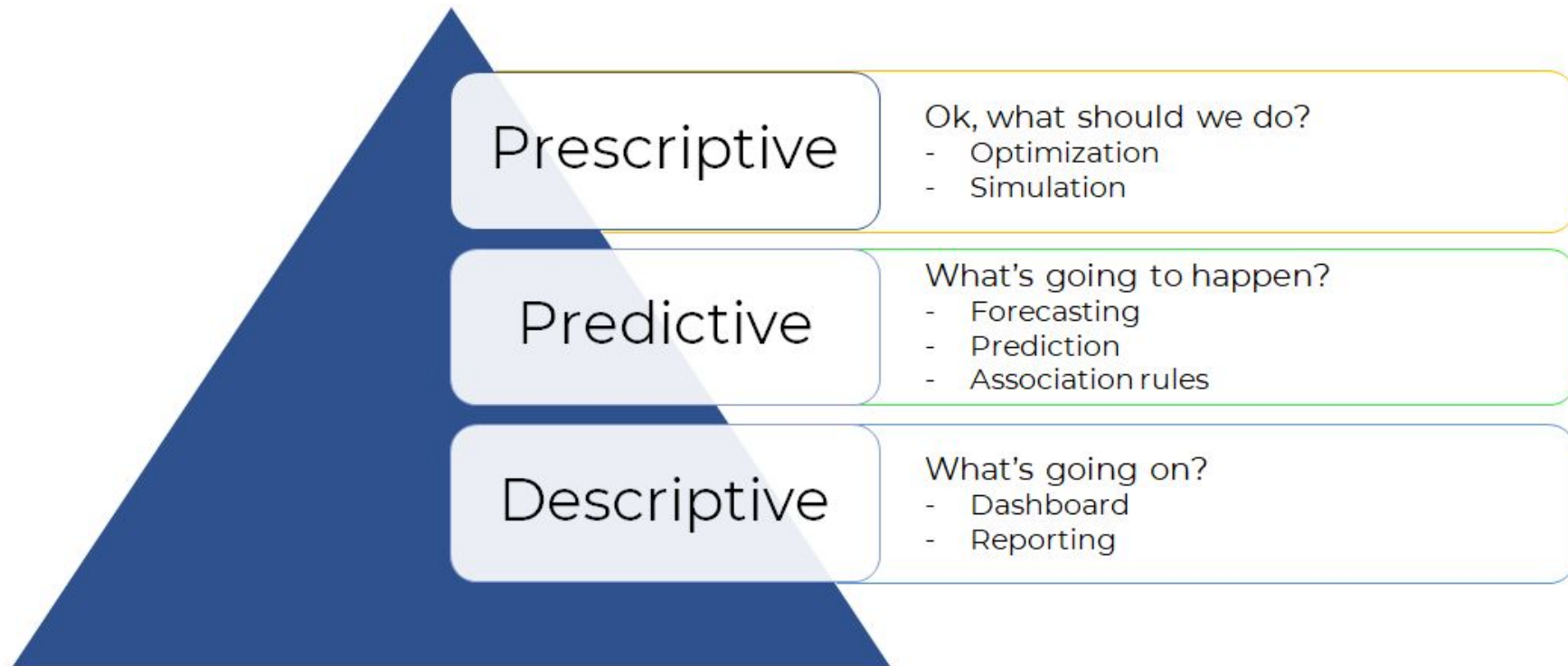




Analisis Prediktif

Level of Analytics



Analisis Prediktif

Analisis prediktif adalah merupakan suatu metode analisis yang digunakan untuk membuat prediksi mengenai kejadian di masa depan.

Analisis prediktif menggunakan hasil yang sudah diketahui sebelumnya untuk mengembangkan (atau melatih) model yang dapat digunakan untuk memprediksi nilai untuk data berbeda atau baru.

