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

1. The value of correlation coefficient will always be:

C) between -1 and 1

2. Which of the following cannot be used for dimensionality reduction?

C) Recursive feature elimination

3. Which of the following is not a kernel in Support Vector Machines?

C) hyperplane

4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?

B) Naïve Bayes Classifier

5. 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?

C) old coefficient of ‘X’ ÷ 2.205

6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?

B) increases

7. Which of the following is not an advantage of using random forest instead of decision trees?

C) Random Forests are easy to interpret

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

8. Which of the following are correct about Principal Components?

B) Principal Components are calculated using unsupervised learning techniques
C) Principal Components are linear combinations of Linear Variables.

9. Which of the following are applications of clustering?

A) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index
B) Identifying loan defaulters in a bank on the basis of previous years' data of loan accounts.
C) Identifying spam or ham emails
D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.

10. Which of the following is(are) hyper parameters of a decision tree?

A) max_depth B) max_features
D) min_samples_leaf

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

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

Outliers are observations in a dataset that are significantly different from the majority of the data. They can occur due to errors, errors in data entry, or due to unexpected events or phenomena. Outliers can have a significant impact on the analysis and interpretation of the data, and it is often important to identify and handle them appropriately.

There are several methods for detecting outliers in a dataset. One such method is the Inter Quartile Range (IQR) method.

The IQR method for outlier detection is based on the interquartile range, which is a measure of the dispersion of a dataset. It is calculated as the difference between the 75th percentile (Q3) and the 25th percentile (Q1) of the data.

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

The primary difference between bagging and boosting algorithms is the way they generate the base models.

Bagging (short for Bootstrap Aggregation) is an ensemble learning method that builds multiple base models, or estimators, using bootstrapped samples of the training data. It is a type of parallel ensemble method, meaning that the base models are trained independently and in parallel.

Boosting is an ensemble learning method that builds multiple base models, or estimators, in a sequential manner. It is a type of sequential ensemble method, meaning that the base models are trained sequentially and the error of the previous models is used to weight the training of the next model.

13. What is adjusted $R_2$ in linear regression. How is it calculated?

Adjusted R2 is a measure of the goodness of fit of a linear regression model. It is an adjusted version of the R2 statistic, which is a commonly used measure of the proportion of variance in the dependent variable that is explained by the model.

Adjusted R2 is used to account for the number of variables in the model and the sample size. It penalizes models that have more variables than necessary, and it is a more reliable measure of the model's fit when the sample size is small.

Adjusted R2 is calculated as follows:

Adjusted R2 = 1 - (1 - R2) * (n - 1) / (n - p - 1)

14. What is the difference between standardisation and normalisation?

Standardization and normalization are two techniques that are used to scale the features of a dataset so that they can be compared on the same scale. They are both commonly used in machine learning to preprocess data before training a model.

Standardization is a technique that scales the features of a dataset so that they have zero mean and unit variance. It is calculated as follows:

Standardized value = (value - mean) / standard deviation

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

Cross-validation is a resampling method that is used to evaluate the performance of a machine learning model. It involves dividing the dataset into a training set and a test set, training the model on the training set, and evaluating the model on the test set. This process is repeated multiple times, and the results are averaged to give a more reliable estimate of the model's performance.There are several types of cross-validation, including k-fold cross-validation, stratified k-fold cross-validation, and leave-one-out cross-validation**.One advantage** of using cross-validation is that it helps to prevent overfitting. By training the model on different subsets of the data and evaluating it on a different subset, cross-validation helps to ensure that the model is not overly optimized to the training data and is able to generalize well to new data.**One disadvantage** of using cross-validation is that it can be computationally expensive, especially when the dataset is large. Cross-validation requires training the model multiple times, which can be time-consuming, especially when using complex models. This can be a problem when working with large datasets and when the model training time is long.