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

**Key tasks involved in machine learning:**

1. **Problem Definition**: Identify the type of problem you are solving (classification, regression, clustering, etc.).
2. **Data Collection**: Gather relevant data from various sources such as databases, APIs, or surveys.
3. **Data Preprocessing**: Clean and prepare the data for modeling by handling missing values, removing outliers, encoding categorical variables, etc.
4. **Feature Engineering**: Create new features or modify existing ones to improve model performance.
5. **Model Selection**: Choose the appropriate algorithm (e.g., linear regression, decision trees, neural networks).
6. **Model Training**: Train the model on a training dataset.
7. **Model Evaluation**: Evaluate the model’s performance using metrics like accuracy, precision, recall, F1-score, etc.
8. **Model Tuning**: Fine-tune the hyperparameters to improve the model’s performance.
9. **Model Deployment**: Deploy the trained model for real-world use.

**Data Preprocessing**: Data preprocessing involves preparing the raw data for analysis or modeling. It includes tasks like:

* Handling missing values (imputation or deletion).
* Normalizing or scaling numerical features.
* Encoding categorical data (e.g., one-hot encoding).
* Removing or handling outliers.
* Splitting the dataset into training and testing sets.

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

* **Quantitative Data**:
  + This data is numerical and can be measured. It represents quantities or amounts and can be used for arithmetic operations (e.g., addition, subtraction).
  + **Types**:
    - **Discrete**: Takes distinct values (e.g., the number of children, count of products sold).
    - **Continuous**: Can take any value within a range (e.g., height, weight, temperature).
  + **Examples**: Age, salary, temperature, distance.
* **Qualitative Data**:
  + This data is non-numeric and represents categories or labels. It describes qualities or characteristics that cannot be measured numerically.
  + **Types**:
    - **Nominal**: Categories without a natural order (e.g., gender, color, nationality).
    - **Ordinal**: Categories with a defined order (e.g., education level, satisfaction rating).
  + **Examples**: Gender, eye color, brand names, marital status.

**Distinction**:

* Quantitative data is numeric and used for performing calculations, whereas qualitative data is descriptive and used to categorize or label items.

**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 (Quantitative)** | **Gender (Qualitative - Nominal)** | **Education Level (Qualitative - Ordinal)** | **Income (Quantitative)** | **Product Purchased (Qualitative - Nominal)** |
| --- | --- | --- | --- | --- | --- |
| 1 | 25 | Female | Bachelor's | 45000 | Laptop |
| 2 | 34 | Male | Master's | 55000 | Phone |
| 3 | 28 | Female | PhD | 70000 | Laptop |
| 4 | 45 | Male | Bachelor's | 60000 | TV |
| 5 | 39 | Female | Master's | 80000 | Tablet |

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

**Causes of Data Issues**:

* **Missing Data**: This can occur due to errors in data collection, loss of data during transfer, or unavailability of some information.
* **Outliers**: Extreme values that deviate significantly from the other data points.
* **Noise**: Random variations or errors in the data that are irrelevant to the target variable.
* **Imbalanced Data**: A dataset where one class is underrepresented, which can lead to biased models.
* **Data Duplication**: Repeated data points that can distort the analysis.
* **Inconsistent Data**: Inconsistencies in formatting or coding in the dataset.

**Ramifications**:

* Poor model performance due to inaccurate or unrepresentative data.
* Incorrect conclusions, as outliers or noise may skew results.
* Misleading predictions when dealing with imbalanced datasets.

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

* **Frequency Distribution**: Count the occurrences of each category to understand the distribution.
  + **Example**: Counting the number of male and female participants in a survey.
* **Bar Charts**: Visualize the frequency of categories.
  + **Example**: A bar chart showing the number of purchases by different product categories (e.g., electronics, clothing).
* **Chi-Square Test**: Test the association between two categorical variables.
  + **Example**: Testing whether gender is associated with product preference (e.g., laptops vs. phones).
* **Cross-tabulation**: Creating a table to display the relationship between two categorical variables.
  + **Example**: A table showing the relationship between customer satisfaction (satisfied, neutral, dissatisfied) and product category (electronics, home goods).

