- The objective of Random forest is to develop model that can make prediction.

RANDOM FOREST CLASSIFIER

- Random Forest is a supervised learning algorithm. It is also called as ensemble decision tree model.
- The objective of random forest classifier is to predict class/label of an unseen observation.

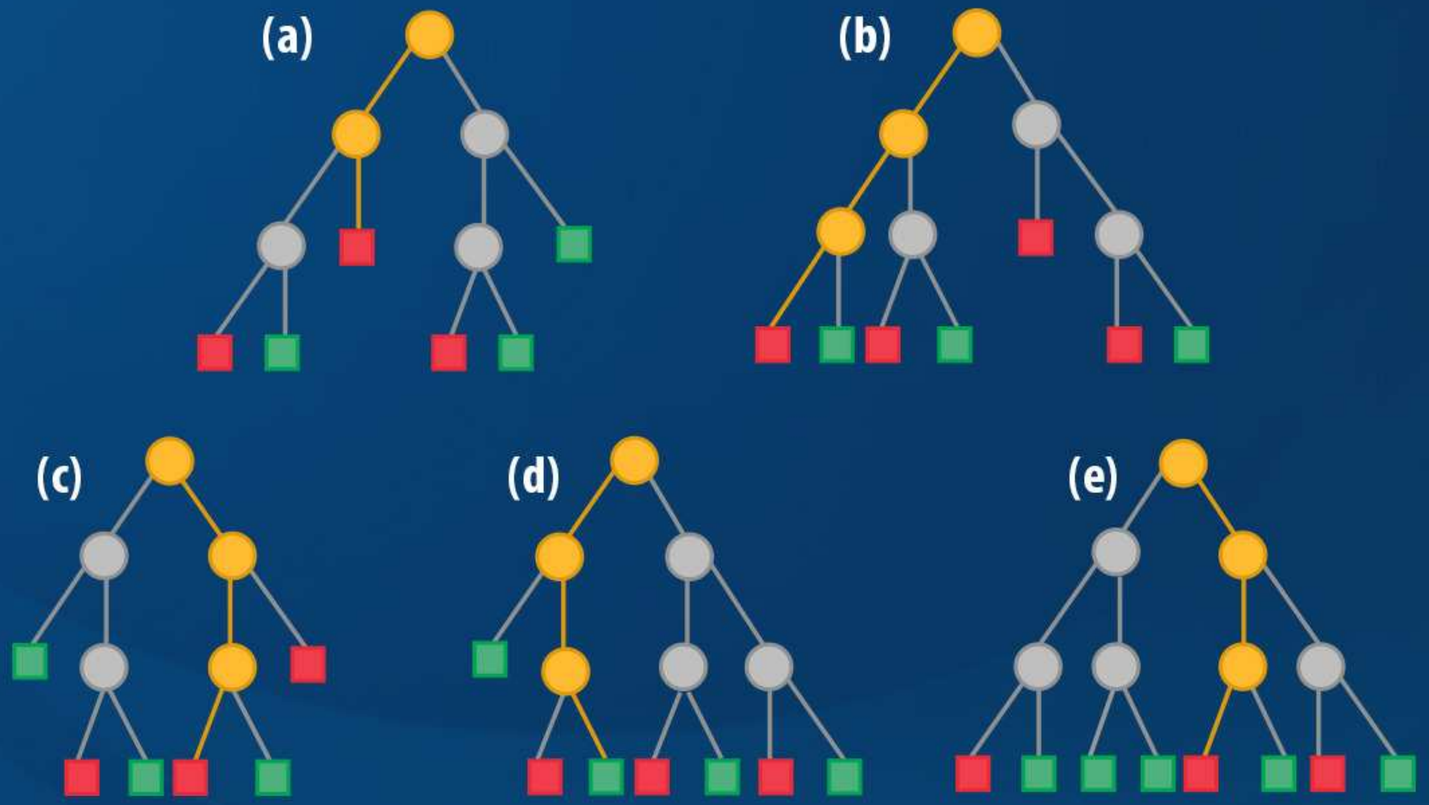


Figure 1: Random forest is collection of decision trees from the same Dataset

FUNDAMENTALS CONCEPTS OF RANDOM FOREST

The development of random forest is based on concepts of:

1. Decision Tree
2. Evaluating methods such as, Hold-out, Cross-validation and Bootstrapping
3. Bagging

Hold-out:

- In this method data set is divided into two parts namely, training and test set.
- Training set is used to train the model.
- Test set is used for evaluating / testing the performance of the model.

Cross-validation:

- In this method data set is divided into k fold.
- The training and test set of each fold is used for training and testing model.

