Assessment 2(AIML)

Section A:Data Wrangling

1. b) Data cleaning and transformation

2. Categorical data is converted into numerical numerical data using techniques like label encoding or one-hot encoding. This is necessary because most machine learning algorithms work with numerical data. Converting categorical variables into numerical form enables data analysis and modeling.

3. Label Encoding assigns a unique numeric value to each category, while One-Hot Encoding creates a new binary column for each category. One-Hot Encoding is preferred when there is no ordinal relationship between the categories.

4. One commonly used method for detecting outliers is the Z-score or Standard Score method. This calculates the number of standard deviations a data point is away from the mean. Data points with Z-scores beyond a certain threshold (e.g., +/- 3) are considered outliers. Identifying outliers is important because they can significantly influence the results of data analysis and model performance.

5. In the Quantile Method, outliers are identified and replaced with the nearest non-outlier value. For example, values below the 25th percentile or above the 75th percentile could be replaced with the respective quantile values.

6. ABox Plot is a graphical representation that displays the distribution of a dataset, including the median, quartiles, and potential outliers. It aids in identifying outliers as they are plotted as individual points beyond the whiskers (which represent the upper and lower boundaries). Box Plots provide a quick visual assessment of the data's central tendency, spread, and the presence of outliers.

Section B: Regression Analysis

7. Linear Regression is employed when predicting a continuous target variable.

8. The two main types of regression are: a) Linear Regression: Predicts a continuous target variable based on a linear combination of independent variables. b) Logistic Regression: Predicts a binary or categorical target variable based on independent variables.

9. Simple Linear Regression is used when there is only one independent variable predicting the target variable. For example, predicting a house's sale price based on its square footage.

10. In Multi Linear Regression, there are typically two or more independent variables involved in predicting the target variable.

11. Polynomial Regression should be used when the relationship between the independent variable(s) and the target variable is non-linear. For example, predicting the price of a car based on its age, where the price may initially decrease slowly but then decrease more rapidly as the car gets older.

12.A higher degree polynomial in Polynomial Regression represents a more complex, non-linear relationship between the variables. As the degree increases, the model can capture more intricate patterns in the data, but it also becomes more prone to overfitting.

13. The key difference between Multi Linear Regression and Polynomial Regression is that Multi Linear Regression assumes a linear relationship between the independent variables and the target variable, while Polynomial Regression can model non-linear relationships by introducing polynomial terms.

14. Multi Multi Linear Regression is most appropriate when there is a linear relationship between the independent variables and the target variable, and the independent variables are not highly correlated with each other (multicollinearity).

15. The primary goal of regression analysis is to model the relationship between independent variable(s) and a target variable, allowing for the prediction of the target variable's value based on the values of the independent variable(s).

Submitted By

Malle Akhila

20HU1A4224

Chebrolu Engineering College