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

Answer 1: Key tasks in machine learning include data collection, data pre-processing, model training, evaluation, and deployment. Data pre-processing involves cleaning and transforming raw data into a suitable format for analysis. This includes handling missing values, outliers, and scaling features.

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

Answer 2:

* Quantitative Data: Numerical data with measurable quantities. Example: Height, weight.
* Qualitative Data: Categorical data with non-numeric values. Example: Colors, gender.

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

Answer 3:

| **ID** | **Age** | **Gender** | **Height** | **Score** |
| --- | --- | --- | --- | --- |
| 1 | 25 | Male | 175.5 | 85 |
| 2 | 30 | Female | 160.0 | 92 |
| 3 | 22 | Male | 180.0 | 78 |

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

Answer 4: Causes include missing values, outliers, noise, and imbalanced data. Ramifications can include biased models, decreased accuracy, and compromised generalization.

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

Answer 5:

* Frequency Distribution: Count of occurrences of each category.
* Bar Chart: Visualization of category frequencies.
* Pie Chart: Proportional representation of category distribution.

Example:

import matplotlib.pyplot as plt

import pandas as pd

data = {'Category': ['A', 'B', 'A', 'C', 'B', 'A']}

df = pd.DataFrame(data)

freq\_dist = df['Category'].value\_counts()

plt.bar(freq\_dist.index, freq\_dist.values)

plt.show()

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

Answer 6: Learning can be impacted as missing values may introduce bias. Imputation methods, such as mean or median replacement, or advanced techniques like predictive modeling, can be employed to handle missing values.

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

Answer 7:

* Deletion: Removing rows or columns with missing values.
* Imputation: Replacing missing values with a statistically derived estimate.
* Interpolation: Estimating missing values based on existing data points.

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

Answer 8: Data pre-processing techniques include normalization, standardization, and encoding.

* Dimensionality Reduction: Reducing the number of features to enhance model efficiency.
* Feature Selection: Choosing the most relevant features for model training.

**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?**

Answer 9:

* Interquartile Range (IQR): Measure of statistical dispersion, calculated as the difference between the third and first quartiles. Criteria for assessment include identifying outliers and assessing data variability.
* Components include the median, quartiles, whiskers, and outliers. The lower whisker surpasses the upper whisker when the data distribution is negatively skewed. Outliers can be identified as points beyond the whiskers.

**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**

Answer 10:

Data collected at regular intervals:

* Time Series Data
* Examples: Stock prices, temperature readings

The gap between the quartiles:

* Interquartile Range (IQR)
* Represents the middle 50% of the data.

Use a cross-tab:

* Cross-tabulation displays the relationships between two categorical variables, 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**

Answer 11:

Data with nominal and ordinal values:

* Nominal: Categories without inherent order (e.g., colors).
* Ordinal: Categories with a meaningful order (e.g., education levels).

Histogram and box plot:

* Histogram: Represents data distribution using bars.
* Box Plot: Displays the distribution's central tendency and spread, emphasizing outliers.

The average and median:

* Average (Mean): Sensitive to outliers, influenced by extreme values.
* Median: Less affected by outliers, represents the middle value.