MACHINE LEARNING

1. The value of correlation coefficient will always be:
Ans : C) between -1 and 1

2. Which of the following cannot be used for dimensionality reduction?
Ans : C) Recursive feature elimination

3. Which of the following is not a kernel in Support Vector Machines?
Ans: C) hyperplane

4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?
Ans: A) Logistic Regression

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?
Ans: A) 2.205 × old coefficient of 'X'

6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?
Ans: B) increases

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

Ans: C) Random Forests are easy to interpret

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

Ans: 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?

Ans: 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?

Ans: B) max_features

D) min_samples_leaf

11. 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.

IQR is the range between the first and the third quartiles namely Q1 and Q3: IQR = Q3 – Q1.

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

Ans: 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.

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

Ans: 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.
Adjusted R squared is calculated by dividing the residual mean square error by the total mean square error.

14. What is the difference between standardisation and normalisation?

Ans: 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.

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

Ans: Cross validation is a technique for assessing how the statistical analysis generalises to an independent data set.
The purpose of cross–validation is to test the ability of a machine learning model to predict new data.
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.