

Random Forest & Ensemble Learning

OBJECTIVE & OUTLINE

Proprietary document of Indonesia AI 2023



Random Forest & Ensemble Learning

Objektif: Memahami algoritma supervised Learning yaitu algoritma Random Forest dan konsep Ensemble Learning.

Outline:

1. Mengenal Decision Tree
2. Konsep Ensemble Learning
3. Algoritma Random Forest

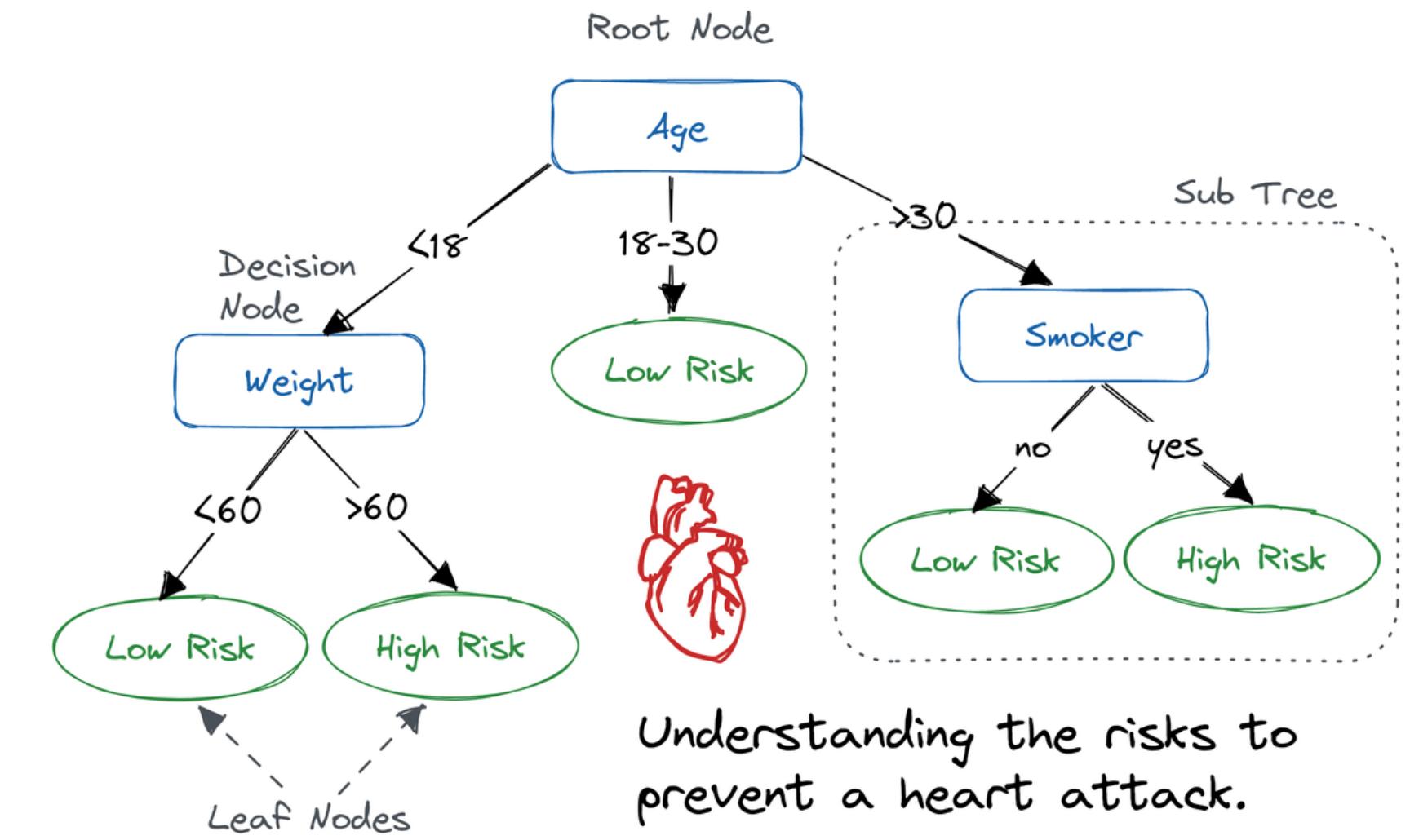
Mengenal Decision Tree

DEFINISI

Proprietary document of Indonesia AI 2023

Decision Tree atau pohon keputusan adalah salah satu algoritma pembelajaran mesin yang digunakan dalam pemrosesan data.

Algoritma ini mewakili serangkaian aturan yang digunakan untuk mengklasifikasikan atau memprediksi output berdasarkan fitur-fiturnya.



