# In Q1 to Q5, only one option is correct, Choose the correct option:

1. In which of the following you can say that the model is overfitting?
   1. High R-squared value for train-set and High R-squared value for test-set.
   2. Low R-squared value for train-set and High R-squared value for test-set.
   3. High R-squared value for train-set and Low R-squared value for test-set.
   4. None of the above
2. Which among the following is a disadvantage of decision trees?
   1. Decision trees are prone to outliers.
   2. Decision trees are highly prone to overfitting.
   3. Decision trees are not easy to interpret
   4. None of the above.
3. Which of the following is an ensemble technique?
   1. SVM B) Logistic Regression

C) Random Forest D) Decision tree

1. Suppose you are building a classification model for detection of a fatal disease where detection of the disease is most important. In this case which of the following metrics you would focus on?
   1. Accuracy B) Sensitivity

C) Precision D) None of the above.

1. The value of AUC (Area under Curve) value for ROC curve of model A is 0.70 and of model B is

0.85. Which of these two models is doing better job in classification?

A) Model A B) Model B

C) both are performing equal D) Data Insufficient

# In Q6 to Q9, more than one options are correct, Choose all the correct options:

1. Which of the following are the regularization technique in Linear Regression??
   1. Ridge B) R-squared

C) MSE D) Lasso

1. Which of the following is not an example of boosting technique?
   1. Adaboost B) Decision Tree

C) Random Forest D) Xgboost.

1. Which of the techniques are used for regularization of Decision Trees?
   1. Pruning B) L2 regularization

C) Restricting the max depth of the tree D) All of the above

1. Which of the following statements is true regarding the Adaboost technique?
   1. We initialize the probabilities of the distribution as 1/n, where n is the number of data-points
   2. A tree in the ensemble focuses more on the data points on which the previous tree was not performing well
   3. It is example of bagging technique
   4. None of the above

# Q10 to Q15 are subjective answer type questions, Answer them briefly.

1. Explain how does the adjusted R-squared penalize the presence of unnecessary predictors in the model?

**Answer:**

Every time you add a independent variable to a model, the R-squared increases, even if the independent variable is insignificant. It never declines. Whereas Adjusted R-squared increases only when independent variable is significant and affects dependent variable

1. Differentiate between Ridge and Lasso Regression.

**Answer:**

|  |  |
| --- | --- |
| **RIDGE** | **LASSO** |
| Ridge Regression is a type of regression algorithm usually considered when there is a high correlation between the independent variables or model parameters. | Lasso Regression algorithm utilizes L1 regularization technique It is taken into consideration when there are more number of features because it automatically performs feature selection. |
| In ridge regression, the penalty is equal to the sum of the squares of the coefficients | In the Lasso, penalty is considered to be the sum of the absolute values of the coefficients |

1. What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modelling?

**Answer:**

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. A VIF of **four or below** is deemed suitable. However, as VIF increases, the less reliable your regression results are going to be.

1. Why do we need to scale the data before feeding it to the train the model?

**Answer:**

To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features.

1. What are the different metrics which are used to check the goodness of fit in linear regression?

**Answer:**

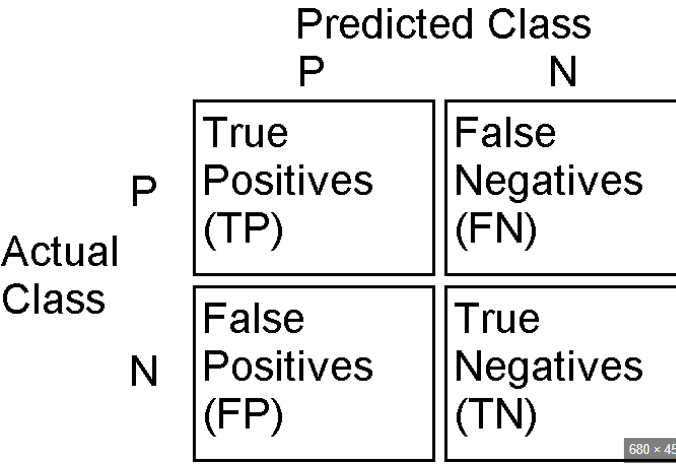
* Mean Squared Error (MSE)
* Root Mean Squared Error (RMSE)
* Mean Absolute Error (MAE)

1. From the following confusion matrix calculate sensitivity, specificity, precision, recall and accuracy.

|  |  |  |
| --- | --- | --- |
| Actual/Predicted | True | False |
| True | 1000 | 50 |
| False | 250 | 1200 |

**Answer:**

**From the above Confusion Matrix**

****

**TP= 1000**

**FP= 250**

**FN= 50**

**TN = 1200**

**SENSITIVITY/RECALL: = (TP)/(TP+FN)**

1000/(1000+50) = 1000/1050 = **0.95**

**SPECIFITY = TN / (TN + FP)**

1200/(1200+250) = 1200/1450 **= 0.83**

**PRECISION = (TP)/(TP+FP)**

1000/(1000+250) = 1000/1250 **= 0.8**

**ACCURACY = (TP+TN)/(TP+TN+FP+FN)**

(1000+1200)/(1000+1200+250+50) = 2200/2500 **= 0.88**