1. What are the key tasks involved in getting ready to work with machine learning modeling?

>>>Data Collection

Data Preprocessing.

Feature Engineering

Data Splitting

Model Selection

Model Training

Hyperparameter Tuning

Model Evaluation

Model Deployment

2. What are the different forms of data used in machine learning? Give a specific example for each of them.

>>>Numeric Data: Data represented by numerical values. Example: Temperature, age, salary.

Categorical Data: Data represented by categories or labels. Example: Gender (Male/Female), color (Red/Blue/Green).

Text Data: Unstructured textual information. Example: Document text, tweets, customer reviews.

3. Distinguish:

1. Numeric vs. categorical attributes

>>>Numeric Attributes: These are attributes with numerical values that can be used for arithmetic operations. Example: Age, salary.

Categorical Attributes: These are attributes with non-numerical values that represent categories or labels. Example: Gender, country.

2. Feature selection vs. dimensionality reduction

>>>Feature Selection: Involves selecting a subset of relevant features from the original feature set to build the model. It aims to improve model performance and reduce complexity. Example: Selecting the most important features based on their correlation with the target variable.

Dimensionality Reduction: Involves reducing the number of features in the dataset while preserving as much information as possible. It helps to mitigate the curse of dimensionality and improve computational efficiency. Example: Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE)

4. Make quick notes on any two of the following:

1. The histogram

>>>Histogram is a graphical representation of the distribution of numeric data.

It consists of bins (intervals) on the x-axis and the frequency or count of data points within each bin on the y-axis.

2. Use a scatter plot

>>>Scatter plot displays individual data points as dots on a two-dimensional plane.

Each dot represents a data point with its values on two different variables.

Useful for visualizing the relationship or correlation between two variables.

Positive correlation: As one variable increases, the other tends to increase.

3.PCA (Personal Computer Aid)

>>>PCA is a dimensionality reduction technique used in data analysis and machine learning.

It transforms data into a new coordinate system while preserving as much variance as possible.

Reduces the dimensionality of data by creating principal components, which are linear combinations of the original features.

5. Why is it necessary to investigate data? Is there a discrepancy in how qualitative and quantitative data are explored?

>>>Investigating Data: Data investigation involves exploring and analyzing datasets to understand their characteristics, patterns, relationships, and potential issues. It helps in making informed decisions during data preprocessing and model building.

Discrepancy between Qualitative and Quantitative Data Exploration: Exploring qualitative (categorical) data involves understanding the distribution of categories, calculating frequencies, and creating bar plots. For quantitative (numeric) data, exploration includes examining summary statistics, creating histograms, scatter plots, and detecting outliers. The methods differ due to the nature of the data.

6. What are the various histogram shapes? What exactly are ‘bins'?

>>>Histogram Shapes: Histograms can take various shapes, including:

Normal (Gaussian) Distribution: Bell-shaped, symmetric.

Skewed Distribution: Either positively (right-skewed) or negatively (left-skewed) skewed.

Bimodal Distribution: Two distinct peaks.

Bins: Bins are intervals into which the range of data is divided in a histogram. They determine the width of each bar and affect the visual representation of the data distribution.

7. How do we deal with data outliers?

>>>Outliers are data points that significantly differ from the rest of the data. To deal with outliers:

Detect Outliers: Use statistical methods (e.g., Z-score, IQR) and visualization techniques (box plots, scatter plots).

Decide Treatment: Depending on the context, outliers can be removed, transformed, or kept after careful consideration of their impact on analysis and modeling.

8. What are the various central inclination measures? Why does mean vary too much from median in certain data sets?

>>>entral Inclination Measures: These include Mean, Median, and Mode. They represent the center or average value of a dataset.

Mean vs. Median: Mean is sensitive to outliers, as it considers all data points. Median is less affected by outliers, making it a better measure of central tendency in skewed distributions.

9. Describe how a scatter plot can be used to investigate bivariate relationships. Is it possible to find outliers using a scatter plot?

>>>Scatter Plot: A scatter plot displays points representing the relationship between two variables. It helps visualize patterns, trends, and correlations between variables.

Outliers: Scatter plots can identify outliers as data points that deviate significantly from the overall pattern.

10. Describe how cross-tabs can be used to figure out how two variables are related.

>>>Cross-Tabs (Cross-Tabulation): A cross-tab is a table that displays the distribution of two categorical variables. It helps identify patterns and relationships between the variables.

Relationships: Cross-tabs show how the distribution of one variable varies with the categories of another variable.