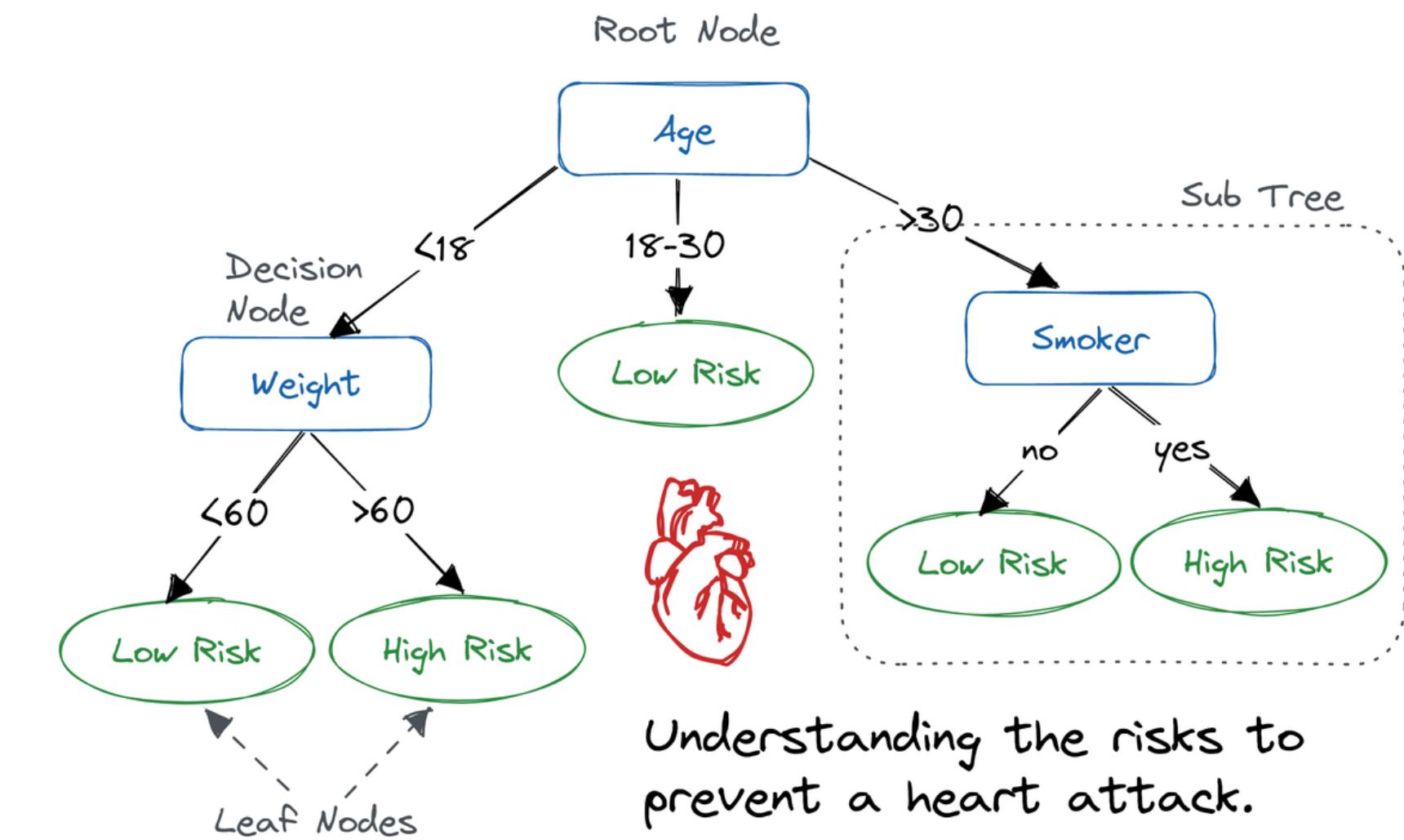
Indonesia AI

Image source: www.datacamp.com

KEUNTUNGAN

Proprietary document of Indonesia AI 2023

1. Mudah dimengerti
2. Mudah diinterpretasikan
3. Multi fungsi (klasifikasi, regresi, clustering)



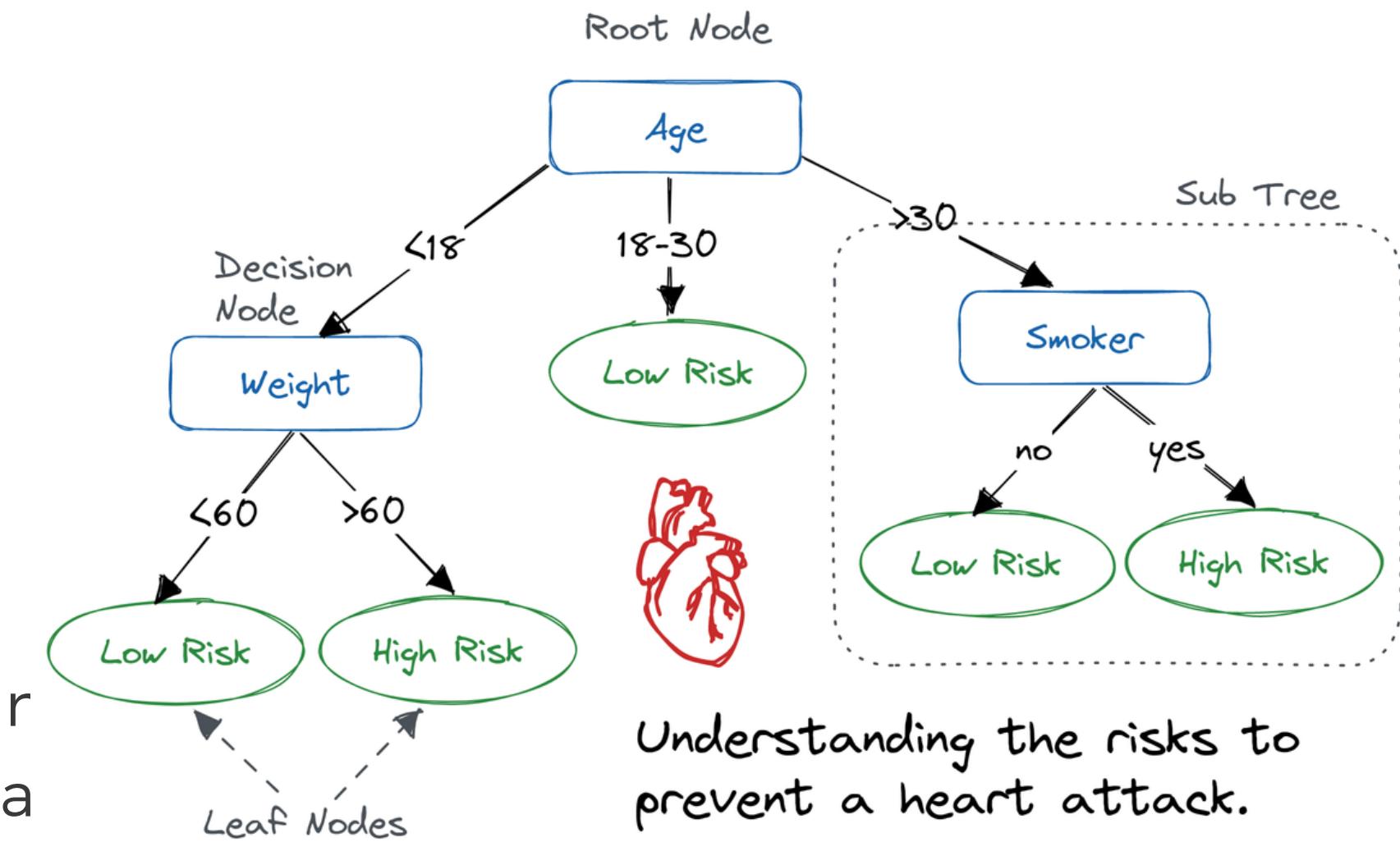
KERUGIAN

Proprietary document of Indonesia AI 2023

ALERT!

Pohon keputusan cenderung overfitting terhadap data training yang dapat menghasilkan performa model yang buruk pada data testing!

Untuk mengatasi hal tersebut, perlu dilakukan penggunaan teknik khusus seperti pruning dan ensemble methods.



Indonesia AI

Image source: www.datacamp.com

INFORMATION THEORY

Proprietary document of Indonesia AI 2023

Sebuah cabang dalam ilmu matematika yang mempelajari kuantifikasi, penyimpanan, dan komunikasi informasi.

