MACHINE LEARNING

1-Ans: **C**

2-Ans: **B**

3-Ans: **C**

4-Ans: **C**

5-Ans: **B**

6-Ans: **A&D**

7-Ans: **B&C**

8-Ans: **A&C**

9-Ans: **A&B**

10-Ans: The adjusted R-squared is a modified version of the R-squared that takes into account the number of predictors in the model. It is calculated as:

Adjusted R-squared = 1 - [(1 - R-squared)*(n - 1)/(n - k - 1)]

where R-squared is the ordinary coefficient of determination, n is the sample size, and k is the number of predictors in the model.The adjusted R-squared penalizes the presence of unnecessary predictors in the model by reducing the value of the adjusted R-squared when additional predictors do not improve the fit of the model. This is because the adjusted R-squared takes into account both the goodness of fit of the model and the complexity of the model (i.e., the number of predictors).

11-Ans: Ridge regression and Lasso regression are both linear regression techniques that are used for regularization, a technique for preventing overfitting in predictive models. While both techniques aim to reduce the complexity of a model and prevent overfitting, they differ in the way they accomplish this goal.

Ridge regression adds a penalty term to the regression coefficients that is proportional to the squared magnitude of the coefficients. The resulting objective function, which is to be minimized, is the sum of the squared error term and the penalty term multiplied by a regularization parameter (lambda), which controls the strength of the penalty. Ridge regression attempts to reduce the impact of each individual feature on the outcome, while still keeping all features in the model.

Lasso regression, on the other hand, adds a penalty term to the regression coefficients that is proportional to the absolute value of the coefficients. The resulting objective function is the sum of the squared error term and the penalty term multiplied by a regularization parameter (lambda). Unlike Ridge regression, Lasso regression aims to set some coefficients to zero, effectively removing some features from the model. This makes Lasso regression useful for feature selection, where the goal is to identify a subset of important features that are most strongly associated with the outcome.

12-Ans: Variance Inflation Factor (VIF) is a statistical measure that is used to identify the extent of multicollinearity (i.e., high correlation) among the independent variables in a regression model. VIF quantifies the amount of variance in the estimated regression coefficients that is caused by the correlation between the independent variables.

The VIF for a particular independent variable is calculated by regressing that variable on all of the other independent variables in the model, and then calculating the ratio of the variance of the coefficient estimates to the variance of the coefficient estimates for a single variable regression model. The formula for VIF is:

VIF = 1 / (1 - R^2)

where R^2 is the coefficient of determination from the regression of the independent variable on the other independent variables.

The suitable value of VIF for a feature to be included in a regression model varies depending on the context and the goals of the analysis. A common rule of thumb is that a VIF value of 1 indicates no correlation between the independent variable and the other independent variables, and values greater than 1 indicate increasing levels of multicollinearity. In general, VIF values of 1-2 are considered low, values between 2-5 are moderate, and values greater than 5 are considered high.

13-Ans: Scaling the data is a common preprocessing step that is often recommended before training machine learning models. There are several reasons why scaling is important:

Prevents numerical instability: When the input features have very different scales, some algorithms may struggle to converge or may take longer to converge. For example, in gradient descent optimization algorithms, the features with large scales will have a much larger influence on the optimization process, which can lead to numerical instability and slow convergence.

Improves model performance: Scaling the data can help to improve the performance of some models, especially those that are based on distance measures, such as k-nearest neighbors and support vector machines. In these models, if the features have different scales, the model may give undue importance to certain features, leading to a suboptimal solution.

Helps interpret model coefficients: Scaling the data makes the coefficients of the model more comparable and easier to interpret. For example, when using linear regression, the coefficients can be interpreted as the change in the outcome variable for a one-unit increase in the input feature. If the input features have different scales, it becomes difficult to compare the coefficients.

Avoids leakage of information: When the data is not scaled, it can lead to information leakage from the test set to the training set, which can lead to overfitting. If the test set has different scales than the training set, the model may perform poorly on the test set, even if it performs well on the training set.

14-Ans: Goodness of fit measures are used to evaluate how well a linear regression model fits the observed data. The choice of which measure to use depends on the specific problem and the goals of the analysis. Some of the most commonly used measures to check the goodness of fit in linear regression include:

R-squared (R^2): R-squared is a measure of how well the regression model fits the data, and it represents the proportion of variance in the dependent variable that is explained by the independent variables. R^2 ranges from 0 to 1, with a value of 1 indicating a perfect fit.

Mean Squared Error (MSE): MSE is a measure of the average squared difference between the predicted and actual values of the dependent variable. It is commonly used to evaluate the accuracy of the model's predictions.

Root Mean Squared Error (RMSE): RMSE is the square root of the MSE, and it represents the average magnitude of the error in the predictions. Like MSE, it is commonly used to evaluate the accuracy of the model's predictions.

Mean Absolute Error (MAE): MAE is a measure of the average absolute difference between the predicted and actual values of the dependent variable. It is similar to RMSE but gives equal weight to all errors, regardless of their magnitude.

Residual Standard Error (RSE): RSE is a measure of the standard deviation of the residuals, which are the differences between the predicted and actual values of the dependent variable. It represents the average amount by which the model's predictions deviate from the actual values.

