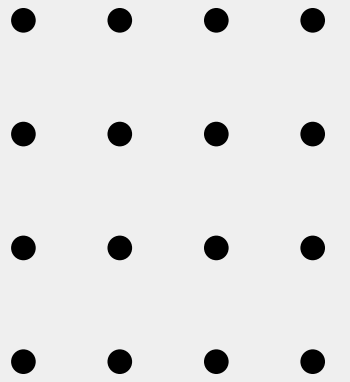


CONTENT



1.) Ensemble Methods

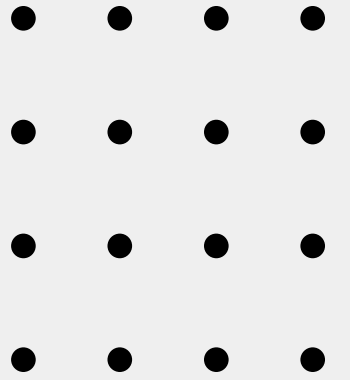
2.) Types: (a) Bagging

(b) Random Forest



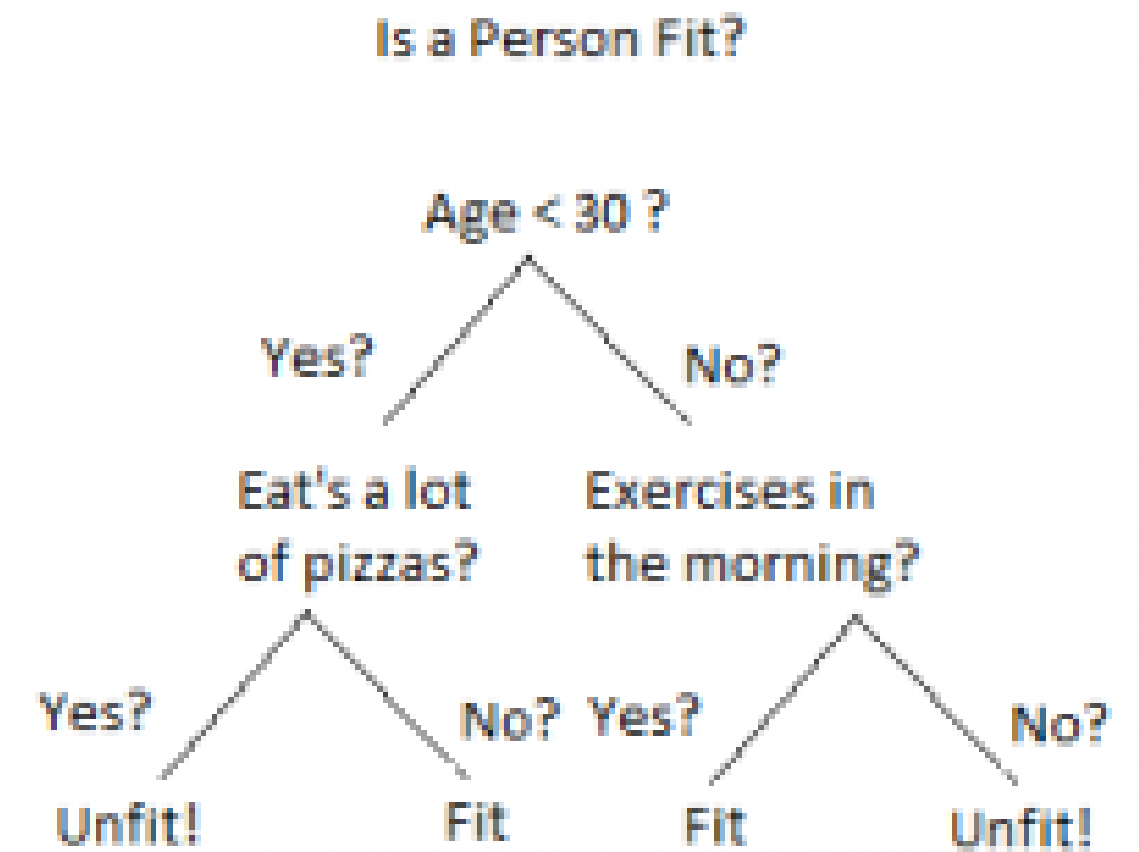
A large orange shape, resembling a stylized 'C' or a partial circle, is positioned on the left side of the slide. A thin black circle is drawn over the top part of this orange shape.

