

Ensembles and its techniques

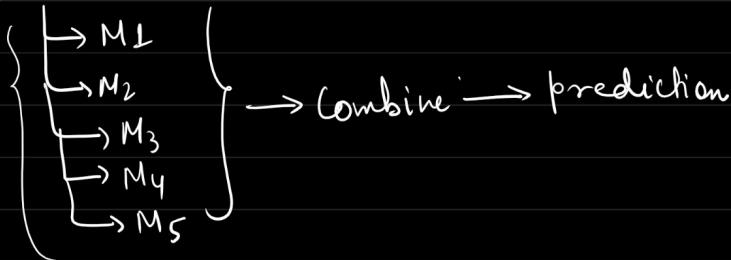
Analogy



① data

Model 1 → train → Prediction

② data



- One person might give you wrong advice
- you will connect to multiple mentors
- Chances of getting wrong is minimized when you connect to multiple person.

Ensembles : Combine multiple models

: prediction - more stable and accurate as compared to individual model.

① (of same Algorithm)

$$\begin{aligned} DT_1 &= (\text{max depth} - 5) \\ DT_2 &= (\text{max depth} - 10) \\ DT_3 &= (\text{max depth} - 12) \end{aligned}$$

② (different Algorithm)

$$\begin{aligned} &\rightarrow \text{Logist Reg} \\ &\rightarrow \text{SVC} \\ &\rightarrow \text{DT C} \end{aligned}$$

* Ensemble : Not necessarily only one type of Algorithm.

Ensemble techniques

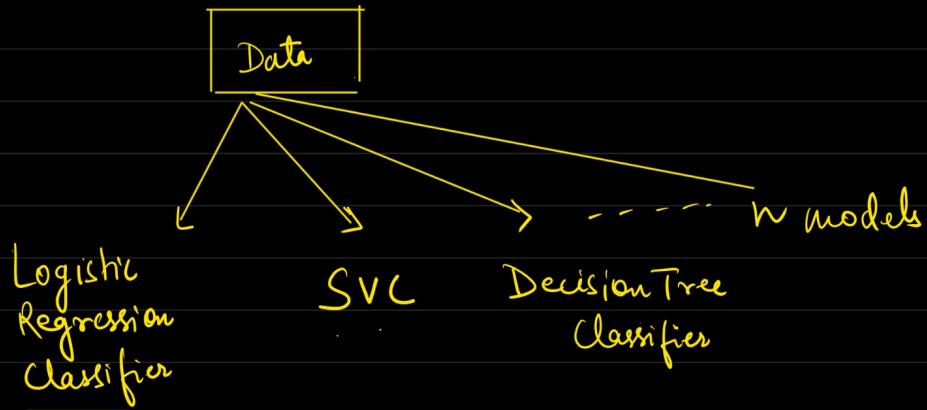
Parallel
technique

(Bagging)

Sequential
technique

(Boosting)

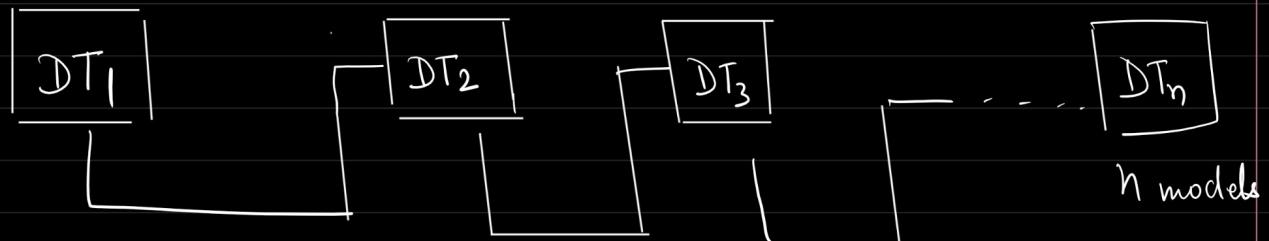
Parallel technique of Ensembles



→ All of the models here are built parallelly and independent of each other.

