BAB 1

The Fundamentals of Machine Learning

1.1. Pengertian Machine Learning

Terdapat beberapa definisi dari Machine Learning berdasarkan beberapa pakar dan beberapa referensi :

A. Machine Learning adalah visi computer yang dapat mempelajari dan meniru kecerdasan manusia. Machine Learning bekerja melalui pengembangan pengetahuan dan keterampilan baru berdasarkan pengalaman baik dengan pengawasan manusia atau tanpa pengawasan dari manusia.

B. Machine Learning adalah desain dan analisis artefak perangkat lunak yang menggunakan masa lalu untuk manjadi panduan pilihan di masa mendatang; Machine Learning belajar dari data.

C. Menurut Arthur Samuel Machine Learning adalah salah satu disiplin ilmu memberi komputer kemampuan untuk belajar tanpa diprogram secara eksplisit

Machine Learning merupakan suatu teknik yang dapat digunakan untuk melakukan ekstrasi serta mempelajari pola dari suatu data, berdasarkan pola tersebut komputer mampu melakukan prediksi dan mampu mengenali suatu trend tertentu. Sehingga hasil atau prediksi kejadian di masa mendatang dapat dihasilkan oleh computer tanpa perlu diprogram secara eksplisit. Contoh Aplikasi yang menerapkan mesin Learning :

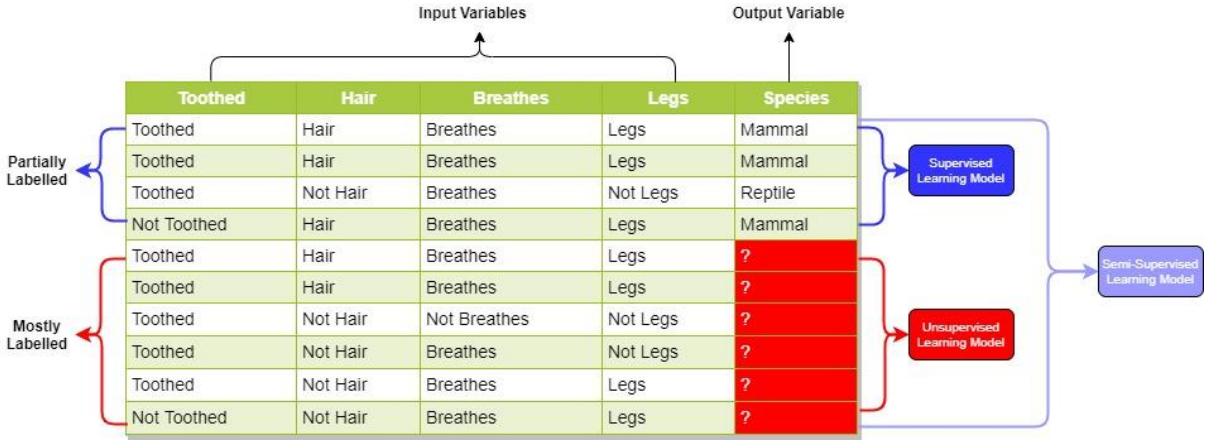
1. Sistem rekomendasi untuk E-commerce
2. Perintah kepada smartphone menggunakan suara
3. Pengaturan suhu otomastis
4. Pengenalan wajah pada aplikasi sosial media
5. Spam filtering pada e-mail
6. Face unlock pada smartphone
7. Prediksi cuaca

1.2. Tipe dari Machine Learning

Terdapat tiga masalah umum yang terdapat pada machine learning, yaitu : supervised learning, semi-supervised learning, and unsupervised learning. Pada supervised learning dibutuhkan dataset berlabel yang akan digunakan untuk memetakan atau menemukan label/class/target dari data yang tidak berlabel. Data berlabel yang digunakan dalam supervised learning disebut data training. Sedangkan data yang akan diprediksi disebut data training. pada teknik supervised learning prediksi variable output yang berupa label kelas / variable target berdasarkan pada data masukan(data training). Untuk melakukan ini, model supervised learning dikembangkan dengan menggunakan data training di mana nilai input dan output telah diketahui sebelumnya.

