1. What are the key tasks involved in getting ready to work with machine learning modelling?

Data Collection: Gather relevant data for the problem at hand.

Data Cleaning: Handle missing values, outliers, and inconsistencies.

Data Exploration: Understand data distributions, patterns, and relationships.

Feature Engineering: Create new features or transform existing ones.

Data Splitting: Divide data into training and testing sets.

Scaling/Normalization: Standardize numerical features if needed.

Model Selection: Choose an appropriate machine learning model.

Model Training: Train the selected model on the training data.

Model Evaluation: Assess the model's performance on test data.

Hyperparameter Tuning: Optimize model parameters for better performance.

Deployment: Implement the model for real-world use.

2. What are the different forms of data used in machine learning? Give a specific example for each of them.

Numeric Data: Involves numerical values (e.g., age, salary).

Categorical Data: Consists of discrete categories (e.g., colour, gender).

Text Data: Involves textual information (e.g., reviews, tweets).

Image Data: Comprises pixel values representing images.

Time Series Data: Sequential data over time (e.g., stock prices).

3. Distinguish:

1. Numeric vs. categorical attributes

Numeric Attributes: Represented by measurable quantities (e.g., height).

Categorical Attributes: Represented by categories or labels (e.g., color).

1. Feature selection vs. dimensionality reduction

Numeric Attributes: Represented by measurable quantities (e.g., height).

Categorical Attributes: Represented by categories or labels (e.g., color).

4. Make quick notes on any two of the following:

1. The histogram

A graphical representation of the distribution of a dataset.

Displays the frequencies of values in intervals (bins).

1. Use a scatter plot

Depicts individual data points on a two-dimensional plane.

Useful for exploring relationships between two variables.

3.PCA (Personal Computer Aid)

5. Why is it necessary to investigate data? Is there a discrepancy in how qualitative and quantitative data are explored?

Necessity: Understand data patterns, identify outliers, and make informed modelling decisions.

Discrepancy: Exploration may differ for qualitative and quantitative data due to their nature.

6. What are the various histogram shapes? What exactly are ‘bins'?

Histogram Shapes:

Normal Distribution: Bell-shaped, symmetric.

Skewed Distribution: Asymmetric, longer tail on one side.

Uniform Distribution: Even distribution of values.

Bins:

Intervals used in a histogram to group data points.

7. How do we deal with data outliers?

Identification: Use statistical methods or visualization techniques.

Treatment: Options include removal, transformation, or imputation.

8. What are the various central inclination measures? Why does mean vary too much from median in certain data sets?

Mean: Sensitive to outliers, affected by extreme values.

Median: Less affected by outliers, represents the central value.

9. Describe how a scatter plot can be used to investigate bivariate relationships. Is it possible to find outliers using a scatter plot?

Mean: Sensitive to outliers, affected by extreme values.

Median: Less affected by outliers, represents the central value.

10. Describe how cross-tabs can be used to figure out how two variables are related.

Usage: Examine the relationship between two categorical variables.

Output: Provides counts or percentages in a table format.