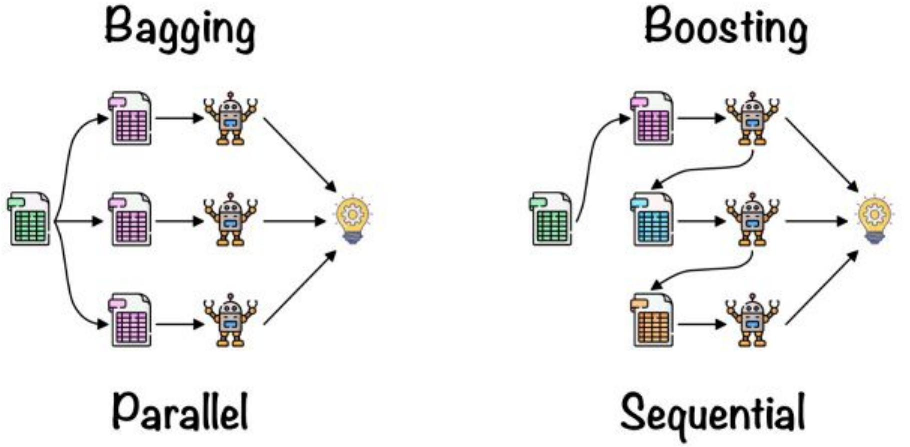
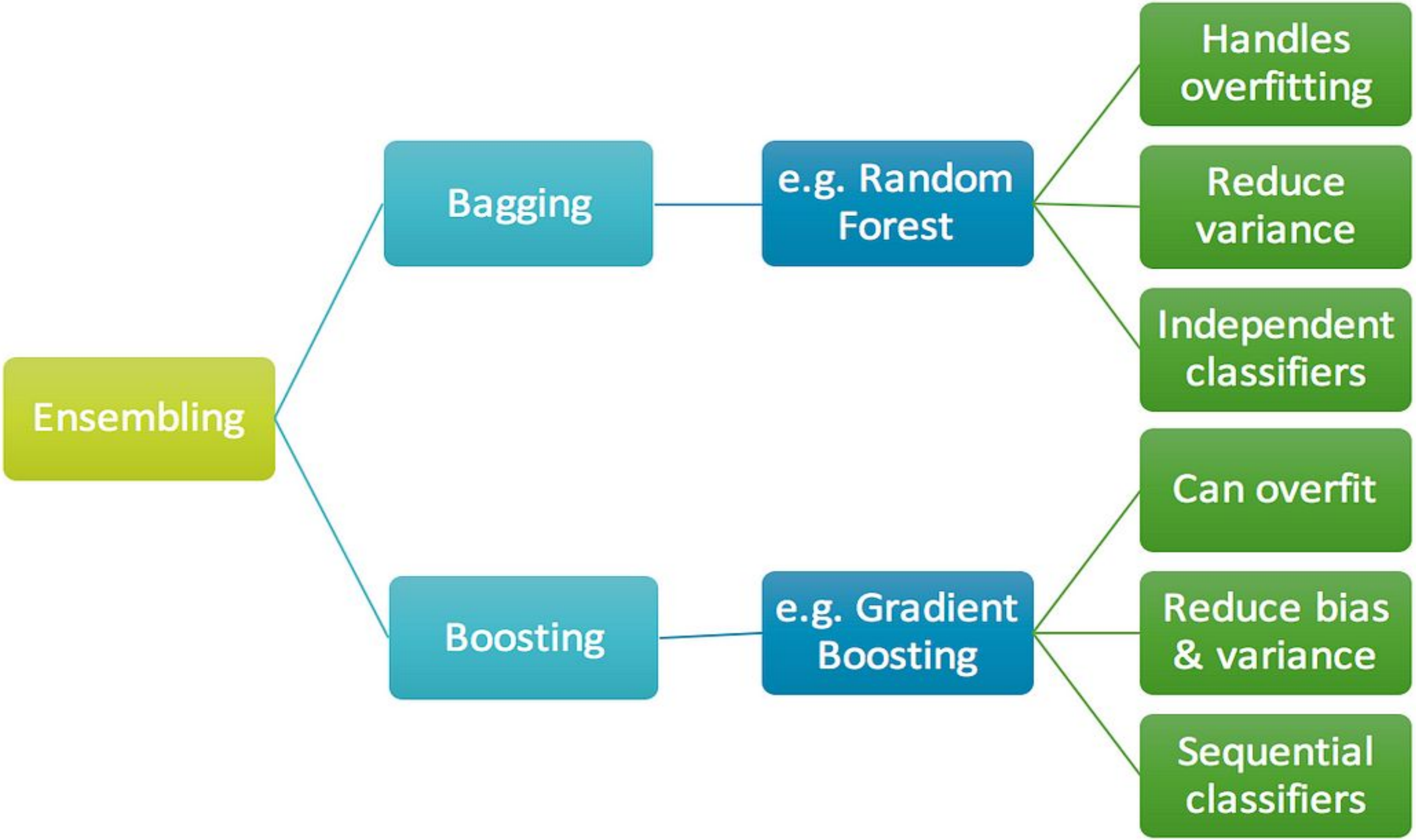
Ensemble learning

Random Forest

Ensemble learning

- Ensemble learning is a machine learning paradigm where multiple models (learners) are trained to solve the same problem.
- By using multiple learners, generalization ability of an ensemble can be much better than single learner.