* Sequential technique of Ensemble

→ All the models are built sequentially and dependent on each other.

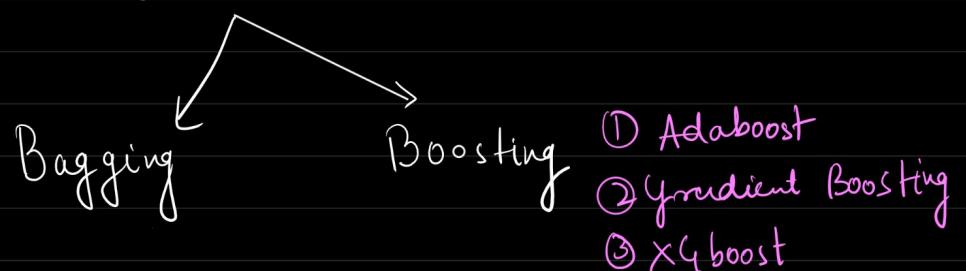


→ learning from mistake.

* Ensembles technique



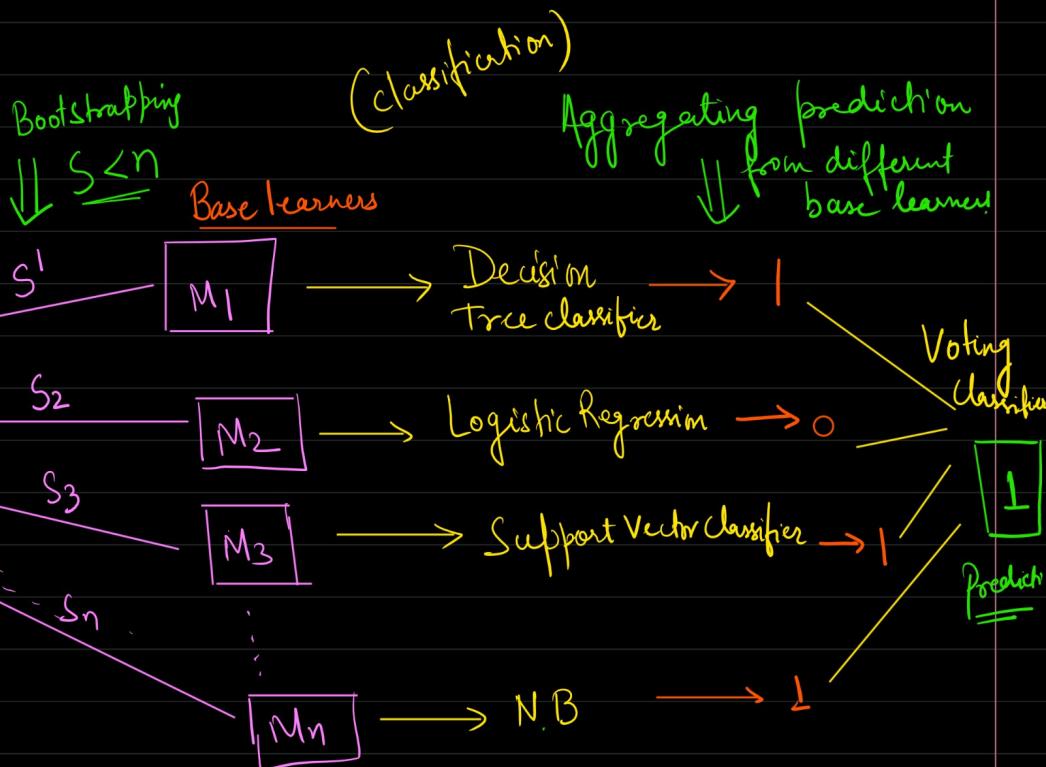
* Ensemble models



- ① Custom Bagging \leftarrow Reg Classifier
 ② Random forest \leftarrow Reg Classifier.

* Bagging Technique

\rightarrow Parallel models



* Samples (subset) for each model is taken with replacement.

