Ensembeing Techniques

Till now we have learnt to use only one ML model at a time

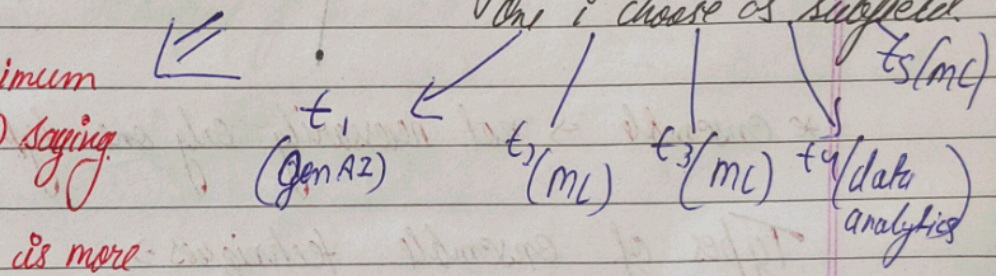
Now, can we make a model integrating more than one ML-algo or using one model to create multiple ml model.

↳ yes we can create

↳ but why we need that ?

* Why we want multiple models ?

Analogy behind it → I go to teacher asked i want a job in data science, which one i choose as subfield.



Now here maximum teachers (3 of 5) saying ML

↳ then ML is more confident field.

Now Similar analogy → we use similar analogy and train multiple models.

m_1, m_2, m_3, m_4, m_5, m_6 → combined prediction and chose one that most of model predicted

↓

by this cost of error ↓

- Ensembles :→ • Combine multiple models
  • Prediction which is more stable and accurate compared to individual models

When we Combine multiple model

↓

It can be of same type / different

of same Algorithm ← → of different algo

DT_1 (max depth = 3)
DT_2 (max depth = 5)
DT_3 (max depth = 7)

Logistic Regression (m_1)
S.V.M.C (m_2)
Decision Tree Classifier (m_3)

* ensemble → not necessarily only one type of algorithm

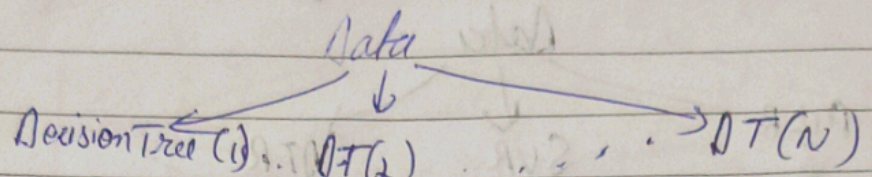
Types of ensemble techniques →

Parallel Technique (Bagging) — one type of model can be used

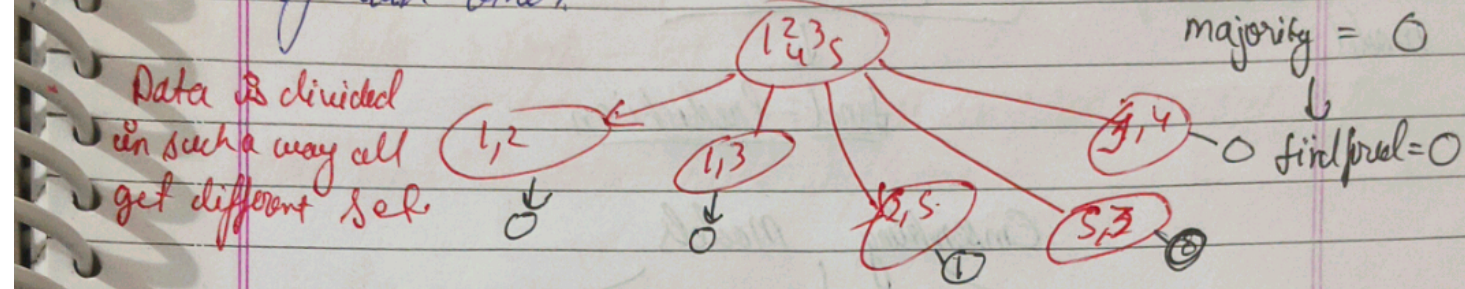
Sequential Technique (Boosting) — one type of model can be used

Parallel + sequential technique (Stacking) → can be used multiple types of model

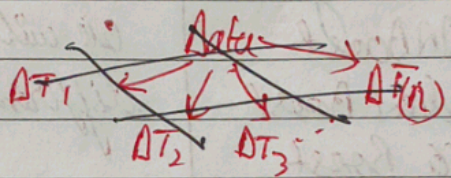
① Parallel Technique of ensemble → Bagging



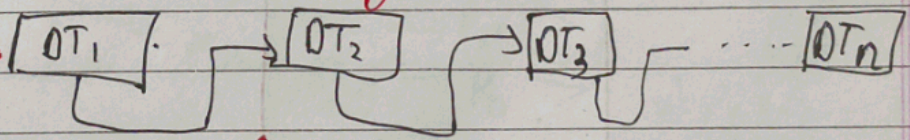
→ All the model are built parallely & independent of each other.



