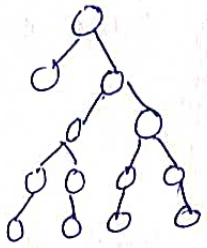
13/4/22. Bagging and Boosting 1. Random-Forest classifier
2. " " regressor

Decision tree :-



Overfitting :- low bias + High Variance :-

In order to avoid that we use

Pre Pruning
Post Pruning

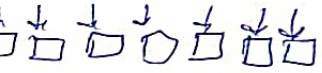
while constructing the decision tree, we use specific hyperparameter like max_depth, min_split, max_feature.

first of all we will construct the decision tree & we will cut the branches.

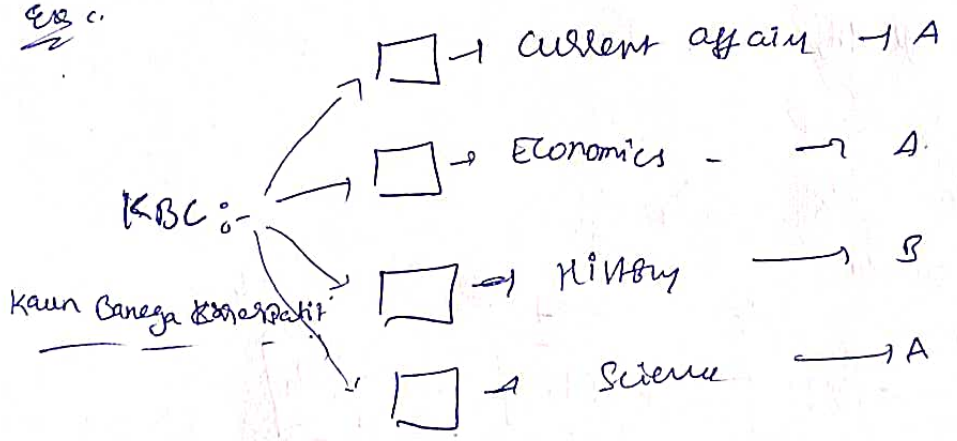
Bagging & Boosting :-

KBC : Amitabh Bachan.

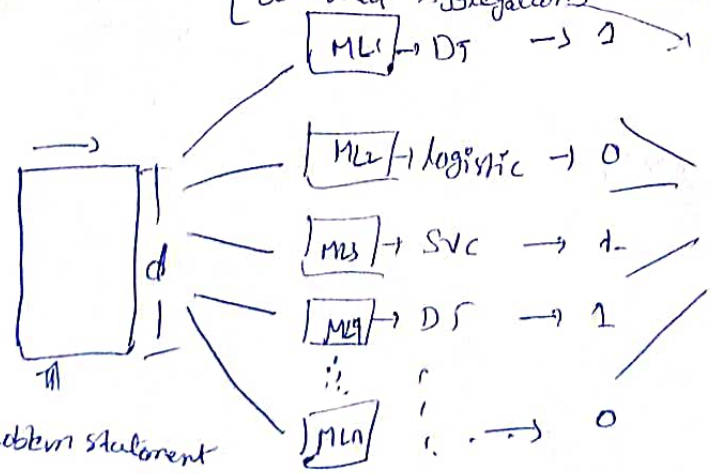
KRISH → Data science
UPSC → Diversified



Ex :-



Bagging :- using ML algorithms Parrallel each other, the final o/p, we will consider majority voting classifier.
[Bootstrap Aggregation]



Ensemble Techniques, when this technique came

⇓

Kaggle, Hackers Ranks

⇓

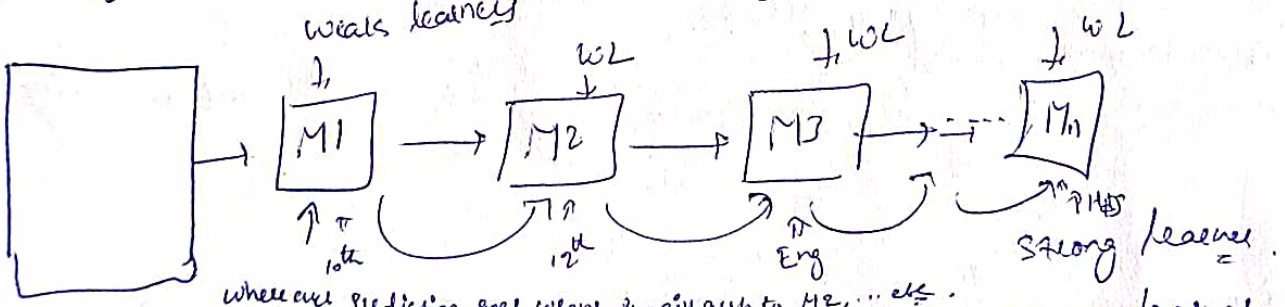
Out performing results.

