Project Summary – Taxi Tip Prediction Using XGBoost

This project focuses on predicting taxi trip tip amounts using machine learning techniques. The dataset used was the 2023 Yellow Taxi Trip Data, containing trip details such as distance, time, passenger count, fare amount, and additional charges like airport fees and congestion surcharges.

The data preprocessing stage involved cleaning and preparing the dataset to ensure quality and accuracy. Missing values in key columns like airport_fee and congestion_surcharge were replaced with zero, while passenger_count null values were imputed using the mean value. Redundant or irrelevant columns such as store_and_fwd_flag were dropped, and the data was shuffled to avoid bias during training.

After preprocessing, the dataset was split into training and testing subsets using an 80-20 ratio. An XGBoost Regressor model was then trained on the cleaned data to predict the tip_amount variable. The model's performance was evaluated using the Mean Squared Error (MSE) metric, which indicated that predictions were accurate within a small margin of error (approximately ±1.14).

Overall, the project successfully demonstrates the complete machine learning pipeline — from data cleaning and preparation to model training and evaluation — showcasing proficiency in Python, pandas, scikit-learn, and XGBoost for real-world regression tasks.