LiverGuard - Disease Predictor

Snehil Jaiswal
snehil22503@iiitd.ac.in

Sneha Nagpal
sneha22502@iiitd.ac.in

Sidak Singh Chahal
sidak22495@iiitd.ac.in

Shubham Kumar Dwivedi
shubham22494@iiitd.ac.in

Abstract

Liver diseases are a leading cause of global mortality, with rising cases linked to increased drug and alcohol use, exposure to toxins, and consumption of processed foods, placing significant strain on healthcare systems. Early detection is critical for effective treatment and improved patient outcomes, and this project aims to develop machine learning models to accurately predict liver diseases, thereby assisting medical professionals in early diagnosis and reducing their workload. Through a comprehensive approach involving data analysis, preprocessing, and the application of advanced techniques such as feature engineering and model evaluation, the study seeks to identify effective predictive algorithms. By leveraging these methodologies, this work establishes a foundation for creating robust tools to enhance clinical decision-making and contribute to the broader field of medical data science.(GitHub)

1. Introduction

Liver diseases are a major global health concern, driven by factors like rising alcohol and drug use, environmental toxins, and processed food consumption. This growing prevalence puts immense pressure on healthcare systems, leading to delayed diagnoses and suboptimal treatment due to limited resources.

Early detection is key to improving patient outcomes, but traditional diagnostic methods are often time-consuming and labor-intensive, compounding the challenges for clinicians. This issue affects not only individual health but also public health and healthcare efficiency.

This study aims to address the need for efficient liver disease prediction by leveraging machine learning algorithms. By evaluating various predictive techniques, the project seeks to improve diagnostic accuracy, enable early detection, and optimize clinical workflows, offering scalable solutions to reduce the global impact of liver diseases.

2. Literature Survey

1. Tokala, Srilatha et al. Liver Disease Prediction and Classification using Machine Learning Techniques, 2023
   This paper applies machine learning algorithms to predict and classify liver disorders, aiming to improve classification accuracy and reduce doctors' workload. Naive Bayes achieved the highest precision, while Logistic Regression and Random Forest performed well in terms of recall.

2. Bhupathi, Deepika et al. Liver disease detection using machine learning techniques, 2022
   Most previous research has focused on analyzing the Liver Patient Records dataset, with limited attention to preprocessing. This study highlights preprocessing as a crucial step and applies various machine learning algorithms using the Liver Disease Prediction (LDP) method, based on the SEMMA framework (Sample, Explore, Modify, Model, Assess).

3. Dataset

3.1. Dataset Description

The dataset is a **Liver Disease Patient Dataset** obtained from Kaggle. The Dataset has 30691 rows and 11 features(columns). The Features are:

- **Age of the patient:** The age of the patient
- **Gender of the patient:** The gender of the patient (male/female)
- **Total Bilirubin:** The total amount of bilirubin in the blood
- **Direct Bilirubin:** The level of direct (conjugated) bilirubin in the blood.
- **Alkphos Alkaline Phosphotase:** The level of alkaline phosphatase enzyme in the blood.

- **Sgpt Alamine Aminotransferase:** The level of SGPT (ALT) enzyme in the blood.
- **Sgot Aspartate Aminotransferase:** The level of SGOT (AST) enzyme in the blood.
- **Total Proteins:** The total amount of proteins in the blood
- **ALB Albumin:** level of albumin in the blood
- **A/G Ratio Albumin and Globulin Ratio:** The ratio of albumin to globulin in the blood
- **Result:**
  - **1:** Liver Disease
  - **2:** Mon Liver Disease

