1. **What are the key tasks that machine learning entails? What does data pre-processing imply?**

**Ans:-**

Machine learning is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms are trained on data, and they learn to identify patterns in the data that can be used to make predictions.

The key tasks of machine learning include:

* **Data pre-processing:** This is the process of cleaning and preparing the data for machine learning algorithms. This includes tasks such as removing outliers, imputing missing values, and transforming the data into a format that the machine learning algorithm can understand.
* **Model training:** This is the process of training the machine learning algorithm on the data. This involves feeding the algorithm the data and allowing it to learn the patterns in the data.
* **Model evaluation:** This is the process of evaluating the performance of the machine learning algorithm. This involves testing the algorithm on a held-out dataset and measuring its accuracy.
* **Model deployment:** This is the process of deploying the machine learning algorithm into production. This involves making the algorithm available to users so that they can use it to make predictions.

Data pre-processing is a critical step in machine learning. It is important to ensure that the data is clean and prepared correctly before it is used to train the machine learning algorithm. If the data is not pre-processed properly, the machine learning algorithm may not be able to learn the patterns in the data and it may not be able to make accurate predictions.

1. **Describe quantitative and qualitative data in depth. Make a distinction between the two.**

**Ans:-**

Quantitative data is data that can be measured and expressed in numbers. It can be either discrete or continuous. Discrete data is data that can only take on a finite number of values. For example, the number of students in a class is discrete data. Continuous data is data that can take on an infinite number of values. For example, the height of a person is continuous data.

Qualitative data is data that cannot be measured or expressed in numbers. It is data that describes qualities or attributes. For example, the color of a car is qualitative data.

The main difference between quantitative and qualitative data is that quantitative data can be measured and expressed in numbers, while qualitative data cannot. Quantitative data is often used in machine learning because it can be easily processed by machine learning algorithms. Qualitative data is often used in natural language processing because it can be used to understand the meaning of text.

1. **Create a basic data collection that includes some sample records. Have at least one attribute from each of the machine learning data types.**

**Ans:-**

1. **What are the various causes of machine learning data issues? What are the ramifications?**

**Ans:-**

There are many different causes of machine learning data issues. Some of the most common causes include:

* **Missing values:** This occurs when there is no data for a particular variable in a record.
* **Outliers:** These are data points that are far outside the range of the other data points.
* **Noise:** This is random variation in the data that can make it difficult to identify patterns.
* **Bias:** This occurs when the data is not representative of the population that the machine learning algorithm is trying to model.

The ramifications of machine learning data issues can be significant. If the data is not clean and prepared correctly, the machine learning algorithm may not be able to learn the patterns in the data and it may not be able to make accurate predictions. This can lead to problems such as financial losses, incorrect medical diagnoses, and incorrect product recommendations.

1. **Demonstrate various approaches to categorical data exploration with appropriate examples.**

**Ans:-**

There are many different approaches to categorical data exploration. Some of the most common approaches include:

* **Frequency tables:** Frequency tables show the number of times each category appears in the data. This can be a useful way to get an overview of the distribution of the categorical data.
* **Bar charts:** Bar charts show the frequency of each category as a bar. This can be a more visually appealing way to represent the frequency of the categorical data.
* **Pie charts:** Pie charts show the frequency of each category as a slice of a pie. This can be a useful way to represent the relative frequencies of the categorical data.
* **Cross-tabs:** Cross-tabs show the frequency of each category in combination with another category. This can be a useful way to explore the relationships between categorical variables.

For example, let's say we have a data set that contains information about the gender and favorite color of a group of people. We can use frequency tables, bar charts, pie charts, and cross-tabs to explore the distribution of gender and favorite color in the data set.

The frequency table for gender would show the number of men and women in the data set. The bar chart for gender would show the number of men and women as bars. The pie chart for gender would show the number of men and women as slices of a pie. The cross-tab for gender and favorite color would show the number of men and women who prefer each color.

1. **How would the learning activity be affected if certain variables have missing values? Having said that, what can be done about it?**

**Ans:-**

Missing values can affect the learning activity in a number of ways. If a large number of values are missing, then the machine learning algorithm may not be able to learn the patterns in the data and it may not be able to make accurate predictions.

There are a number of things that can be done about missing values. One option is to simply ignore the records with missing values. However, this can lead to bias in the data, as the records with missing values may be different from the records with complete data.

Another option is to impute the missing values. This means filling in the missing values with estimates. There are a number of different imputation methods available, such as mean imputation, median imputation, and multiple imputation.

The best approach to dealing with missing values depends on the specific data set and the machine learning algorithm that is being used. However, it is important to consider the potential impact of missing values on the learning activity before making a decision about how to deal with them.

1. **Describe the various methods for dealing with missing data values in depth.**

**Answer:-**

There are a number of different methods for dealing with missing data values. Some of the most common methods include:

* **Mean imputation:** This is the simplest method for dealing with missing data. It involves replacing the missing values with the mean of the observed values.
* **Median imputation:** This is similar to mean imputation, but it uses the median of the observed values instead of the mean.
* **Mode imputation:** This method replaces the missing values with the most frequent value in the data set.
* **Multiple imputation:** This is a more sophisticated method for dealing with missing data. It involves creating multiple versions of the data set, each with a different imputation method. This allows the machine learning algorithm to learn from the different versions of the data set and to make more accurate predictions.