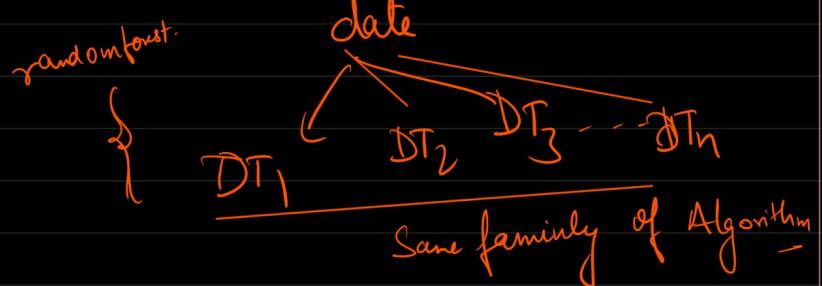
Bagging

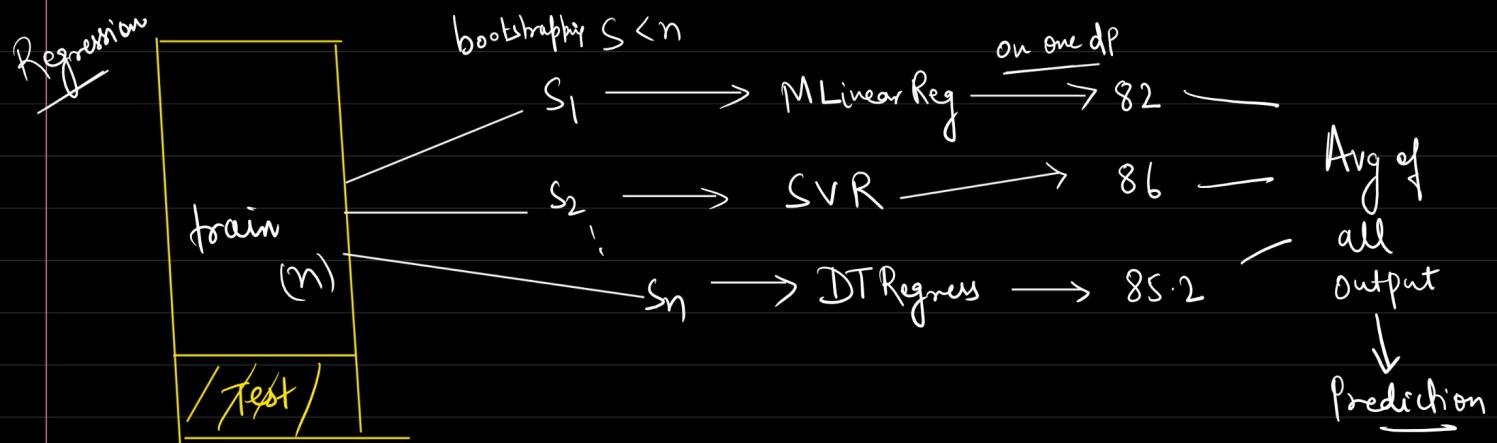
Bootstrap Aggregating

different samples with replacement

Bootstrap sample

* Custom bagging technique





→ Custom Bagging Regressor / classifier.

