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

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

3. Create a basic data collection that includes some sample records. Have at least one attribute from

each of the machine learning data types.

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

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

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

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

8. What are the various data pre-processing techniques? Explain dimensionality reduction and

function selection in a few words.

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?

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

11. Make a comparison between:

1. Data with nominal and ordinal values

2. Histogram and box plot

3. The average and median

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

**Key Tasks in Machine Learning:**

1. **Problem Definition:** Clearly define the problem you want to solve and the objectives of your model.
2. **Data Collection:** Gather relevant and sufficient data from various sources.
3. **Data Pre-processing:** Clean and prepare the data by handling missing values, removing duplicates, and transforming features.
4. **Feature Engineering:** Create and select relevant features that improve model performance.
5. **Model Selection:** Choose the appropriate machine learning algorithm based on the problem type (e.g., classification, regression).
6. **Model Training:** Fit the model to the training data by adjusting its parameters.
7. **Model Evaluation:** Assess the model’s performance using validation metrics and adjust as needed.
8. **Hyperparameter Tuning:** Optimize the model by tuning hyperparameters to enhance performance.
9. **Deployment:** Implement the model in a real-world application.
10. **Monitoring and Maintenance:** Continuously monitor the model’s performance and update it as necessary.

**Data Pre-processing:**

* **Definition:** The process of preparing raw data for analysis by cleaning, transforming, and organizing it to improve the quality and usability of the data.
* **Includes:** Handling missing values, removing outliers, normalizing data, encoding categorical variables, and scaling features.

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

**Quantitative Data:**

* **Definition:** Data that is numerical and can be measured or counted. It represents quantities and allows for mathematical operations.
* **Types:**
  + **Discrete Data:** Countable data (e.g., number of students, number of cars).
  + **Continuous Data:** Measurable data with infinite possible values within a range (e.g., height, weight).
* **Example:** The annual sales revenue of a company ($5,000,000) or the height of a person (175 cm).

**Qualitative Data:**

* **Definition:** Data that is descriptive and categorical, representing qualities or characteristics. It cannot be measured numerically.
* **Types:**
  + **Nominal Data:** Categories without a natural order (e.g., color, gender).
  + **Ordinal Data:** Categories with a meaningful order (e.g., education level, satisfaction ratings).
* **Example:** A customer's feedback on service quality ("Excellent", "Good", "Fair") or the type of product purchased ("Electronics", "Clothing").

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

**Sample Data Collection:**

| **Customer ID** | **Age** | **Gender** | **Purchase Amount** | **Purchase Date** |
| --- | --- | --- | --- | --- |
| 001 | 25 | Female | 120.50 | 2024-08-15 |
| 002 | 34 | Male | 450.75 | 2024-08-16 |
| 003 | 29 | Female | 320.00 | 2024-08-17 |
| 004 | 40 | Male | 220.30 | 2024-08-18 |
| 005 | 22 | Female | 150.00 | 2024-08-19 |

* **Customer ID:** Categorical (Nominal)
* **Age:** Quantitative (Continuous)
* **Gender:** Categorical (Nominal)
* **Purchase Amount:** Quantitative (Continuous)
* **Purchase Date:** Temporal Data (Date/Time)

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

**Causes of Machine Learning Data Issues:**

1. **Missing Data:** Incomplete records due to data not being collected or recorded.
   * **Ramifications:** Can lead to biased or inaccurate model predictions. May require imputation or removal of records.
2. **Outliers:** Extreme values that deviate significantly from the majority of the data.
   * **Ramifications:** Can skew the results and affect model performance. May need to be identified and handled appropriately.
3. **Inconsistent Data:** Data that does not follow a consistent format or standard.
   * **Ramifications:** Can result in errors during processing or analysis. Requires standardization and cleaning.
4. **Noisy Data:** Data with errors or random variations.
   * **Ramifications:** Can obscure patterns and reduce model accuracy. Requires smoothing or filtering.
5. **Unbalanced Data:** Unequal distribution of classes or categories in classification tasks.
   * **Ramifications:** Can lead to biased models that favor the majority class. Requires techniques like resampling or reweighting.

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

**Approaches to Categorical Data Exploration:**

1. **Frequency Distribution:**
   * **Example:** Counting the number of occurrences of each category in a dataset (e.g., the number of customers in each age group).
2. **Bar Charts:**
   * **Example:** Visualizing the frequency of different categories (e.g., a bar chart showing the number of products sold by type).
3. **Cross-Tabs (Contingency Tables):**
   * **Example:** Analyzing the relationship between two categorical variables (e.g., the relationship between gender and preferred product category).
4. **Pie Charts:**
   * **Example:** Showing the proportion of each category in relation to the whole (e.g., a pie chart depicting market share of different brands).
5. **Chi-Square Test:**
   * **Example:** Testing the independence of two categorical variables (e.g., testing if there's an association between education level and employment status).

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

**Impact of Missing Values:**

* **Learning Activity Impact:**
  + **Model Accuracy:** Missing values can lead to incomplete training data, affecting model accuracy and performance.
  + **Bias:** May introduce bias if the missing values are not random and are related to the outcome variable.