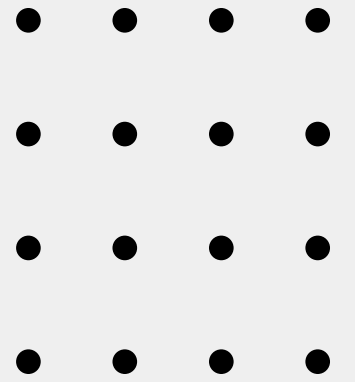
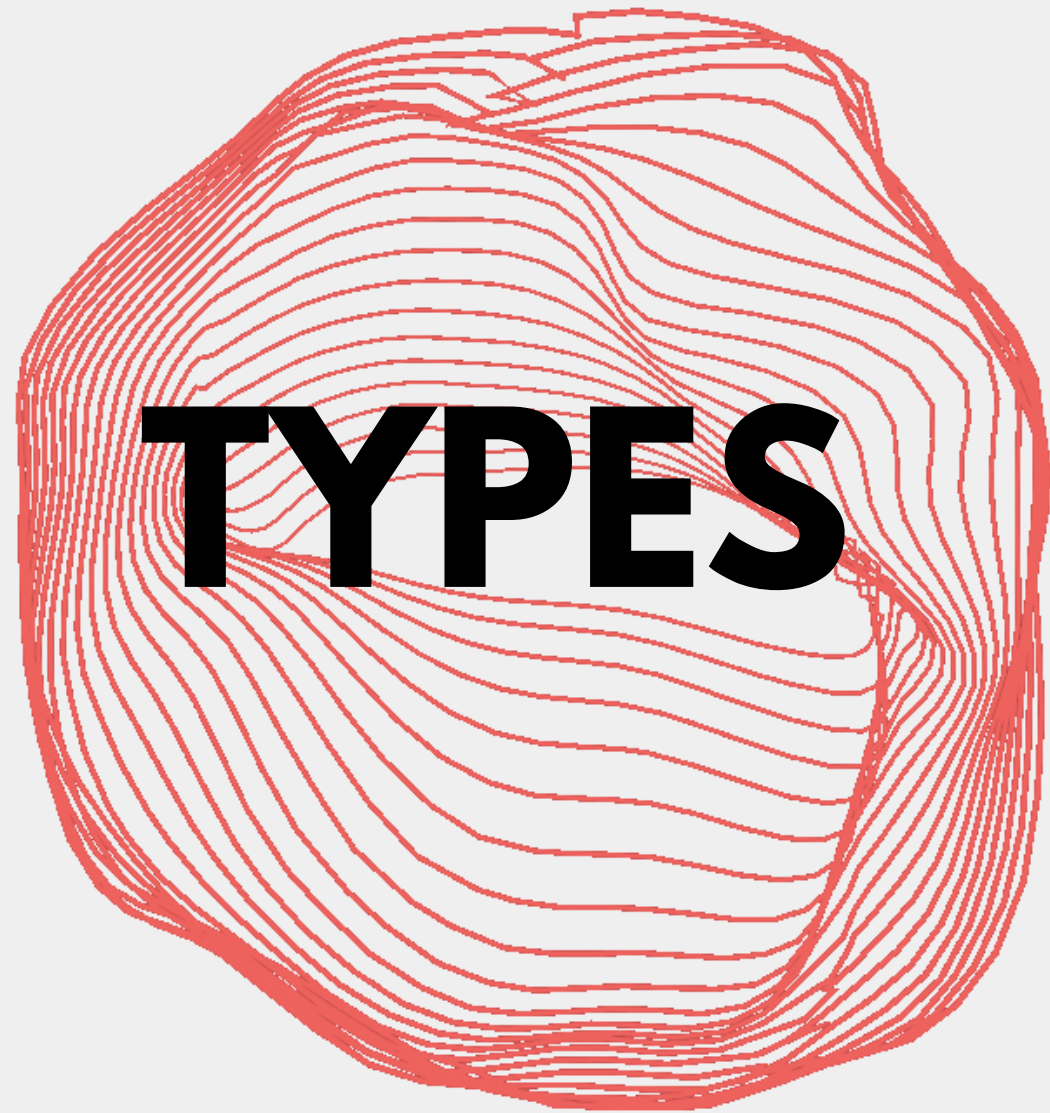
Ensemble Methods



Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model.

