Machine Learning Assignment Ans

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. B) Decision Tree
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. Adjusted R-squared is a modified version of R-squared, which considers the number of predictors in the model. It penalizes unnecessary predictors in the model by subtracting a penalty term from R-squared. This penalty term increases as the number of predictors in the model increases or as the sample size decreases and reduces the value of the adjusted R-squared. A higher adjusted R-squared indicates a better model fit while taking into account the number of predictors.

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

The main difference between Ridge and Lasso Regression is how they apply regularization. Ridge Regression applies L2 regularization, which adds a penalty term proportional to the square of the coefficients. This penalty term shrinks the coefficient values towards zero but does not set any coefficients exactly to zero. As a result, Ridge Regression can be useful when all the predictors are potentially important.

On the other hand, Lasso Regression applies L1 regularization, which adds a penalty term proportional to the absolute value of the coefficients. This penalty term shrinks the coefficient values towards zero and sets some of them exactly to zero. As a result, Lasso Regression can be useful when there are many predictors. However, only a few of them are likely to be important. The Lasso method can also be used for feature selection, as it can automatically eliminate some of the less important predictors from the model.

12. VIF stands for Variance Inflation Factor. VIF is used to detect multicollinearity in regression analysis. Multicollinearity occurs when a regression model's two or more independent variables are highly correlated. Multicollinearity can cause instability in the regression coefficients and make it difficult to interpret the results.

VIF measures the amount of multicollinearity in a set of predictors. A VIF of 1 indicates no multicollinearity, while a VIF greater than 1 indicates the presence of multicollinearity. Generally, a VIF greater than 5 or 10 is considered high and indicates significant multicollinearity.

As a rule of thumb, removing the predictors with a high VIF value (usually VIF greater than 5 or 10) from the model can cause problems with the interpretation of the regression coefficients and the overall model fit. It is usually better to remove the least important variables with high VIF values and re-run the regression model to check for improved model performance.

13. Scaling the data is essential because it can help machine learning algorithms work better. Many algorithms are designed to work with data with similar ranges, and scaling can help achieve this. Scaling can also improve the performance of some machine learning models by reducing the effect of differences in feature ranges. Finally, scaling can make it easier to compare the relative importance of the different features. Different scaling methods are available, and the choice of method depends on the specific data and the requirements of the machine learning algorithm being used.

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

1. R-squared: R-squared measures the proportion of variance in the dependent variable explained by the independent variables. It ranges from 0 to 1, with higher values indicating a better fit.
2. Adjusted R-squared: Adjusted R-squared is similar to R-squared, but it considers the number of independent variables in the model. It penalizes the addition of unnecessary variables, which can lead to overfitting.
3. Mean Squared Error (MSE): MSE measures the average squared difference between the predicted and actual values. Lower values of MSE indicate a better fit.
4. Root Mean Squared Error (RMSE): RMSE is the square root of the MSE, and it is in the same units as the dependent variable. It provides a measure of the average error in the predicted values.
5. Mean Absolute Error (MAE): MAE measures the average absolute difference between the predicted and actual values. It is less sensitive to outliers than MSE.
6. Residual plots: Residual plots are graphical representations of the differences between predicted and actual values. They can be used to identify patterns in the residuals, which can indicate problems with the model.

15. True Positive (TP) = 1000

False Positive (FP) = 250

False Negative (FN) = 50

True Negative (TN) = 1200

Sensitivity (True Positive Rate) = TP / (TP + FN) = 1000 / (1000 + 50) = 0.9524

Specificity (True Negative Rate) = TN / (FP + TN) = 1200 / (250 + 1200) = 0.8276

Precision = TP / (TP + FP) = 1000 / (1000 + 250) = 0.8

Recall = TP / (TP + FN) = 1000 / (1000 + 50) = 0.9524

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