Data Science

Specialization

A Mathematic Driven Approach

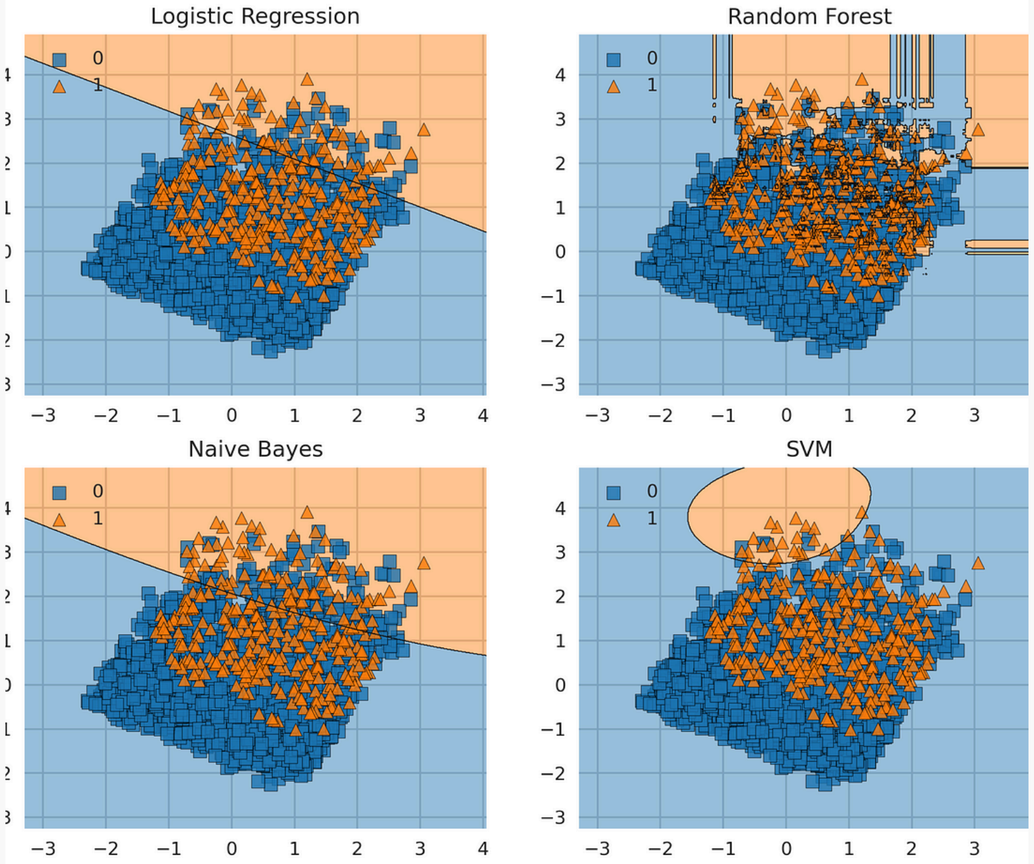
Machine Learning

What is machine learning?

Machine learning is a field of artificial intelligence that enables systems to learn patterns from data and improve performance on tasks without explicit programming.

How does it work?

Machine Learning



Create a Dataset

- Create or utilize an existing set of data (labeled or unlabeled) and separate it on training and test data

Evaluate the model using their formula

- Depending on the selected algorithm, evaluate the data with it's formula

Evaluate the outcome with the cost function

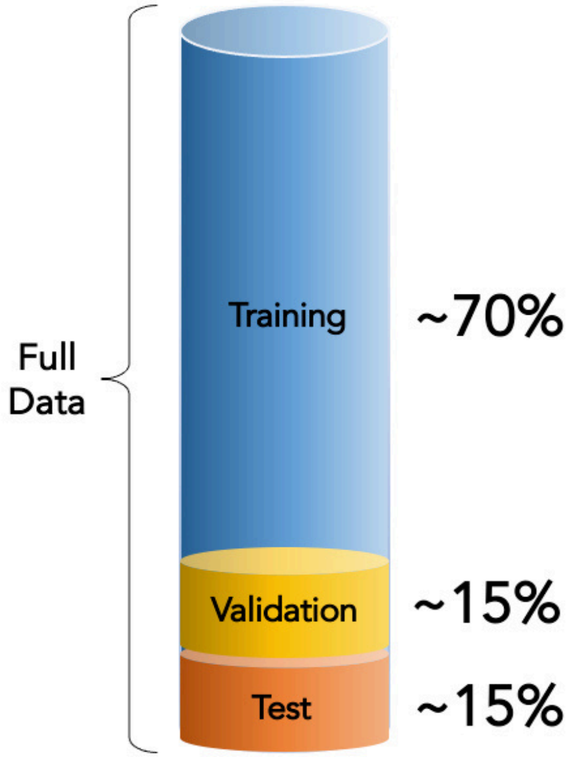
- Analize the efficiency of the model utilizing the cost function for each formula

