MACHINE LEARNING

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

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

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

   A) Lasso Regularisation B) PCA

   C) Recursive feature elimination D) Ridge Regularisation

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

   A) linear B) Radial Basis Function

   C) hyperplane D) polynomial

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

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

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

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

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

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

    C) n_estimators D) min_samples_leaf

MACHINE LEARNING

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

A value that "lies outside" (is much smaller or larger than) most of the other values in a set of data. IQR (Interquartile Range) is the difference between the third and the first quartile of a distribution (or the 75th percentile minus the 25th percentile). It is a measure of how wide our distribution is since this range contains half of the points of the dataset. It's very useful to make an idea of the shape of the distribution.

12. What is the primary difference between bagging and boosting algorithms?
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?
It measures the proportion of variation explained by only those independent variables that really help in explaining the dependent variable. It penalizes you for adding independent variable that do not help in predicting the dependent variable
Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size.

14. What is the difference between standardisation and normalisation?

Normalisation	**Standardisation**
Scaling is done by the highest and the lowest values.	Scaling is done by mean and 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
It is applied when we are not sure about the data distribution	It is used when the data is Gaussian or normally distributed

15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.
Cross-validation is a statistical method used to estimate the performance (or accuracy) of machine learning models. It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited.
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