Ensembling

Bagging

e.g. Random Forest

Handles overfitting

Reduce variance

Independent classifiers

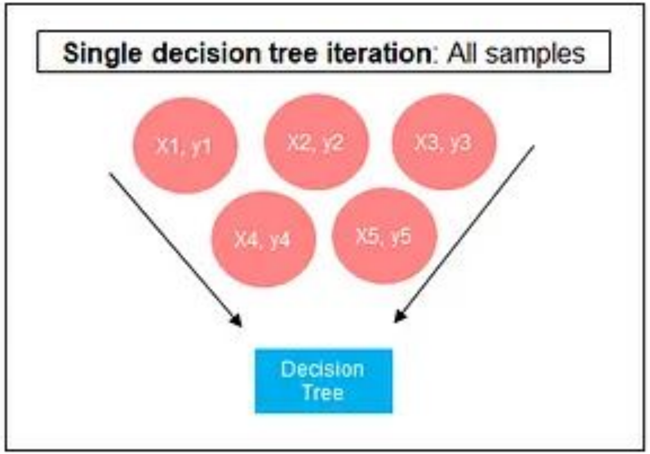
Boosting

e.g. Gradient Boosting

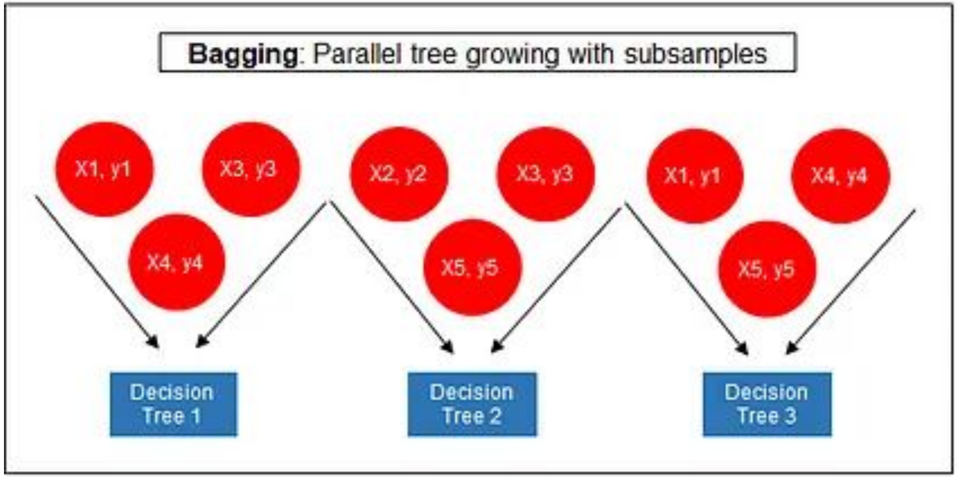
Can overfit

Reduce bias & variance

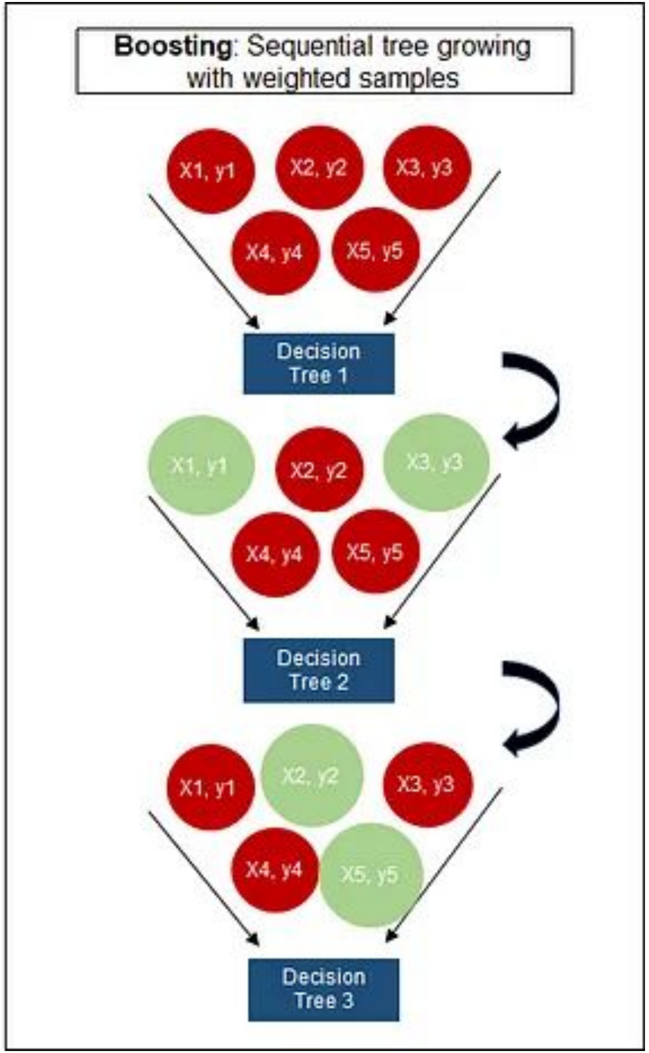
Sequential classifiers