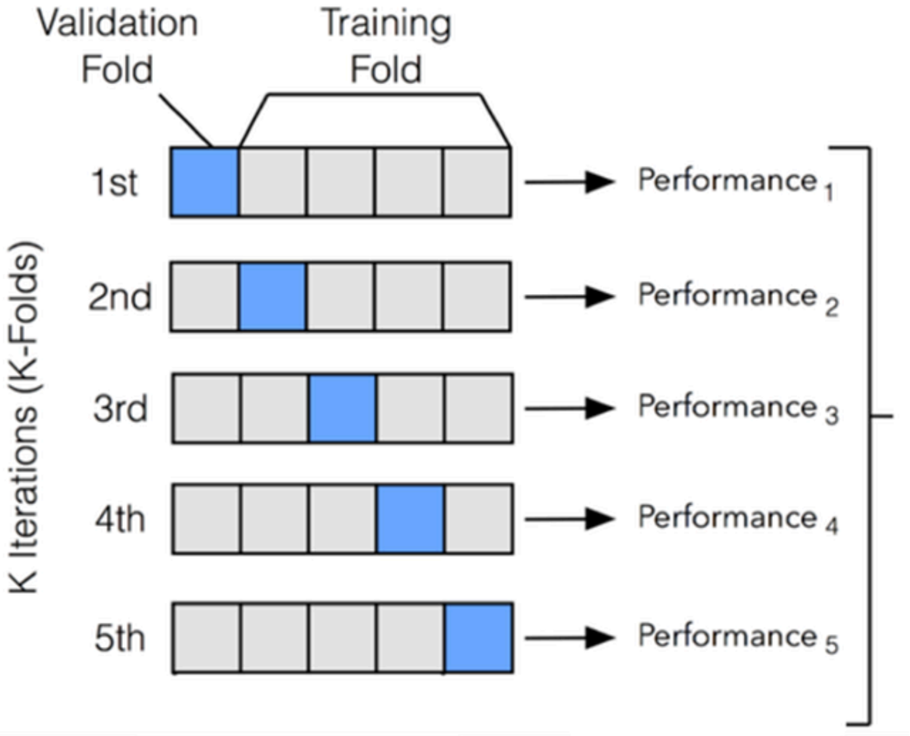
Data

Training and test splits

How does it work?

The separation of training and test sets involves splitting the dataset into two subsets: the training set, used to train the model, and the test set, used to evaluate its performance. This helps assess generalization and avoid overfitting.

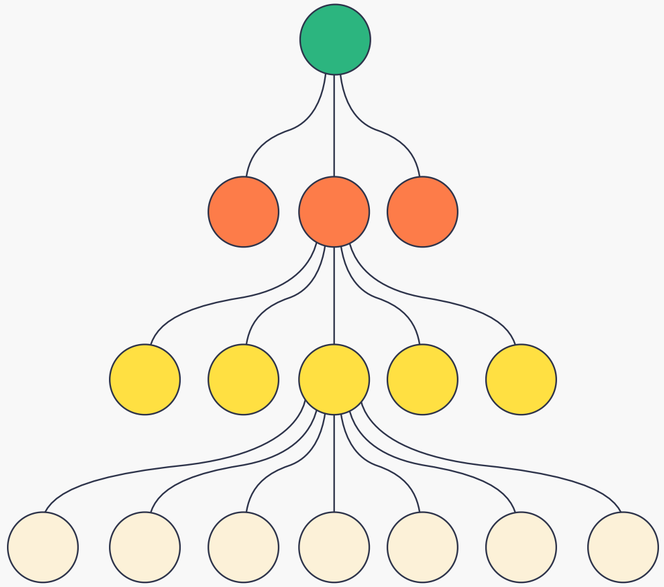




The 2 types of learning

Supervised Learning

Supervised learning is a machine learning approach where models are trained on labeled data, meaning input-output pairs. The algorithm learns to map inputs to correct outputs, enabling it to make predictions on new, unseen data. Examples include classification (e.g., spam detection) and regression (e.g., predicting house prices).



Unsupervised Learning

Unsupervised learning is a machine learning approach where models analyze and find patterns in unlabeled data without predefined outputs. It’s used for clustering (e.g., grouping customers) and dimensionality reduction (e.g., simplifying datasets). The algorithm identifies hidden structures, relationships, or distributions within the data.

