ASSIGNMENT - 4

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

Ans: Key Tasks in Preparing for Machine Learning Modeling:

Data Collection: Gathering relevant data from various sources.

Data Preprocessing: Cleaning data, handling missing values, normalization, etc.

Feature Engineering: Selecting, creating, or transforming features for modeling.

Model Selection: Choosing appropriate algorithms for the specific task.

Training: Feeding data into the model to learn patterns.

Evaluation: Assessing model performance and refining if needed.

Deployment: Implementing the model in real-world scenarios.

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

Ans: Different Forms of Data Used in Machine Learning:

Numerical Data: Examples include age, height, temperature (e.g., age of individuals in a dataset).

Categorical Data: Represents categories, like gender, color, or types of cars (e.g., Sedan, SUV).

Text Data: Includes reviews, tweets, or documents (e.g., product reviews in an e-commerce dataset).

Image Data: Pixel values representing images, such as images of handwritten digits in a digit recognition dataset.

Time-Series Data: Sequential data collected over time, like stock prices over a period.

3. Distinguish:

* Numeric vs. categorical attributes
* Feature selection vs. dimensionality reduction

Ans: Numeric vs. Categorical Attributes:

* Numeric attributes are measurable quantities (e.g., age, height).
* Categorical attributes represent discrete categories (e.g., gender, color).

Feature Selection vs. Dimensionality Reduction:

* Feature selection involves choosing relevant features for the model.
* Dimensionality reduction aims to reduce features while retaining important information.

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

* The histogram
* Use a scatter plot
* PCA (Personal Computer Aid)

Ans:

The histogram:

A histogram is a graphical representation of the distribution of numerical data. It consists of bars that show the frequency or count of data points falling within different intervals, known as bins.

Use a scatter plot:

A scatter plot is a graph that displays the relationship between two numerical variables or dimensions. Each data point represents an observation, with one variable on the x-axis and another on the y-axis.

PCA (Personal Computer Aid)

PCA is a dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional form while preserving the most important information.

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

Ans: Investigating data is crucial because it helps us make informed decisions, uncover patterns, identify trends, and understand various phenomena. It allows us to derive insights, validate hypotheses, and solve problems effectively.

Qualitative Data: This type of data deals with non-numeric information, such as opinions, observations, and descriptions. Exploring qualitative data often involves techniques like content analysis, thematic analysis, or narrative analysis. It focuses on understanding context, exploring meanings, and identifying patterns in non-numeric information. The exploration is more subjective, requiring interpretation and a deeper understanding of the data's context.

Quantitative Data: Quantitative data, on the other hand, deals with numerical information and lends itself well to statistical analysis. Exploring quantitative data involves techniques such as statistical tests, regression analysis, and data visualization. It's more objective in nature, focusing on numerical relationships, correlations, and trends within the data.

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

Ans: Histograms can take various shapes, including normal (bell-shaped), skewed (positively or negatively), uniform, and bimodal, among others. Histograms display the distribution of data by dividing it into intervals called "bins" along the x-axis.

Bins represent ranges of values, and the height of each bin corresponds to the frequency of data points falling within that range.

7. How do we deal with data outliers?

Ans: Outliers are data points significantly different from other observations in a dataset. Strategies to handle outliers include removing them if they're errors, transforming the data, using robust statistical measures (like median instead of mean), or conducting separate analyses with and without outliers to observe their impact.

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

Ans: Measures of central tendency include mean, median, and mode. Mean is sensitive to outliers because it takes into account every data point, while median is less affected by outliers because it's the middle value when the data is arranged in order. In certain datasets with extreme values or skewed distributions, the mean can vary significantly from the median.

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 displays the relationship between two variables by placing data points on a Cartesian plane. It's used to visualize patterns and relationships between variables. Outliers can sometimes be visually identified on a scatter plot as data points that lie far away from the general pattern or trend formed by other points.

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

Ans: Cross-tabulation (or cross-tabs) is a statistical tool used to examine the relationship between two categorical variables. It creates a table where the intersections show the frequency or distribution of data for each combination of categories in the variables. It helps identify patterns, dependencies, or correlations between the variables.