Machine Learning Worksheet 4

1. C)
2. C)
3. C)
4. A)
5. B)
6. D)
7. C)
8. A) and C)
9. A) and D)
10. A), B) and D)
11. An *outlier* is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations.
    In the **IQR Method,** we use the system of percentiles to define quartiles in the data.
    Q1 represents the 25th percentile of the data.
    Q2 represents the 50th percentile of the data.
    Q3 represents the 75th percentile of the data.
    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.

12. Bagging means that you take bootstrap samples (with replacement) of your data set and each sample trains a (potentially) weak learner. Boosting, on the other hand, uses all data to train each learner, but instances that were misclassified by the previous learners are given more weight so that subsequent learners give more focus to them during training.

13. Adjusted R-Squared measures the proportion of variation explained by only those independent variables that really help in explaining the dependent variable. Adjusted R-Squared can be calculated mathematically in terms of sum of squares. The only difference between R-square and Adjusted R-square equation is degree of freedom.
    Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size.

14. **Standardization**
    Standardization (or **Z-score normalization**) is the process of rescaling the features so that they'll have the properties of a Gaussian distribution with
    μ=0 and σ=1

where μ is the mean and σ is the standard deviation from the mean; standard scores (also called ***z*** scores) of the samples are calculated as follows: z= (x-u/σ)
Normalization often also simply called **Min-Max scaling** basically shrinks the range of the data such that the range is fixed between 0 and 1 (or -1 to 1 if there are negative values). It works better for cases in which the standardization might not work so well. If the distribution is not Gaussian or the standard deviation is very small, the min-max scaler works better.

15. **Cross validation** is a statistical method or a resampling procedure used to evaluate the skill of machine learning models on a limited data sample."
Advantages of Cross Validation

1. 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.

2. Hyperparameter Tuning: Cross Validation helps in finding the optimal value of hyperparameters to increase the efficiency of the algorithm.
Disadvantages of Cross Validation

1. 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.

2. Needs Expensive Computation: Cross Validation is computationally very expensive in terms of processing power required.