Penggunaan Analisis Prediktif

1. Mendeteksi kesalahan/kecurangan
2. Mengurangi Resiko
3. Manajemen Operasi

Prediksi menggunakan Machine Learning

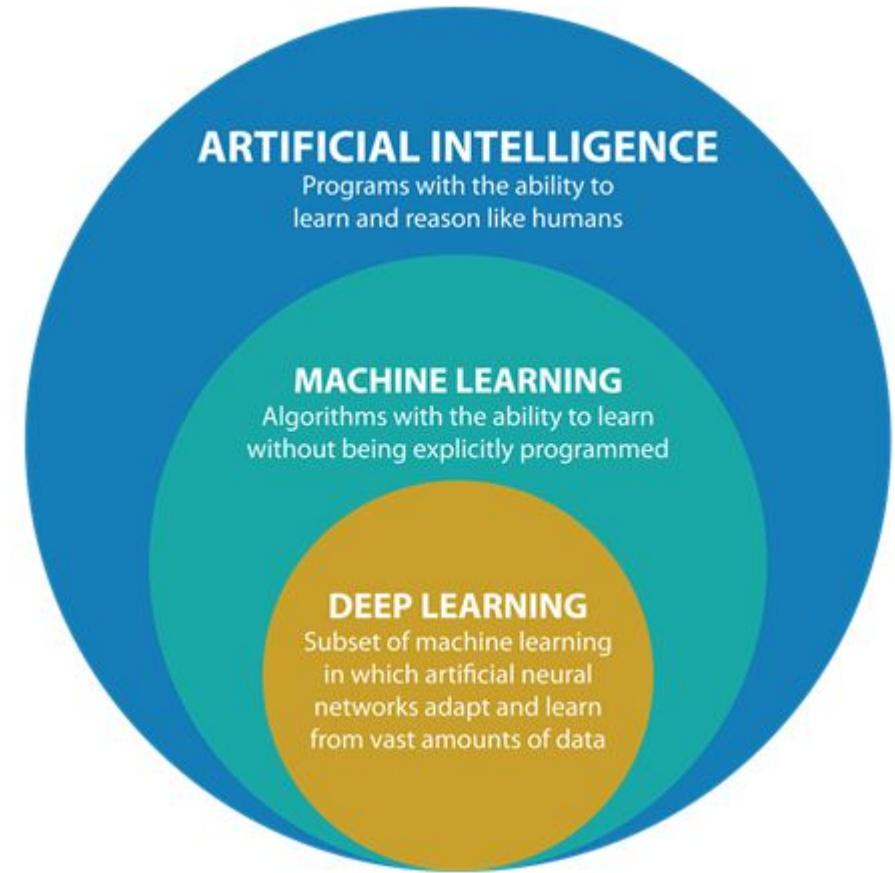
Machine Learning

Machine Learning merupakan cabang ilmu dari Artificial Intelligence (Kecerdasan Buatan) yang berfokus pada pengembangan sistem yang mampu belajar "sendiri" (Tidak harus di program secara manual oleh manusia)

- Traditional Programming



- Machine Learning



Melihat Pola

Lama bekerja (Tahun)	Gaji (Rupiah)
2	3.000.000
4	6.000.000
6	9.000.000
10	15.000.000
12	24.000.000
14	28.000.000

Melihat Pola

Lama bekerja (Tahun)	Gaji (Rupiah)
2	3.000.000
4	6.000.000
6	9.000.000
10	15.000.000
12	24.000.000
14	28.000.000

```
if (experience <= 10)
{ salary = experience * 1.5 * 1000000}
else if(experience >10)
{ salary = experience * 2 * 1000000}
```


Melihat Pola

Lama bekerja	Level Pekerjaan	Pendidikan Terakhir	Gaji (Rupiah)
2	3	Yes	4.500.000
4	3	No	6.000.000
6	4	No	7.500.000
10	5	Yes	18.000.000
12	5	No	15.000.000
14	6	No	18.000.000

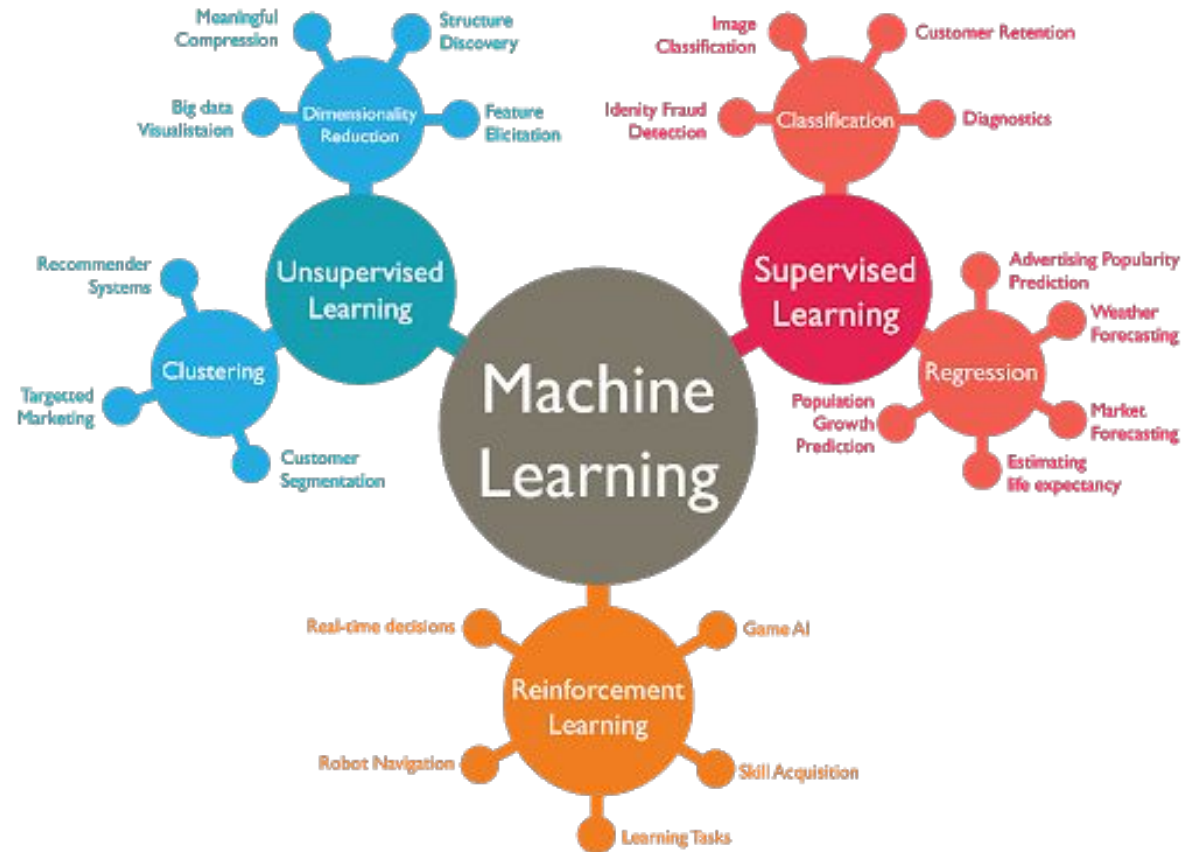
Melihat Pola

Lama bekerja	Level Pekerjaan	Pendidikan Terakhir	Gaji (Rupiah)
2	3	Yes	4.500.000
4	3	No	6.000.000
6	4	No	7.500.000
10	5	Yes	18.000.000
12	5	No	15.000.000
14	6	No	18.000.000

