Data Collection and Preprocessing Phase

Date	10/07/2024
Team ID	team-739866
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	Project: Predicting Liver Cirrhosis Disease using Machine Learning Objective: 1. Develop a predictive model that can accurately identify patients at risk of developing liver cirrhosis. 2. Utilize a combination of demographic, medical history, laboratory results, and ultrasound findings to predict the likelihood of liver cirrhosis.

	3. Improve early detection and diagnosis of liver cirrhosis, enabling timely interventions and potentially improving patient outcomes. 4. Explore the relationships between various risk factors and liver cirrhosis to inform future research and clinical decision-making.
Data Collection Plan	Collected the data from kaggle HealthCare dataset
Raw Data Sources Identified	1. Electronic Health Records (EHRs) - Source: Healthcare providers, hospitals, and clinics - Content: Patient demographics, medical history, medications, lab results, and clinical notes 2. Hospital Information Systems (HIS) - Source: Hospital databases - Content: Patient admissions, diagnoses, treatments, outcomes, and discharge summaries 3. Laboratory Information Systems (LIS) - Source: Laboratory databases - Content: Lab test results, including liver function tests, complete blood counts, and other relevant parameters 4. Radiology Information Systems (RIS) - Source: Radiology departments

	- Content: Imaging reports, including liver scans, biopsies, and other relevant radiology results 5. Patient Questionnaires and Surveys - Source: Patient self-reports - Content: Lifestyle habits, alcohol consumption, family history, and other relevant factors 6. Public Health Databases - Source: National and international health organizations - Content: Aggregated data on liver disease prevalence, risk factors, and outcomes 7. Research Studies and Clinical Trials - Source: Academic institutions, research centers, and clinical trial databases - Content: Data from studies on liver cirrhosis, including genetic and molecular markers 8. Healthcare Claims Data - Source: Insurance companies and healthcare payers - Content: Patient outcomes, treatments, and resource utilization data

	These raw data sources will be used to extract relevant features and build a comprehensive dataset for training and validating the machine learning model.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
HealthCare	Collection of patient details	Link: https://www.kaggle.com/datasets/bhavanipriya222/liver-cirrhosis-prediction	CSV	XX GB	Public
…	…	…	…	…	…