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

Ans:- A machine learning task is the type of prediction or inference being made, based on the problem or question that is being asked, and the available data. For example, the classification task assigns data to categories, and the clustering task groups data according to similarity. Data preprocessing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms

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

Ans:- Quantitative data is anything that can be counted or measured; it refers to numerical data. Qualitative data is descriptive, referring to things that can be observed but not measured—such as colors or emotions

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

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

Ans:- Noisy data, dirty data, and incomplete data are the quintessential enemies of ideal Machine Learning. The definition of a ramification is an effect that came from a specific action. An example of a ramification is having trouble getting a job after quitting college

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

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

Ans:- Even in a well-designed and controlled study, missing data occurs in almost all research. Missing data can reduce the statistical power of a study and can produce biased estimates, leading to invalid conclusions.

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

* Ans:- Listwise or case deletion. ...
* Pairwise deletion. ...
* Mean substitution. ...
* Regression imputation. ...
* Last observation carried forward. ...
* Maximum likelihood. ...
* Expectation-Maximization. ...
* Multiple imputation

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

* Ans:- Data Cleaning/Cleansing. Cleaning “dirty” data. Real-world data tend to be incomplete, noisy, and inconsistent.
* Data Integration. Combining data from multiple sources.
* Data Transformation. Constructing data cube.
* Data Reduction. Reducing representation of data set

9.

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

Ans:- The interquartile range (IQR) is a measure of variability, based on dividing a data set into quartiles The interquartile range is calculated in much the same way as the range. All you do to find it is subtract the first quartile from the third quartile: IQR = Q3 – Q1. The interquartile range shows how the data is spread about the median

* 1. 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?

Ans:- A box and whisker plot—also called a box plot—displays the five-number summary of a set of data. When reviewing a box plot, an outlier is defined as a data point that is located outside the whiskers of the box plot

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

* 1. Data collected at regular intervals

Ans:- Interval data, also called an integer, is defined as a data type which is measured along a scale, in which each point is placed at equal distance from one another

* 1. The gap between the quartiles

Ans:- The interquartile range or IQR is the range of the middle half of a set of data. It is the difference between the upper quartile and the lower quartile.

* 1. Use a cross-tab

Ans:- Cross-tabulation (also cross-tabulation or crosstab) is one of the most useful analytical tools and a mainstay of the market research industry. Cross-tabulation analysis, also known as contingency table analysis, is most often used to analyze categorical (nominal measurement scale) data.

11. Make a comparison between:

1. Data with nominal and ordinal values

Ans:- Nominal data is classified without a natural order or rank, whereas ordinal data has a predetermined or natural order. On the other hand, numerical or quantitative data will always be a number that can be measured

1. Histogram and box plot

Ans:- A box and whisker plot is defined as a graphical method of displaying variation in a set of data. In most cases, a histogram analysis provides a sufficient display, but a box and whisker plot can provide additional detail while allowing multiple sets of data to be displayed in the same graph

1. The average and median

Ans:- The mean (average) of a data set is found by adding all numbers in the data set and then dividing by the number of values in the set. The median is the middle value when a data set is ordered from least to greatest