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

There are several key tasks involved in getting ready to work with machine learning modeling. Here are some of the most important ones:

Define the problem: The first step is to clearly define the problem that you are trying to solve. This will help you choose the appropriate machine learning algorithm and data to use.

Collect and preprocess the data: Collecting and preprocessing data is a crucial task in machine learning modeling. You need to ensure that the data is relevant, accurate, and clean, as well as properly formatted and organized for analysis.

Choose the right algorithm: Choosing the right machine learning algorithm is critical to the success of your model. You need to evaluate which algorithms are suitable for the type of problem you are trying to solve.

Train and validate the model: Once you have chosen the algorithm, you need to train and validate the model using your data. This involves splitting your data into training and testing sets, fitting the model to the training data, and evaluating its performance on the testing data.

Tune the model: You may need to fine-tune the model by adjusting its parameters to optimize its performance on the testing data.

Deploy the model: Once the model is trained and validated, you can deploy it to make predictions on new data. You will also need to monitor the model's performance and update it as necessary.

Overall, getting ready to work with machine learning modeling involves a careful and thorough approach to data collection, algorithm selection, model training, and performance evaluation. It requires a strong foundation in statistical analysis, programming, and data processing, as well as a deep understanding of the underlying machine learning principles.

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

There are several forms of data used in machine learning, and each form has its own unique properties and characteristics. Here are some of the most common forms of data used in machine learning, along with a specific example for each of them:

Numerical Data: Numerical data is a type of data that is expressed in numerical form, such as age, weight, or temperature. Numerical data is typically used for regression analysis, where the goal is to predict a continuous numerical value. An example of numerical data is the height of individuals in a population.

Categorical Data: Categorical data is a type of data that represents different categories or labels, such as colors, gender, or occupation. Categorical data is typically used for classification analysis, where the goal is to predict which category an observation belongs to. An example of categorical data is the type of car owned by individuals in a population.

Text Data: Text data is a type of data that consists of words, sentences, and paragraphs. Text data is typically used for natural language processing (NLP) tasks, such as sentiment analysis, topic modeling, or text classification. An example of text data is product reviews written by customers.

Image Data: Image data is a type of data that represents visual information in the form of pixels, such as photographs, drawings, or maps. Image data is typically used for computer vision tasks, such as object detection, image segmentation, or facial recognition. An example of image data is satellite imagery used to identify changes in land use patterns over time.

Time Series Data: Time series data is a type of data that represents a sequence of observations over time, such as stock prices, weather patterns, or web traffic. Time series data is typically used for forecasting tasks, where the goal is to predict future values based on historical patterns. An example of time series data is the monthly sales figures for a retail store over the past year.

Overall, the different forms of data used in machine learning are diverse and varied, and each requires a unique set of techniques and tools to analyze and model effectively.

3. Distinguish:

1. Numeric vs. categorical attributes

1. Feature selection vs. dimensionality reduction

Numeric vs. categorical attributes:

Numeric attributes are those that have a quantitative or numerical value, such as height, weight, or age. Numeric attributes can be further divided into continuous or discrete, where continuous attributes can take on any value within a range, while discrete attributes have a finite number of possible values.

Categorical attributes, on the other hand, are those that represent different categories or labels, such as colors, gender, or occupation. Categorical attributes can be nominal or ordinal, where nominal attributes have no intrinsic order, while ordinal attributes have a natural order.

The main difference between numeric and categorical attributes is the type of information they represent. Numeric attributes are useful for numerical analysis, while categorical attributes are useful for classification or grouping.

Feature selection vs. dimensionality reduction:

Feature selection and dimensionality reduction are two techniques used to reduce the number of features in a dataset, but they differ in their approach and goals.

Feature selection is the process of selecting a subset of relevant features from a larger set of features. The goal of feature selection is to reduce the dimensionality of the data while retaining the most informative features. Feature selection can be done by using statistical tests, correlation analysis, or model-based methods.

Dimensionality reduction, on the other hand, is the process of transforming high-dimensional data into a lower-dimensional space while preserving as much of the original information as possible. The goal of dimensionality reduction is to reduce the complexity of the data and remove redundant or irrelevant features. Dimensionality reduction can be done by using techniques such as principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), or autoencoders.

