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

Ans: A machine learning task is the type of prediction or inference being made, based on the problem or question that is being asked, and the available data. For example, the classification task assigns data to categories, and the clustering task groups data according to similarity.

Data preprocessing is essential before its actual use. Data preprocessing is the concept of changing the raw data into a clean data set. The dataset is preprocessed in order to check missing values, noisy data, and other inconsistencies before executing it to the algorithm.

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

Ans: **Quantitative data** refers to any information that can be quantified. If it can be counted or measured, and given a numerical value, it’s quantitative data.

### Types of quantitative data (with examples)

Quantitative data is either **discrete** or **continuous**:

* **Discrete quantitative data** takes on fixed numerical values and cannot be broken down further. An example of discrete data is when you count something, such as the number of people in a room. If you count 32 people, this is fixed and finite.
* **Continuous quantitative data** can be placed on a continuum and infinitely broken down into smaller units. It can take any value; for example, a piece of string can be 20.4cm in length, or the room temperature can be 30.8 degrees.

Unlike quantitative data, **Qualitative data** cannot be measured or counted. It’s descriptive, expressed in terms of language rather than numerical values.

### Types of qualitative data (with examples)

Qualitative data may be classified as **nominal** or **ordinal**:

* **Nominal data** is used to label or categorize certain variables without giving them any type of quantitative value. For example, if you were collecting data about your target audience, you might want to know where they live. Are they based in the UK, the USA, Asia, or Australia? Each of these geographical classifications count as nominal data. Another simple example could be the use of labels like “blue,” “brown,” and “green” to describe eye color.
* **Ordinal data** is when the categories used to classify your qualitative data fall into a natural order or hierarchy. For example, if you wanted to explore customer satisfaction, you might ask each customer to select whether their experience with your product was “poor,” “satisfactory,” “good,” or “outstanding.” It’s clear that “outstanding” is better than “poor,” but there’s no way of measuring or quantifying the “distance” between the two categories.

**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:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name | Age (Years) | Weight (kg) | Gender | Graade | Nationality |
| abc | 18 | 56.2 | M | A | India |
| xyz | 22 | 59.8 | M | A+ | Australia |
| pqr | 29 | 63.4 | F | AA | USA |
| jkl | 19 | 78.5 | M | B+ | Spain |

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

Ans: Various ML data issues are -

1. Lack Of Quality Data

2. Understanding Which Processes Need Automation

3. Inadequate Infrastructure

4. Lack of Skilled Resources

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

Ans: **Categorical Data** can be of two types – (i) **Nominal Data** and (ii) **Ordinal Data**

In case of Nominal Data we can use **Label Encoder** and for Ordinal Data we can use **Mapping** to convert them into Numeric equivallent.

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

Ans: **It is important to handle the missing values appropriately.**

* Most of the machine learning algorithms fail if the dataset contains missing values. However, algorithms like K-nearest and Naive Bayes support data with missing values.
* You may end up building a biased machine learning model which will lead to incorrect results if the missing values are not handled properly.
* Missing data can lead to a lack of precision in the statistical analysis.

**There are 2 primary ways of handling missing values:**

1. Deleting the Missing values  
    a. Deleting the entire row  
    b. Deleting the entire column
2. Imputing the Missing Values  
   **Imputing Missing Values For Numerical Features**  
    a. Replacing With Arbitrary Value  
    b. Replacing With Mean  
    c. Replacing With Mode  
    d. Replacing With Median  
    e. Replacing with previous value – Forward fill  
    f. Replacing with next value – Backward fill  
    g. Interpolation  
    **Imputing Missing Values For Categorical Features** a. Impute the Most Frequent Value  
    b. Impute the Value “missing”, which treats it as a Separate Category

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

Ans:

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

Ans:

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

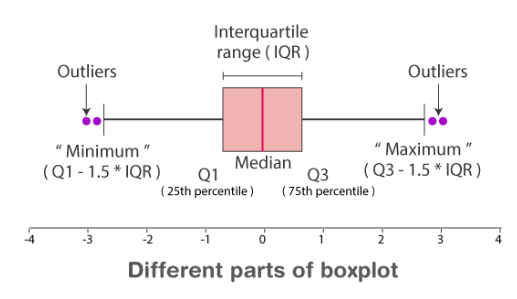
Ans: **The difference between the upper and lower quartile is known as the interquartile range. The formula for the interquartile range is given below**

**Interquartile range = Upper Quartile – Lower Quartile = Q3 – Q1**

where Q1 is the first quartile and Q3is the third quartile of the series.

**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:A box and whisker plot—also called a box plot—displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum. In a box plot, we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median.



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

**1. Data collected at regular intervals**

Ans:

**2. The gap between the quartiles**

Ans:

**3. Use a cross-tab**

Ans:

**11. Make a comparison between:**

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

Ans:

**2. Histogram and box plot**

Ans:

**3. The average and median**

Ans: