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

Machine learning involves several key tasks such as data collection, data cleaning, data pre-processing, feature selection, model building, model training, and model evaluation. Data pre-processing refers to the process of cleaning and transforming raw data into a format that can be used for machine learning. It includes tasks such as data cleaning, data normalization, data transformation, and data reduction.

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

Quantitative data is numerical data that can be measured and expressed using numerical values, such as height, weight, age, and temperature. Qualitative data is non-numerical data that is descriptive in nature and cannot be measured using numbers, such as color, taste, and smell. The main difference between the two is that quantitative data can be measured and expressed numerically, while qualitative data cannot.

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

Sample data collection:

Name Age Gender Income Marital Status

John 32 Male 50000 Married

Sarah 25 Female 30000 Single

Michael 45 Male 75000 Divorced

Samanth 28 Female 40000 Married

David 40 Male 60000 Single

The data types used in this data collection are:

Nominal: Gender, Marital Status

Ordinal: None

Interval: Age, Income

Ratio: None

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

Various causes of machine learning data issues include missing values, inconsistent formatting, outliers, noisy data, and imbalanced data. The ramifications of these issues can be severe, as they can lead to inaccurate predictions, biased models, and reduced model performance.

5.Demonstrate various approaches to categorical data exploration with appropriate examples.

Various approaches to categorical data exploration include frequency tables, bar charts, pie charts, and stacked bar charts. For example, a frequency table can be used to show the number of occurrences of each category in a dataset, while a bar chart can be used to visualize the relative frequencies of each category.

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

If certain variables have missing values, the learning activity can be affected in several ways. It can lead to biased models, inaccurate predictions, and reduced model performance. To address this issue, missing values can be imputed using various techniques such as mean imputation, median imputation, and regression imputation.

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

Various methods for dealing with missing data values include listwise deletion, pairwise deletion, mean imputation, median imputation, mode imputation, regression imputation, and multiple imputation. Listwise deletion involves removing entire cases with missing values, while pairwise deletion involves using available data for each pairwise comparison. Imputation techniques involve estimating missing values based on the available data.

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

Various data pre-processing techniques include data cleaning, data normalization, data transformation, and data reduction. Dimensionality reduction refers to the process of reducing the number of features in a dataset, while function selection involves selecting the most relevant features for a particular task.

9.

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

The IQR, or interquartile range, is a measure of the spread of a dataset. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1). The criteria used to assess the IQR depend on the context of the data being analyzed. Generally, a larger IQR indicates greater variability in the data, while a smaller IQR indicates less 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?

A box plot is a graphical representation of the distribution of a dataset. It consists of a box that represents the interquartile range (IQR) of the data, and whiskers that extend to the lowest and highest observations within a certain range. The median of the data is shown as a line within the box. Outliers are represented by points outside the whiskers.

The lower whisker of a box plot will surpass the upper whisker in length when the median is closer to the upper quartile (Q3) than the lower quartile (Q1), indicating that the data is skewed towards the upper end of the range.

Box plots can be used to identify outliers by looking for points outside the whiskers, which are typically defined as 1.5 times the IQR. Outliers may indicate errors in the data, or they may be genuine extreme values that should be analyzed separately.

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

Data collected at regular intervals

Data collected at regular intervals is known as time series data. It is used to analyze patterns and trends over time, such as stock prices, weather patterns, and website traffic. Time series data is typically analyzed using techniques such as autocorrelation, moving averages, and time series forecasting.

The gap between the quartiles

The gap between the quartiles is known as the interquartile range (IQR), and is a measure of the spread of a dataset. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1). The IQR is used to identify the middle 50% of the data and is a useful tool for detecting outliers and assessing the variability of a dataset.