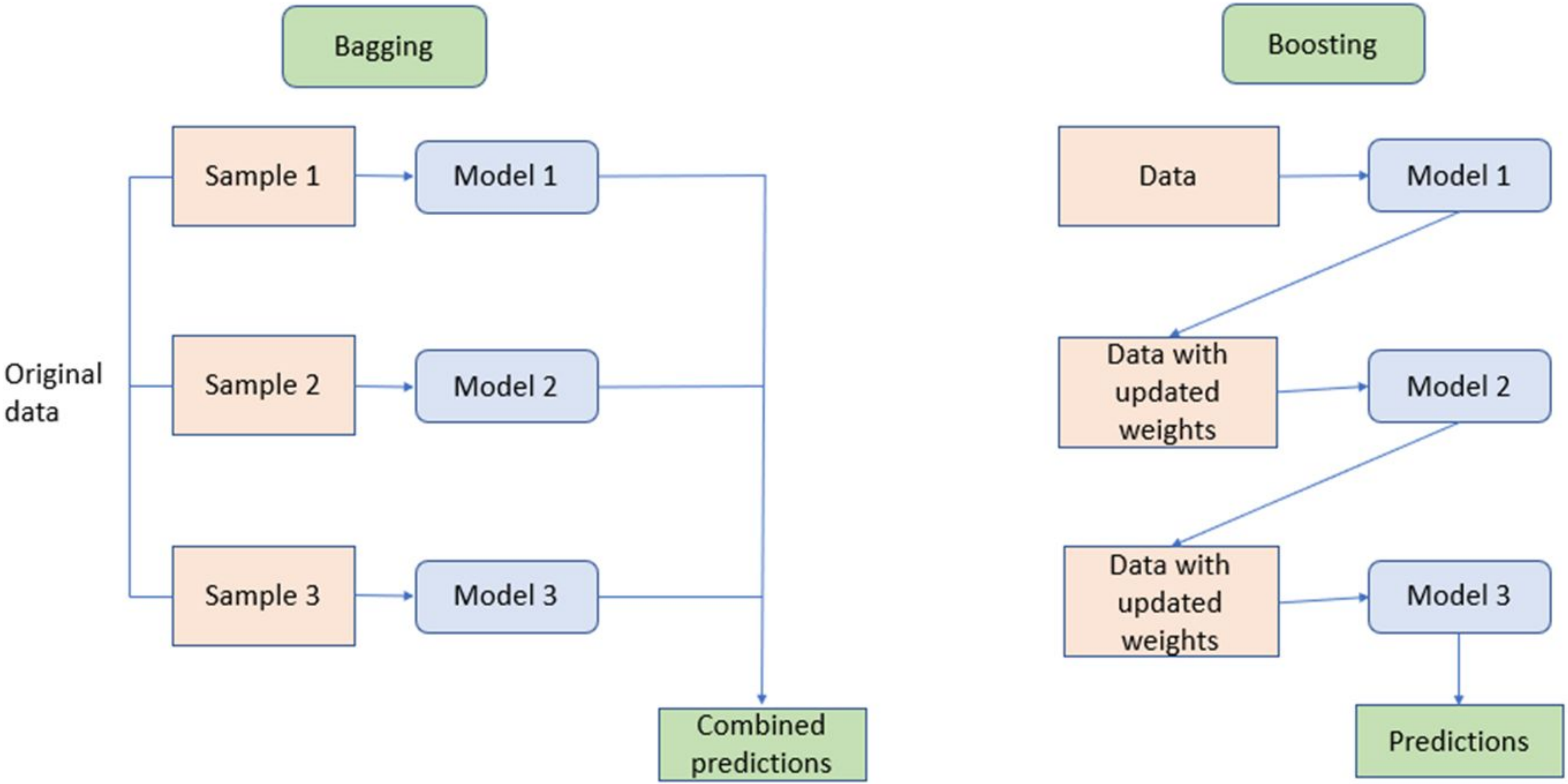
Single decision tree iteration: All samples



Bagging: Parallel tree growing with subsamples



Boosting: Sequential tree growing with weighted samples



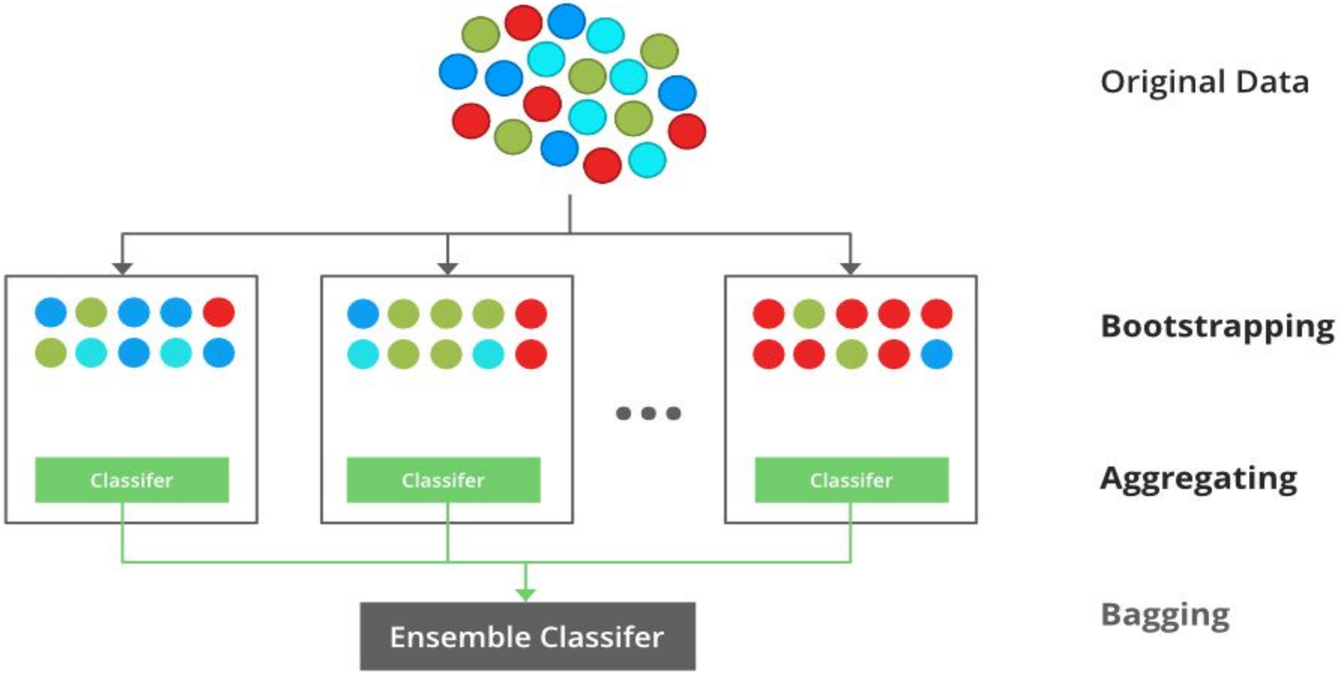
★ Bagging is used mainly in reducing variance and it is a parallel process