3.2. Dataset Visualization

3.2.1 Histogram

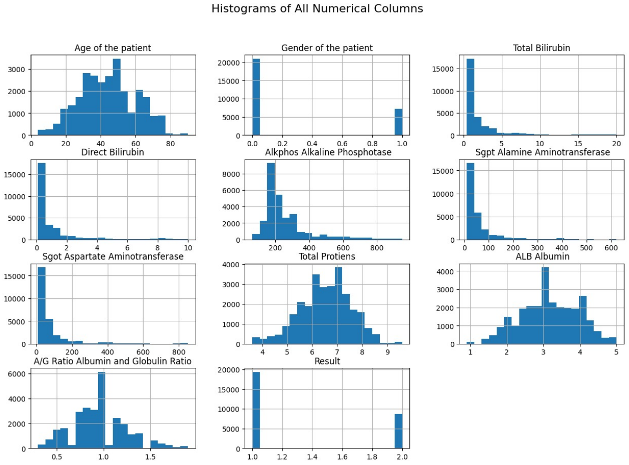

Figure 1. Histograms of all Numerical Columns

3.2.2 Correlation Heatmap

Total Bilirubin and Direct Bilirubin have a strong positive correlation (0.95), indicating they are highly related. SGPT and SGOT show a strong positive correlation (0.79), and ALB Albumin and Total Proteins are also strongly correlated (0.8). The ”Result” column/row represents the correlation between features and liver disease prediction. Total Bilirubin, Direct Bilirubin, and Alkphos Alkaline Phosphatase exhibit a negative correlation with the Result, suggesting higher values may indicate a negative outcome. A/G Ratio and ALB Albumin show a slightly positive correlation with the Result.

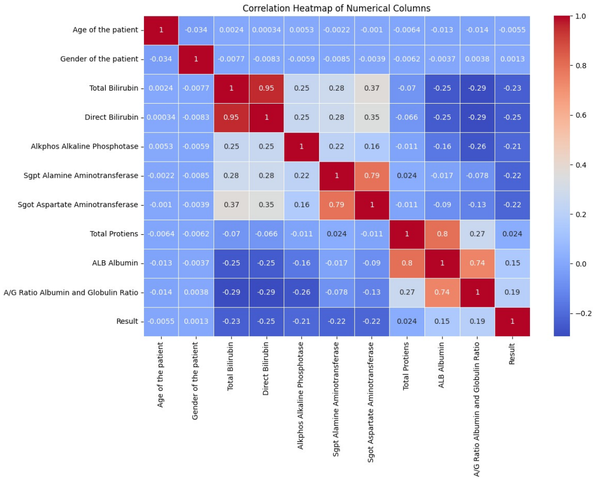

Figure 2. Correlation Heatmap of the Features

3.2.3 Box Plots

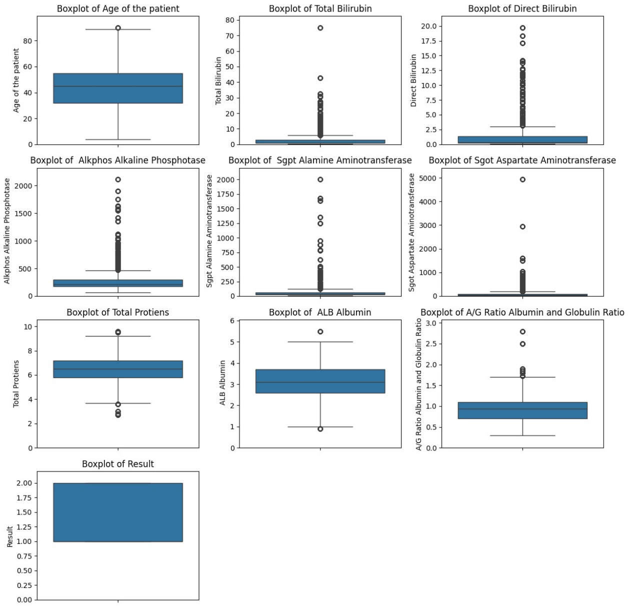

Figure 3. Box plots

The patient population is predominantly middle-aged (median 45-50 years), with a broad age range. Liver function markers (Bilirubin, Alkaline Phosphatase, SGPT, SGOT) are right-skewed with many high outliers, suggesting most patients have normal values but some show liver dysfunction. Protein markers (Total Proteins, Albumin, A/G Ratio) have more symmetric distributions with fewer outliers, indicating stable protein levels, while the Result appears to be a binary classification (disease presence/absence).

3.3. Data Pre-Processing

3.3.1 Handling Null Values

For all Numerical Columns we replaced all the Null values with their mean.

