Date: **21/04/24**

SMART INTERNZ - APSCHE

AI / ML Training

Assignment: Data Wrangling and Regression Analysis

Instructions: Answer the following questions to the best of your ability. Provide concise explanations where necessary.
Section A: Data Wrangling (Questions 1-6)

1. What is the primary objective of data wrangling?
   a) Data visualization
   b) Data cleaning and transformation ✔
   c) Statistical analysis
   d) Machine learning modeling

Explanation: Data wrangling, also known as data munging, is the process of cleaning, structuring, and enriching raw data into a more suitable format for analysis. Its primary objective is to ensure that the data is clean, consistent, and relevant for analysis by removing errors, handling missing values, transforming data types, and restructuring data as needed.

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

One common technique to convert categorical data into numerical data is Label Encoding. In Label Encoding, each category is assigned a unique numerical value. For example, if there are three categories: "red," "blue," and "green," they might be encoded as 1, 2, and 3, respectively. This technique helps in data analysis by allowing machine learning algorithms to operate on categorical data, which usually require numerical inputs. However, it should be noted that Label Encoding may introduce ordinality where there is none, potentially leading to incorrect model assumptions.

3. How does LabelEncoding differ from OneHotEncoding?

LabelEncoding assigns a unique numerical value to each category in the feature, whereas OneHotEncoding creates binary dummy variables for each category. In LabelEncoding, ordinality is introduced, implying an order among the categories, which may not always be accurate. On the other hand, OneHotEncoding eliminates this issue by representing each category as a separate binary feature, where 1 indicates the presence of the category and 0 indicates absence. OneHotEncoding is often preferred when there is no ordinal relationship among the categories.

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

One commonly used method for detecting outliers is the Z-score method. In this method, the Z-score of each data point is calculated by subtracting the mean of the dataset and then dividing by the standard deviation. Data points with a Z-score above a certain threshold (typically $|Z| > 3$) are considered outliers. It is important to identify outliers because they can significantly affect the results of statistical analyses and machine learning models. Outliers can skew the distribution of the data, distort estimates of central tendency and variability, and reduce the accuracy and generalizability of predictive models.

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

The Quantile Method involves setting thresholds based on the quantiles of the data distribution. Outliers are identified by comparing data points against these thresholds. For example, outliers can be defined as data points that fall below the first quartile (Q1) minus 1.5 times the interquartile range (IQR) or above the third quartile (Q3) plus 1.5 times the IQR. Outliers identified using this method can be either removed from the dataset or treated separately based on the context of the analysis.

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

A Box Plot, also known as a box-and-whisker plot, is a graphical representation of the distribution of a dataset, particularly its central tendency, variability, and skewness. It consists of a box that spans the interquartile range (IQR) of the data, with a line inside representing the median, and "whiskers" extending from the box to the minimum and maximum values within a certain range. Box Plots are useful in identifying potential outliers because they visually display the spread of the data and highlight any data points that fall outside the whiskers, which are typically defined based on some multiple of the IQR. Outliers can be easily identified as individual points beyond the whiskers of the box plot. Therefore, Box Plots provide a quick and intuitive way to detect outliers and understand the distribution of the data.

Section B: Regression Analysis (Questions 7-15)

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

When predicting a continuous target variable, linear regression is used. This method assumes a linear relationship between the independent variables and the target variable, represented by a straight line equation. It's commonly employed for tasks like predicting sales figures, stock prices, or temperature readings based on one or more explanatory variables.

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

The two main types of regression are:

Linear Regression: Linear regression establishes a linear relationship between the independent variables and the dependent variable. It assumes that the relationship between the variables can be expressed by a straight line equation.

Non-linear Regression: Non-linear regression is used when the relationship between the independent and dependent variables is not linear. It involves fitting a curve or a nonlinear function to the data points.

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

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

Example Scenario: Predicting a student's exam score based on the number of hours spent studying. Here, the independent variable is the number of hours studied, and the dependent variable is the exam score.

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

In Multi Linear Regression, multiple independent variables are involved, typically more than one.

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

Polynomial Regression should be utilized when the relationship between the independent and dependent variables is non-linear. It is suitable for capturing curved relationships.

Scenario: Predicting house prices based on square footage. In this scenario, the relationship between house prices and square footage might not be linear, as larger houses might have disproportionately higher prices. Polynomial regression can capture this non-linear relationship better than simple linear regression.

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

A higher degree polynomial in Polynomial Regression represents a curve of higher order. For example, a quadratic (degree 2) polynomial introduces a squared term.

Increasing the degree of the polynomial increases the complexity of the model. Higher-degree polynomials can fit the training data more closely, but they also increase the risk of overfitting, where the model captures noise in the data rather than the underlying pattern.

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

The key difference between Multi Linear Regression and Polynomial Regression lies in the nature of the relationship between the independent and dependent variables.

Multi Linear Regression involves linear relationships between multiple independent variables and the dependent variable. Polynomial Regression, on the other hand, can capture non-linear relationships by introducing polynomial terms (e.g., squared terms, cubic terms) into the model.

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

Multi Linear Regression is most appropriate when there are multiple independent variables that have a linear relationship with the dependent variable. It is suitable for scenarios where the outcome is influenced by multiple factors simultaneously.

Example Scenario: Predicting a person's salary based on their education level, years of experience, and age. Here, each independent variable (education level, years of experience, age) contributes linearly to the prediction of salary.

15. What is the primary goal of regression analysis?

The primary goal of regression analysis is to understand the relationship between the dependent variable (response) and one or more independent variables (predictors). It aims to predict the value of the dependent variable based on the values of the independent variables, as well as to assess the strength and significance of the relationships.