

Data Collection and Preprocessing Phase

Date	12 March 2024
Team ID	SWTID1720089323
Project Title	Ecommerce Shipping 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 project involves analyzing an e-commerce dataset with features such as shipping mode, customer details, product details, and discount offers to build a predictive model that determines if an order will be delivered on time. The objectives include data preprocessing (handling missing values, encoding categorical variables, scaling, and normalization), feature engineering (creating and selecting relevant features), model building (training and optimizing various models like RandomForest, K-Nearest Neighbors, and Logistic Regression), and model evaluation (using metrics like Accuracy, F1 Score, Recall, and Precision to select the best model). The project also aims to interpret results to provide actionable insights and develop a user interface for practical application. By achieving these objectives, the project aims to provide a reliable tool for e-commerce companies to anticipate delivery delays and ensure timely deliveries, thus improving customer satisfaction and operational efficiency.

Data Collection Plan	<ul style="list-style-type: none"> E-commerce Transaction Records and Customer Databases: These sources provide detailed information on orders, customer demographics, and purchasing behavior. Exploring and considering the available recollected data sources from the online data sources like Kaggle, GitHub etc.
Raw Data Sources Identified	The raw data sources for this project include datasets obtained from Kaggle and UCI, popular platforms for data science competitions and repositories. The provided sample data represents a subset of the collected information, encompassing variables such as shipping mode, customer details, product details, and discount offers for machine learning analysis.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle Dataset	The dataset comprises applicant details (shipping mode, customer details, product specifics, and discount offers). These variables are essential for conducting machine learning analyses aimed at predicting shipping times and optimizing e-commerce operations.	https://www.kaggle.com/datasets/prachi13/customer-analytics/data	CSV	431 kB	Public