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 have become one of the leading causes of death globally, with cases steadily increasing due to rising drug and alcohol use, exposure to toxic substances, and consumption of processed foods. Early detection is critical for effective treatment, but the growing number of patients places a significant burden on healthcare systems. In this study, we propose the development of machine learning algorithms to accurately predict liver diseases based on patient data. These predictive models can assist medical professionals by providing early diagnoses and reducing their workload, thereby enabling more efficient and targeted healthcare interventions. Our approach involves evaluating various classification techniques to identify the most effective algorithm for liver disease prediction, aiming to contribute to both clinical practice and the broader field of medical data science.

1. Introduction

Please follow the steps outlined below when submitting your manuscript to the IEEE Computer Society Press. This style guide now has several important modifications (for example, you are no longer warned against the use of sticky tape to attach your artwork to the paper), so all authors should read this new version.

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 using patient data. Chronic liver disease is analyzed with various classifiers, improving classification accuracy and reducing doctors' workload. The results showed that Naive Bayes achieved the highest precision compared to the other algorithms examined. Additionally, the Logistic Regression and Random Forest algorithms were found to have good results when recall was considered

2. Bhupathi, Deepika et al. Liver disease detection using machine learning techniques, 2022
   Most previous research has focused primarily on the analysis of the Liver Patient Records dataset, with little attention given to the preprocessing stage. This study addresses that gap by emphasizing the importance of preprocessing as a critical step in data analysis. Additionally, various machine learning algorithms are applied in this research. The proposed Liver Disease Prediction (LDP) method follows the SEMMA framework (Santos Azevedo, 2005), which stands for Sample, Explore, Modify, Model, and Assess (Azevedo Santos, 2008)

3. Dataset

3.1. Dataset Description

The dataset is a **Liver Disease Patient Dataset** obtained from Kaggle. The Dataset has 30691 rows and 10 features(columns). The 11th column tells us whether the patient has liver disease or not, where '1' denotes liver patient and '2' denotes Non-liver patient.
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 or 2 , classifies patient as a liver patient or a Mon liver patient

3.2. Dataset Visualization

3.2.1 Histogram

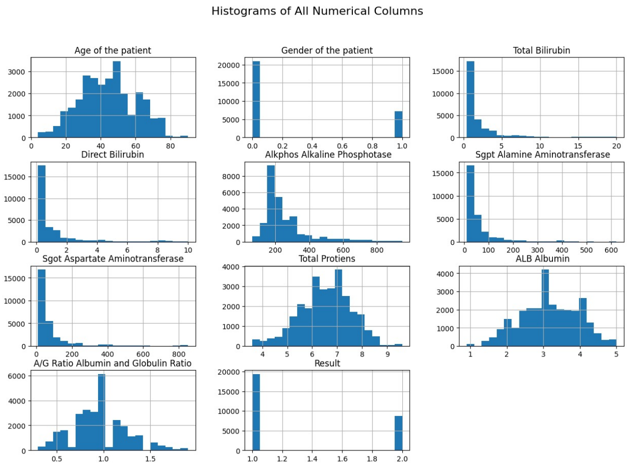


Figure 1. Histograms of all Numerical Columns

3.2.2 Correlation Heatmap

Total Bilirubin and Direct Bilirubin show a very strong positive correlation (0.95), meaning these two variables are highly related. SGPT (Alamine Aminotransferase) and SGOT (Aspartate Aminotransferase) have a strong positive correlation (0.79), indicating that liver enzyme levels tend to rise together. ALB Albumin and Total Proteins also have a strong positive correlation (0.8), reflecting their biological relationship. **The column/row labeled "Result" represents the correlation between each feature and the target (liver disease prediction)**. **Total Bilirubin, Direct Bilirubin, and Alkphos Alkaline Phosphatase** have a **negative correlation with the Result**, which suggests that higher values of these features might be associated with a negative outcome (or vice versa). **A/G Ratio and ALB Albumin** show a slightly positive correlation with the Result.

