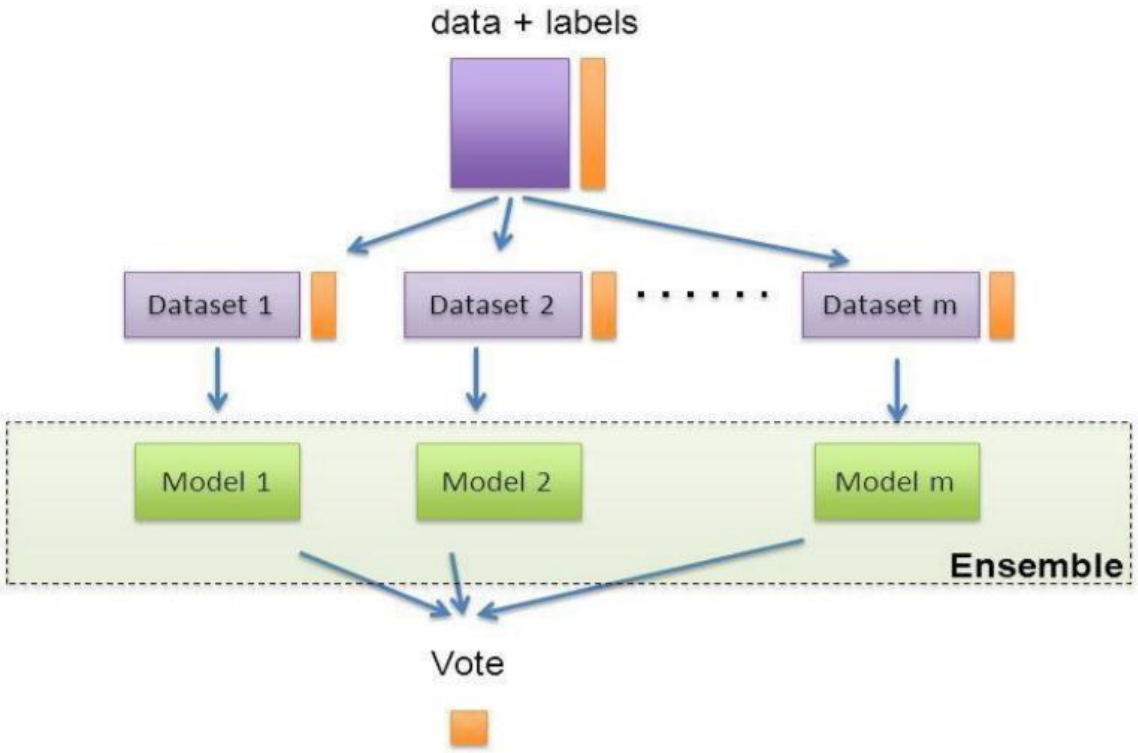
Ensemble Learning

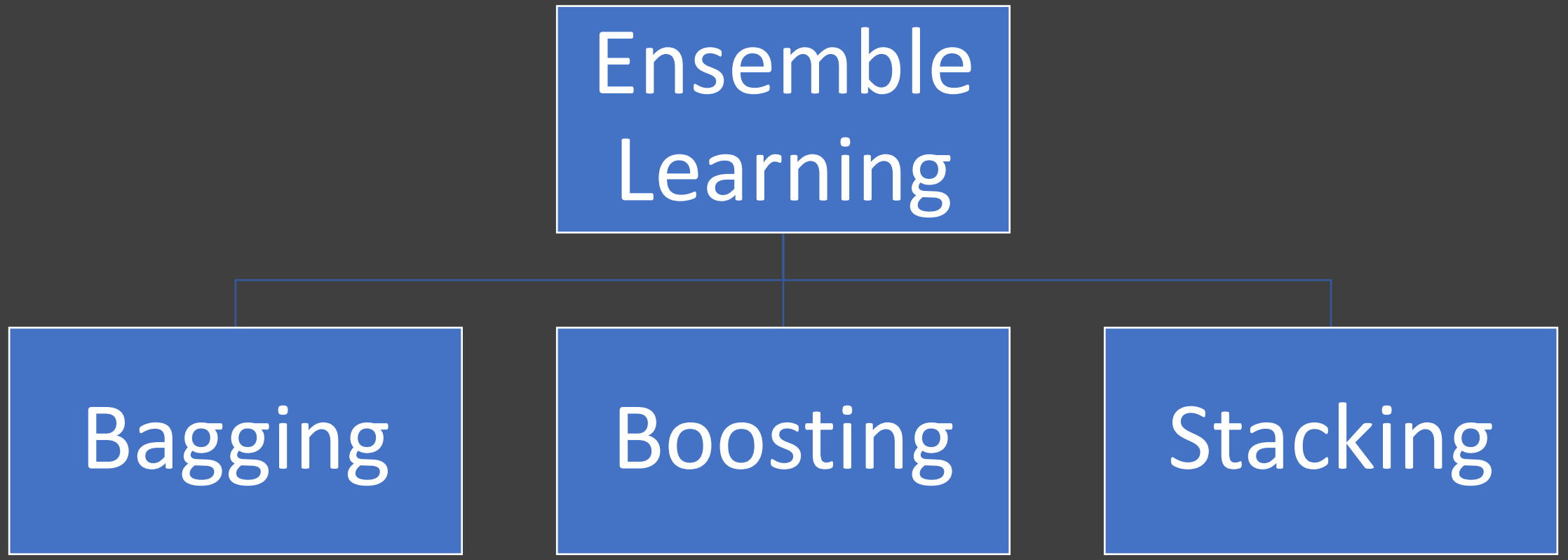
Anshu Pandey

Ensemble Learning

- Ensemble learning is technique that creates multiple models and then combines them to produce improved results

