WORKSHEET -4 (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-lineardecision boundaries?**
   Ans: D) Support Vector 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?**
   Ans: C) old coefficient of 'X' ÷ 2.205

6. **As we increase the number of estimators in ADABOOST Classifier, what happens tothe accuracy of the model?**
   Ans: C) decreases

7. **Which of the following is not an advantage of using random forest instead ofdecision 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

   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: A) max_depth B) max_features D) min_samples_leaf

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

    Ans: Outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In other words, they are the unusual values in a dataset.

    Inter Quartile Range (IQR) has the formula IQR=Q3-Q1. IQR is the range between first and third quartile range. The datapoints which falls below Q1-1.5 and above Q3+1.5 are outliers.

12. **What is the primary difference between bagging and boosting algorithms?**
    Ans:
    BAGGING - The result is obtained by averaging the responses of the N learners.
    BOOSTING - It assigns a second set of weights, this time for the N classifiers in order to takea weighted average of their estimates.

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

    Ans: Adjusted R2 is a modified version of Rsquared that has been adjusted for the number of predictors in the model. Adjusted Rsquared value can be calculated based on value of Rsquared. Every time you add an independent variable to a model, R squared increases, even if the independent variable is insignificant.

14. **What is the difference between standardisation and normalisation?**

    Ans: Normalization rescales the value into a range of [0,1].

    Standardization rescales data to have a mean of 0 and a std deviation of 1.

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

    Ans: Cross validation is a technique for assessing how statistical analysis generalises to an independent dataset. It is a technique for evaluating machine learning models by training several models on subsets of the available input data and evaluating them on complementary subset of data.

    Advantage –It make use of all datapoints and hence it is low bias.

    Disadvantage-It leads to higher variation in testing model as we are testing against one data point.