# 1 In Q1 to Q7, only one option is correct, Choose the correct option:

1. The value of correlation coefficient will always be:
   1. between 0 and 1 B) greater than -1

C) between -1 and 1 D) between 0 and -1

Answer- C) between -1 and 1

1. Which of the following cannot be used for dimensionality reduction?
   1. Lasso Regularisation B) PCA

C) Recursive feature elimination D) Ridge Regularisation

Answer- C) Recursive feature elimination

1. Which of the following is not a kernel in Support Vector Machines?
   1. linear B) Radial Basis Function

C) hyperplane D) polynomial

Answer- Linear

1. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?
   1. Logistic Regression B) Naïve Bayes Classifier

C) Decision Tree Classifier D) Support Vector Classifier

Answer- A) Logistic Regression

1. In a Linear Regression problem, ‘X’ is independent variable and ‘Y’ is dependent variable, where ‘X’ represents weight in pounds. If you convert the unit of ‘X’ to kilograms, then new coefficient of ‘X’ will be?

(1 kilogram = 2.205 pounds)

* 1. 2.205 × old coefficient of ‘X’ B) same as old coefficient of ‘X’

C) old coefficient of ‘X’ ÷ 2.205 D) Cannot be determined

Answer – D) Cannot be determined

1. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?
   1. remains same B) increases

C) decreases D) none of the above

Answer- B) Increases

1. Which of the following is not an advantage of using random forest instead of decision trees?
   1. Random Forests reduce overfitting
   2. Random Forests explains more variance in data then decision trees
   3. Random Forests are easy to interpret
   4. Random Forests provide a reliable feature importance estimate

Answer- C) Random Forest are easy to interpret

# In Q8 to Q10, more than one options are correct, Choose all the correct options:

1. Which of the following are correct about Principal Components?
   1. Principal Components are calculated using supervised learning techniques
   2. Principal Components are calculated using unsupervised learning techniques
   3. Principal Components are linear combinations of Linear Variables.
   4. All of the above

Answer- B) AND C)

1. Which of the following are applications of clustering?
   1. Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index
   2. Identifying loan defaulters in a bank on the basis of previous years’ data of loan accounts.
   3. Identifying spam or ham emails
   4. Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.

Answer- B) and C)

1. Which of the following is(are) hyper parameters of a decision tree?
   1. max\_depth B) max\_features

C) n\_estimators D) min\_samples\_leaf

Answer- A)- max\_depth ,B)- max\_features and D) min\_samples\_leaf



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

1. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.

Answer- An outlier is **an observation that lies an abnormal distance from other values in a random sample from a population**. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal.

The IQR **describes the middle 50% of values when ordered from lowest to highest**. To find the interquartile range (IQR), ​first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1.

1. What is the primary difference between bagging and boosting algorithms?

Answer - **Bagging**is a homogeneous weak learners’ model that learns from each other independently in parallel and combines them for determining the model average.

**Boosting**is also a homogeneous weak learners’ model but works differently from Bagging. In this model, learners learn sequentially and adaptively to improve model predictions of a learning algorithm.

1. What is adjusted R2 in linear regression. How is it calculated?

Answer - Adjusted R2 is **a corrected goodness-of-fit (model accuracy) measure for linear models**. It identifies the percentage of variance in the target field that is explained by the input or inputs. R2 tends to optimistically estimate the fit of the linear regression.

**R 2 = 1 − sum squared regression (SSR) total sum of squares (SST) , = 1 − ∑ ( y i − y i ^ ) 2 ∑ ( y i − y ¯ ) 2** . The sum squared regression is the sum of the residuals squared, and the total sum of squares is the sum of the distance the data is away from the mean all squared.

1. What is the difference between standardisation and normalisation?

Answer- Normalisation - Normalization of data is a type of Feature scaling and is only required when the data distribution is unknown or the data doesn't have Gaussian Distribution. This type of scaling technique is used when the data has a diversified scope and the algorithms on which the data are being trained do not make presumptions about the data distribution such as Artificial Neural Network.

Standardization- Standardized data is usually preferred when the data is being used for multivariate analysis i.e. when we want all the variables of comparable units. It is usually applied when the data has a bell curve i.e. it has gaussian distribution. No this isn't always true but is considered more effective when applied to Gaussian distribution. This technique comes in handy when the data has varying ratios and the algorithms used, make assumptions about the data distribution like Logistic Regression, Linear Discriminant Analysis, etc.

1. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.

Answer- Cross Validation in Machine Learning is a great technique to deal with overfitting problem in various algorithms. Instead of training our model on one training dataset, we train our model on many datasets. Below are some of the advantages and disadvantages of Cross Validation in Machine Learning:  
  
**Advantages of Cross Validation**  
  
**1. Reduces Overfitting:** In Cross Validation, we split the dataset into multiple folds and train the algorithm on different folds. This prevents our model from overfitting the training dataset. So, in this way, the model attains the generalization capabilities which is a good sign of a robust algorithm.  
  
**Note:**Chances of overfitting are less if the dataset is large. So, Cross Validation may not be required at all in the situation where we have sufficient data available.  
  
**2. Hyperparameter Tuning:** Cross Validation helps in finding the optimal value of hyperparameters to increase the efficiency of the algorithm.

**Disadvantages of Cross Validation**  
  
**1. Increases Training Time:** Cross Validation drastically increases the training time. Earlier you had to train your model only on one training set, but with Cross Validation you have to train your model on multiple training sets.   
  
For example, if you go with 5 Fold Cross Validation, you need to do 5 rounds of training each on different 4/5 of available data. And this is for only one choice of hyperparameters. If you have multiple choice of parameters, then the training period will shoot too high.  
  
**2. Needs Expensive Computation:** Cross Validation is computationally very expensive in terms of processing power required.