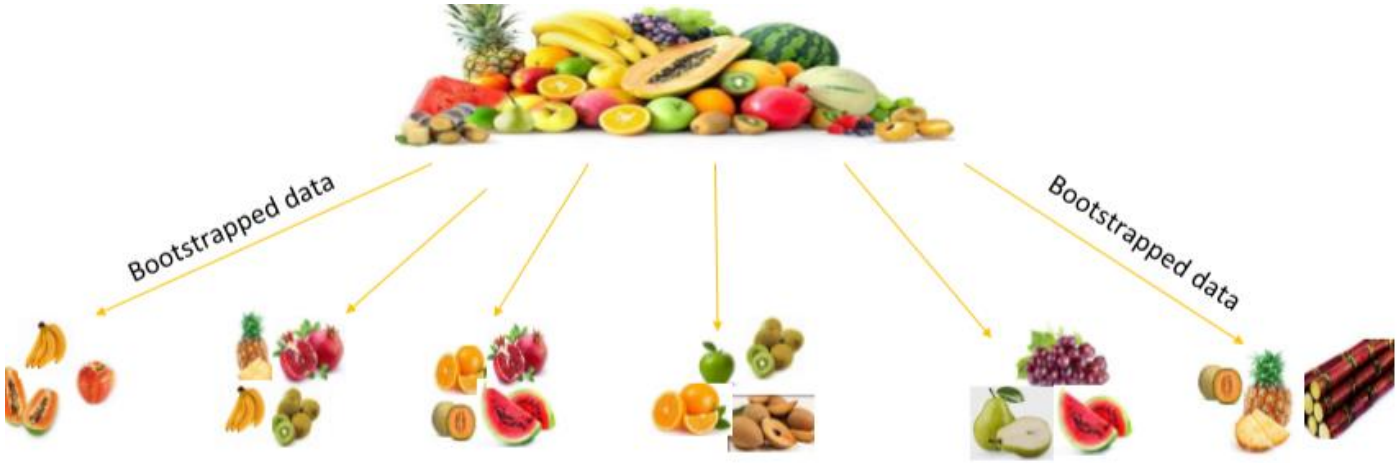


Types of Ensemble Models



Bagging

- Bagging also called Bootstrapped aggregation.
- Bootstrap in statistics denotes sampling with replacement.