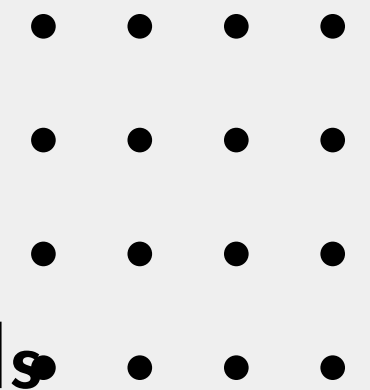
DECISION TREE





(a) Bagging
(b) Random Forest

DESCRIPTION

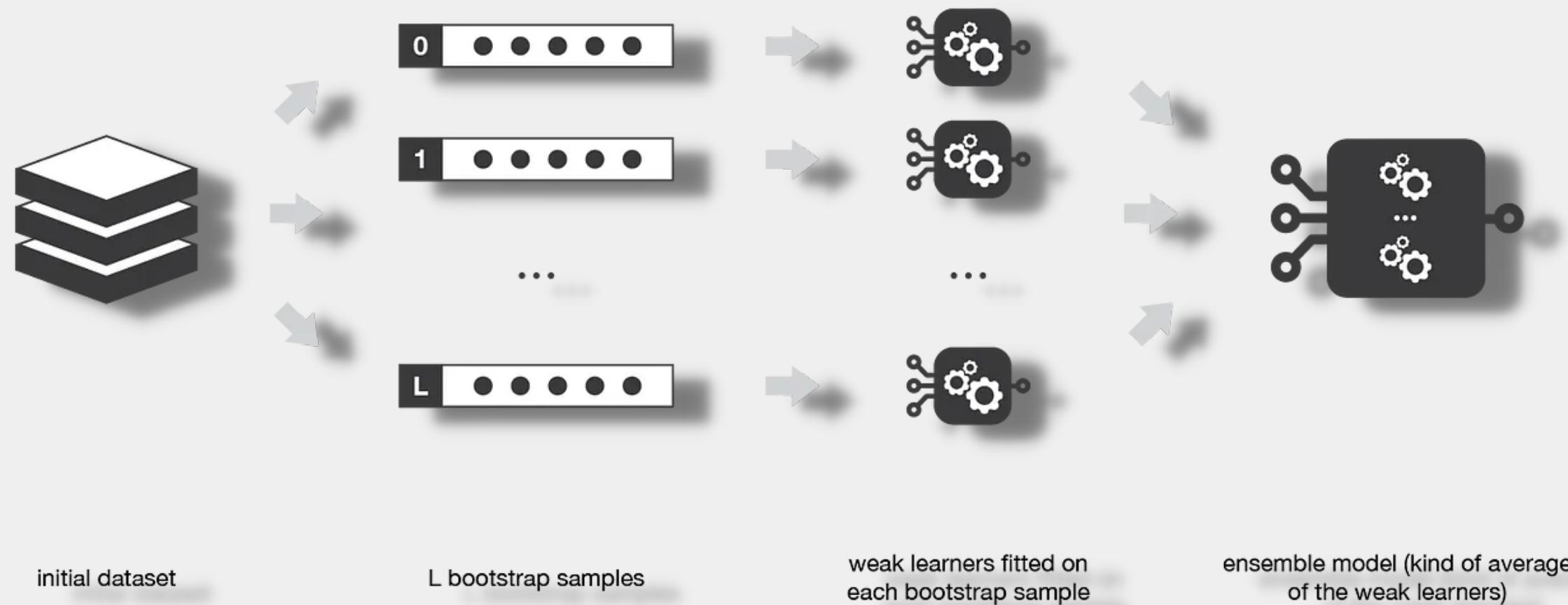


Bagging is a type of Ensemble Method in Machine Learning. The term "Bagging" stands for Bootstrapped Aggregating, which is a technique used to improve the stability and accuracy of machine learning algorithms.

In bagging, the original training set is randomly partitioned into multiple subsets of equal size, with replacement. Each subset is used to train a separate model, and the final prediction is made by combining the predictions from all the individual models. This can be done through simple averaging for regression problems or majority voting for classification problems.

A decorative graphic consisting of several concentric orange circles.

Bagging



The idea behind bagging is to reduce the variance of the model by combining the predictions of multiple models trained on different subsets of the data. By doing this, bagging helps to reduce overfitting and improve the overall accuracy of the model.

Bagging is commonly used with decision trees, but can be applied to other algorithms as well. It is a simple and effective technique that is often used as a baseline for comparison with more complex ensemble methods, such as Random Forest and Boosting.



Example

Suppose we have a dataset of customer behavior data, and we want to build a model that predicts whether a customer will make a purchase or not. We can use decision trees as the base algorithm and apply bagging to improve its accuracy.

