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?

- Random Forests reduce overfitting
- Random Forests explains more variance in data then decision trees
- Random Forests are easy to interpret
- 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?

- Principal Components are calculated using supervised learning techniques
- Principal Components are calculated using unsupervised learning techniques
- Principal Components are linear combinations of Linear Variables.
- All of the above

9. Which of the following are applications of clustering?

- Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index
- Identifying loan defaulters in a bank on the basis of previous years' data of loan accounts.
- Identifying spam or ham emails
- 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

FLIP ROBO

Q11 to Q15 are subjective answer type questions, Answer them briefly.

11. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.

In simple terms, an outlier is an extremely high or extremely low data point relative to the nearest data point and the rest of the neighboring co-existing values in a data graph or dataset you're working with.Outliers are extreme values that stand out greatly from the overall pattern of values in a dataset or graph.Below, on the far left of the graph, there is an outlier.

One common way to find outliers in a dataset is to use the interquartile range. The interquartile range, often abbreviated IQR, is **the difference between the 25th percentile (Q1) and the 75th percentile (Q3) in a dataset**. It measures the spread of the middle 50% of values. One popular method is to declare an observation to be an outlier if it has a value 1.5 times greater than the IQR or 1.5 times less than the IQR.

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

Bagging

It is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It decreases the variance and helps to avoid overfitting. It is usually applied to decision tree methods. Bagging is a special case of the model averaging approach.

Boosting is an ensemble modeling technique that attempts to build a strong classifier from the number of weak classifiers. It is done by building a model by using weak models in series. Firstly, a model is built from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models is added.

13. What is adjusted R^2 in linear regression. How is it calculated?

R2 shows how well terms (data points) fit a curve or line. Adjusted R2 also indicates how well terms fit a curve or line, but adjusts for the number of terms in a model. If you add more and more useless variables to a model, adjusted r-squared will decrease.

Adjusted R squared is calculated by dividing the residual mean square error by the total mean square error (which is the sample variance of the target field). The result is then subtracted from 1. Adjusted R2 is always less than or equal to R2.

14. What is the difference between standardisation and normalisation?

Normalization is used when the data doesn't have Gaussian distribution whereas Standardization is used on data having Gaussian distribution.

Normalization scales in a range of [0,1] or [-1,1]. Standardization is not bounded by range.

Normalization is highly affected by outliers. Standardization is slightly affected by outliers.

Normalization is considered when the algorithms do not make assumptions about the data distribution. Standardization is used when algorithms make assumptions about the data distribution.

15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.

Cross validation is a technique for assessing how the statistical analysis generalises to an independent data set.It is a technique for evaluating machine learning models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data.

Advantage-Using cross-validation, there are high chances that we can detect over-fitting with ease

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.