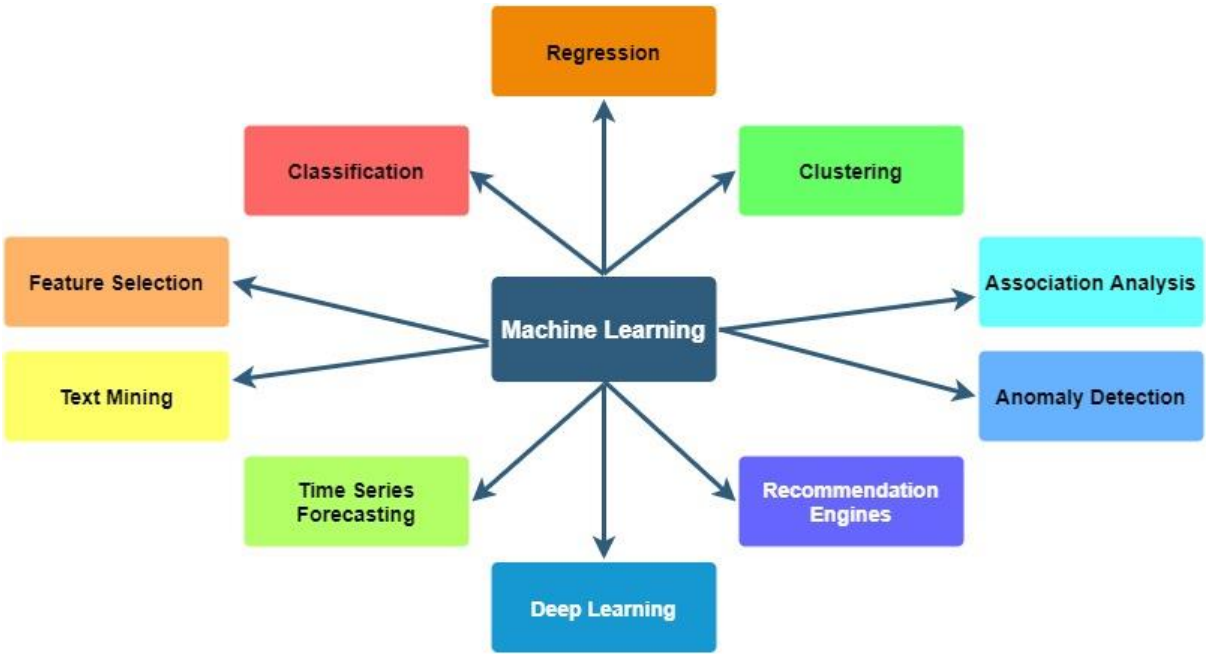
Pada unsupervised learning belum diketahui label/class dari suatu dataset. Teknik unsupervised learning akan mencari pola-pola data yang tersembuyi dari dataset yang tidak berlabel tersebut. Tujuan dari unsupervised learning adalah mencari pola-pola data berdasarkan kemiripan data.

semi-supervised learning terletak pada tengah-tengah diantar supervised learning dan unsupervised learning dimana data yang digunakan pada semi-supervised adalah data dari kedua Teknik supervised learning dan unsupervised learning. pada Teknik semi-supervised ini beberapa data adalah data berlabel akan tetapi Sebagian besar data adalah data yang tidak berlabel. Bisa juga data yang digunakan pada semi-supervised learning adalah data yang seluruhnya tidak berlabel sehingga diperlukan pencarian label secara otomatis (missal menggunakan clustering). Pada kategori ini biasanya data yang digunakan adalah sedikit kemudian diciptakan data tambahan baik menggunakan supervised learning maupun unsupervised learning. kemudian baru dibuat model dari data tambahan tersebut. Gambar 1 merupakan ilustrasi dari supervised learning, semi-supervised learning, dan unsupervised learning.



Gambar 1 Supervised, Semi-Supervised and Unsupervised Learning

Machine Learning dapat dikategorikan ke dalam tugas-tugas seperti: Klasifikasi, regresi, analisis asosiasi, pengelompokan, deteksi anomali, mesin rekomendasi, feature selection, peramalan deret waktu, deep learning, dan text mining seperti yang diilustrasikan pada Gambar 2.