② Sequential technique of ensemble → Boosting



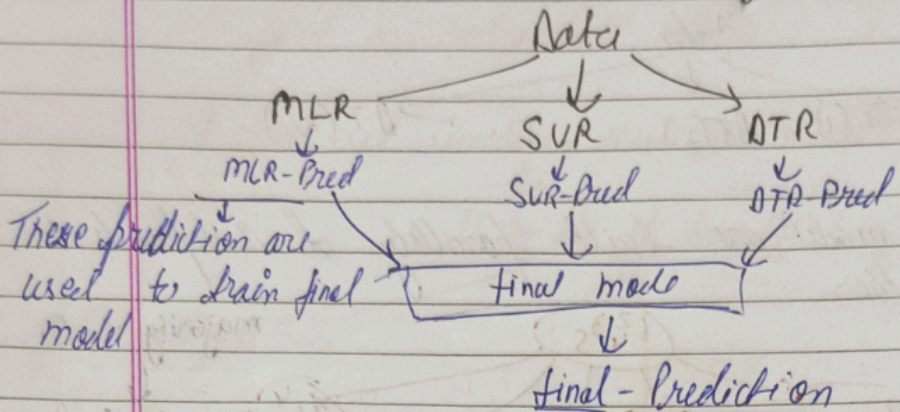
- all model are buildd sequentially and dependent on each other.



- learning from mistake of previous model

→ 1→ DT_1 → trained → then the dp's misclassified are trained on DT_2 ... so upto n^{th} model

* Parallel + Sequential → (Stacking)



Ensembling Models

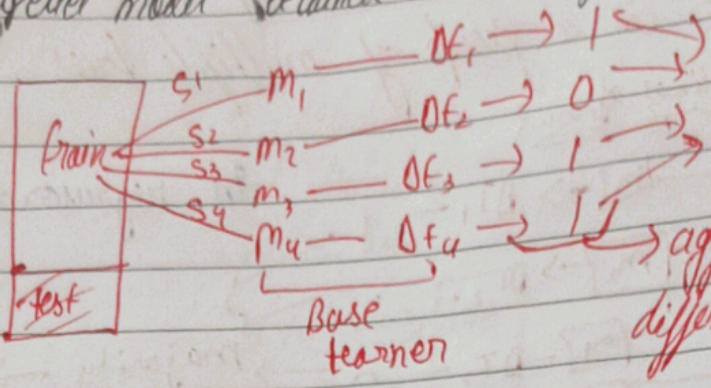
Bagging
- Random forest (Regressor, Classifier)
- Custom Bagging (diff. algo)

Boosting
- ADABoost
- Gradient Boost
- XG Boost
- CatuBoost

Stacking
We will see example of different models

* Bagging technique (with classification problem)

→ Parallel model trained → used for Bagging model



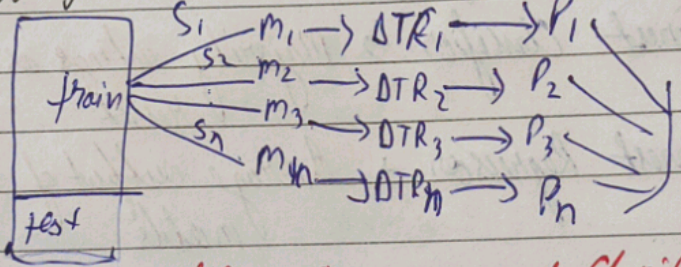
* Data → Train – test

↳ Samples → there samplese (subset) for each model is taken with replacement

Bagging → aggrigation

Bootstrapping → diff. samples with replacement.

