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

The main tasks that need to acquired before approaching with machine learning modelling is: 1) The problem statement, 2)What is the need of using ML in the dataset, 3) How to apply the ML algorithm.

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

Different forms of data used in Machine Learning are:

i) Structured Data: Structured data refers to data that is organized and arranged in a well-defined format, typically in tabular form with rows and columns.

ii) Unstructured Data: Unstructured data refers to data that does not have a well-defined format or organization. It includes text documents, emails, images, audio recordings, video files, social media posts, and more.

iii) Semi-structured Data: Semi-structured data falls between structured and unstructured data. It contains some organizational structure or metadata that provides limited schema or tags. Examples of semi-structured data include XML files, JSON, etc.

iv) Time Series Data: Time series data consists of observations collected over time, where each data point is associated with a specific timestamp.

v) Image and Video Data: Image and video data involve visual information in the form of pixels. Image data consists of two-dimensional arrays of pixel values, while video data is a sequence of images over time.

vi) Textual Data: Textual data involves written or natural language text. It can be in the form of documents, articles, social media posts, customer reviews, or any other textual content.

vii) Graph Data: Graph data represents entities and their relationships using nodes and edges. It is used to model complex relationships and networks. Examples include social networks, web graphs, biological networks, and recommendation systems.

3. Distinguish:

1. Numeric vs. categorical attributes

Numeric data are the types of data that have continuous value or the data-type can be of integer or float data type.

Categorical attributes: these attributes are used to classify a particular discrete data. The discrete data may be nominal, binomial or ordinal kind of data. Basically we get these datatype in the form of objects. These data types are encoded to numerical data types for better understanding and result.

2. Feature selection vs. dimensionality reduction

Feature selection is a kind of technique where certain criterion is used or analyzed to arrive at some conclusion where some of the independent feature are not important to the problem or are heavily correlated with other. This technique is helpful in controlling over fitting of data. Exampe: L1 (Lasso) regularisation.

Whereas dimensionality reduction is a technique where we reduce the number of input features or variables while preserving the most important information. It aims to simplify complex datasets, improve computational efficiency, and mitigate the curse of dimensionality. Example: PCA (Principal Component Analysis)

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

1. The histogram: Histograms provide a visual summary of the underlying data distribution, allowing us to understand the frequency or density of values within specific ranges.

2. Use a scatter plot: A scatter plot is a type of data visualization that displays the relationship between two continuous variables. Scatter plots are useful for identifying patterns, trends, and correlations between 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?

Investigating data is essential because it allows us to gain insights into the characteristics of the data, identify patterns and relationships, detect outliers or errors, and ultimately make informed decisions based on the information provided by the data. Investigating data can help us uncover hidden trends, explain phenomena, and test hypotheses, and can be critical to ensuring the validity and reliability of any conclusions or decisions based on the data. While there may be some differences in how qualitative and quantitative data are explored, both require careful examination and analysis to gain meaningful insights and draw accurate conclusions.

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

Various histogram shapes are: Normal Distribution, Skewed Distribution, Bimodal Distribution, Multimodal Distribution, Uniform Distribution and Exponential Distribution

7. How do we deal with data outliers?

Data outliers are either removed or capping with similar data by analysing the feature and dataset. In some cases, for numerical data we perform or cap the outliers by medians where as for categorical we use mode. Another approach is to transform the data using mathematical functions.

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

The mean ,median and mode are central inclination measures that represent different aspects of the data. The mean is influenced by all values and is sensitive to outliers, resulting in potential variations from the median in datasets with extreme values or skewed distributions. The median, being less affected by extreme values and robust to outliers, provides a more resistant measure of central tendency in such cases. The choice between mean and median should be made based on the characteristics of the dataset and the analysis objectives.

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

scatter plots provide a visual representation of the relationship between two continuous variables. They allow for the assessment of the relationship type, correlation strength, identification of outliers, detection of clusters or subgroups, and the potential for further analysis and modeling. The visual nature of scatter plots facilitates a comprehensive understanding of the bivariate relationship in the data.

Scatter plots can help identify outliers, which are data points that deviate significantly from the general pattern or trend of the other points. Outliers can be detected as individual points that are noticeably distant from the main cluster or trend of the data points. By visually examining the scatter plot, we can identify observations that appear unusually high or low compared to the majority of the data.

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

Cross-tabs provide a comprehensive view of the relationship between two categorical variables by organizing the data into a table format. They allow for the assessment of association, the calculation of proportions or percentages, statistical testing, and visualizations. Cross-tabs are a useful tool for exploring and understanding the relationship between categorical variables and identifying patterns or dependencies within the data.