Data Collection and Preprocessing Phase

Data Collection Plan & Raw Data Sources Identification Report:

Data Collection Plan:

Date	18 June 2024
Team ID	739634
Project Title	Flight Delays Prediction Using Machine Learning
Maximum Marks	2 Marks

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.

Raw Data Sources Report:

Section	Description
Project Overview	The machine learning project aims to predict flight delays based on various factors such as flight details, weather conditions, and air traffic control data. Using a dataset with features such as departure time, arrival time, weather conditions, and more, the objective is to build a model that accurately predicts flight delays, facilitating efficient scheduling and enhanced customer satisfaction.
Data Collection Plan	- Search for datasets related to flight schedules, delays, weather conditions, and airport operations. - Use APIs from airlines and weather services to gather real-time data - Prioritize datasets that include comprehensive historical delay data and weather information.
Raw Data Sources Identified	The raw data sources for this project include datasets obtained from Kaggle & UCI, the popular platforms for data science competitions and repositories. The provided sample data represents a subset of the collected information, encompassing variables such as, YEAR,QUARTER ,MONTH,DAY_OF_MONTH,DAY_OF_WEEK,UNIQUE_CARRIER,TAIL_NUM ,FL_NUM,ORIGIN_AIRPORT_ID ORIGIN,DEST_AIRPORT_ID,DEST,CRS_DEP_TIME,DEP_TIME,DEP_DELAY,DEP_DEL15,CRS_ARR_TIME,ARR_TIME,ARR_DELAY ARR_DEL15,CANCELLED,DIVERTED,CRS_ELAPSED_TIME,ACTU

	AL_ELAPSED_TIME,DISTANCE				
Source Name	**Description**	**Location/URL**	**Format**	**Size**	**Access Permissions**
Kaggle.com,data.gov,UCI	The dataset comprises applicant details (YEAR,QUARTER , MONTH,DAY_OF_MONTH,DAY_OF_WEEK,UNIQUE_CARRIER,TAIL_NUM,FL_NUM,ORIGIN_AIRPORT_ID,ORIGIN,DEST_AIRPORT_ID,DEST,CRS_DEP_TIME, DEP_TIME,DEP_DELAY,DEP_DEL15,CRS_ARR_TIME,ARR_TIME ,ARR_DELAY ARR_DEL15,CANCELLED,DIVERTED,CRS_ELAPSED_TIME ,ACTUAL_ELAPSED_TIME,DISTANCE), and flight delays outcomes.	https://drive.google.com/file/d/1HNYx6fX5hvRDX43egcAAUsrQ9sccv4AR/view	CSV	15 kB	Public