`Salary = Experience * Magic_Number_1 + JobLevel * Magic_Number_2 + Skill * Magic_Number_3 + Magic_Number_4`

Tipe Machine Learning

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning



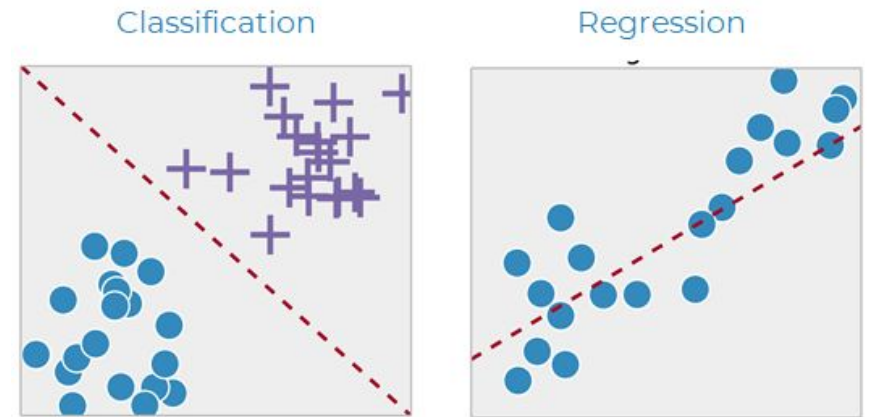
Supervised Learning

Supervised Learning merupakan tipe pembelajaran yang mirip seperti Guru memberikan contoh kepada siswa.

Pada supervised learning, mesin sudah diberikan informasi berupa label apa yang akan diprediksi.

Supervised Learning ini umum digunakan untuk melakukan 2 tipe analisis:

- **Regresi:** Model mencari output berupa data numerik
- **Klasifikasi:** Model mencari output berupa kelas yang sesuai dengan data input.



Supervised Learning



Fitur:

- Bentuk: Lonjong
- Warna: Kuning
- Tekstur: Lembut



Fitur:

- Bentuk: Bulat
- Warna: Merah
- Tekstur: Keras

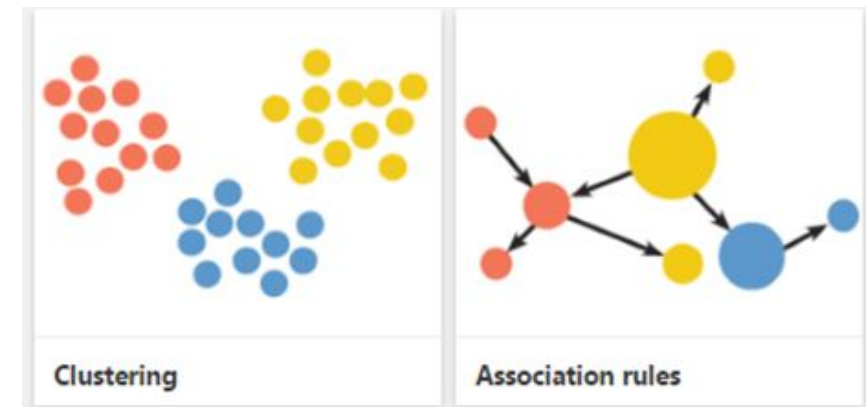
Bentuk	Warna	Tekstur	Buah
Lonjong	Kuning	Lembut	Pisang
Lonjong	Kuning	Lembut	Pisang
Bulat	Kuning	Keras	Apel

Unsupervised Learning

Unsupervised Learning merupakan tipe pembelajaran yang mirip seperti proses siswa belajar sendiri dengan melihat kesamaan yang ada. Pada unsupervised learning, mesin tidak diberikan informasi berupa label apa yang akan diprediksi.

Unsupervised Learning ini umum digunakan untuk menyelesaikan 2 tipe masalah:

- **Clustering:** Mengelompokkan data berdasarkan persamaan karakteristik
- **Association:** Menemukan asosiasi di antara item dalam data transaksi yang besar



Unsupervised Learning



Bentuk	Warna	Tekstur	Buah
Lonjong	Kuning	Lembut	1
Lonjong	Kuning	Lembut	1
Bulat	Kuning	Keras	2