Bootstrap:

- In this method, the training dataset is randomly selected with replacement.

INTRODUCTION TO BOOTSTRAPPING

Bootstrapping is a sampling technique where, a new sample data set is created by random sampling with replacement policy using original data set.

Figure 2 illustrates the idea of bootstrapping.

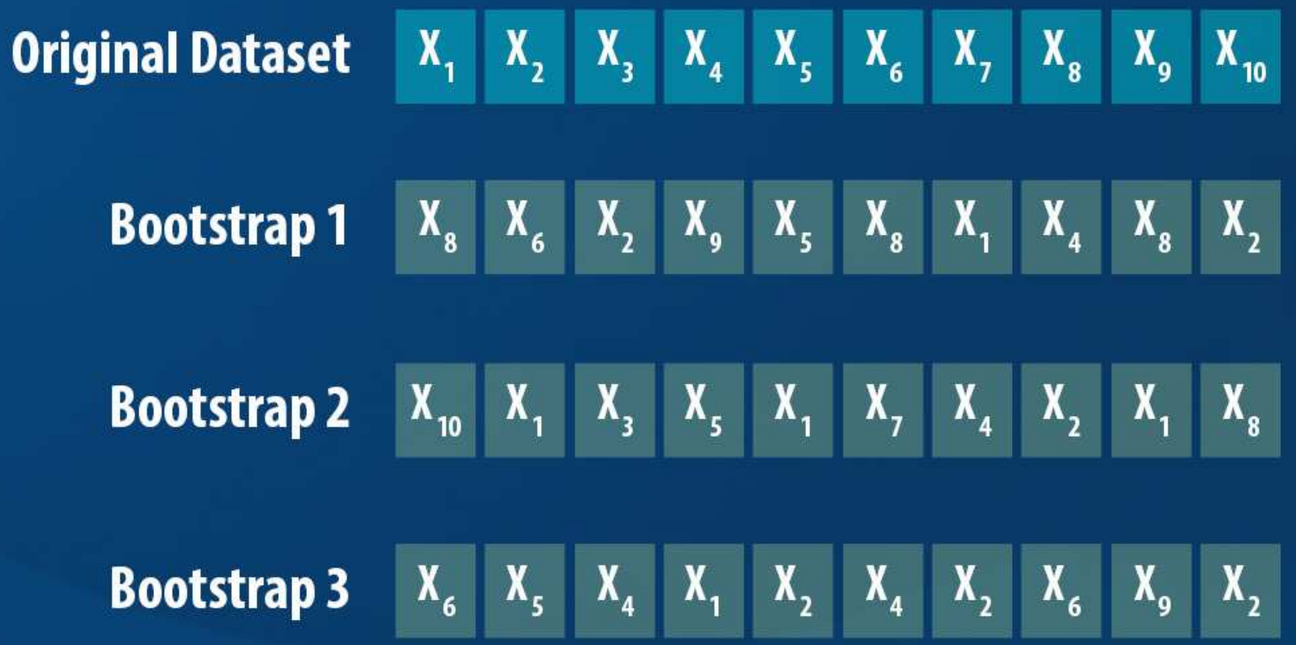


Figure 2: Three new data samples are created using original data set by Bootstrapping

Advantage of Bootstrap

- It develops more robust, stable and consistent models in case of small data sets.

CONFIGURATION OF BOOTSTRAPPING

There are two parameters that must be chosen when performing the bootstrap:

- **The size of the sample.**

  In Machine learning, the sample size is usually kept as the original dataset. It is for the reason to provide the learning model with the same number of observations as given in original data set.

- **Number of repetitions of the procedure to perform.**

  The number of repetitions must be large enough to ensure that meaningful statistics, such as the mean, standard deviation of original data set is captured.

INTRODUCTION TO BAGGING

- Bootstrapping with aggregation to make decision is called bagging.
- Figure 3 illustrates the process of Bagging.