Dalam konteks pohon keputusan atau decision tree, **information theory** dapat digunakan untuk memilih fitur atau variabel mana yang paling relevan atau informatif dalam memprediksi target atau keluaran.

Play Golf	
Yes	No
9	5


$$\begin{aligned} \text{Entropy}(\text{PlayGolf}) &= \text{Entropy}(5, 9) \\ &= \text{Entropy}(0.36, 0.64) \\ &= -(0.36 \log_2 0.36) - (0.64 \log_2 0.64) \\ &= 0.94 \end{aligned}$$

INFORMATION GAIN

Proprietary document of Indonesia AI 2023

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

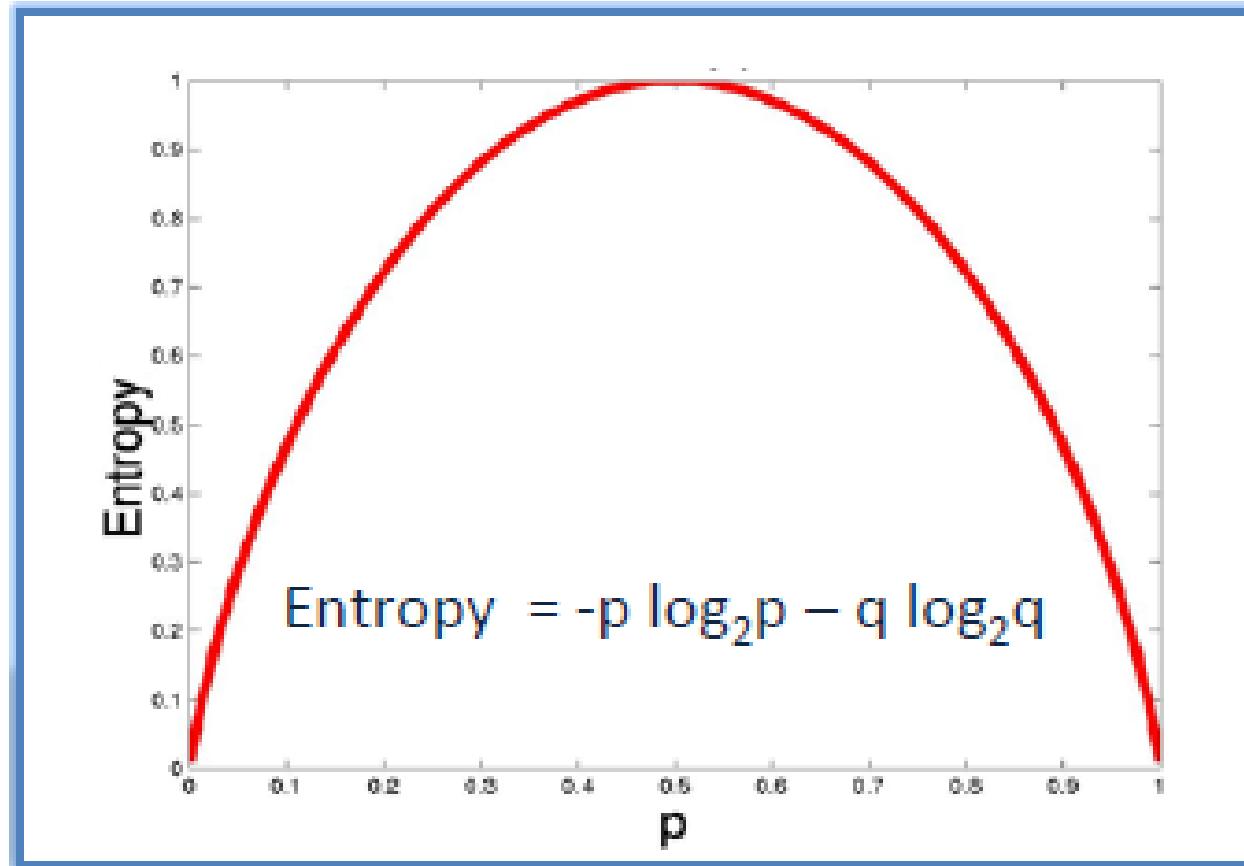
$$\begin{aligned}G(\text{PlayGolf}, \text{Outlook}) &= E(\text{PlayGolf}) - E(\text{PlayGolf}, \text{Outlook}) \\&= 0.940 - 0.693 = 0.247\end{aligned}$$