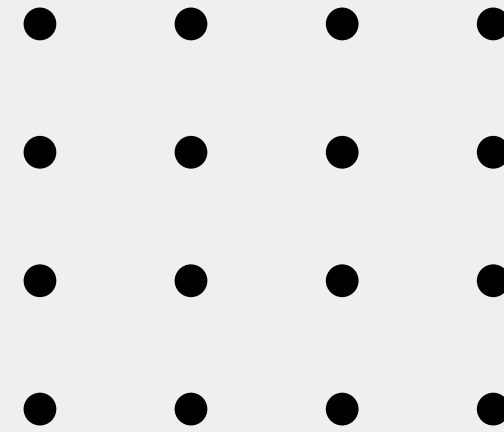
Step 1: Split the dataset into subsets: We'll randomly partition the dataset into 5 subsets of equal size, with replacement.

Step 2: Train 5 separate decision tree models: We'll use each subset to train a separate decision tree model.

Step 3: Make predictions: We'll use each decision tree model to predict a given customer.

Step 4: Combine predictions: Finally, we'll combine the predictions from all 5 decision tree models to make the final prediction. For example, we can take the majority vote on all 5 models to decide whether the customer will make a purchase or not.

By combining the predictions of multiple models trained on different subsets of the data, bagging helps reduce overfitting and improve the model's overall accuracy. This is because the variance of the combined model is reduced by the diversity of the individual models.

