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

-There are few regularly utilised tasks involved while working with machine learning modelling –

task 1: Collecting Data

task 2: Preparing the data ie analysis

task 3: Choose the model or algorithm

task 4 Train your machine model

task 5: Evaluation of that trained model

task 6: Parameter Tuning for better accuracy if required

task 7: Predicting the outcome.

After collecting the data from various options like a database or web scraping we basically go for analysing data and while analyzing the pre-processing comes into the picture so in this data preprocessing we clean the data we try to remove the missing values we try to remove the outliers we check for any kind of correlation between features as well as we analyze the data set so that we can extract more insights from the data. Data processing is all about processing the data before sending it to a model for training in such a way that the model should give a better accuracy on the given data set as well as the data set should be a well-structured and balanced the format

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

- Quantitative data consists of data that is present in the form of numbers numerics etc quantitative data consists of two types ie distinct quantitative data and continuous quantitative data . Examples are the height of a person, the weight of a person, etc.

Qualitative data of consists of data that is present in the form of a category the categorical data are nothing but called qualitative data these are present in the form of classes or groups for example gender, classes of flowers.

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

- There are few of data related issues which could be highlighted such as :-

1.Lack of quality of data – The data either is not well structured and balanced and it requires little bit of work to achieve in a good format.

2. Huge amount of data – when it comes to huge number of data , ML algorithms fail to execute as they take huge amount of time and accuracy would be later on stage.

3. Data present in another forms – Data present in videos or images format could be treated well with Deep learning but ML algorithm would not be helpful in this cause.

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

- There are many approaches which we can follow for exploring categorical data’s few of them are:-

1. Unique value count :- This helps us to identify how many unique values of categories or data are present in data set.

2. frequency count :- This helps us to identify frequency of data occurred in the dataset.

3. Correlation between two specific categorical columns

4. Bar charts or pie charts could be helpful in exploring categorical data.

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

- If there are missing values in the data set then many of the machine learning algorithms fail to give a proper result for a better model there are few algorithms that could handle by themselves like k-nearest neighbors but many machine learning models fail to give a better accuracy If the data set contains missing values in it so it's very important to handle missing values . There are a few ways that come across after examining the missing values

1. the data would not give a proper insight ie proper statistical Insight

2. it could provide Less number of samples for analyzing the data as well as the computation would be totally wrong if there are missing values in the data set.

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

- There are many methods to handle the missing values present in the data set and a few of them are:-

1. Imputation method -

Mean - Using the mean method we can handle these missing values that is we can replace all the missing values with the average of the feature for all the values present in that feature.

Mode - Mode is a type of method where we can replace all the missing values with the most frequent values that occurred in that feature.

2. Dropping of the missing values - In a few of the cases if the missing values are too less in the column or feature then we can directly delete that entire feature.

3. We can replace these missing values with random labels or values which can be used across the feature

4. If the feature contains category variables then the missing values can be replaced with the help of the most frequently occurring category values.

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

- Before passing the data set to the model for training purposes it is very important to preprocess our data and here are a few of the steps which include the preprocessing techniques in them.

Data Cleaning - Cleaning of the data that is making the data structured and balanced form is very important is a part of preprocessing technique. It usually consists of labeling the names with the columns or indexes removing the unwanted characters or any other not required features if present.

Dimensionality Reduction. - Reducing the data set to a particular dimension is must most important method of preprocessing this could be done with the help of PCA which is Principal component analysis.

Feature Engineering. - Feature engineering is nothing but an analysis of data where we can take out all the insights present in that data.

Sampling Data. - sampling is the practice of analyzing a subset of all data in order to uncover meaningful information in the larger data set.

Data Transformation. - The obtained data would not be in the same or a normally distributed form which would contain a lot of noise in it so smoothening the data is the most important preprocessing technique. here the data is transformed to a normal distribution with the help of smoothening process or it could be converted to any other form which suits the model.

Imbalanced Data - Handling the imbalance data is nothing but called as handling the missing values which are present in the data set there are various methods of handling the missing values it could be done with the help of the Imputation method or it could be with the help of replacing the Nan values with some other variables or it could be with the deletion of features.

**Feature selection vs. dimensionality reduction**

- Both of the methods can be used to get or extract a proper data which could be provided to the model. In Feature selection , we drop the irrelevant features from the dataset and select only those which are required ie subset of dataset for model training but in case of dimensionality reduction we are transforming the data or features to the lower dimension. PCA ie Principle of component analysis could be used to reduce the dimensionality of the data.