Salah satu metode yang umum digunakan dalam information theory untuk pohon keputusan adalah gain information atau gain informasi. **Gain informasi** adalah pengukuran jumlah informasi yang diberikan oleh suatu fitur atau simpul dalam pohon keputusan.

Semakin tinggi gain informasi, semakin baik fitur tersebut dalam membedakan target atau keluaran!

ENTROPY

Proprietary document of Indonesia AI 2023



Entropy adalah ukuran dari ketidakpastian atau keacakan dalam sebuah set data. Entropi digunakan dalam perhitungan Information Gain pada pohon keputusan sebagai ukuran kuantitatif dari informasi yang diberikan oleh simpul atau fitur dalam memprediksi target atau keluaran.

$$\text{Entropy} = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$$

ENTROPY

Proprietary document of Indonesia AI 2023

Entropy dengan frekuensi tabel 1 atribut

$$E(S) = \sum_{i=1}^c -p_i \log_2 p_i$$

Play Golf	
Yes	No
9	5

Entropy(PlayGolf) = Entropy (5,9)
= Entropy (0.36, 0.64)
= $-(0.36 \log_2 0.36) - (0.64 \log_2 0.64)$
= 0.94



Entropy dengan frekuensi tabel 2 atribut

$$E(T, X) = \sum_{c \in X} P(c)E(c)$$

		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
				14

$E(\text{PlayGolf}, \text{Outlook}) = P(\text{Sunny})*E(3,2) + P(\text{Overcast})*E(4,0) + P(\text{Rainy})*E(2,3)$
= $(5/14)*0.971 + (4/14)*0.0 + (5/14)*0.971$
= 0.693

