**ML Assignment 1**

**1) What are the 3 stages to build the hypotheses or model in machine learning?**

Stage 1: Identify the independent variables and the dependent variable and the relationship that exists between the independent and dependent variable.

Stage 2: Identify the machine learning algorithm that can be used for the given scenario in stage 1 depending on the identified relationship between the independent variables and the dependent variable.

Stage 3: Run the machine learning algorithm on the training data set to generate the machine learning model. Then run the model on the test data set to calculate and fine tune the model to identify the best fit/optimal values for the parameters of the function that the learning algorithm is solving for achieving the desired accuracy.

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

In Supervised learning, the independent variables(input) and the dependent variable(output) are already known. The standard approach is as follows:

- Divide the dataset into two sets - Training set and the test set.

- Identify and run the Machine learning algorithm on the training data set to get the optimized parameters for the coefficients and constants of the algorithmic function.

- Run the algorithm on the test data and check the accuracy of the algorithm on the test data set.

- Further optimize the accuracy of the model if required till the desired range is achieved.

**3. What is Training set and Test set?**

The training set is the data set which is used by the Machine learning algorithm to come up with the machine learning model. The test data set is the data set on which the machine learning model that is generated in the training stage is used to evaluate and achieve the required accuracy of the model.

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

Not covered yet

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

Overfitting can be avoided by removing the correlation between the input variables. Another reason for overfitting is when the independent variables have different scales. This problem can be solved by using the standard scaler function to standardize the data set.