# WORKSHEET 04 MACHINE LEARNING

## ANSWER KEY

1. (C) between -1 and 1
2. (B) PCA
3. (A) linear
4. (A) Logistic Regression
5. (C) old coefficient of ‘X’ ÷ 2.205
6. (B) increases
7. (A) Random Forests reduce overfitting
8. (D) All of the above
9. (B) Identifying loan defaulters in a bank on the basis of previous years’ data of loan accounts.
   (C) Identifying spam or ham emails.

10. (A) max_depth (B) max_features

11. An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

IQR is the range between the first and the third quartiles namely Q1 and Q3: IQR = Q3 – Q1. The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers.

**Example:**

Assume the data 6, 2, 1, 5, 4, 3, 50. If these values represent the number of chapatis eaten in lunch, then 50 is clearly an outlier.

12.

| **Bagging** | **Boosting** |
|---|---|
| Various training data subsets are randomly drawn with replacement from the whole training dataset. | Each new subset contains the components that were misclassified by previous models. |
| Bagging attempts to tackle the over-fitting issue. | Boosting tries to reduce bias. |

13. 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. It never declines.

14. In Normalisation, the change in values is that they are at a standard scale without distorting the differences in the values. Whereas, Standardisation assumes that the

dataset is in Gaussian distribution and measures the variable at different scales, making all the variables equally contribute to the analysis.

15. 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. In cross-validation, you make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate.

An advantage of using this method is that we make use of all data points and hence it is low bias. The major drawback of this method is that it leads to higher variation in the testing model as we are testing against one data point.

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