Gambar 2 Kategori Machine Learning

Tasks	Description	Algorithms	Examples
Classification	Melakukan prediksi terhadap suatu data termasuk kedalam label yang mana. Prediksi didasarkan pada data yang telah memiliki label/kelas	Decision trees, neural networks, Bayesian models, induction rules, k-nearest neighbours	Deteksi dokumen plagiasi dan bukan plagiasi
Regression	Melakukan prediksi nilai label target (nilai numerik) dari suatu titik data.	Linear regression, logistic regression	prediksi harga mobil berdasarkan engine-size, body-style, horsepower
Anomaly detection	Melakukan prediksi apakah data termasuk outlier dibantdngkan dengan kumpulan dataset yang lain.	Distance-based, density-based, LOF	Deteksi penipuan transaksi kartu kredit, deteksi gangguan jaringan
Time series forecasting	peramalan berdasarkan perilaku data masa lampau untuk diproyeksikan ke masa depan	Exponential smoothing, ARIMA, regression	Peramalan penjualan, peramalan produksi
Clustering	Menemukan sebuah kumpulan data atau objek yang memiliki kemiripan satu sama lain di dalam kumpulan atau kelompok tersebut, dan berbeda dengan objek di kelompok lain	k-Means, density-based clustering (e.g., DBSCAN)	Menemukan segmentasi pelanggan berdasarkan transaksi yang dilakukan
Association analysis	Identifikasi hubungan antara satu set item berdasarkan data transaksi	FP-growth algorithm, a priori algorithm	Menemukan peluang penjaualan barang berdasarkan histori transaksi yang dilakukan pembeli
Recommendation engines	Memprediksi preferensi item untuk pengguna	Collaborative filtering, content-based filtering, hybrid recommenders	Menemukan film paling direkomendasikan untuk pengguna

1.3.Training data, testing data AND VALIDATION DATA

Pada Machine Learning pembuatan model tentunya membutuhkan data. Data ini disebut sebagai DATASET yaitu Sekumpulan data yang digunakan dalam machine learning. Umumnya dataset bisa dibagi menjadi dua yaitu data training (data latih) dan data testing (data uji). Data Training adalah data yang digunakan untuk melatih model machine learning yang dibangun. Sedangkan Data Testing adalah sekumpulan data yang digunakan untuk menguji atau mengevaluasi kinerja dari model yang sudah dilatih. Pada Umumnya dataset ditampilkan dalam bentuk tabel erdiri dari baris dan kolom. Bagian Nama Kolom merupakan FEATURE atau VARIABEL data yang selanjutnya akan dianalisa, sedangkan pada setiap barisnya merupakan DATA POINT/OBSERVATION/EXAMPLE. Pada machine learning seringkali akan memprediksi suatu kategori/nilai numerik hal ini disebut LABEL/CLASS/TARGET.

Dalam statistika/matematika, LABEL/CLASS/TARGET ini dinamakan dengan Dependent Variabel, dan FEATURE adalah Independent Variabel.

Berikut merupakan Beberapa istilah yang digunakan dalam machine learning :

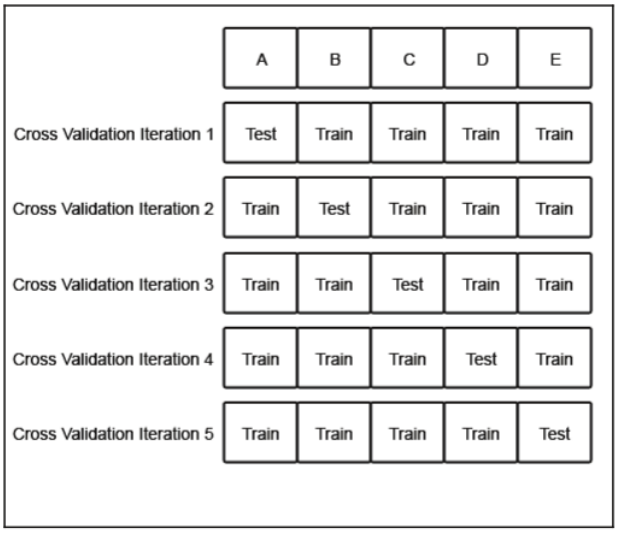
Istilah	**Pengertian**
Dataset	sekumpulan data yang akan diolah
Training Dataset	data yang digunakan untuk melatih model machine learning
Test Dataset	data yang digunakan untuk menguji model yang telah dilatih
Algorithm	pendekatan yang digunakan untuk membangun model tersebut, seperti seperti Decision Tree, K-NN, Linear Regression, Random Fores, dan SVM
Model	output/hasil dari proses latih dan uji pada dataset yang kita ,miliki
Feature	penyajian tabel dataset bagian kolom
Variabel	penyajian tabel dataset bagian baris
Label/Class/Target	hal yang menjadi target atau yang diprediksi dengan machine learning
Dependent Variable	penamaan dalam ilmu statistika dan matematika untuk label/class/target
Independent Variable	penamaan dalam ilmu statistika dan matematika untuk Feature

