MACHINE LEARNING WORKSHEET - 1

Ans 1 - C

Ans 2 - C

Ans 3 - C

Ans 4 - B

Ans 5 - C

Ans 6 - B

Ans 7 - C

Ans 8 - A,B

Ans 9 - A,D

Ans 10 – A,B,D

Ans 11 –

An outlier in a dataset is a data point that lies outside the overall pattern in the data distribution of the dataset.

IQR – Inter Quartile Range method is used to measure variability by dividing a dataset into quartiles. The data is sorted and split into 4 equal parts.

* Q1 represent the 25th percentile of data.
* Q2 represent the 50th percentile of data.
* Q3 represent the 75th percentile of data.

IQR = Q3 – Q1

The data points which fall below Q1-1.5IQR or above Q3+1.5IQR are the outliers.

Ans 12 –

BAGGING:

* The N learners are built independently for bagging.
* The final decision is made by equally weighted averaging the N learners.
* Samples drawn from the training dataset are replaced back into training set.

BOOSTING:

* The Boosting tries to add new models that do well where previous models fail.
* The final decision is made by weighted averaging the N learners while giving more weight to those with better performance on training data.
* Samples drawn from training dataset are not replaced back into training set.

Ans 13 –

The adjusted R-squared is used to compare the goodness-of-fit for models that contains different numbers of independent variables. Adjusted R-squared can be calculated based on the value of R-squared, number of predictors and total sample size.

Adjusted R-squared = 1 – (1 – R-Squared) (N – 1) / (N – p – 1)

Where,

R-Squared = R- Squared ,

P = number of predictors,

N = Total sample size.

Ans 14 –

* Normalization is a feature scaling technique where values are rescaled so that they end up ranging between 0 and 1.
* Standardization is a feature scaling technique where values are rescaled so that they have mean value of 0 and standard deviation of 1.

Ans 15 –

Cross – Validation is a technique in which we train the model using a subset of the dataset and then test the model using the complimentary subset of dataset.

ADVANTAGES:

* It reduces over-fitting.
* It helps in finding the best value of hyperparameters.

DISADVANTAGES:

* High training time.
* It is computationally expensive.