Types of algorithms

Decision & Classification

Classification (Clustering)

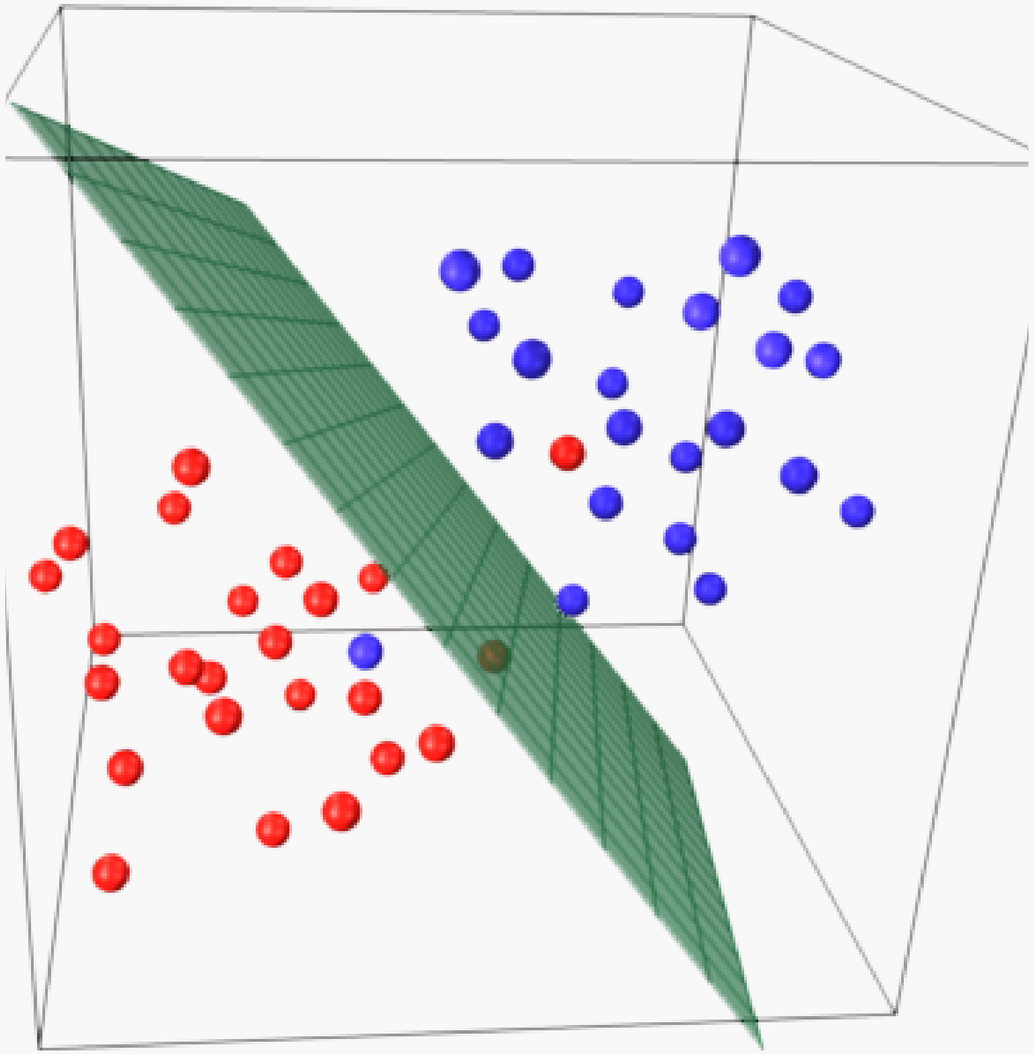
For supervised learning

What are clustering algorithms?

Clustering algorithms group similar data points into clusters based on their features, identifying patterns or structures in unlabeled datasets for analysis or decision-making.

Let’s check some

Support Vector Machine



Effective in High Dimensions

- Handles high-dimensional data well and works efficiently with a large number of features.

Works with Small Datasets

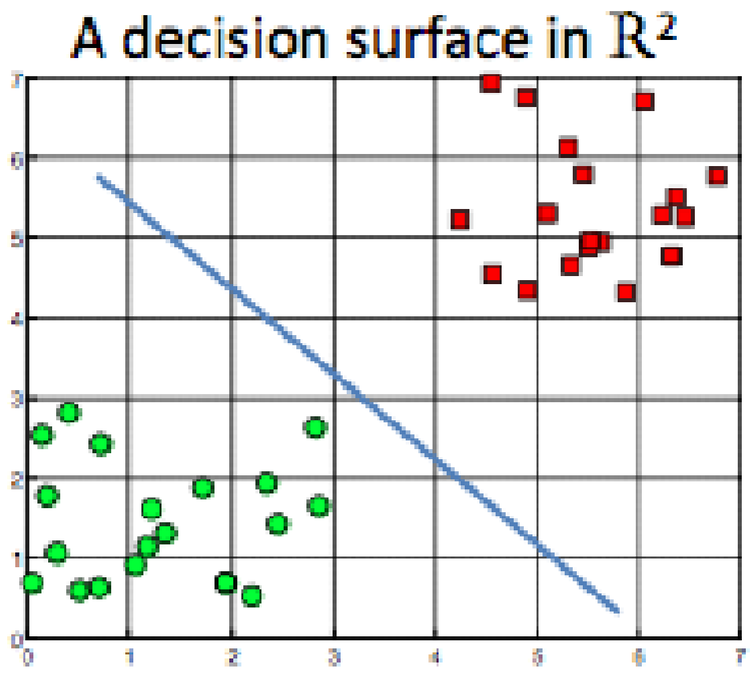
- SVMs are effective even with limited data, provided it's well-structured and properly labeled.

