

Assignment-4

Name: Shukla Kushang Akshay

Email ID: kushangashukla1@gmail.com

Mobile No.: +91 8511350065

Enrollment No.: 221130107024

College Name: 113-Sal College of Engineering

Branch: Computer Engineering

Semester- 6th

Faculty Code: T00111- Mikin Dagli Sir

- **Project Report:**

- Predicting Order Delivery Time Using Product Specifications.
- Drive Link for the Project:
<https://drive.google.com/drive/folders/16if5chK0x0YpiP-r8y6UCI1U1Uc2bW8Q?usp=sharing>

- **Introduction:**

- In this project, we aim to predict the delivery time for new product orders based on product specifications and other related factors. The goal is to develop a predictive model that accurately estimates the delivery time for new product orders.

- **Objective:**

- Build a predictive model to estimate delivery time for new product orders.
- Perform exploratory data analysis (EDA) to identify trends and patterns.
- Handle missing values, create new features, and clean the dataset.
- Evaluate the performance of multiple machine learning models.

- **Dataset Overview:**

- The following datasets were used in the project:
 - olist_orders_dataset.csv
 - olist_order_items_dataset.csv
 - olist_order_reviews_dataset.csv

- olist_order_payments_dataset.csv
- olist_products_dataset.csv
- olist_sellers_dataset.csv
- product_category_name_translation.csv
- olist_customers_dataset.csv
- olist_geolocation_dataset.csv

- **Data Cleaning:**

- Handled missing values using median and mode.
- Dropped rows with critical missing values.
- Converted timestamps to datetime format.
- Ensured numeric fields are of type float or int.

- **Feature Engineering:**

- Delivery Time = Difference between delivery and purchase date.
- Payment Ratio = $\text{payment_value} \div \max(\text{payment_value})$.
- Review Sentiment = Based on review score.
- Price Range = Categorized based on price.

- **Model Development:**

- Models Used:
 - Linear Regression
 - Random Forest
 - Gradient Boosting Regressor
 - Decision Tree
 - XGBoost

- **Model Results:**

- Linear Regression: $R^2 = 0.0031$, $MSE = 0.0850$
- Random Forest: $R^2 = 0.26$, $MSE = 0.74$
- Gradient Boosting: $R^2 = 0.2571$, $MSE = 0.7412$, $MAE = 0.6686$
- Decision Tree: $R^2 = 0.59$, $RMSE=157.69$, $MSE = 24863.56$
- XGBoost: $R^2 = 0.82$, $RMSE=103.65$, $MSE = 10743.32$

- **Conclusion:**

- XGBoost was the best-performing model with an R^2 score of 0.82 and MSE of 10743.32. The newly created features significantly improved the model's predictive capability.