★ Boosting is used mainly in reducing bias and it is a sequential process

Ensemble Learning

- **Bagging** and **Boosting** are two types of **Ensemble Learning**. These two decrease the **variance** of a single estimate as they combine several estimates from different models. So the result may be a model with higher stability.
- **Bagging**: It is a homogeneous weak learners’ model that learns from each other independently in parallel and combines them for determining the model average.
- **Boosting**: It is also a homogeneous weak learners’ model but works differently from Bagging. In this model, learners learn sequentially and adaptively to improve model predictions of a learning algorithm.

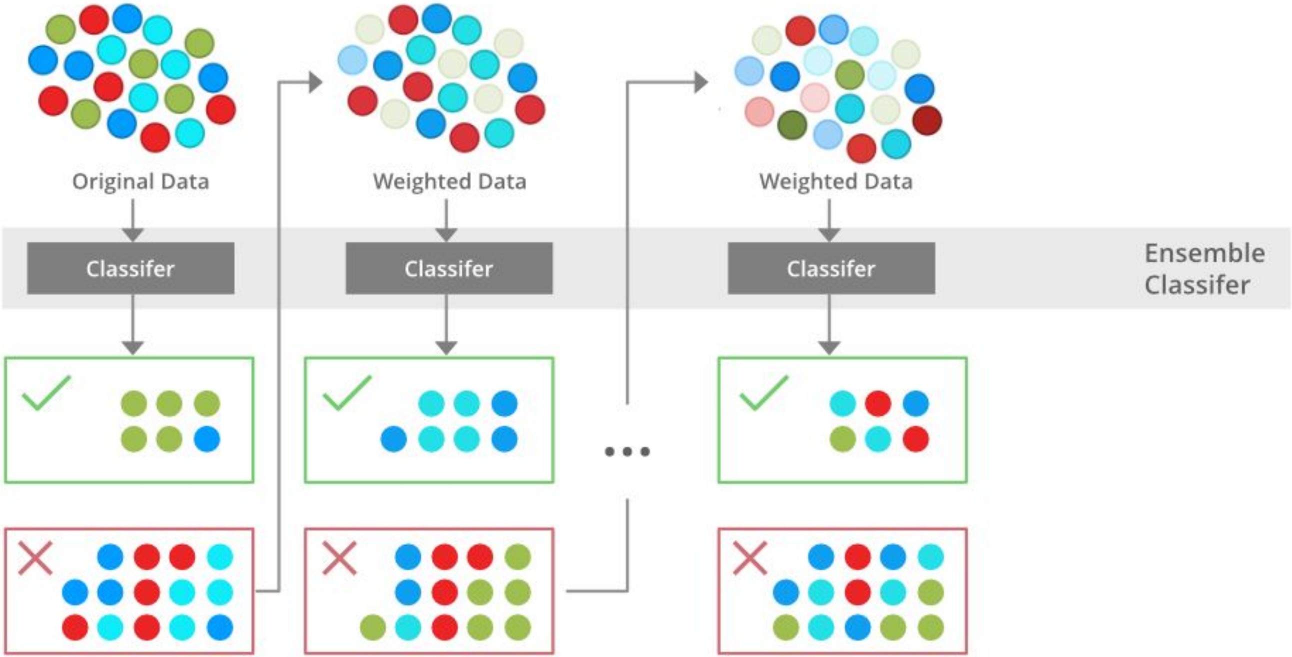
Bagging **in Machine Learning**



Bagging **in Machine Learning**

- **B**ootstrap **A**ggregating, also known as bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression.
- It decreases the variance and helps to avoid overfitting. It is usually applied to decision tree methods.
- Bagging is a special case of the model averaging approach.