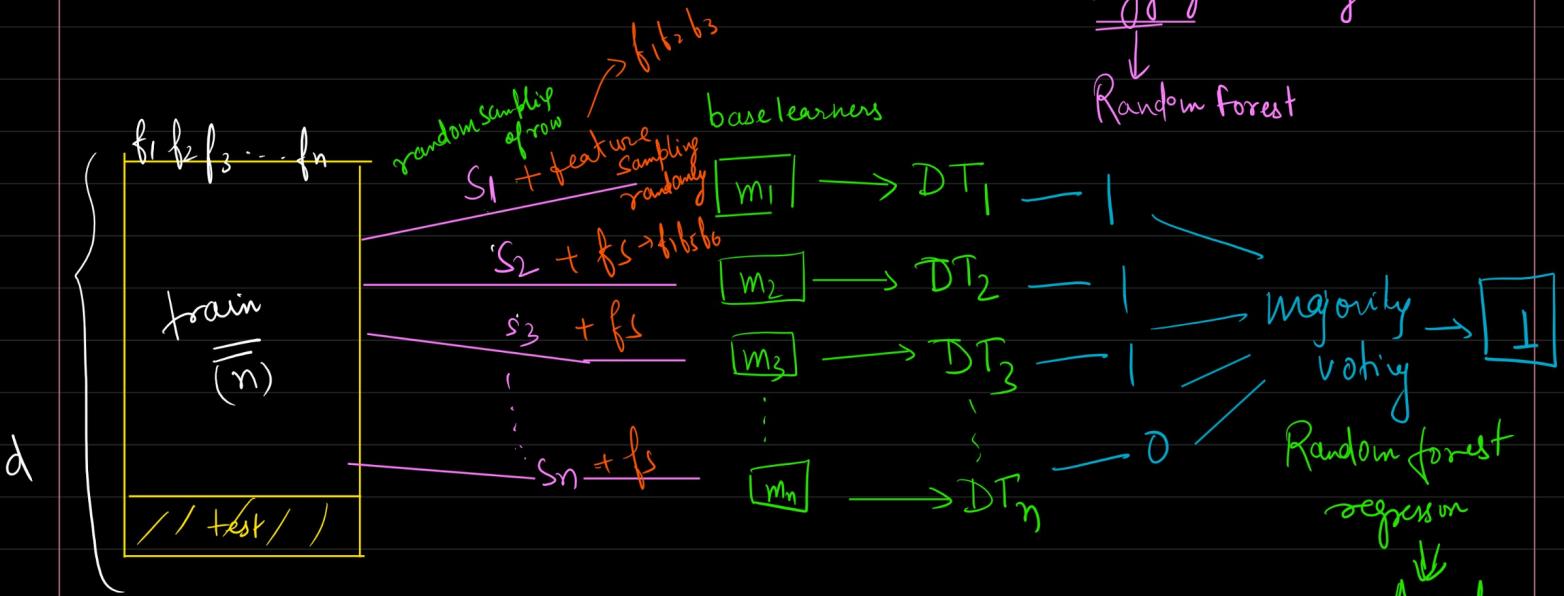
→ Pipeline / Column Transformer (missingValue, Encoding, Scaling)