The main difference between feature selection and dimensionality reduction is their approach. Feature selection is a subset selection approach, while dimensionality reduction is a feature extraction approach. Feature selection keeps the original features and selects the most informative ones, while dimensionality reduction creates new features that are a combination of the original ones.

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

1. The histogram

2. Use a scatter plot

3.PCA (Personal Computer Aid)

Feature selection vs. dimensionality reduction:

Feature selection and dimensionality reduction are two techniques used to reduce the number of features in a dataset, but they differ in their approach and goals.

Feature selection is the process of selecting a subset of relevant features from a larger set of features. The goal of feature selection is to reduce the dimensionality of the data while retaining the most informative features. Feature selection can be done by using statistical tests, correlation analysis, or model-based methods.

Dimensionality reduction, on the other hand, is the process of transforming high-dimensional data into a lower-dimensional space while preserving as much of the original information as possible. The goal of dimensionality reduction is to reduce the complexity of the data and remove redundant or irrelevant features. Dimensionality reduction can be done by using techniques such as principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), or autoencoders.

The main difference between feature selection and dimensionality reduction is their approach. Feature selection is a subset selection approach, while dimensionality reduction is a feature extraction approach. Feature selection keeps the original features and selects the most informative ones, while dimensionality reduction creates new features that are a combination of the original ones.

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

It is necessary to investigate data because it allows us to understand the characteristics of the dataset, identify patterns, trends, and relationships, and make informed decisions based on the data. By exploring data, we can also detect errors, outliers, and anomalies that could affect the validity of the analysis

Yes, there is a discrepancy in how qualitative and quantitative data are explored. Quantitative data is usually explored using statistical techniques and numerical summaries, while qualitative data is typically explored using techniques such as content analysis and thematic analysis. These differences in exploration techniques arise due to the nature of the data.

Quantitative data consists of numerical values that can be measured and analyzed using mathematical or statistical tools. Quantitative data can be analyzed using various statistical techniques such as regression analysis, ANOVA, and hypothesis testing. Exploratory data analysis for quantitative data typically involves using graphical tools and numerical summaries to understand the distribution, central tendency, and variability of the data.

Qualitative data, on the other hand, consists of non-numerical data such as text, images, or audio recordings. Qualitative data cannot be measured using mathematical or statistical tools, and it requires a different set of techniques for analysis. Qualitative data can be analyzed using content analysis, thematic analysis, or narrative analysis to identify themes, patterns, and categories within the data. Exploratory data analysis for qualitative data typically involves identifying and coding themes within the data and using visual aids such as word clouds or network diagrams to summarize the main themes and relationships between them.

Therefore, there is a discrepancy in how qualitative and quantitative data are explored due to their different natures and the methods required to analyze them effectively.

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

Histograms can take on different shapes depending on the distribution of the data being plotted. Some common histogram shapes include:

Normal distribution: This is a bell-shaped curve where the data is symmetrical around the mean, with the majority of the data clustered around the center.

Skewed distribution: This is a distribution where the data is not symmetrical around the mean, and the tail of the distribution is longer on one side than the other. Skewed distributions can be either left-skewed (negative skewness) or right-skewed (positive skewness).

Bimodal distribution: This is a distribution with two peaks, indicating that the data may come from two different populations.

Uniform distribution: This is a distribution where the data is evenly spread across the range of values, with no peaks or troughs.

Bins, also known as intervals or buckets, are the intervals into which the data is divided in a histogram. Each bin represents a range of values, and the height of the bar in the histogram represents the frequency of data points falling within that range. The number of bins used in a histogram can affect the visual interpretation of the data. Too few bins may hide important patterns in the data, while too many bins may create noise in the visualization. Choosing an appropriate number of bins is often a matter of trial and error, and depends on the size of the dataset and the characteristics of the data being plotted.

1. How do we deal with data outliers?

Creating boundaries: One method for dealing with outliers is to create boundaries or thresholds beyond which any data point is considered an outlier. This can be done by setting a minimum and maximum value for the variable based on prior knowledge or statistical methods. Any data point that falls outside the established boundaries can be considered an outlier and dealt with accordingly.