Boosting in Machine Learning



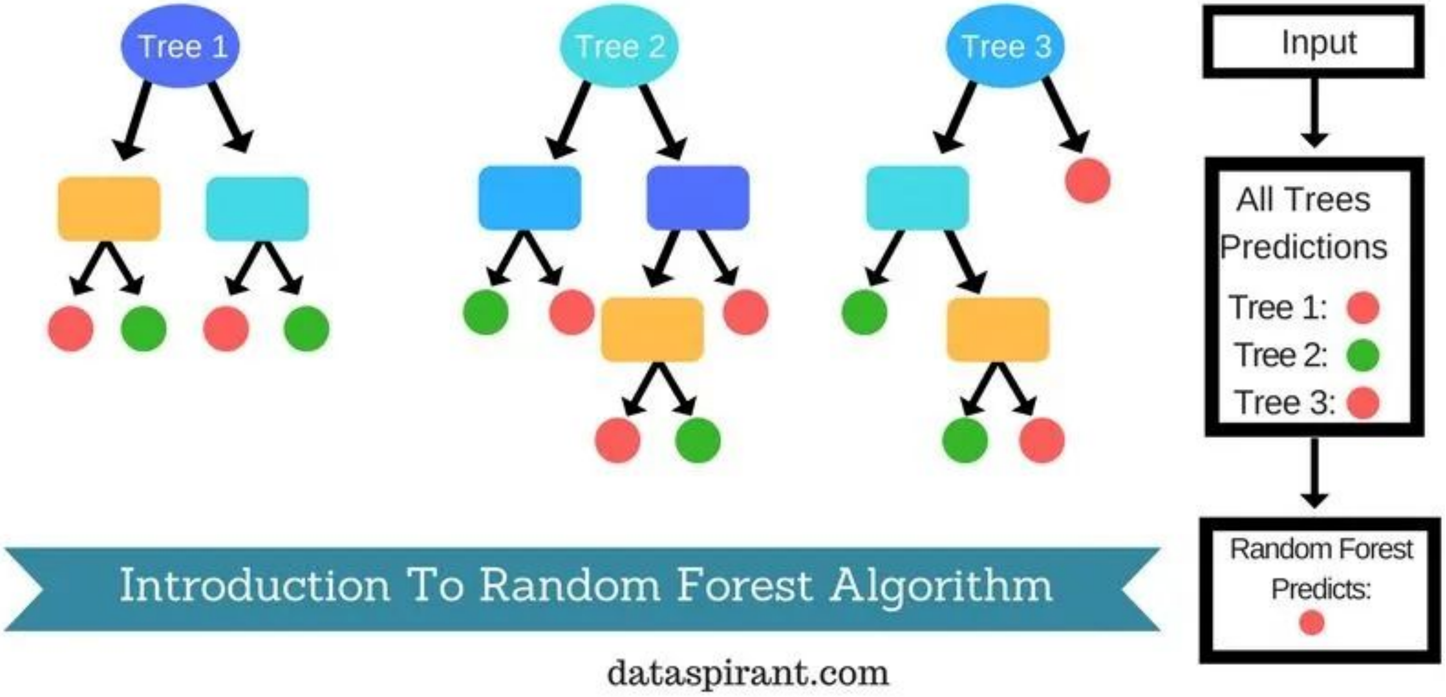
Boosting

- Boosting is the **ensemble learning method** where we build multiple weak learners (same algorithms) in a **SEQUENTIAL** manner.
- All these **weak learners** take the previous models' feedback to improve their power in accurately predicting the missclassified classes.
- Boosting algorithms are one of the best-performing algorithms among all the other Machine Learning algorithms with the best performance and higher accuracies.
- All the boosting algorithms work on the basis of learning from the errors of the previous model trained and tried avoiding the same mistakes made by the previously trained weak learning algorithm.

S.NO	Bagging	Boosting
1.	The simplest way of combining predictions that belong to the same type.	A way of combining predictions that belong to the different types.
2.	Aim to decrease variance, not bias.	Aim to decrease bias, not variance.
3.	Each model receives equal weight.	Models are weighted according to their performance.
4.	Each model is built independently.	New models are influenced by the performance of previously built models.
5.	Different training data subsets are selected using row sampling with replacement and random sampling methods from the entire training dataset.	Every new subset contains the elements that were misclassified by previous models.
6.	Bagging tries to solve the over-fitting problem.	Boosting tries to reduce bias.
7.	If the classifier is unstable (high variance), then apply bagging.	If the classifier is stable and simple (high bias) the apply boosting.
8.	In this base classifiers are trained parallelly.	In this base classifiers are trained sequentially.
9	Example: The Random forest model uses Bagging.	Example: The AdaBoost uses Boosting techniques