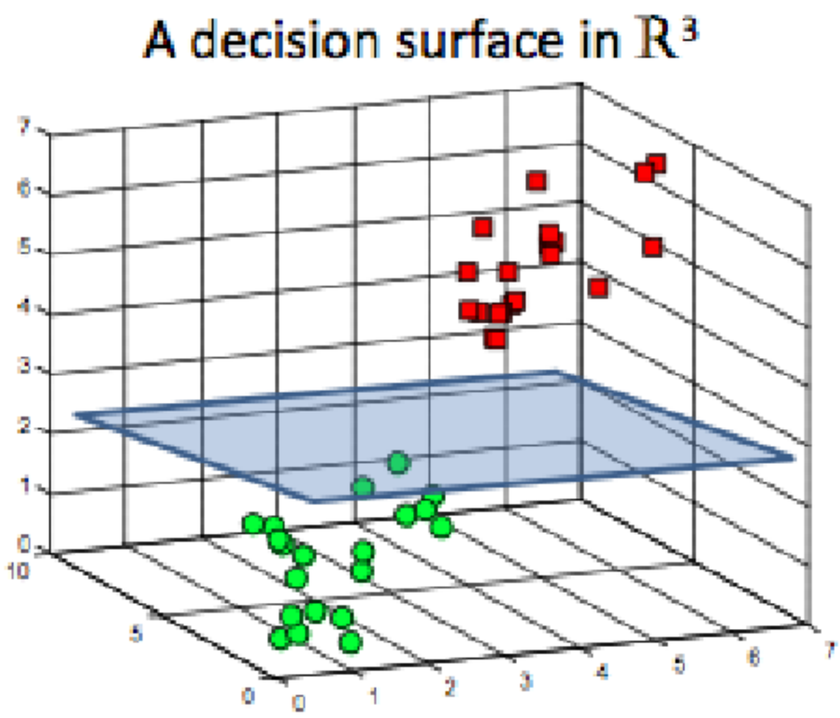
Versatile

- Supports linear and nonlinear classification through kernel functions.

How does it work?

Support Vector Machines classify data by finding the optimal hyperplane that maximizes the margin between classes. Using support vectors and kernel functions, they handle both linear and nonlinear data, ensuring effective separation for accurate predictions.





How can we do it?

Using This:

$$J(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i d_i (w^T x_i + b) + \sum_{i=1}^{N} \alpha_i$$

$$\textit{At the optimum } \frac{\partial J}{\partial w} = 0 \textit{ and } \frac{\partial J}{\partial b} = 0$$

$$\Rightarrow w_o = \sum_{i=1}^{N} \alpha_i d_i x_i \textit{ and } \sum_{i=1}^{N} \alpha_i d_i = 0$$

