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

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

It involves below steps:

Getting the dataset

Importing libraries

Importing datasets

Finding Missing Data

Encoding Categorical Data

Splitting dataset into training and test set

Feature scaling

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

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

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

Many different methodologies can be used for data collection and analysis. Most are based around a core set of basic tools. These include interviews, focus group discussions, observation, photography, video, surveys, questionnaires and case studies.

Attribute:

It can be seen as a data field that represents the characteristics or features of a data object. For a customer, object attributes can be customer Id, address, etc. We can say that a set of attributes used to describe a given object are known as attribute vector or feature vector.

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

Noisy data, dirty data, and incomplete data are the quintessential enemies of ideal Machine Learning. The solution to this conundrum is to take the time to evaluate and scope data with meticulous data governance, data integration, and data exploration until you get clear data.

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

Categorical data are data recorded about units on variables which take values in a discrete set of categories. Examples of categorical variables are gender, citizenship, or number of children.

Visualizing Multivariate Categorical Data

* Prerequisites.
* Bar plots of contingency tables.
* Balloon plot.
* Mosaic plot.
* Correspondence analysis.

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

Missing values present in the dataset can impact the performance of the model by creating a bias in the dataset. This bias can create a lack of relatability and trustworthiness in the dataset. The loss in values might contain crucial insights or information for model development.

The cause of missing values can be data corruption or failure to record data. The handling of missing data is very important during the pre-processing of the dataset as many machine learning algorithms do not support missing values.

How to Handle Missing Data in Machine Learning: 5 Techniques

* Deductive Imputation
* Mean/Median/Mode Imputation.
* Regression Imputation.
* Stochastic Regression Imputation

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

How to Handle Missing Data in Machine Learning: 5 Techniques

* Deductive Imputation
* Mean/Median/Mode Imputation.
* Regression Imputation.
* Stochastic Regression Imputation

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

They are Data Cleaning/Cleansing, Data Integration, Data Transformation, and Data Reduction.

Feature selection is simply selecting and excluding given features without changing them. Dimensionality reduction transforms features into a lower dimension

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

In descriptive statistics, the interquartile range, also called the midspread, middle 50%, or H‑spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, IQR = Q₃ − Q₁.

* Unlike total range, the interquartile range has a breakdown point of 25%, and is thus often preferred to the total range.
* The IQR is used to build box plots, simple graphical representations of a probability distribution.
* The IQR is used in businesses as a marker for their income rates.
* For a symmetric distribution (where the median equals the midhinge, the average of the first and third quartiles), half the IQR equals the median absolute deviation (MAD).
* The median is the corresponding measure of central tendency.
* The IQR can be used to identify outliers (see below). The IQR also may indicate the skewness of the dataset.
* The quartile deviation or semi-interquartile range is defined as half the IQR.

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?

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

1. Data collected at regular intervals

2. The gap between the quartiles

3. Use a cross-tab

1. Make a comparison between:

1. Data with nominal and ordinal values

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.

2. Histogram and box plot

Histograms are preferred to determine the underlying probability distribution of a data. Box plots on the other hand are more useful when comparing between several data sets. They are less detailed than histograms and take up less space.

3. The average and median

The average is the arithmetic mean of a set of numbers. The median is a numeric value that separates the higher half of a set from the lower half

The term average is the sum of all the numbers divided by the total number of values in the set. The term mean is finding of the average of a sample data. Average is finding the central value in math, whereas mean is finding the central value in statistics.

The median is the middle number in a sorted, ascending or descending.