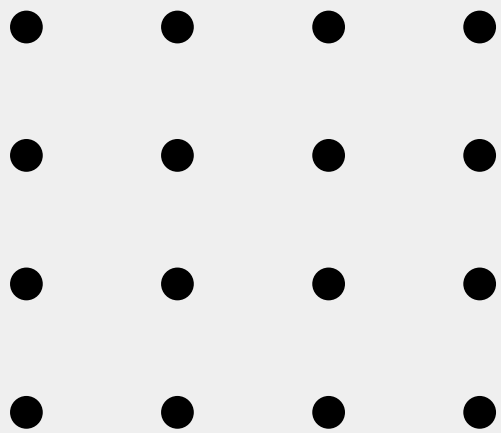


RANDOM FOREST


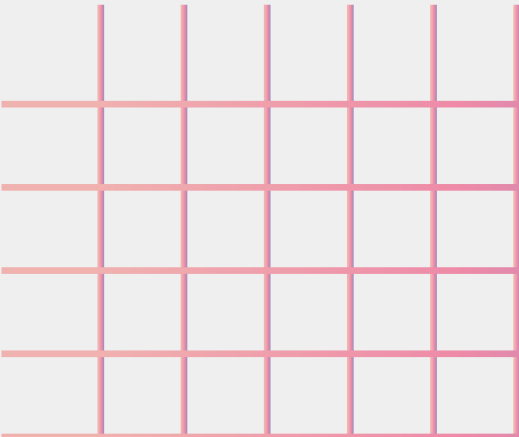




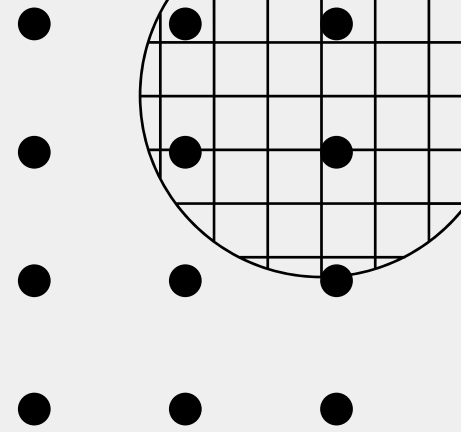
DESCRIPTION



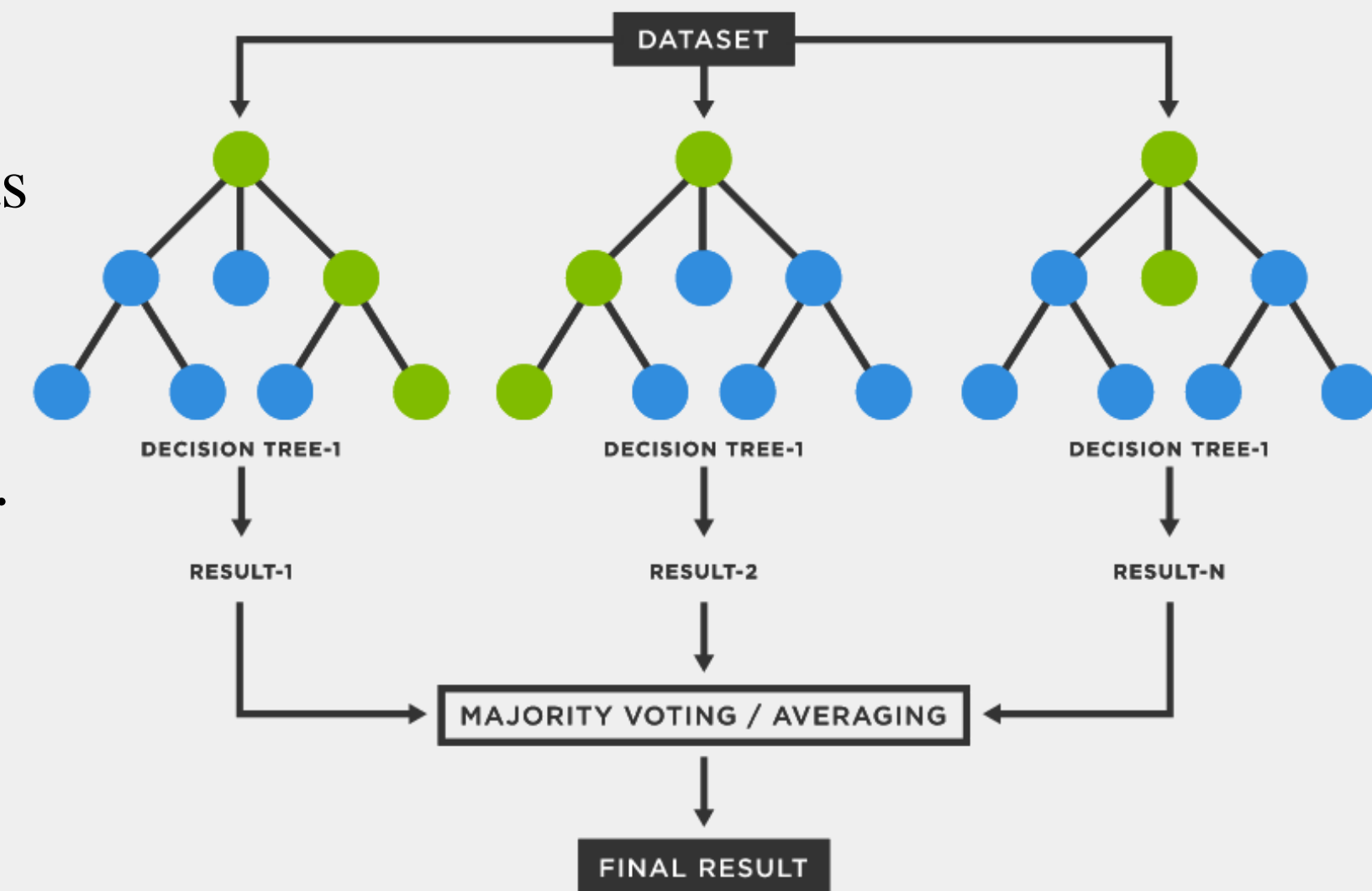
>Random forest is a popular machine learning algorithm used for classification and regression tasks. It works by combining multiple decision trees that are trained on different subsets of the features and samples in the dataset. The trees are constructed using random feature selection and bootstrapped samples, which reduces overfitting and improves the generalization of the model. The final prediction is made by averaging the predictions of all the trees. Random forest is widely used in various fields such as finance, healthcare, and natural language processing

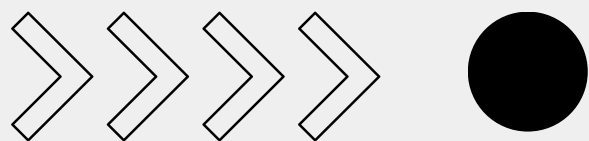


BASIC WORKING



1. Select the number of decision trees to include in the forest, as well as any other hyperparameters (such as the maximum depth of each tree).
2. For each decision tree, randomly select a subset of the training data to use for building that tree (this is known as bootstrapping).
3. For each split in each decision tree, randomly select a subset of the features to consider as potential split points.
4. Build each decision tree using the bootstrapped training data and the selected features.
5. To make a prediction for a new data point, pass that data point down each decision tree in the forest, and use the average (for regression) or mode (for classification) of the predicted values from all trees as the final prediction.





