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

**Answer:**

**Key Tasks:**

* Data collection
* Data cleaning and preprocessing
* Feature engineering and selection
* Model selection
* Model training and validation
* Model evaluation and deployment

**Data Preprocessing:**

* Handling missing values
* Removing duplicates
* Scaling/normalization
* Encoding categorical variables
* Handling outliers

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

**Answer:**

**Quantitative Data:**

* Measurable and numerical (e.g., age, height, weight)
* Can be continuous (e.g., temperature) or discrete (e.g., number of students)

**Qualitative Data:**

* Descriptive and categorical (e.g., color, gender)
* Can be nominal (e.g., nationality) or ordinal (e.g., satisfaction level)

**Difference:** Quantitative data deals with numbers, while qualitative data deals with categories.

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

**Answer:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Id** | **Name** | **Age** | **Income Lavel** | **Join Date** |
| **1** | **Amin** | **24** | **Low** | **2023-05-10** |
| **2** | **Lisa** | **30** | **high** | **2022-08-15** |
| **3** | **Sunny** | **28** | **Medium** | **2021-12-20** |

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

**Answer:**

**Causes:**

* Missing data
* Incorrect data
* Duplicates
* Outliers
* Data inconsistency

**Ramifications:**

* Poor model performance
* Biased predictions
* Misinterpretation of results

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

**Answer:**

**Frequency tables:** Count occurrences (e.g., Male: 40, Female: 60)

**Bar charts**: Visualize category distribution

**Cross-tabulation:** Compare relationships (e.g., gender vs. purchase behavior)

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

**Answer:**

**Effect:**

* Incomplete learning
* Biased outcomes
* Reduced accuracy

**Solutions:**

* Remove missing data (if minimal)
* Impute values (mean, median, mode)
* Use machine learning algorithms to predict missing values

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

**Answer:**

* Deletion (listwise, pairwise)
* Mean/Median/Mode imputation
* Forward/Backward fill (for time-series)
* Predictive modeling (KNN, regression)

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

**Answer:**

* Handling missing data
* Encoding categorical variables
* Scaling/normalization
* Feature selection
* Dimensionality reduction

**Feature Selection vs. Dimensionality Reduction:**

**Feature selection:** Retains important features

**Dimensionality reduction:** Transforms features into lower dimensions (e.g., PCA)

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

**Answer: IQR (Interquartile Range)**

**Formula:** IQR = Q3 - Q1

Measures statistical dispersion

**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: Box Plot Components:**

* Median (Q2)
* Q1 (25th percentile)
* Q3 (75th percentile)
* Whiskers (min & max values)
* Outliers (beyond 1.5 \* IQR)

**When Lower Whisker > Upper Whisker?**

* When the data is skewed to the right.

**Box Plot for Outliers:**

* Points beyond whiskers indicate outliers.

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

**1. Data collected at regular intervals**

* Time-series data (e.g., hourly temperature)

**2. The gap between the quartiles**

* Also called IQR, represents data spread

**3. Use a cross-tab**

* Helps in analysing categorical relationships

**1. Make a comparison between:**

1. Data with nominal and ordinal values

2. Histogram and box plot

3. The average and median

**Answer:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Comparison** | **Data Type 1** | **Data Type 2** | **Difference** |
| **Nominal vs. Ordinal** | Names, categories (e.g., colors) | Ranked categories (e.g., satisfaction) | Ordinal data has an order, nominal doesn’t. |
| **Histogram vs. Box Plot** | Shows frequency distribution | Shows data spread & outliers | Histogram is for distribution; box plot is for statistical summary. |
| **Mean vs. Median** | Average of values | Middle value in sorted data | Mean is affected by outliers, median is not. |