MACHINE LEANING- WORKSHEET – 4

1. Question -1 - Answer - **C**
2. Question -2 - Answer - **B**
3. Question -3 - Answer - **A**
4. Question -4 - Answer - **A**
5. Question -5 - Answer - **A**
6. Question -6 - Answer - **B**
7. Question -7 - Answer - **A**
8. Question -8 - Answer - **B,C**
9. Question -9 - Answer - **A,B,D**
10. Question -10 - Answer - **A ,D**

11. **Outliers** are an important part of a dataset. They can hold useful information about your data.

    Outliers can give helpful insights into the data you're studying, and they can have an effect on statistical results. This can potentially help you discover inconsistencies and detect any errors in your statistical processes.

    An outlier is a piece of data that is an abnormal distance from other points. In other words, its data that lies outside the other values in the set. If you had Pinocchio in a class of children, the length of his nose compared to the other children would be an outlier.

    INTER QUARTILE METHOD:

    Any set of data can be described by its five-number summary. These five numbers, which give you the information you need to find patterns and outliers, consist of (in ascending order):

- The minimum or lowest value of the dataset
- The first quartile Q1, which represents a quarter of the way through the list of all data
- The median of the data set, which represents the midpoint of the whole list of data
- The third quartile Q3, which represents three-quarters of the way through the list of all data
- The maximum or highest value of the data set.

12.

SR. NO.	BAGGING	BOOSTING
1.	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.
2.	Bagging attempts to tackle the over-fitting issue.	Boosting tries to reduce bias.
3.	If the classifier is unstable (high variance), then we need to apply bagging.	If the classifier is steady and straight forward, then we need to apply boosting.
4.	Very model receives an equal weight.	Models are weighted by their performance.
5.	Objective to decrease variance, not bias.	Objective to decrease bias, not variance.
6.	It is the easiest way of connecting predictions that belong to the same type.	It is a way of connecting predictions that belong to the different types.

13. Adjusted R2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs. R2 tends to optimistically estimate the fit of the linear regression.

Adjusted R Squared: [1 (1-R) 2 * (N-1)]
Formula [(N-K-1)]

Using Correlation Coefficient:

Correlation Coefficient = Σ [(X – Xm) * (Y – Ym)] / √ [Σ (X – Xm)2 * Σ (Y – Ym)2]\

Where:

X – Data points in data set X
Y – Data points in data set Y
Xm– Mean of data set X
Ym– Mean of data set Y

So
R2 = (Correlation Coefficient) 2
Adjusted R Squared = 1 – [((1 – R2) * (n – 1)) / (n – k – 1)]

Where:

n – Number of points in your data set.
k – Number of independent variables in the model, excluding the constant

2. Using Regression outputs

R2 = Explained Variation / Total Variation

R2 = MSS / TSS

R2= (TSS – RSS) / TSS

Where:

TSS – Total Sum of Squares = Σ (Yi – Ym) 2
MSS – Model Sum of Squares = Σ (Y^ – Ym) 2
RSS – Residual Sum of Squares =Σ (Yi – Y^) 2

Adjusted R Squared = 1 – [((1 – R2) * (n – 1)) / (n – k – 1)]

14.

Normalisation	Standardisation
Scaling is done by the highest and the lowest values.	Scaling is done by mean and standard deviation.
It is applied when the features are of separate scales.	It is applied when we verify zero mean and unit standard deviation.
Scales range from 0 to 1	Not bounded
Affected by outliers	Less affected by outliers
It is applied when we are not sure about the data distribution	It is used when the data is Gaussian or normally distributed
It is also known as Scaling Normalization	It is also known as Z-Score

15. Cross-validation is a statistical method used to estimate the performance of machine learning models. It is used to protect against over fitting 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 of the data, run the analysis on each fold, and then average the overall error estimate. Cross-Validation is one of the key topics around testing your learning models. Although the subject is widely known, I still find some misconceptions cover some of its aspects. When we train a model, we split the dataset into two main sets: training and testing.

Advantages of Cross Validation:

1. Reduces over fitting: In Cross Validation, we split the dataset into multiple folds and train the algorithm on different folds. This prevents our model from over fitting the training dataset. So, in this way, the model attains the generalization capabilities which are a good sign of a robust algorithm.

2. Hyper parameter Tuning**:** Cross Validation helps in finding the optimal value of hyper parameters 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.