



Project Based Internship

Machine Learning

Supervised Learning



Daftar Isi

Supervised Learning	3
Algoritma Supervised Learning	4
Regression	5
Classification	9
Hyperparameter Tuning	11
Use Case	12
References	15

Supervised Learning

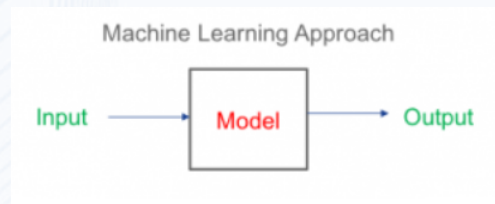


Figure 1: Machine learning Approach

Supervised Machine Learning, atau lebih umumnya disebut Supervised Learning adalah salah satu jenis Machine Learning yang paling umum digunakan. Supervised Learning mengacu pada algoritma yang mempelajari pemetaan x ke y atau pemetaan Input (masukan) ke Output (keluaran). Karakteristik utama dari Supervised Learning adalah Anda memberikan contoh kepada algoritma pembelajaran untuk dipelajari, ini termasuk jawaban yang benar. Dilanjutkan dengan mempelajari hubungan antara data yang tersedia dengan label/target yang tepat ketika proses training. Kemudian setelah training, model supervised learning akan mengambil data baru sebagai input dan kemudian memprediksi label dari data tersebut berdasarkan data yang digunakan saat training.

Mari kita lihat beberapa contoh. Jika masukan x adalah email dan keluaran y adalah email ini, spam atau bukan spam, ini memberikan Anda penyaring spam. Atau jika masukan berupa klip audio dan tugas algoritma adalah mengeluarkan transkripsi teks, maka ini adalah pengenalan ucapan. Atau jika Anda ingin memasukkan bahasa Inggris dan menghasilkan terjemahan ke bahasa Spanyol, Arab, Hindi, Cina, Jepang, atau terjemahan bahasa lainnya, maka itu adalah terjemahan mesin. Atau bentuk paling menguntungkan dari Supervised Learning saat ini mungkin digunakan dalam periklanan online. Hampir semua platform iklan online besar memiliki algoritma

pembelajaran yang memasukkan beberapa informasi tentang iklan dan beberapa informasi tentang Anda, lalu mencoba untuk mencari tahu apakah Anda akan mengklik iklan tersebut atau tidak. Karena dengan menampilkan iklan yang sedikit lebih mungkin Anda mengklik, bagi platform iklan online besar, setiap klik adalah pendapatan, ini benar-benar menghasilkan banyak pendapatan bagi perusahaan-perusahaan ini.

Atau ambil contoh manufaktur. Anda bisa menggunakan algoritma pembelajaran yang mengambil gambar produk yang diproduksi sebagai masukan, katakanlah ponsel yang baru saja keluar dari garis produksi, dan algoritma pembelajaran menghasilkan apakah ada goresan, penyok, atau cacat lain dalam produk tersebut. Ini disebut inspeksi visual dan membantu produsen mengurangi atau mencegah cacat dalam produk mereka. Dalam semua aplikasi ini, Anda akan pertama kali melatih model Anda dengan contoh masukan x dan jawaban yang benar, yaitu label y . Setelah model belajar dari pasangan masukan, keluaran, atau pasangan x dan y ini, mereka kemudian dapat menggunakan masukan x yang sama sekali baru, sesuatu yang belum pernah dilihat sebelumnya, dan mencoba menghasilkan keluaran y yang sesuai.

Algoritma Supervised Learning

Umumnya, algoritma supervised learning dapat dibedakan menjadi dua jenis:

- Klasifikasi, yaitu jenis supervised learning di mana tipe data dari variabel label/target adalah diskrit.
- Regresi, yaitu jenis supervised learning di mana tipe data dari variabel label/target adalah numerikal kontinu.

Masing-masing jenis supervised learning tersebut menggunakan algoritma yang berbeda dan juga metrik yang digunakan untuk mengukur performa antar keduanya juga akan berbeda. Contoh metrik performa yang digunakan pada regresi adalah Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE). Sedangkan, contoh metrik performa yang dipakai untuk mengukur performa klasifikasi adalah akurasi, precision, recall, AUC.

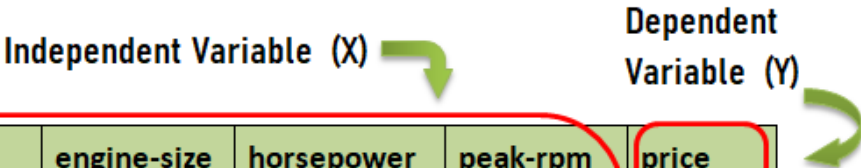
Ada banyak algoritma yang dapat digunakan untuk melakukan training model supervised learning. Perlu diperhatikan bahwa algoritma yang dipakai dalam Klasifikasi dan Regresi berbeda.