**Easy to determine
feature importance**

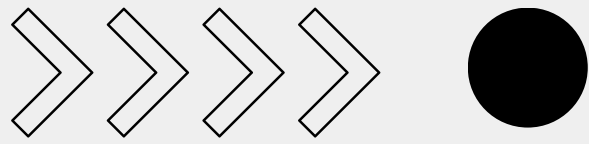
**Provides
flexibility**

PROS

**Ability to handle
large datasets**

**Reduced risk
of overfitting**





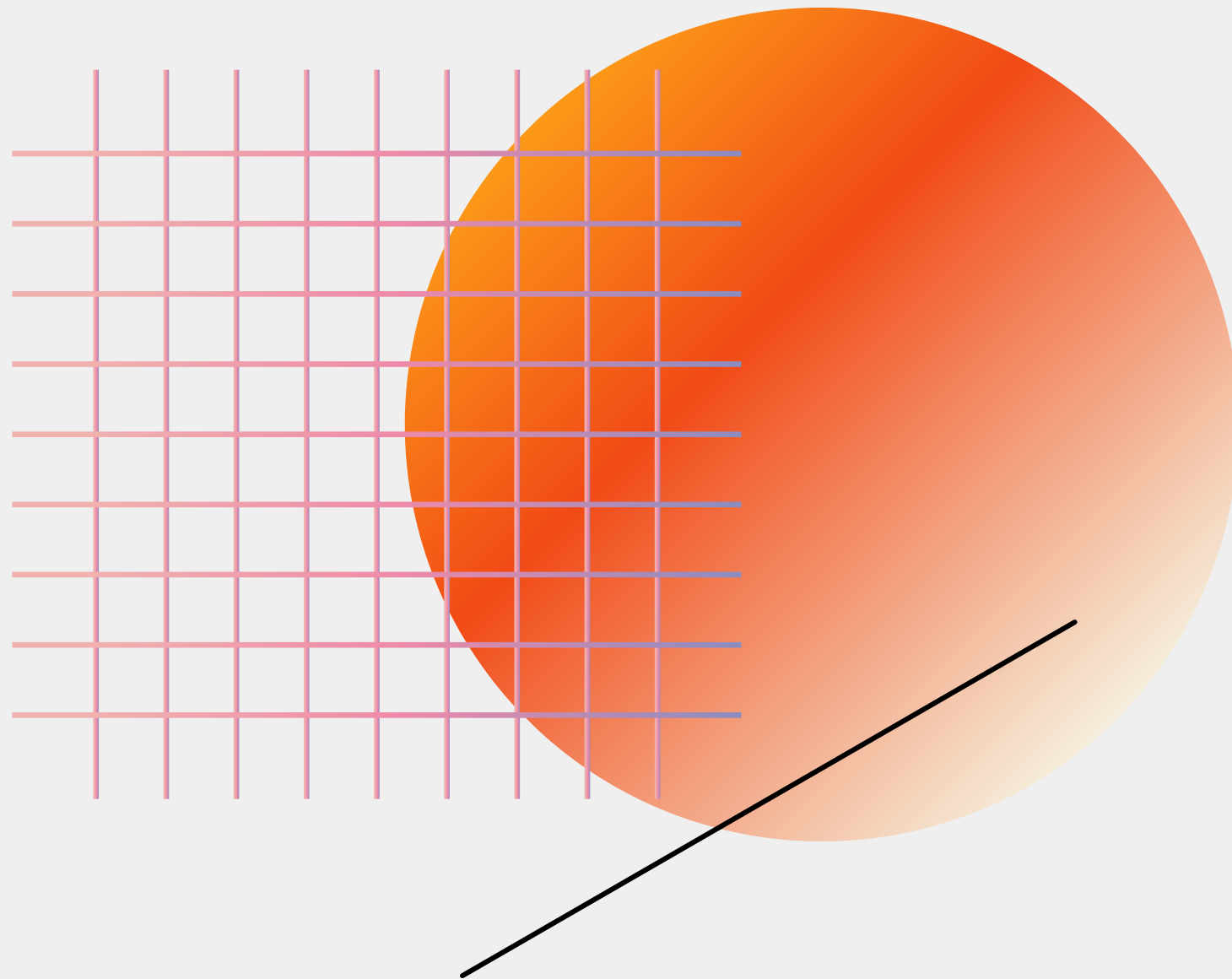
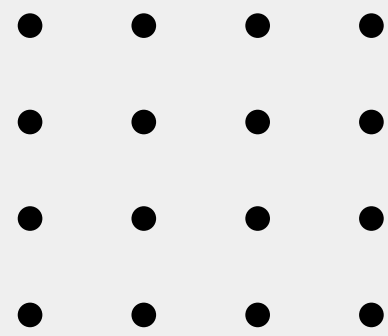
**Time-
consuming
process**

More complex

CONS

**Requires more
resources**





Feature Selection:

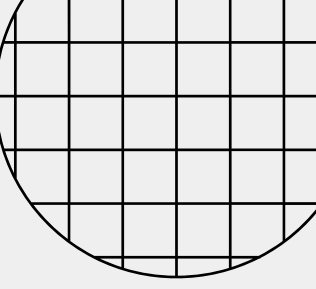
>Feature refers to column(s)!

The sole purpose of 'Feature Selection' is to reduce the number of columns so that the classification can get easier.

