**E-commerce Data Analysis Report**

**Overview**

The project focuses on analyzing data from an international e-commerce company specializing in electronic products. The aim is to derive key insights from the customer database using advanced machine learning techniques to study customer behavior and product delivery performance.

**Dataset Description**

The dataset contains 10,999 observations with 12 variables:

* **ID**: Customer ID number
* **Warehouse block**: Warehouse divided into blocks (A, B, C, D, E)
* **Mode of shipment**: Shipping method (Ship, Flight, Road)
* **Customer care calls**: Number of customer service calls
* **Customer rating**: Customer rating (1 to 5)
* **Cost of the product**: Cost in USD
* **Prior purchases**: Number of previous purchases
* **Product importance**: Product priority (low, medium, high)
* **Gender**: Customer gender (Male, Female)
* **Discount offered**: Discount on the product
* **Weight in grams**: Product weight
* **Reached on time**: Delivery status (1 = Not on time, 0 = On time)

**1. Data Exploration and Cleaning**

**Missing Values and Duplicates**

* **Missing Values**: No missing values were found.
* **Duplicates**: No duplicate rows were present.

**Summary Statistics**

* **Number of Rows and Columns**: 10,999 rows and 12 columns.
* **Average Customer Calls**: 4 calls.
* **Average Customer Rating**: 2.9.
* **Average Cost of Product**: $210.
* **Basic Statistics**: Refer to the dataset description.

**Data Unique Values and Counts**

* **Shipment Modes**:
  + Ship: 7,462
  + Flight: 1,777
  + Road: 1,760
* **Delivery Status**:
  + On Time: 4,436
  + Late: 6,563

**2. Data Visualization**

**Customer Ratings by Gender**

* **Finding**: Females rated products slightly lower than males.

**Prior Purchases by Gender**

* **Finding**: Most customers made 3 prior purchases.

**Customer Care Calls by Gender**

* **Finding**: Most customers made 3 to 4 calls.

**Prior Purchases vs. Delivery Time**

* **Finding**: Orders frequently did not reach on time.

**Customer Ratings vs. Delivery Time**

* **Finding**: Delivery delays had minimal impact on customer ratings.

**Product Importance**

* **Finding**: Most products are categorized as low priority.

**Product Importance vs. Delivery Time**

* **Finding**: High-priority products also experienced delays.

**Delivery Status Count**

* **Finding**: More than 6,000 products were delayed.

**Mode of Shipment vs. Delivery Time**

* **Finding**: Shipments, whether on time or delayed, were mostly by ship.

**Cost of Products**

* **Finding**: Most products are priced around $270.

**Discounts Offered**

* **Finding**: Discounts are usually less than 10%.

**Customer Care Calls vs. Delivery Time**

* **Finding**: Increased customer care calls correspond with late deliveries.

**Product Weight Distribution**

* **Finding**: Product weight distribution shows density across various weights.

**4. Model Building and Evaluation**

**Comparing the model accuracy of the models used**

* The accuracy of the Decision tree model is 0.6575757575757576 i.e. 65.75%
* The accuracy of the Random Forest classifier model is 0.6619834710743802 i.e. 66.19%
* The accuracy of the KNN model is 0.6586776859504132 i.e. 65.86%
* Accuracy after tuning 0.6741046831955922 i.e. 67.41%

**Finding :** The Random Forest Classifier model outperformed the other models in predicting on-time deliveries at 66.19 % python

Random Forest's performance was further improved using RandomizedSearchCV, leading to a higher accuracy of 67.41%.

**Conclusion:**

This project successfully explored and analyzed the E-commerce dataset. Key insights were derived through visualizations, and predictive models were built to forecast delivery outcomes. Random Forest proved to be the most effective model, and hyperparameter tuning further enhanced its accuracy.