**ASSIGNMENT-2**

1.**What is the primary objective of data wrangling?**

* Data Visualization
* Data cleaning and transformation
* Statistical analysis
* Machine learning modelling

Ans) Data clening and transformation

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

Ans) One common technique to convert categorical data into numerical data is “one-hot-encoding”. This method represents each category as a binary vector, where each element corresponds to a unique category. It helps in data analysis by enabling machine learning algorithms to work with categorical variables, as they often requires numerical inputs.one hot encoding prevents the model from assuming ordinal realtionships between categories and ensures meaningful representation of categorical information in the data set.

3.**How does Label Encoding differ from one hot encoding?**

Ans) Label Encoding and one hot encoding are two different techniques are used in the machine learning to handle categorical data.

1. Label Encoding:

* Assigns a unique numerical label to each category in a categorical variable.
* Sutiable for ordinal data, Where there is an inherent order among the categorical.
* May introduce unintended ordinal relationships between categories that don’t actually exists.

1. One-Hot-Encoding:

* Represents each category with a binary vector.
* Creates a new binary column for each category and only one of these column is” hot” for a given row.
* sutiable for nominal data, where there is no inherent order among the categorical.
* Avoids introducing artificial ordinal relationships.

In the summary, label encoding uses integers to represents categories, potentially implying an order, while one hot encoding creates binary columns to represents each category independently , avoiding any implied order. The choice between the two depends on the nature of the data and requirements of the machine learning model.

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

Ans) One commonly used methos for detecting Outliers is the Z-score. It involves calculating the standard deviations its is from the mean. Points with high absolute Z-score are considered outliers.

Identifying outliers is important because they can significantly influence statistical analyses, leading to skewed results or in accurate interpretations. outliers may indicate errors in data collection measurement issues,or highlight unique patterns that warrant further investigation. Removing or addressing outlier can enhance the robustness and reliability of data analyses.

5.**Explian how outliers are handled using the Quantile method?**

Ans) The Quantile method involves setting thresholds based on Quantile. typically the lower and upper percentile of the data distribution. Outlier are then identified and handled based on these thresholds. Here’s basic explanation of the process:

1. Define Quantiles: Choose specifics percentiles, such as the 25thand 75thpercentiles to establish the range within in which most data points lie.
2. Calculate interquartile range: find the IQR by subtracting the lower quartile from the upper quartile.
3. Set Thresholds: Establish lower and upper bound for identifying outliers often using a multiplier. data points outside these bounds are considered outliers.
4. Identifying and Handle Outliers: points beyond the defined thresholds are flagged as outliers. Depending on the analysis, outlier can be addressed by either excluding them transforming them, or applying statistical methods that are less sensitive to extreme values.

The Quantile method is robust and less sensitive to extreme values than some other outlier detection techniques making it a useful approach in various data analysis scenarios.

**6.Discuss the significance of a box plot in data analysis. how does ii aid in identifying potential outliers?**

Ans) A Box plot, also knowns as a box-and-whisker plot, is a valuable tool in the analaysis for visualizing the distribution and central tendency of a dataset.it consists of a rectangular “Box ”that represents the interquartile range and whiskers that extend to the minimum and maximum values. Here’s how it aids in identifying the potential outliers:

1. Outlier Detection: Box plots help identify potential outliers by highlighting the data points that fall beyonds the whiskers. outliers are often considered values the lie more than 1.5 times the IQR above the upper quartile or below the lower quqrtile.
2. Central Tendency: The position of the median within the box provides insight into the central tendency of the data. if the median is not centerd, it might indicates skewness or asymmetry in the distribution.
3. Symmetry and Skewness : box plots visually reveal the symmetry or skewness of a distribution. if one whisker is longer than the other, it suggests asymmetry, aiding in understanding the overall shapes of the data.

In summary ,box plots provides a concise summary of key statistical properties, facilating the identification of potential outlier and offering a quick visual overview of the distribution of data.

7.**What type of regression is employees when predicting continuous target variables?**

Ans) Linear Regression is commonly used in the when predicting continuous target variables.

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

Ans) The two main types of regressions are:

1. Linear Regression:

* Purpose: predicting a continuous dependent variable based on one or more independent variables.
* Assumption: Assume a linear relationships between the independent and dependent.
* Equations: Represents a straight line(y=mx+b),where y is the dependent variables,x is the independent variable, m is the slope, and b is the intercept.

1. Logistic Regression:

* Purpose: Used the binary classifications problems where the outcome is categorical and has two classes.
* Nature: Estimates the probability that given input belongs to a particular category.
* Equation: Employs the logistics functions to map the output between 0 and 1, represented probabilities.

These regression types serve different purposes and are applied based on the nature of the data and the problem at the hand.

9.**When would you use simple linear regression? provide an example scenario**?

Ans) simple linear regression is used when you want to understand the linear relationships between two variables. it’s relationships when there is a single independent variables predicting a dependent variables.

For examples, if you want to predict a student’s exam score based on the number of hours they studied you could use simple linear regression. The model would help you estimate how the exam the score changes for each additional hours to study.

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

Ans) In multi linear regression, there are typically more than two dependent variables involved.the model considered the relationship between a dependent variables and two or more independent variables.

11.**When should polynomial Regression be utilized? Provide a scenario polynomial regression would be preferable over simple regression**?

Ans) regression is suitable when the relationships between the independent and dependent variables is nonlinear uses it when the data suggest a curve pattern rather than a straight line. For example, predicting houses prices based on squares ,footage alone might benefit from polynomial regression if there’s a clear curvature in data, capturing more complex relationship than simple linear regression can handle.

12. **What does a higher degree polynomial Regression? How does it affect the model’s complexity**?

Ans) In polynomial Regression, a higher degree polynomial corresponds to more intricate curves or surfaces that can better fit complex patterns in the data. however ,increasing the degree also elevates the model’s complexity, potentially leading to overfitting, where the model fits the training data too closely and struggles to generalize to new data .therfore, choosing an appropriate degree is crucial to balance the trade off between fiting the training data well maintain good generalization.

13.**Higlights the key differences between multi linear regression and polynomial regression?**

Ans)The main distinction between Multiple Linear regression and polynomial regression is that multiple regression deals with linear relationships involving multiple predictors, whereas polynomial regression accommdates non linear relationships by introducing polynomials terms enabling it to capture more than intricate patterns in the data.

14**.Explian Scenario in which multi linear regression is the most appropriate regression techniques**?

Ans) Multi linear regression is most appropriate when you have a dependent variable influenced by tow or more independent variables. This techniques is suitable when the relationships between the variables is assumed to be linear and you want to model the impact of multiple predictors on the response variable simultaneously.

15.**What is the primary goal of the regression analysis?**

Ans) The primary goal of the regression analysis is to examine and quantify the relationships between a dependent variables and one or more independent variables. its aims to understand how the changes are associated with changes in the dependent variables, allowing for prediction and inferences.