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

**Ans:** Key Tasks in Machine Learning:

Data Collection: Gathering relevant data from various sources.

Data Preprocessing: Cleaning, transforming, and preparing data for analysis.

Feature Engineering: Selecting, creating, or transforming features to improve model performance.

Model Selection: Choosing an appropriate machine learning algorithm or model.

Model Training: Training the selected model on the training dataset.

Model Evaluation: Assessing the model's performance using appropriate metrics.

Hyperparameter Tuning: Optimizing model settings for better results.

Model Deployment: Deploying the trained model for making predictions or decisions.

Data Preprocessing: Data preprocessing involves tasks such as handling missing values, dealing with outliers, scaling or normalizing features, and encoding categorical variables to ensure the data is suitable for machine learning models.

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

**Ans:** Quantitative and Qualitative Data:

Quantitative Data: Quantitative data consists of numerical values that can be measured and quantified. It is continuous or discrete. Examples include age, temperature, and income.

Qualitative Data: Qualitative data, also known as categorical data, consists of categories or labels that represent different groups or classes. It is divided into two types:

Nominal Data: Categories have no inherent order, such as colors or vehicle types.

Ordinal Data: Categories have a meaningful order, but the intervals between them may not be uniform, such as education levels (e.g., high school, bachelor's, master's).

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

**Ans:** Basic Data Collection:

Here is a sample data collection with various data types:

Numeric (Quantitative):

Age: [25, 30, 22, 35, 28]

Income: [50,000, 60,000, 45,000, 70,000, 55,000]

Categorical (Qualitative):

Gender: [Male, Female, Female, Male, Male]

Education Level (Ordinal): [Bachelor's, Master's, High School, PhD, Bachelor's]

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

**Ans:** Causes of Machine Learning Data Issues:

Incomplete Data: Missing values or incomplete records.

Noisy Data: Data with errors or outliers.

Imbalanced Data: Unequal distribution of classes in classification problems.

Biased Data: Data that is not representative of the entire population.

Data Scaling: Data with different scales or units.

Irrelevant Features: Features that do not contribute to the prediction.

Ramifications can include inaccurate models, biased predictions, and reduced model performance.

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

**Ans:** Approaches to Categorical Data Exploration:

Frequency Tables: Count the occurrences of each category.

Bar Charts: Visualize the distribution of categorical data.

Cross-Tabulations: Analyze relationships between two categorical variables.

One-Hot Encoding: Convert categorical data into binary columns for modeling.

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

**Ans:** Impact of Missing Values on Learning:

Missing values can lead to biased or inaccurate models.

Learning algorithms may not handle missing data, resulting in errors.

Missing data can affect model training and evaluation.

Handling Missing Values: Options include imputation (replacing missing values with estimates), removal of affected samples or features, or using algorithms that handle missing data.

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

**Ans:** Methods for Dealing with Missing Data:

Imputation: Replacing missing values with estimates (e.g., mean, median, mode).

Deletion: Removing samples or features with missing values.

Interpolation: Estimating missing values based on neighboring data points.

Advanced Imputation: Using machine learning models to predict missing values.

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

**Ans:** Data Preprocessing Techniques:

Dimensionality Reduction: Reducing the number of features while retaining important information. Example techniques include Principal Component Analysis (PCA) and feature selection.

Feature Selection: Selecting a subset of relevant features to improve model efficiency and reduce overfitting.

**9.i. What is the IQR? What criteria are used to assess it?**

**ii. 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?**

**Ans:** i. IQR (Interquartile Range): The IQR is a measure of statistical dispersion and represents the range between the first quartile (25th percentile) and the third quartile (75th percentile) of a dataset. It assesses the spread of the middle 50% of the data.

ii. Box Plot Components:

Box: Represents the IQR, with the central line indicating the median.

Whiskers: Lines extending from the box, typically 1.5 times the IQR in length. Outliers may be shown beyond the whiskers.

Lower Whisker Surpassing Upper Whisker: This occurs when the data is skewed or has outliers in the lower range. It indicates an asymmetric distribution.

Box Plots for Identifying Outliers: Outliers are data points beyond the whiskers and can be visually identified on the plot.

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

**1. Data collected at regular intervals**

**2. The gap between the quartiles**

**3. Use a cross-tab**

**Ans:**1.Data Collected at Regular Intervals: Data collected at equal time intervals (e.g., hourly, daily) is often used for time series analysis and forecasting.

2.Gap Between the Quartiles: The gap between the first quartile (Q1) and the third quartile (Q3) represents the IQR and provides information about data spread.

3.Use a Cross-Tab: Cross-tabulation is used to analyze relationships between two categorical variables by creating a contingency table showing their joint frequencies.

**11. Make a comparison between:**

**1. Data with nominal and ordinal values**

**2. Histogram and box plot**

**3. The average and median**

**Ans:** Comparisons:

1.Data with Nominal and Ordinal Values: Both are categorical data, but ordinal values have a meaningful order, whereas nominal values do not.

2.Histogram and Box Plot: Histograms display data distribution and frequency, while box plots show quartiles and outliers.

3.Average and Median: Both are measures of central tendency, but the average (mean) is affected by outliers, while the median is robust to outliers.