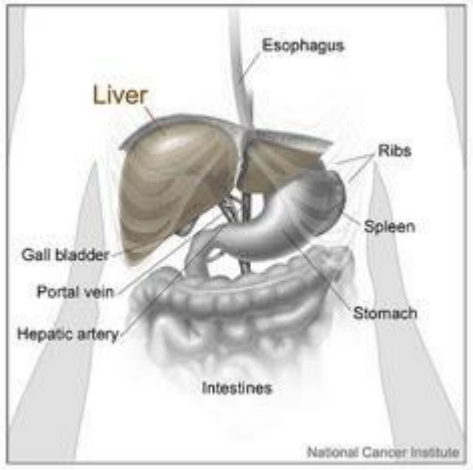
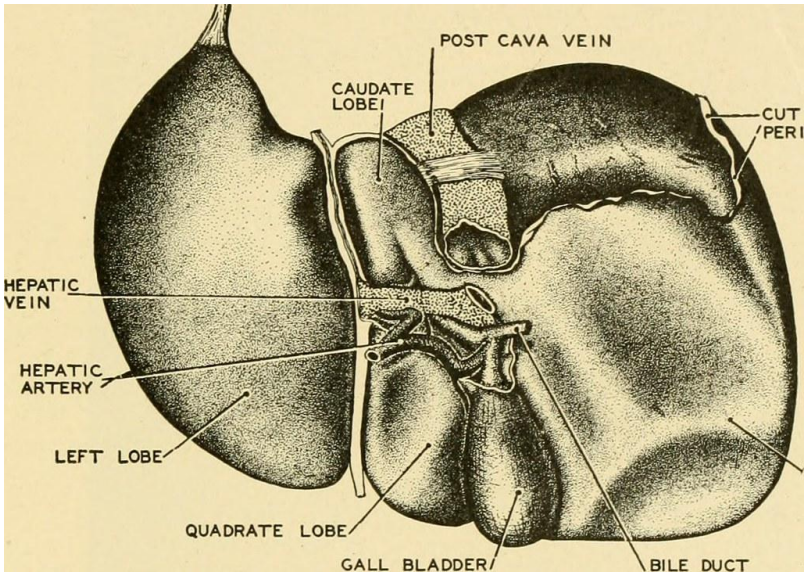
Capstone 2: Indian Liver Patient Records Analysis

Humayra Azra

The problem

Liver disease is a significant public health issue, particularly in India, where early detection and accurate diagnosis can improve patient outcomes. However, traditional diagnostic methods rely on expert interpretation of liver function test results, which can be time-consuming and prone to human error.





Project Overview

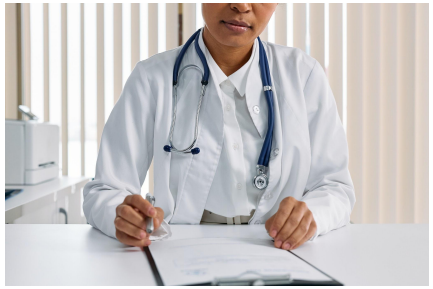
This project aims to leverage data science techniques to improve the diagnosis of liver disease using the Indian Liver Patient dataset. By employing data preprocessing, exploratory data analysis (EDA), and machine learning models, we seek to identify key biomarkers associated with liver disease and develop a predictive model that can assist healthcare professionals in making more accurate and efficient diagnoses.

Our Target Audience

Public Health Organizations

World Health Organization

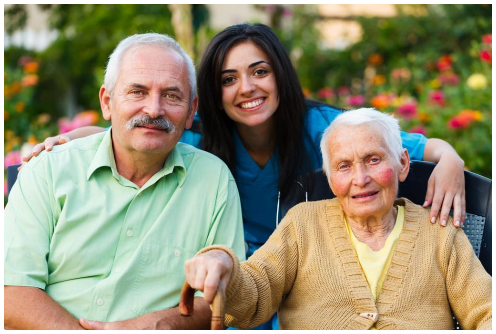
Healthcare Professionals



Hospitals & Clinics

Health Insurance Providers

Patients & Their Families



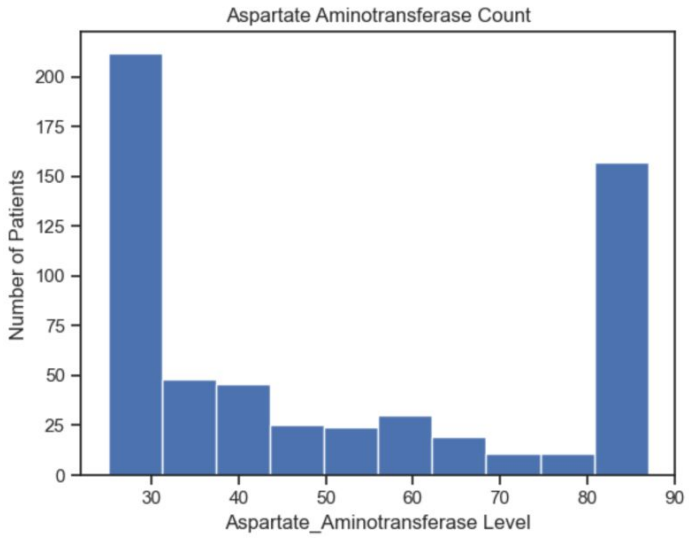
Medical Researchers & Data Scientists

Pharmaceutical Companies

Causes and What We Can Monitor

Liver Function Tests (LFTs) – Biomarkers in Blood Tests:

