In [ ]:

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


```

In [ ]:

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

The Three Stages for building the hypothesis are as follows

a. Model Building /DataEngineering : These include following
    1. Understanding the Business /Domain ,This will help to understand and identify correctly the
       problem that is occuring ,that needs to be solved and build a case for the same.
    2. Once the problem is identified , then identify the data sources from which data can be obtained
       for solving the problem, the Data can be structured data as databases or unstructured like emails,
       documents,pictures etc.

       Once the Data is obtained , then preprocess/normalize the Data by processes like Data Cleaning
       (filling missing values, ...), Data integration, data transformation.
       The normalized data is then separated into Training Set and Testing Set.

    3. Developing or selecting a model/Algorithim , which helps to solve the use case .For this The
       training Data Set is used and model is build is build for it .
       Once the model is built , it is validated with the Validate set.


b. Model Testing/Data Intelligence :
    In these stage,You test the buildModel with the testing Data Set, several times  and analyse the predicted
    results , if they are within the predicted score then model is performing as needed , but if
    accuracy of predicted score is not good , then identify the problem ,rectify,finetune the model to
    improve score , else some other algorithms are tried .

c. Model Appling/Deployment :
    In this stage , once the model is working as predicted and sufficent accuracy, then apply or deploy
    the model in the business environment  .



2. What is the standard approach to supervised learning?

   The Standard approach for Supervised learning is :
   PreProcess the Raw Data and transform it into Data Features .The Data is split into training and
   Test Data . The Algorithm is then trained with the training data and labels for these DataFeatures.
   The Model is tuned to get the better accuracy for different algoritm parameters.
   The Fine tuned Model is then tested to predict Labels for the Test Data.
   Based on the Training  it has achieved , it appropriately labels the new Data.

```

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

    The Raw Data is preprocessed by Data Cleaning, Data integretion , Data transformation processes like normalizatic
    The Processed Data is divided into Two sets , Training and Test Set in the ratio 70:30 percent.

    Training Set is the Data that is used for training the model/algorithm and hence the name.
    The training set includes labels that by which the Model learns their mapping and this will help
    it to make predictions of labels  for the new Data.

    Once the Learning by the model is done , the Test set (the subset of input normalised data) is used
    to test the model and predict the labels for the new data. The predictions made by the test Set
    helps to give the accuracy of the predictions done by the Model.


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

    The general principle of ensemble is , it is supervised meta machine learning model which combines
    the predictions of many small models to improve on the accuracy of predictions and make a robust ,
    generalized model.

    Ensemble is achieved by methods like bagging , boosting , stacking , to address the issues like statistical, comp
    and representational respectively

    Bagging is a type of ensemble where many models/estimators are build independently on subsets of the data
    and then their predictions are averaged. The result /prediction of bagging model is better /more accurate
    then each of the sub models. This method addresses the statistical issue faced by ensemble model by limiting
    the variance between the data points without much affecting the bias.
    eg is random Forest

    Boosting is a type of ensemble where model/estimators are build sequentially and thesubsequent  model
    works on the weaknesses of the previous model to make a better predicting model i.e by combining many
    weak models it makes a stronger model .This method addresses the issue of computational faced by reducing
    the bias of the data points without affecting the variance between the data points.
    eg. Adaboost

5. How can you avoid overfitting ?

    Overfitting in the model occurs when we train it with lot of data , in doing so , the model not only learns from
    but also from the noise or inaccuracies in the data . It starts considering the errors as the actual data and
    hence becomes overfitted , resulting in wrongly predicting new data .
    To avoid overfitting , one needs to study the behaviours of data , if it is  linear , then use linear algoritm
    and when non linear , then use paramters in the model classifier like max_features, max_dept, min_samples_leaf
    to limit overfitting

```

In [ ]:

```

```