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.
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.

Ans-11 The data which is away from +3 Std normal data taey are called Outliers.We can check outliers with the help of z-score method using box plot and with IQR method

IQR = Q3-Q1

Ans-12 Bagging tries to tacle overfitting problem and boosting tries to reduce bias from model.

Ans-13 Adjusted R^2 calculated based on R^2 squared number of indepedndent varivble to total sampl size

R^2 =COEFFICIENT of determination = EXPECTED variation / Total variation

Ans-14 Standardisation means a dataset to have a mean of 0 and std is 1 but to Normalization means a dataset each value fall between 0 and 1.

Ans-15 Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set

Advantages- More accurate, More efficient

Disadvantages- The advantage of cross validation is it is a lengthy process for every iteration