Winsorization: Winsorization is a technique that replaces extreme values with less extreme values. The most extreme values are replaced with the next highest or lowest value in the dataset. Winsorization can be performed on one or both tails of the distribution.

Imputation: Imputation involves replacing missing or extreme values with a new value. Imputation can be done using statistical methods such as mean imputation or regression imputation, or using more sophisticated methods such as decision trees or clustering.

Transformation: Transformation involves applying a mathematical function to the data to make it more normally distributed. This can help to reduce the impact of extreme values on the analysis. Common transformations include logarithmic or square root transformations.

Using robust statistical methods: Robust statistical methods are less sensitive to outliers than traditional statistical methods. Examples of robust statistical methods include the median and trimmed mean.

It is important to choose an appropriate method for dealing with outliers based on the nature of the data and the goals of the analysis. Additionally, it is important to carefully document any methods used for dealing with outliers in order to ensure transparency and reproducibility of the analysis.

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

Central tendency measures are used to describe the central or typical value of a dataset. The three most commonly used measures of central tendency are the mean, median, and mode.

Mean: The mean is the sum of all the values in a dataset divided by the total number of values. It is commonly used when the data is normally distributed and has no extreme values.

Median: The median is the middle value in a dataset when the values are arranged in order. It is useful when the data has extreme values or is skewed.

Mode: The mode is the most frequent value in a dataset. It is useful when the data has a clear peak or mode.

The mean and median are often used together to get a more complete picture of the central tendency of the data. In some datasets, the mean varies significantly from the median. This can occur when the data has extreme values, also known as outliers. Outliers can significantly affect the mean, pulling it towards the direction of the outliers. On the other hand, the median is not affected by extreme values, as it only takes the middle value of the dataset.

For example, consider a dataset with values {2, 3, 4, 5, 50}. The mean of this dataset is (2+3+4+5+50)/5 = 12.8, while the median is 4. The mean is much larger than the median because the outlier value of 50 pulls the mean towards the right.

Therefore, it is important to use both the mean and median when analyzing datasets, especially when the data has extreme values or is skewed. The choice of which central tendency measure to use depends on the nature of the data and the research question being addressed.

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

A scatter plot is a type of visualization used to explore the relationship between two variables in a dataset. It consists of a set of data points plotted on a two-dimensional graph, with one variable on the x-axis and the other variable on the y-axis.

Scatter plots are particularly useful for investigating bivariate relationships, which are relationships between two variables. By plotting the two variables against each other, patterns or trends in the data can be visualized. For example, a scatter plot can be used to determine whether there is a positive or negative relationship between two variables, or whether there is no relationship at all.

In addition to exploring the relationship between two variables, scatter plots can also be used to identify outliers. An outlier is a data point that is significantly different from other data points in the dataset. In a scatter plot, an outlier can be identified as a data point that is located far away from the other data points. Outliers can be important to identify because they can have a significant impact on the results of statistical analyses.

To identify outliers using a scatter plot, it is important to visually examine the plot and look for any data points that are located far away from the other data points. However, it is important to note that the identification of outliers using a scatter plot is a subjective process, and there is no objective rule for determining what constitutes an outlier. In addition, the presence of outliers in a dataset may require special attention when conducting statistical analysis, as outliers can significantly affect the results of analyses such as regression.

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

Cross-tabulation, also known as contingency table analysis, is a statistical method used to explore the relationship between two categorical variables. It involves creating a table that shows the frequency distribution of one variable broken down by the categories of the other variable.

To use cross-tabs to figure out how two variables are related, follow these steps:

Determine the two categorical variables of interest.

Create a contingency table that shows the frequency distribution of one variable by the categories of the other variable. For example, if the two variables are gender and occupation, the contingency table will show the number of males and females in each occupation category.

Analyze the contingency table to identify any patterns or relationships between the two variables. This can be done by examining the frequencies of each category, calculating percentages, or using statistical tests such as the chi-square test.

Interpret the results of the analysis. Based on the patterns or relationships identified, it may be possible to draw conclusions about how the two variables are related.

