Worksheet 4 ML answers

1. C) The value of the correlation coefficient will always be between -1 and 1
2. C) Recursive feature elimination
3. A) Linear
4. A) Logistic Regression
5. D) Cannot be determined
6. B) Increases
7. C) Random forests are easy to interpret
8. B) Principal components are calculated using unsupervised learning techniques

   C) Principal components are linear combinations of linear variables
9. 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.

   D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.
10. A) max_depth

    D) min_samples_leaf

11. Outlier is that abnormal part of a distinct data which lies at a point that is very uncommon and unnatural when it comes to the various other data points that are present in the data. IQR or the inter quantile range is the difference between Q1 and Q3.

    Data points that fall below Q1-1.5IQR and above Q3+1.5IQR are considered to be outliers.

12. The primary difference between Bagging and Boosting is that,

    Bagging is the simplest way of combining the predictions that belong to the same type and also focuses on decreasing variance not bias whereas in Boosting it combines predictions that belong to different data type and focuses on decreasing bias not variance.

13. R2 is the measure of accuracy of a linear regression model. It can be calculated by using the value of r squared, number of independent variables and total sample size.

14. In normalization the values are at a standard scale without distorting the difference in values. Whereas in standardization it assumes that the data set is in gaussian distribution and therefore makes all the variables contribute to the analysis equally. Normalization is more sensitive to outliers compared to standardization.

15. Cross validation is a statistical algorithm or method that evaluates and learns about the data by comparing the data into two segments. One segment of data is used to learn or train about the model whereas the other segment helps in validating the data.

Cross validation helps in finding the optimal value of the hyperparameters which thereby increases the efficiency of the model. Whereas the training time of the model increases drastically in cross validation making it as a big disadvantage of this model.