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

A1. Before starting with machine learning modeling, there are several key tasks that need to be undertaken. Some of these tasks include:

1. Data Collection: Collecting and gathering relevant data from various sources is the first and most crucial step in machine learning. Data should be relevant to the problem being solved and should be in a format that can be easily used by the machine learning algorithm.
2. Data Cleaning and Pre-processing: Raw data often contains errors, noise, or missing values that can affect the machine learning model's accuracy. Data cleaning and pre-processing involve identifying and correcting these errors, removing irrelevant or redundant data, and transforming the data into a format suitable for machine learning.
3. Data Exploration and Visualization: Exploring the data can help to identify patterns, relationships, and trends that can be used to create a more accurate machine learning model. Data visualization tools can be used to create graphs, charts, and other visual representations of the data.
4. Feature Engineering: Feature engineering is the process of selecting and transforming relevant variables or features in the data that can help the machine learning model to learn and make accurate predictions.
5. Model Selection: There are various machine learning algorithms available, and selecting the right algorithm is crucial to achieving the best results. The selection of the model should be based on the problem statement, the size of the dataset, and the type of data.
6. Model Training: Model training involves feeding the algorithm with the dataset and iteratively adjusting the model parameters until the algorithm's performance is satisfactory.
7. Model Evaluation: Model evaluation involves measuring the performance of the machine learning model using various metrics such as accuracy, precision, recall, F1 score, and AUC-ROC.
8. Model Optimization: After evaluating the model, it may be necessary to optimize it further by tuning hyperparameters, adjusting the model architecture, or changing the training data.

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

A2. There are several forms of data used in machine learning, including:

1. Numerical data: This type of data is represented by numbers and can be either continuous or discrete. An example of continuous numerical data is the temperature of a city on a given day, while an example of discrete numerical data is the number of people in a movie theater.
2. Categorical data: This type of data consists of categories or labels and can be either nominal or ordinal. Nominal categorical data has no intrinsic order, such as the colors of a car, while ordinal categorical data has an inherent order, such as the grades of a student (A, B, C, D, F).
3. Text data: This type of data consists of text documents, such as emails, articles, and social media posts. Text data is often processed through natural language processing techniques to extract meaningful information.
4. Image data: This type of data consists of images or videos. Image data is often processed through computer vision techniques to extract useful features.
5. Time series data: This type of data consists of a sequence of observations recorded over time, such as stock prices, weather patterns, or sensor readings.

3. Distinguish:

1. Numeric vs. categorical attributes -

Numeric attributes are variables that contain numerical data, such as age, weight, height, and temperature. These variables are typically continuous or discrete and can be measured using various units of measurement.

Categorical attributes, on the other hand, are variables that contain data in a categorical or nominal form, such as color, gender, or occupation. These variables are usually represented as labels or codes and do not have a specific numerical value.

2. Feature selection vs. dimensionality reduction

Feature selection involves selecting a subset of the original features in the dataset that are most relevant to the task at hand. This is done to reduce the complexity of the model, improve accuracy, and speed up the learning process. Examples of feature selection techniques include correlation-based feature selection, mutual information-based feature selection, and recursive feature elimination.

Dimensionality reduction, on the other hand, involves transforming the original features into a lower-dimensional space while preserving the most important information. This is done to reduce the complexity of the model, improve interpretability, and prevent overfitting. Examples of dimensionality reduction techniques include principal component analysis (PCA), linear discriminant analysis (LDA), and t-distributed stochastic neighbor embedding (t-SNE).

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

1. The histogram - The histogram is a graphical representation of the distribution of data. It is a type of bar chart that divides a range of continuous values into a series of intervals, or "bins," and displays the frequency of the data falling into each bin. The x-axis represents the range of values, and the y-axis represents the frequency of the data falling into each bin. Histograms are commonly used in data analysis and visualization to understand the distribution and frequency of the data.

2. Use a scatter plot - A scatter plot is a graphical representation of data points in two dimensions. It is used to visualize the relationship between two variables. Each data point is plotted as a point on the x-y plane, with the x-axis representing one variable and the y-axis representing the other. Scatter plots are commonly used in data analysis to identify trends and patterns in the data and to identify outliers.

3.PCA (Personal Computer Aid) - PCA (Principal Component Analysis) is a technique used to reduce the dimensionality of large data sets while retaining as much of the variability in the data as possible. PCA transforms the data into a new coordinate system that is aligned with the principal components of the data. The principal components are the directions in the data with the highest variance. PCA is commonly used in data analysis and machine learning to reduce the number of features in the data, which can help to improve model accuracy and reduce overfitting.

