Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /

# Random Forest

Before, we touch Random forest, Let's talk about Bagging and Boosting.

## Bagging:

- Also known as bootstrap aggregation. It is an ensemble learning model, that is commonly used to reduce variance within a noisy dataset. In bagging, a random sample of data in a training set is selected with replacement - meaning that the individual data points can be chosen more than once.

Example:

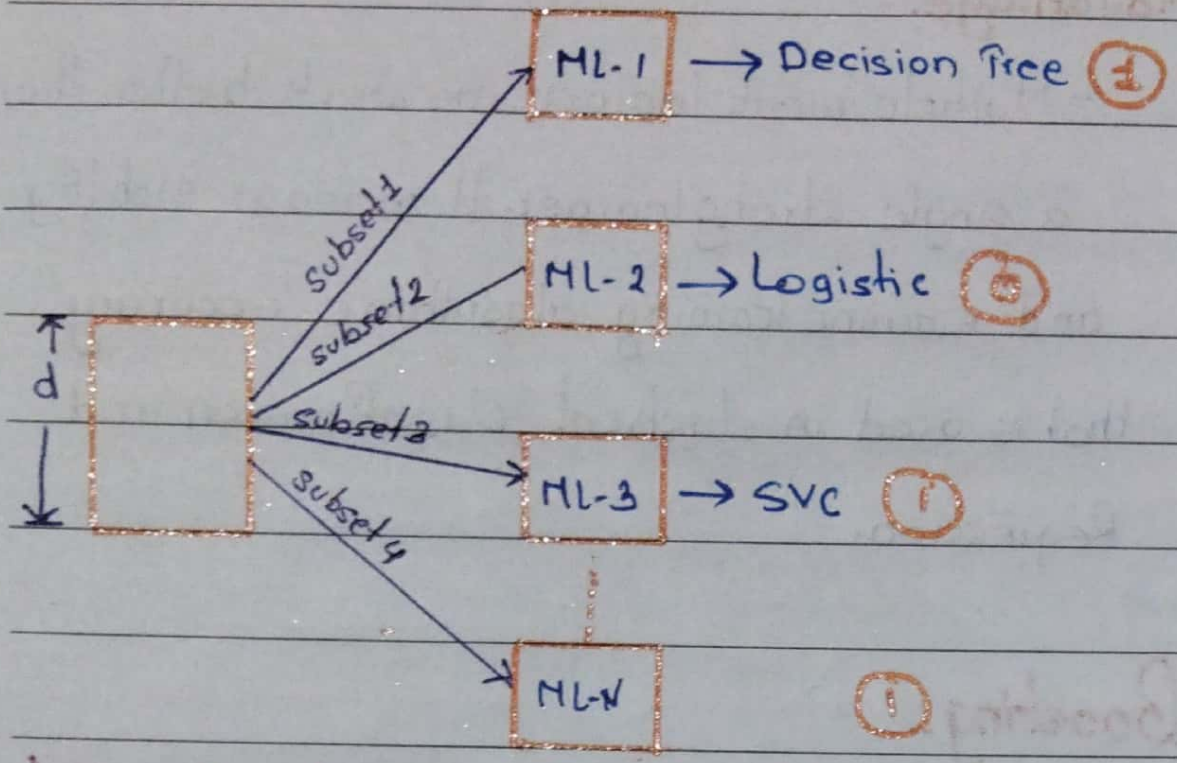
Let us say, we have "d datasets". We are taking 'n' number of sample. So, for each subset we are training them with different machine learning algorithm.

Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /

## For classification Problem




"Many algorithms are used parallely, not sequentially."

Final output will be the:- Major Voting classifier

Final verdict: [1, 0, 1, ] → 1

## For Regression Problem

Final output :- Mean or Average.

Bagging Based Algorithms:-

1. Random Forest Regressor
2. Random Forest classifier

Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /

## Advantage:-

- Multiple weak learner can work better than a single strong learner. It increase stability and machine learning algorithm's accuracy that is used in stastical classification and Regression.