Any question guys ~

Konsep Ensemble Learning

DEFINISI

Proprietary document of Indonesia AI 2023



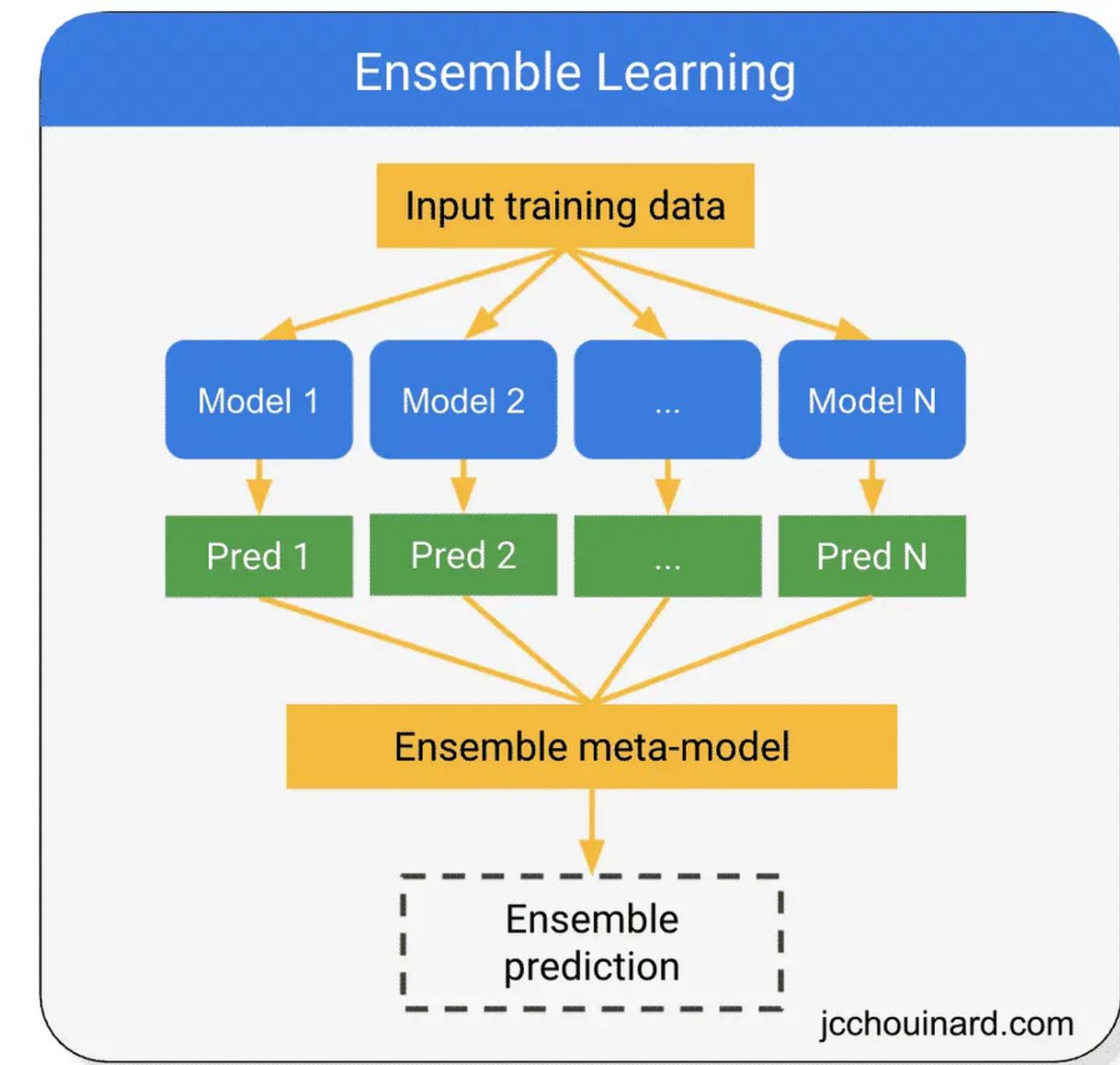
Ensemble Learning adalah sebuah teknik dalam machine learning yang menggabungkan beberapa model atau algoritma untuk meningkatkan kinerja atau akurasi prediksi.

Ensemble learning bertujuan untuk mengurangi varians dan meningkatkan kinerja prediksi dengan menggabungkan hasil prediksi dari beberapa model atau algoritma yang berbeda.

