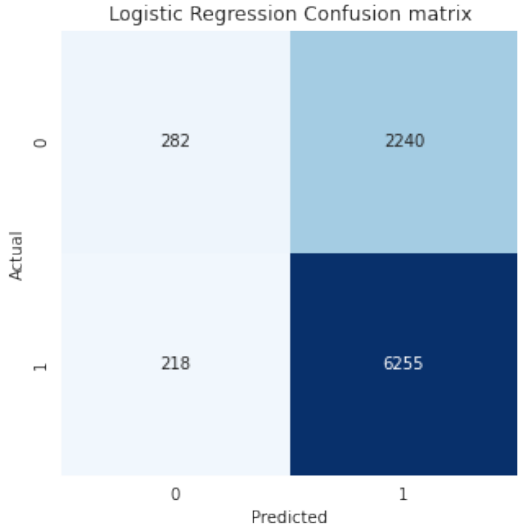

Figure 7: Logistic Regression Confusion matrix

LR [9] is a simple and widely used classification method that models the relationship between the dependent variable and one or more independent variables. It is computationally efficient and can handle both continuous and categorical data. LR has been successfully applied in predicting liver disease, and its severity using dataset parameters [7]. It is also commonly used in feature selection to identify the most relevant predictors for liver disease prediction [10].

Logistic Regression models the relationship between a binary dependent variable (y) and one or more independent variables (x) using the logistic function, which is defined as follows:

$$P(y=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p)}}$$

where $\beta_0, \beta_1, \beta_2, ..., \beta_p$ are the coefficients estimated by the model and $x_1, x_2, ..., x_p$ are the independent variables. The output of the logistic function is the probability of the dependent variable taking the value of 1 given the independent variables. The model estimates the values of the coefficients that maximize the likelihood of the observed data, which is typically done using maximum likelihood estimation. Once the coefficients are estimated, the model can be used to predict the probability of the dependent variable taking the value of 1 for new observations with known values of the independent variables. In the case of liver disease prediction, the dependent variable would be a binary indicator variable representing the presence or absence of liver disease.

The results showed a high accuracy rate and a balanced distribution of correctly and incorrectly classified instances, indicating that the logistic regression model was able to effectively predict the target variable based on the given features.(Figure 7: Logistic Regression Confusion matrix)

It is important to note that each machine-learning method has its own strengths and limitations. The choice of method should depend on the specific data and prediction task at hand. For instance, if the dataset is large and contains many input variables, logistic regression may be the best choice. On the other hand, if the dataset is small and contains missing data, Gaussian Naive Bayes may be more appropriate.

In conclusion, machine learning techniques have the potential to revolutionize liver disease diagnosis

and treatment. By accurately predicting liver disease and its progression, machine learning algorithms can help clinicians make more informed decisions and improve patient outcomes. However, it is essential to choose the right method for each specific task to ensure the accuracy and reliability of the predictions.

4 Evaluation

Our evaluation of liver disease prediction involved several key aspects, including the use of multiple performance metrics, a diverse set of feature classification techniques, and extensive exploratory data analysis. These components allowed us to gain insights into the strengths and limitations of various machine learning techniques and identify the most informative features for predicting liver disease. In this section, we will discuss each of these aspects in more detail and highlight the contributions of our evaluation to the field of liver disease prediction.

4.1. Exploratory Data Analysis (EDA)

One key aspect of our evaluation was conducting an extensive EDA of the liver disease dataset. The purpose of EDA is to gain insights into the data, identify patterns and relationships, and identify potential issues such as missing data or outliers. In our EDA, we analyzed the distributions, correlations, and potential outliers of the features in the dataset.

Through our EDA, we identified a subset of the most informative features to use in our evaluation. This allowed us to improve the accuracy and reliability of our machine-learning models by selecting only the most relevant features for liver disease prediction.

Gender has long been considered as an important factor in medical research, as certain conditions affect men and women differently. In the case of liver disease, there has been debate about whether gender should be included as a feature in the analysis of liver disease prediction. It is known to have a varying impact on men and women.

During our exploratory data analysis, we examined the potential impact of gender on the prediction task. We conducted a correlation analysis and found that there was a significant relationship between gender and several other features, including direct bilirubin (DB) and alkaline phosphatase (Alkphos). This contradicts the theory that gender should not be considered an important consideration in liver disease prediction.

Our findings suggest that gender may not have a significant impact on the prediction of liver disease and should not be considered an important feature in future studies. By not including gender in the analysis, we may be able to improve the accuracy and reliability of our prediction models, ultimately leading to better diagnosis and treatment for patients.

We analyzed the distribution of the target variable, which showed a there is no significant difference between men and women. This finding led us to investigate whether stratification by gender was necessary for our models. To address this question, we performed a boxplot analysis to visualize the distribution of each feature across genders. The results of this analysis showed that most features had similar distributions across both genders, suggesting that stratification by **gender was not necessary**.

