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

ANS.Some key tasks involved in getting ready to work with machine learning modeling include defining the problem, collecting and preparing data, selecting a model and evaluation metric, training and tuning the model, and testing and deploying the model.

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

ANS.The different forms of data used in machine learning include numerical data (e.g. height, weight), categorical data (e.g. gender, color), ordinal data (e.g. rating scale), and text data (e.g. reviews). An example of numerical data would be a person's age, an example of categorical data would be a person's occupation, an example of ordinal data would be a product rating on a scale of 1-5, and an example of text data would be a product review.

3.Distinguish:

ANS.

Numeric vs. categorical attributes

Numeric attributes are continuous or discrete numerical values, while categorical attributes are values that fall into distinct categories or groups. An example of a numeric attribute is a person's age, while an example of a categorical attribute is a person's gender.

Feature selection vs. dimensionality reduction

Feature selection involves selecting a subset of the original features, while dimensionality reduction involves transforming the original features into a smaller set of new features. Feature selection aims to improve model performance by removing irrelevant or redundant features, while dimensionality reduction aims to reduce the complexity of the model.

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

ANS.

The histogram

A histogram is a graphical representation of the distribution of a dataset. It consists of a set of bars that represent the frequency of observations falling within a particular range, called a bin. Histograms are useful for visualizing the shape of the distribution, including its central tendency and spread.

Use a scatter plot

A scatter plot is a graphical representation of the relationship between two variables. It consists of a set of points that represent the values of the two variables for each observation. Scatter plots are useful for identifying patterns and relationships between variables, including any outliers.

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 understand the characteristics and patterns of the data, identify any issues or anomalies, and select appropriate methods for analysis. There may be some differences in how qualitative and quantitative data are explored, as qualitative data often requires more subjective interpretation and may be less structured than quantitative data.

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

ANS.The various histogram shapes include symmetric, skewed left, skewed right, bimodal, and uniform. Bins are the intervals used to group the data in a histogram, which determines the width of the bars and the granularity of the representation.

7.How do we deal with data outliers?

ANS.There are various methods for dealing with data outliers, including removing them, transforming them, or treating them as a separate category. The choice of method depends on the nature and extent of the outliers and the goals of the analysis.

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

ANS.The various central inclination measures include the mean, median, and mode. The mean is the average value of the data, the median is the middle value when the data is ordered, and the mode is the most common value in the data. The mean can vary too much from the median in certain data sets that have outliers or a skewed distribution, which can disproportionately affect the mean.

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 can be used to investigate bivariate relationships by plotting the values of two variables on a two-dimensional graph, where each point represents an observation. By examining the pattern of points, it is possible to identify any relationship between the variables, such as a positive or negative correlation. Outliers can also be identified in a scatter plot as points that are significantly distant from the other points in the plot.

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

ANS.Cross-tabs, or cross-tabulations, can be used to examine the relationship between two variables by creating a table that displays the frequency distribution of the two variables. Each cell in the table represents the number or percentage of observations that have a particular combination of values for the two variables. By examining the distribution of values in each cell, it is possible to identify any patterns or relationships between the two variables, such as an association or dependency.