Data Collection and Preprocessing Phase

Date	9 July 2024
Team ID	SWTID1720084775
Project Title	Ecommerce Shipping Prediction using Machine Learning
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification

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	We're building a machine learning model to predict e-commerce delivery times. This will optimize logistics, potentially reducing costs, and improve customer satisfaction by setting realistic expectations and allowing for proactive communication about delays. Overall, it enhances the customer experience and gives us a competitive edge.
Data Collection Plan	The dataset includes information collected from Kaggle.
Raw Data Sources Identified	Developing a predictive model using customer and order details to forecast delivery reliability (1 for late, 0 for on time). This approach leverages customer history, order specifics like processing times and shipping methods, and product details to optimize logistics and

	improve future delivery predictions. The goal is to boost customer satisfaction by ensuring timely deliveries based on comprehensive data analysis.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Dataset 1	The "Customer Analytics" dataset on Kaggle provides comprehensive information about customer transactions, including the following attributes: customer ID, warehouse block (A, B, C, D, F), mode of shipment (Ship, Flight, Road), number of customer care calls, customer rating (1 to 5), cost of the product (in USD), number of prior purchases, product	https://www.kaggle.com/datasets/prachi13/customer-analytics?select=Train.csv	CSV	440.46 kB	Public

	importance (low, medium, high), gender, discount offered, product weight (in grams), and whether the product reached on time (1 for not on time, 0 for on time). This dataset is ideal for analyzing various aspects of customer service and logistics.				