**Machine Learning Assignment 4**

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

Ans-) Key tasks involved in preparing to work with machine learning modeling include understanding the problem to be solved, collecting and cleaning the data, selecting appropriate features, choosing a machine learning model, and evaluating the model's performance.

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

them.

Ans-) The different forms of data used in machine learning include structured, unstructured, and semi-structured data. An example of structured data is a database of customer information, unstructured data could be text from social media, and semi-structured data could be a combination of both, such as email threads.

3. Distinguish:

1. Numeric vs. categorical attributes

Numeric attributes represent quantities or numerical values, while categorical attributes represent categories or labels. Numeric attributes can be further divided into discrete or continuous values.

2. Feature selection vs. dimensionality reduction

Feature selection involves choosing a subset of relevant features from a dataset, while dimensionality reduction involves reducing the number of features in a dataset by transforming them into a smaller set of representative features.

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

1. The histogram

A histogram is a graphical representation of the distribution of a dataset, where the x-axis represents the values and the y-axis represents the frequency of occurrence.

2. Use a scatter plot

A scatter plot is a graphical representation of the relationship between two variables, where the x-axis represents one variable and the y-axis represents another. Each point on the plot represents a single data point.

3.PCA (Principal Component Analysis)

PCA (Principal Component Analysis) is a statistical technique used to reduce the dimensionality of a large data set while retaining as much variance as possible. It involves transforming the original variables into a new set of variables called principal components, which are linearly uncorrelated and ordered by the amount of variance they capture. PCA is often used in data preprocessing to simplify subsequent analyses or visualizations. It can also be used for data compression, noise reduction, and pattern recognition. The acronym "PCA" does not stand for "Personal Computer Aid" - it is simply short for "Principal Component Analysis".

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

data are explored?

Ans-) It is necessary to investigate data to understand its characteristics and identify any anomalies, such as missing values or outliers. Qualitative and quantitative data are explored differently, with qualitative data often being analyzed through thematic analysis and quantitative data through statistical methods.

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

Ans-) The various histogram shapes include normal, skewed left, skewed right, bimodal, and uniform. Bins are intervals on the x-axis that divide the data into groups or ranges.

7. How do we deal with data outliers?

Ans-) Outliers can be dealt with by removing them from the dataset, replacing them with another value, or transforming the data.

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

certain data sets?

Ans-) The various central inclination measures include mean, median, and mode. Mean varies too much from the median in certain data sets because it is sensitive to outliers.

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

outliers using a scatter plot?

Ans-) A scatter plot can be used to investigate bivariate relationships by plotting the data points and observing any patterns or trends. Outliers can be identified by points that lie far away from the general pattern of the data.

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

Ans-) Cross-tabs can be used to figure out how two variables are related by creating a table that displays the frequency of occurrence of each combination of variables. The table can be analyzed to identify any patterns or trends in the data.