Topik hari ini: Supervised Learning

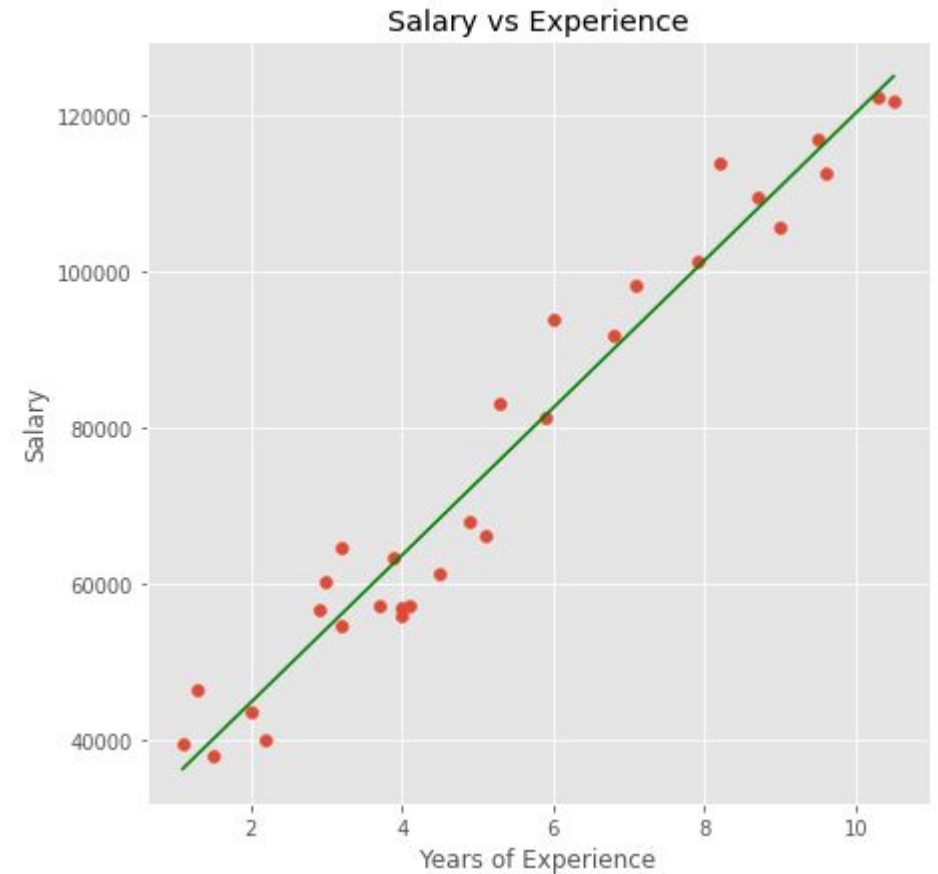


Regresi

Analisis regresi adalah suatu bentuk teknik pemodelan prediktif yang menyelidiki hubungan antara variabel dependen (target) dan variabel independen (prediktor).

Analisis regresi umum digunakan untuk:

- Memprediksi nilai variabel dependen berdasarkan nilai setidaknya satu variabel independen
- Menjelaskan dampak perubahan dalam variabel independen terhadap variabel dependen

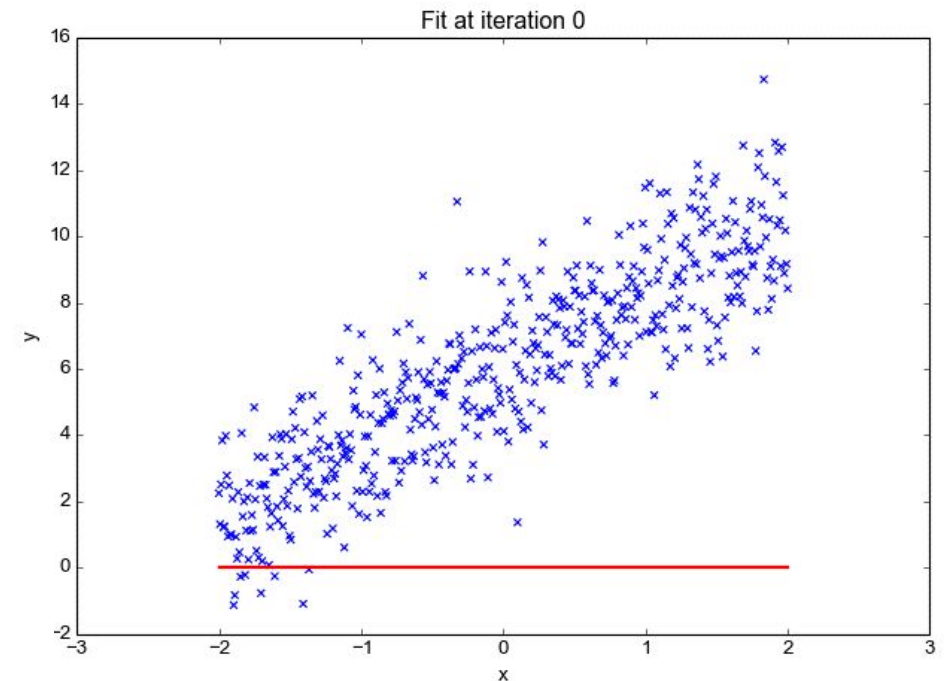


Regresi Linear

Regresi linier bertujuan untuk memprediksi nilai variabel dependen (y) berdasarkan variabel independen tertentu (x). Teknik regresi ini mencari hubungan linier antara x (input) dan y (output).

$$y = m \cdot x + b$$

$$y = \text{slope} \cdot x + \text{intercept}$$

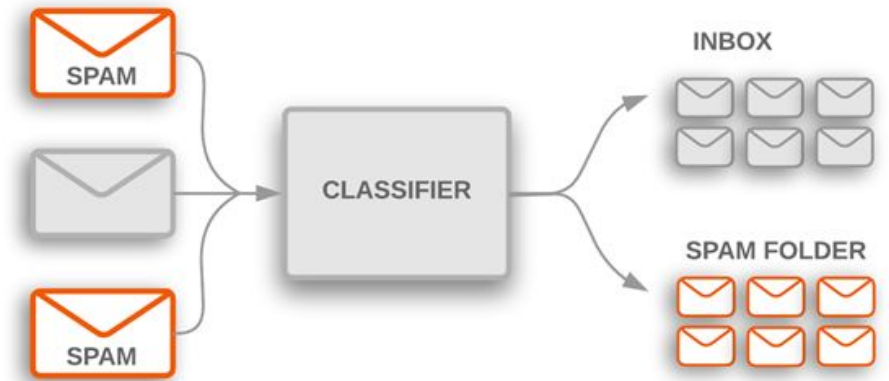


Klasifikasi

Analisis klasifikasi adalah suatu bentuk pemodelan yang mencari output/target kelas yang sesuai berdasarkan input yang diberikan.