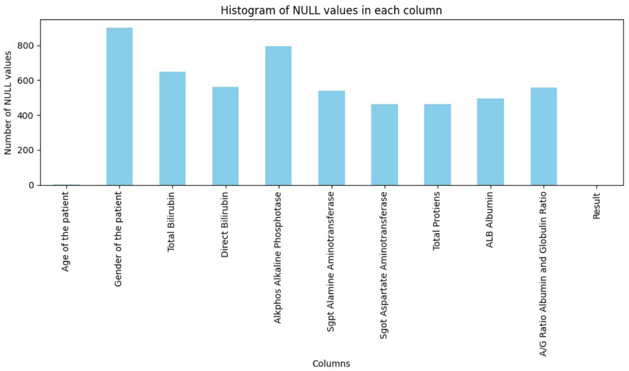

Figure 4. Plot of NULL values in each feature

3.3.2 Label Encoding of categorical Column

- There was only one Categorical column in our dataset i.e Gender of the Patient. For handling Null values in gender we replaced it by the most frequent Gender(Mode of the data) which was 'Male'. To apply label encoding we mapped Male to '0' and Female to '1'.

- The Labels in the original dataset were 1 and 2 representing Liver patient and Mon liver patient respectively. To simplify the execution of ML models we mapped the labels $2 \rightarrow 0$ and $1 \rightarrow 1$.

3.3.3 Splitting Data

We split the data into Train:Val:test :: 70:10:20

3.3.4 Handling Outliers

We used the Z-Score to remove outliers from the Data. Any data points which were **3 Z-Scores** away were imputed from the database. After removing the Outliers the no. of data points were **28133**.

4. Methodology

4.1. Handling Data Discrepancies

We observed a major data imbalance in the labels of our original dataset. There were approximately 20,000 values mapping to 1 and 8,000 to 0. To avoid Bias towards the majority class in our model and improve model generalization

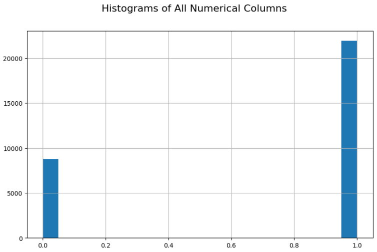

Figure 5. Data Imbalance

we fixed the imbalance using **BootStrapping**. The final Ratio was $18000(1) : 15000(0)$

4.2. Feature Engineering

4.2.1 Feature Extraction

The feature extraction process began by refining numerical attributes, including handling column names to ensure consistency. Polynomial features were generated for the bilirubin attributes, capturing higher-order interactions between Total Bilirubin and Direct Bilirubin, with the resulting features added to the dataset. Additional derived features were created, such as the Albumin to Proteins Ratio, which provides insights into protein composition. To address skewness in certain attributes, log transformation was applied to Alkphos Alkaline Phosphatase.

4.2.2 Feature Selection

Feature selection was carried out using Recursive Feature Elimination (RFE) with a Logistic Regression model as the estimator. The RFE process iteratively assessed the importance of features and ranked them based on their contribution to predicting the target variable. The top six features were selected based on their significance, reducing dimensionality and ensuring the model focused on the most informative predictors. The dataset was then refined to include only the selected features for subsequent modeling tasks.

4.3. Evaluating ML Models

The modeling process began with simpler algorithms, such as Gaussian Naive Bayes and Logistic Regression, to establish baseline performance. Stratified K-Fold cross-validation ensured robust evaluation, while metrics like accuracy, F1 score, RMSE, and classification reports highlighted each model's strengths and limitations in capturing data relationships. Gaussian Naive Bayes offered probabilistic insights, and Logistic Regression provided interpretability with linear decision boundaries.

Subsequently, ensemble methods like Random Forest, Decision Tree, and XGBoost were employed to exploit nonlinear interactions, with XGBoost excelling in performance through gradient boosting. Cross-validation measured consistency, and advanced models such as Support Vector Machines (SVM) and Multi-Layer Perceptrons (MLP) were explored for complex patterns in high-dimensional data. This iterative approach refined the understanding of dataset characteristics and guided optimal model selection.

