**MACHINE LEARNING ASSIGNMENT WORKSHEET-4**

**Q-1:- (C) Between -1 and 1.**

**Q-2:- (C)**  **Recursive Feature Elimination.**

**Q-3:- (A) Linear.**

**Q-4:- (A) Logestic Regression.**

**Q-5:- Doubt in this.**

**Q-6:- (B) Increases.**

**Q-7:- (B) Random Forests explains more variance in data then**

**decision trees.**

**Q-8:-(B) Principal Components are calculated using unsupervised learning techniques.**

**(C) Principal Components are linear combinations of Linear Variables.**

**Q-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.**

**C) Identifying spam or ham emails.**

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

**Q-10:- (A) max\_depth.**

**(D) min\_samples\_leaf.**

**Q-11:-** An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

The interquartile range defines the difference between the third and the first quartile. Quartiles are the partitioned values that divide the whole series into 4 equal parts. There are 3 quartiles, the First Quartile is denoted by Q1known as the lower quartile, the second Quartile is denoted by Q2 and the third Quartile is denoted by Q3 known as the upper quartile. Therefore, the interquartile range is equal to the

upper quartile - lower quartile.

**Q-12:-** The main difference between the two is:-

**Bagging**: It is a ensemble technique and a homogeneous weak learners’ model that learns from each other independently in parallel and combines them for determining the model average.

**Boosting**: It is also a ensemble technique and a homogeneous weak learners’ model but works differently from Bagging. In this model, learners learn sequentially and adaptively to improve model predictions of a learning algorithm.

**Q-13:-**

Adjusted R2 measures the proportion of variation explained by only those independent variables that really help in explaining the dependent variable. It penalizes you for adding independent variable that do not help in predicting the dependent variable.

Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size.

Adjusted R2 :- **1- (1-R2) (N-1)**

**N-p-1**

**Q-14:-**

**Normalization**means rescaling the values into a range of [0,1].

**Standardization** means rescaling the data to have a mean of 0 and a standard deviation of 1 (unit variance).

The goal of standardization is to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information.

**Q-15:-** Cross-validation is a technique that allows us to utilize our training data better for training and evaluating the model performance . For example, while using cross-validation, we will effectively use complete data for training the model.

Cross-validation also helps in finding the best hyperparameter for the model.

The advantage of using Cross-Validation method is that we make use of all data points and hence it is low bias.

The major drawback of this method is that it leads to higher variation in the testing model as we are testing against one data point.