## Boosting:

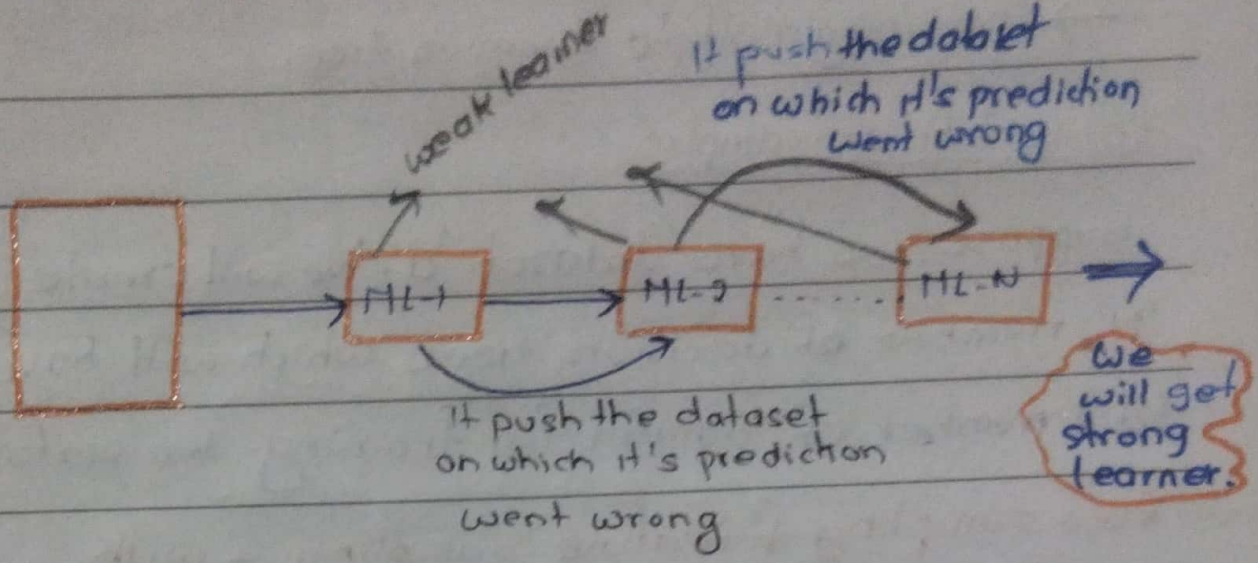
Boosting is an ensemble modeling technique that attempts to build a strong classifier from the number of weak classifiers. It is done by building a model by using weak models in series. Firstly, a model is built from the training data. Then, the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added untill either the complete training

Mo Tu We Th Fr Sa Su

Memo No. _____

Date __/__/__

data set is predicted correctly or the maximum number of model is added.




Algorithm based on boosting:-

① Adaboost Regressor and classifier

② Gradient boost Regressor and classifier.

③ Xgboost Regressor and classifier.

