Data Collection and Preprocessing Phase

Date	4th June 2024
Team ID	SWTID1720175375
Project Title	Prediction and analysis of liver patient data using ML
Maximum Marks	2 Marks

Data Collection Plan

Section	Description
Project Overview	The coexistence of liver diseases poses significant clinical challenges, requiring effective predictive models for early detection and intervention. In this study, we employed decision tree and logistic regression algorithms to predict the likelihood of liver disease in individuals diagnosed . Distinct datasets were utilized, for liver disease prediction, containing relevant clinical attributes. Through rigorous experimentation and evaluation, our models demonstrated promising performance in identifying the presence of liver disease in individuals. In the ever-evolving field of healthcare, predicting and preventing liver diseases have become paramount to ensuring the well-being of individuals and communities. Today, we will delve into two powerful machine learning techniques, Logistic Regression and Decision Tree, which have shown significant potential in predicting the likelihood of these diseases. Logistic Regression is a statistical

	method that allows us to model the relationship between predictor variables and a binary outcome, such as the presence or absence of liver diseases. This technique is particularly useful when we want to understand the effect of various factors on the probability of a specific disease. Decision Trees, on the other hand, are a non-parametric method used for both classification and regression tasks. They work by recursively splitting the data into subsets based on the most significant predictor variables, thus creating a tree- like model that can be easily interpreted and understood. In the context of predicting liver diseases, decision trees can help identify the most important risk factors and provide a visual representation of the decision-making process. Combining these two techniques can lead to more accurate and robust predictions, as well as a deeper understanding of the complex interplay between various risk factors and the likelihood of developing liver diseases.
Data Collection Plan	This dataset has contained 12 parameters where we choose 11 parameters for our further analysis and 1 parameter as a target class Such as, 1. Age of the patient 2. Gender of the patient 3. Total Bilirubin 4. Direct Bilirubin 5. Alkaline Phosphatase 6. Alanine Aminotransferase 7. Aspartate Aminotransferase 8. Total Proteins 9. Albumin 10. Albumin and Globulin Ratio 11. Cholesterol 12. Dataset: field used to split the data into two sets (patient with liver disease, or no disease)

Raw Data Sources Identified	Liver dataset contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India. Number of instances in dataset is 583 whereas 75.64% male patients and 24.36% are female patients and total of attributes is 12. The "Dataset" column is a class label used to divide groups into liver patient (liver disease) or not (no disease). This data set contains 441 male patient records and 142 female patient records. Any patient whose age exceeded 89 is listed as being of age "90".

Raw Data Sources

Source Name	Description	Location/URL	Format	Size	Access Permissions
liver-patient-analysis	The dataset covers a wide range of patient ages (from 13 to 85 years old) and genders (both male and female). This suggests the data represents a diverse population of liver patients.	https://github.com/AbhishekMali21/LIVER-PATIENT-ANALYSIS/blob/master/liver_patient.csv	CSV	24 KB	Public
indian-liver-patient	The dataset includes a comprehensive set of blood test results, such as Total	https://www.kaggle.com/datasets/uciml/indian-liver-patient-records	CSV	23 KB	Public

	Bilirubin, Direct Bilirubin, Alkaline Phosphotase, Alamine Aminotransferase, Aspartate Aminotransferase, Total Proteins, Albumin, and Albumin and Globulin Ratio. This provides a detailed medical profile for each patient.				
…	…	…	…	…	…