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

Solution - Key tasks for machine learning are given below as:

Gathering Data - Data Gathering is the first step of the machine learning life cycle. The goal of this step is to identify and obtain all data-related problems.

Data preparation - After collecting the data, we need to prepare it for further steps. Data

preparation is a step where we put our data into a suitable place and prepare it to use in our machine learning training.

Data Wrangling - Data wrangling is the process of cleaning and converting raw data into a useable format.

Analyse Data - The aim of this step is to build a machine learning model to analyse the data using various analytical techniques and review the outcome.

Train the model - In this step we train our model to improve its performance for better outcome of the problem.

Test the model - In this step, we check for the accuracy of our model by providing a test dataset to it.

Deployment - The last step of machine learning life cycle is deployment, where we deploy the model in the real-world system

Data pre-processing is the concept of changing the raw data into a clean data set. The dataset is pre-processed to check missing values, noisy data, and other inconsistencies before executing it to the algorithm.

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

Solution - Qualitative data describes qualities or characteristics. It is collected using questionnaires, interviews, or observation, and frequently appears in narrative form.

Quantitative data is data that can be counted or measured in numerical values. The two main types of quantitative data are discrete data and continuous data

Quantitative data refers to any information that can be quantified, counted or measured, and given a numerical value. Qualitative data is descriptive in nature, expressed in terms of language rather than numerical values.

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

Solution –

Here's a basic data collection with sample records that include at least one attribute from each of the machine learning data types**.** In this example, we have a data collection with five records. Each record has attributes representing different data types

Record	Numeric	Categorical	Text	Date
1	25	Male	"Lorem ipsum dolor sit amet"	2022-10-15
2	35.5	Female	"Sed ut perspiciatis unde omnis iste"	2023-02-28
3	42	Non-binary	"Nemo enim ipsam voluptatem"	2023-05-10
4	18.9	Male	"At vero eos et accusamus et iusto odio"	2023-01-07
5	67.2	Female	"Excepteur sint occaecat cupidatat non proident"	2022-11-20

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

Solution – Inadequate Training Data: The major issue that comes while using machine learning algorithms is the lack of quality as well as quantity of data.

Noisy Data- It is responsible for an inaccurate prediction that affects the decision as well as accuracy in classification tasks.

Incorrect data- It is also responsible for faulty programming and results obtained in machine learning models. Hence, incorrect data may affect the accuracy of the results also.

Generalizing of output data- Sometimes, it is also found that generalizing output data becomes complex, which results in comparatively poor future actions.

Overfitting: Whenever a machine learning model is trained with a huge amount of data, it starts capturing noise and inaccurate data into the training data set. It negatively affects the performance of the model.

Underfitting: Underfitting is just the opposite of overfitting. Whenever a machine learning model is trained with fewer amounts of data, and as a result, it provides incomplete and inaccurate data and destroys the accuracy of the machine learning model.

Slow implementations and results: However, machine learning models are highly efficient in producing accurate results but are time-consuming.

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

Solution – Categorical data represent characteristics that one can observe and sort into groups. If this data happens to be numerical, then the numbers would not have any mathematical meaning or proper order.

Bar chart: Bar charts use rectangular bars to plot qualitative data against its quantity.

Pie chart: Pie charts are circular graphs in which various slices have different arc lengths depending on its quantity.

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

Solution – Many machine learning algorithms fail if the dataset contains missing values. You may end up building a biased machine learning model, leading to incorrect results if the missing values are not handled properly.

To rectify the above issue either we can delete the entire data having missing values or impute it with another replaced value.

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

Solution – When dealing with missing data, data scientists can use two primary methods to solve the error: imputation or the removal of data.

The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.

Deletion

There are two primary methods for deleting data when dealing with missing data: listwise and dropping variables.

Listwise - In this method, all data for an observation that has one or more missing values are deleted. The analysis is run only on observations that have a complete set of data.

Pairwise - Pairwise deletion assumes data are missing completely at random but all the cases with data, even those with missing data, are used in the analysis.

Dropping Variables - If data is missing for more than 60% of the it may be wise to discard it if the variable is insignificant**.**

Imputation

Instead of deletion, data scientists have multiple solutions to impute the value of missing data.

Mean, Median and Mode - This is one of the most common methods of imputing values when dealing with missing data. In cases where there are a small number of missing observations.

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

Solution - In an ideal world, your dataset would be perfect and without any problems. Unfortunately, real-world data will always present some issues that you'll need to address.

Various data pre-processing techniques are Data Cleaning, Dimensionality Reduction, Feature Engineering, Sampling Data, Data Transformation, Imbalanced Data

9.

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

Solution - The interquartile range (IQR) measures the spread of the middle half of your data. It is the range for the middle 50% of your sample.

We can find the interquartile range or IQR in four simple steps:

- Order the data from least to greatest
- Find the median
- Calculate the median of both the lower and upper half of the data
- The IQR is the difference between the upper and lower medians

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?**

Solution - 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.

A variation of the box and whisker plot restricts the length of the whiskers to a maximum of 1.5 times the interquartile range. That is, the whisker reaches the value that is the furthest from the centre while still being inside 1.5 times the interquartile range from the lower or upper quartile

In a boxplot, the IQR is the box portion between the first and the third quartile. IQR can be used to calculate the lower and upper bounds of the data, which helps identify outliers.

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

1. **Data collected at regular intervals**

Solution - Interval recording documents whether a behaviour occurred during a particular period. To determine this, an observation period is divided into brief intervals. At the end of each of these, the observer records whether a behaviour has occurred

2. **The gap between the quartiles**

Solution - The interquartile range is the difference between upper and lower quartiles. The semi-interquartile range is half the interquartile range.

3. **Use a crosstab**

Solution - Cross tabulation (crosstab) is a useful analysis tool commonly used to compare the results for one or more variables with the results of another variable. It is used with data on a nominal scale, where variables are named or labelled with no specific order.

1. Make a comparison between:

1. **Data with nominal and ordinal values**

Solution - The main differences between Nominal Data and Ordinal Data are: While Nominal Data is classified without any intrinsic ordering or rank, Ordinal Data has some predetermined or natural order. Nominal data is qualitative or categorical data, while Ordinal data is considered “in-between” qualitative and quantitative data.

2. **Histogram and box plot**

Solution - 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.

3. **The average and median**

Solution - 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.