# Data Collection and Preprocessing Phase

| | |
|---|---|
| Date | 24 June 2025 |
| Team ID | SWTID1749708868 |
| Project Title | Revolutionizing Liver Care : Predicting Liver Cirrhosis Using Advanced Machine Learning Techniques |
| Maximum Marks | 2 Marks |

## Data Collection Plan & Raw Data Sources Identification Template

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 Template

| Section | Description |
|---|---|
| Project Overview | This project focuses on developing a machine learning–driven web application that predicts the risk of liver cirrhosis in patients by analyzing medical history, laboratory data, and lifestyle factors. By integrating regression-based models into a user-friendly Flask interface, the system provides timely risk assessments that support healthcare providers in making early and informed clinical decisions. The application is designed to assist in screening, monitoring disease progression, and optimizing treatment strategies—contributing to more proactive and personalized liver care within resource-constrained healthcare environments.. |

| | |
|---|---|
| Data Collection Plan | • **Primary Dataset Acquisition from Kaggle**: Identify and download relevant liver disease datasets from Kaggle (e.g., the "Indian Liver Patient Dataset" or similar repositories) that include patient attributes such as bilirubin levels, albumin, enzymes, and diagnosis status—ensuring the data aligns with clinical indicators of cirrhosis.<br>• **Dataset Validation and Preprocessing**: Review Kaggle dataset metadata for credibility, assess data completeness, and clean the dataset by handling null values, encoding categorical features, and standardizing formats to prepare it for effective exploratory analysis and model training. |
| Raw Data Sources Identified | The dataset used in this project was sourced from Kaggle, a widely trusted platform hosting diverse, high-quality open datasets for machine learning and data analysis. |

**Raw Data Sources Template**

| Source Name | Description | Location/URL | Format | Size | Access Permissions |
|---|---|---|---|---|---|
| Dataset 1 | The dataset from Kaggle includes patient medical records with liver health indicators like bilirubin, enzymes, albumin, and diagnosis status, structured for machine learning analysis. | https://www.kaggle.com/datasets/bhavanipriya222/liver-cirrhosis-prediction | excel | 214 KB | Public |