Contoh Aplikasi:

- Memprediksi seorang customer akan “Churn” atau “Tidak Churn”
- Mengidentifikasi sebuah email merupakan sebuah “Spam” atau “Bukan Spam”



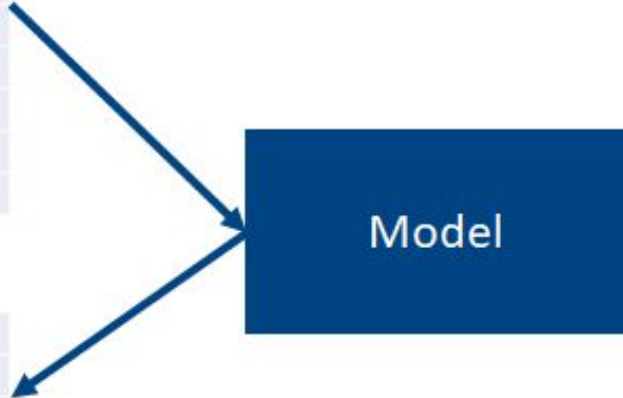
Klasifikasi

Algoritma Popular:

- Decision Tree
- SVM
- Naïve Bayes
- KNN
- Random Forest

No	Tipe Langganan	Jumlah Tagihan	Lama Berlangganan	Churn
1	Biasa	1250	3	Churn
2	Partnership	2000	2	Tidak
3	Partnership	1500	12	Tidak
4	Biasa	800	3	Churn
5	Biasa	95	4	Tidak
6	Partnership	2000	5	Tidak
7	Biasa	220	6	Tidak
8	Biasa	600	2	Churn
9	Biasa	700	1	Churn

10	Biasa	1200	7	
11	Partnership	750	6	
12	Biasa	500	8	
13	Partnership	600	10	



Model

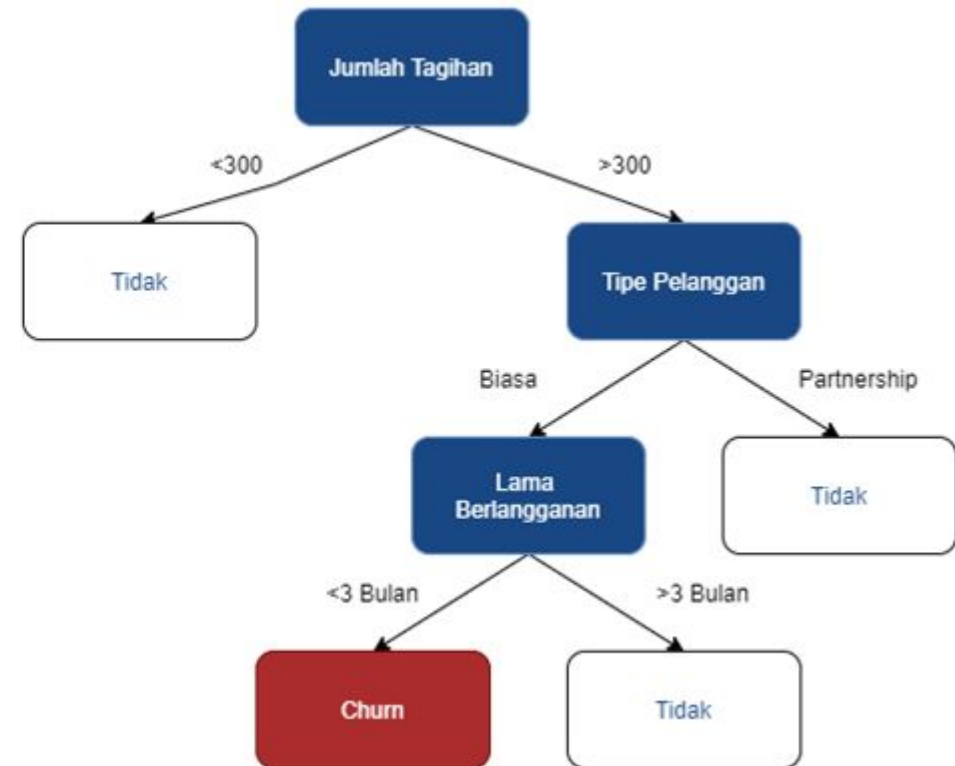
Decision Tree

Decision tree adalah model prediksi menggunakan struktur pohon atau struktur berhirarki.

Konsep Dasar dari decision tree adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan.

