ASSIGNMENT – 4 MACHINE LEARNING

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. B) Sensitivity
5. B) Model B
6. A) Ridge and D) Lasso
7. C) Random Forest
8. D) All of the above
9. B) A tree in the ensemble focuses more on the data points on which the previous tree was not performing well
10. The adjusted R-squared penalizes the presence of unnecessary predictors in the model by reducing the R-squared value in proportion to the number of extra predictors. This is because the adjusted R-squared takes into account the number of predictors in the model, meaning that the higher the number of predictors, the lower the adjusted R-squared value will be. Thus, the adjusted R-squared acts as a penalty for adding unnecessary predictors to the model, as it reduces the overall accuracy of the model.
11. Ridge Regression: Ridge regression is a type of linear regression in which the coefficients of the model are estimated by minimizing the residual sum of squares with the addition of an L2 regularization term. The regularization term helps to reduce the complexity of the model and therefore helps to reduce overfitting. It works by adding a penalty term to the cost function, which helps to regularize the parameters and reduce the variance in the estimates.

    Lasso Regression: Lasso regression is another type of linear regression in which the coefficients of the model are estimated by minimizing the residual sum of squares with the addition of an L1 regularization term. The L1 regularization term helps to reduce the complexity of the model and therefore helps to reduce overfitting. It works by adding a penalty term to the cost function, which helps to regularize the parameters and reduce the variance in the estimates. Unlike ridge regression, lasso regression can set some of the coefficients of the model to zero and thus remove them from the model, making it simpler and easier to interpret.

12. VIF stands for Variance Inflation Factor. It is a measure used to assess how much the variance of a model parameter is inflated due to multicollinearity in the model. A VIF of 1 indicates that there is no multicollinearity, while a VIF greater than 1 indicates that there is multicollinearity present in the model. A suitable value of VIF for a feature to be included in a regression modelling is less than 5.
13. Scaling the data is important because it ensures that all the features are on the same scale, which helps the model to learn faster and more accurately. Without scaling, the features with larger values will dominate the model and make it harder for the model to learn from the other features. Additionally, some machine learning algorithms (such as K-nearest neighbors) require that all the features be on the same scale for the algorithm to work correctly.
14. 1. R-Squared: R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination. 2. Adjusted R-Squared: Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. 3. F-Statistic: The F-statistic is a measure of how significant the overall fit of the regression model is. The F-statistic is calculated from the sums of squares of the regression and the error and has a chi-square distribution. 4. Akaike Information Criterion (AIC): The AIC is an estimator of the relative quality of statistical models for a given set of data. It is based on the likelihood function and it provides a means for model selection. 5. Root Mean Square Error (RMSE): The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model
15. Sensitivity = 1000/1250 = 80% Specificity = 1200/1450 = 82.76% Precision = 1000/1050 = 95.23% Recall = 1000/1250 = 80% Accuracy = (1000 + 1200)/ (1000 + 50 + 250 + 1200) = 93.10%