RANDOM FOREST


farooqeddy.com

What is Random Forests ?

An ensemble classifier using many decission tree models

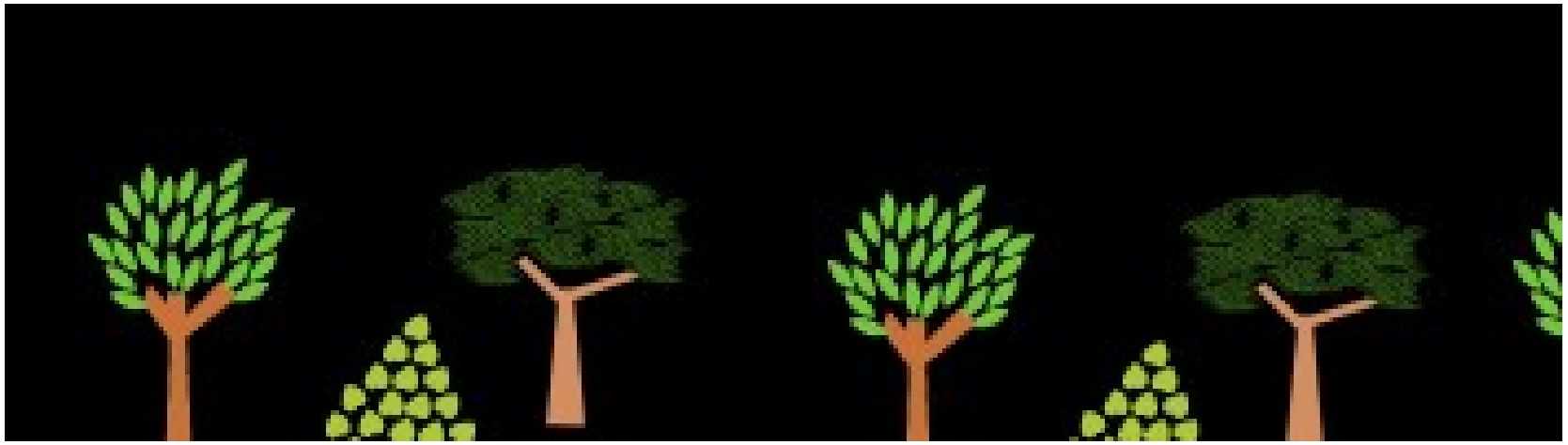
What are ensemble models ?

Ensemble models combine the results from different models
The result from an ensemble model is usually better than the result from one of the individual models.

Basic Principals of Random Forest Method

- As the trees are based on random selection of data as well as variables,these are random tree

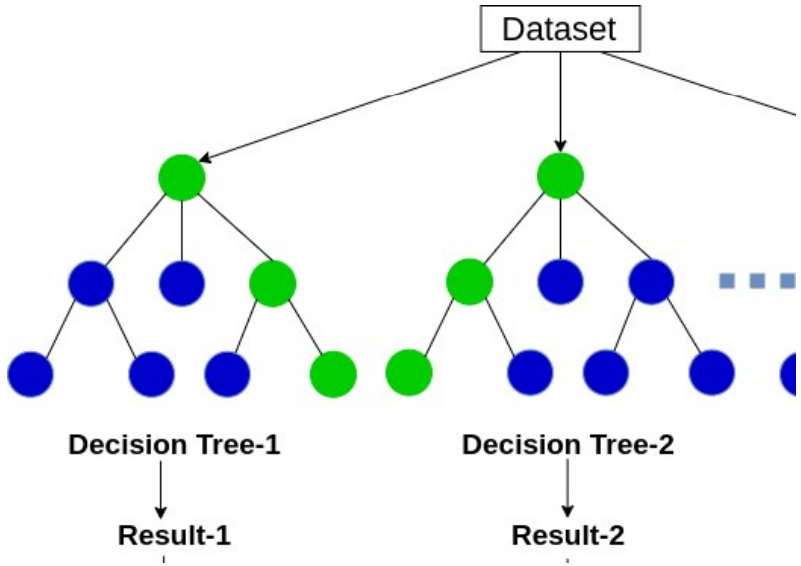
- Many Such random trees leads to random forest



Basic Principals of Random Forest Method

- It develops lots of decission tree based on random selection of data and random selection of variables.

- It provides the class of dependent variable based on many trees.



Features of Random Forests

✓It is unexcelled in accuracy among current algorithms.

✓It runs efficiently on large data bases.

✓It can handle thousands of input variables without variable deletion.

✓It gives estimates of what variable are important in the classification

✓It generates an internal unbiased estimate of the generalization error as the forest building progresses.

Features of Random Forests

✓It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.

