Session 19 - Assignment

1. **What are the three stages to build the hypotheses or model in machine learning?**

   - Model building
   - Model testing
   - Applying the model

2. **What is the standard approach to supervised learning?**

   The standard approach to supervised learning is to split the dataset into a training set and a test set.

3. **What is Training set and Test set?**
   - Separating data into training and testing sets is an important part of machine learning.
   - **Training set** is an examples given to the learner,
   - **Test set** is used to test the accuracy of the hypotheses generated by the learner, and it is the set of data held back from the learner.
   - Typically, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing.
   - Analysis Services randomly samples the data to help ensure that the testing and training sets are similar.
   - By using similar data for training and testing, you can minimize the effects of data discrepancies and better understand the characteristics of the model.
   - After a model has been processed by using the training set, you test the model by making predictions against the test set. Because the data in the testing set already contains known values for the attribute that you want to predict, it is easy to determine whether the model's guesses are correct.

4. **What is the general principle of an ensemble method and what is bagging and boosting in ensemble method?**
   - Ensemble learning helps improve machine learning results by combining several models. This approach allows the production of better predictive performance compared to 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.

5. How can you avoid overfitting?

- By using a lot of data overfitting can be avoided.
- Overfitting happens relatively as you have a small dataset, and you try to learn from it.
- But 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 this method the dataset splits into two section, testing and training datasets, the testing dataset will only test the model while, in training dataset, the datapoints will come up with the model.