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

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

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

Ans] C

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

C) Recursive feature elimination D) Ridge Regularisation

Ans)B

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

C) hyperplane D) polynomial

Ans) C

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

C) Decision Tree Classifier D) Support Vector Classifier

Ans)A

* 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)

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

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

Ans) C

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

C) decreases D) none of the above

Ans) B

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

Ans) b

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

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

Ans) b

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

Ans) A and D

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

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

Ans) D



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

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

Outliers:

The outliers may suggest experimental errors, variability in a measurement, or an anomaly. The age of a person may wrongly be recorded as 200 rather than 20 Years. Such an outlier should definitely be discarded from the dataset.

However, not all outliers are bad. Some outliers signify that data is significantly different from others. For example, it may indicate an anomaly like bank fraud or a rare disease.

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.

* What is the primary difference between bagging and boosting algorithms?

Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions. Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.

* 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.

* What is the difference between standardisation and normalisation?

In Normalisation, the change in values is that they are at a standard scale without distorting the differences in the values. Whereas, Standardisation assumes that the dataset is in Gaussian distribution and measures the variable at different scales, making all the variables equally contribute to the analysis.

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

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.