BEBERAPA TEKNIK ENSEMBLE LEARNING

Proprietary document of Indonesia AI 2023

1. Bagging (Bootstrap Aggregating)
2. Boosting
3. Stacking



Indonesia AI

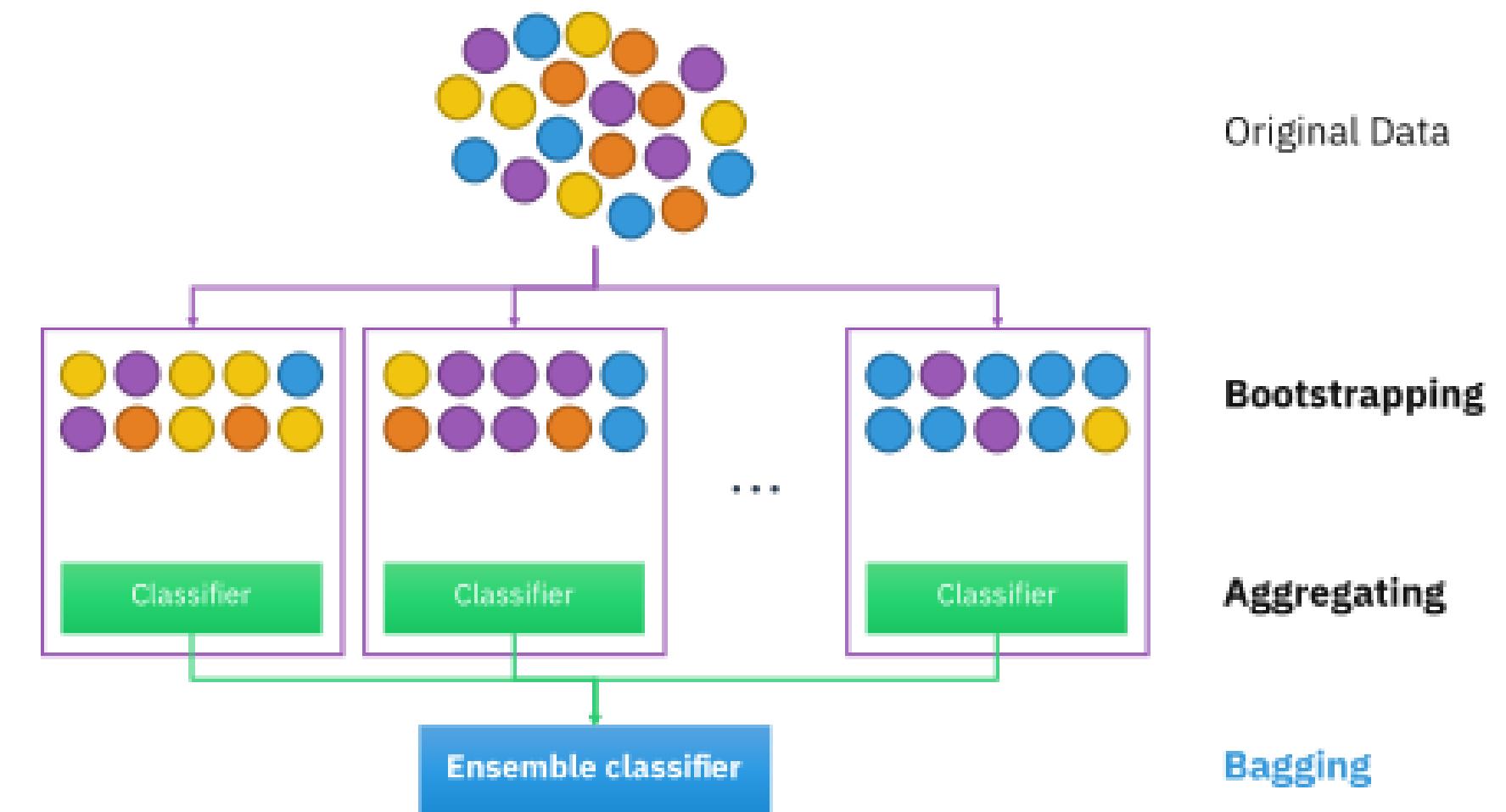
Image source: www.jcchouinard.com

jcchouinard.com

BAĞGINİNG (BOOTSTRAP AGGREGATING)

Proprietary document of Indonesia AI 2023

Teknik ini memanfaatkan bootstrapping (sampling dengan pengembalian) pada set data asli untuk membuat beberapa sampel set data yang berbeda, kemudian membuat model pada setiap sampel dan menggabungkan hasil prediksi dari semua model untuk mencapai hasil prediksi yang lebih akurat.



Indonesia AI

Image source: en.wikipedia.org

BAĞGINİNG (BOOTSTRAP AĞGREGATING)

Proprietary document of Indonesia AI 2023

Tiga contoh algoritma yang menggunakan konsep Bagging:

