MACHINE LEARNING

1. The value of correlation coefficient will always be:
A) between 0 and 1 B) greater than -1
C) between -1 and 1 D) between 0 and -1
ANS:- between -1 and 1

2. Which of the following cannot be used for dimensionality reduction?
A) Lasso Regularisation B) PCA
C) Recursive feature elimination D) Ridge Regularisation
ANS:- PCA

3. Which of the following is not a kernel in Support Vector Machines?
A) linear B) Radial Basis Function
C) hyperplane D) polynomial
ANS:-A)linear

4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?
A) Logistic Regression B) Naïve Bayes Classifier
C) Decision Tree Classifier D) Support Vector Classifier
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?
(1 kilogram = 2.205 pounds)
A) 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:-

6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?
A) remains same B) increases
C) decreases D) none of the above
ANS:- B) increases

7. Which of the following is not an advantage of using random forest instead of decision trees?
A) Random Forests reduce overfitting
B) Random Forests explains more variance in data then decision trees
C) Random Forests are easy to interpret
D) Random Forests provide a reliable feature importance estimate
ANS:- 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?

A)Principal Components are calculated using supervised learning techniques
B) Principal Components are calculated using unsupervised learning techniques
C) Principal Components are linear combinations of Linear Variables.
D) All of the above
ANS:- D) All of the above

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

10. Which of the following is(are) hyper parameters of a decision tree?
A) max_depth B) max_features
C) n_estimators D) min_samples_leaf
ANS:- A) max_depth

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.

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.

We can use the IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3. Any values that fall outside of this fence are considered outliers. To build this fence we take 1.5 times the IQR and then subtract this value from Q1 and add this value to Q3.

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 R_2 in linear regression. How is it calculated?

ANS:-

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 evaluating ML models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data. Use cross-validation to detect overfitting, ie, failing to generalize a pattern

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.
12. What is the primary difference between bagging and boosting algorithms?
13. What is adjusted R_2 in linear regression. How is it calculated?
14. What is the difference between standardisation and normalisation?
15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.