Q1.What are the key tasks that machine learning entails? What does data pre-processing imply?

Key Tasks in Machine Learning and Data Pre-processing: Key Tasks in Machine Learning:

* Data Collection
* Data Pre-processing
* Feature Engineering
* Model Selection
* Model Training
* Model Evaluation
* Hyperparameter Tuning
* Model Deployment
* Monitoring and Maintenance

Data Pre-processing: Data pre-processing involves cleaning, transforming, and organizing raw data into a usable format for analysis. It includes tasks like handling missing values, dealing with outliers, scaling, normalization, encoding categorical variables, and reducing dimensionality.

Q2. Describe quantitative and qualitative data in depth. Make a distinction between the two.

**Quantitative vs. Qualitative Data:** Quantitative Data: Numerical measurements that represent quantities. Examples: Age, Height, Income. Qualitative Data: Descriptive, categorical, or label-based information. Examples: Gender, Color, Region. Distinction: Quantitative data can be measured and subjected to mathematical operations, while qualitative data represents categories or labels.

Q3. Create a basic data collection that includes some sample records. Have at least one attribute from each of the machine learning data types.

Basic Data Collection with Sample Records:

| ID | Age | Gender | Income | Color |
| --- | --- | --- | --- | --- |
| 1 | 25 | Male | $50000 | Blue |
| 2 | 30 | Female | $60000 | Red |
| 3 | 22 | Male | $45000 | Green |
| 4 | 28 | Female | $55000 | Blue |

Q4. What are the various causes of machine learning data issues? What are the ramifications?

**Causes of Data Issues and Ramifications:** Causes: Missing values, outliers, noisy data, inconsistent formats, duplicate records. Ramifications: Reduced accuracy, biased results, poor model performance, incorrect insights.

Q5. Demonstrate various approaches to categorical data exploration with appropriate examples.

Approaches to Categorical Data Exploration:

* Frequency Distribution: Count of occurrences of each category.
* Bar Chart: Visual representation of category frequencies.
* Pie Chart: Showing proportion of each category in a circle chart.

Q6. How would the learning activity be affected if certain variables have missing values? Having said that, what can be done about it?

Learning Activity and Missing Values:

* Missing values can lead to biased analysis and inaccurate predictions.
* Handling: Impute with mean/median, drop rows/columns, use advanced techniques like regression, or treat as a separate category.

Q7. Describe the various methods for dealing with missing data values in depth.

Methods for Dealing with Missing Data:

* Mean/Median Imputation
* Forward/Backward Fill
* Interpolation
* Model-based Imputation

Q8. What are the various data pre-processing techniques? Explain dimensionality reduction and function selection in a few words.

Data Pre-processing Techniques:

* Dimensionality Reduction: Reducing the number of features while preserving important information.
* Feature Selection: Selecting relevant features based on their importance.

Q9. 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?

Box Plot and IQR:

* IQR (Interquartile Range): Range between the 75th (Q3) and 25th (Q1) percentiles.
* Box Plot Components: Median, quartiles, whiskers, and outliers.
* Lower Whisker surpasses upper whisker when data is negatively skewed.
* Box plots identify outliers beyond the whiskers.

Q10. 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

Brief Notes:

* Data Collected at Regular Intervals: Time series data collected at fixed time intervals.
* Gap Between Quartiles: IQR represents the range containing the middle 50% of data.
* Use a Cross-Tab: Cross-tabulation displays relationships between categorical variables.

Q11. Make a comparison between:

1. Data with nominal and ordinal values

2. Histogram and box plot

3. The average and median

* Data with Nominal and Ordinal Values: Nominal has no order, ordinal has a defined order.
* Histogram and Box Plot: Histogram shows frequency distribution, box plot displays quartiles and outliers.
* Average and Median: Average is affected by outliers, median is not.