## **Project Report: Predicting Customer Reviews with Random Forest Classification**

**1. Introduction**

This project aims to develop a machine learning model to predict customer reviews for products sold on an e-commerce platform. The model will analyze various factors related to the order and product details to identify patterns that influence customer satisfaction. This information can be valuable for businesses to improve product offerings, personalize recommendations, and enhance the overall customer experience.

**2. Data Acquisition and Preprocessing**

The project utilizes a collection of datasets provided by an e-commerce platform, including:

* Orders: Information about customer purchases, including order ID, product ID, seller ID, and order value.
* Products: Details about products offered, including product ID, price, and category.
* Customers: Customer information such as demographic data and location.
* Additional datasets: Order items, payments, reviews, and geolocation data.

These datasets are merged into a single comprehensive dataset to capture the relationships between various entities involved in a customer's purchase and review behavior. Following the merge, data cleaning is performed to address missing values.

**3. Feature Engineering and Target Variable**

* **Target Variable:** Reviews are converted into a binary classification task. Reviews with a score of 4 or higher are considered positive and assigned a value of 1, while others are labeled as negative and assigned a value of 0.
* **Features:** Selected features from the merged dataset include product ID, seller ID, product price, freight value, and product category name.
* **Categorical Encoding:** Product category names are converted into one-hot encoded features to represent them numerically for the model.

**4. Model Selection and Training**

* **Model Choice:** A Random Forest Classifier is chosen due to its robustness to overfitting, ability to handle mixed-type features, and interpretability of results.
* **Data Split:** The data is divided into training and testing sets using a split ratio of 80/20. The training set is used to train the model, and the testing set evaluates its performance on unseen data.
* **Normalization:** StandardScaler is applied to standardize the features in the training and testing sets, ensuring all features contribute equally to the model's predictions.
* **Model Training:** The Random Forest Classifier is trained on the prepared training data (features and target variables).

**5. Model Evaluation**

* **Evaluation Metrics:** The trained model's performance is evaluated using metrics like accuracy score and classification report.
* **Accuracy Score:** This metric measures the percentage of correct predictions made by the model on the testing set.
* **Classification Report:** This provides a detailed breakdown of the model's performance, including precision, recall, F1-score, and support for each class (positive and negative reviews).

**6. Conclusion**

This project demonstrates the use of machine learning to predict customer reviews for e-commerce products. By analyzing various factors related to the purchase and product details, the model can identify trends that influence customer satisfaction. The results of this project can be used to enhance product offerings, personalize recommendations, and improve the overall customer experience.

**7. Future Work**

* Explore the use of different machine learning algorithms for comparison with the Random Forest Classifier.
* Implement feature selection techniques to identify the most impactful features on review prediction.
* Investigate techniques for handling imbalanced datasets if the distribution of positive and negative reviews is uneven.
* Integrate the model into a web application for real-time review prediction and recommendation generation.