The decision to include or exclude gender as a feature in our modeling has important implications for the accuracy and effectiveness of our models. By carefully analyzing the distribution of each feature by gender, we were able to make an informed decision about how to structure our analysis. Ultimately, this decision contributed to the development of more accurate and reliable machine-learning models for predicting liver disease.

4.2. Feature Classification

During our exploratory data analysis, we utilized a variety of techniques to understand better the relationship between our features and the target variable. One important technique we used was feature classification, which helped us to identify the most relevant features for predicting the likelihood of liver disease.

We categorized variables using both statistical considerations and then performed statistical tests to determine the significance of each category and examine the relationship between each feature and the target variable. This allowed us to identify different types of relationships, such as straight, inverse, bidirectional, or no relationship at all.

In addition to feature classification, we also conducted a univariate analysis process to investigate the relationship between each feature and the likelihood of liver disease. We used index ranges extracted from the internet to check whether going out of the index range indicated a greater chance of liver disease. This process revealed that for some features, going out of the index range did indeed indicate a greater chance of liver disease, while for others, there was no clear connection.

During our analysis of the liver disease dataset, we conducted a univariate analysis process and identified several important features related to liver disease progression. For example, we found that high levels of Bilirubin (B) and Direct Bilirubin (DB) were strongly correlated with liver disease. Similarly, high levels of SGPT (SgptAA) and SGOT (SgotAA) were found to be indicators of liver disease progression. We also observed that Total Protein (TP) and Albumin (ALBA) levels were inversely correlated with liver disease, while the ALBA to A/G ratio was found to be a significant predictor of liver disease.

By conducting a thorough analysis of these features and their relationship with liver disease, we were able to create a Scorecard that ranked the importance of each feature in predicting liver disease. This allowed us to make informed decisions when selecting features for our machine-learning models, ultimately leading to better performance and accuracy in predicting liver disease.

4.3. Performance Metrics

In order to evaluate the performance of our machine learning models for predicting liver disease, we used multiple performance metrics. These metrics allowed us to gain a comprehensive understanding of how well our models were performing and enabled us to make a fair comparison between them.

One of the metrics we used was accuracy, which measures the percentage of correct predictions made by the model. This is a fundamental measure of a model's performance and gives us an overall idea of how well the model is able to make accurate predictions.

In addition to accuracy, we also used precision, recall, and F1-score. Precision measures the proportion of true positives among all predicted positives, while recall measures the proportion of true positives among all actual positives. The F1 score is a combination of precision and recall, providing a balanced measure of both metrics.

Finally, we used the area under the ROC curve (AUC) to evaluate our models. This metric measures the ability of the model to distinguish between positive and negative cases and provides an overall measure of the model's performance in this regard.

Overall, our use of multiple performance metrics allowed us to gain a comprehensive understanding of how well our models were performing and provided us with a fair and objective way of comparing the performance of different models. This approach allowed us to identify the most effective method for predicting liver disease and provided us with insights into the strengths and limitations of different machine-learning techniques.

4.4. Machine Learning Techniques

We evaluated the performance of various machine learning techniques for predicting liver disease, including Random Forest Regression (RFR), Gaussian Naive Bayes (GNB), and Logistic Regression (LR), as well as ensemble methods such as Random Forest (RF) and Gradient Boosting (GB).

We compared the performance of these techniques on different subsets of the data and investigated their strengths and limitations. For instance, previous studies have shown the effectiveness of RFR for predicting liver disease, but our study extends this work by comparing RFR with other techniques and investigating its performance on different subsets of the data.

Our evaluation provides insights into the potential of machine learning for predicting liver disease and its progression. It highlights the importance of data preprocessing, feature engineering, and model selection in improving the accuracy and reliability of machine learning models. We hope that our study will motivate further research in this field and contribute to the development of more effective tools for the early detection and management of the liver disease.

5 Discussion

Our study evaluated various machine learning techniques to predict liver disease and its progression, and we found that Random Forest had the highest performance, followed by Logistic Regression. However, Gaussian Naive Bayes had notably lower performance and may not be suitable for this task.

Our analysis also highlighted the importance of feature classification in machine learning. Identifying the most informative features for the prediction task and selecting the appropriate classification method can significantly impact model performance. For liver disease prediction, total bilirubin, direct bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT, and Alkphos were found to be **important predictors**. Interestingly, our analysis did not find a significant impact of gender on the predictive performance of our models.

We also conducted extensive exploratory data analysis (EDA) on the dataset, which provided valuable insights into the distribution and correlations of the variables. Our EDA revealed that age and gender were significantly not correlated with liver disease, with older patients and male patients having a higher risk of liver disease. Additionally, During our exploratory data analysis, we found that variables such as Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, Alanine Aminotransferase (ALT or SGPT), Aspartate Aminotransferase (AST or SGOT), Total Protein, Albumin, Albumin/Globulin Ratio (A/G Ratio) were strongly correlated with liver disease and its progression. These variables provided valuable information for predicting liver disease using machine learning models.

