**Machine Learning Assignment 6**

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 D) Lasso
7. 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 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 is a modified version of the R-squared value, which adjusts for the number of predictors in the model. It penalizes the presence of unnecessary predictors in the model by reducing the R-squared value if a new predictor does not improve the model's performance significantly. The adjusted R-squared increases only if the new predictor adds significant value to the model, thus avoiding the issue of overfitting.
11. Ridge and Lasso Regression are both regularization techniques used in Linear Regression to reduce overfitting. The key difference between the two techniques is the penalty term used. Ridge Regression uses L2 regularization, which adds the sum of squared coefficients multiplied by a penalty factor to the cost function, whereas Lasso Regression uses L1 regularization, which adds the absolute value of the coefficients multiplied by a penalty factor to the cost function.
12. VIF (Variance Inflation Factor) is a measure used to identify the presence of multicollinearity in the data. It measures the degree to which the variance of the regression coefficient is inflated due to the correlation of the predictor variable with other predictor variables. A suitable value of VIF for a feature to be included in a regression model is 1 or less, as values greater than 1 indicate the presence of high multicollinearity.
13. Scaling the data before feeding it to the model is important to ensure that all the features are on the same scale and have equal importance in the model. If the features are not scaled, then features with higher magnitudes will dominate the model, and features with lower magnitudes will be ignored. Scaling helps in improving the model's accuracy and stability, and it also reduces the training time.
14. The different metrics used to check the goodness of fit in linear regression are:

* R-squared value: measures the proportion of variance in the dependent variable explained by the independent variables.
* Adjusted R-squared value: a modified version of the R-squared value that adjusts for the number of predictors in the model.
* Mean squared error (MSE): measures the average squared difference between the predicted and actual values.
* Root mean squared error (RMSE): measures the square root of the average squared difference between the predicted and actual values.
* Mean absolute error (MAE): measures the average absolute difference between the predicted and actual values.

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

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

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

Recall (same as Sensitivity) = TP / (TP + FN) = 0.8

Accuracy = (TP + TN) / (TP + TN + FP + FN) = (1000 + 1200) / (1000 + 1200 + 50 + 250) = 0.88