Data Collection and Preprocessing Phase

Date	15 July 2024
Team ID	740080
Project Title	Flight Delay 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 Machine Learning project aims to predict flight delay based on the customer information. Using a dataset with features such as flight number, day of month, day of week, month, arrival time, departure time, actual arrival time, origin and destination the objective is to build a model that accurately classifies flight status(will be on time or flight will be delay) facilitating efficient and informed decision making in the flight delay process.
Data Collection Plan	- Search for datasets related to flight delays , airports delays. - Prioritize the datasets with diverse demographic information.
Raw Data Sources Identified	The raw data sources for this project include datasets obtained from kaggle.com, & UCI, the popular platforms for repositories.

	The provided sample dataset represents a subse of collected information , encompassing variables like YEAR, DAY OF MONTH, DAY OF WEEK, ORIGIN, DESTINATION, ARRIVAL TIME, DEPARTURE TIME , ACTUAL CRS_TIME.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle.com, data.gov, UCL	The dataset comprises of applicant details (YEAR, QUARTER, DAY OF WEEK, Day OF MONTH, MONTH, ORIGIN, DESTINATION,ARRIVAL TIME,DEPARTURE TIME,TAIL_NUM, FL_NUM,ORIGIN_AIRPORT_ID,DEST_AIRPORT_ID,UNIQUE_CARRIER,	flightdata.csv - Google Drive	CSV	15KB	PUBLIC

	CRS_DEPT_TIME, DEP_TIME,ARR_TIME.				