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

Ans:- Step1: Collect Data. ...

* Step 2: Prepare the data. ...
* Step 3: Choose the model. ...
* Step 4 Train your machine model. ...
* Step 5: Evaluation. ...
* Step 6: Parameter Tuning. ...
* Step 7: Prediction or Inference

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

* Ans:- Most data can be categorized into 4 basic types from a Machine Learning perspective: **numerical data, categorical data, time-series data, and text**. Image recognition. Image recognition is a well-known and widespread example of machine learning in the real world. ...
* Speech recognition. Machine learning can translate speech into text. ...
* Medical diagnosis. ...
* Statistical arbitrage. ...
* Predictive analytics. ...
* Extraction.

3. Distinguish:

* 1. Numeric vs. categorical attributes

Ans:- A categorical variable is a category or type. For example, hair color is a categorical value or hometown is a categorical variable. A numerical variable is a variable where the measurement or number has a numerical meaning

* 1. Feature selection vs. dimensionality reduction

Ans:- The basic difference is that **PCA transforms features but feature selection selects features without transforming them**. PCA is a dimensionality reduction method but not feature selection method. They all are good for feature selection. Greed algorithm and rankers are also better

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

1. The histogram

Ans:- A histogram is a graphical representation that **organizes a group of data points into user-specified ranges**. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.

1. Use a scatter plot

Ans:- Scatter plots are used to **plot data points on a horizontal and a vertical axis in the attempt to show how much one variable is affected by another**. Each row in the data table is represented by a marker whose position depends on its values in the columns set on the X and Y axes.

3.PCA (Personal Computer Aid)

Ans:-PCA forms the basis of **multivariate data analysis** based on projection methods. The most important use of PCA is to represent a multivariate data table as smaller set of variables (summary indices) in order to observe trends, jumps, clusters and outliers

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

Ans:- If your data set is messy, building models will not help you to solve your problem. What will happen is “garbage in, garbage out.” In order to build a powerful machine learning algorithm. We need to explore and **understand** our data set before we define a predictive task and solve it . The quantitative and qualitative studies **do not measure the same**, Therefore, their findings cannot be reported in the same manner

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

Ans:- **Bell-shaped**: A bell-shaped picture, shown below, usually presents a normal distribution. Bimodal: A bimodal shape, shown below, has two peaks. If this shape occurs, the two sources should be separated and analyzed separately. Skewed right: Some histograms will show a skewed distribution to the right, as shown below 1 : **a box, frame, crib, or enclosed place used for** storage. 2 chiefly British : a can for trash or garbage : dustbin The woman retreated to a desk, dropping the leaflets in the bin as though she couldn't bear to hold them anymore

7. How do we deal with data outliers?

Ans:- When you decide to remove outliers, document the excluded data points and explain your reasoning. You must be **able to attribute a specific cause for removing outliers**. Another approach is to perform the analysis with and without these observations and discuss the differences

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

Ans:- The deviation from vertical, irrespective of compass direction, **expressed in degrees**. Inclination is measured initially with a pendulum mechanism, and confirmed with MWD accelerometers or gyroscopes.

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

Ans:- The scatter plot is a fundamental tool for looking at bivariate data. It **shows the important characteristics of the data** and can be used to decide what model may describe the relationship between the variables.

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

Ans:- Cross tabulation is a method to quantitatively analyze the relationship between multiple variables. Also known as contingency tables or cross tabs, cross tabulation groups variables to understand **the correlation between different variables**.