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 & D) Lasso

7. 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

10. Adjusted R-squared is a statistical measure that assesses the goodness of fit of a regression model. It is similar to R-squared, which measures the proportion of variation in the dependent variable that is explained by the independent variables in the model. However, adjusted R-squared has an additional feature that penalizes the presence of unnecessary predictors in the model.

In regression analysis, it is possible to add more independent variables to the model, which can increase the R-squared value even if the new variables are not significant or do not contribute to the explanation of the dependent variable. The adjusted R-squared adjusts the R-squared value for the number of independent variables in the model, penalizing the inclusion of unnecessary predictors.

The formula for adjusted R-squared is:

Adjusted R-squared = 1 - [(1-R-squared)*(n-1)/(n-p-1)]

where n is the number of observations and p is the number of independent variables in the model.

The penalty term (n-p-1) in the denominator of the formula for adjusted R-squared increases as the number of independent variables in the model increases. As a result, if a model includes unnecessary predictors, the adjusted R-squared will be lower than the R-squared value, indicating that the model does not fit the data as well as it could. This encourages the use of simpler models with fewer predictors, which can improve the interpretability and generalizability of the results.

In summary, the adjusted R-squared penalizes the presence of unnecessary predictors in the model by adjusting the R-squared value for the number of independent variables, providing a more accurate measure of the model's goodness of fit.

11. The main difference between Ridge and LASSO Regression is that if ridge regression can shrink the coefficient close to 0 so that all predictor variables are retained. Whereas LASSO can shrink the coefficient to exactly 0 so that LASSO can select and discard the predictor variables that have the right coefficient of 0.

12.VIF stands for Variance Inflation Factor, which is a measure of the degree of multicollinearity between two or more predictor variables in a regression model. It quantifies how much the variance of the estimated regression coefficients is increased due to multicollinearity.

The VIF for each predictor variable is calculated by regressing it against all other predictor variables in the model and then computing the ratio of the variance of the coefficient estimate to its expected variance if the predictor variable were not correlated with the other predictors. VIF values greater than 1 indicate that the predictor variable is highly correlated with the other predictors in the model and may cause problems with the stability and interpretability of the regression coefficients.

There is no hard and fast rule for the maximum acceptable value of VIF, as it depends on the context of the data and the specific objectives of the analysis. However, as a general guideline, a VIF value of 1 indicates that there is no multicollinearity between the predictor variable and other variables, while a VIF value of 5 or higher suggests that there is a high degree of multicollinearity that may require corrective action.

A common threshold for deciding whether a predictor variable should be included in a regression model is to use a VIF cutoff value of 2 or 2.5. If a predictor variable has a VIF value above this threshold, it may be considered for removal from the model to reduce multicollinearity and improve the stability and interpretability of the regression coefficients.

In summary, VIF is a measure of the degree of multicollinearity between predictor variables in a regression model. A VIF value greater than 1 indicates the presence of multicollinearity, and a VIF value of 5 or higher suggests a high degree of multicollinearity that may require corrective action. A commonly used VIF cutoff value is 2 or 2.5 to decide whether to include a predictor variable in a regression model.

13.Scaling the data before training a machine learning model is an important step in the data preprocessing pipeline. There are several reasons why data scaling is necessary:

1.Helps improve the performance of certain machine learning algorithms: Some machine learning algorithms, such as k-nearest neighbors and support vector machines, are based on distances between data points. If the features in the data are not on the same scale, features with larger scales will dominate the distance calculations and can lead to biased results. Scaling the data to a common scale can help mitigate this issue and improve the performance of these algorithms.

2.Helps speed up convergence of optimization algorithms: Many optimization algorithms used in machine learning, such as gradient descent, are designed to work best when the data is centered around zero with a similar range. Scaling the data to a common range can help optimization algorithms converge faster and more accurately.

3.Helps prevent numerical instability: Some machine learning algorithms, such as principal component analysis and clustering, rely on eigenvalue decomposition, which can be numerically unstable if the data is not scaled properly. Scaling the data can help prevent numerical instability and improve the stability of these algorithms.

4.Facilitates comparison of feature importance: When the features in the data are on different scales, it can be difficult to compare their relative importance. Scaling the data can help make the features comparable and facilitate the interpretation of feature importance.

In summary, scaling the data before training a machine learning model is important to improve the performance of certain algorithms, speed up convergence of optimization algorithms, prevent numerical instability, and facilitate the comparison of feature importance.

14.There are several metrics that can be used to check the goodness of fit of a linear regression model. Some of the most commonly used metrics are:

R-squared (R^2),Adjusted R-squared,Mean Squared Error,Root Mean Squared Error,Mean Absolute Error ,Residual plots

15. The confusion matrix is as follows

Actual/Predicted	True	False
True	1000	50
False	250	1200

Using this confusion matrix, we can calculate the following evaluation metrics:

1.Sensitivity (recall): the proportion of actual positives (True) that are correctly identified by the model.

Sensitivity = true positives / (true positives + false negatives) = 1000 / (1000 + 250) = 0.8

2.Specificity: the proportion of actual negatives (False) that are correctly identified by the model.

Specificity = true negatives / (true negatives + false positives) = 1200 / (1200 + 50) = 0.96

3.Precision: the proportion of predicted positives that are correctly identified by the model.

Precision = true positives / (true positives + false positives) = 1000 / (1000 + 50) = 0.95

4.Accuracy: the proportion of correct predictions made by the model.

Accuracy = (true positives + true negatives) / (true positives + false positives + true negatives + false negatives) = (1000 + 1200) / (1000 + 50 + 250 + 1200) = 0.88

In summary, the sensitivity is 0.8, specificity is 0.96, precision is 0.95, recall is 0.8, and accuracy is 0.88.