Beberapa contoh algoritma supervised learning adalah sebagai berikut:

- Linear Regression (untuk regresi)
- Logistic Regression (untuk klasifikasi)
- Support Vector Machine
- Decision Tree
- Random Forest
- XGBoost

Regression

Regresi merupakan sebuah metode analisis yang bertujuan untuk mengidentifikasi hubungan antara dua variabel atau lebih. Tujuan utama dari regresi adalah menemukan fungsi yang dapat memodelkan data dengan cara meminimalkan kesalahan atau perbedaan antara nilai prediksi dan nilai sebenarnya.



Independent Variable (X)				Dependent Variable (Y)
body-style	engine-size	horsepower	peak-rpm	price
convertible	130	111	5000	13495
convertible	130	111	5000	16500
hatchback	152	154	5000	16500
sedan	109	102	5500	13950
sedan	136	115	5500	17450
wagon	136	110	5500	18920
hatchback	131	160	5500	?

Figure 2: Independent and Dependent Variable

Teknik regresi termasuk dalam Supervised Learning yang digunakan untuk memprediksi nilai yang bersifat kontinu. Contohnya, mari perhatikan potongan data dari kumpulan data mobil. Namun, perlu diingat bahwa kumpulan data mobil yang sebenarnya memiliki 26 variabel, sedangkan contoh ini hanya mempertimbangkan beberapa di antaranya, seperti gaya bodi, ukuran mesin, tenaga kuda, putaran per menit puncak, dan harga.

Dalam regresi, ada dua jenis variabel yang berperan, yaitu variabel terikat dan variabel bebas. Variabel terikat merupakan variabel yang ingin kita prediksi atau pelajari, sedangkan variabel bebas adalah variabel yang digunakan untuk menjelaskan atau mempengaruhi nilai target pada variabel terikat.

Umumnya, variabel bebas dilambangkan dengan X, sedangkan variabel terikat dilambangkan dengan Y. Hal yang perlu diperhatikan dalam kasus regresi adalah bahwa nilai dari variabel terikat (Y) harus bersifat kontinu, bukan diskrit. Sementara

itu, variabel bebas (X) bisa memiliki nilai kontinu maupun kategorikal, seperti jenis mobil seperti sedan, hatchback, wagon, atau konvertibel.

Dalam rangka memprediksi harga mobil, langkah pertama adalah membuat model regresi berdasarkan data yang telah ada. Setelah model tersebut selesai dibangun, kita dapat menggunakannya untuk memprediksi harga mobil dengan menggunakan data baru.

Pada dasarnya, ada dua tipe model regresi yaitu *simple regression* (regresi sederhana), dan *multiple regression* (regresi berganda).

Simple regression adalah ketika hanya satu variabel independen yang digunakan untuk memprediksi dependen variabel, bisa berupa linear (*simple linear regression*) maupun non-linear (*simple non-linear regression*). Misalnya memprediksi harga mobil hanya dengan berdasarkan *engine-size* saja.

Multiple regression adalah ketika ada lebih dari satu variabel independen yang digunakan untuk memprediksi variabel dependen. Misalnya memprediksi harga mobil berdasarkan *engine-size*, *body-style*, *horsepower*, dan sebagainya.

Linearitas regresi ditentukan berdasarkan sifat hubungan antara variabel independen dan dependen. Sama seperti *simple regression*, *multiple regression* juga bisa berupa linear maupun non-linear.

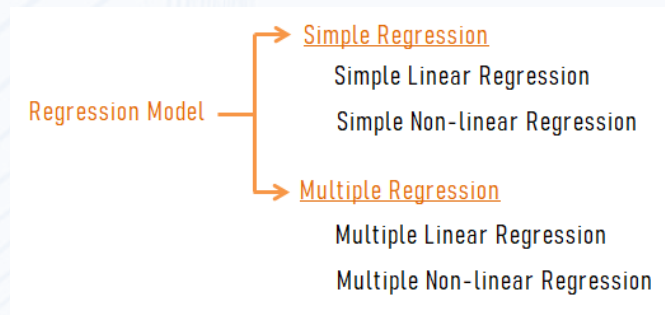


Figure 3: Regression Model

Algoritma regresi

Linear regression merupakan salah satu algoritma regresi yang paling populer. Kenyataannya, bukan hanya *linear regression* saja yang dapat kita gunakan untuk pemodelan regresi, tetapi ada banyak algoritma lainnya yang dapat kita coba, di antaranya:

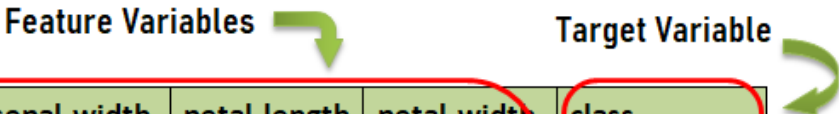
- [Linear Regression](#)
- [Lasso Regression and Ridge Regression](#)
- [Decision Tree Regression](#)
- [Neural Network Regression](#)

Algoritma-algoritma yang disebutkan di atas hanya sebagian dari teknik yang dapat digunakan untuk kasus regresi. Seiring dengan proses belajar, mungkin kita akan menemukan serta dapat mengeksplorasi teknik-teknik regresi lainnya.

