

Random forest

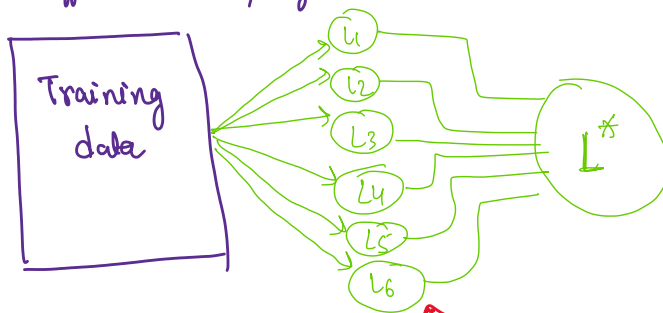
Thursday, 18 April 2024 11:28 AM

Ensemble Learning: Ensemble learning helps improve machine learning results by combining several models.

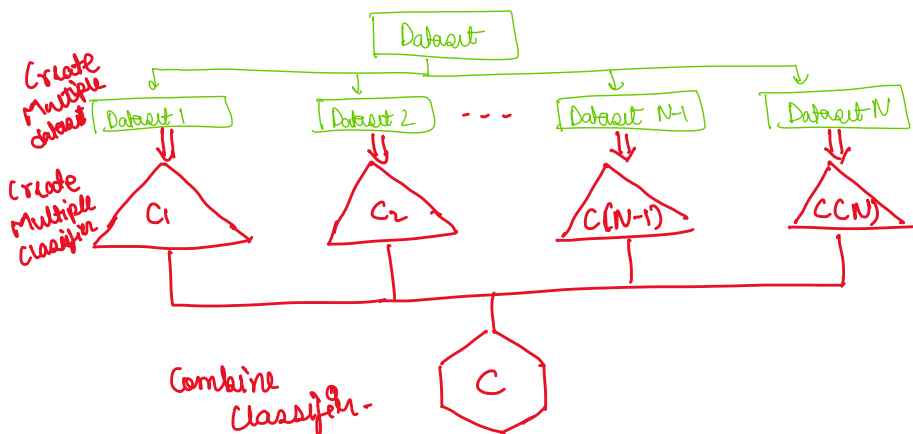
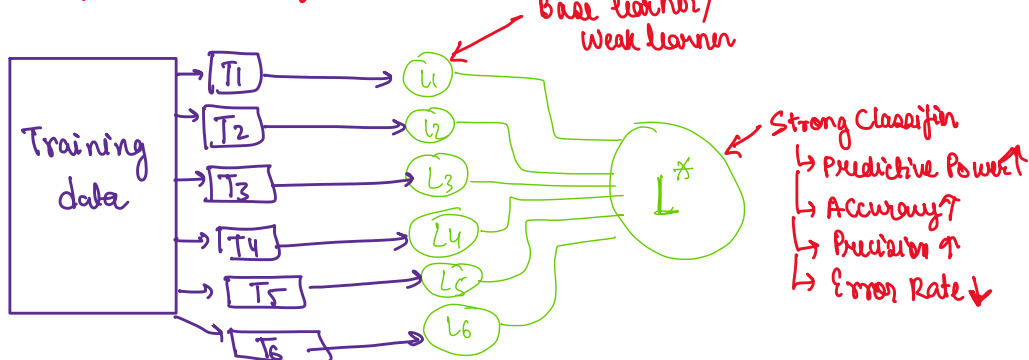
This approach allows the production of better predictive performance compared to single model.

Basic idea is to learn a set of classifiers (experts) and to allow them to vote.

*** Different Model/Algorithm**



Different Training dataset:



**** Random forests or random decision forests** is an ensemble learning method for classification, regression.

**** Random decision forests** correct for decision trees habit of overfitting to their training dataset.

Step 1: Create a Bootstrap Dataset from Original data by Randomly choosing data (repetition is allowed)

AD2	UHV	SML	Section
A	B	C	32
B	C	A	8
C	A	A	8
A	A	D	32

Bootstrap
dataset

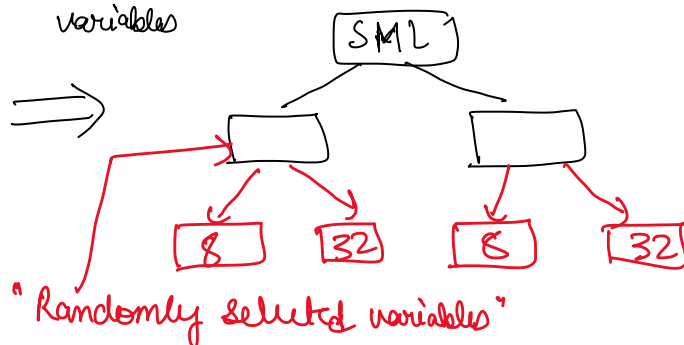
AD2	UHV	SML	Section
A	A	D	32
B	C	A	8
C	A	A	8
A	A	D	32

Duplicate Entry

Step 2: Create Randomized Decision Tree from Bootstrap Dataset, But only use a Random subset of variables (or columns) at each step.

AD2	UHV	SML	Section
A	A	D	32
B	C	A	8
C	A	A	8
A	A	D	32

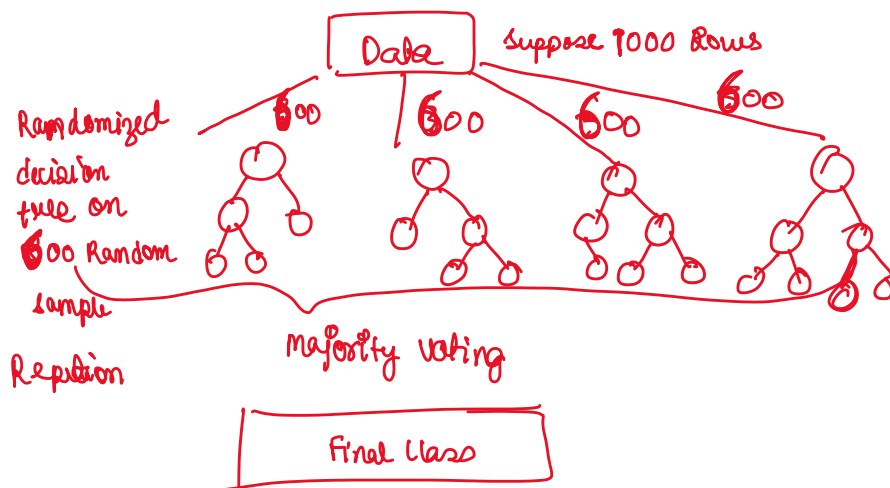
Randomly Selected variables



Now, go back to Step 1 and repeat: Make a new bootstrapped dataset and build a tree considering a subset of variables at each step.

Ideally, you do this 100's of times, this results in wide variety of trees.


Step 3: Finally output of the random forest is the Class selected by most trees.



The variety is what makes random forests more effective than individual decision tree

* Bootstrapping the data plus using the aggregate to make decision is called Bagging

optimizing the random forest

- ① Build a Random forest
- ② Estimate the accuracy of a Random forest
- change the number of variables used per step...
- 

Do this for a bunch of times and then choose the one that is most accurate.