The best method for dealing with missing data values depends on the specific data set and the machine learning algorithm that is being used. However, it is important to consider the potential impact of missing values on the learning activity before making a decision about how to deal with them.

1. **What are the various data pre-processing techniques? Explain dimensionality reduction and function selection in a few words.**

**Answer:-**

Data pre-processing is the process of cleaning and preparing data for machine learning algorithms. This includes tasks such as:

* **Missing value imputation:** This involves filling in missing values with estimates.
* **Outlier detection:** This involves identifying data points that are far outside the range of the other data points.
* **Feature scaling:** This involves normalizing the data so that all of the features have a similar scale.
* **Dimensionality reduction:** This involves reducing the number of features in the data set.
* **Function selection:** This involves selecting the most important features for the machine learning algorithm.

Dimensionality reduction is the process of reducing the number of features in a data set. This can be done by using techniques such as principal component analysis (PCA) and feature selection. Function selection is the process of selecting the most important features for the machine learning algorithm. This can be done by using techniques such as recursive feature elimination (RFE) and stepwise selection.

**9.**

**i. What is the IQR? What criteria are used to assess it?**

**Answer:-**

The interquartile range (IQR) is a measure of variability in a data set. It is calculated by finding the difference between the third and first quartiles. The IQR is a useful measure of variability because it is not affected by outliers.

The criteria used to assess the IQR are:

* The IQR should be large enough to capture the majority of the data points in the data set.
* The IQR should be small enough to be a useful measure of variability.

**ii. Describe the various components of a box plot in detail? When will the lower whisker surpass the upper whisker in length? How can box plots be used to identify outliers?**

**Answer:-**

A box plot is a graphical representation of the distribution of data. It shows the median, quartiles, and outliers in a data set.

The components of a box plot are:

* **The median:** The median is the middle value in a sorted data set. It divides the data set into two equal halves.
* **The first quartile (Q1):** The first quartile is the value below which 25% of the data points lie.
* **The third quartile (Q3):** The third quartile is the value above which 75% of the data points lie.
* **The interquartile range (IQR):** The IQR is the difference between Q3 and Q1.
* **The lower whisker:** The lower whisker extends from Q1 to the smallest data point that is within 1.5 IQRs of Q1.
* **The upper whisker:** The upper whisker extends from Q3 to the largest data point that is within 1.5 IQRs of Q3.

The lower whisker will surpass the upper whisker in length when there are outliers in the data set that are located below Q1 or above Q3.

Box plots can be used to identify outliers by looking for data points that fall outside the whiskers. Outliers are data points that are far outside the range of the other data points. They can be caused by measurement errors, data entry errors, or genuine anomalies.

**10. Make brief notes on any two of the following:**

* 1. **Data collected at regular intervals**

**Answer:-**

Data collected at regular intervals is called **time series data**. Time series data is often used to track changes over time, such as the price of a stock or the number of visitors to a website.

Time series data can be analyzed using a variety of statistical techniques, such as **simple moving averages**, **exponential smoothing**, and **ARIMA models**. These techniques can be used to forecast future values of the time series or to identify trends and patterns in the data.

* 1. **The gap between the quartiles**

**Answer:-**

The gap between the quartiles is called the **interquartile range (IQR)**. The IQR is a measure of variability in a data set. It is calculated by finding the difference between the third and first quartiles. The IQR is a useful measure of variability because it is not affected by outliers.

The IQR can be used to identify outliers in a data set. Outliers are data points that are far outside the range of the other data points. They can be caused by measurement errors, data entry errors, or genuine anomalies.

* 1. **Use a cross-tab**

**Answer:-**

A cross-tab is a table that shows the frequency of each combination of values of two variables. It is a way of summarizing the relationship between two variables.

Cross-tabs can be used to answer the following questions:

* How are the two variables related?
* Are the two variables independent?
* What is the strength of the relationship between the two variables?

Cross-tabs can be used to find outliers by looking for cells with very low or very high frequencies. Outliers can be caused by measurement errors, data entry errors, or genuine anomalies.

**11. Make a comparison between:**

* 1. **Data with nominal and ordinal values**

**Answer:-**

* **Nominal data** is data that is used to classify or categorize observations. It has no inherent order or ranking. For example, the data "male" and "female" are nominal data.
* **Ordinal data** is data that has a natural order or ranking. However, the difference between two ordinal data points is not necessarily meaningful. For example, the data "low", "medium", and "high" are ordinal data.

The main difference between nominal and ordinal data is that nominal data has no inherent order, while ordinal data has a natural order. This means that nominal data cannot be meaningfully sorted or compared, while ordinal data can be.

* 1. **Histogram and box plot**

**Answer:-**

* **Histogram** is a graphical representation of the distribution of data. It shows the frequency of each value in the data set.
* **Box plot** is a graphical representation of the distribution of data. It shows the median, quartiles, and outliers in the data set.

The main difference between a histogram and a box plot is that a histogram shows the frequency of each value in the data set, while a box plot shows the median, quartiles, and outliers in the data set.

* 1. **The average and median**

**Answer:-**

* **The average** is the sum of the values in a data set divided by the number of values in the data set.
* **The median** is the middle value in a sorted data set.

The main difference between the average and the median is that the average is affected by outliers, while the median is not. This means that the median is a more robust measure of central tendency than the average.