Machine Learning Worksheet Answers

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. A) SVM B) Logistic Regression C) Random Forest
4. B) Sensitivity
5. B) Model B
6. A) Ridge B) Lasso
7. A) Adaboost
8. A) Pruning B) Restricting the max depth of the tree
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 adjusting the R-squared value for the number of predictors. When the number of predictors increases, the R-squared value also increases even if the predictors are not useful. The adjusted R-squared corrects for this by subtracting a penalty term for each predictor from the R-squared value.

11. Ridge and Lasso are both regularization techniques used in linear regression to prevent overfitting. The main difference between the two is in how they add the penalty term to the cost function. Ridge regression adds a penalty term to the cost function that is the square of the magnitude of the coefficients, while Lasso regression adds a penalty term that is the absolute value of the magnitude of the coefficients.

12. VIF (Variance Inflation Factor) is a measure of how much the variance of the estimated regression coefficients are increased because of multicollinearity (correlation) among the independent variables. A VIF of 1 indicates no multicollinearity, while a VIF greater than 1 indicates that there is multicollinearity. A suitable value of VIF for a feature to be included in a regression modeling is less than 5.

13. Scaling the data before feeding it to the model is important because it ensures that all features are on the same scale. This is necessary because the algorithm uses the scale of the feature to assign weight to it, so if the scales are different, the algorithm will assign more weight to the feature with a larger scale. This can cause the algorithm to be skewed and perform poorly. Additionally, many machine learning algorithms, such as those used in neural networks, require input data to be in a specific range, such as -1 to 1, which is achieved by scaling the data.

14. Different metrics used to check the goodness of fit in linear regression are:

    - R-squared: a measure of how well the model fits the data. It ranges between 0 and 1, where 1 indicates a perfect fit.
    - Mean Squared Error (MSE): a measure of the average of the squared differences between the predicted and actual values.
    - Root Mean Squared Error (RMSE): the square root of the MSE, which gives the same unit of measurement as the data.
    - Mean Absolute Error (MAE): a measure of the average of the absolute differences between the predicted and actual values.

15.

    - Sensitivity (True Positive Rate) = 1000 / (1000 + 50) = 0.95
    - Specificity (True Negative Rate) = 1200 / (1200 + 250) = 0.82
    - Precision = 1000 / (1000 + 250) = 0.8
    - Recall (Sensitivity) = 1000 / (1000 + 50) = 0.95
    - Accuracy = (1000 + 1200) / (1000 + 50 + 250 + 1200) = 0.8