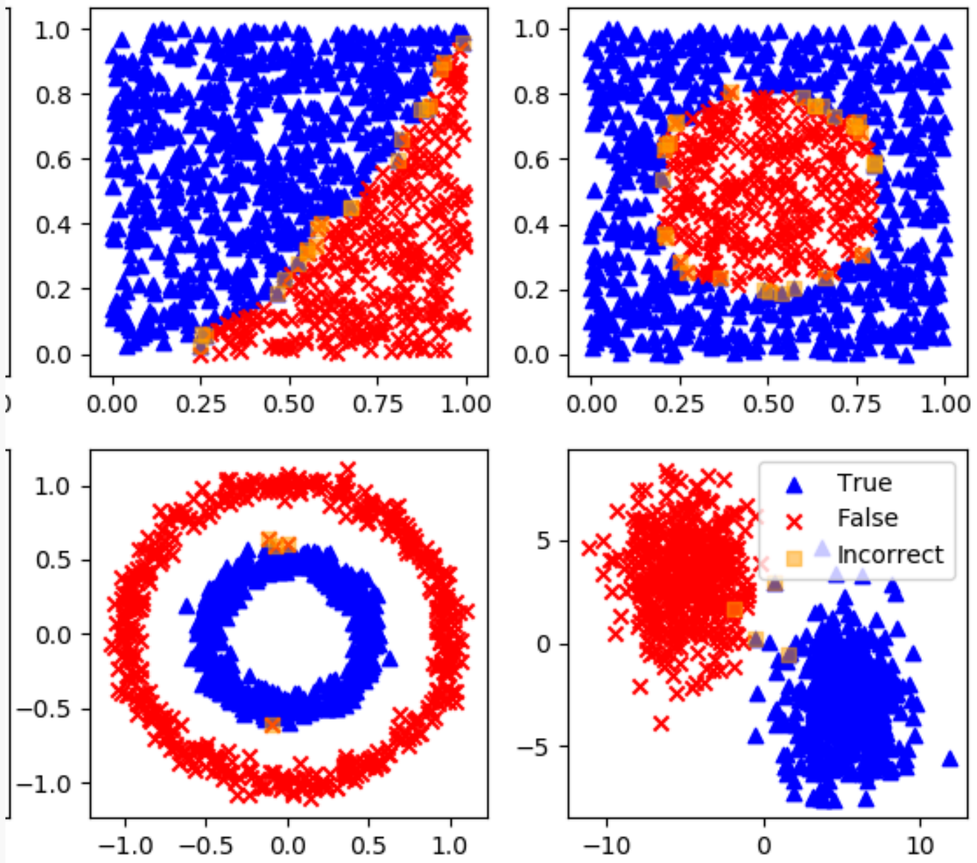
We don't need to suffer

Let's Try It

(With Scikit-Learn)

Let's check another one

Random Forest



Handles Missing Data

- Can handle missing values by averaging or selecting the most common value.

Robust to Noise

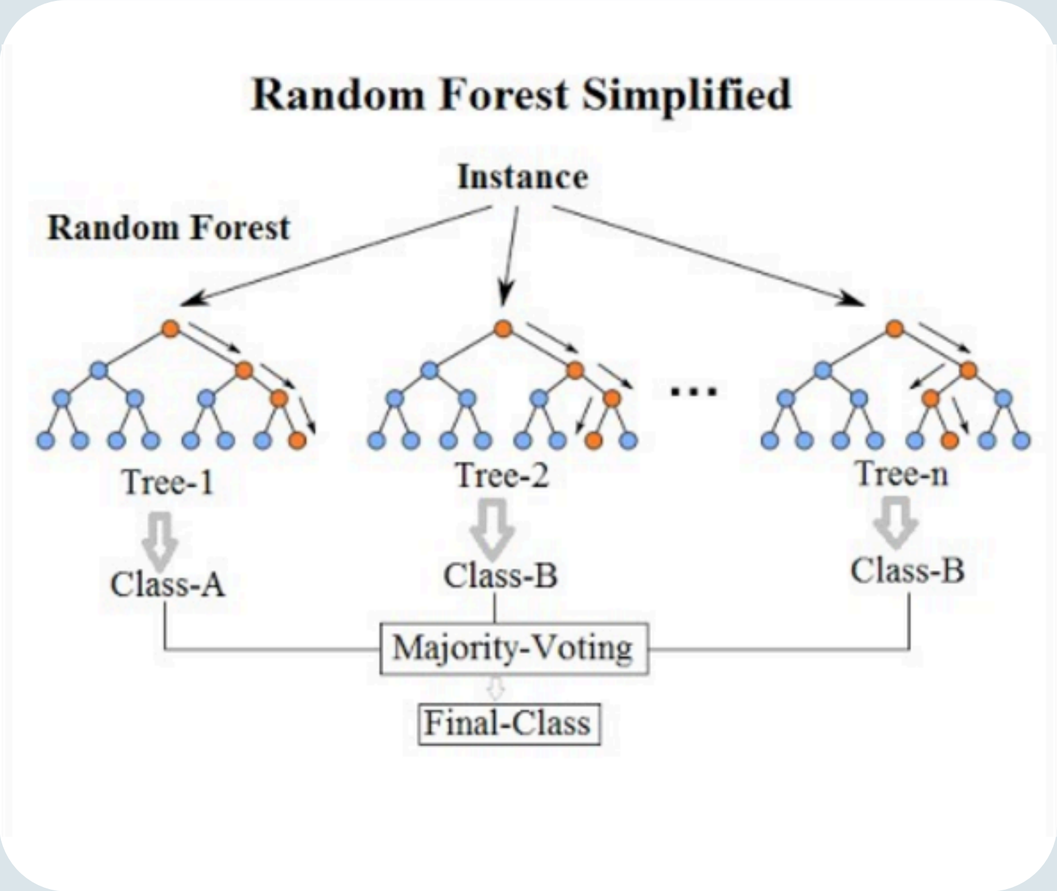
- Performs well even with noisy data by averaging over multiple trees, reducing the impact of outliers.

