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

Key tasks in machine learning:

- Data Collection: Gathering relevant data.

- Data Pre-processing: Cleaning, transforming, and organizing data.

- Feature Engineering: Selecting or creating relevant features for modeling.

- Model Selection: Choosing an appropriate algorithm or model.

- Model Training: Training the chosen model on data.

- Model Evaluation: Assessing model performance.

- Model Deployment: Putting the trained model to use.

Data Pre-processing: Data pre-processing involves cleaning and preparing raw data for analysis. It includes handling missing values, dealing with outliers, scaling or normalizing data, and encoding categorical variables.

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

Quantitative Data: Numerical data with measurable values. It can be discrete (countable) or continuous (infinitely divisible). Examples: age, height, temperature.

Qualitative Data: Categorical data that represents characteristics or qualities. It can be nominal (unordered categories) or ordinal (ordered categories). Examples: gender, color, education level.

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | **Age** | **Gender** | **Income** | **Education** |
| **1** | **20** | **M** | **10 LPA** | **BSc** |
| **2** | **25** | **F** | **15 LPA** | **BTech** |
| **3** | **30** | **M** | **25 LPA** | **BTech** |

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

Causes of data issues:

- Missing Data: Incomplete records.

- Outliers: Extreme values that don't fit the pattern.

- Noisy Data: Data with errors or inconsistencies.

- Imbalanced Data: Unequal distribution of classes.

Ramifications: These issues can lead to biased or inaccurate models, reduced predictive power, and poor generalization.

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

- Frequency Table: Count the occurrences of each category.

- Bar Chart: Visualize category frequencies as bars.

- Pie Chart: Show the proportion of each category in a circle chart.

Example: Exploring the "Gender" variable in a dataset.

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

Learning can be affected as models may struggle to learn from incomplete data, leading to biased or less accurate results. Missing data can also introduce noise or distort patterns.

Handling Missing Data:

- Removing Rows: If missing data is small, remove affected rows.

- Imputation: Fill missing values with estimates (mean, median, mode).

- Advanced Methods: Use machine learning to predict missing values.

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

- Mean/Median Imputation: Replace missing values with the mean or median of the variable.

- Mode Imputation: Replace missing values with the most frequent value.

- Predictive Imputation: Use regression or other models to predict missing values.

- Removing Rows: If missing values are significant, remove corresponding rows.

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

Data pre-processing techniques:

- Normalization/Scaling: Rescale features to a common range.

- Encoding Categorical Data: Convert categorical variables to numerical values.

- Handling Missing Data: Impute or remove missing values.

Dimensionality Reduction: Reducing the number of features while retaining essential information to improve efficiency and prevent overfitting.

Feature Selection: Selecting a subset of relevant features to improve model performance and reduce complexity.

9.

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

- IQR (Interquartile Range) is the range between the 25th and 75th percentiles in a dataset.

- It helps identify outliers and assesses the spread of data. Outliers can be detected using the IQR rule: data points outside 1.5 times the IQR.

**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?**

- A box plot consists of a box (interquartile range), a horizontal line (median), and two whiskers (minimum and maximum values within a certain range).

- The lower whisker may surpass the upper whisker when the data distribution is skewed, causing a longer tail on the lower side.

- Outliers are identified as individual points beyond the whiskers.

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

1. Data collected at regular intervals:

- Data collected with consistent time or space intervals.

- Examples: Temperature recordings every hour, stock prices at the end of each trading day.

2. The gap between the quartiles:

- The gap between the first quartile (25th percentile) and the third quartile (75th percentile) in a dataset.

- It's a measure of the spread or variability of data within the middle 50% range.

3. Use a cross-tab:

- Cross-tabulation shows the relationship between two categorical variables.

- It presents data in a table format, displaying the frequency distribution of one variable against another.

**11. Make a comparison between:**

1. Data with nominal and ordinal values:

- Nominal values have no inherent order (e.g., colors).

- Ordinal values have a meaningful order (e.g., education levels: high school < bachelor's < master's).

2. Histogram and box plot:

- Histogram displays the frequency distribution of a continuous variable.

- Box plot summarizes the distribution's central tendency, spread, and outliers.

3. The average and median:

- Both are measures of central tendency.

- Average (mean) is the sum of values divided by the count.

- Median is the middle value when data is sorted; it's less affected by outliers.