1. Random Forest
2. KNN
3. SVM

Indonesia AI

Image source: en.wikipedia.org

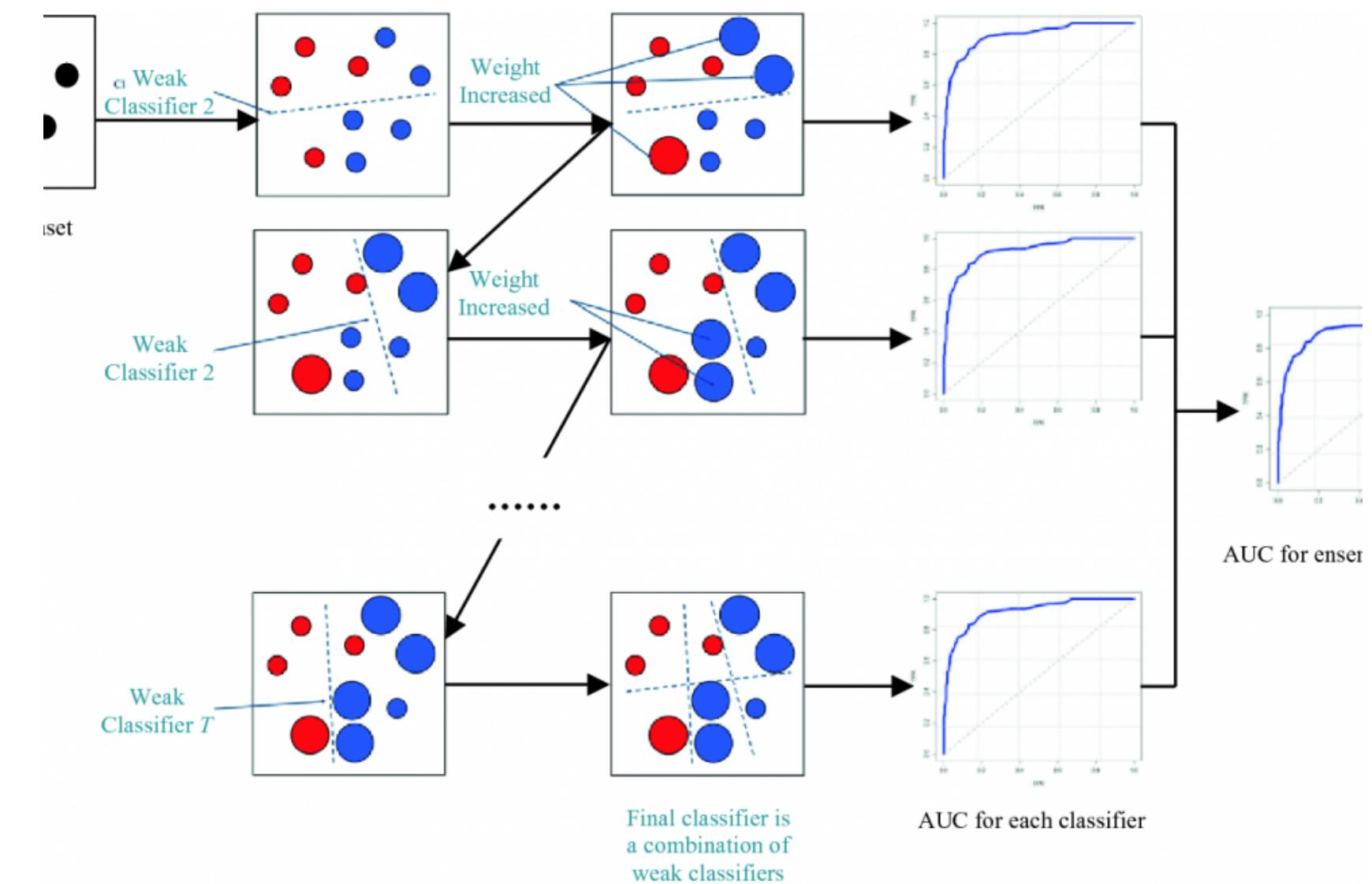
BOOSTING

Proprietary document of Indonesia AI 2023

Teknik ini memperkuat model yang lemah (weak learner) secara bertahap dan menggabungkan prediksi dari setiap model yang lebih kuat (stronger) untuk mencapai akurasi prediksi yang lebih baik.

Contoh algoritma:

1. AdaBoost
2. Gradient Boosting
3. XGBoost



Indonesia AI

Image source: <https://datascience.eu>

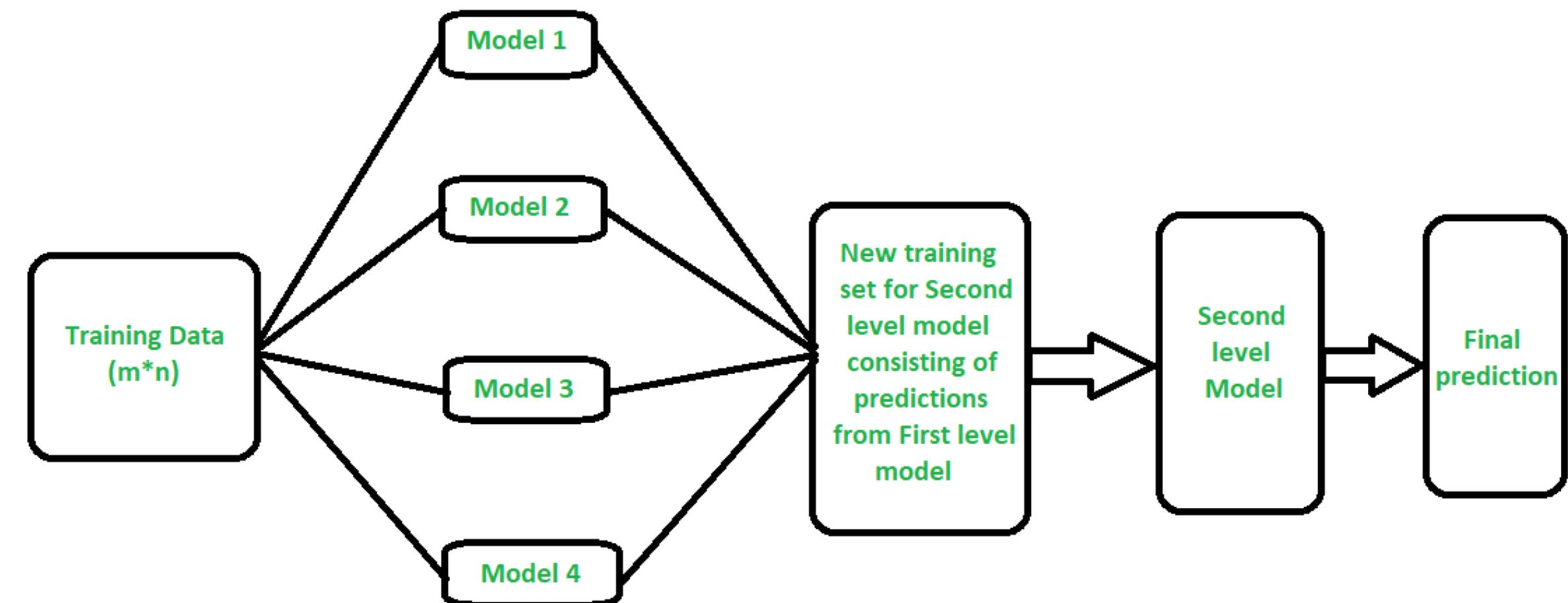
STACKING

Proprietary document of Indonesia AI 2023

Teknik ini menggabungkan hasil prediksi dari beberapa model atau algoritma yang berbeda, kemudian menggabungkannya menjadi input untuk model yang lebih tinggi atau yang disebut meta-model.

Contoh algoritma:

1. Super Learner
2. Stacked Generalization
3. Blending



Indonesia AI

Image source: www.geeksforgeeks.org

Any question guys ~

— Algoritma Random Forest

DEFINISI

Proprietary document of Indonesia AI 2023



Random Forest adalah sebuah algoritma machine learning yang digunakan untuk masalah klasifikasi, regresi, dan pengelompokan data (clustering).

Random Forest bekerja dengan membangun sejumlah besar pohon keputusan secara acak, di mana setiap pohon keputusan dibangun dengan sub-sampel data yang dipilih secara acak dan subset fitur yang dipilih secara acak.

