**MACHINE LEARNING – WORKSHEET 4**

Solution:1 – C

Solution:2 – B

Solution:3 – C

Solution:4 – A

Solution:5 – C

Solution:6– D

Solution:7 – C

Solution:8 – A & C

Solution:9 – C & D

Solution:10 – A & B & D

**Solution:11** – outliers:- The outliers may suggest experimental errors, variability in a measurement, or an anomaly. The age of a person may wrongly be recorded as 200 rather than 20 Years. Such an outlier should definitely be discarded from the dataset.

However, not all outliers are bad. Some outliers signify that data is significantly different from others. For example, it may indicate an anomaly like bank fraud or a rare disease.

IQR:- IQR is used to measure variability by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts.

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.

**Solution:12** – Difference between Bagging and Boosting:-

1. Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions.
2. Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.
3. In Bagging, each model receives an equal weight. In Boosting, models are weighed based on their performance.
4. Models are built independently in Bagging. New models are affected by a previously built model’s performance in Boosting.

**Solution:13** – R-squared (R2) is an important statistical measure which is a regression model that represents the proportion of the difference or variance in statistical terms for a dependent variable which can be explained by an independent variable or variables. In short, it determines how well data will fit the regression model.

Formula:-

R Squared=r^2

Where:

**r = n (∑xy) – ∑x ∑y / √ [n\* (∑x2 – (∑x)2)] \* [n\* (∑y2 – (∑y)2)]**

r = The Correlation coefficient

n = number in the given dataset

x = first variable in the context

y = second variable

**Solution:14** –

| S.NO. | Normalisation | Standardisation |
| --- | --- | --- |
| 1. | Minimum and maximum value of features are used for scaling | Mean and standard deviation is used for scaling. |
| 2. | It is used when features are of different scales. | It is used when we want to ensure zero mean and unit standard deviation. |
| 3. | Scales values between [0, 1] or [-1, 1]. | It is not bounded to a certain range. |
| 4. | It is really affected by outliers. | It is much less affected by outliers. |
| 5. | Scikit-Learn provides a transformer called MinMaxScaler for Normalization. | Scikit-Learn provides a transformer called StandardScaler for standardization. |
| 6. | It is useful when we don’t know about the distribution | It is useful when the feature distribution is Normal or Gaussian. |
| 7. | It is a often called as Scaling Normalization | It is a often called as Z-Score Normalization. |

**Solution 15:** In machine learning, we couldn’t fit the model on the training data and can’t say that the model will work accurately for the real data. For this, we must assure that our model got the correct patterns from the data, and it is not getting up too much noise. For this purpose, we use the cross-validation technique.

Cross-Validation:

Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

Advantage of Cross validation:

**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.

Disadvantage of Cross validation:

**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.

For example, if you go with 5 Fold Cross Validation, you need to do 5 rounds of training each on different 4/5 of available data. And this is for only one choice of hyperparameters. If you have multiple choice of parameters, then the training period will shoot too high.