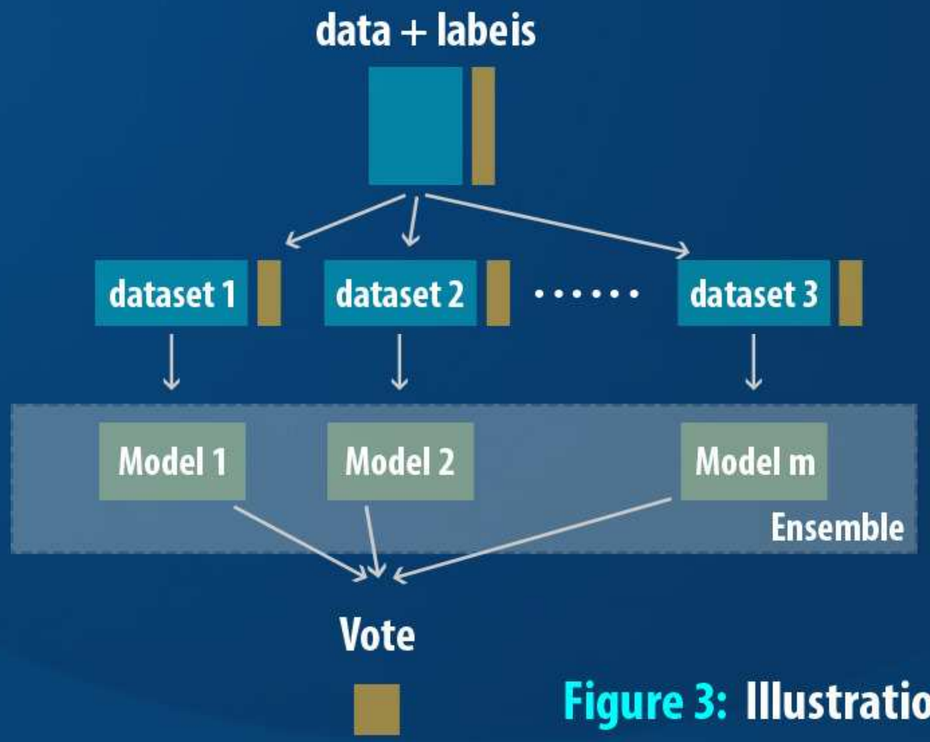
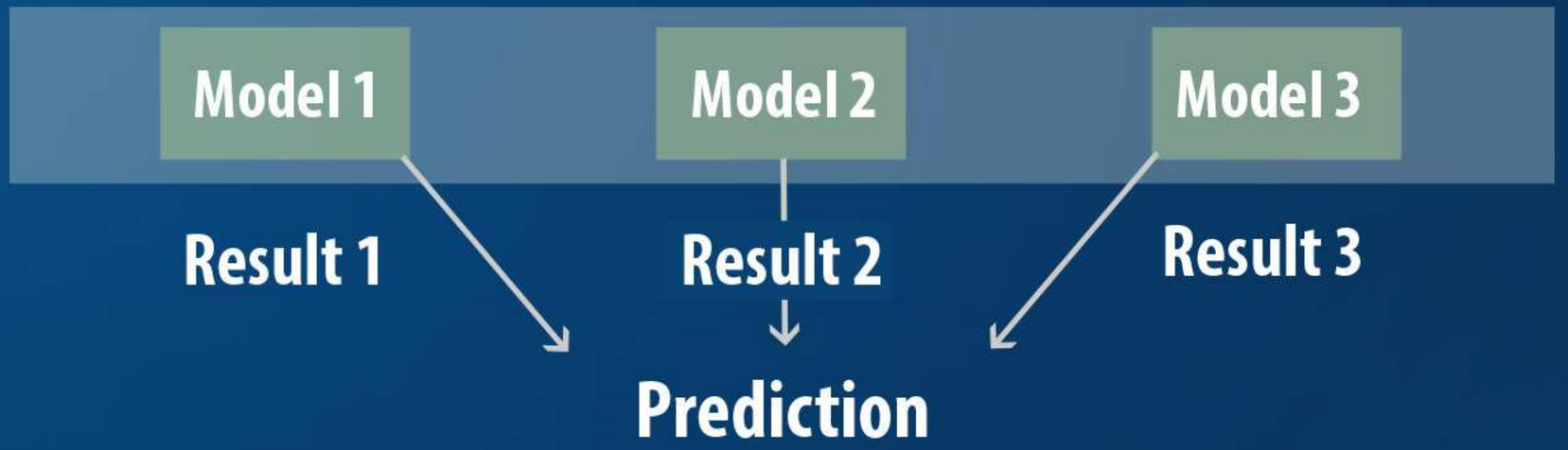


Figure 3: Illustration of Bagging

- An ensemble method is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model.

INTRODUCTION TO BAGGING



Random Forest helps to make prediction from collection of models rather than one model.

ADVANTAGES OF BOOTSTRAPPING AND BAGGING

- The models based on bootstrapping are more robust and stable. The robustness is achieved by using multiple sample data sets and then testing multiple models. In one sense, we get variety of data samples to test our model.
- It works for small to large data sets.
- Bootstrapping overcomes problem of "overfitting". Since we are testing model with several new samples created using bootstrap, the problem of overfitting is reduced to minimum.
- Using bagging, biases controlled.