ASSIGNMENT:
DATA WRANGLING &
REGRESSION ANALYSIS

AI / ML TRAINING **DATE: 21/02/24**

SECTION -A: DATA WRANGLING:

1. What is the primary objective of data wrangling?

a) Data visualization

b) Data cleaning and transformation

c) Statistical analysis

d) Machine learning modelling

Ans) b) Data cleaning and transformation

2. Explain the technique used to convert categorical data into numerical data. How does it help in data analysis?

Ans)

Categorical data is converted into numerical data through techniques like One-Hot Encoding or Label Encoding. One-Hot Encoding creates binary columns for each category, while Label Encoding assigns a unique numerical label to each category. This conversion helps in incorporating categorical information into machine learning models that require numerical input.

3. How does LabelEncoding differ from OneHot Encoding?

Ans)

Label Encoding assigns a unique numerical label to each category, converting categorical data into ordinal numerical data. OneHotEncoding, on the other hand, creates binary columns for each category, representing the presence or absence of that category. The main difference lies in the representation of the data.

4. Describe a commonly used method for detecting outliers in a dataset. Why is it important to identify outliers?

Ans)

The commonly used method is Z-score. it is important for Identifying outliers is crucial as they can significantly impact statistical analyses and machine learning models. Outliers may distort results, affect model performance, and lead to inaccurate predictions.

5. Explain how outliers are handled using the Quantile Method.

Ans)

The Quantile Method involves defining a range (interquartile range) based on the dataset's quantiles. Data points outside this range are considered outliers and can be treated by removing them or applying transformations. This method provides a robust way to identify and handle outliers.

6. **Discuss the significance of a Box Plot in data analysis. How does it aid in identifying potential outliers?**

Ans)

A Box Plot visually represents the distribution of data, displaying quartiles and potential outliers. It consists of a box (interquartile range), median line, and "whiskers" extending to the most extreme data points within a certain range. Outliers are often identified as points beyond the whiskers, aiding in their visual detection.

SECTION B: REGRESSION ANALYSIS

7. What type of regression is employed when predicting a continuous target variable?

Ans)

Linear Regression :

Linear Regression is a statistical method and a fundamental machine learning algorithm used for modeling the relationship between a dependent variable (also known as the target or outcome) and one or more independent variables (predictors or features). The relationship is assumed to be linear, meaning it can be represented by a straight line.

8. **Identify and explain the two main types of regression.**

Ans)

a) Linear Regression: Used for predicting a continuous target variable based on one or more independent variables.

b) Logistic Regression: Used for binary classification problems, predicting the probability of an event occurring.

9. When would you use Simple Linear Regression? Provide an example scenario.

Ans)

Simple Linear Regression is used when there is a linear relationship between a single independent variable and the dependent variable.

Example: Predicting a student's exam score based on the number of hours spent studying.

10. In Multi Linear Regression, how many independent variables are typically involved?

Ans)

Multi Linear Regression involves predicting a target variable based on two or more independent variables.

11. When should Polynomial Regression be utilized? Provide a scenario where Polynomial Regression would be preferable over Simple Linear Regression.

Ans)

Polynomial Regression is suitable when the relationship between the independent and dependent variables is nonlinear. Example: Predicting the price of a house based on its area, where the relationship isn't strictly linear.

12. What does a higher degree polynomial represent in Polynomial Regression? How does it affect the model's complexity?

Ans)

A higher degree polynomial in Polynomial Regression introduces more curvature to the model. Increasing the degree increases the model's complexity, allowing it to fit the training data more closely. However, a very high degree may lead to overfitting.

13. Highlight the key difference between Multi Linear Regression and Polynomial Regression.

Ans)

The key difference lies in the nature of the relationship between independent and dependent variables. Multi Linear Regression deals with linear relationships, while Polynomial Regression accommodates nonlinear relationships by using polynomial functions.

14. Explain the scenario in which Multi Linear Regression is the most appropriate regression technique.

Ans)

Multi Linear Regression is suitable when predicting a target variable based on two or more independent variables, assuming a linear relationship among them. For example, predicting house prices based on features like area, number of bedrooms, and location**.**

15. What is the primary goal of regression analysis?

Ans)

The primary goal of regression analysis is to model and analyze the relationship between a dependent variable and one or more independent variables. It aims to understand the pattern of the data, make predictions, and infer the strength and nature of the relationships.