**Handling Missing Values:**

1. **Removal:** Exclude records or features with missing values if the missing data is minimal.
2. **Imputation:** Replace missing values with statistical measures like mean, median, or mode.
3. **Prediction:** Use machine learning algorithms to predict and fill in missing values based on other data.
4. **Indicator Variable:** Create a binary indicator variable to flag missing values.

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

**Methods for Dealing with Missing Data:**

1. **Deletion:**
   * **Listwise Deletion:** Remove entire records with missing values. Suitable if missing data is minimal and random.
   * **Pairwise Deletion:** Use available data without removing entire records, typically used in correlation calculations.
2. **Imputation:**
   * **Mean/Median/Mode Imputation:** Replace missing values with the mean, median, or mode of the observed data.
   * **K-Nearest Neighbors (KNN) Imputation:** Use the average values from the nearest neighbors in feature space to impute missing data.
   * **Multiple Imputation:** Generate several imputed datasets and combine results to account for uncertainty.
3. **Prediction Models:**
   * **Regression Imputation:** Predict missing values using regression models based on other features.
   * **Machine Learning Algorithms:** Use algorithms like Random Forest or XGBoost to predict missing values.
4. **Data Augmentation:**
   * **Synthetic Data Generation:** Create synthetic data points to fill gaps based on observed patterns.
5. **Indicator Variables:**
   * **Missingness Indicator:** Create a new binary feature indicating whether data is missing, allowing the model to learn patterns related to missing data.

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

**Data Pre-processing Techniques:**

1. **Data Cleaning:** Handling missing values, removing duplicates, and correcting errors.
2. **Data Transformation:** Normalizing, scaling, and encoding categorical variables.
3. **Feature Engineering:** Creating new features or modifying existing ones to improve model performance.
4. **Data Integration:** Combining data from multiple sources into a unified format.
5. **Data Reduction:** Reducing the size of data through techniques like dimensionality reduction and feature selection.

**Dimensionality Reduction:**

* **Definition:** The process of reducing the number of features in a dataset while preserving as much information as possible. Techniques include Principal Component Analysis (PCA) and t-SNE.

**Feature Selection:**

* **Definition:** The process of selecting a subset of relevant features from the original set, aiming to improve model performance and reduce overfitting. Techniques include recursive feature elimination and feature importance scoring.

### **9.**

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

* **IQR (Interquartile Range):** A measure of statistical dispersion, calculated as the difference between the first quartile (Q1) and the third quartile (Q3) in a dataset.
* **Criteria for Assessment:** IQR is used to identify outliers and understand the spread of the middle 50% of the data. Outliers are typically defined as values below Q1−1.5×IQRQ1 - 1.5 \times \text{IQR}Q1−1.5×IQR or above Q3+1.5×IQRQ3 + 1.5 \times \text{IQR}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:**
  + **Box:** Represents the interquartile range (IQR) between the first quartile (Q1) and third quartile (Q3).
  + **Median Line:** A line inside the box representing the median value.
  + **Whiskers:** Lines extending from the box to the minimum and maximum values within 1.5 times the IQR from Q1 and Q3.
  + **Outliers:** Points outside the whiskers, often marked with individual dots or symbols.
* **Whisker Length Discrepancy:** The lower whisker can surpass the upper whisker in length when the data is negatively skewed, meaning there are more extreme values on the lower side of the distribution.
* **Identifying Outliers:** Outliers are values that fall outside the whiskers of the box plot, indicating they are significantly different from the majority of the data.

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

**1. Data Collected at Regular Intervals:**

* **Definition:** Data collected at consistent, evenly spaced time intervals.
* **Example:** Daily temperature readings, monthly sales figures.

**2. The Gap Between the Quartiles:**

* **Definition:** The distance between the first quartile (Q1) and the third quartile (Q3), also known as the interquartile range (IQR).
* **Use:** Measures the spread of the middle 50% of the data and helps identify data dispersion and variability.

**3. Use a Cross-Tab:**

* **Definition:** A table that displays the frequency distribution of two or more categorical variables.
* **Example:** A cross-tab showing the number of customers who purchased different product categories across different age groups.

### **11. Make a comparison between:**

**1. Data with Nominal and Ordinal Values**

* **Nominal Data:** Categories without a meaningful order (e.g., colors, gender).
* **Ordinal Data:** Categories with a meaningful order or ranking (e.g., customer satisfaction levels, education degrees).

**2. Histogram and Box Plot**

* **Histogram:** Displays the frequency distribution of numerical data with bars representing the count of data points within intervals (bins). Used to visualize data distribution and density.
* **Box Plot:** Shows data distribution through quartiles and highlights outliers. Provides a summary of the central tendency, dispersion, and skewness of the data.

**3. The Average and Median**

* **Average (Mean):** The sum of all values divided by the number of values. Sensitive to extreme values (outliers).
* **Median:** The middle value in a sorted dataset. Less affected by outliers and provides a better measure of central tendency for skewed distributions.