Evaluasi Performa dari Model Machine Learning

Selama pengembangan terutama pada saat terdapat data pelatihan yang tidak mencukupi dapat dilakukan Teknik cross validation untuk melakukan pelatihan data dan evaluasi pada data yang sama. Pada gambar 3 dapat dilihat bahwa data dibagi menjadi lima bagian / partisi. Dimana model dilatih menggunakan empat partisi data dan salah satu bagian data digunakan sebagai data testing untuk evaluasi. Selanjutnya proses akan dilakukan berulang-ulang dengan menggunakan data testing yang berbeda-beda sesuai dengan partisinya. Penggunana Teknik cross validation memiliki perkiraan performa yang lebih baik dibandingkan evaluasi menggunakan satu pelatihan / pengujian terpisah.

Perhatikan gambar 3, pada gambar 3 dataset dibagi menjadi 5 buah partisi yang sama kemudian diberi label A, B, C, D, dan E. pada iterasi pertama model dilatih

menggunakan data pada partisis B sampai E dan data latih yang digunakan adalah data pada partisi A. pada iterasi berikutnya model dilatihi dengan menggunakan data pada partisi A, C, D, dan E, kemudian diuji dengan menggunakan data pada partisi B. Partisi diputar hingga model dilatih dan diuji pada semua partisi.



Gambar 3. 5-folds Cross validation

BIAS AND VARIANCE

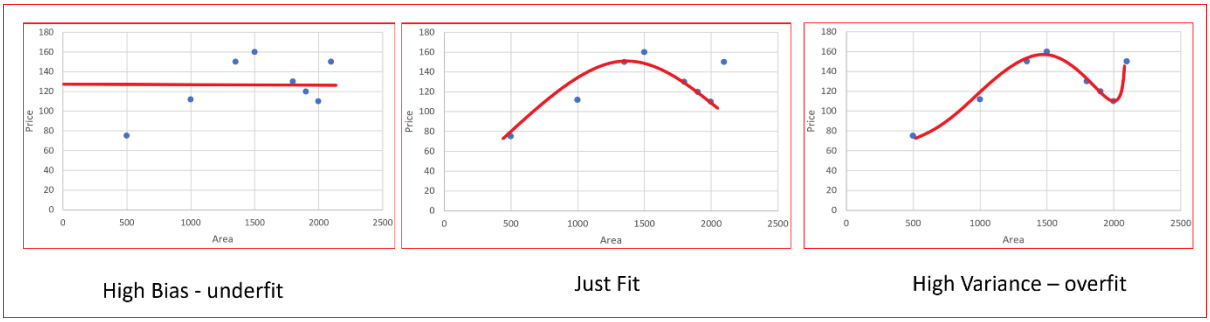
Banyak metrik dapat digunakan untuk mengukur apakah suatu program berjalan secara lebih efektif atau tidak dalam melaksanakan tugasnya. Pada supervised learning terdapat beberapa metrik yang dapat digunakan untuk mengujur kinerja dari kesalahan prediksi. Ada dua penyebab mendasar dari kesalahan dalam melakukan prediksi yaitu bias dan variance.

Bias

Bias seringkali terjadi dalam development sistem machine learning. Bias disebut juga error pada data training yaitu seberapa jauh perbedaan antara hasil prediksi dari model Machine Learning yang dikembangkan dengan data nilai yang sebenarnya. Jika Jika rata-rata nilai prediksi jauh dari nilai sebenarnya maka menyebabkan bias yang tinggi. Ketika sebuah model memiliki bias yang tinggi maka hal tersebut mengimplikasikan bahwa model tersebut terlalu sederhana dan tidak menangkap kompleksitas data sehingga akan terjadi underfitting.

