MACHINE LEARNING WORKSHEET – 4

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: A) linear

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: C) decreases

7. Which of the following is not an advantage of using random forest instead of decision trees?
   Ans: A) Random Forests reduce overfitting

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: B) Identifying loan defaulters in a bank on the basis of previous years' data of loan accounts.

C) Identifying spam or ham emails

10. Which of the following is(are) hyper parameters of a decision tree?
Ans: B) max_features
D) min_sample_leaf

11. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.
Ans: An Outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

The difference between Q3 and Q1 is called the Inter-Quartile Range or IQR. IQR = Q3-Q1.

Any data point less than the Lower Bound or more than the Upper Bound is considered as an outlier.
Lower Bound : (Q1 – 1.5 * IQR)
Upper Bound : (Q3 + 1.5 * IQR)

- Q1 represents the 25th Percentile of the data.
- Q2 represents the 50th Percentile of the data.
- Q3 represents the 75th Percentile of the data.

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

Bagging	**Boosting**
Bagging is the simplest way of combining predictions that belong to the same type	Boosting is a way of combining predictions that belong to the different types.
Bagging aims to decrease variance, not bias	Boosting aims to decrease bias, not variance.
Baggiing each model receives equal weight	Boosting models are weighted according to their performance.
Bagging each model is built independently	Boosting new models are influenced by performance of previously built models.
Bagging tries to solve over-fitting problem	Boosting tries to reduce bias.

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

Ans: Adjusted R squared is the 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, Rsquared increases, even if the independent variable is insignificant.

Adjusted Rsquared value can be calculated based on value of r-squared, number of independent variables(predictors), total sample size. Every time you add a independent variable to a model, the R-squared increases, even if the independent variable is insignificant. It never declines.

The adjusted R-squared is a modified version of R-squared that adjusts for the number of predictors in a regression model. It is calculated as:

Adjusted R2 = 1-[(1-R2)*(n-1)/(n-k-1)]

14. What is the difference between standardisation and normalisation?

Ans:

Normalisation	**Standardisation**
Scaling is done by the highest and the lowest values.	Scaling is done by mean and standard deviation.
It is applied when the features are of separate scales.	It is applied when we verify zero mean and unit standard deviation.
Scales range from 0 to 1	Not bounded
Affected by outliers	Less affected by outliers
It is also known as Scaling Normalization	It is also known as Z-Score

15. 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.

Advantages : Cross-Validation is a very powerful tool. It helps us better use our data, and it gives us much more information about our algorithm performance.

Disadvantages : 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.