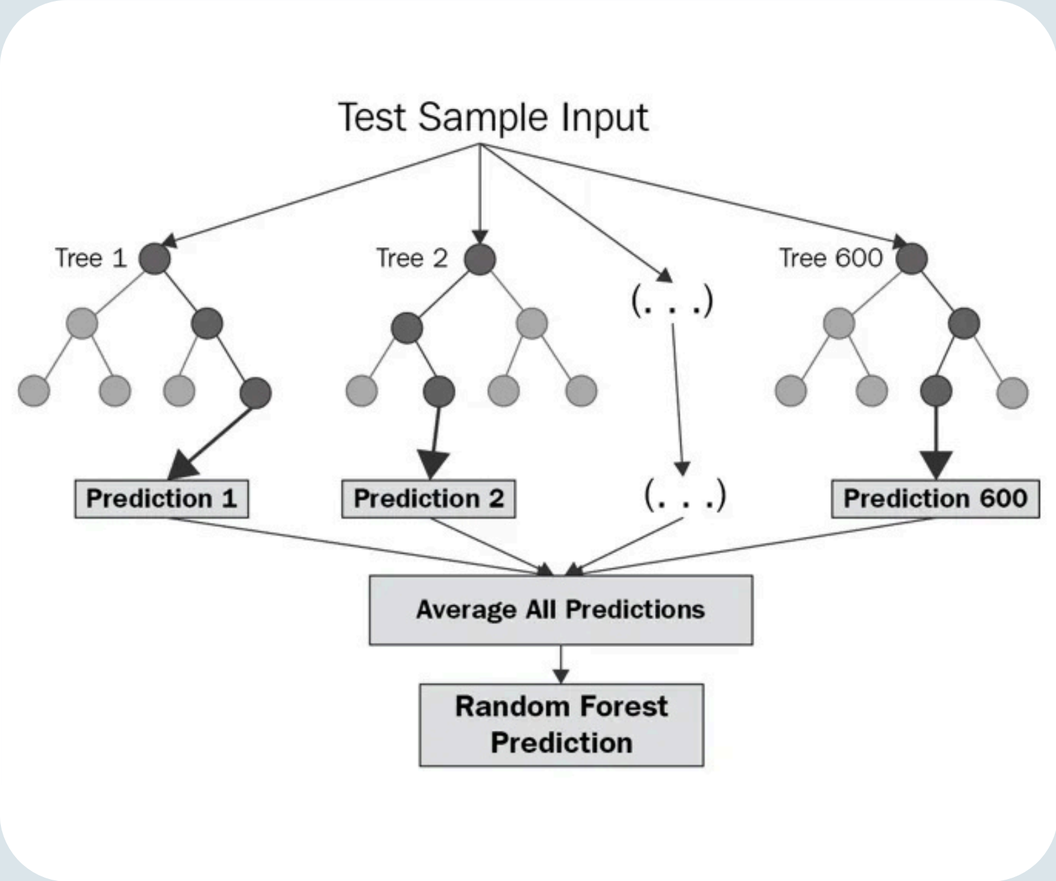
Flexible

- Works for both classification and regression tasks with diverse data types.

How does it work?

Random Forest works by constructing multiple decision trees during training. Each tree is trained on a random subset of the data, and their predictions are averaged (regression) or voted on (classification) to improve accuracy and reduce overfitting.





How can we do it?

Using This:

$$S = \begin{bmatrix} f_{A1} & f_{B1} & f_{C1} & C_1 \\ \vdots & & \vdots & \\ f_{AN} & f_{BN} & f_{CN} & C_N \end{bmatrix}$$

Create random subsets

$$S_1 = \begin{bmatrix} f_{A12} & f_{B12} & f_{C12} & C_{12} \\ f_{A15} & f_{B15} & f_{C15} & C_{15} \\ \vdots & & \vdots & \\ f_{A35} & f_{B35} & f_{C35} & C_{35} \end{bmatrix} \quad S_2 = \begin{bmatrix} f_{A2} & f_{B2} & f_{C2} & C_2 \\ f_{A6} & f_{B6} & f_{C6} & C_6 \\ \vdots & & \vdots & \\ f_{A20} & f_{B20} & f_{C20} & C_{20} \end{bmatrix}$$

$$S_M = \begin{bmatrix} f_{A4} & f_{B4} & f_{C4} & C_4 \\ f_{A9} & f_{B9} & f_{C9} & C_9 \\ \vdots & & \vdots & \\ f_{A12} & f_{B12} & f_{C12} & C_{12} \end{bmatrix}$$