Random Forest

Random Forest



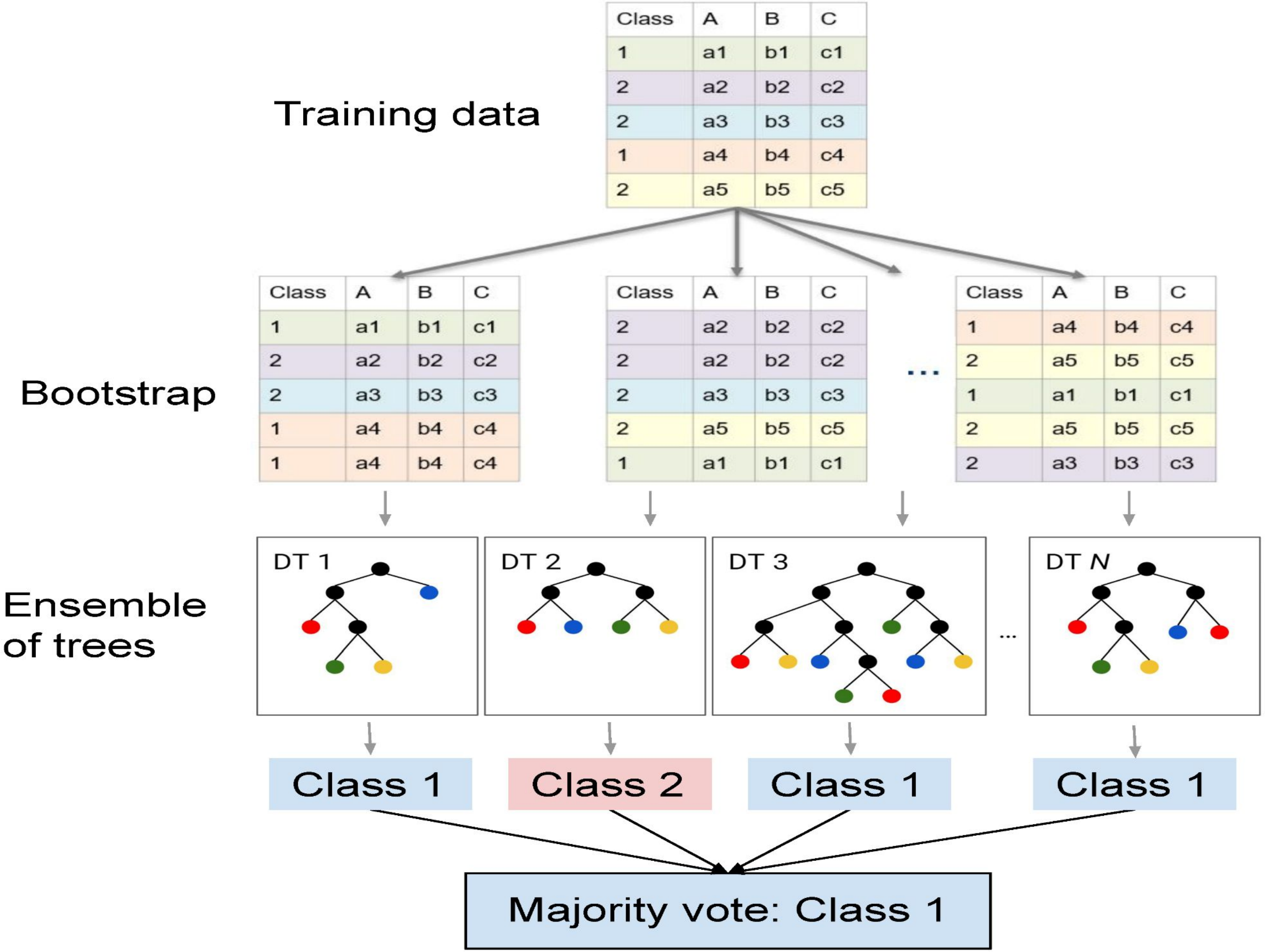
Random Forest

- The random forest algorithm is a supervised classification algorithm.
- As the name suggests, this algorithm creates the forest with a number of trees.
- In general, the **more trees in the forest** the more robust the forest looks like.
- In the same way in the random forest classifier, the **higher the number** of trees in the forest gives **the high the accuracy** results.
- In comparison, the random forest algorithm randomly selects observations and features to build several decision trees and then averages the results.

Why Random forest algorithm

- The same **random forest algorithm** or the random forest classifier can use for both classification and the regression task.
- Random forest classifier will **handle the missing** values.
- When we have **more trees** in the forest, a random forest classifier won’t **overfit** the model.
- The random forest algorithm can be used for feature engineering.
    - This means identifying the most important features out of the available features from the training dataset.

Bagging



Training data

Class	A	B	C
1	a1	b1	c1
2	a2	b2	c2
2	a3	b3	c3
1	a4	b4	c4
2	a5	b5	c5

Bootstrap

Class	A	B	C
1	a1	b1	c1
2	a2	b2	c2
2	a3	b3	c3
1	a4	b4	c4
1	a4	b4	c4

Class	A	B	C
2	a2	b2	c2
2	a2	b2	c2
2	a3	b3	c3
2	a5	b5	c5
1	a1	b1	c1

...