5. Result and Analysis

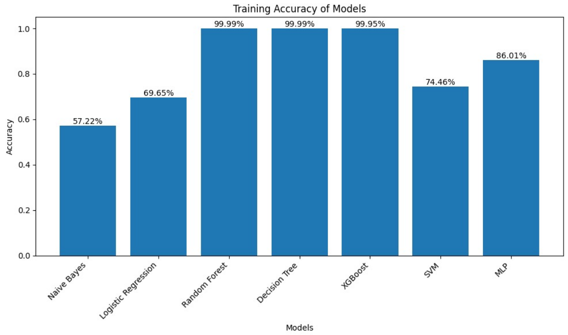

Figure 6. Training Accuracy of Models

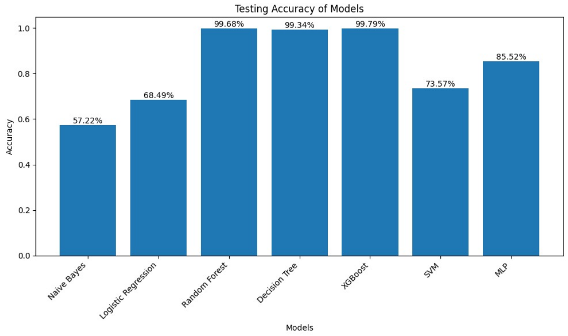

Figure 7. Testing Accuracy of Models

Ensemble learning methods, such as **Random Forest and XGBoost**, achieved the highest accuracies of 99.68% and 99.79%, respectively. These models perform better for several reasons:

1. Ability to Handle Non-linear Relationships: Ensemble methods can capture non-linear patterns in the data. When a dataset has even slight non-linearities, these models are more adept at adapting to the data.

2. Feature Importance and Interactions: XGBoost excels in ranking features based on their importance, helping to leverage the most influential variables in the dataset more effectively.

3. Dataset Simplicity: The high performance of ensemble methods suggests that the dataset contains clear, separable patterns. These models leverage this simplicity by efficiently learning these patterns through multiple decision boundaries.

Complex models like SVM (73.57%) and MLP (85.52%) performed moderately, likely due to the dataset's simplicity and the models' sensitivity to hyperparameters. Simpler models, such as Naive Bayes (57.22%) and Logistic Regression (68.49%), struggled due to their assumptions of linear relationships and feature independence, which made them less suited for non-linearities and complex interactions.

6. Conclusion

6.1. Learnings from the Project

In our liver disease prediction project, we gained skills in data preprocessing, feature selection, and addressing class imbalance. We focused on selecting and tuning models, particularly ensemble methods, for improved performance. The project emphasized the importance of interpretability, ethical considerations, and the role of machine learning in early diagnosis and treatment planning in healthcare.

6.2. Key Results

Ensemble and heuristic methods, such as Random Forest and XGBoost, perform well on simple datasets because they efficiently capture clear patterns through multiple decision boundaries and are robust to noise and feature interactions. In contrast, complex models like SVM and MLP, which are designed for intricate data patterns, struggle with simpler datasets due to their sensitivity to hyperparameters and tendency to overfit, leading to lower performance.

6.3. Limitations and Challenges

One limitation of our liver disease prediction project was the availability of imbalanced data, which could affect model accuracy and generalization. Additionally, the dataset's incomplete nature required significant preprocessing, in achieving optimal feature extraction. Furthermore the computational cost and time constraints of complex models like ensemble methods and neural networks were a factor to consider. Lastly, the model's performance could be limited by the representativeness of the dataset, as it may not fully capture the wide variety of cases seen in real-world medical scenarios.

6.4. Contribution

Snehil Jaiswal: Dataset Analysis and Pre-Processing,Literature Review and Report Writing
Sneha Nagpal: Model Evaluation and Comparison, EDA and Feature Extraction
Sidak Singh Chahal: Literature Review, Inference of Results and PPT
Shubham Kumar Dwivedi: EDA, Graph Plotting and PPT

7. References

- Tokala, Srilatha et al. Liver Disease Prediction and Classification using Machine Learning Techniques, 2023

- Bhupathi, Deepika et al. Liver disease detection using machine learning techniques, 2022

- https://ieeexplore.ieee.org/document/9074368
- https://seaborn.pydata.org/tutorial/introduction.html
- https://www.w3schools.com/python/matplotlib_intro.asp
- ScienceDirect
- https://ieeexplore.ieee.org/document/9787574
- https://www.sciencedirect.com/science/article/pii/S1877050
- Gaurav K et al. Human Disease Prediction using Machine Learning Techniques and Real-life Parameters, 2023
- https://www.geeksforgeeks.org/disease-prediction-using-machine-learning/
- https://github.com/yaswanthpalaghat/Disease-prediction-using-Machine-Learning
- Rohan Volety ,Liver Disease Prediction Using Machine Learning, 2024