Variance

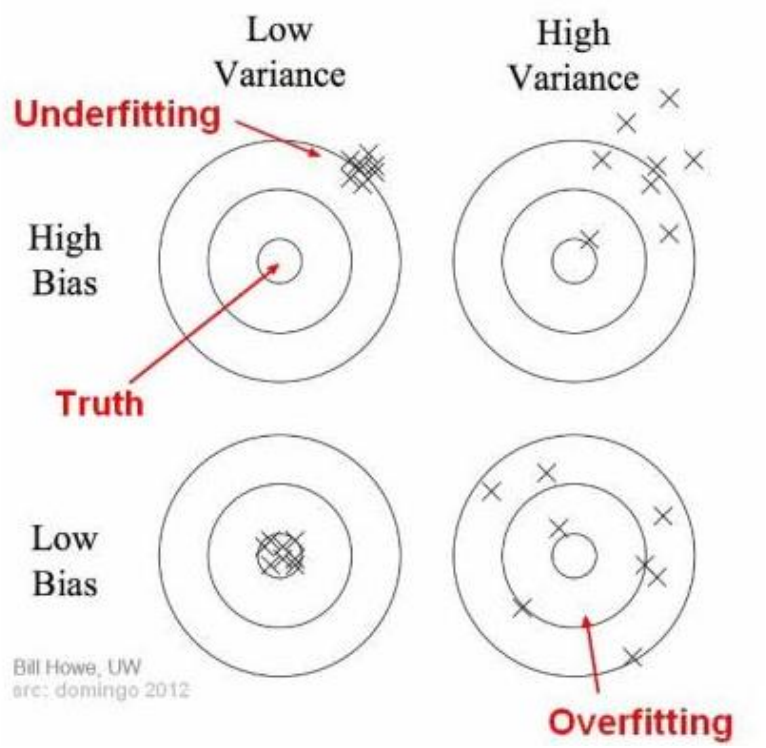
Varians terjadi saat model berperforma baik pada data set yang dilatih, tetapi model tersebut tidak berfungsi dengan baik pada data uji. Variance menjukkan informasi persebaran data yang diprediksi dibandingkan dengan nilai yang sebenarnya. Model yang memiliki data dengan variance tinggi sangat memperhatikan hanya pada data latih daja. High variance model, memiliki performa yang baik pada data latih. Akan Tetapi jika diberikan data baru yang belum pernah ditemukan pada data model, model tersebut tidak akan memberikan hasil yang baik terhadap data yang baru. Sehingga dapat menyebabkan nilai prediksi yang keliru. Variance tinggi menyebabkan overfitting yang artinya bahwa terdapat noise yang ada dalam data pelatihan sehingga model yang dibagun juga mempelajari noise yanga ada pada data latih tersebut.



Ilustrasi Bias Dan Variance

Anak panah dan papan dart dapat digunakan untuk melakukan visualisasi variance dan bias. Setiap anak panah dianalogikan dengan prediksi, dan dilemparkan oleh model yang dilatih pada kumpulan data yang berbeda setiap saat. Model dengan bias tinggi tetapi variance rendah akan melempar anak panah yang akan berkerumun rapat, tetapi mungkin jauh dari lingkar sasaran (sasaran utama). Model dengan bias tinggi dan varians tinggi akan melempar anak panah ke seluruh papan mengakibatkan anak panah jauh dari sasaran dan dari data satu sama lain. Model dengan bias rendah dan varians tinggi akan melempar anak panah yang mungkin

tidak terkumpul dengan baik tetapi dekat dengan sasaran. Akhirnya, model dengan bias rendah dan varians rendah akan melempar anak panah yang berkerumun erat di sekitar sasaran. Gambar 4 merupakan visualisasi dari bias dan variance.



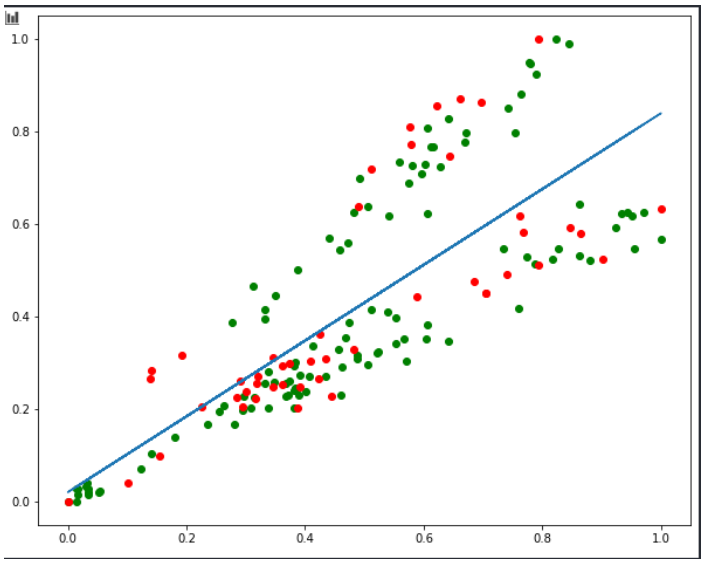
Gambar 4

Data yang memiliki Bias yang tinggi dengan variance yang rendah akan menjadi underfitting. Sementara jika dengan bias tinggi dan juga high variance menjadikan prediksi sangat tidak tepat. Jika biasnya rendah dan variancenya tinggi akan menimbulkan overfitting dimana dengan data train, perform baik tapi ketika diberikan data baru, tidak dapat memprediksi. Pastinya yang paling baik jika bias rendah dan variance rendah.