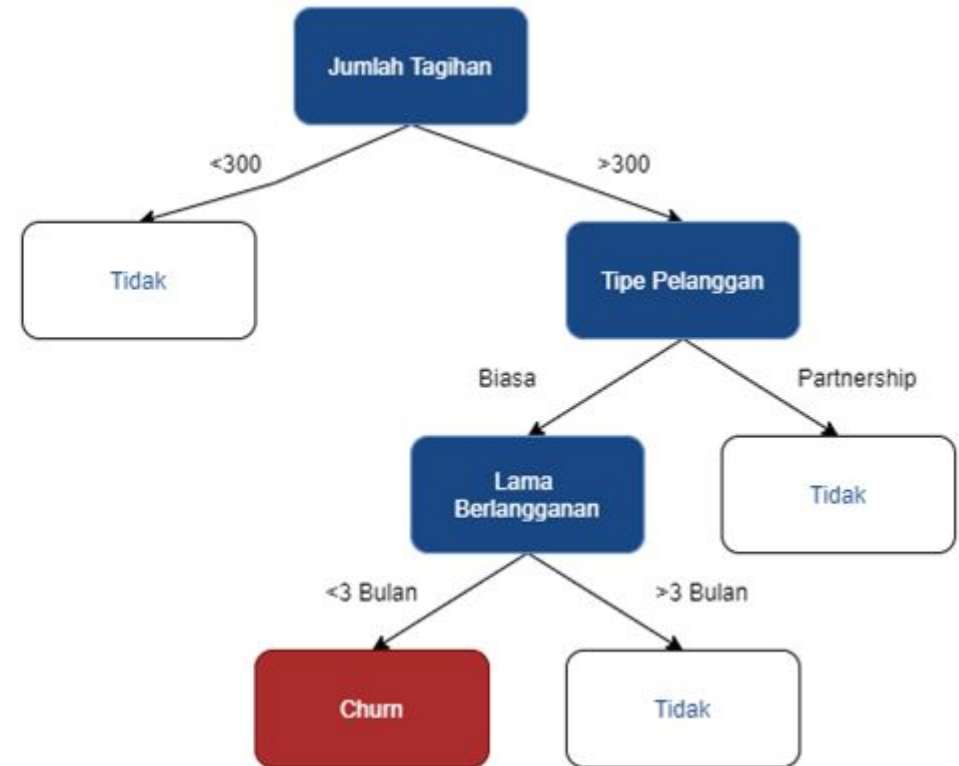
Decision Tree

No	Tipe Langganan	Jumlah Tagihan	Lama Berlangganan	Churn
1	Biasa	1250	3	Churn
2	Partnership	2000	2	Tidak
3	Partnership	1500	12	Tidak
4	Biasa	800	3	Churn
5	Biasa	95	4	Tidak
6	Partnership	2000	5	Tidak
7	Biasa	220	6	Tidak
8	Biasa	600	2	Churn
9	Biasa	700	1	Churn