We don't need to suffer

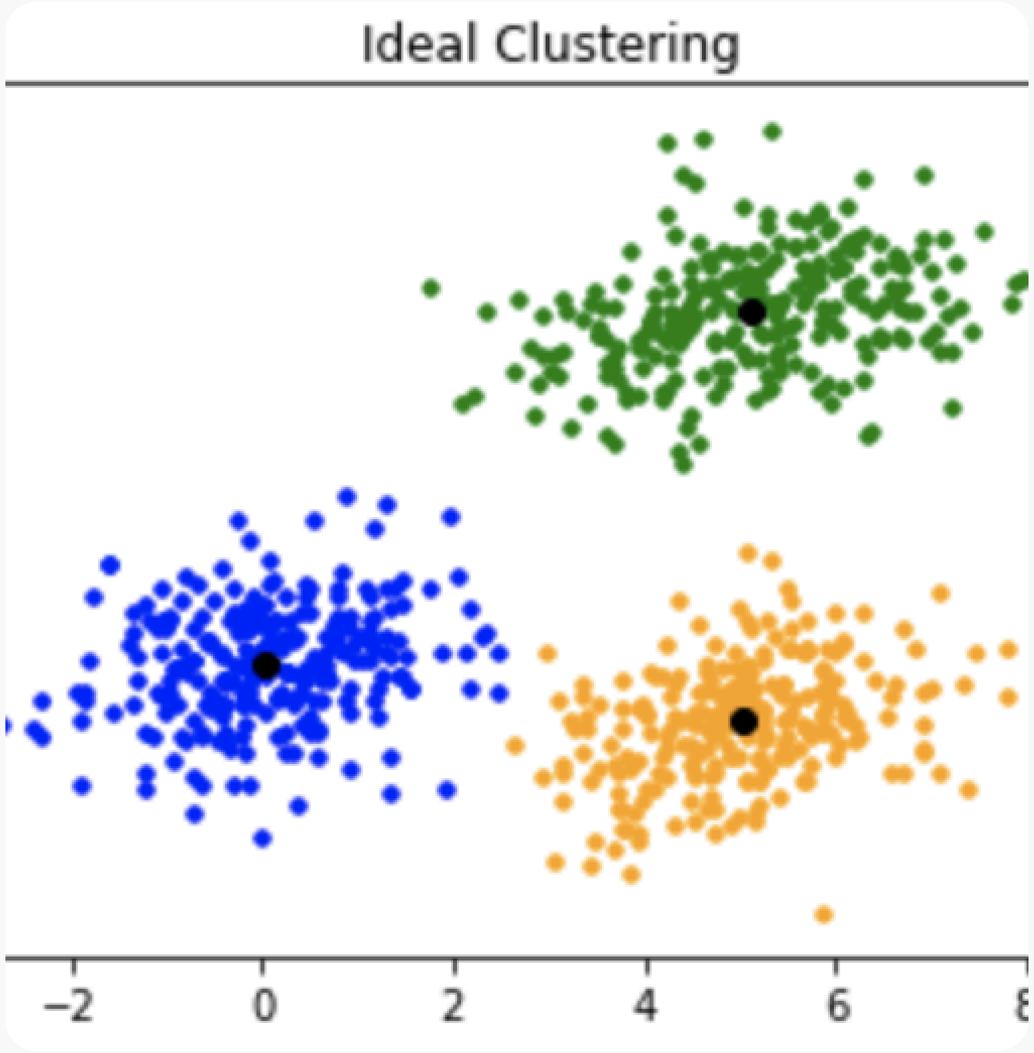
Let's Try It

(With Scikit-Learn)

Classification (Clustering)

For unsupervised learning

K-Means



Unsupervised Learning

- Groups data into clusters without labeled outcomes.

Efficient

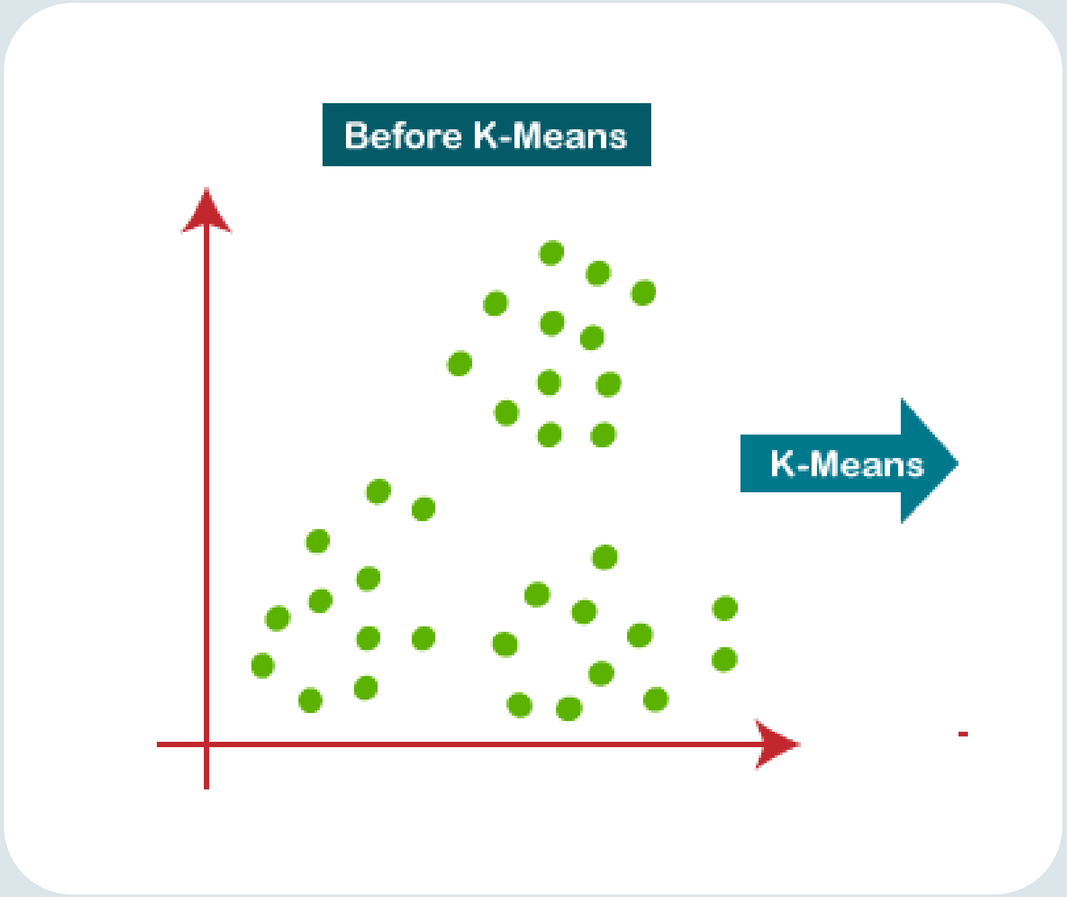
- Works well with large datasets and scales efficiently.