* Bagging technique ~~on~~ for regression



~~final~~ final pred

↓

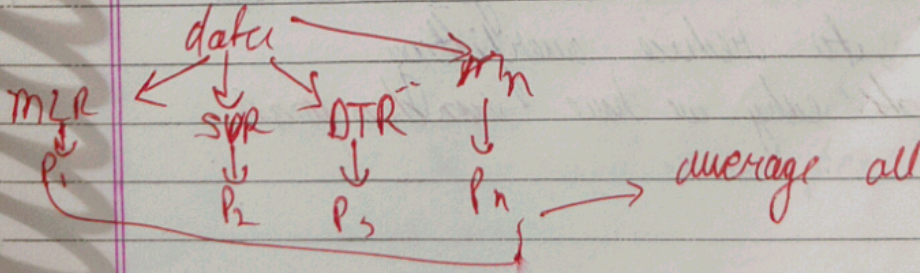
avg of all output

↓

$\Rightarrow \frac{P_1 + P_2 \cdots P_n}{n}$

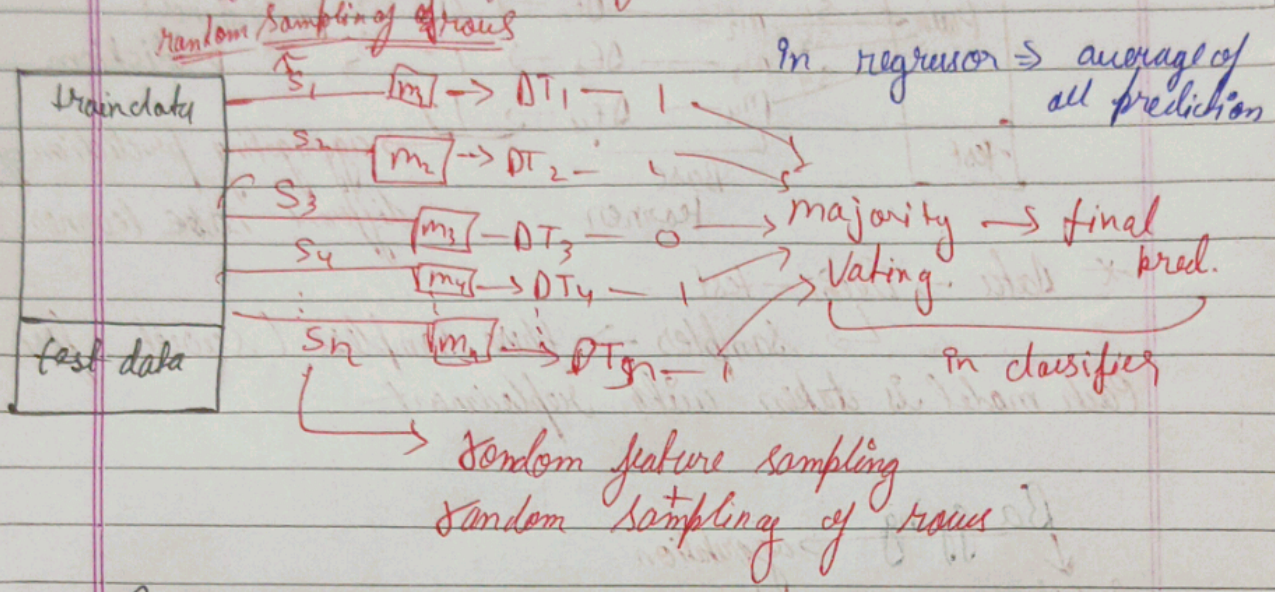
every thing is same as of classification problem

* Custom Bagging technique / pseudo Bagging tech.



Random Forest Classifier and Regressor

Here forest is a group of multiple trees



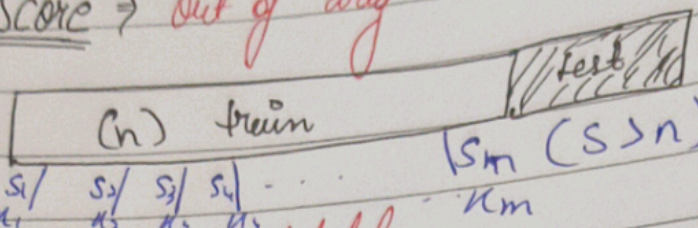
Random Forest → multiple decision tree models in parallel
→ the rows and features will be randomly sampled

(1) Random Forest Classifier → Majority votings as predicted result

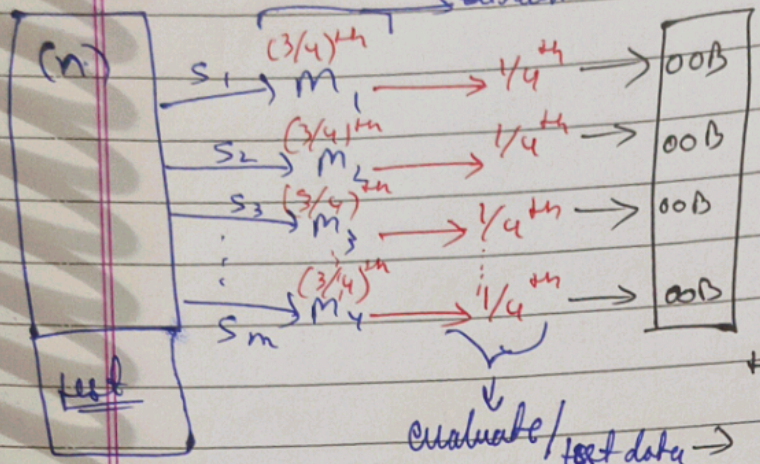
(2) Random Forest Regressor → Average output of all the models

Why Random Forest if we have Decision Tree →?
↳ DT is greedy algo, it will keep splitting until all are memorised for this we have ensemble to reduce overfitting
↳ thats why we have random forest

OOB Score → out of bag score



Samples : randomly selected rows and features (with replacement)



this 1/4 data is nothing but out of bag sample

OOB sample → part of train data, not used in model training for individual models/decision tree.

→ now this act as an validation data for each specific individual DT.

→ {OOB Score 1, OOB Score 2 . . . OOB score m}

→ average of all OOB Score ⇒ OOB Score

↓

act as a validation score for random forest.

↓

if in training itself if oob score is low, model is not performing well on train data.