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

A - Key tasks in Machine Learning are as follows.

1. Data gathering

2. Data Pre-processing

3. Exploratory Data Analysis

4. Model selection

5. Training Machine Learning Models

Regression

Classification

Clustering

6. Testing

7. Monitoring

8. Model Retraining

Data Pre-processing:

Steps involved in Data Pre-processing are-

1. Data Cleansing:

A .Handle Missing Data

1. Outlier detection
2. Data Transformation:
3. Normalisation
4. Feature selection
5. Data reduction:
6. Dimensionality Reduction(PCA)

Question2-difference between qualitative and quantative data

**Quantitative data** refers to any information that can be quantified — that is, numbers. If it can be counted or measured, and given a numerical value, it's quantitative in nature. Think of it as a measuring stick.

Quantitative variables can tell you "how many," "how much," or "how often."

**Qualitative data** analysis describes information and cannot be measured or counted. It refers to the words or labels used to describe certain characteristics or traits.

You would turn to qualitative data to answer the "why?" or "how?" questions.

Question4-What are the various causes of machine learning data issues? What are the ramifications**?**

1. Poor Quality of Data
2. Under fitting of Training Data: when data is too small for model training( we try to fit under fiting jeans).very less accuracy on both training and testing datasets

To overcome under fitting issue:

* 1. Maximise training time
  2. Enhance model complexity
  3. Add more data features
  4. Reduce regular parameters

1. Over fitting of training Data- 100% accuracy on training data but very less on testing data.

To overcome over fitting issue:

* 1. Analyse data to utmost perfection
  2. Use **data augmentation** technique
  3. Remove outlier
  4. Select model with lesser features

1. Machine Learning is a very complex process with high chances of error if data is not analysed properly
2. Lack of training data
3. Slow implementation
4. Algo changes as data grows

Question5. Demonstrate various approaches to categorical data exploration with appropriate examples.

Answers- categorical Data- for example gender, blood group etc

Data analysis can be done by mode or drawing graph which helps to identify maximum occurences of a single record.

Question6. How would the learning activity be affected **if** certain variables have missing values? Having said that, what can be done about it**?**

Answer – if missing values are not handled properly then we might end up building biased model with incorrect predictions

Ways to handle missing values

1. Drop data
2. Data imputation using mean ,median,mode
3. Data imputation with KNN imputer

question8. What are the various data pre**-**processing techniques? Explain dimensionality reduction **and** function selection **in** a few words**.**

Answer- data pre-processing techniques has been explained above

Dimensionality reduction is a technique that is used to reduce features of data without losing important information.

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

IQR is interquartile range it is used to detect outlier , any datapoint falling beyond this range is outlier.

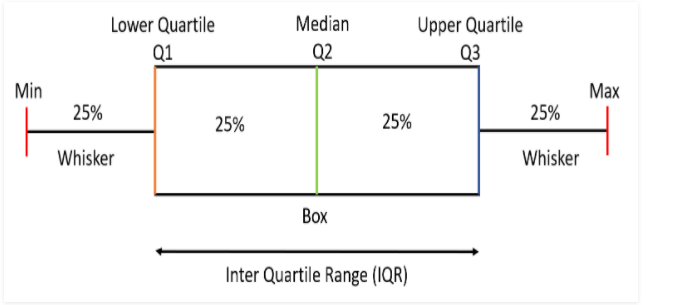
Formulae is given by-q3-q1

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 outlier

Box Plot: It is a type of chart that depicts a group of numerical data through their quartiles. It is a simple way to visualize the shape of our data

Components of boxplot-



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

11. 1. Make a comparison between:

1. Data with nominal and ordinal values

Ordinal values have order like first second third, nominal data doesnot – like male ,female or bloodgroup

2. Histogram and box plot- histogram gives max occurrence of data point whereas boxplot helps in outlier detection.

3. The average and median-

Median is the middle value

Average is simple sum/number