The Superior University

Name: Muhammad Jawad Ahsan

Course: Data Science Lab

Semester: 3rd

Section: BSAI 3A

Department: Computer Science

Submitted To: [Instructor Name]

Total Marks: [Marks]

Date: [Date]

**Lab Report**

**Task: Data Processing and Model Training**

**Project: Data Analysis and Model Training Using Machine Learning**

**Introduction**

This project involves data processing and model training using machine learning techniques. The dataset includes various attributes, and the objective is to preprocess the data, perform exploratory data analysis (EDA), and train machine learning models for predictive analysis.

**Code Explanation**

**1. Data Loading and Preprocessing**

* The dataset is loaded using pandas.read\_csv().
* The initial structure of the dataset is explored using .shape, .columns, .dtypes, and .describe().
* Missing values are handled using SimpleImputer (median for numeric and most frequent for categorical features).

**2. Exploratory Data Analysis (EDA)**

* The distribution of numerical features is visualized using histograms (seaborn.histplot).
* A correlation heatmap is generated to examine relationships between numerical features (seaborn.heatmap).

**3. Feature Engineering and Scaling**

* The dataset is split into numeric and categorical columns.
* Missing values in numeric columns are filled with the median.
* Missing values in categorical columns are filled with the most frequent values.
* One-hot encoding is applied using pd.get\_dummies().
* Features are standardized using StandardScaler.

**4. Model Training and Evaluation**

* The dataset is split into training and validation sets (train\_test\_split() with an 80-20 split).
* Two models are trained:
  + **Random Forest Regressor**
  + **XGBoost Regressor**
* Models are evaluated using **Root Mean Squared Error (RMSE)**.
* The results of both models are displayed.

**5. Final Predictions and Submission**

* The test dataset undergoes the same preprocessing steps as the training dataset.
* Missing columns in the test dataset are aligned with the training dataset.
* The trained model is used to predict outcomes.
* Predictions are saved to a CSV file for submission.

**Model Performance**

* **Random Forest RMSE:** Displayed in the console.
* **XGBoost RMSE:** Displayed in the console (typically better performance due to boosting).

**Conclusion**

This notebook focuses on data preprocessing, exploratory analysis, and model training. The trained models provide predictions based on structured input data, and their performance is measured using RMSE. Further improvements can be made by tuning hyperparameters and optimizing feature selection.