# 1.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

**Building a Machine Learning Application**

Building ML applications is an iterative process that involves a sequence of steps. To build an ML application, follow these general steps:

1. Frame the core ML problem(s) in terms of what is observed and what answer you want the model to predict.
2. Collect, clean, and prepare data to make it suitable for consumption by ML model training algorithms. Visualize and analyze the data to run sanity checks to validate the quality of the data and to understand the data.
3. Often, the raw data (input variables) and answer (target) are not represented in a way that can be used to train a highly predictive model. Therefore, you typically should attempt to construct more predictive input representations or features from the raw variables.
4. Feed the resulting features to the learning algorithm to build models and evaluate the quality of the models on data that was held out from model building.
   1. **Formulating the Problem**

The first step in machine learning is to decide what you want to predict, which is known as the label or target answer. Imagine a scenario in which you want to manufacture products, but your decision to manufacture each product depends on its number of potential sales. In this scenario, you want to predict how many times each product will be purchased (predict number of sales). There are multiple ways to define this problem by using machine learning. Choosing how to define the problem depends on your use case or business need.

# Collecting Labeled Data

ML problems start with data—preferably, lots of data (examples or observations) for which you already know the target answer. Data for which you already know the target answer is calledlabeled data. In supervised ML, the algorithm teaches itself to learn from the labeled examples that we provide.

Each example/observation in your data must contain two elements:

* The target – The answer that you want to predict. You provide data that is labeled with the target (correct answer) to the ML algorithm to learn from. Then, you will use the trained ML model to predict this answer on data for which you do not know the target answer.
* Variables/features – These are attributes of the example that can be used to identify patterns to predict the target answer.

# Analyzing Your Data

Before feeding your labeled data to an ML algorithm, it is a good practice to inspect your data to identify issues and gain insights about the data you are using. The predictive power of your model will only be as good as the data you feed it.

# Feature Processing

After getting to know your data through data summaries and visualizations, you want to transform your variables further to make them more meaningful. This is known as feature processing. For example, say you have a variable that captures the date and time at which an event occurred. This date and time will never occur again and hence won’t be useful to predict your target. However, if this variable is transformed into features that represent the hour of the day, the day of the week, and the month, these variables could be useful to learn if the event tends to happen at a particular hour, weekday, or month. Such feature processing to form more generalizable data points to learn from can provide significant improvements to the predictive models.

# Splitting the Data into Training and Evaluation Data

The fundamental goal of ML is to generalize beyond the data instances used to train models. We want to evaluate the model to estimate the quality of its pattern generalization for data the model has not been trained on

# Training the Model

You are now ready to provide the ML algorithm (that is, the learning algorithm) with training data. The algorithm will learn from the training data patterns that map the variables to the target, and it will output a model that captures these relationships. The ML model can then be used to get predictions on new data for which you do not know the target answer.

# Evaluating Model Accuracy

The goal of the ML model is to learn patterns that generalize well for unseen data of just memorizing the data that it was shown during training. Once you have a model, it is important to check if your model is performing well on unseen examples that you have not used for training the model. To do this, you use the model to predict the answer on the evaluation dataset (held out data) and then compare the predicted target to the actual answer (ground truth).

# 2. 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.

# 3. What is Training set and Test set?

In various areas of information science like machine learning, a set of data is used to discover the potentially predictive relationship known as ‘Training Set’. Training set is an examples 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 example held back from the learner. Training set are distinct from Test set.

**Splitting the Data into Training and Evaluation Data**

The fundamental goal of ML is to *generalize* beyond the data instances used to train models. We want to evaluate the model to estimate the quality of its pattern generalization for data the model has not been trained on. However, because future instances have unknown target values and we cannot check the accuracy of our predictions for future instances now, we need to use some of the data that we already know the answer for as a proxy for future data. Evaluating the model with the same data that was used for training is not useful, because it rewards models that can “remember” the training data, as opposed to generalizing from it.

A common strategy is to take all available labeled data, and split it into training and evaluation subsets, usually with a ratio of 70-80 percent for training and 20-30 percent for evaluation. The ML system uses the training data to train models to see patterns, and uses the evaluation data to evaluate the predictive quality of the trained model. The ML system evaluates predictive performance by comparing predictions on the evaluation data set with true values (known as ground truth) using a variety of metrics. Usually, you use the “best” model on the evaluation subset to make predictions on future instances for which you do not know the target answer.

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

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.

# How can you avoid overfitting ?

In machine learning, when a statistical model describes random error or noise instead of underlying relationship ‘overfitting’ occurs.  When a model is excessively complex, overfitting is normally observed, because of having too many parameters with respect to the number of training data types. The model exhibits poor performance which has been overfit.