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

**Answer:**

* Model building
* Model testing
* Applying the Model

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

**Answer:**

To splits the dataset 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.

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

**Answer:**

The testing dataset will only test the model while, in training dataset, the datapoints will come up with the model. 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 examples held back from the learner. Training set are distinct from Test set.

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

**Answer:**

The general principle of an ensemble method is to combine the predictions of several models built with a given learning algorithm to improve robustness over a single model. This approach allows the production of better predictive performance compared to a single model.

Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model to **decrease** **variance** (bagging), decrease bias (boosting), or **improve predictions** (stacking).

Bagging is a method in ensemble for improving unstable estimation or classification schemes. Bagging both can reduce errors by reducing the variance term.

Boosting method are used sequentially to reduce the bias of the combined model. Boosting can reduce errors by reducing the variance term.

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

Answer:

The most obvious consequence of overfitting is poor performance on the validation dataset. Other includes

* A function that is overfitted is likely to request more information about each item in the validation dataset than does the optimal function; gathering this additional unneeded data can be expensive or error-prone, especially if each individual piece of information must be gathered by human observation and manual data-entry.
* A more complex, overfitted function is likely to be less portable than a simple one. At one extreme, a one-variable linear regression is so portable that, if necessary, it could even be done by hand. At the other extreme are models that can be reproduced only by exactly duplicating the original modeler's entire setup, making reuse or scientific reproduction difficult.

**Methods to overcome overfitting**

Cross Validation – K Fold Cross Validation: We split the date into K different subsets. K-1 subsets are used to train the data and leave the last subset as test data. Then average the model against each fold and then finalize the model. after that is tested against test data. If validation accuracy is lower, do regularization else repeat regularization until no overfitting

Ensembling - is machine learning methods for combining predictions from multiple separate models. *Bagging* attempts to reduce the chance overfitting complex models.