Adjusted R-squared: Adjusted R-squared is a modified version of R-squared that takes into account the number of independent variables in the model. It penalizes the use of unnecessary variables in the model and provides a better estimate of how well the model will generalize to new data.

15-Ans: True Positives (TP) = 1000

False Positives (FP) = 50

False Negatives (FN) = 250

True Negatives (TN) = 1200

Sensitivity (or Recall) = TP / (TP + FN) = 1000 / (1000 + 250) = 0.8

Specificity = TN / (TN + FP) = 1200 / (1200 + 50) = 0.96

Precision = TP / (TP + FP) = 1000 / (1000 + 50) = 0.95

Accuracy = (TP + TN) / (TP + FP + TN + FN) = (1000 + 1200) / (1000 + 50 + 250 + 1200) = 0.904

So the sensitivity (or recall) is 0.8, the specificity is 0.96, the precision is 0.95, and the accuracy is 0.904.

STATISTICS

1-Ans: **C**

2-Ans: **A**

3-Ans: **A**

4-Ans: **C**

5-Ans: **C**

6-Ans: **B**

7-Ans: **C**

8-Ans: **B**

9-Ans: **B**

10-Ans: Histograms and box plots are graphical representations for the frequency of numeric data values. They aim to describe the data and explore the central tendency and variability before using advanced statistical analysis techniques.Both histograms and box plots allow to visually assess the central tendency, the amount of variation in the data as well as the presence of gaps, outliers or unusual data points.

11-Ans: Selecting metrics involves identifying the key performance indicators (KPIs) that are most important for measuring the success of your project, product, or business. Here are some steps to follow when selecting metrics:

Identify your goals, Brainstorm potential metrics, Evaluate each metric, Choose a small set of metrics, and Monitor and adjust

12-Ans: We need to perform a statistical test first. There are steps to follow :

a. Formulate your null and alternative hypotheses

b. Choose an appropriate statistical test

c. Determine the level of significance

d. Conduct the statistical test

e. Interpret the results

It's important to note that statistical significance does not necessarily imply practical significance or meaningfulness. Therefore, it's essential to consider the effect size and other relevant factors when interpreting the results of a statistical test.

13-Ans: Examples of data that doesnot have a Gaussian distribution, nor log-normal are Power-law distributions, Bimodal distributions, Poisson distributions, Exponential distributions, Weibull distributions,Gamma distributions,

14-Ans: The median is often a better measure than the mean for skewed distributions or datasets that contain outliers. Here's an example:

If we calculate the mean salary, it will be significantly higher than the majority of the salaries and may not accurately represent the typical salary in the company. This is because the mean is affected by extreme values, such as the high earners.

On the other hand, the median is not affected by extreme values, and it represents the value that separates the top 50% of salaries from the bottom 50%. In this example, the median salary would be close to $50,000, which is a better representation of the typical salary in the company.

Therefore, in cases where the distribution is skewed or contains outliers, the median is a more robust measure than the mean, as it is less influenced by extreme values and gives a better indication of the typical value in the dataset.

15-Ans: In statistics, the likelihood is a function that measures the probability of observing a set of data given a specific set of parameter values in a statistical model. In other words, the likelihood function tells us how likely it is that the data we have observed would be generated by a particular set of parameter values in the model. The likelihood function is denoted by L(θ|X), where θ represents the set of model parameters, and X represents the observed data. The likelihood function is calculated by assuming that the parameter values are fixed and that the data are random. It is typically defined as the joint probability density function of the data, given the parameter values. The maximum likelihood estimation (MLE) is a method that uses the likelihood function to estimate the most likely values of the parameters in the model that generated the data. The MLE seeks to find the parameter values that maximize the likelihood function. In other words, it finds the parameter values that make the observed data most probable.

SQL WORKSHEET

1-Ans: **A&B&C&D**

2-Ans: **A&B&C&D**

3-Ans: **B**

4-Ans: **C**

5-Ans: **B**

6-Ans: **B**

7-Ans: **A**

8-Ans: **C**

9-Ans: **B**

10-Ans: **A**

11-Ans: **Denormalization** is a database optimization technique where we add redundant data in the database to get rid of the complex join operations. This is done to speed up database access speed. Denormalization is done after normalization for improving the performance of the database. The data from one table is included in another table to reduce the number of joins in the query and hence helps in speeding up the performance.

12-Ans: A **database cursor** can be thought of as a pointer to a specific row within a query result. The pointer can be moved from one row to the next. Depending on the type of cursor, you may even be able to move it to the previous row.

13-Ans: SQL commands are mainly categorized into five categories such as

- DDL – Data Definition Language.
- DQL – Data Query Language.
- DML – Data Manipulation Language.
- DCL – Data Control Language.
- TCL – Transaction Control Language.

14-Ans: **Constraints** are used to limit the type of data that can go into a table. This ensures the accuracy and reliability of the data in the table. If there is any violation between the constraint and the data action, the action is aborted.

15-Ans: **Auto Increment** allows us to automatically generate values in a numeric column upon row insertion.