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

sol: The key tasks are: classification, regression and clustering.

Data-preprocessing implies that processing the data before feeding it to the ml algorithm. doing this will make learning easy and accurate and result in a better model.Some of the steps involved in this are data transformation, reducing the noice, normalisation , regularisation, etc.

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

sol: Quantitative data is the data that be counter or measured like annual income, temperature, price of a xyz item, etc where as on the other hand we have qualitative data which cannot be measured, it is observed like color, emotions, etc.

Quantitative data is analyzed using statistical analysis in which we play with numbers and try to to get the insights , where as qualitative data is analyzed by clustering the data into categories.

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

sol:

name(object) id(int64) increase(float) inIndia(boolean)

arnav 5678 67.9 True

radha 5676 89.9 False

shaam 5677 37.9 True

arun 5675 79.9 False

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

sol: Some of the causes are lack of quality data, too much noice in the data, a huge amount of null values present in the data, etc. The ramifications are majorly due to noice, and resulting into a model which understands the data in a way we do not want.

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

sol: We can visualize different graphs to understand the data better , getting the insights of the data like how many null values . Bar charts and pie charts are great tools for comparing two or more categorical values against each other. They just represent the number of things in a category.

For example, if you want to display the number of workers in a company, the outcomes can be presented on a pie chart or on a bar graph.Sometimes, it is difficult to distinguish between categorical and quantitative data.Quantitative data is measured and expressed numerically. It has numerical meaning and is used in calculations and arithmetic.

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

sol: If the missing values are very less compared to the data available , we can easily replace the missing values with the place holder and can train our model. like if there there is a salary column in the data with say, 7000 points and 10-100 are missing points , so we try to replace these missing values with either mean or the mode of the data.

but if the feature is having a lot of missing values , then it will be difficult to process and we cannot do anything except dropping that column.

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

Model-based progressive imputation uses previously imputed missing values to predict other missing values. Additional methods include Stochastic Regression Imputation, Multiple Imputations , Datawig, Hot-Deck imputation, Extrapolation, Interpolation, Listwise Deletion.

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

sol: data cleaning, resolving the null values, standardisation, dropping columns, checking for duplicates are the some of the basic pre-processing techniques.

Dimensionality reduction is a technique in which we try to reduce the dimensions of the data .PCA is the most common dimensionality reduction algorithm which reduce the dimensions by transforming the data and changing the axis of the data.

9.

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

sol: IQR is the inter quartile range which is difference between two consecutive quartile and use for outlier detection.

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?

sol: in box plot we have median , first quartile, second quartile , third quartile and also shows the outliers in the data.

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

1. Data collected at regular intervals

sol:A **time series** refers to any data which is collected at regular intervals over time. We often analyze time series data with a time series plot. The frequency of points in a time series plot can vary between plots based on the context, but there should be an equal gap between the points on a time series plot.

2. The gap between the quartiles

sol: In statistics, a quartile is a type of quantile which divides the number of data points into four parts, or quarters, of more-or-less equal size. the gap between quartiles are the IQR .

3. Use a cross-tab

sol: we can compute cross-tabulation of two or more different factors.