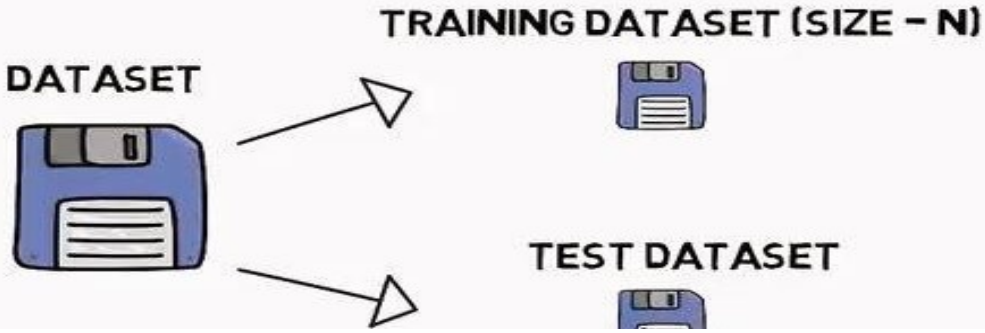
✓It has methods for balancing error in class population unbalanced datasets.

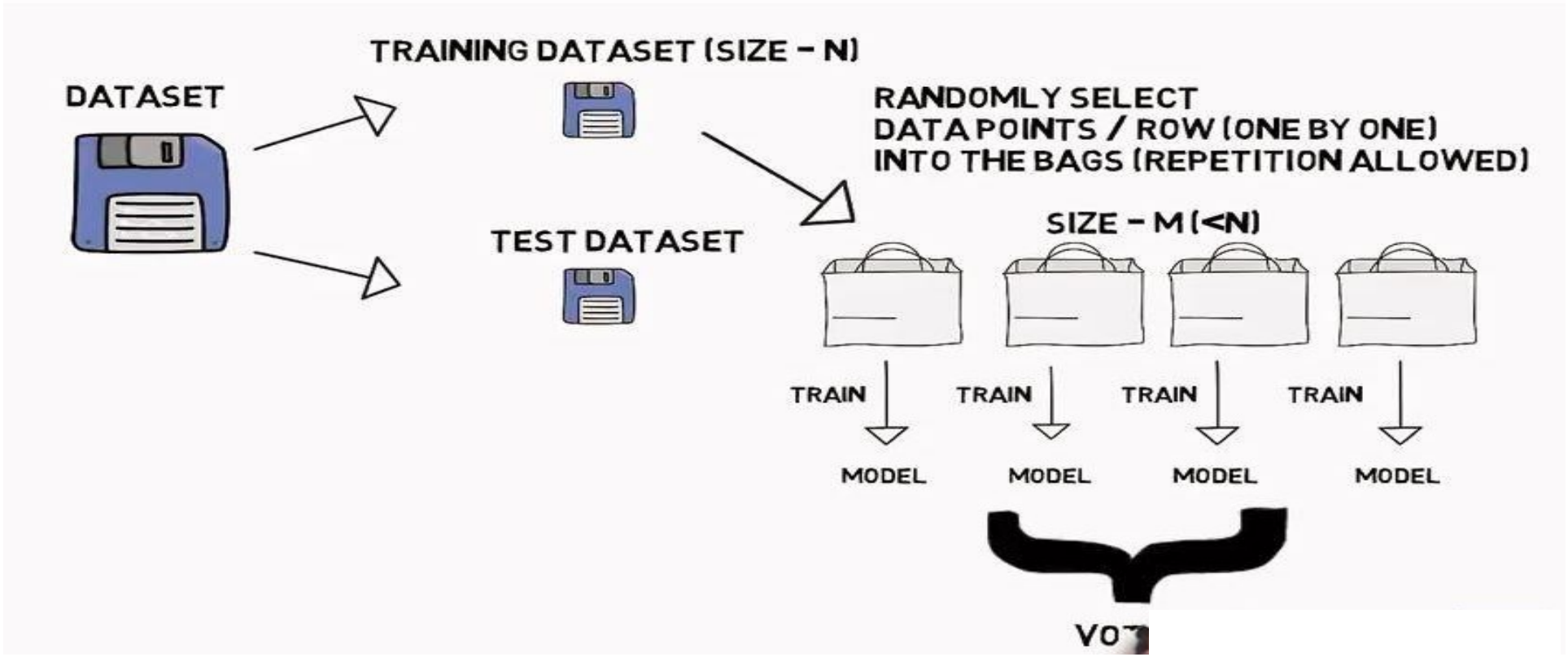
✓Generated forests can be saved for future use on other data.

✓Prototypes are computed that give information about the relation between the variables and the classification.

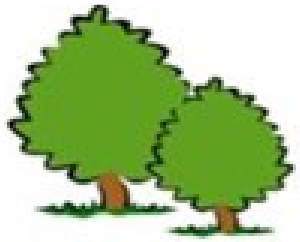
BOOTSTRAP AGGREGATING (BAGGING)

MULTIPLE MODELS OF SAME LEARNING ALGORITHM TRAINED WITH SUBSETS OF DATASET RANDOMLY PICKED FROM THE TRAINING DATASET.




