**Problem​ ​Statement**

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

**Solution**

a)      Model building

b)      Model testing

c)       Applying the model

* What is the standard approach to supervised learning?

**Solution**

Supervised learning is where we have input variables (x) and an output variable (Y) and we use an algorithm to learn the mapping function from the input to the output.

Y = f(X)

The goal is to approximate the mapping function so well that when we have new input data (x) that we can predict the output variables (Y) for that data.

* What is Training set and Test set?

**Solution**

In a dataset a training set is implemented to build up a model, while a test set is to validate the model built. Data points in the training set are excluded from the test set. we run calculations on the training set to determine various coefficients. we can then use the testing set to check how well the predictions do on a wider set of data, and that gives you information about false positives and false negatives.

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

**Solution**

The general principle of ensemble methods is to construct a linear combination of some model fitting method, instead of using a single fit of the method

**Bagging**: It is the method to decrease the variance of model by generating additional data for training from your original data set using combinations with repetitions to produce multisets of the same size as your original data.

**Boosting**: It helps to calculate the predict the target variables using different models and then average the result( may be using a weighted average approach).

* How can you avoid over-fitting ?

**Solution**

Over fitting can be reduce via cross validation techniques e.g. k fold validation

K-fold cross validation is performed as per the following steps:

1. Partition the original training data set into k equal subsets. Each subset is called a fold. Let the folds be named as f1, f2, …, fk .
2. For i = 1 to i = k
3. Keep the fold fi as Validation set and keep all the remaining k-1 folds in the Cross validation training set.
4. Train your machine learning model using the cross validation training set and calculate the accuracy of your model by validating the predicted results against the validation set.
5. Estimate the accuracy of your machine learning model by averaging the accuracies derived in all the k cases of cross validation.

In the k-fold cross validation method, all the entries in the original training data set are used for both training as well as validation. Also, each entry is used for validation just once.