Concept Note:

(E-Commerce Product Delivery Prediction)

Objective:

The aim of this project to predict whether the product from an e-commerce company will reach on time or not. This project also analyses various factors that affect the delivery of the product as well as studies the customer behaviour.

Rationale:

IN the era of the big data, businesses are generating and collecting the vast amounts of the data from various source such as transactions, customer interactions, social media etc. Here from E-commerce datasheet which we use for the predicting the delivery time. the data of from various variable use here . Data analytics offers powerful tools and methodologies to extract valuable information, enabling businesses to make informed decisions.

Methodology:

1. Data Collection and Preprocessing:

- **Data Sources:** Gather data from website called KAGGLE as csv formatted datasheet.
- **Data Cleaning:** Handle missing values, outliers, and inconsistencies through data cleaning techniques.

2. Data Analysis:

- **Exploratory Data Analysis (EDA)**: This involves summarizing the main characteristics of the data, often with visual methods. EDA helps in understanding the distribution of data, detecting outliers, and identifying relationships between variables.
- **Descriptive Statistics**: Basic statistical techniques such as mean, median, mode, and standard deviation are used to describe the central tendency and variability of the dataset.
- **Predictive Analysis**: Machine learning models are applied to predict the likelihood of on-time delivery. This involves using supervised learning techniques where the model is trained on historical data to predict future outcomes.

- **Correlation Analysis**: Identifying and quantifying the relationships between different variables in the dataset. This helps in understanding how different factors are interrelated and influence delivery times.
- **Feature Engineering**: Creating new features or modifying existing ones to improve the performance of machine learning models. This includes selecting the most relevant features for prediction.

3. Model Development:

Algorithm Selection:

- Regression: If predicting continuous variables, algorithms like Linear Regression, Ridge Regression, or Lasso Regression might be chosen.
- Classification: For predicting categorical outcomes (e.g., on-time vs. late delivery), algorithms like Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, or Support Vector Machines (SVM) can be used.
- Clustering: For grouping similar data points, clustering algorithms like K-Means, Hierarchical Clustering, or DBSCAN might be utilized.

Training and Validation:

- Dataset Splitting: Divide the dataset into training and validation sets, typically using a split ratio like 70:30 or 80:20.
- Training Models: Train the selected algorithms on the training set, allowing them to learn patterns and relationships in the data.
- Evaluation: Assess the models' performance on the validation set using metrics appropriate to the problem (e.g., accuracy, precision, recall, F1-score for classification; mean squared error or R-squared for regression).

Model Tuning:

- Hyperparameter Optimization: Use techniques like Grid Search, Random Search, or Bayesian Optimization to find the best hyperparameters for the models.
- Cross-Validation: Implement k-fold cross-validation to ensure the models are robust and perform well across different subsets of the data.
- Regularization: Apply regularization techniques like L1 or L2 regularization to prevent overfitting and improve model generalization.
- Feature Selection: Identify and select the most important features that contribute to the model's performance, potentially reducing dimensionality and improving accuracy.

4. Implementation and Testing:

Customer Care Calls:

- Increased customer care calls are associated with delays in delivery. Customers tend to call more frequently when there are delays, indicating that they are anxious about the status of their orders.
- Visual Insight: The graph showing the number of customer care calls reveals that as the number of calls increases, the difference between on-time and late deliveries decreases.

Customer Rating and Prior Purchases:

- Higher customer ratings correlate with on-time deliveries.
- Customers with more prior purchases are more likely to receive their products on time, suggesting loyalty and satisfaction with the service.
- Visual Insight: Graphs demonstrate that higher customer ratings and more prior purchases are linked to a higher count of on-time deliveries.

Discount Offered:

- Products with higher discounts (more than 10%) are more likely to be delivered on time compared to those with lower discounts (0-10%).
- Visual Insight: The distribution of discounts shows that products with higher discounts tend to be delivered on time more frequently.

Product Properties:

- Product weight and cost significantly impact delivery times. Heavier products and those costing more than $250 are more likely to be delivered late.
- Visual Insight: The violin plots and count plots illustrate the distribution and impact of weight and cost on delivery times.

Logistics and Shipment:

- The mode of shipment and the warehouse block do not have a significant impact on the timely delivery of products.
- Visual Insight: Graphs comparing warehouse blocks and modes of shipment with delivery times show no significant difference in on-time and late deliveries.

Gender:

- There is no significant difference in delivery times based on customer gender.
- Visual Insight: The count plot for gender shows an equal distribution of on-time and late deliveries across both genders.

Correlation Matrix:

- The heatmap indicates a positive correlation between the cost of the product and the number of customer care calls, highlighting that customers are more concerned about the delivery of higher-cost products.

Model Building:

- Four machine learning models were considered: Random Forest Classifier, Decision Tree Classifier, Logistic Regression, and K Nearest Neighbours.
- The Random Forest Classifier with hyperparameter tuning through GridSearchCV showed the best performance with a training accuracy of approximately 72.53%.

5. Deployment:

The Random Forest model results for the E-Commerce Product Delivery Prediction project include the best parameters obtained from GridSearchCV and the corresponding training accuracy.

Best Parameters for Random Forest Classifier:

Criterion: Gini

Max Depth: 8

Min Samples Leaf: 8

Min Samples Split: 2

Random State: 42

Training Accuracy:

Training accuracy: 0.7253 (approximately 72.53%)

For the Decision Tree Classifier:

Best Parameters for Decision Tree Classifier:

Criterion: Gini

Max Depth: 6

Min Samples Leaf: 6

Min Samples Split: 2

Random State: 0

Training Accuracy:

Training accuracy: 0.6913 (approximately 69.13%)

These results indicate that both models have been fine-tuned using GridSearchCV to find the optimal parameters. The Random Forest model achieved a slightly higher training accuracy compared to the Decision Tree model.

Dataset:

The data set contains the 10999 rows and 12 variables.

Variable	Description
ID	ID Number of Customers
Warehouse_block	The Company have big Warehouse which is divided into 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_Y.N	It is the target variable, where 1 Indicates that the product has NOT reached on time and 0 indicates it has reached on time

Expected Outcomes:

- Identification of key business drivers and their impact on performance.
- Enhanced ability to delivery time for the customer.
- Improved decision-making processes through data-driven insights.
- Increased customer satisfaction through personalized experiences and services

Conclusion:

- The aim of the project was to predict whether the product from an e-commerce company will reach on time or not. This project also analyses various factors that affect the delivery of the product as well as studies the customer behaviour.
- From the exploratory data analysis, I found that the product weight and cost have an impact on the product delivery. Where product that weighs between 2500 - 3500 grams and having cost less than 250 dollars had higher rate of being delivered on time. Most of the products were shipped from warehouse F though ship, so it is quite possible that warehouse F is close to a seaport. The customer's behaviour also helps in predicting the timely delivery of the product. The more the customer calls, higher the chances the product delivery is delayed. Interestingly, the customers who have done more prior purchases have higher count of products delivered on time and this is the reason that they are purchasing again from the company. The products that have 0-10% discount have higher count of products delivered late, whereas products that have discount more than 10% have higher count of products delivered on time.
- Coming to the machine learning models, the decision tree classifier as the highest accuracy among the other models, with accuracy of 69%.
- The random forest classifier and logistic regression had accuracy of 68% and 67% respectively. The K Nearest Neighbours had the lowest accuracy of 65%.