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

**ANSWER:** . Key tasks involved in getting ready to work with machine learning modeling:

a) Data Collection: Gather the relevant data needed for the machine learning task. This may involve collecting data from various sources, such as databases, APIs, or online repositories.

b) Data Preprocessing: Clean the data by handling missing values, dealing with outliers, and addressing inconsistencies or errors. This step may also involve data normalization, feature scaling, and handling categorical variables.

c) Feature Engineering: Select or create the most informative features for the machine learning model. This may include transforming variables, creating new features, or reducing dimensionality.

d) Data Splitting: Divide the dataset into training, validation, and testing sets. The training set is used to train the model, the validation set is used for hyperparameter tuning and model selection, and the testing set is used to evaluate the final model's performance.

e) Model Training: Select an appropriate machine learning algorithm or model, and train it using the training dataset. This involves feeding the input features and corresponding output labels to the model and optimizing its parameters.

f) Model Evaluation: Assess the performance of the trained model using evaluation metrics and techniques such as accuracy, precision, recall, F1 score, or mean squared error, depending on the specific problem.

g) Model Deployment: Once the model is trained and evaluated, it can be deployed for prediction or inference on new, unseen data.

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

**ANSWER:** . Different forms of data used in machine learning:

a) Numerical Data: This includes continuous or discrete numerical values. An example would be a dataset of house prices, where each house is represented by its price, which is a numeric value.

b) Categorical Data: This includes discrete values that represent categories or labels. For example, a dataset of animals could have a categorical variable indicating the species of each animal, such as "dog," "cat," or "bird."

c) Textual Data: This includes unstructured text data, such as articles, reviews, or tweets. Natural Language Processing techniques are used to process and analyze this type of data.

d) Image Data: This includes pixel values representing images. Each pixel's intensity or color can be used as input features. For instance, a dataset of handwritten digits, where each digit image is represented by its pixel values.

e) Time-Series Data: This represents data collected over time, such as stock prices or temperature readings. It has a temporal component and can be used to predict future values or patterns.

**3. Distinguish:**

**1. Numeric vs. categorical attributes**

**ANSWER:** Numeric attributes represent continuous or discrete numerical values and can be subjected to mathematical operations like addition, subtraction, or averaging. Categorical attributes represent discrete categories or labels and don't have inherent mathematical meaning. For example, age is a numeric attribute (e.g., 25, 30), while gender is a categorical attribute (e.g., male, female).

**2. Feature selection vs. dimensionality reduction**

**ANSWER:** Feature selection involves selecting a subset of the original features to use in the machine learning model. The goal is to choose the most informative features that contribute the most to the model's performance. Dimensionality reduction, on the other hand, aims to transform the original high-dimensional feature space into a lower-dimensional representation. This can be done through techniques like Principal Component Analysis (PCA) or t-SNE (t-Distributed Stochastic Neighbor Embedding).

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

**1. The histogram**

**ANSWER:** A histogram is a graphical representation of the distribution of a dataset. It consists of a set of bins along the x-axis, which represent intervals of the data's range, and the y-axis represents the frequency or count of observations falling within each bin. It helps visualize the shape, central tendency, and spread of the data distribution.

**2. Use a scatter plot**

**ANSWER:** A scatter plot is a graph that displays the relationship between two continuous variables. Each data point is represented by a dot on the plot, where the x-axis corresponds to one variable and the y-axis corresponds to the other. Scatter plots are useful for identifying patterns, trends, or relationships between variables, such as correlation or clustering.

**3.PCA (Personal Computer Aid)**

**ANSWER:** PCA is a dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional representation. It aims to find the principal components, which are new orthogonal axes that capture the maximum variance in the data. PCA can be used for data visualization, noise reduction, and feature extraction.

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

**ANSWER:** It is necessary to investigate data because it helps in understanding its characteristics, identifying patterns, and making informed decisions about feature selection, preprocessing steps, and model selection. Both qualitative and quantitative data can be explored, but the approaches may differ. Qualitative data exploration may involve techniques like content analysis, thematic coding, or sentiment analysis. Quantitative data exploration often includes summary statistics, visualization techniques (e.g., histograms, scatter plots), and correlation analysis.

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

**ANSWER:** Various histogram shapes are:

a) Normal distribution (bell-shaped): The data is symmetrically distributed around the mean, forming a bell-shaped curve.

b) Skewed to the left (negatively skewed): The majority of the data is concentrated on the right side, with a long tail stretching to the left.

c) Skewed to the right (positively skewed): The majority of the data is concentrated on the left side, with a long tail stretching to the right.

d) Bimodal distribution: The data has two distinct peaks, indicating the presence of two separate modes or subpopulations.

Bins in a histogram represent intervals or ranges along the x-axis, dividing the data's range into smaller segments. The frequency or count of observations falling within each bin is represented by the height of the corresponding bar on the y-axis.

**7. How do we deal with data outliers?**

**ANSWER:** Dealing with data outliers can be done through several approaches:

a) Removing outliers: Outliers can be removed from the dataset if they are determined to be erroneous or extreme values that don't represent the underlying distribution accurately. However, caution must be exercised to ensure the removal doesn't introduce bias or affect the overall data representation.

b) Transforming outliers: Outliers can be transformed using statistical techniques like winsorization or log transformation. These methods bring extreme values closer to the mean or other central tendency measures, reducing their impact.

c) Treating outliers as a separate category: In some cases, outliers may contain valuable information or represent distinct phenomena. They can be kept as a separate category or considered separately during analysis.

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

**ANSWER:** Various measures of central inclination:

a) Mean: The mean is the average of a set of values and is calculated by summing all the values and dividing by the number of observations. However, the mean can be sensitive to extreme values or outliers, leading to its deviation from the median in certain datasets.

b) Median: The median is the middle value in a sorted dataset. It is less influenced by extreme values or outliers compared to the mean and provides a measure of central tendency that is robust to such values.

c) Mode: The mode represents the most frequently occurring value or values in a dataset. It is particularly useful for categorical data or discrete variables.

The mean can vary significantly from the median in certain datasets when extreme values or outliers are present, as the mean takes into account the values' magnitudes and can be pulled towards those extreme values. The median, on the other hand, is less affected by extreme values and provides a better representation of the data's central tendency.

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

**ANSWER:** A scatter plot can be used to investigate bivariate relationships by plotting two variables against each other. It helps visualize the correlation or relationship between the variables

If the points in the scatter plot show a clear pattern or trend, it indicates a relationship between the variables. Outliers can also be identified as points that deviate significantly from the overall pattern of the scatter plot.

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

**ANSWER:** Cross-tabs, or cross-tabulation tables, are used to examine the relationship between two categorical variables. They present the frequency or count of observations falling into various categories of both variables, displayed in a tabular format. By analyzing the counts in each cell of the table, patterns or dependencies between the two variables can be identified. Cross-tabs are helpful in understanding how the categories of one variable are distributed among the categories of another variable and can provide insights into their relationship.