1. What are the key tasks that machine learning entails? What does data pre-processing imply?

Machine Learning Tasks:

Data Collection: Gather relevant data for analysis.

Data Cleaning: Handle missing values, outliers, and inconsistencies.

Data Exploration: Understand data distributions, patterns, and relationships.

Feature Engineering: Create new features or transform existing ones.

Data Splitting: Divide data into training and testing sets.

Model Selection: Choose an appropriate machine learning model.

Model Training: Train the selected model on the training data.

Model Evaluation: Assess the model's performance on test data.

Hyperparameter Tuning: Optimize model parameters for better performance.

Deployment: Implement the model for real-world use.

Data Pre-processing:

Data Cleaning: Handling missing values, outliers, and errors.

Data Transformation: Scaling, normalization, and encoding categorical variables.

Data Reduction: Reducing dimensionality to improve model efficiency.

Data Integration: Combining data from multiple sources.

Data Discretization: Converting continuous data into discrete bins.

2. Describe quantitative and qualitative data in depth. Make a distinction between the two.

Quantitative Data:

Definition: Numerical, measurable values.

Examples: Height, weight, income.

Qualitative Data:

Definition: Categorical, non-numeric values.

Examples: Color, gender, country.

3. Create a basic data collection that includes some sample records. Have at least one attribute from each of the machine learning data types.

Sample Records:

ID Name Age Income Gender

1 John 28 50000 Male

2 Alice 35 60000 Female

3 Bob 22 45000 Male

4. What are the various causes of machine learning data issues? What are the ramifications?

Causes:

Missing values, outliers, errors, inconsistent formats.

Ramifications:

Biased models, inaccurate predictions, reduced model performance.

5. Demonstrate various approaches to categorical data exploration with appropriate examples.

Approaches:

Frequency Counts: Count occurrences of each category.

Bar Charts: Visualize the distribution of categories.

Cross-Tabulation: Explore relationships between two categorical variables.

6. How would the learning activity be affected if certain variables have missing values? Having said that, what can be done about it?

Impact on Learning:

May lead to biased models, reduced accuracy.

Handling:

Imputation (mean, median, mode), deletion, advanced imputation methods.

7. Describe the various methods for dealing with missing data values in depth.

Deletion:

Remove records or features with missing values.

Imputation:

Fill missing values with statistical measures (mean, median, mode).

Advanced Techniques:

Predict missing values using machine learning models.

8. What are the various data pre-processing techniques? Explain dimensionality reduction and function selection in a few words.

Deletion:

Remove records or features with missing values.

Imputation:

Fill missing values with statistical measures (mean, median, mode).

Advanced Techniques:

Predict missing values using machine learning models.

9.

1. What is the IQR? What criteria are used to assess it?

Definition: Range between the first quartile (Q1) and the third quartile (Q3).

Assessment Criteria: Outliers lie outside the range [Q1 - 1.5 \* IQR, Q3 + 1.5 \* IQR].

1. Describe the various components of a box plot in detail? When will the lower whisker surpass the upper whisker in length? How can box plots be used to identify outliers?

Box: Represents the IQR (Q1 to Q3).

Whiskers: Extend from the box to the smallest and largest non-outlier data points.

Median Line: Line inside the box represents the median.

Outliers: Data points beyond the whiskers.

10. Make brief notes on any two of the following:

1. Data collected at regular intervals

Time series data collected at uniform time intervals (e.g., daily stock prices).

1. The gap between the quartiles

Represents the range between the first quartile (Q1) and the third quartile (Q3).

3. Use a cross-tab

1. Make a comparison between:

1. Data with nominal and ordinal values

Nominal: No inherent order (e.g., colors).

Ordinal: Has a meaningful order (e.g., education levels).

1. Histogram and box plot

Histogram: Represents data distribution; bars indicate frequency.

Box Plot: Displays the median, quartiles, and outliers.

1. The average and median

Average (Mean): Sensitive to outliers, influenced by extreme values.

Median: Less affected by outliers, represents the central value.