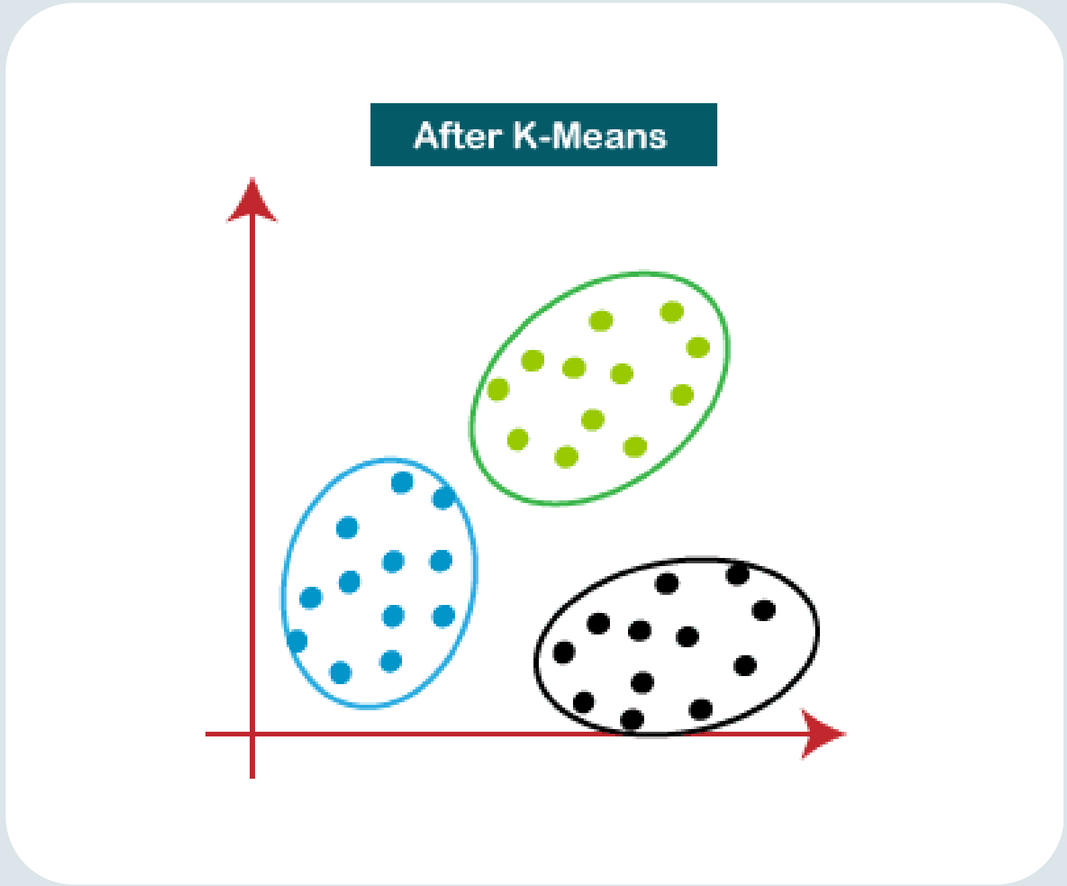
Simple and Fast

- Uses iterative refinement to minimize intra-cluster variance, making it computationally fast for many applications.

How does it work?

K-means clustering works by randomly initializing cluster centroids, then iteratively assigning each data point to the nearest centroid and recalculating the centroids based on the mean of assigned points, repeating until convergence for optimal cluster grouping.





How can we do it?

Using This:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{m} w_{ik} \, \|x_i - c_k\|^2$$

$w_{ik} = 0$ *if the data point does not belong to the cluster*

$w_{ik} = 1$ *if the data point belongs to the cluster*

We don't need to suffer

Let's Try It

(With Scikit-Learn)

Decision

What are decision algorithms?

Decision algorithms are processes or rules used to make choices or classify data, often in machine learning, by evaluating input conditions to produce specific outcomes or predictions.

We will see this soon

Our project will be a neural network for
human-like decision making

Thank you