- **Total Bilirubin & Direct Bilirubin**
- **Alkaline Phosphatase (ALP)**
- **Alanine Aminotransferase (ALT) & Aspartate Aminotransferase (AST)**
- **Total Proteins & Albumin**
- **Albumin-to-Globulin Ratio (A/G Ratio)**



Common Causes:

1. **Lifestyle Factors**
2. **Infections**
3. **Genetic & Autoimmune Conditions**
4. **Toxins & Medications**
5. **Metabolic Disorders**

Significant Findings

Patient Demographics:

- **Average age** of patients is calculated and visualized.
- Most patients fall in the **45-65 age range**, suggesting liver issues are more common in middle-aged individuals.
- The data shows a **higher number of male patients** compared to females.

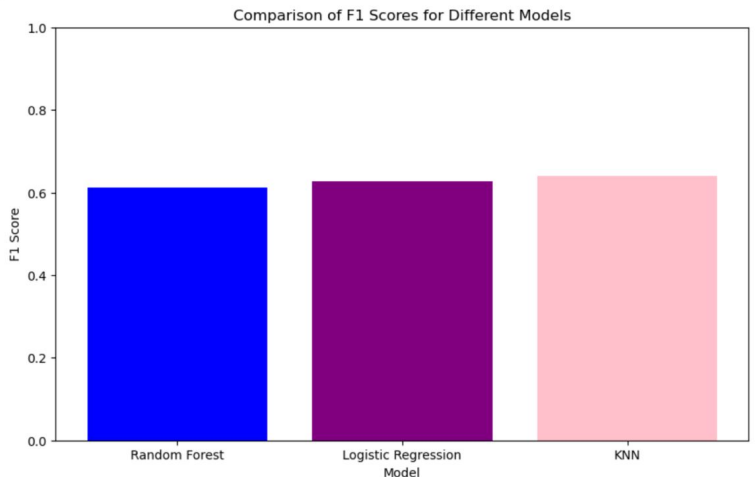
Liver Disease Diagnosis:

- The dataset labels patients as either having or not having liver disease.
- A **larger proportion of patients have liver disease** compared to those without.

Outlier Detection:

- Several outliers identified in bilirubin and enzyme levels.
- Boxplots and IQR-based statistics are used to visualize and manage outliers.

Machine Learning Models



Based on the weighted F1 scores alone, the KNN model performed better than the Random Forest model and the Logistic Regression model.

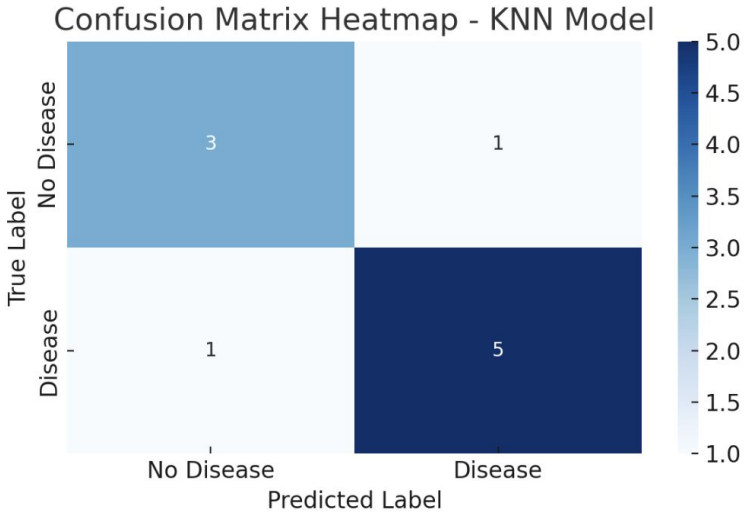
KNN Overview

```
Accuracy: 0.73
Classification Report:
              precision    recall  f1-score   support

           1       0.73      0.99      0.84        83
           2       0.75      0.09      0.16        34

    accuracy                           0.73       117
   macro avg       0.74      0.54      0.50       117
weighted avg       0.73      0.73      0.64       117

Confusion Matrix:
[[82  1]
 [31  3]]
0.6394649627732334
```



Future Improvements

Contribute to Research or Public Health

- Compare your dataset with others (e.g., global liver patient records)
- Analyze trends based on age, gender, or region
- Publish a **case study** or **data story** for awareness

Combine with Other Data Sources

- Merge with other clinical data:
  - Diet, alcohol use, medications
  - Imaging or genetic info (if available)
- Enables **multi-modal modeling**: richer predictions, deeper insights

Additional modeling tunine

- Try **XGBoost**, **LightGBM**, or **SVM**
- Tune thresholds to increase recall for high-risk patients

Next steps

Clinical Decision Support

- Assist doctors with:
  - Risk scores
  - Treatment suggestions
  - Flagging unusual results

Group patients by:

- Disease severity
- Response to treatment
- Demographics or risk level

Medication Risk Prediction

- Predict **adverse drug reactions** using:
  - Medical history
  - Lab results
  - Demographics

Health App or Dashboard Development

- Build apps that:
  - Monitor chronic conditions (liver, diabetes, BP)
  - Alert patients or caregivers of changes
  - Let users **track progress over time**

Public Health Insights

- Use EDA and models to:
  - Identify disease trends across regions
  - Guide resource allocation
  - Support awareness and prevention programs