ML UBER EXP 1

What is the goal of this assignment?
To predict Uber ride fares based on pickup and drop-off locations using linear regression and random forest models.

What is data preprocessing, and why is it essential?
Data preprocessing involves cleaning and formatting raw data to make it suitable for machine learning, enhancing model accuracy and efficiency.

Explain what outliers are in data analysis.
Outliers are data points that deviate significantly from the rest, potentially indicating errors or unusual observations.

What is linear regression?
Linear regression is a statistical method to predict a continuous variable based on the linear relationship between dependent and independent variables.

Describe random forest regression.
Random forest regression is an ensemble method that averages predictions from multiple decision trees for better accuracy and reduced overfitting.

What is underfitting?

Underfitting occurs when a model is too simple to capture the underlying data patterns, resulting in poor performance.

What is overfitting?

Overfitting happens when a model learns noise or random error in the training data rather than the actual pattern, reducing its generalization ability.

What are the reasons for overfitting?

Reasons include a complex model, noisy data, too many features, or small training set.

How to prevent overfitting?

Techniques to prevent overfitting include regularization, cross-validation, pruning (for trees), and using simpler models.

How is correlation useful in this assignment?
Correlation helps identify relationships between variables, crucial for selecting features in predictive modeling.

What is the Haversine formula used for?
It calculates the shortest distance between two points on a sphere, often used in navigation.

How does the boxplot help in data analysis?
A boxplot visualizes data distribution and identifies outliers by displaying minimum, maximum, and quartiles.

What are the main steps of data preprocessing?
The steps include data collection, missing value handling, encoding categorical data, splitting the dataset, and feature scaling.

What is the role of feature scaling in preprocessing?
Feature scaling standardizes data, ensuring that variables have similar scales to improve model performance.

How does a random forest avoid overfitting?
By averaging predictions from multiple decision trees, random forests reduce the risk of overfitting to specific training data points.

What does the R^2 score represent?
R^2 indicates the proportion of variance in the dependent variable explained by the model, with values closer to 1 being preferable.

What is RMSE, and why is it used?
Root Mean Squared Error (RMSE) measures prediction errors; lower values indicate better model performance.

How does one identify outliers using a boxplot?
Data points outside the “whiskers” (minimum and maximum limits) of a boxplot are potential outliers.

What are the types of outliers?
Outliers include global, collective, and contextual, each differing by their deviation and context in data.

Why are libraries like NumPy and Pandas important for data science?
They offer functions for handling arrays and data manipulation, streamlining data preprocessing and analysis.

Explain contextual outliers with an example.
Contextual outliers deviate based on context, such as high temperatures during winter months being unusual.

Why is model evaluation necessary?
It assesses how well a model performs, helping refine and choose models based on accuracy and error metrics.

What types of visualizations can Matplotlib generate?
It can create line, bar, scatter, and histogram plots, aiding in data interpretation.

How can linear regression predict fare prices?
By establishing a linear relationship between fare and influencing factors, linear regression can estimate prices.

Question: What is Mean Squared Error (MSE), and why is it important?
Mean Squared Error (MSE) measures the average squared difference between predicted and actual values, helping to quantify prediction accuracy. Lower MSE values indicate better model performance, as they signify smaller errors in predictions.

Question: What are the types of outliers, and how do they differ?

Global Outliers: These are individual data points that deviate significantly from the entire dataset. For example, in Uber fare data, an unusually high fare might be a global outlier.

Collective Outliers: These occur when a group of data points together behaves abnormally, although each point individually may not be unusual. For instance, a series of high fares at midnight could be a collective outlier.

Contextual Outliers: These depend on the data context; a value is considered an outlier only under certain conditions. An example would be high temperatures in winter, which may be typical in summer but unusual for winter.