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

A:

Data preprocessing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models.

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

A:

Quantitative data is numbers-based, countable, or measurable. Qualitative data is interpretation-based, descriptive, and relating to language. Quantitative data tells us how many, how much, or how often in calculations. Qualitative data can help us to understand why, how, or what happened behind certain behaviors.

Quantitative research deals with numbers and statistics, while qualitative research deals with words and meanings. Quantitative methods allow you to systematically measure variables and test hypotheses. Qualitative methods allow you to explore concepts and experiences in more detail.

3. Create a basic data collection that includes some sample records. Have at least one attribute from

each of the machine learning data types.

A:

Data can come in many forms, but machine learning models rely on four primary data types. These include numerical data, categorical data, time series data, and text data.

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

A:

* Commonly used Algorithms in Machine Learning
* Inadequate Training Data
* Poor quality of data
* Non-representative training data
* Overfitting and Underfitting
* Monitoring and maintenance
* Getting bad recommendations
* Lack of skilled resources.

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

A:

* Value Counts
* Group by
* Cross tab
* Pivot Table
* Encoding Categorical Features

6. How would the learning activity be affected if certain variables have missing values? Having said

that, what can be done about it?

A:

Many machine learning algorithms fail if the dataset contains missing values. However, algorithms like K-nearest and Naive Bayes support data with missing values. You may end up building a biased machine learning model, leading to incorrect results if the missing values are not handled properly.

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

A:

There are 2 primary ways of handling missing values:

1. Deleting the Missing values
2. Imputing the Missing Values

#### Deleting the Missing value

Generally, this approach is not recommended. It is one of the quick and dirty techniques one can use to deal with missing values. If the missing value is of the type Missing Not At Random (MNAR), then it should not be deleted.

#### Imputing the Missing Value

There are many imputation methods for replacing the missing values. You can use different python libraries such as Pandas, and Sci-kit Learn to do this. Let’s go through some of the ways of replacing the missing values(mean, mode, median)

8. What are the various data pre-processing techniques? Explain dimensionality reduction and

feature selection in a few words.

A:

* Data Cleaning.
* Dimensionality Reduction.
* Feature Engineering.
* Sampling Data.
* Data Transformation.
* Imbalanced Data.

**Dimensionality reduction, or dimension reduction**:

is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its [intrinsic dimension](https://en.wikipedia.org/wiki/Intrinsic_dimension).

**Feature Selection** is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve.

9

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

A:

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.

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:

A box and whisker plot—also called a box plot—displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum. In a box plot, we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median.

Outliers: An extreme value that lies outside the lower or upper whiskers is designated an “Outlier”, i.e. values less than (Q1 – 1.5\*IQR) or greater than (Q3 + 1.5\*IQR)

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

1. Data collected at regular intervals

A:

Interval data is measured along a scale consisting of a certain number of values in a more theoretical sense. Each value is separated from the other at an equal distance. T

The distances which are between the values are termed intervals.

2. The gap between the quartiles

A:

It is defined as the difference between the 75th and 25th percentiles of the data. To calculate the IQR, the data set is divided into quartiles, or four rank-ordered even parts via linear interpolation.

3. Use a cross-tab

1. Make a comparison between:

1. Data with nominal and ordinal values

A:

Ordinal data is data that can be ranked or ordered. Examples include data taken from a poll or survey. Nominal data is data that can be made to fit various categories. Examples include whether an animal is a mammal, fish, reptile, amphibian, or bird.

2. Histogram and box plot

A:

WHAT ARE HISTOGRAMS AND BOX PLOTS? Histograms are a special kind of bar graph that shows a bar for a range of data values instead of a single value. A box plot is a data display that draws a box over a number line to show the interquartile range of the data.

1. The average and median

A:

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.