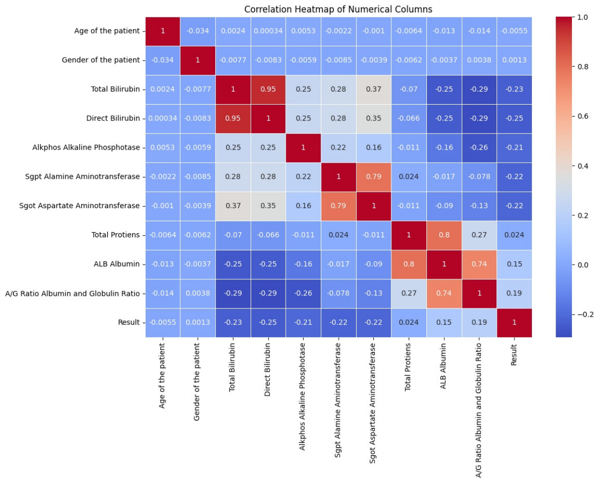


Figure 2. Correlation Heatmap of the Features

3.2.3 Box Plots

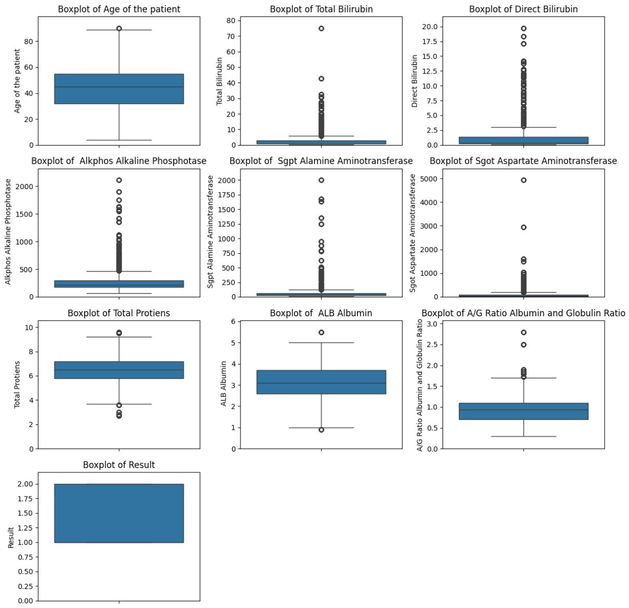


Figure 3. Box plots

Patient population is middle-aged centered (median 45-50 years), with a wide age spread from young to elderly patients. Liver function markers (Bilirubin, Alkaline Phosphatase, SGPT, SGOT) show strongly right-skewed distributions with numerous high outliers, suggesting most patients have normal values but some show significant liver dysfunction. Protein markers (Total Proteins, Albumin, A/G Ratio) show more symmetric distributions with fewer outliers, indicating relatively stable protein levels across the population, while the Result appears to be a binary classification (likely 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.

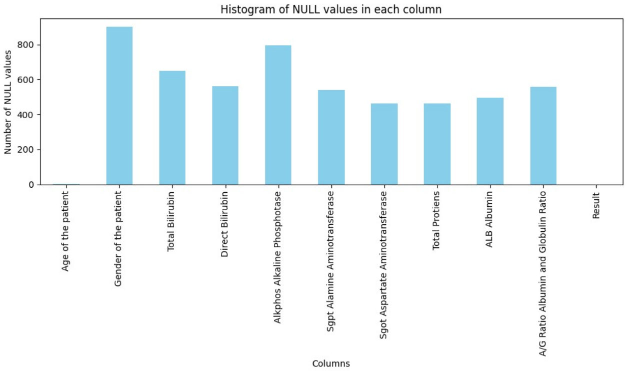


Figure 4. Plot of NULL values in each feature

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'.

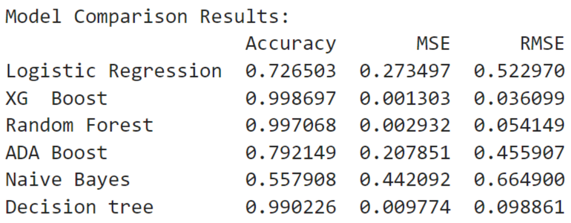
3.3.3 Handling Outliers

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.

4. Methodology

The main aim of our model is to predict liver disease based on various medical parameters and blood test results. Our methodology begins with comprehensive data preprocessing and Exploratory Data Analysis (EDA), where we analyze distributions, correlations, and handle outliers in our medical dataset. Following the preprocessing, we then split our dataset into an 80-20 ratio for training and testing respectively. Further , we implement and compare multiple machine learning algorithms including Logistic Regression, followed by more complex algorithms such as Random Forest Classification, Naive Bayes, Decision Trees, XGBoost, AdaBoost. To enhance our model's robustness and handle potential data discrepancies in the future, we plan to implement bootstrapping techniques. The final phase will involve selecting the best performing model based on both accuracy and computational efficiency, followed by fine-tuning it with optimal hyperparameters for deployment in real-world medical scenarios.