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

- IQR stands for interquartile range so the difference between upper quartile and lower quarter is nothing but called IQR.  the range defines our maximum data present inside that quarter range. As per the notation format, the upper quartile is Q3 and the lower quartile is Q1. Larger values indicate that the central portion of your data spread out further.

Criteria :-

Inter quartile range that is IQR can be accessed with the help of a boxplot ie, especially in cases where the data set consists of outliers. The data set can be plotted with the help of a box plot with a lower fence and high fence including the Ranges and specifying the outliers apart from the given range so that we can take out the interquartile range off the data and the difference between the upper quartile range and the lower quartile range is nothing but called as IQR.

IQR = Q3 - Q1

For example lets consider a list of values ranging from (for even sample data)

34, 56, 67, 67,88,43,23,68,78,64

Here we can divide them into two halfs

First half - 34, 56, 67, 67,88

Second half - 43,23,68,78,64

So , In first half middle value is 67 = Q1 and in second half middle value ie 68 = Q3

Difference between these values is nothing but 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?**

- To find the outliers or to plot the graph which includes range in it we use boxplot.

Box plot is plotted with the help of ranges. The first range is called as first quartile or lower quartile and is denoted by Q1. The second range is called as second quartile or upper quartile and is denoted by Q3. Here Q1 is 25th percentile and Q3 is 75th percentile of the data.

The Difference between these two values is called as IQR ie Inter quartile range.

The lower whisker surpass the upper whisker in length when Each whisker extends to the furthest data point in each wing that is within 1.5 times the IQR. Any data point further than that distance is considered an outlier, and is marked with a dot.

The outliers can be figured out with both minimum and maximum value of data in boxplot.

First minimum = Q1-(1.5 . IQR)

Maximum = Q3+(1.5 . IQR)

Any values present after or before these values are considered as outliers.

**9. Make brief notes on any two of the following:**

**1. Data collected at regular intervals**

**-** Interval data is a type of data which is measured along a scale, in which each point is placed at an equal distance (interval) from one another. Interval data is one of the two types of discrete data. An example of interval data is the data collected on a thermometer

**2. The gap between the quartiles**

**-**The gap between quartiles is called as IQR ie Interquartile range as well as the Difference between these two values is called as IQR ie Inter quartile range.. The first range is called as first quartile or lower quartile and is denoted by Q1. The second range is called as second quartile or upper quartile and is denoted by Q3. Here Q1 is 25th percentile and Q3 is 75th percentile of the data.

**3. Use a cross-tab**

**10. Make a comparison between:**

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

Nominal data -These categories have corresponding numbers allotted for analysis of collected data. For example, a person’s gender, ethnicity, hair color etc. are considered to be data for a nominal scale.

Ordinal data – This involves arranging information in a specific order, i.e. in comparison to one another and “rank” each parameter (variable). For example, after a customer shops from a retail outlet, he/she is asked to fill out a kiosk survey: “On a scale of 1-5, how was your shopping experience?”

**2. Histogram and box plot**

- Histogram - A Graphical technique which can be used to display the data in the form bars of different heights which include bins in them. The data can be displayed in such a manner that each bar consists of a range of values in it. So, it the bar height is more then the bar consists of more values in it. A histogram displays the shape and spread of continuous sample data.

A histogram when plotted do look like few of shapes listed below:-

a. Bell shaped Histogram – a kind of plot which is bell in shape and basically looks like normal distribution.

b. Skewed right or skewed left – A plot which is skewed either from right side or from left side. So basically most the data range values are present either at right side or left side.

c. Uniform – A Plot where data is distributed uniformly

d. Random - A plot where data is distributed randomnly.

Bins – Bins are nothing but called as intervals or classes present in the range of data. A histogram displays numerical data by grouping data into "bins" of equal width. Each bin is plotted as a bar whose height corresponds to how many data points are in that bin.

Boxplot - Box plot is plotted with the help of ranges. The first range is called as first quartile or lower quartile and is denoted by Q1. The second range is called as second quartile or upper quartile and is denoted by Q3. Here Q1 is 25th percentile and Q3 is 75th percentile of the data.

The Difference between these two values is called as IQR ie Inter quartile range.

The lower whisker surpass the upper whisker in length when Each whisker extends to the furthest data point in each wing that is within 1.5 times the IQR. Any data point further than that distance is considered an outlier, and is marked with a dot.

**3. The average and median**

- Average – The average is defined as the mean value which is equal to the ratio of the sum of the number of a given set of values to the total number of values present in the set.

Ex – 1,2,3,4,5

Avergare of above values = 1+2+3+4+5 = 3

5

Median - The median is the middle number in an ordered data set.

Ex **-** the median of 3, 5, 7, 9

(5+7)/2 = 6.