* Random Forest Classifier and Regressor

↓
Forest → multiple trees

Ensembles

Bagging boosting
↓
Random forest



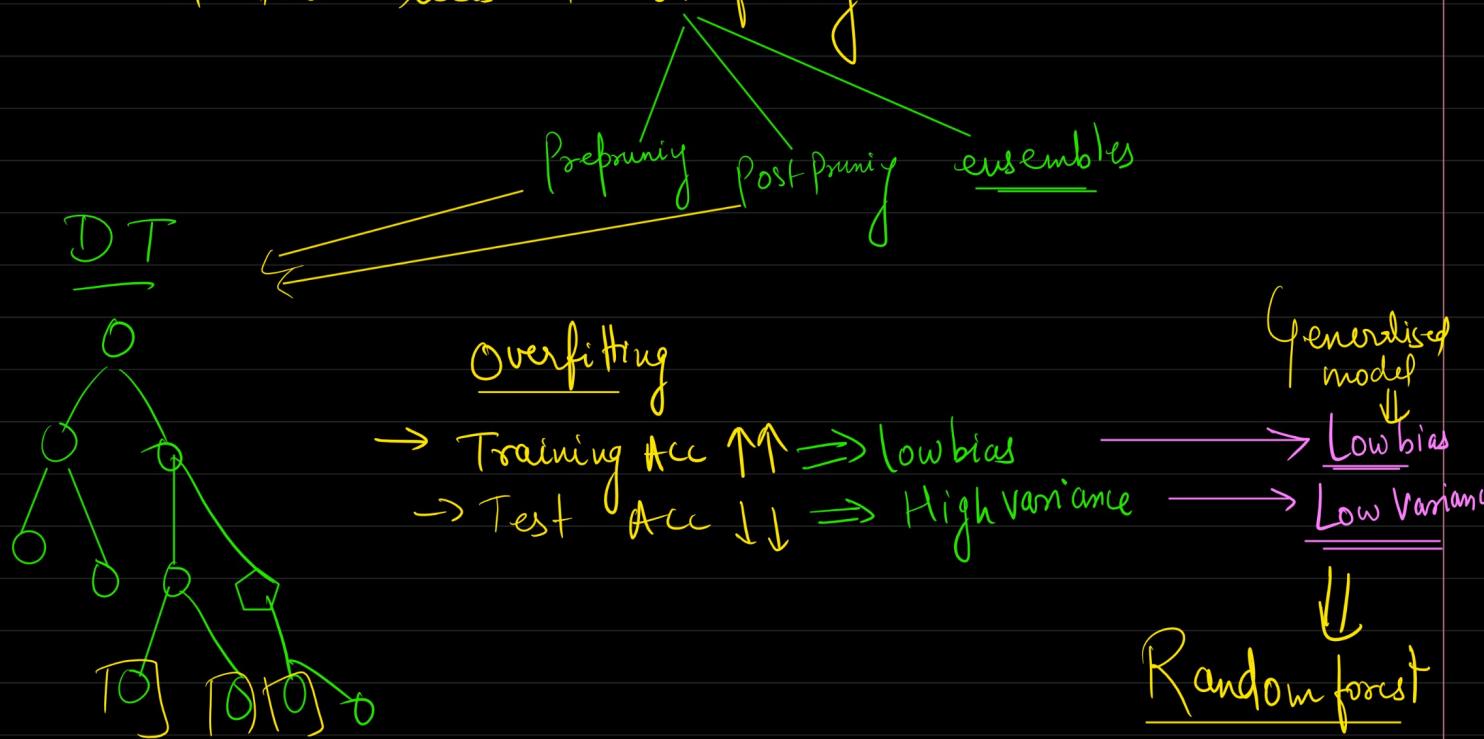
Random forest → multiple decision tree model
in Parallel

→ The rows and features
will be randomly sampled.

- * Random forest classifier \rightarrow Majority Voting as the predicted result
- * Random forest Regressor \rightarrow Average of all the models.

* If Decision tree, then why Random forest?

\rightarrow DT is a greedy Algorithm. It will keep splitting until all the DT's are memorized and this leads to overfitting.



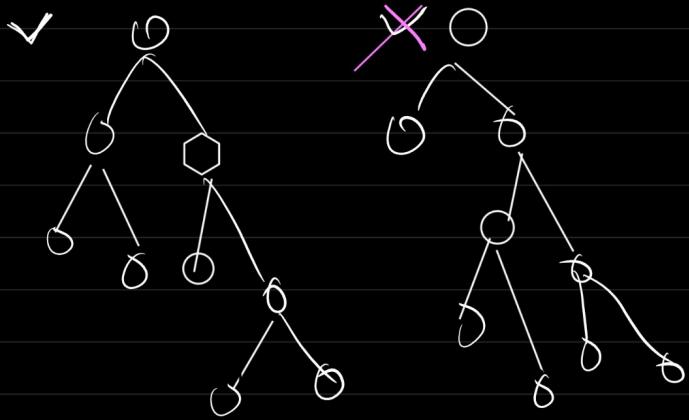
* Random forest Reduces the Variance (reduces Overfitting)
 \hookrightarrow give more robust/generalized model!

\rightarrow in RF, random Sampling of rows and features are done.

\Downarrow
 data is splitted into small chunks (randomly)

\Downarrow
 due to random subset of rows and features,
 each of the subset is different representation
 of itself.

\rightarrow Each of random subset trains a different model.



⇒ The feature (row which was causing an overfitted model might not be contributing to overfitting in other DT models of a random forest.