5. Result and Analysis



```
Model Comparison Results:
                     Accuracy       MSE      RMSE
Logistic Regression  0.726503  0.273497  0.522970
XG  Boost            0.998697  0.001303  0.036099
Random Forest        0.997068  0.002932  0.054149
ADA Boost            0.792149  0.207851  0.455907
Naive Bayes          0.557908  0.442092  0.664900
Decision tree        0.990226  0.009774  0.098861
```

Figure 5. result analysis of different models

XG Boost emerges as the top performer with an outstanding accuracy of 99.87% and the lowest error rates (MSE: 0.001303, RMSE: 0.036099). These exceptional metrics indicate that XG Boost has successfully captured the complex patterns in the liver disease data, making it the most reliable choice for prediction.
Random Forest Following closely, Random Forest achieves an impressive 99.71% accuracy with minimal error rates (MSE: 0.002932, RMSE: 0.054149). This strong performance demonstrates Random Forest's robust ensemble learning capabilities, making it a reliable alternative to XG Boost.
Decision Tree The Decision Tree model shows remarkable performance with 99.02% accuracy and reasonable error metrics (MSE: 0.009774, RMSE: 0.098861). This performance is particularly noteworthy given its simpler structure, suggesting it could be a good choice when model interpretability is prioritized.
ADA Boost shows moderate performance with 79.21% accuracy and higher error rates (MSE: 0.207851, RMSE: 0.455907). While not competing with the top performers, it still provides better results than traditional algorithms and might improve with further parameter tuning.
Logistic Regression with an accuracy of 72.65% and higher error rates (MSE: 0.273497, RMSE: 0.522970), Logistic Regression demonstrates limited capability in handling the complexity of our dataset. It serves better as a baseline model rather than a final solution.
Naive Bayes shows the weakest performance with 55.79% accuracy and the highest error rates (MSE: 0.442092, RMSE: 0.664900). These metrics suggest that the model's assumptions might not align well with the underlying data

patterns, making it unsuitable for this specific prediction task.

6. Conclusion

- **Learning from project:** In this project focused on liver disease prediction using medical data, we gained valuable experience in developing a comprehensive healthcare analytics pipeline. We realized the critical importance of accurate medical data preprocessing and utilized various statistical techniques to handle the diverse range of biochemical markers and patient parameters. Through extensive data visualization using matplotlib and seaborn, we analyzed the distributions and correlations of liver function tests, protein markers, and other vital medical indicators. Our model comparison demonstrated that XG Boost significantly outperformed other algorithms with 0.9987 accuracy, followed closely by Random Forest at 0.9971 and Decision Tree at 0.9902. The stark performance difference between ensemble methods and traditional algorithms like Naive Bayes (0.5579) highlighted the complexity of medical diagnosis patterns. Through this project, we developed essential skills in handling sensitive medical data, implementing various machine learning algorithms, and understanding the importance of model selection in healthcare predictions. Additionally, we learned about the potential of boosting techniques to enhance prediction accuracy in medical diagnosis, setting a foundation for future work in healthcare analytics.

- **Future Work:** To enhance our liver disease prediction model's robustness and clinical applicability, several improvements are planned. We aim to address potential overfitting issues by implementing more rigorous cross-validation techniques and regularization methods, ensuring our model's generalizability across diverse patient populations. Bootstrapping will be implemented to handle data imbalances and generate more reliable confidence intervals for our predictions, particularly important given the varying distributions of liver function parameters. We plan to conduct extensive hyperparameter tuning using advanced techniques like Bayesian optimization and randomized search, moving beyond our current grid search approach to find optimal model configurations.

- **Member Contribution:**
  Sneha Nagpal: Model Evaluation and Comparison, EDA and Report Writing.
  Snehil Jaiswal: Dataset Analysis and Pre-Processing, Literature Review and Report Writing.
  Sidak Singh Chahal: Literature Review, Data Analysis and Pre-Processing and PPT Making.
  Shubham Kumar Dwivedi: Model Evaluation, EDA, Graph Plotting and PPT Making

7. References

- 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