Class	A	B	C
1	a4	b4	c4
2	a5	b5	c5
1	a1	b1	c1
2	a5	b5	c5
2	a3	b3	c3

Ensemble of trees

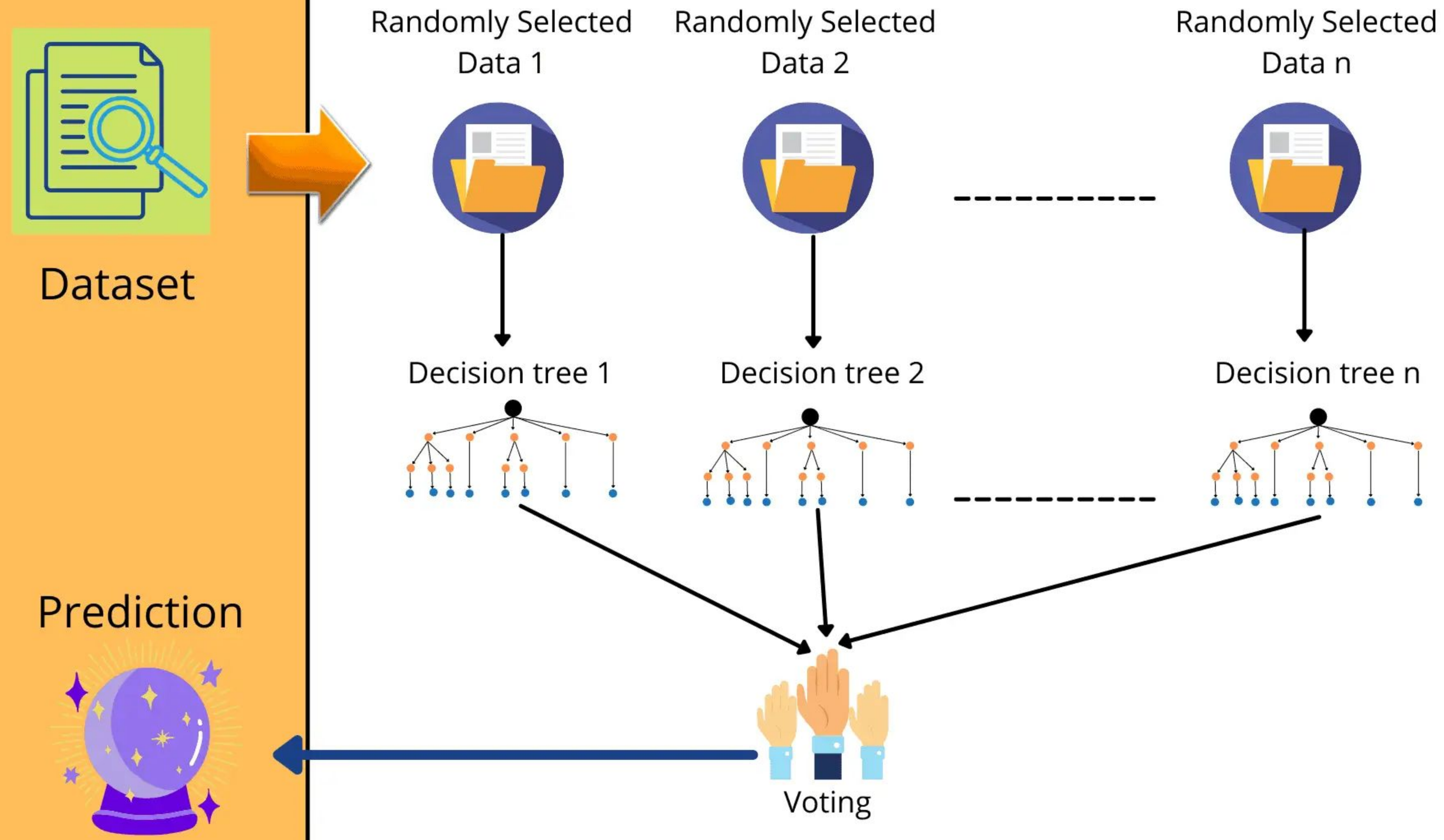
Class 1

Class 2

Class 1

Class 1

Majority vote: Class 1



Dataset

Randomly Selected Data 1

Randomly Selected Data 2

Randomly Selected Data n

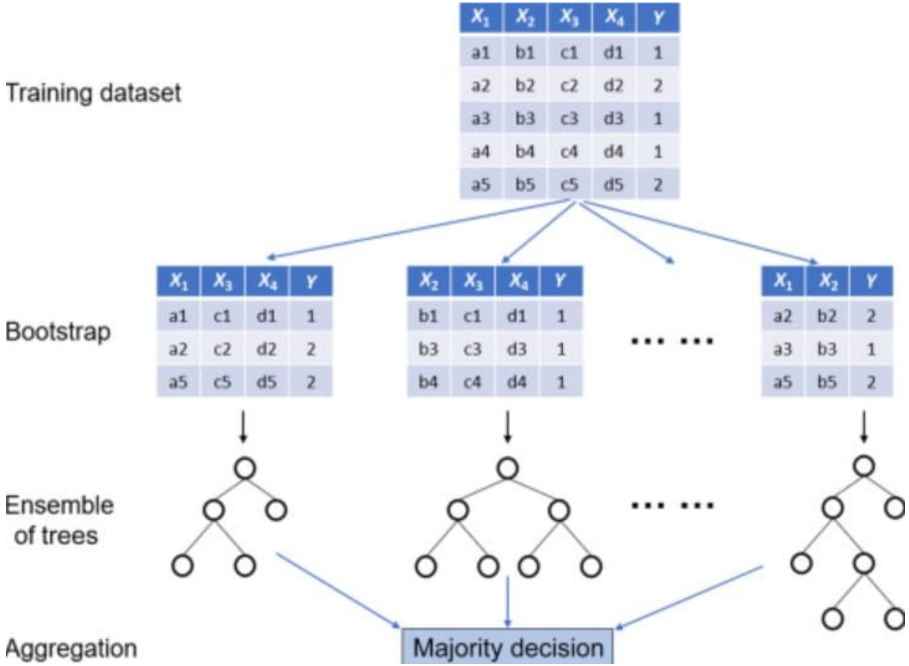
Decision tree 1

Decision tree 2

Decision tree n

Voting

Prediction



Training dataset

X_1	X_2	X_3	X_4	Y
a1	b1	c1	d1	1
a2	b2	c2	d2	2
a3	b3	c3	d3	1
a4	b4	c4	d4	1
a5	b5	c5	d5	2