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

Missing values can reduce the amount of usable data, introduce bias, and affect model accuracy. If important variables are missing, the model may fail to capture the true relationship between the features and the target.

**What can be done**:

* **Imputation**: Replace missing values with the mean, median, mode, or a predicted value based on other data points.
* **Data Removal**: Remove rows or columns with too many missing values.
* **Modeling Methods for Missing Data**: Use algorithms that can handle missing data (e.g., decision trees).

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

1. **Imputation**:
   * Replace missing values with the mean, median, or mode for numerical data.
   * Use the most frequent value for categorical data.
   * Predict missing values using algorithms like k-NN, regression, or other machine learning models.
2. **Data Removal**:
   * Remove rows that have missing values (if the number of missing values is small).
   * Drop columns that have too many missing values (e.g., if a feature has more than 50% missing values, consider removing it).
3. **Forward or Backward Fill**:
   * Use the last known value (forward fill) or the next available value (backward fill) to fill missing data for time series.
4. **Multiple Imputation**:
   * Impute missing values multiple times, creating several plausible datasets, and combine the results to account for uncertainty.

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

* **Data Preprocessing Techniques**:
  + **Normalization/Standardization**: Scale numerical values to a standard range or distribution.
  + **Encoding Categorical Variables**: Convert categorical data into numeric format (e.g., one-hot encoding).
  + **Handling Missing Values**: Fill or remove missing values.
  + **Outlier Detection**: Identify and manage outliers.
  + **Feature Scaling**: Ensure that all features are on a comparable scale.
* **Dimensionality Reduction**:
  + Reducing the number of features in the dataset while retaining important information, e.g., through PCA (Principal Component Analysis).
* **Feature Selection**:
  + Identifying and selecting the most important features for training the model, reducing overfitting and improving model efficiency.

**9.**

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

* **IQR (Interquartile Range)** is a measure of statistical dispersion, representing the range between the 25th percentile (Q1) and the 75th percentile (Q3) of the data.
* **Assessment**: IQR helps in identifying the spread of the middle 50% of the data. It is used to detect outliers by applying the formula:
  + Outliers = Q1 - 1.5 \* IQR or Q3 + 1.5 \* 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?**

* **Components of a Box Plot**:
  + **Box**: Represents the interquartile range (IQR), with the lower and upper edges marking Q1 and Q3, respectively.
  + **Median**: The line inside the box representing the 50th percentile.
  + **Whiskers**: Lines extending from the box to the minimum and maximum values within 1.5 \* IQR from the quartiles.
  + **Outliers**: Points beyond the whiskers, often shown as dots.
* **When the lower whisker surpasses the upper whisker**: This typically occurs in **negatively skewed distributions**, where more data points are clustered at higher values, creating a longer lower whisker.
* **Identifying outliers**: Any data points beyond the whiskers are considered outliers.

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

1. **Data collected at regular intervals**:
   * Data collected at consistent intervals over time, like hourly or daily measurements. It's useful for time series analysis.
2. **The gap between the quartiles**:
   * The difference between the 1st quartile (Q1) and the 3rd quartile (Q3) represents the interquartile range (IQR), showing the spread of the middle 50% of the data.
3. **Use a cross-tab**:

* A **cross-tabulation** is a tool used to examine the relationship between two categorical variables by displaying the frequency distribution of variables in a matrix form.

**11. Make a comparison between:**

1. **Data with nominal and ordinal values**:
   * **Nominal**: Categories without a meaningful order (e.g., color, gender).
   * **Ordinal**: Categories with a meaningful order (e.g., education level, satisfaction).
2. **Histogram and box plot**:
   * **Histogram**: A graphical representation of the distribution of numerical data, showing frequency by bins.
   * **Box Plot**: A graphical representation of the data distribution using quartiles, highlighting outliers and spread.
3. **The average and median**:
   * **Average (Mean)**: The sum of all values divided by the number of values; sensitive to outliers.
   * **Median**: The middle value when data is sorted; less affected by outliers.