Data Collection and Preprocessing Phase

Date	15 June 2024
Team ID	740770
Project Title	Disease prediction using machine learning
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	The Disease Prediction model project aims to develop a machine learning tool to predict the likelihood of a specific disease based on patient data such as age, gender, medical history, and symptoms. The project involves collecting and preprocessing data, developing and evaluating a predictive model using algorithms like logistic regression or decision trees, and deploying the model in a user-friendly application for initial use by healthcare professionals. The goal is to provide an early detection mechanism that improves patient outcomes and supports preventive healthcare measures.
Data Collection Plan	*__Methods__: Extract data from existing EHR systems, conduct health surveys, or gather data from health apps. *__Techniques__: API integration for EHRs, online survey tools, data

	extraction scripts.
Raw Data Sources Identified	Gathered a dataset from Kaggle containing patient information such as age, gender, symptoms, and medical history for disease prediction. The dataset includes features relevant for building and training the prediction model, enabling accurate risk assessments and analysis

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle Dataset	The dataset compriscs list of symptoms and target column	https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning	CSV	1348 kb	Public