MACHINE LEARNING

ASSIGNMENT NO – 4

Q1. C)

Q2. B)

Q3. C)

Q4. B)

Q5. B)

Q6. A) , D)

Q7. B)

Q8. A) , C)

Q9. B)

Q10. The adjusted R-squared penalizes the presence of unnecessary predictors in the model by taking into account the number of predictors in the model and the sample size. The adjusted R-squared will decrease as the number of predictors in the model increase, even if the predictors are not contributing to the model fit. This means that adding unnecessary predictors will result in a decrease in the adjusted R-squared, providing a penalization for the presence of those predictors. By using the adjusted R-squared, we can determine the number of predictors that should be included in the model to achieve the best balance between model fit and parsimony.

Q11. Ridge Regression and Lasso Regression are both regularization techniques used to prevent overfitting in linear regression.

1. Ridge Regression: It adds a penalty term to the least squares objective function to reduce the magnitude of the coefficients. The penalty term is equal to the square of the magnitude of the coefficients multiplied by a constant alpha. This results in the optimization problem that encourages coefficients to be close to zero but does not set any coefficient exactly to zero.
2. Lasso Regression: Lasso stands for Least Absolute Shrinkage and Selection Operator. It adds a penalty term to the least squares objective function that is equal to the absolute value of the magnitude of the coefficients multiplied by a constant alpha. This results in some coefficients being set exactly to zero, effectively performing feature selection and reducing the number of predictors in the model.

In summary, Ridge Regression shrinks the coefficients, but does not set any coefficient to zero, while Lasso Regression not only shrinks the coefficients, but also sets some coefficients to zero, resulting in sparse solutions.

Q12. VIF (Variance Inflation Factor) is a measure of collinearity among the predictors in a multiple regression model. VIF calculates the amount of increase in the variance of the estimated regression coefficients caused by the presence of the predictor variables. The VIF value ranges from 1 to infinity. A VIF value of 1 indicates that there is no collinearity among the predictors.

A suitable value of a VIF for a feature to be included in a regression modelling is usually considered to be below 4 or 5. A VIF greater than 4 or 5 indicates that the predictor is highly correlated with other predictors in the model and might not add much new information. In this case, it is recommended to remove the predictor with the highest VIF to reduce collinearity. The process can be repeated until all VIF values are below 4 or 5.

Q13. Scaling the data before training a model is important because most of the machine learning algorithms are based on Euclidean distance and use the scale of the data to calculate the distances. If the scale of the data is not the same, then some predictors will dominate the others and may result in biased or incorrect results.

For example, in linear regression, the coefficients are estimated based on the minimization of the sum of squared residuals. If the scale of one predictor is much larger than the scale of other predictors, then the squared residuals from that predictor will dominate the sum, leading to a bias in the coefficients.

Moreover, many algorithms, such as k-nearest neighbours, k-means clustering, or support vector machines, are sensitive to the scale of the data and can be influenced by the presence of large-scale features. Scaling the data ensures that each feature has a similar contribution to the final result and can lead to improved performance of the model.

In conclusion, scaling the data is important to ensure that all features are on a similar scale, which can prevent bias in the model, improve the interpretability of the results, and lead to better performance.

Q14. The following are the common metrics used to check the goodness of fit in linear regression:

1. Mean Squared Error (MSE): It is the average of the squared differences between the actual and predicted values. A lower MSE value indicates a better fit.
2. Root Mean Squared Error (RMSE): It is the square root of MSE and represents the average magnitude of the error.
3. R-squared: It is a measure of the proportion of variance in the dependent variable that is explained by the independent variables in the model. A higher R-squared value indicates a better fit.
4. Adjusted R-squared: It is a modified version of R-squared that adjusts for the number of predictors in the model. It is useful in comparison between models with different number of predictors.
5. Mean Absolute Error (MAE): It is the average of the absolute differences between the actual and predicted values.

These metrics help to evaluate the performance of the linear regression model and to determine the goodness of fit of the model. However, it's important to choose the appropriate metric based on the specific use case and the data at hand.

Q15. Sensitivity (True Positive Rate) = 1000 / (1000 + 250) = 0.8

Specificity (True Negative Rate) = 1200 / (1200 + 50) = 0.96

Precision (Positive Predictive Value) = 1000 / (1000 + 50) = 0.95

Recall (Sensitivity, True Positive Rate) = 1000 / (1000 + 250) = 0.8

Accuracy = (True Positives + True Negatives) / Total = (1000 + 1200) / (1000 + 50 + 250 + 1200) = 0.91