2.​ Problem Statement

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

a)      Model building 🡪 Involves (Data Preparation, Training and Test set generation, Algorithm Training)

b)      Model testing 🡪 Involves (Prediction and Evaluation of Test Data)

c)       Applying the model 🡪 Involves (Deployment and Monitoring)

1. What is the standard approach to supervised learning?

The standard approach to supervised learning is to split the set of example into the training set and the test.

1. What is Training set and Test set?

**Training set** is a data set used to train the model. Specific features are picked up from the training set for training purpose.

**Test set** is a data set used to measure how well the model performs at making predictions on the test data based on the training.

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

Ensemble learning is used when you build component classifiers that are more accurate and independent from each other. 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 over fitting?

The possibility of over fitting exists as the criteria used for training the model is not the same as the criteria used to judge the efficacy of a model. By using a lot of data over fitting can be avoided, over fitting 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 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