Kondisi Underfiiting (low variance – high bias)

scatter plot yang berwarna hijau merupakan data training dan scatter plot yang berwarna merah merupakan data testing. Bisa dilihat visualisasi data pada gambar 5

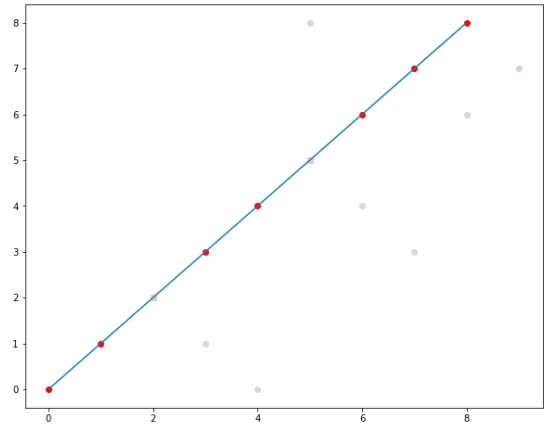
bahwa data training maupun testing menjauhi garis regresi, Seharusnya data yang baik adalah data yang semakin mendekati garis fungsi regresi.



Gambar 5 ilustrasi underfitting

Kondisi Overvitting (high varians – low bias)

Scatter plot yang berwarna merah merupakan data training dan scatter plot yang berwarna abu-abu adalah data testing, bisa dilihat bahwa pada gambar 6 data training sempurna segaris dengan garis regresinya namun jika kita perhatikan data testingnya sangat jauh dari garis tersebut.

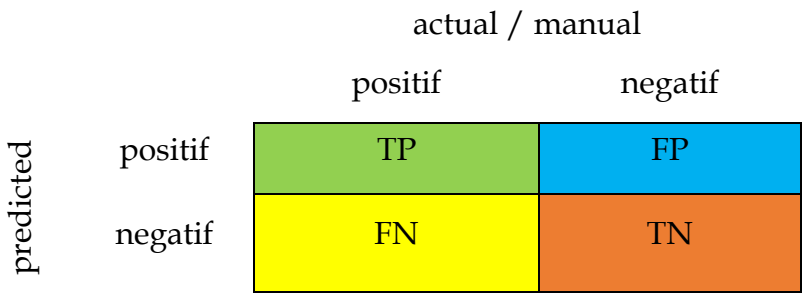


Gambar 6 ilustrasi overfitting

Mengatasi data dengan bias tinggi

-

Tahap terakhir dari membuat model adalah melakukan evaluasi dari hasil model yang telah dibangun. Untuk melakukan evalusi performa dari metode evalusi yang dapat digunakan adalah menggunakan Confusion Matrix. Confusion Matrix merepresentasikan perbandingan prediksi dihasilkan oleh algoritma machine learning dengan LABEL sebenarnya yang terdapat pada dataset.



True Positive (TP): Kondisi dimana hasil sistem memberikan hasil prediksi Positif dan hasil dari actual data juga positif

True Negative (TN): Kondisi dimana hasil sistem memberikan hasil prediksi negatif dan hasil dari data aktualnya juga menghasilkan negatif

False Positive (FP): Kondisi dimana hasil sistem memberikan hasil prediksi Positif, tetapi ternyata dari actual data adalah negatif

False Negatif (FN): Kondisi dimana hasil sistem memberikan hasil prediksi negatif akan tetapi hasil dari data aktualnya menghasilkan positif.

Untuk menghasilkan nilai akurasi, presisi, recall, dan f-measure adalah dengan menggunakan rumus berikut ini :

$$\text{Accuracy} = \frac{(TP + TN)}{(TP+FP+FN+TN)}$$

$$\text{Precision} = \frac{TP}{(TP+FP)}$$

$$\text{Recall} = \frac{TP}{(TP+FN)}$$

$$\text{F1 Score} = 2 * \frac{(\text{Presisi} * \text{recall})}{(\text{recall}+\text{Presisi})}$$