# 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

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

C) Recursive feature elimination D) Ridge Regularisation

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

C) hyperplane D) polynomial

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

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

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

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

# 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
2. 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.
3. 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

# 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.
2. What is the primary difference between bagging and boosting algorithms?
3. What is adjusted R2 in linear regression. How is it calculated?
4. What is the difference between standardisation and normalisation?
5. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.

11. Any data points that differ from the overall pattern of the dataset is known as outliers.

For example, marks of a student maybe recorded as 800 instead of 80. This is an error in the data and should be eliminated from the dataset. However, all outliers should not be discarded, some outlier can actually assist with identifying data that is unique from the others. For example, I can help identify insurance fraud., as fraud transactions stand out from normal transactions.

Outliers can deeply impact mean and standard deviation of the dataset which can statistically produce erroneous results. Most machine learning algorithms require the absence of outlier to produce accuracy, hence it is desirable to detect and remove outliers.

IQR is used to **measure variability** by segregating a data set into four equal parts, known as quartiles - Q1, Q2, Q3.

* Q1 represents the 25th percentile
* Q2 represents the 50th percentile
* Q3 represents the 75th percentile

IQR is the range between the first and the third quartiles(Q1 and Q3: *IQR = Q3 – Q1)*. The data points which fall below *Q1 – 1.5 IQR* or above *Q3 + 1.5 IQR* are outliers.

**Example:**  
If you look at the following data: 6, 4, 1, 5, 3, 60.

These values represent the number of chocolates eaten after lunch. In this case, 60 is clearly an outlier.

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

The primary difference between the two is that the objective of bagging is to decrease variance, not bias, whereas for boosting the objective is to decrease bias, not variance. If the classifier is unstable (high variance), then bagging is used. In case the classifier is steady and straightforward (high bias), then we need to apply boosting. Each is constructed independently in case of bagging, whereas for boosting, the models will be impacted by the performance of the previous model.

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

# 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. It always increases as the number of effects are included in the model. Adjusted R2 attempts to correct for this overestimation. Adjusted R2 might decrease if a specific effect does not improve the model.

Adjusted R squared is calculated by dividing the residual mean square error by the total mean square error (which is the sample variance of the target field). The result is then subtracted from 1.

Adjusted R2 is always less than or equal to R2. A value of 1 indicates a model that perfectly predicts values in the target field. A value that is less than or equal to 0 indicates a model that has no predictive value. In the real world, adjusted R2 lies between these values.

Unlike R-squared, the **Adjusted R-squared** would penalize you for adding features which are not useful for predicting the target. It takes into account the number of **independent variables** used for predicting the target variable.