bagging. { in regression we will take avg / mean.
in classification :- Majority Voting

Bagging :-

1. Random forest classifier
2. Random forest regressor.

Boosting :- we create models sequentially.



where ever prediction goes wrong, it will pass to M2, ... etc.

combine multiple weak learners to create strong learner;

1. Adaboost Regressor / classifier
2. Gradient-boost "
3. XGBoost "

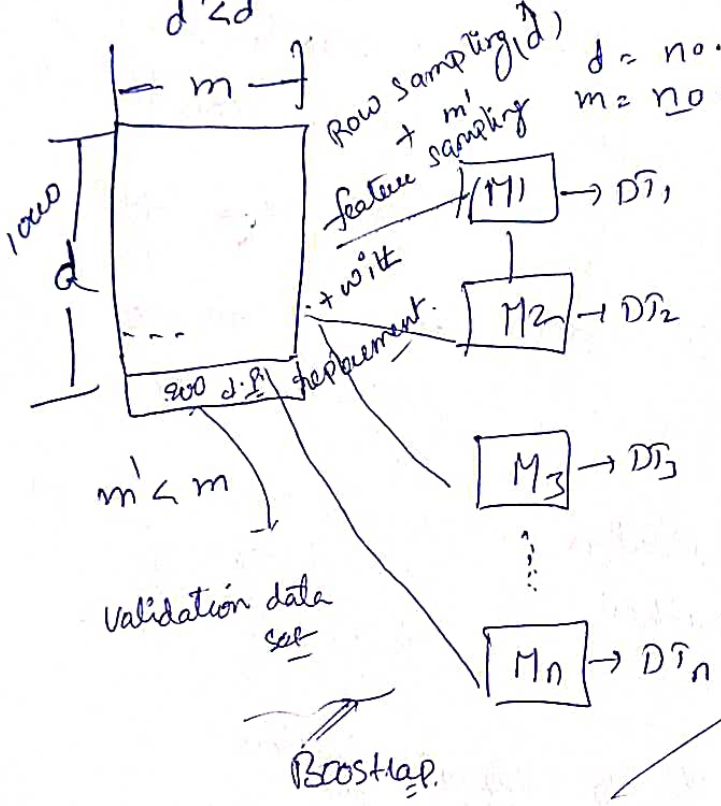
Xtream Gradient boost

Random forest classification & Regression :- Ensemble techniques.

↓↓ bagging & boosting

Combine many models

combine multiple DT's.



M_1 here we consider majority voting

Row sampling :- Sample of row's

M_2 = Row Sampling + feature sampling.

Aggregation.

Average in case of regression.

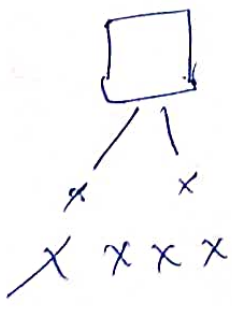
New test data

→ it's a classification Task.

DT's → Overfitting → Prepruning (or) Post Pruning.

1. Low bias → Training
2. High variance → we need to get Low Variance.

using Prepruning / Post Pruning Process.



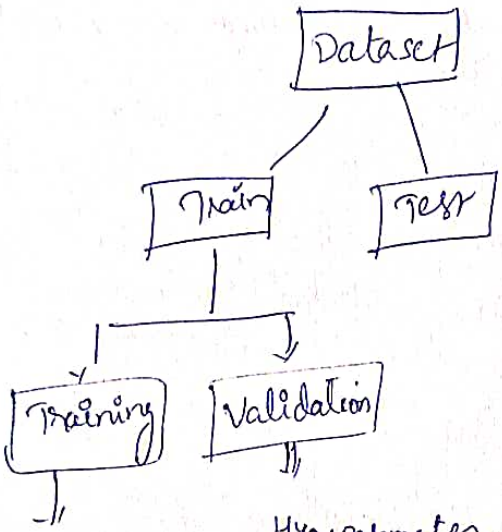
For huge data sets we can't see the DT properly.

generalised models.

Classification :- DT's will get trained using Entropy, gini.

Regression :- DT's " " by Average, mean, threshold MSE

Out of Bag Evaluation:



Model Traing

Hyperparameter tuning } Cross validation
Grid search cv

OOB → There will be a scenario, some of the data Points will never get selected.

those called as Validation data Points.

It will produce some score, it's called

↓

OOB score = TRUE

In random forest we can see

oob_score = TRUE

use this validation data

oob score = 83%

Out of Bag Error ⇒ Validation

1 − oob_score = 1 − 83 = 0.17%

in

Assignment :-

Use bagging classifier and regressor voting classifier, regressor and random forest classifier, regressor on household consumption data & census data