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

Key task for getting ready to work with machine learning modelling are:

* Data collection: Collecting data from various sources like api, database, local, etc.
* Data preprocessing: Data should be cleaned. Removing nan values, handling missing values, etc.
* Feature selection and engineering: Select relevant features
* Splitting data: Splitting data into Training set, testing set, and validation/development set.
* Model selection: Selecting the appropriate machine learning algorithm.
* Model training: Training the model
* Model evaluation: Evaluating the model
* Hyperparameter tuning: Hypertune the parameter to achieve the goal
* Model Deployment: Deploy the model
* Maintenance: Maintain the model by continuously monitoring it.

Data preprocessing refers to cleaning, transforming, and preparing the raw data before feeding to the machine learning model.

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

Quantitative data refers to numerical data, data that can be measured or counted. Quantities, amount, size, etc. Quantitative data is often analyzed using statistical techniques such as mean, median, mode. Quantitative data can be further divided into two

* Discrete data: Specific data or whole number, example students in the class, total bottles
* Continuous data: Any values within range or interval, for example temperature, height, decimal values.

Qualitative data refers to non-numerical data, data that can not be measured numerically. It deals with qualities, properties or attributes, etc. Data that is collected through interviews, observations, surveys, etc.

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

Record 1:

Name: John Smith

Age: 20

Gender: Male

GPA: 3.5

Major: Computer Science

Graduation Year: 2023

Record 2:

Name: Jane Doe

Age: 22

Gender: Female

GPA: 3.8

Major: Mathematics

Graduation Year: 2022

Record 3:

Name: David Johnson

Age: 19

Gender: Male

GPA: 3.2

Major: Psychology

Graduation Year: 2024

Record 4:

Name: Sarah Williams

Age: 21

Gender: Female

GPA: 3.6

Major: Business Administration

Graduation Year: 2023

Name is qualitative data type, age is quantitative(Numerical) data type, gender is qualitative(categorical) type, GPA is quantitative(Numerical) type, major is Qualitative(categorical) type, Graduation is Quantitative(Numerical) type

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

Various cause and ramifications are:

* Insufficient data- Reduce accuracy
* Biased data- Biased results
* Missing data- Poor generalizations
* Noisy data-Increases errors
* Imbalance data- Causes erros

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

Various approaches of categorical data exploration are:

* Frequency distribution- Calculating frequency of each category, use barplot, pie chart for visualization. Example-Customer preview of a product with a categorical variable with ratings excellent, good, average and poor, we can calculate the frequency of each ratings
* Cross-Tabulation- Analyze relationships between two categorical variable by creating cross-tabulation. Example: Consider students' performance in a class, with two categorical variables "Gender" and "Grade" (A, B, C). We can create a cross-tabulation table to examine the distribution of grades across different genders and identify any patterns.
* Chi-Square-Determine the association or independence between two categorical variable using the chi-square test of independence. This test calculates the expected and observed frequencies and compares them. Example: Survey data with one categorical variable "Occupation" (Engineer, Teacher, Doctor) and another categorical variable is "Job Satisfaction" (Satisfied, Neutral, Dissatisfied). We can perform a chi-square test to determine if there is a significant association between occupation and job satisfaction.

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

It can cause bias in analysis, data loss, and model performance would fail. To handle missing values we can impute data (mean, median, mode, etc any other summary statistics), regression based imputation

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

Various method to handle missing values are:

* Deletion method
* Mean/median/mode imputation
* Regression-based imputation- Predicting missing values using regression models.
* K-NN imputation- estimating the missing values based on the values of the nearest neighbors.
* Advance machine learning techniques- Models such as random forest, gradient boosting, deep learning model can be utilized to impute missing values.

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

Various data pre-processing techniques are:

* Data cleaning
* Data intefration
* Data transformation
* Feature extraction
* Data encoding
* Data splitting

Dimensionality Reduction is technique used to reduce high-dimension data to low-dimension data, while preserving the most relevant information

Feature Selection is the process of selecting a subset of the most relevant features from the original features. It aims to improve the accuracy of the model by reducing noise, preventing overfitting.

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

IQR stands for interquartile range, It is a statistical measure used the spread or dispersion of data. It is calculated as the difference between the quartile 3 and quartile 1. Criteria used to assess it are:

* Range should be central 50%
* Outliers
* Skewness- IQR can provide skewnesss of the data
* Box plot- IQR can be used to visualize using box blot

**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?**

Components of box plot are:

* Minimum: The lowest data point within 1.5 times the interquartile range (IQR) below the first quartile (Q1).
* First quartile (Q1): The median of the lower half of the dataset.
* Median: The middle value of the dataset.
* Third quartile (Q3): The median of the upper half of the dataset.
* Maximum: The highest data point within 1.5 times the IQR above the third quartile (Q3).
* Interquartile range (IQR): The range between the first quartile (Q1) and the third quartile (Q3). It represents the spread of the middle 50% of the data.
* Whiskers: The lines that extend from the box to the minimum and maximum values that are not considered outliers.
* Outliers: Data points that fall below the lower whisker or above the upper whisker.

The length of the lower whisker will surpass the upper whisker when the lower quartile (Q1) is less than the negative difference between the third quartile (Q3) and the median. This indicates a skewed distribution with a longer tail on the lower end.

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

* **Data collected at regular intervals**

Data collected at regular intervals refers to the collection of data points at consistent and predefined time intervals. This type of data collection is commonly used in various domains such as finance, weather monitoring, and stock market analysis. It allows for the analysis of patterns, trends, and changes over time. The regular intervals ensure a standardized and structured approach to data collection, enabling comparisons and the application of time series analysis techniques. Examples include daily temperature readings or hourly stock market prices.

* **The gap between the quartiles**

The gap between the quartiles in a dataset is known as the interquartile range (IQR). It is a measure of the spread or dispersion of the data. The IQR is calculated as the difference between the third quartile (Q3) and the first quartile (Q1). It represents the range within which the middle 50% of the data lies. A larger IQR indicates a greater spread of data points, while a smaller IQR suggests a more clustered distribution. The IQR is commonly used in box plots to identify outliers based on the "1.5 times the IQR" rule.

* **Use a cross-tab**

A cross-tab, also known as a contingency table, is a tabular representation of the relationship between two categorical variables. It displays the frequency distribution of the variables and allows for the comparison of their values. Cross-tabs are useful for exploring the association between variables and identifying any pattern. They provide a clear visual representation of how the variables are related and can help in making data-driven decisions. For example, a cross-tab between gender and product preference can reveal if there is a difference in preference between males and females.