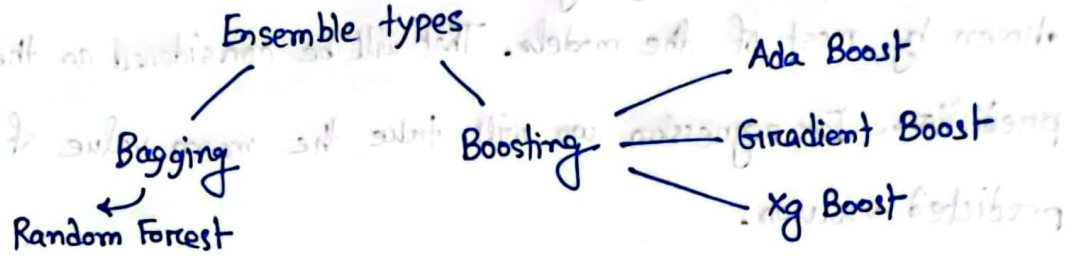
PW Skills (After Naive Bayes)

Ensemble techniques and Bagging:

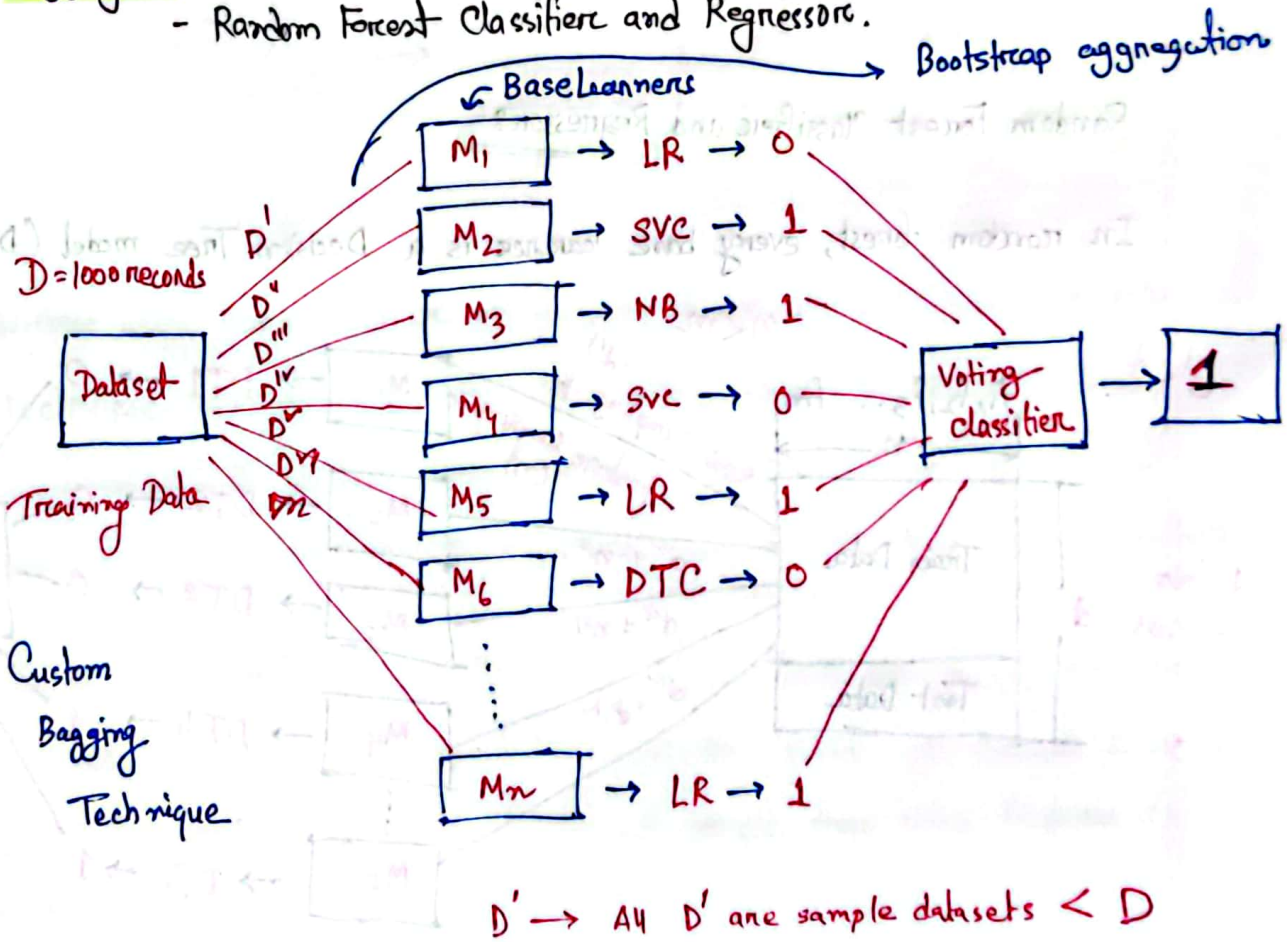
① What is Ensemble?

- Combining multiple models together to get a better accuracy for a particular problem statement.



① Bagging:

- Random Forest Classifier and Regressor.



