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. B) Logistic Regression
4. A) Accuracy
5. B) Model B
6. A) Ridge , D) Lasso
7. A) Adaboost , B) Decision Tree
8. A) Pruning , C) Restricting the max depth of the tree
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 penalizes the presence of unnecessary predictors in the model by decreasing the overall R-squared value. This is because the adjusted R-squared takes into account the number of predictors in the model, thus if there are too many predictors, the R-squared value will be penalized. This encourages model parsimony and discourages the use of unnecessary predictors in the model.
11. Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean.

    Ridge Regression is usually considered when there is a high correlation between the independent variables or model parameters. As the value of correlation increases the least square estimates evaluates unbiased values. But if the collinearity in the dataset is very high, there can be some bias value. Therefore, we create a bias matrix in the equation of Ridge Regression algorithm.
12. VIF stands for Variance Inflation Factor. It is a measure of the extent to which the variance of a given feature is increased due to the existence of other correlated features in a regression model. The suitable value of VIF for a feature to be included in a regression modelling is typically between 1 and 5.
13. Scaling data is important because it helps to ensure that all of the features of the data have a similar range of values. This allows the model to learn more efficiently, as the features will be more evenly distributed and thus the model will not be biased towards any particular feature. Additionally, scaling the data can help to reduce the effect of outliers and can help to make the model more robust.
14. The different metrics used to check for the goodness of fit in linear regression are:
    1. R-squared (R2): This metric measures the amount of variance in the dependent variable that is explained by the model. It's value ranges from 0 to 1, with higher values indicating better model fit.
    2. Adjusted R-squared (Adj. R2): This is a modified version of R2 that takes into account the number of predictors in the model. It adjusts the R2 value to account for the complexity of the model.
    3. Root Mean Squared Error (RMSE): This metric measures the average magnitude of the errors in a set of predictions. Lower values indicate better model fit.
    4. Mean Absolute Error (MAE): This metric measures the average magnitude of the errors in a set of predictions, but it is not squared like the RMSE. Lower values indicate better model fit.
    5. Akaike Information Criterion (AIC): This metric measures the quality of a model by taking into account the complexity of the model and the amount of data used to fit the model. Lower values indicate better model fit.
15. Sensitivity = 1000/(1000 + 50) = 0.95
    Specificity = 1200/(250 + 1200) = 0.83
    Precision = 1000/(1000 + 250) = 0.8
    Recall = 1000/(1000 + 50) = 0.95

Accuracy = (1000 + 1200)/(1000 + 50 + 250 +1200) = 0.88