DEFINISI

Proprietary document of Indonesia AI 2023

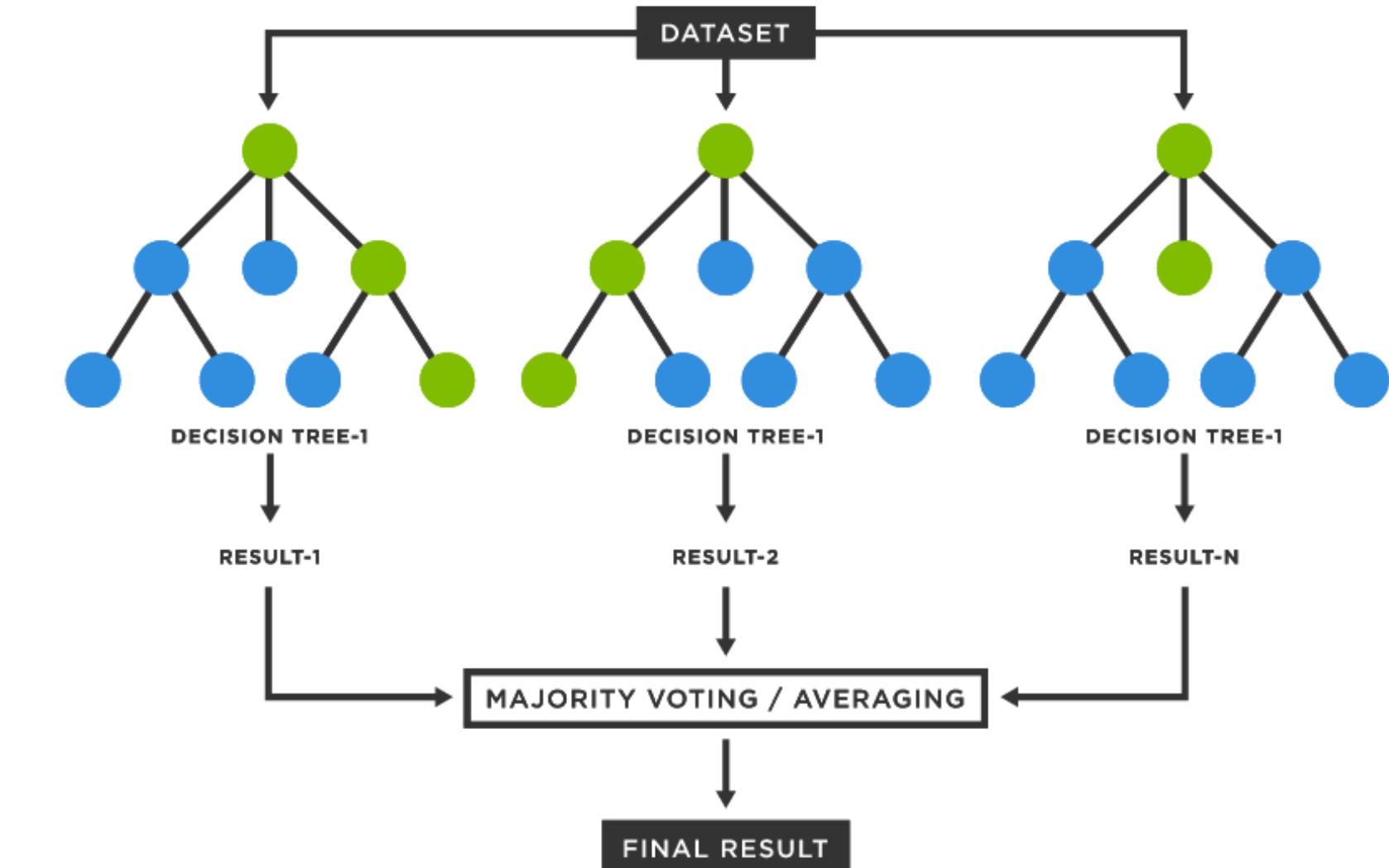
Hasil prediksi dari setiap pohon diambil dan diambil rata-ratanya (untuk regresi) atau melalui voting (untuk klasifikasi), untuk menghasilkan prediksi akhir.

ISU OVERFITTING

Proprietary document of Indonesia AI 2023

Setiap pohon keputusan dibangun secara independen, dan tidak saling terkait satu sama lain. Hal ini memungkinkan Random Forest untuk menghindari overfitting dan meningkatkan akurasi prediksi.

Random Forest juga memiliki kemampuan untuk mengekstrak pentingnya setiap fitur dalam data, sehingga dapat digunakan untuk pemilihan fitur.



Indonesia AI

Image source: www.tibco.com



KEUNGGULAN RANDOM FOREST

Proprietary document of Indonesia AI 2023

- **Multifungsi:** Cocok untuk klasifikasi maupun regresi
- **Robust:** Akurasi yang baik dan tidak mudah terpengaruh ketika diuji dengan data test
- **Reliable:** Klasifikasi Random Forest dapat menangani *missing value* dan mempertahankan akurasi dari kebanyakan data
- **Baseline:** Random Forest akan digunakan sebagai model dasar untuk jenis proyek apa pun di industri

Any question guys ~

Terimakasih!