Data Collection and Preprocessing Phase

Date	9 July 2024
Team ID	SWTID1720013031
Project Title	Prediction and Analysis of Liver Patient Data Using Machine Learning
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

Data Collection Plan

Section	Description
Project Overview	The Liver Disease Prediction project aims to develop a machine learning-based web application that predicts the likelihood of liver disease in individuals based on various health parameters. By leveraging multiple classification algorithms such as SVM, Random Forest, K-Nearest Neighbors, and Logistic Regression, the application provides a comprehensive analysis of user-submitted health data including age, gender, and biochemical markers. The primary objective is to facilitate early detection and intervention, thereby improving patient outcomes and reducing healthcare costs associated with liver disease management.

Data Collection Plan	The data for the Liver Disease Prediction project is sourced from the "Indian Liver Patient Records" dataset available on Kaggle. This dataset comprises various health parameters collected from individuals, which are essential for building predictive models for liver disease. The dataset includes features such as age, gender, and several biochemical markers related to liver function. You can access the dataset from the following link: 'https://www.kaggle.com/datasets/uciml/indian-liver-patient-records'
Raw Data Sources Identified	The primary raw data source for the Liver Disease Prediction project is the "Indian Liver Patient Records" dataset from Kaggle. This dataset contains health-related parameters from liver patients, including demographic information (age and gender) and various biochemical markers such as 'Total_Bilirubin', 'Direct_Bilirubin', 'Alkaline_Phosphotase', 'Alamine_Aminotransferase', 'Aspartate_Aminotransferase', 'Total_Protiens', 'Albumin', 'Albumin_and_Globulin_Ratio'.

Raw Data Sources

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle Dataset	The dataset comprises details/columns: ['Age', 'Gender', 'Total_Bilirubin', 'Direct_Bilirubin','Alkaline_ Phosphotase',	https://www.kaggle.com/datasets/uciml/indian-liver-patient-records	CSV	23.3KB	Public

	‘Alamine_Aminotransferase’, ‘Aspartate_Aminotransferase’. ‘Total_Proteins’, ‘Albumin’, ‘Albumin_and_Globulin_Ratio’, ‘Dataset’]				