# 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.

Ans: Outliers are those data points that are significantly different from the rest of the dataset. They are often abnormal observations that skew the data distribution, and arise due to inconsistent data entry, or erroneous observations.

**Significance of outliers:**

* Outliers badly affect mean and standard deviation of the dataset. These may statistically give erroneous results.
* Most machine learning algorithms do not work well in the presence of outlier. So it is desirable to detect and remove outliers.
* Outliers are highly useful in anomaly detection like fraud detection where the fraud transactions are very different from normal transactions.

**Inter Quartile Range (IQR) Method:**

IQR is used to **measure variability** by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts.

* Q1 represents the 25th percentile of the data.
* Q2 represents the 50th percentile of the data.
* Q3 represents the 75th percentile of the data.

If a dataset has *2n / 2n+1* data points, then  
Q1 = median of the dataset.  
Q2 = median of n smallest data points.  
Q3 = median of n highest data points.

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

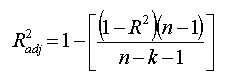
1. What is the primary difference between bagging and boosting algorithms?

Ans: Difference between Bagging and Boosting:

|  |  |
| --- | --- |
| **Bagging** | **Boosting** |
| Bagging attempts to tackle the over-fitting issue. | Boosting tries to reduce bias. |
| If the classifier is unstable (high variance), then we need to apply bagging. | If the classifier is steady and straightforward (high bias), then we need to apply boosting |
| Every model receives an equal weight. | Models are weighted by their performance. |
| Objective to decrease variance, not bias. | Objective to decrease bias, not variance. |
| It is the easiest way of connecting predictions that belong to the same type. | It is a way of connecting predictions that belong to the different types. |
| Every model is constructed independently. | New models are affected by the performance of the previously developed model. |
| |  |  | | --- | --- | | Different training data subsets are selected using row sampling with replacement and random sampling methods from the entire training dataset. |  | | Every new subset contains the elements that were misclassified by previous models. |
| |  |  | | --- | --- | | In this base classifiers are trained parallelly. |  | |  |  |  | | In this base classifiers are trained sequentially. |
| |  |  | | --- | --- | | Example: The Random forest model uses Bagging. |  | | Example: The AdaBoost uses Boosting techniques |

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

**Ans**: R2 shows how well terms (data points) fit a curve or line. Adjusted R2 also indicates how well terms fit a curve or line, but adjusts for the number of terms in a model. If you add more and more useless variables to a model, adjusted r-squared will decrease. If you add more **useful** variables, adjusted r-squared will increase.

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 will always be less than or equal to R2. The formula is:  
[](https://www.statisticshowto.com/wp-content/uploads/2013/09/r-squared-adjusted.jpg)  
where:

* N is the number of points in your data sample.
* K is the number of independent regressors, i.e. the number of [variables](https://www.statisticshowto.com/probability-and-statistics/types-of-variables/) in your model, excluding the [constant](https://www.calculushowto.com/constant-term-definition/).

If you already know R2 then it’s a fairly simple formula to work.

1. What is the difference between standardisation and normalisation?

**Ans: Difference between Normalization and standardization:**

|  |  |
| --- | --- |
| Normalization | Standardization |
| |  |  | | --- | --- | | * Minimum and maximum value of features are used for scaling |  | | Mean and standard deviation is used for scaling. |
| |  |  | | --- | --- | | * It is used when features are of different scales. |  | | It is used when we want to ensure zero mean and unit standard deviation. |
| |  |  | | --- | --- | | * Scales values between [0, 1] or [-1, 1]. |  | | It is not bounded to a certain range. |
| |  |  | | --- | --- | | * It is really affected by outliers. |  | | It is much less affected by outliers. |
| |  |  | | --- | --- | | * Scikit-Learn provides a transformer called MinMaxScaler for Normalization. | . | | Scikit-Learn provides a transformer called StandardScaler for standardization |
| |  | | --- | | * This transformation squishes the n-dimensional data into an n-dimensional unit hypercube. | | It translates the data to the mean vector of original data to the origin and squishes or expands. |
| |  |  | | --- | --- | | * It is useful when we don’t know about the distribution |  | | It is useful when the feature distribution is Normal or Gaussian. |
| * It is an often called as Scaling Normalization  |  |  | | --- | --- | |  |  | | It is a often called as Z-Score Normalization. |

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

**Ans**: Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model.

It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited.

The disadvantage of this method is that the training algorithm has to be rerun from scratch k times, which means it takes k times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set k different times.