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 Quality Report Template

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

Data Source	Data Quality Issue	Severity	Resolution Plan
indian-liver-patient	Data Quality: The dataset contains some duplicate rows, which should be identified and removed to ensure data integrity.	Moderate	The dataset has a large number of rows (over 500) and a wide range of values for the different features, indicating it may be a comprehensive dataset. However, there are a few rows with duplicate values for some patients, which could indicate data quality issues that need to be addressed

indian-liver-patient	Imbalanced Classes: The target variable (liver disease diagnosis) is imbalanced, with more instances of liver disease (class 1) than no liver disease (class 2). This may require techniques like oversampling or undersampling to balance the classes.	Low	The "Dataset" column indicates that the target variable (liver disease diagnosis) is imbalanced, with 1 representing liver disease and 2 representing no liver disease. Imbalanced target variables can pose challenges for machine learning models and may require techniques like oversampling or undersampling to address
indian-liver-patient	Outliers: The dataset contains some features with a wide range of values, indicating the potential presence of outliers that should be identified and handled appropriately.	Moderate	Some of the feature values, such as Alamine_Aminotransferase, Aspartate_Aminotransferase, and Total_Bilirubin, appear to have a wide range, indicating the potential presence of outliers. Outliers can significantly impact model performance and should be identified and handled appropriately.