MACHINE LEARNING WORKSHEET_SET 4

1.C) High R-squared value for train-set and Low R-squared value for test-set.

2.B) Decision trees are highly prone to overfitting.

3.C) Random Forest

4.A) Accuracy

5.B) Model B

6.A) Ridge, D) Lasso

7.B) Decision Tree C) Random Forest

8.D) All of the above

9.A) We initialize the probabilities of the distribution as 1/n, where n is the number of data-points

10.Explain how does the adjusted R-squared penalize the presence of unnecessary predictors in the model?

Ans: The adjusted R-squared is a modified version of R-squared that accounts for predictors that are not significant in a regression model. In other words, the adjusted R-squared shows whether adding additional predictors improve a regression model or not. To understand adjusted R-squared, an understanding of R-squared is required.

- The adjusted R-squared is a modified version of R-squared that adjusts for predictors that are not significant in a regression model.
- Compared to a model with additional input variables, a lower adjusted R-squared indicates that the additional input variables are not adding value to the model.
- Compared to a model with additional input variables, a higher adjusted R-squared indicates that the additional input variables are adding value to the model.

11) Differentiate between Ridge and Lasso Regression.

Ans: Similar to the lasso regression, ridge regression puts a similar constraint on the coefficients by introducing a penalty factor. However, **while lasso regression takes the magnitude of the coefficients, ridge regression takes the square**. Ridge regression is also referred to as L2 Regularization.

What is the advantage of lasso regression over ridge regression?

One obvious advantage of lasso regression over ridge regression, is that **it produces simpler and more interpretable models that incorporate only a reduced set of the predictors**. However, neither ridge regression nor the lasso will universally dominate the other.

12) What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modelling?

Ans: A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

A variance inflation factor is a tool to help identify the degree of multicollinearity. Multiple regression is used when a person wants to test the effect of multiple variables on a particular outcome. The dependent variable is the outcome that is being acted upon by the independent variables—the inputs into the model. Multicollinearity exists when there is a linear relationship, or correlation, between one or more of the independent variables or inputs.

13) Why do we need to scale the data before feeding it to the train the model?

Ans: **To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features**, we scale the data before feeding it to the model.