TRAINING DATASET (SIZE – N)
DATASET
RANDOMLY SELECT
DATA POINTS / ROW (ONE BY ONE)
INTO THE BAGS (REPETITION ALLOWED)
SIZE – M (<N)
TEST DATASET
TRAIN
TRAIN
TRAIN
TRAIN
MODEL
MODEL
MODEL
MODEL

✓ A different subset of the training data are selected (~2/3), with replacement, to train each tree



Random Forest Algorithm

- Random forest is an ensemble of decision tree algorithms.

- It is an extension of bootstrap aggregation (bagging) of decision trees and can be used for classification and regression problems.

- In bagging, a number of decision trees are created where each tree is created from a different bootstrap sample of the training dataset.

- A bootstrap sample is a sample of the training dataset where a sample may appear more than once in the sample, referred to as sampling with replacement.

- Bagging is an effective ensemble algorithm as each decision tree is fit on a slightly different training dataset, and in turn, has a slightly different performance. Unlike normal decision tree models, such as classification and regression trees (CART), trees used in the ensemble are unpruned, making them slightly over fit to the training dataset.

- This is desirable as it helps to make each tree more different and have less correlated predictions or prediction errors.
- Predictions from the trees are averaged across all decision trees resulting in better performance than any single tree in the model.

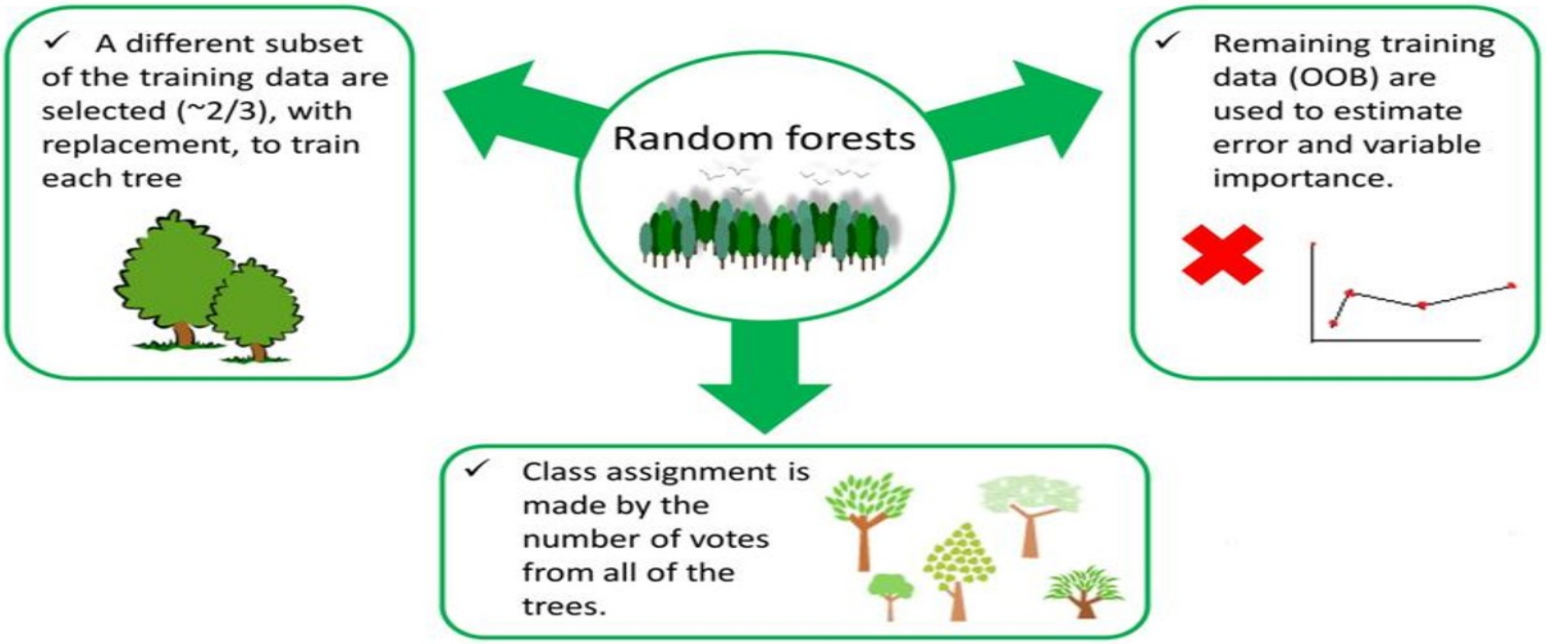
"Each model in the ensemble is then used to generate a prediction for a new sample and these m predictions are averaged to give the forest's prediction"

Regression: Prediction is the average prediction across the decision trees.

Classification: Prediction is the majority vote class label predicted across the decision trees.


RANDOM F
1) 1. Many trees are created using random subsets
of f res and bootstrapped da

How Random Forests Work?



RANDOM FOREST
Case Studies