Each subset of bootstrapped data is called bags of data.

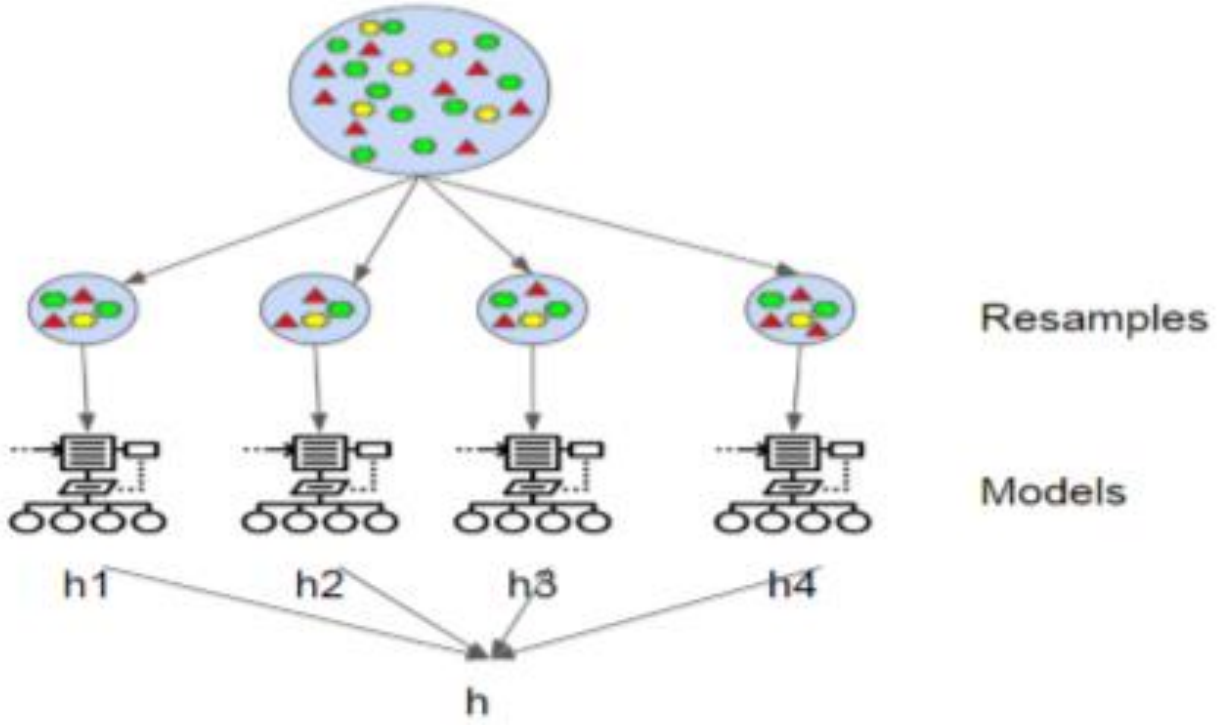
Random Forest

Random Forest algorithm is example of bagging technique.

Random forest is collection of many decision trees

Random Forest

- Random forest (or random forests) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees.



Why Random Forest is called Random?

There are two types of randomness in RF algorithm—

1. Row level
2. Column level

Working of Random Forest

1. Assume number of cases in the training set is N. Then, sample of these N cases is taken at random but *with replacement*. This sample will be the training set for growing the tree.
2. If there are M input variables, a number m<M is specified such that at each node, m variables are selected at random out of the M. The best split on these m is used to split the node. The value of m is held constant while we grow the forest.
3. Each tree is grown to the largest extent possible and there is no pruning.
4. Predict new data by aggregating the predictions of the ntree trees (i.e., majority votes for classification, average for regression).

Validation of Random Forest with Out of Bag data

- Some data points don’t go into the bag for training.
- Those data are called Out of Bag data/sample.
- If this is the entire training data, then only 2/3 of the data is used for model building in Random Forest.
- This 2/3 of the data goes into the bag.

