*Session 7: Assignment 2*

**Table of Contents**

1. Introduction
2. Problem Statement
3. Output

1. Introduction

This assignment will help you to consolidate the concepts learnt in the session.

1. Problem Statement
2. What are the three stages to build the hypotheses or model in machine learning?
3. What is the standard approach to supervised learning?
4. What is Training set and Test set?
5. What is the general principle of an ensemble method and what is bagging and boosting in ensemble method?
6. How can you avoid overfitting?

**Note: The solution shared through Github should contain the source code used and the screenshot of the output.**

**3. Solution:**

1. The three stages to build the hypotheses or model in machine learning are:
2. Model building
3. Model testing
4. Applying the model
5. The standard approach to supervised learning is to split the set of examples into the training set and the test set.
6. A set of data is used to discover the potentially predictive relationship known as ‘Training Set’. Training set is an example given to the learner, while Test set is used to test the accuracy of the hypotheses generated by the learner, and it is the set of examples held back from the learner. Training set are distinct from Test set.
7. The general principle of an ensemble method is to combine the predictions of several models built with a given learning algorithm to improve robustness over a single model. Bagging is a method in ensemble for improving unstable estimation or classification schemes. While boosting method are used sequentially to reduce the bias of the combined model. Boosting and Bagging both can reduce errors by reducing the variance term.
8. The problem of overfitting can be addressed by reducing after it has learnt to remove some of the detail it has picked up. Overfitting happens relatively as you have a small dataset and try to learn from it or that models the training data too well. If you have a small database and you are forced to come with a model based on that. In such situation, you can use a technique known as cross validation. In Cross validation method the dataset splits into two sections, testing and training datasets, the testing dataset will only test the model while, in training dataset, the data points will come up with the model. In this technique, a model is usually given a dataset of a known data on which training (training data set) is run and a dataset of unknown data against which the model is tested. The idea of cross validation is to define a dataset to “test” the model in the training phase.