**Session14-Assignment1**

1. What are the three stages to build the hypotheses or model in machine learning?
2. Creating and Building the model.
3. Train the model using Training Data.
4. Applying the model to Test Data.
5. What is the standard approach to supervised learning?

Standard approaches for supervised learning are like

1. Use labeled data set which is Collected and applying all Data Preparation processes.
2. Identify the models like based on type of learning as Regression, Classification, and anomaly detection.
3. From the data separate Training and Test Data, apply the model on Training data, on stratification on training data with maximum result apply to test data.
4. What is Training set and Test set?

Book Definition: In various areas of information science like machine learning, 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 example held back from the learner. Training set is distinct from Test set.

From the data collected for applying machine learning concept, data has to be separated in order to train the model for the specific learning. Training Set is normally 80% proportion of the total data and on training the data we have to come up with perfect or near perfect results i.e error should be less. Test Set is usually used to evaluating the mode which is trained using training set. On training set also error should be less avoiding Overfitting.

1. What is the general principle of an ensemble method and what is bagging and boosting in ensemble method?

Book Definition: The general principle of an ensemble method is to combine the predictions of several models built with a given learning algorithm in order 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.

1. **How can you avoid overfitting?**

Book Definition: 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.

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.