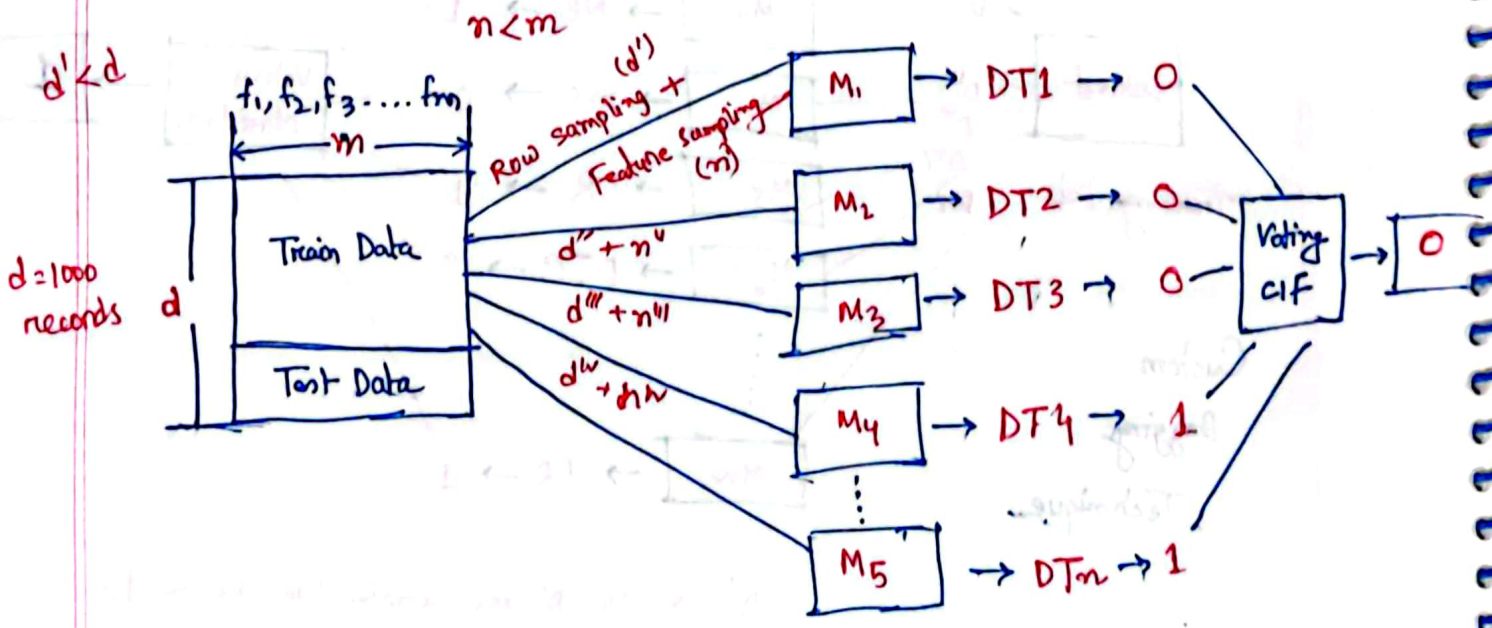
D' → All D' are sample datasets < D

In Bagging technique, From our training dataset, we send sample data for each model (There can be many number of models). Each model will be train on different samples (Bootstrap aggregation) comes from the training dataset and get trained. Then for every model (some model can be used (Algorithms) more than once) we will do prediction. Then we for classification we we will check for majority prediction means what value (0 or 1) is chosen by most of the models. That will be considered as the main prediction. For regression we will take the mean value of all the predicted values.

That's what the bagging technique is.

Random Forest Classifier and Regressor:

In random forest, every base learner is a Decision Tree model (DT)



For Test data →

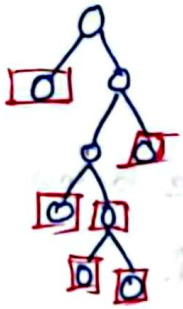
For classification → output → Mgonity voting classifier

For regression → output → Average value of all model predicted values.

Why should we use random forest instead of Decision Tree?

We know that in decision Tree algorithm, we mainly ~~desir~~ continue dividing nodes, till we find the leaf nodes. And without prepruning and postpruning our model gets overe fitted.

Decision Tree



→ Overfitting

Training Acc ↑↑ → Low Bias

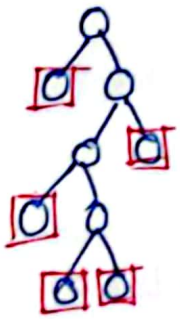
Test Acc ↓↓ → High Variance

Without using prepruning or postpruning techniques, the random forest technique can eliminate the overfitting issue. As we are using many number of decision Tree models, for each sample dataset, that each model get trained, they became perfect with that sample dataset

This is to mention that, in random forest, we provide sample rows and as well as samples features (means we don't provide all features to each model. We provide n number of different sample from total features to each model).

So, each model gets to train with it's sample data and sample features very well. Then each model become a specialist in a particular area, So when we combine all the models and ~~then~~ get an output that provides a really good accuracy for any new test dataset. So random ~~tree~~ Forest work in this way

Decision Tree



→ Overfitting —converted by→ Random Forest ↘

Trainning Acc ↑↑ → Low Bias → ↑↑ → Low Bias

Test Acc ↓↓ → High Variance → ↑↑ → Low Variance

So, Random forest convert the high variance to low variance. For that, model becomes a generalized model.

low Bias
low Variance } Generalized Model → Good accuracy for both train and test data

We can also apply post pruning and pre pruning to get ~~more lot~~ better outcome.

Out of Bag Score Decision Tree:

In random forest we provide sample row data and sample feature data to each and every model after performing row sampling and feature sampling. Sometimes what happens is, while doing that, some row dataset become unused. They don't go to any model for training because of random sampling. It's not mendatory that every model will get unique ~~row~~ sample rows and features. They can be same or partially same. So, in random Forest model, there is a parameter called "oob_score". If we make "oob_score" to True, then the model take the unused data as a ~~val~~ validation dataset and The model can get trained as on them and provide validation accuracy. That validation accuracy is called "oob_score". With that we can know, how our model is performing.