**Q1.** The value of correlation coefficient will always be?

**ANS.** Between 1 and -1

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

**ANS**. Recursive feature elimination

**Q3.** Which of the following is not a kernel in Support Vector Machines?

**ANS.** RBF

**Q4.** Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?

**ANS.** SVC

**Q5.** 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?

**ANS.** 2.205 × old coefficient of ‘X’

**Q6.** As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?

**ANS.** Remains same

**Q7.** Which of the following is not an advantage of using random forest instead of decision trees?

**ANS.** Random Forests reduce overfitting

**Q8.** Which of the following are correct about Principal Components?

**ANS.** All of the above

**Q9.** Which of the following are applications of clustering?

**ANS.** B and C

**Q10.** Which of the following is(are) hyper parameters of a decision tree?

**ANS.**A and D

**Q11**. What are outliers? Explain the Inter Quartile Range(IQR) method for outlier detection.

**ANS.** Outliers are population or data which is far away from other data eg.in 2,3,4,20. 20 is an outlier.

A commonly used rule says that a data point is an outlier if it is more than 1.5.In particular, calculating the inter-quartile range (IQR) and using its multiples can help us define the outliers. The **IQR** = **Q3** - **Q1**, where **Q1** is the first quartile, and **Q3** the third quartile.

**Q12.** What is the primary difference between bagging and boosting algorithms?

**ANS. Bagging** is a way to decrease the variance **in the** prediction by generating additional data **for** training from dataset using combinations with repetitions to produce multi-sets **of** the original data. **Boosting** is an iterative technique which adjusts the weight **of** an observation based on the last classification.

**Q13.** What is adjusted R2 in logistic regression. How is it calculated?

**ANS.** The **adjusted R**-**squared** is a modified version of **R**-**squared** that has been **adjusted** for the number of predictors in the model. The **adjusted R**-**squared** increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. **Adjusted R-squared** value can be **calculated** based on value of **r-squared**, number of independent variables (predictors), total sample size. Every time you add a independent variable to a model, the **R-squared** increases, even if the independent variable is insignificant.

**Q14.** What is the difference between standardisation and normalisation?

**ANS.** The terms **normalization** and **standardization** are sometimes used interchangeably, but they usually refer to **different** things. **Normalization** usually means to scale a variable to have a values  **between** 0 and 1, while **standardization** transforms data to have a mean **of** zero and a standard deviation **of** 1.

**Q15**. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation

**ANS**. **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 . The **advantage** of this method is that the proportion of the **validation** or training split is not dependent on the number of folds (K-**fold** test). However, there is **a disadvantage** as well. There are chances that you might miss out some observations whereas you might select some observations more than once.