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

A5. It is essential to investigate data to gain insights and identify patterns, trends, and relationships between variables. This process helps in making informed decisions and predictions based on the available data. Exploring data also helps in detecting and correcting errors and anomalies that could affect the accuracy of the analysis.

There is a difference in how qualitative and quantitative data are explored. Qualitative data is typically explored through techniques such as content analysis, which involves examining the text or narrative data to identify common themes or patterns. On the other hand, quantitative data is explored through statistical analysis, which involves using mathematical and statistical methods to identify relationships and patterns between variables. While there are some similarities in the methods used to explore both types of data, the approach is generally different due to the nature of the data being analyzed.

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

A6. Histograms are a graphical representation of the distribution of data, and they are useful for understanding the underlying structure of data. A histogram depicts the number of data points that fall within certain ranges or "bins" of values. Bins are the intervals on the x-axis of a histogram that represent ranges of values.

There are several types of histogram shapes that are commonly observed:

1. Bell-shaped: This is the most common shape, and it resembles a bell curve. The data is symmetrically distributed around the mean, with fewer data points in the tails.
2. Skewed: In a skewed histogram, the data is not symmetrically distributed around the mean. If the tail of the distribution is longer on the left, the histogram is said to be left-skewed or negatively skewed. If the tail is longer on the right, it is said to be right-skewed or positively skewed.
3. Bimodal: A bimodal histogram has two peaks, indicating that there are two different groups or populations within the data.
4. Uniform: In a uniform histogram, the data is evenly distributed across all bins.

It is necessary to investigate data in order to identify patterns, trends, and relationships within the data. This helps to uncover insights and inform decision-making. The methods used to explore data may differ depending on whether the data is qualitative or quantitative, but the overall goal remains the same. Qualitative data may be explored through methods such as content analysis, while quantitative data may be explored through statistical analysis and data visualization.

7. How do we deal with data outliers?

A7. Outliers are data points that are significantly different from the rest of the data. They can be caused by measurement errors, data processing errors, or other anomalies. Dealing with outliers is important because they can skew the results of data analysis and machine learning models. Here are some ways to deal with outliers:

1. Remove the outliers: In some cases, the best way to deal with outliers is to simply remove them from the dataset. This approach works well if the outliers are clearly erroneous and do not represent any meaningful information.
2. Transform the data: Another approach is to transform the data by applying a mathematical function that reduces the effect of outliers. For example, taking the log of a variable can help to reduce the effect of extreme values.
3. Winsorization: Winsorization is a method that replaces the extreme values with less extreme values. For example, we can replace the top 5% of the data with the value at the 95th percentile.
4. Robust statistics: Robust statistics methods are designed to be less sensitive to outliers. For example, using the median instead of the mean to calculate central tendency can reduce the effect of outliers.

It is important to note that dealing with outliers requires a careful understanding of the data and the problem at hand. Also, the approach used to deal with outliers will depend on the specific application and the goals of the analysis.

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

A8. The various central inclination measures are mean, median, and mode. Mean is the average value of a data set, while the median is the middle value when the data is arranged in ascending or descending order. The mode is the value that occurs most frequently in the data set.

In certain data sets, the mean varies too much from the median because of the presence of outliers or extreme values. Outliers can significantly impact the mean value, pulling it towards their extreme values. In such cases, the median may be a more representative measure of central tendency. Additionally, the distribution of the data, whether it is skewed or symmetric, can also affect the difference between mean and median.

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

A9.   
A scatter plot is a type of data visualization that is used to investigate the relationship between two variables. It is created by plotting one variable on the x-axis and the other variable on the y-axis, and then placing a point on the graph for each observation in the data set. By examining the pattern of the points, we can identify whether there is a linear, nonlinear, or no relationship between the variables.

Scatter plots can also be used to identify outliers, which are observations that fall far away from the majority of the data points. Outliers can be identified as points that are located far from the main cluster of points, and they can have a significant impact on the slope and fit of a linear regression line. By visually inspecting a scatter plot, we can identify potential outliers and investigate them further to determine whether they are valid data points or errors in measurement.

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

A10. Cross-tabulation, also known as contingency table analysis, is a statistical method for exploring the relationship between two categorical variables. It is a method for summarizing categorical data using a table that shows the frequency counts of various combinations of two or more categorical variables.

Cross-tabulation is useful for identifying patterns or relationships between two variables. It allows you to compare the frequencies of one variable across the categories of another variable. This information can be presented in a variety of ways, such as using percentages or proportions to describe the relationship.

For example, let's say we want to investigate how gender is related to car preferences. We can create a cross-tabulation table that shows the frequency counts of males and females for each type of car. This table will help us to identify if there is a gender preference for a particular type of car.