ASSIGNMENT - 5

1. What are the key tasks that machine learning entails? What does data pre-processing imply?

Ans: These include tasks like data collection, data cleaning, feature selection, model training, evaluation, and deployment. It involves algorithms and methodologies that allow systems to learn patterns and make predictions or decisions based on data.

This involves cleaning, transforming, and organizing raw data into a format suitable for analysis. It includes tasks like handling missing values, normalization, standardization, encoding categorical variables, and feature scaling.

2. Describe quantitative and qualitative data in depth. Make a distinction between the two.

Ans: Quantitative Data: This refers to numerical data that can be measured and expressed with numbers. Examples include height, weight, income, etc. It's analyzed using statistical methods.

Qualitative Data: Also known as categorical or non-numeric data, qualitative data describes qualities or characteristics. Examples include gender, color, or occupation. It's often descriptive and analyzed using techniques like content analysis or thematic analysis.

3. Create a basic data collection that includes some sample records. Have at least one attribute from each of the machine learning data types.

Ans:

| Name | Age | Gender | Height | Weight | Income |
| --- | --- | --- | --- | --- | --- |
| Alice | 25 | female | 165 | 60 | $50,000 |
| Bob | 30 | Male | 180 | 75 | $70,000 |
| Claire | 22 | Female | 160 | 65 | $45,000 |

4. What are the various causes of machine learning data issues? What are the ramifications?

Ans: Causes include incomplete data, noisy data, biased data, irrelevant features, and overfitting. Ramifications include inaccurate models, biased predictions, and poor generalization to new data.

5. Demonstrate various approaches to categorical data exploration with appropriate examples.

Ans: Exploring categorical data involves frequency distributions, bar charts, pie charts, and cross-tabulations. For example, creating a bar chart to visualize the distribution of genders in a dataset or a pie chart showing the proportion of different categories in a variable.

6. How would the learning activity be affected if certain variables have missing values? Having said that, what can be done about it?

Ans: Missing values can reduce the quality of models as they may lead to biased estimates or reduced predictive accuracy. It can also affect the performance of some machine learning algorithms.

7. Describe the various methods for dealing with missing data values in depth.

Ans: Handling missing data can involve techniques like imputation (replacing missing values with estimates like mean, median, or mode), deletion of rows or columns with missing values, or using advanced algorithms that handle missingness during training.

8. What are the various data pre-processing techniques? Explain dimensionality reduction and

function selection in a few words.

Ans: Data Pre-processing Techniques:

* Data pre-processing involves techniques like normalization, standardization, encoding categorical variables, and handling missing values.
* Dimensionality reduction involves reducing the number of features in a dataset, while feature selection involves choosing the most relevant features for modeling purposes. These techniques help improve model efficiency and reduce computational complexity.

9. i. What is the IQR? What criteria are used to assess it?

Ans: The IQR is a measure of statistical dispersion and represents the range between the first (Q1) and third (Q3) quartiles in a dataset. It is calculated as IQR = Q3 - Q1. The IQR provides insight into the spread or variability of the middle 50% of the data.

Criteria to Assess IQR:

* Variability: A larger IQR indicates greater variability within the middle 50% of the data.
* Outliers: IQR is used in identifying potential outliers in a dataset. Data points lying significantly below Q1 - 1.5 \* IQR or above Q3 + 1.5 \* IQR are often considered outliers.

ii. Describe the various components of a box plot in detail? When will the lower whisker

surpass the upper whisker in length? How can box plots be used to identify outliers?

Ans: Components of a Box Plot:

* Median (Q2): The line inside the box represents the median.
* Box: The box represents the IQR, with its lower edge at Q1 and upper edge at Q3.
* Whiskers: These extend from the box to the smallest/largest values within 1.5 \* IQR from the quartiles.
* Outliers: Individual points beyond the whiskers are shown as individual points on the plot and might be considered outliers.

The lower whisker surpassing the upper whisker in length can occur when the lower quartile (Q1) is close to the minimum value and there are numerous data points that fall above the upper quartile (Q3). This creates a longer lower whisker compared to the upper one.

Box plots can identify outliers by:

* Any data points beyond Q3 + 1.5 \* IQR or below Q1 - 1.5 \* IQR are considered potential outliers.
* Points lying outside the whiskers (the lines that extend from the box) are displayed individually on the plot and are often flagged as outliers.

10. Make brief notes on any two of the following:

1. Data collected at regular intervals

Ans: Data Collected at Regular Intervals:

* Time Series Data: Data collected at regular intervals, such as hourly, daily, monthly, etc., constitutes time series data.
* Characteristics: It exhibits patterns and trends over time, making it valuable for forecasting and trend analysis.
* Applications: Commonly found in financial markets, weather data, sales records, and stock prices.
* Analysis: Methods such as moving averages, trend analysis, and seasonality assessment are used to analyze and interpret this data.
* Challenges: Seasonal variations, trend identification, and handling missing values are essential considerations when working with this data type.

1. The gap between the quartiles

Ans: The Gap Between the Quartiles:

* Interquartile Range (IQR): The gap between the quartiles (Q3 - Q1) is known as the Interquartile Range.
* Significance: It represents the middle 50% of the data, giving a measure of the spread or dispersion.
* Robustness: The IQR is robust against outliers compared to the range (max - min) because it focuses on the central portion of the dataset.
* Outlier Detection: Data points lying significantly beyond 1.5 times the IQR from either Q1 or Q3 are often flagged as potential outliers.
* Analysis: Helps in understanding the variability within the central portion of the data distribution without being heavily influenced by extreme values.

1. Use a cross-tab

Ans: Use a Cross-Tab:

* Definition: Cross-tabulation, or cross-tab, is a statistical tool used to examine the relationship between two categorical variables.
* Tabular Representation: It creates a table that displays the frequency or distribution of data for each combination of categories in the variables.
* Application: Helps in exploring and understanding associations or dependencies between categorical variables.
* Example: For instance, in a survey dataset, a cross-tab can show the relationship between Gender (Male/Female) and Voting Preference (Candidate A/B/C) by displaying the count or percentage of males and females voting for each candidate.
* Insight Generation: Identifies patterns, trends, or correlations between categorical variables, aiding in decision-making or further analysis in various fields like marketing, sociology, or medicine.