Classification

Seperti regresi, klasifikasi juga termasuk dalam kategori pembelajaran berawasi. Klasifikasi merupakan metode untuk mengelompokkan atau mengkategorikan sejumlah item yang belum diberi label ke dalam sejumlah kelas diskrit.

Klasifikasi berusaha memahami relasi antara kumpulan fitur dan variabel target. Dalam klasifikasi, variabel target memiliki bentuk kategori.



Feature Variables				Target Variable
sepal-length	sepal-width	petal-length	petal-width	class
5.1	3.5	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
5.9	3.2	4.8	1.8	Iris-versicolor
6.8	2.8	4.8	1.4	Iris-versicolor
6.9	3.2	5.7	2.3	Iris-virginica
7.4	2.8	6.1	1.9	Iris-virginica
6.2	2.8	4.8	1.8	?

Figure 4: Feature and Target Variable

Elemen-elemen variabel target dan fitur dalam konteks klasifikasi terlihat dalam contoh tabel di atas. Klasifikasi memiliki dua jenis variabel, yaitu variabel target dan variabel fitur. Paralel dengan regresi, perbedaan utama antara klasifikasi dan regresi terletak pada nilai variabel target. Pada klasifikasi, variabel target harus berupa kategori atau nilai diskrit. Data baru yang sudah diberi label akan dikelompokkan ke dalam salah satu kategori yang ada dalam variabel target.

Di dalam machine learning, klasifikasi dibagi menjadi dua jenis yaitu *Binary classification* dan *Multi-class classification*.

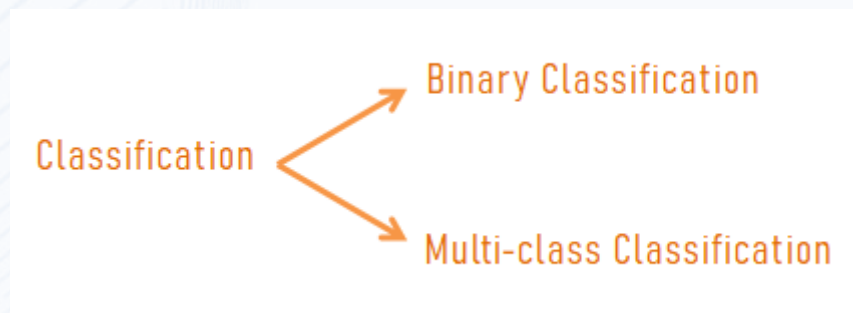


Figure 5: Classification Types

Binary classification adalah jika kategori dalam target variabel hanya ada dua, misalnya 0 dan 1, Yes dan No, X dan Y, dan sebagainya. Misalnya melihat kemungkinan nasabah bank akan mengambil pinjaman atau tidak.

Sebaliknya, *Multi-class classification* memiliki lebih dari dua kategori pada variabel targetnya. Contoh dataset Iris di atas termasuk jenis *multi-class classification* karena memiliki tiga kategori yaitu Iris-setosa, Iris-versicolor, dan Iris-virginica.

Algoritma klasifikasi

Untuk menyelesaikan masalah klasifikasi, kita dapat menggunakan algoritma-algoritma berikut ini, di antaranya:

- [Decision Tree](#)
- [Naïve Bayes](#)
- [K-Nearest Neighbor \(KNN\)](#)
- [Logistic Regression](#)
- [Support Vector Machines \(SVM\)](#)
- [Neural Network](#)

Perlu dicatat bahwa, tidak seperti namanya, *Logistic regression* bukan merupakan algoritma untuk menyelesaikan masalah regresi, melainkan klasifikasi.

Hyperparameter Tuning

Ketika membangun suatu model machine learning, kita akan diperhadapkan dengan pilihan-pilihan untuk mendesain struktur dari model machine learning tersebut. Sering kali, pembuat model tidak akan langsung tahu secara pasti struktur model seperti apa yang terbaik sehingga dibutuhkan percobaan untuk mengubah parameter-parameter model. Parameter yang mendefinisikan struktur model machine learning disebut hyperparameter, sementara itu hyperparameter tuning adalah proses pencarian parameter-parameter yang paling sesuai untuk menciptakan suatu struktur model machine learning yang terbaik.

