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

**ANSWER:** Key tasks in machine learning are:

Data collection and preprocessing

Feature selection and engineering

Model training and evaluation

Model deployment and monitoring

Data pre-processing: Data pre-processing refers to the steps taken to clean, transform, and organize raw data before it can be used for machine learning. This may involve handling missing values, removing outliers, normalizing or scaling features, encoding categorical variables, and splitting the data into training and testing sets.

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

**ANSWER:** Quantitative data: Quantitative data is numerical data that can be measured or counted. It deals with quantities and values that can be expressed in terms of numbers. Examples include age, height, weight, temperature, and income.

Qualitative data: Qualitative data is descriptive data that cannot be measured or counted numerically. It deals with qualities, characteristics, or attributes. Examples include colors, opinions, preferences, and categorical labels.

The distinction between the two are:

The main distinction is that quantitative data can be expressed numerically and involves measurable quantities, while qualitative data is descriptive and involves non-numerical attributes.

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

**ANSWER:** Causes of machine learning data issues are:

Missing data

Outliers

Imbalanced data

Irrelevant features

Data duplications or inconsistencies

Ramifications of data issues are that Data issues can lead to biased or inaccurate models, reduced performance, misleading insights, and unreliable predictions.

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

**ANSWER:** Approaches to categorical data exploration are:

Frequency analysis: Counting the occurrences of each category.

Cross-tabulation: Analyzing the relationship between two categorical variables.

Bar plots: Visualizing the distribution of categorical variables.

Chi-squared test: Assessing the association between 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?**

**ANSWER:** Impact of missing values on learning activity:

Missing values can lead to biased or incomplete analysis, reduced model performance, and inaccurate predictions. It can also cause issues in feature selection and introduce noise in the data.

Dealing with missing values:

Removal: Remove records or features with missing values.

Imputation: Fill in missing values with estimated values (e.g., mean, median, regression imputation).

Advanced techniques: Use algorithms like k-nearest neighbors (KNN) or expectation-maximization (EM) to impute missing values.

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

**ANSWER:** Methods for dealing with missing data values:

Deletion: Remove the records or features with missing values. This can be done by removing the entire record (listwise deletion) or excluding specific features (column-wise deletion). However, this approach may result in loss of valuable information if the missing values are not randomly distributed.

Mean/median imputation: Replace missing values with the mean or median of the available values for that feature. This method assumes that the missing values have a similar distribution as the observed values.

Mode imputation: Replace missing values with the mode (most frequent value) of the feature. This is suitable for categorical variables.

Regression imputation: Predict missing values using regression models based on other variables.

K-nearest neighbors (KNN) imputation: Estimate missing values by averaging the values of the nearest neighbors in the feature space.

Multiple imputation: Generate multiple plausible imputations to account for uncertainty in missing data. This involves creating multiple complete datasets, each with different imputed values, and then analyzing them using appropriate methods.

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

**ANSWER:** Data pre-processing techniques:

Data cleaning: Handling missing values, correcting errors, and removing outliers.

Data transformation: Normalizing or scaling numerical features to a common scale, such as min-max scaling or z-score normalization.

Encoding categorical variables: Converting categorical data into numerical representations, such as one-hot encoding or label encoding.

Dimensionality reduction: Reducing the number of features while preserving important information. Techniques like Principal Component Analysis (PCA) or feature selection methods can be used.

Feature selection: Selecting a subset of relevant features based on their importance or relevance to the target variable.

Dimensionality reduction: It aims to reduce the number of features in a dataset while preserving important information. It helps in simplifying models, reducing computational complexity, and avoiding the curse of dimensionality.

Function selection: It refers to selecting a subset of functions or models that are most appropriate for a given task. It involves choosing the right algorithm or model to solve a specific problem based on its characteristics, performance, and suitability.

**9.**

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

**ANSWER:** IQR (Interquartile Range):

The IQR is a measure of statistical dispersion, representing the range between the first quartile (Q1) and the third quartile (Q3) in a dataset. It provides information about the spread and variability of the middle 50% of the data.

Criteria used to assess IQR:

Outliers: Values below Q1 - 1.5 \* IQR or above Q3 + 1.5 \* IQR are considered outliers.

Skewness: If the IQR is skewed, it indicates asymmetry in the distribution of data.

**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?**

Components of a box plot:

Median: The line inside the box represents the median value.

Box: The box represents the IQR, with the lower and upper quartiles

Whiskers: Lines (often represented as vertical lines or whiskers) extend from the box's edges to the data points within a certain range. They can represent minimum and maximum values or be determined by specific criteria.

Outliers: Data points outside the whiskers are considered outliers.

The lower whisker surpasses the upper whisker when the distribution is skewed to the left (negatively skewed).

Box plots can be used to identify outliers by plotting data points outside the whiskers. Outliers are typically defined as points that are located more than 1.5 times the IQR away from the upper or lower quartile.

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

**1. Data collected at regular intervals:** This refers to data collected at fixed time intervals, such as daily, weekly, or monthly measurements. It allows for studying trends, patterns, and seasonal variations in the data over time.

**2. The gap between the quartiles:** The gap between the quartiles, also known as the interquartile range (IQR), represents the spread or dispersion of the middle 50% of the data. It provides information about the variability within the data set.

**1. Make a comparison between:**

**1. Data with nominal and ordinal values:** Nominal data represents categories or labels without any inherent order, while ordinal data represents categories with a natural ordering or ranking. Nominal data can be represented by labels or names, while ordinal data can be ranked or ordered.

**2. Histogram and box plot:** Histograms display the distribution of continuous or discrete variables by dividing the data into bins and showing the frequency or count within each bin. Box plots summarize the distribution of a variable using quartiles, median, and outliers, providing information about the spread, central tendency, and skewness of the data.

**3. The average and median:** Both the average and median are measures of central tendency. The average (mean) is calculated by summing all values and dividing by the number of observations. The median is the middle value when the data is sorted. While the average considers all values, the median is less influenced by extreme values and is more robust to outliers.