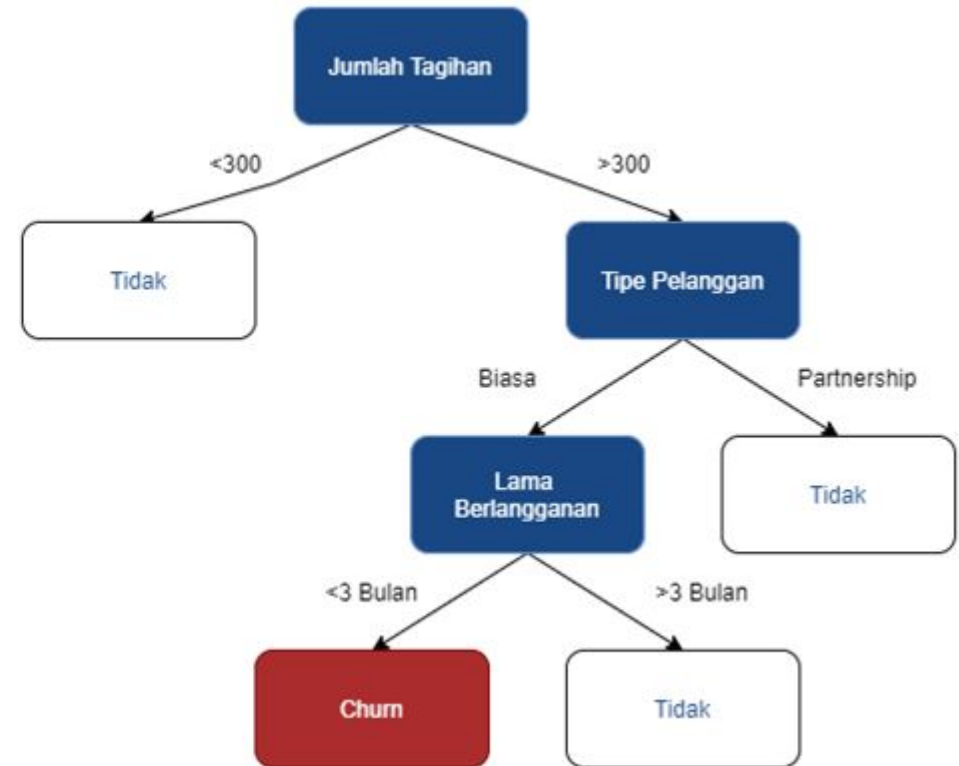
Pengaplikasian Decision Tree

10	Biasa	1200	7	?
11	Partnership	750	6	?
12	Biasa	500	8	?
13	Partnership	600	10	?

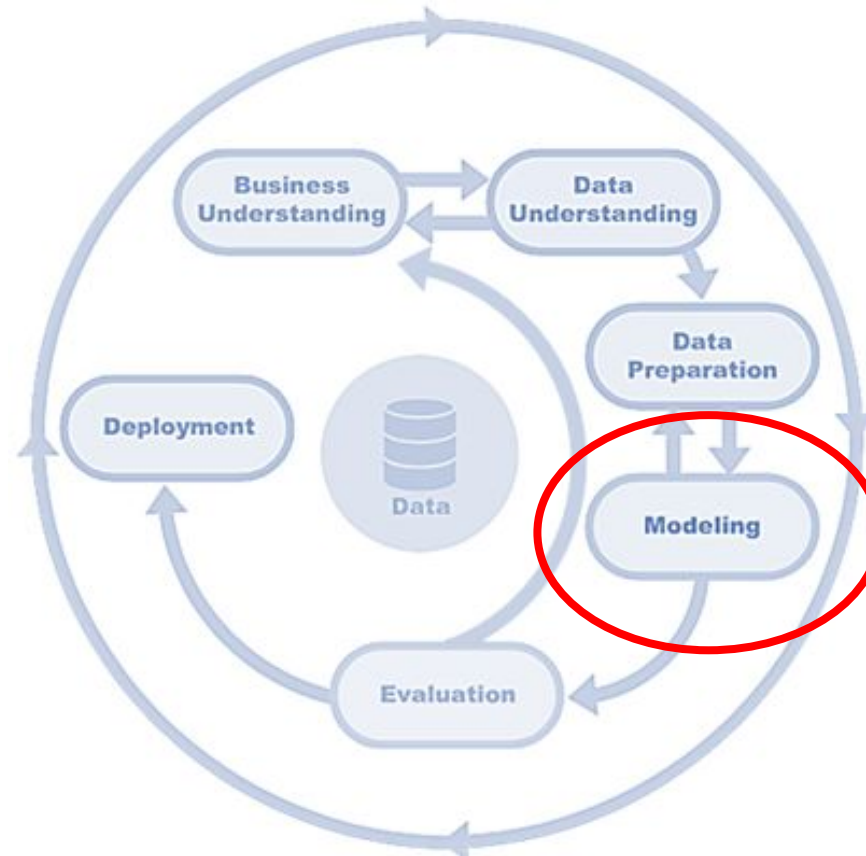


Pengaplikasian Decision Tree

10	Biasa	1200	7	?
11	Partnership	750	6	?
12	Biasa	500	8	?
13	Partnership	600	10	?



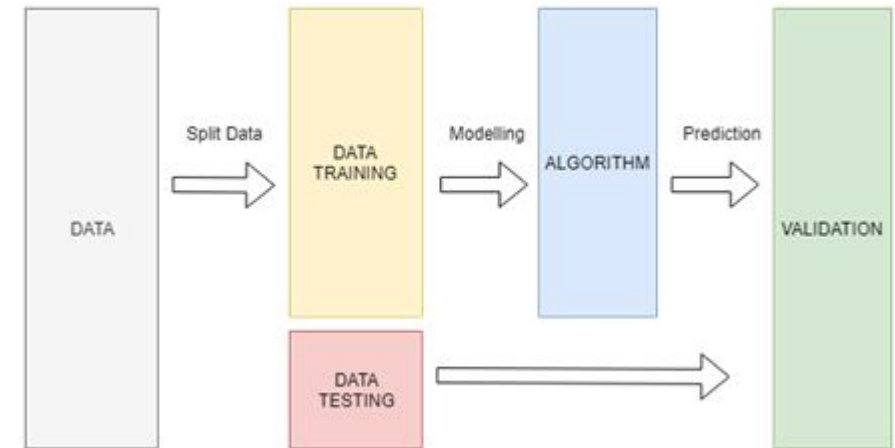
Machine Learning dalam CRISP-DM



Modeling

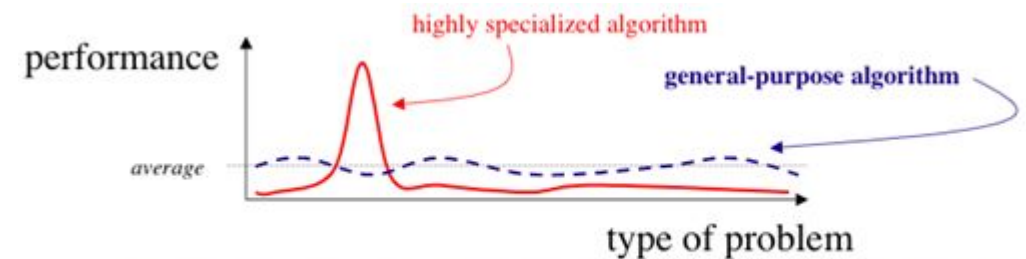
Membagi data Training dan Testing

- Umumnya dengan perbandingan 70:30 (70 untuk data Testing dan 30 untuk data training)
- Jika data cukup besar dapat menggunakan perbandingan 80:10 atau 90:10



Memilih Algoritma

- No free lunch theorem
- No algorithm is the best, No universal best algorithm.
- Tidak ada satu algoritma yang dapat bekerja sangat baik untuk setiap masalah



Source: <https://towardsdatascience.com/a-blog-about-lunch-and-data-science-how-there-is-no-such-a-thing-as-free-lunch-e46fd57c7f27>

Evaluasi Model

- Bertujuan untuk membantu kita mencari model yang dapat merepresentasikan data kita, dan dapat bekerja dengan baik kedepannya
- Biasanya dilakukan dengan mengkomparasi hasil prediksi dengan nilai yang sebenarnya.

Evaluasi Model Regresi

- **MAE (Mean Absolute Error)** : Mengukur perbedaan mutlak antara nilai aktual atau sebenarnya dan nilai yang diprediksi. Beda mutlak artinya jika hasilnya bertanda negatif, maka diabaikan.
- **MSE (Mean Square Error)** : Mengukur rata-rata kuadrat kesalahan — rata-rata selisih kuadrat antara nilai prediksi dengan nilai sebenarnya.
- **RMSE (Root Mean Square Error)** : Akar kuadrat dari MSE. Root kuadrat digunakan untuk membuat skala kesalahan menjadi sama dengan skala asli.