Contoh metode dalam melakukan hyperparameter tuning yang paling populer adalah sebagai berikut:

1. Grid Search, yaitu hyperparameter tuning yang dilakukan dengan menyediakan sekumpulan nilai-nilai hyperparameter yang hendak diuji, kemudian mesin akan melakukan pencarian terhadap seluruh kombinasi nilai yang disediakan dan mengembalikan informasi kombinasi mana yang terbaik.
2. Random Search, yaitu hyperparameter tuning yang dilakukan dengan hanya menyediakan distribusi statistik dari calon nilai-nilai hyperparameter yang hendak diuji (berbeda dengan Grid Search yang menyediakan nilai-nilai diskrit).

Use Case

Background & Problem Statement:

Kamu adalah seorang data scientist di IDX Partners dan kamu saat ini sedang melakukan pemodelan Machine Learning. Data yang akan kamu gunakan adalah data terkait data customer yang telah churn dan yang masih tidak churn (nilainya 0 jika tidak churn dan nilainya 1 jika churn) dari perusahaan.

Data tersebut memiliki 6 kolom yaitu gender, age, days_employed, is_married, churn, have_house

Jelaskan langkah-langkah yang akan kamu gunakan untuk melakukan pemodelan ML! Jelaskan apakah kamu akan menggunakan classification atau regression dan model apa yang akan kamu gunakan, lalu identifikasi mana features dan target!

Identifikasi Features dan Target:

- Features (Variabel Independen):
 - gender
 - age
 - days_employed
 - is_married
 - have_house
- Target (Variabel Dependen):
- churn (0 jika tidak churn, 1 jika churn)

Jenis Model:

Karena ini adalah masalah klasifikasi dengan variabel target biner, saya mungkin akan mencoba menggunakan Logistic Regression, Decision Trees, atau Random Forest.

Tahapan Pengerjaan:

- **Pemahaman Data:**
 - Memahami struktur data dan arti dari setiap kolom.
 - Menilai apakah terdapat missing values dan bagaimana menanganinya.
 - Mengeksplorasi statistik deskriptif dari data untuk mendapatkan pemahaman awal.
- **Preprocessing Data:**
 - Melakukan encoding pada variabel kategorikal seperti gender, is_married, dan have_house.
 - Menangani missing values jika ada, misalnya dengan imputasi atau penghapusan baris/kolom tertentu.
 - Menstandarisasi skala jika diperlukan, terutama pada kolom seperti age dan days_employed.
- **Pemilihan Fitur (Feature Selection):**
 - Menganalisis korelasi antar fitur untuk mengidentifikasi ketergantungan.
 - Memilih fitur-fitur yang memiliki pengaruh signifikan terhadap variabel target churn.
- **Pembagian Data:**

- Membagi data menjadi set pelatihan (training set) dan set pengujian (testing set) untuk mengukur kinerja model secara objektif.
- Pemilihan Model:
 - Karena targetnya adalah untuk memprediksi apakah pelanggan akan churn atau tidak (0 atau 1), ini adalah masalah klasifikasi. Oleh karena itu, saya akan menggunakan algoritma klasifikasi.
 - Algoritma klasifikasi yang umum digunakan termasuk Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), dan Neural Networks.
 - Pemilihan model akan bergantung pada karakteristik data dan kebutuhan bisnis.
- Pelatihan Model:
 - Melatih beberapa model pada set pelatihan dan menilai kinerjanya menggunakan metrik yang relevan seperti akurasi, presisi, recall, dan F1-score.
 - Melakukan penyetelan parameter (hyperparameter tuning) jika diperlukan untuk meningkatkan kinerja model.
- Evaluasi Model:
 - Menggunakan set pengujian untuk mengevaluasi performa model yang telah dilatih.
 - Memeriksa metrik evaluasi untuk memastikan bahwa model memiliki kinerja yang baik dalam memprediksi churn.

References

<https://towardsdatascience.com/a-brief-introduction-to-supervised-learning-54a3e3932590>

<https://medium.com/beridata/supervised-dan-unsupervised-learning-apa-perbedaanya-e64acc0f5f79>

<https://www.vpslabs.net/supervised-learning/>

<https://ilmudatapy.com/apa-itu-regresi-klasifikasi-dan-clustering-klasterisasi/>

<https://towardsdatascience.com/logistic-regression-explained-9ee73cede081>

<https://towardsdatascience.com/decision-trees-explained-3ec41632ceb6>

<https://towardsdatascience.com/boosting-algorithms-explained-d38f56ef3f30>

<https://machinelearningmastery.com/regression-metrics-for-machine-learning/>

<https://towardsdatascience.com/the-5-classification-evaluation-metrics-you-must-know-aa97784ff226>

<https://cloud.google.com/ai-platform/training/docs/hyperparameter-tuning-overview>



<https://www.jeremyjordan.me/hyperparameter-tuning/>