Data Collection and Preprocessing Phase

Date	20 July 2024
Team ID	SWTID1720110595
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	Ecommerce shipping prediction is the process of estimating the whether the product reached on time. which is based on various factors such as the origin and destination of the package, the shipping method selected by the customer, the carrier used for shipping, and any potential delays or issues that may arise during the shipping process. Machine learning models can be used to make accurate predictions about shipping times based on historical data and real-time updates from carriers. These models may take into account factors such as weather conditions, traffic, and other external factors that can impact delivery times.

Data Collection Plan	There are many popular open sources for collecting the data. Eg: kaggle.com, UCI repository, etc. In this project we have used .csv data. This data is downloaded from kaggle.com Link: https://www.kaggle.com/datasets/prachi13/customer-analytics?select=Train.csv
Raw Data Sources Identified	The data contains the following information: • **ID:** ID Number of Customers. • **Warehouse block:** The Company have big Warehouse which is divided in to block such as A,B,C,D,E. • **Mode of shipment:**The Company Ships the products in multiple way such as Ship, Flight and Road. • **Customer care calls:** The number of calls made from enquiry for enquiry of the shipment. • **Customer rating:** The company has rated from every customer. 1 is the lowest (Worst), 5 is the highest (Best). • **Cost of the product:** Cost of the Product in US Dollars. • **Prior purchases:** The Number of Prior Purchase. • **Product importance:** The company has categorized the product in the various parameter such as low, medium, high. • **Gender:** Male and Female. • **Discount offered:** Discount offered on that specific product. • **Weight in gms:** It is the weight in grams. • **Reached on time:** It is the target variable, where 1 Indicates that the product has NOT reached on time and 0 indicates it has reached on time.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle Datset	The dataset contains • ID • Warehouse_block • Mode_of_Shipment • Customer_care_calls • Customer_rating	https://www.kaggle.com/datasets/prachi13/customer-analytics?select=Train.csv	CSV	440.46 kB	Public

	• Cost_of_the_Product • Prior_purchases • Product_importance • Gender • Discount_offered • Weight_in_gms • Reached_on_Time_Y_N				