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}|$$

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2}$$

Evaluasi Model Klasifikasi

Evaluasi model klasifikasi umumnya menggunakan Confusion matrix atau yang dikenal juga dengan nama eror matrix. Confusion Matrix merupakan suatu metode validasi untuk mengukur akurasi prediksi suatu model dengan membandingkan nilai sebenarnya dengan nilai yang diprediksi oleh model.

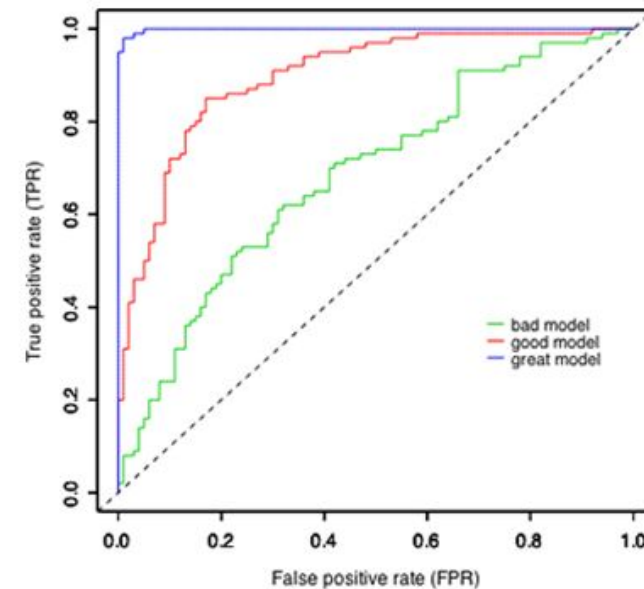
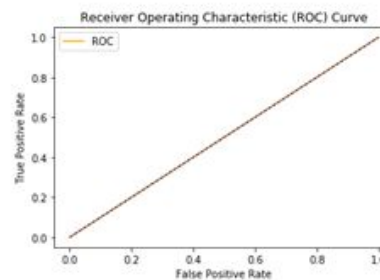
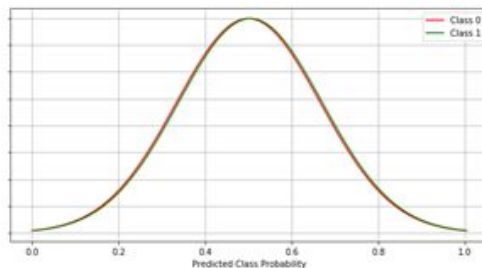
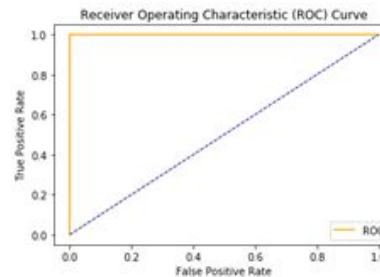
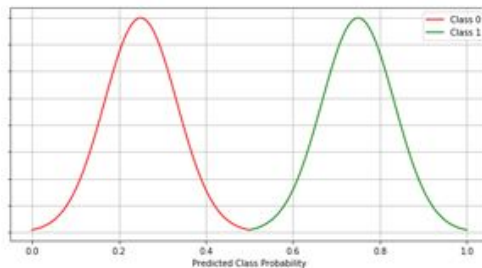
Metric dalam Confusion Matrix:

- **Accuracy** adalah proporsi dari jumlah total prediksi yang benar.
- **Sensitivity/Recall/True Positive Rate** adalah proporsi jumlah benar yang diprediksi sebagai benar.
- **Precision** adalah proporsi jumlah benar pada kelas positive

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

ROC

ROC (Receiver Operating Characteristic) adalah kurva probabilitas untuk kelas yang berbeda. ROC memberi tahu kita seberapa baik model untuk membedakan kelas yang diberikan, dalam hal probabilitas yang diprediksi





Praktek

California Housing Prices

Investasi properti/rumah merupakan salah satu instrumen investasi yang cukup populer. Sebelum melakukan investasi, seorang investor perlu untuk mengetahui harga seharusnya properti tersebut dengan melihat properti dengan spesifikasi sejenis, apakah properti yang akan dibeli tersebut overprice atau underprice. Sehingga, investor dapat menentukan apakah investasi tersebut merupakan investasi yang tepat.

Pada praktek kali ini, kita akan membuat model prediktif untuk memprediksi median harga rumah berdasarkan data lokasi rumah, spesifikasi rumah, dan demografi tetangga/penduduk sekitar.

Data yang digunakan adalah data California Housing Prices yang merupakan data sensus yang dilakukan pada tahun 1990.

Telco Customer Churn

Customer churn merupakan istilah yang digunakan untuk menjelaskan bahwa pelanggan telah menghentikan layanan pada provider tertentu, dan kemungkinan berpindah ke provider lain. Padahal, menjaga agar pelanggan tidak “churn” sangatlah penting bagi sebuah perusahaan. Menurut buku *Leading on The Edge of Chaos* karangan Emmett C. Murphy dan Mark A. Murphy, memperoleh customer baru biayanya lima kali lipat dibandingkan dengan memuaskan dan mempertahankan customer lama.

Di praktek kali ini kita akan mencoba untuk membuat model prediktif untuk memprediksi apakah seorang pelanggan akan melakukan “churn” atau tidak berdasarkan data historis.

kita akan menggunakan data pelanggan dari sebuah perusahaan telekomunikasi yang berisikan profil pelanggan, riwayat langganan pelanggan, dan informasi apakah pelanggan tersebut churn atau tidak.