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

Ans- There are several key tasks involved in getting ready to work with machine learning model including 1.Defining the problem 2. Gathering and cleaning data 3. Exploring and visualizing data 4. Selecting and engineering features 5. Training and validating the model 6. Tuning and optimizing the model 7. Deploying and maintaining the model.

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

Ans- There are generally three forms of data used in machine learning:

1. Numerical Data: This type of data consists of numeric values and can be further classified as continuous or discrete. Continuous data is measured over a continuous range and can take any value within that range (e.g., height, weight, temperature). Discrete data is counted and has a finite set of possible values (e.g., number of children, number of cars).

Example: The temperature of a city on a particular day.

1. Categorical data: This type of data includes discrete values that represent categories, such as gender, occupation, or race. Categorical data can be nominal or ordinal. Nominal data has no order, while ordinal data has a natural order. For example, in a survey, the gender and occupation of respondents would be categorical data.
2. Text Data: This type of data consists of textual information and can be further classified as unstructured or semi-structured. Unstructured text data does not have any predefined structure (e.g., emails, social media posts, news articles). Semi-structured text data has a defined structure or schema (e.g., medical records, financial statements).

Example: Customer reviews of a product on an e-commerce website.

**3. Distinguish:**

**1. Numeric vs. categorical attributes**

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

Ans- 1.Numeric vs. categorical attributes:

Numeric attributes represent numerical values such as age, height, weight, temperature, etc., while categorical attributes represent non-numerical values such as gender, color, type, category, etc. Numeric attributes can be measured on a continuous scale, while categorical attributes represent discrete values that cannot be ordered or compared directly.

In data analysis and machine learning, it is important to understand the type of attribute being analyzed, as different algorithms and techniques are applied to handle numeric and categorical data. For example, statistical models typically require numerical data, while decision trees and rule-based models can handle both numerical and categorical data.

2.Feature selection vs. dimensionality reduction:

Feature selection and dimensionality reduction are both techniques used to reduce the number of features or variables in a dataset. However, they differ in their approach and purpose.

Feature selection refers to the process of selecting a subset of the most important features or variables from a larger set of features. The goal of feature selection is to simplify the model and reduce the risk of overfitting. The selected features are often based on their relevance to the target variable and their ability to improve the model's accuracy.

Dimensionality reduction, on the other hand, refers to the process of transforming the original set of features into a smaller set of new features while preserving most of the relevant information. The goal of dimensionality reduction is to eliminate redundant features that may lead to computational complexity and improve the model's performance.

In summary, feature selection aims to identify and keep the most important features, while dimensionality reduction aims to transform the original features into a smaller set of more informative features. Both techniques can help improve the model's accuracy and efficiency, and the choice between them depends on the specific goals and characteristics of the dataset being analyzed.

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

**1. The histogram**

**2. Use a scatter plot**

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

Ans-1.The histogram:

A histogram is a graphical representation of the distribution of numerical data. It is used to show how many data points fall within a certain range or bin. The x-axis of the histogram represents the range of values being considered, and the y-axis represents the frequency of data points falling within that range. Histograms can be used to identify the shape of the distribution, identify outliers, and assess the central tendency and spread of the data.

2.Use a scatter plot:

A scatter plot is a graphical representation of the relationship between two variables. It is used to visualize how one variable affects the other. The x-axis represents one variable, and the y-axis represents the other variable. Each data point is plotted as a point on the graph, and the pattern of the points can reveal the strength and direction of the relationship between the variables. Scatter plots are useful for identifying trends, identifying outliers, and assessing the strength and direction of the relationship between variables.

3.PCA :

PCA (Principal Component Analysis) is a dimensionality reduction technique used to transform a high-dimensional dataset into a lower-dimensional dataset while retaining most of the original information. It is a mathematical algorithm that identifies the most important features or variables in the data and generates new variables (called principal components) that capture the most variation in the data. PCA can be used to simplify complex datasets, visualize high-dimensional data, and improve the performance of machine learning algorithms

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

Ans-Investigating data is necessary to gain insights and understanding of the underlying patterns, relationships, and characteristics of the data. It helps to identify trends, outliers, and potential issues that may affect the accuracy and validity of the results obtained from the data. By analyzing and exploring data, one can make informed decisions, develop effective strategies, and solve problems.

There is a discrepancy in how qualitative and quantitative data are explored. Qualitative data, such as interviews, observations, and open-ended survey responses, are typically analyzed through a process of coding and categorizing data into themes or patterns. This approach is often more exploratory and subjective, and it involves interpreting and making sense of the data in a holistic manner.

In contrast, quantitative data, such as numerical measurements and counts, are typically analyzed using statistical methods and techniques. This approach involves testing hypotheses and making objective inferences based on the data. Quantitative data exploration often focuses on measures of central tendency, variability, and relationships between variables.Overall, both qualitative and quantitative data exploration are important for gaining a comprehensive understanding of the data. While the approaches may differ, the ultimate goal is to uncover meaningful insights that can inform decision-making and problem-solving

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

Ans- Histograms can have various shapes, and the shape of a histogram can provide insights into the underlying distribution of the data. Some common histogram shapes are Normal distribution, Skewed distribution, Bimodal distribution, Uniform distribution.

Bins are intervals or ranges of values used to group the data in a histogram. The x-axis of a histogram is divided into these intervals or bins, and the frequency of data falling within each bin is represented by the height of the bar above the bin. The number and size of the bins can affect the shape and appearance of the histogram and should be chosen carefully based on the characteristics of the data and the purpose of the analysis. A small number of bins may result in a loss of detail, while a large number of bins may result in overfitting or noise in the data.

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

Ans- Outliers are data points that are significantly different from other data points in a dataset. Outliers can occur due to measurement error, data entry errors, or genuine variation in the data. Dealing with outliers is important because they can affect the accuracy and validity of statistical analyses and machine learning algorithms. Here are some ways to deal with data outliers such as Remove outliers, Identify the cause, Transform data, Use robust statistical methods, keep outliers and report them separately.

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

Ans- There are several measures of central tendency used to summarize a dataset :

1.Mean 2.Median and 3.Mode

The mean can vary significantly from the median in certain datasets when there are extreme values or outliers present in the data. This is because the mean is sensitive to extreme values, while the median is not.

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

Ans- A scatter plot is a graph that is used to investigate the relationship between two variables in a dataset. It displays the values of one variable on the horizontal axis and the values of the other variable on the vertical axis. Each data point is represented as a dot on the plot, and the pattern of dots can reveal the nature of the relationship between the variables.

Scatter plots are useful for investigating bivariate relationships because they allow us to see if there is a pattern or trend between the two variables. We can look for positive or negative associations between the variables (when one variable increases, does the other increase or decrease?), as well as the strength and direction of the relationship. A scatter plot can also be used to identify outliers, which are data points that are significantly different from the others in the dataset. Outliers can be identified as points that fall far away from the general pattern of the other data points.

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

Ans- Cross-tabulation, or crosstab for short, is a statistical tool used to examine the relationship between two categorical variables. To create a cross-tab, the data is organized into a table with one variable on the rows and the other variable on the columns. The intersection of each row and column represents the frequency or count of the combination of the two variables.

Once the cross-tab is created, it can be used to explore the relationship between the two variables. The table can be analyzed to identify patterns and trends in the data, and to calculate various summary statistics, such as marginal distributions and conditional probabilities. For example, suppose we have data on the gender and occupation of a group of people, and we want to explore the relationship between these two variables.

By analyzing the cross-tab, we can gain insights into the relationship between the two variables and identify any patterns or trends in the data.