FLIP ROBO

MACHINE LEARNING ASSIGNMENT 6

1. C
2. B
3. C
4. C
5. B
6. A&D
7. B & C
8. A & C
9. A & B
10. The adjusted R-squared is a modified version of the R-squared that penalizes the addition of unnecessary predictors to the model. As more predictors are added to the model, the R-squared value will typically increase, even if the predictors are not useful. The adjusted R-squared accounts for this by adjusting the R-squared value based on the number of predictors in the model. It is a more reliable indicator of the goodness of fit of a model and is useful in comparing models with different numbers of predictors.
11. Ridge and Lasso regression are both regularization techniques used to prevent overfitting in linear regression. Ridge regression adds a penalty term to the cost function that is proportional to the square of the magnitude of the coefficients. This helps to shrink the coefficients of less important predictors towards zero, but does not set any coefficients exactly to zero. Lasso regression, on the other hand, adds a penalty term to the cost function that is proportional to the absolute value of the coefficients. This can result in some coefficients being set exactly to zero, effectively performing feature selection.

12. **VIF stands for Variance Inflation Factor; it is a measure of how much the variance of the estimated regression coefficients are increased because of collinearity. A high VIF indicates that the corresponding predictor is highly correlated with one or more of the other predictors. A VIF value of 1 indicates that there is no correlation between this predictor and any other predictors, while a value greater than 1 indicates that there is correlation. A suitable value of VIF for a feature to be included in a regression modeling is typically less than 5 or 10.**
13. **Scaling the data is important before training a model because many machine learning algorithms use distance based calculations. So if the data is not scaled, then the algorithm will be sensitive to the scale of the data. For example, if one feature is measured in kilometers and another feature is measured in meters, then the algorithm will be biased towards the feature measured in kilometers. Scaling the data ensures that all the features are on the same scale, which leads to a fair comparison of the importance of each feature.**
14. **There are several metrics used to check the goodness of fit in linear regression, some examples are:**

<u>R-squared:</u> R-squared measures the proportion of variation in the dependent variable that is explained by the independent variables. It ranges from 0 to 1, where 1 indicates a perfect fit.
<u>Mean Squared Error (MSE):</u> MSE is the average of the square of the residuals, it measures the average difference between the predicted values and the true values.
<u>Root Mean Squared Error (RMSE):</u> it is the square root of MSE and it gives the error in the same unit as the response variable.
Mean Absolute Error (MAE): it is the mean of the absolute values of the residuals.
<u>Adjusted R-squared:</u> It is a modified version of the R-squared that penalizes the addition of unnecessary predictors to the model.

15. **Sensitivity (also known as recall) = True Positives / (True Positives + False Negatives) = 1000 / (1000 + 250) = 0.8**
    Specificity = True negatives / (True negatives + False positives) = 1200 / (1200 + 50) = 0.96
    Precision = True Positives / (True Positives + False Positives) = 1000 / (1000 + 50) = 0.95
    Recall = True Positives / (True Positives + False Negatives) = 1000 / (1000 + 250) = 0.8
    Accuracy = (True Positives + True negatives) / (Total) = (1000 + 1200) / (1000+50+250+1200) = 0.89
    Note: sensitivity and recall are the same.