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

Ans:- between -1 and 1 (C)

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

Ans:- Ridge Regularisation (D)

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

Ans :- hyperplane (C)

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 :- Support Vector Classifier (D)

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' willbe? (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 :- same as old coefficient of 'X' (B)

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

Ans :- decreases (C)

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 : - Random Forests explains more variance in data then decision trees (B)

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 :- Principal Components are calculated using unsupervised learning techniques (B)
Principal Components are linear combinations of Linear Variables (C)

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 sugarlevels.

   Ans :- Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index (A)
   Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels. (D)

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 :- max_depth (A)
max_features (B)
n_estimators (C)
min_samples_leaf (D)

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.

11. **Ans :-**
Data points that differ from the rest of the dataset are classified as outliers. The data distribution is frequently skewed by these anomalous observations, which are frequently the result of inaccurate observations or invalid data entry. Outliers are values that are drastically out of the normal, either abnormally low or unusually high, and their presence can frequently affect the outcomes of statistical analysis performed on the dataset. This can result in models that are less effective and useful.

The difference between a distribution's third and first quartile is known as the interquartile range (or the 75th percentile minus the 25th percentile). Given that half of the dataset's points fall inside this range, it serves as a measure of how wide our distribution is. Making an idea of the shape of the distribution is quite helpful. If a point meets one of the criteria below, we classify it as an outlier:

- It is higher than the 75th percentile + 1.5 IQR (Upper Bound)
- It is lower than the 25th percentile - 1.5 IQR (Lower Bound)

An outlier is any data point that lies outside either the Lower Bound or the Upper Bound. According to the theory, a point is considered "odd" and should be classified as an outlier if it deviates too much from the 75th percentile (or from the 25th percentile). The IQR is the order of magnitude of such a distance.

Example : A sample of 15 school student was chosen at random and given a survey. They were asked, "how many teddies do you own?". The following answers were provided: 0, 2, 8, 8, 12, 0, 10, 10, 11, 12, 12, 14, 15, 40 and 25.
Here,
median= 11,
Q1= 8
Q3= 14
IQR= Q3-Q1
= 14-8
= 6

Upper Bound = Q3 + 1.5 * IQR
= 14+1.5*6
= 23

Lower Bound = Q1 - 1.5 * IQR
= 8 - 1.5*6
= -1
Therefore, outliers are 25, 40.

12. **Ans:-**
There are two techniques that are used to perform ensemble decision tree, bagging and boosting. When a decision tree's variance needs to be reduced, bagging is used. Here, the idea is to divide the training sample, which is selected at random with replacement, into a few smaller data sets. We now have an ensemble of different models as each collection of subset data is used to create its decision trees. It is more effective than using just one decision tree to use the average of all the assumptions from many trees.

In other hand, Boosting is an ensemble method for creating a set of predictors. The goal is to solve the net errors from the preceding trees when we fit successive trees, typically drawn from random samples. By consolidating the complete set, weak learners are finally transformed into better performing models whenever a certain input is misclassified by theory. This increases the weight of the input so that the next hypothesis is more likely to categories the input properly.

In Bagging, every model receives an equal weight but in Boosting, models are weighted by their performance.

In Bagging, every model is constructed independently but in Boosting, new models are affected by the performance of the previously developed model.

13. **Ans:-**
Adjusted R-Squared calculates the percentage of variance that can be explained by merely the independent variables that have a significant impact on the explanation of the dependent variable. If you include independent variables that have no bearing on predicting the dependent variable, you will be penalized. Sum of squares calculations can be used to determine it mathematically.

In the below equation, dft is the degrees of freedom n– 1 of the estimate of the population variance of the dependent variable, and dfe is the degrees of freedom n – p – 1 of the estimate of the underlying population error variance.

$$\bar{R}^2 = 1 - \frac{SS_{\text{res}}/\text{df}_e}{SS_{\text{tot}}/\text{df}_t}$$

Adjusted R-squared value can be calculated based on value of R-squared, number of independent variables (predictors), total sample size.

$$R^2_{\text{adjusted}} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where
R^2 = sample R-square
p = Number of predictors
N = Total sample size.

14. **Ans:-**

When the data does not fit to the criteria of the Gaussian distribution, normalization is a suitable technique. It can be applied to algorithms like K-Nearest Neighbors and neural networks that don't assume data distribution. On the other hand, when the dataset has a Gaussian distribution, standardization is advantageous. Because Standardization has no boundary range, it is unaffected by the dataset's outliers, unlike Normalization.

Depending on the issue and the machine learning technique, Normalization or Standardization may be applied. To employ normalization or standardization, there are no set rules. It is possible to compare the two by modelling the dataset after it has been normalized or standardized.

Normalization has scales range from 0 to 1 on other hand Standardization has no bound.

Normalization is affected by outliers but Standardization is less affected by outliers.

15.**Ans :-**

Cross validation is a method for evaluating how well a statistical analysis generalizes to a different data set. It is a method for assessing machine learning models that involves training various models on different subsets of the input data and then comparing the results. There is a good likelihood that we can identify over-fitting with easy using cross-validation. In cross-validation, we can make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate.

Advantage: - Reduces Overfitting: In Cross Validation, we split the dataset into multiple folds and train the algorithm on different folds. This prevents our model from overfitting the training dataset. So, in this way, the model attains the generalization capabilities which is a good sign of a robust algorithm.

Disadvantage: - Increases Training Time: Cross Validation drastically increases the training time. Earlier you had to train your model only on one training set, but with Cross Validation you have to train your model on multiple training sets.