>Also feature selection is done to reduce the complexity of the dataset.

>But during feature selection we have to keep in mind, not to drop the important columns otherwise it may lead to information loss and the model would be considered as an Overfitting model.(TE = 0)

Feature Selection:

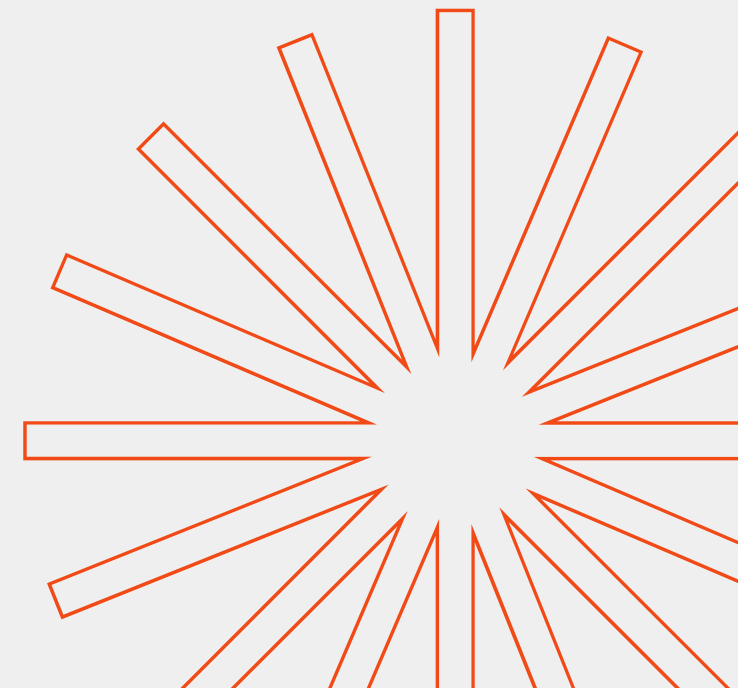


Definition:

- > A process that chooses an optimal subset of features according to an objective function:

Objective:

- > To reduce dimensionality(columns) and remove the noise(values that are not required)
- > To improve the mining performance
 - > **Speed of learning:** If we reduce the number of features then the speed of learning (by the machine learning model) will increase.
 - > **Predictive accuracy:** The ratio of predictive value matching the actual value increases.
 - > **Simplicity and Comprehensibility of mined results**





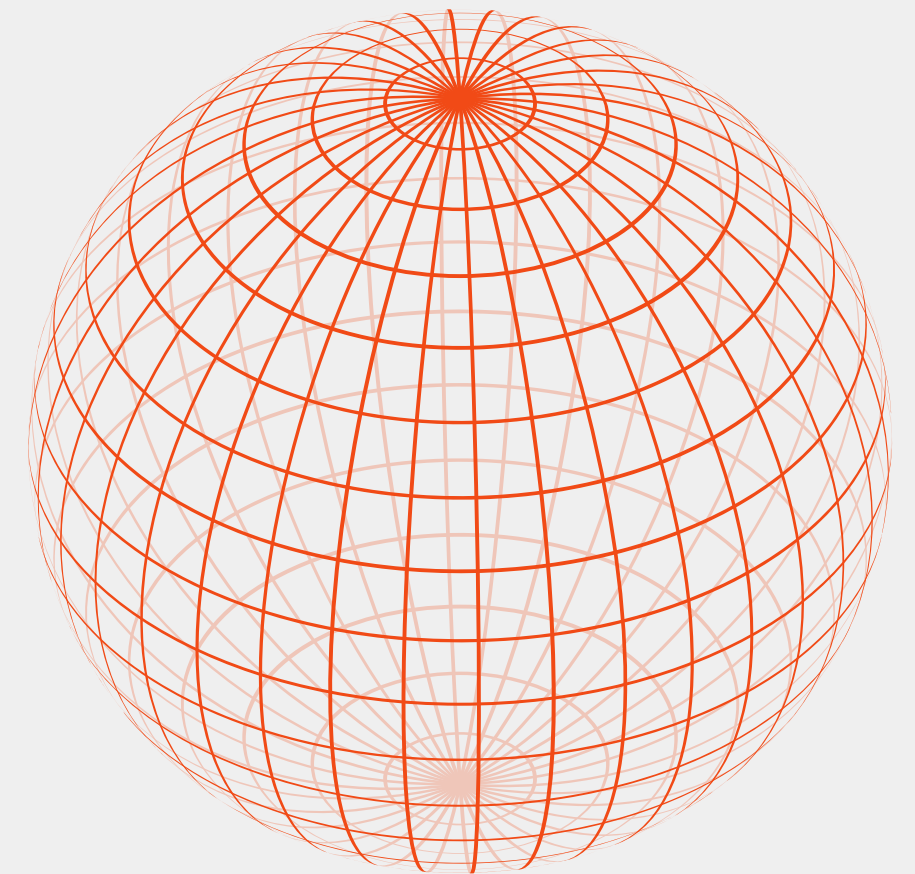
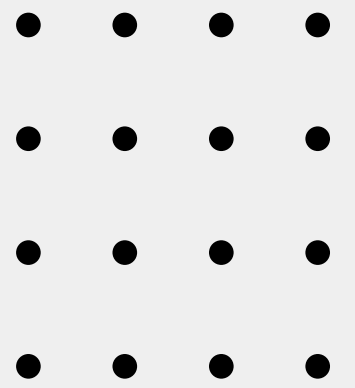
Feature Selection Models:

Filter Model:

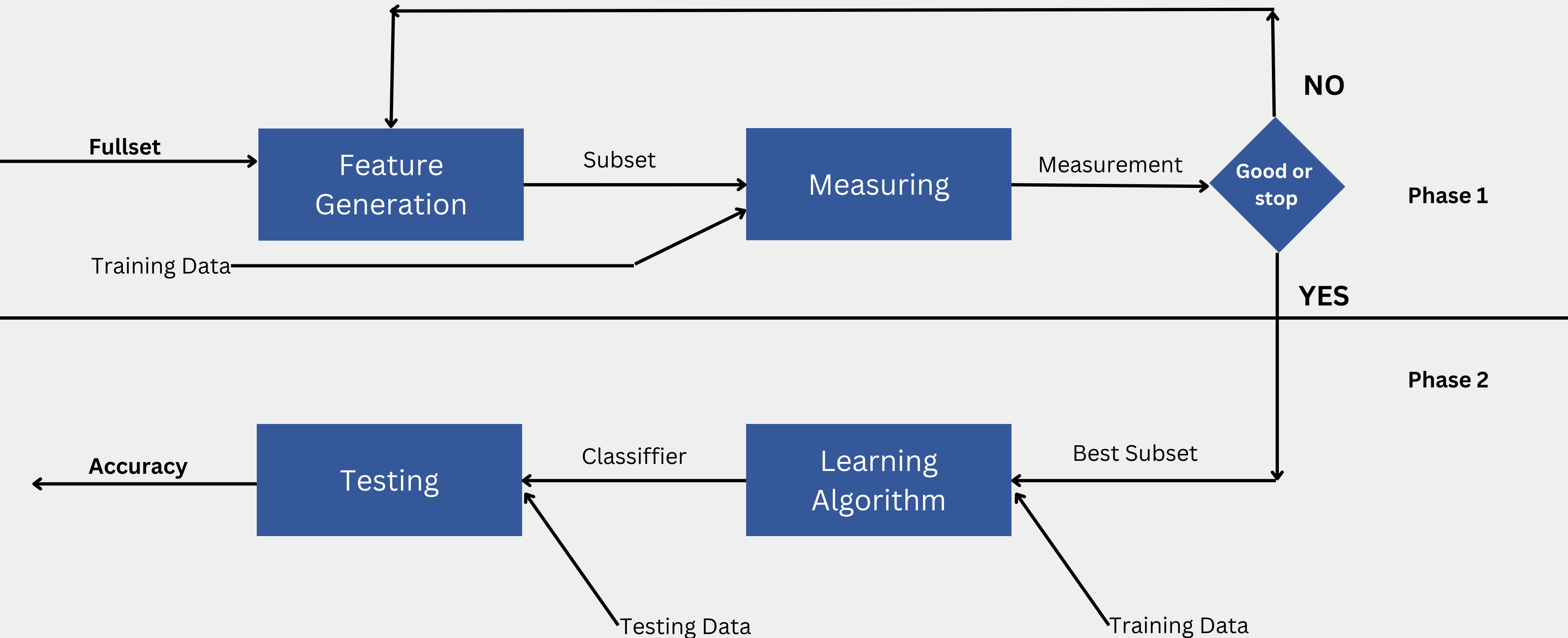
- > Separating feature selection from classifier learning.
- > Relying on general characteristics of data (information, distance, dependence, etc.)
- > No bias towards any learning algorithm.
- > Since this model does not depend on any classification algorithm.

Wrapper Model:

- > Relying on a predetermined Classification algorithm.
- > Using Predictive Accuracy as goodness measure.
- > High Accuracy but computationally very expensive.



Filter Model



Wrapper Model