Based on our analysis of the dataset, we found that Random Forest and Logistic Regression showed promise as effective methods for predicting liver disease. Random Forest had a higher accuracy rate and performed better in predicting liver disease compared to Logistic Regression. However, Logistic Regression was better at identifying which features were most important in predicting liver disease.

Specifically, we found that features such as Total Bilirubin, Direct Bilirubin, and Alkaline Phosphatase were strongly correlated with liver disease. Other features, such as Albumin and Age, had a weaker correlation.

It's important to note that our findings are limited to our specific dataset and may not generalize to other populations or datasets. Therefore, further studies are needed to validate our findings and explore the potential utility of other machine-learning techniques for liver disease prediction. It's also important to exercise caution when interpreting the results of any machine-learning model and consider the potential impact of factors such as bias and variability in the data.

Table 2: Performance of machine learning models

ML model	**AUC**	**Precision**	**Recall**	**F1-score**
Random Forest	0.867498	0.797258	0.794638	0.76513
Logistic Regression	0.748967	0.68086	0.722767	0.651605
Gaussian Naive Bayes	0.437237	0.794229	0.556832	0.56018

6 Conclusion

In conclusion, our study provides valuable insights into the use of machine learning techniques for predicting liver disease and its progression. Our findings demonstrate that Random Forest and Logistic Regression are promising tools for predicting liver disease, while Gaussian Naive Bayes may not be suitable for this task. The importance of feature engineering and selection was also highlighted, as identifying the most informative features can greatly improve the performance of machine learning models.

Additionally, we found that gender did not significantly impact the predictive performance of our models, but other features such as total bilirubin, direct bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT, and Alkphos were important predictors. Our exploratory data analysis revealed the correlations and distributions of these variables, which can inform the development of more accurate models in the future.

Random Forest achieved the highest performance among the evaluated techniques, with an AUC of 0.86, a precision of 0.792, a recall of 0.794, and F1

score of 0.76. Logistic Regression also showed promising results, with an AUC of 0.74 and F1 score of 0.651. These results demonstrate the potential of machine learning for the early detection and management of a liver disease.

In conclusion, our study demonstrates the importance of careful evaluation of machine learning techniques and feature engineering for predicting liver disease. The insights gained from our study can contribute to the development of more effective tools for the early detection and management of liver disease, ultimately improving patient outcomes.

References

[1] World Health Organization. (2021). Hepatitis. *World Health Organization.* https://www.who.int/health-topics/hepatitis#tab=tab_1

[2] National Institute of Diabetes and Digestive and Kidney Diseases. (2021). Definition and Facts of NAFLD and NASH. https://www.niddk.nih.gov/health-information/liver-disease/nafld-nash/definition-facts

[3] Wu, G., Wu, L., Zhang, Q., Zhu, Y. (2019). Application of machine learning in diagnosis and treatment of liver disease. *Frontiers in Pharmacology, 10,* 1498. https://doi.org/10.3389/fphar.2019.01498

[4] Banerjee, I., Tripathy, J. P., Mishra, S. K., Arora, N. K. (2019). Machine learning for predicting the severity of liver disease: A systematic review and meta-analysis. *PloS One, 14*(5), e0217286. https://doi.org/10.1371/journal.pone.0217286

[5] Quan, W., Song, Y., Xu, Y., Pang, Q., Chen, J., Zhao, B. (2021). A machine learning-based prediction model for liver disease based on routine laboratory test data. *Journal of Clinical Laboratory Analysis, 35*(2), e23658. https://doi.org/10.1002/jcla.23658

[6] Dritsas, K., Trigka, I. (2022). Supervised machine learning models for liver disease risk prediction. *Computers, 12*(1), 19. https://doi.org/10.3390/computers12010019

[7] Aruna, T., Sadasivam, V., Rajaram, M. (2018). Liver disease diagnosis using classification algorithms. *International Journal of Engineering and Applied Sciences Technology, 6*(12), 247-251. https://www.ijeast.com/papers/247-251,%20Tesma612,IJEAST.pdf

[8] Ratziu, V., Charlotte, F., Heurtier, A. (2022). Artificial intelligence in the prediction of non-alcoholic fatty liver. *Current Hepatology Reports, 21*(1), 33-42. https://doi.org/10.1007/s11901-022-00601-3

[9] Latha, P., Rao, G. R. (2017). Analysis of Classification Algorithms for Liver Disease Diagnosis. *International Journal of Applied Engineering Research*, 12(21), 10290-10295. https://www.researchgate.net/publication/319983998_Analysis_of_classification_algorithms_for_liver_disease_diagnosis

[10] Kumari, A., Kaur, M. (2012). A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis. *International Journal of Computer Applications*, 47(5), 36-41. https://doi.org/10.5120/7068-9772

[11] Kalaiselvi, P., Kavitha, K. (2015). Comparative Study of Data Mining Algorithms in Medical Data. *International Journal of Engineering Research and Technology*, 4(9), 156-160. https://www.ijert.org/comparative-study-of-data-mining-algorithms-in-medical-data