Passenge	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, M	male	22	1	0	A/5 2117	7.25		S
2	1	1	Cumings,	female	38	1	0	PC 17599	71.28	C85	C
3	1	3	Heikkinen	female	26	0	0	STON/O:	7.925		S
4	1	1	Futrelle, M	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr.	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mi	male		0	0	330877	8.458		Q
7	0	1	McCarthy,	male	54	0	0	17463	51.86	E46	S
8	0	3	Palsson, M	male	2	3	1	349909	21.08		S
9	1	3	Johnson, I	female	27	0	2	347742	11.13		S
10	1	2	Nasser, M	female	14	1	0	237736	30.07		C
11	1	3	Sandstron	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, M	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercc	male	20	0	0	A/5. 215	8.05		S
14	0	3	Anderssor	male	39	1	5	347082	31.28		S
15	0	3	Vestrom,	female	14	0	0	350406	7.854		S
16	1	2	Hewlett, M	female	55	0	0	248706	16		S
17	0	3	Rice, Mast	male	2	4	1	382652	29.13		Q
18	1	2	Williams,	male		0	0	244373	13		S
19	0	3	Vander Pl	female	31	1	0	345763	18		S

OOB Error

The prediction accuracy on out of bag(OOB) data is called Out of Bag score. 1- Out of Bag score= Out of Bag error

Smaller the OOB error better the model.

Features and Advantages of Random Forest

Random Forest is one of the most powerful Machine Learning algorithms.

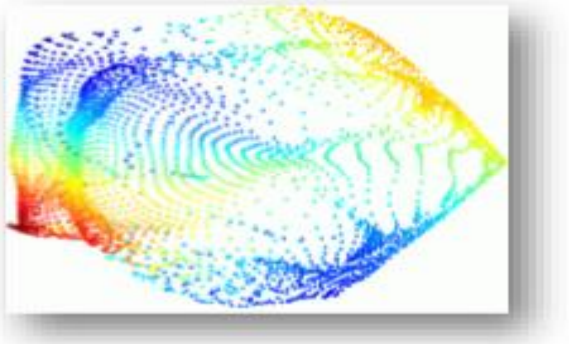
Accuracy

It provides excellent accuracy both for classification & regression.

Handling large scale data

Efficiently handles large scale data making it scalable based on increasing data size

Handling high dimensional data



Takes only important features without deleting unwanted ones.

Hyperparameters of Random Forest

INHERITED FROM DECISION TREE	
criterion : *string, optional (default="gini")*	The function to measure the quality of a split. Supported criteria are "gini" for the Gini impurity and "entropy" for the information gain.
max_depth : *int or None, optional (default=None)*	The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.
min_samples_split : *int, float, optional (default=2)*	The minimum number of samples required to split an internal node
min_samples_leaf : *int, float, optional (default=1)*	The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least min_samples_leaf training samples in each of the left and right branches. This may have the effect of smoothing the model, especially in regression.
max_features : *int, float, string or None, optional (default=None)*	The number of features to consider when looking for the best split: •If "auto", then max_features=sqrt(n_features). •If "sqrt", then max_features=sqrt(n_features). •If "log2", then max_features=log2(n_features). •If None, then max_features=n_features.
max_leaf_nodes : *int or None, optional (default=None)*	Grow a tree with max_leaf_nodes in best-first fashion. Best nodes are defined as relative reduction in impurity. If None then unlimited number of leaf nodes.
min_impurity_decrease : *float, optional (default=0.)*	A node will be split if this split induces a decrease of the impurity greater than or equal to this value.
class_weight : *dict, list of dicts, "balanced", default=None*	Weights associated with classes in the form {class_label: weight}. If not given, all classes are supposed to have weight one.

Hyperparameters of Random Forest

RANDOM FOREST	
n_estimators	integer, optional (default=10) The number of trees in the forest.
oob_score	bool (default=False),if true, it returns OOB score

Case Study