Bootstrap

X_1	X_3	X_4	Y
a1	c1	d1	1
a2	c2	d2	2
a5	c5	d5	2

X_2	X_3	X_4	Y
b1	c1	d1	1
b3	c3	d3	1
b4	c4	d4	1

…… ……

X_1	X_2	Y
a2	b2	2
a3	b3	1
a5	b5	2

Ensemble of trees

…… ……

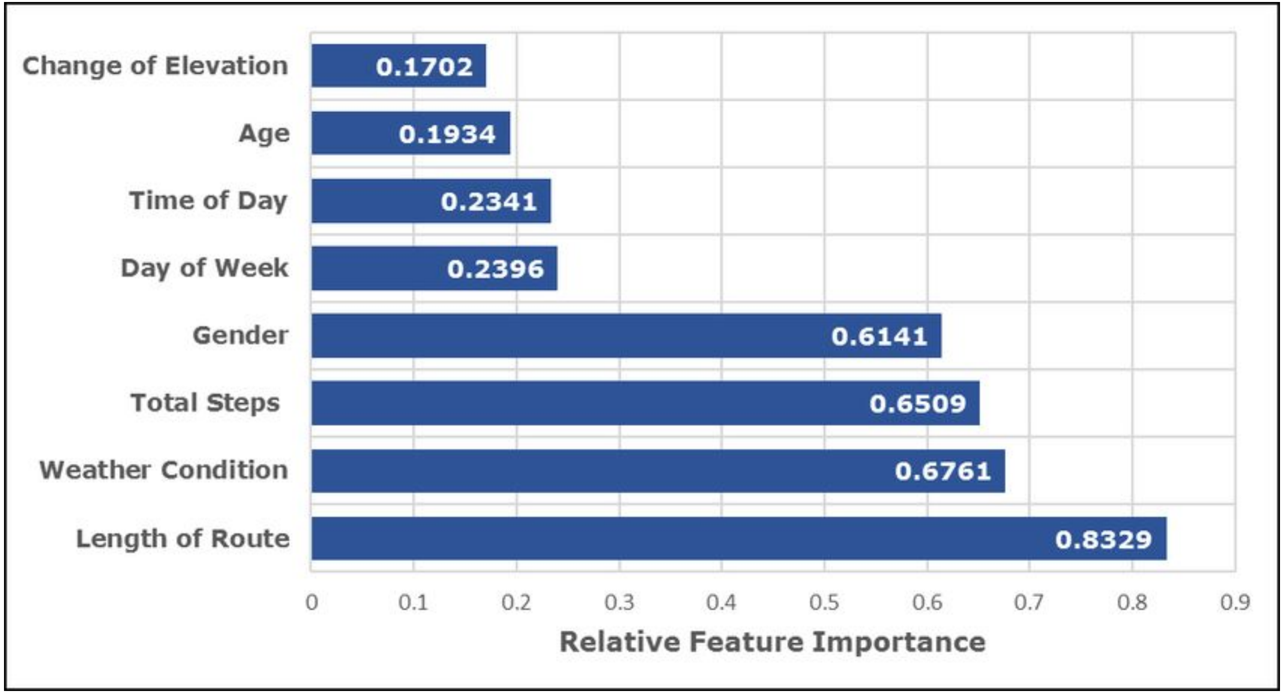
Aggregation

Majority decision

Random Forest

1. Takes the **test features** and use the rules of each randomly created decision tree to predict the oucome and stores the predicted outcome (target)
2. Calculate the **votes** for each predicted target.
3. Consider the **high voted** predicted target as the **final prediction** from the random forest algorithm.

Feature Importance

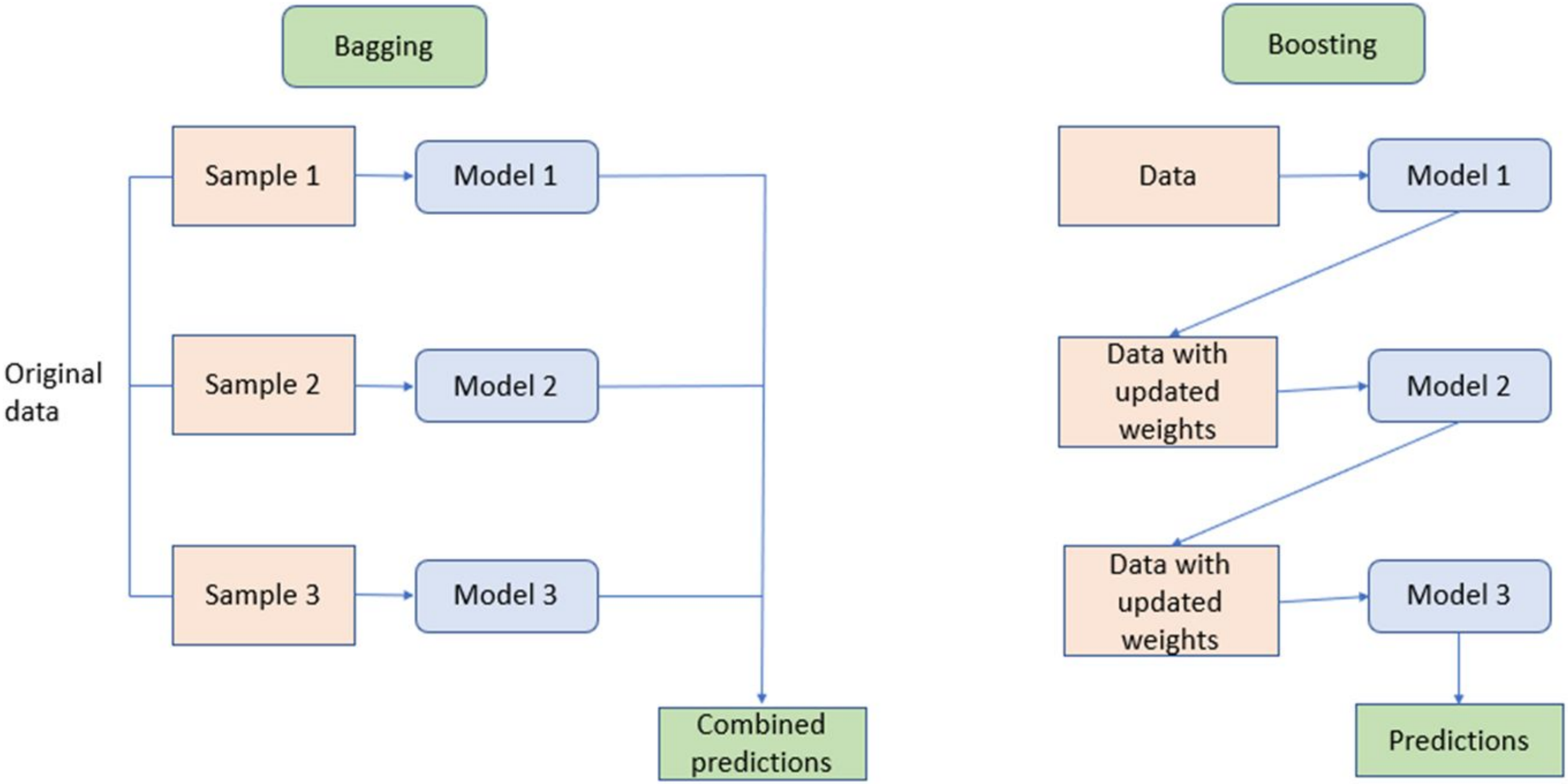


RANDOM FOREST DISADVANTAGES

- Increased accuracy requires more trees
- More trees slow down model

Hyperparameter RANDOM FOREST IN Sklearn

- Firstly, there is the **n_estimators** hyperparameter, which is just the number of trees the algorithm builds before taking the maximum voting or taking the averages of predictions
- Another important hyperparameter is **max_features,** which is the maximum number of features random forest considers to split a node.
- The last important hyperparameter is **min_sample_leaf.** This determines the minimum number of leafs required to split an internal node.



★ Bagging is used mainly in reducing variance and it is a parallel process

★ Boosting is used mainly in reducing bias and it is a sequential process