↳ ~~Extreme~~ Gradient Boost

Memo No. _____

Date / /

Mo Tu We Th Fr Sa Su

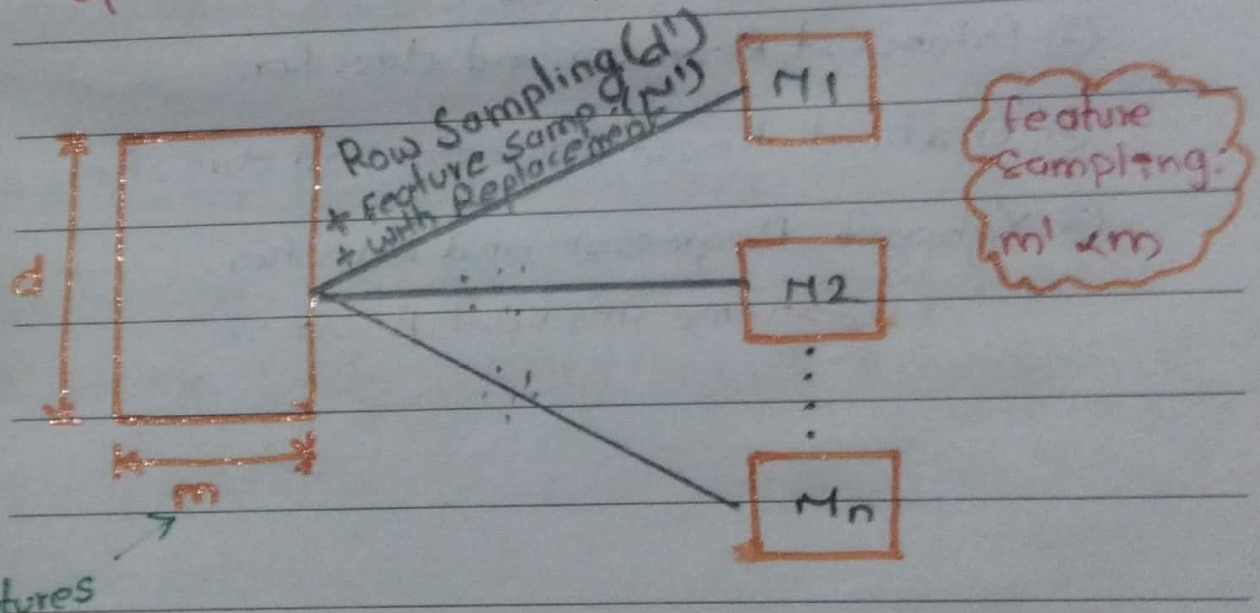
# Random Forest Classification & Regression

- Ensemble Technique.

→ Combining multiple decision tree.

Let us take an example:

Suppose, we have dataset 'd'. We will create 'N' number of decision trees which will have (d') number of dataset for training. We perform Row sampling + feature Sampling + with Replacement such that $d' < d$.




Every model perform computation on the basis of entropy and whatever we studied in decision tree.

Memo No. _____

Date / /

## Problem with Single decision tree:

→ When one single model of decision tree got overfitted.

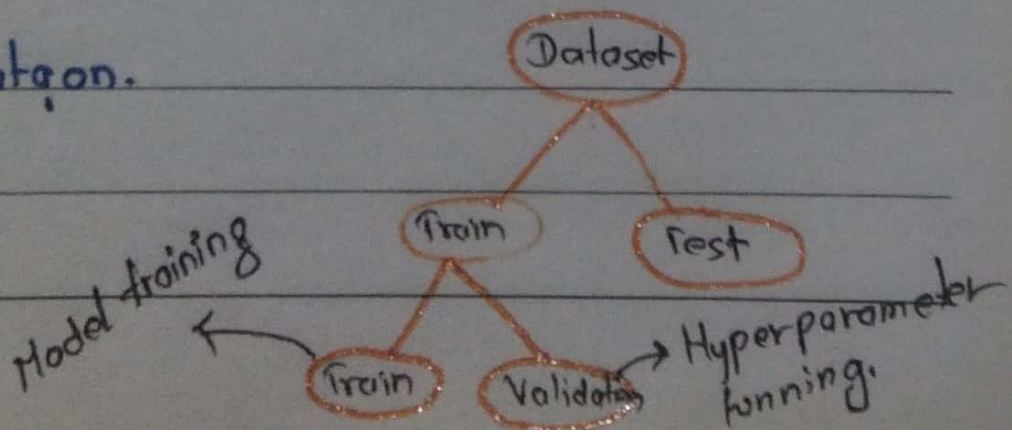
- Low bias
- High variance — we want to make it → Low variance

We can do post-prunning and pre-prunning, but this is very difficult task when we have huge datasets.

But, when we have multiple datasets, taking averages or taking majority voting classification, we can ~~use to many~~ convert high variance → low variance

Even though few of decision tree get overfitted, most of tree won't. Thus, the majority voting will balance the high variance to low variance.

While sampling, some datasets may never get ~~data~~ get chance, these dataset will be used in validation.



Memo No. _____

Mo Tu We Th Fr Sa Su

Date / /

## Out of bag Score:

While making the samples, data points were chosen randomly with replacement, and the data points which fail to be a part of that particular sample are known as out-of-bag points. These dataset are used for validation purpose. and whatever the accuracy we get is known as out of bag score.

2/3rd of datapoints are selected for training and 1/3rd for validation purpose statically

Out of bag error = 1 - OOB Score