MACHINE LEARNING

MEMBANGUN MODEL -2

(Regresi Linier dan Non- Linier, Decision Tree, SVR, Random Forest Regression)

Magister Teknik Informatika

Dr. Chairani, S.Kom., M.Eng

IIB DARMAJAYA, 2023/2024

Subject

Kursus ini akan menjelaskan regresi dan bagaimana membangun model (regresi), yaitu:

a. menyiapkan parameter model,

b. menggunakan tools pemodelan,

selanjutnya menjelaskan algoritma dan menggunakan Regresi Linier dan Non-Linier, Decision Tree, SVR, Random Forest Regression dan performansi regresi dengan Python dan Scikit-learn.

Capaian Pembelajaran

Peserta dapat menjelaskan, menyiapkan, dan mengimplementasikan model regresi dengan algoritma:

A. Regresi Linier sederhana dan variabel jamak
B. Regresi Polinomial dan Non-linier
C. Support Vector Regression
D. Decision Tree Regression
E. Random Forest Regression

Beserta pengukuran performansinya menggunakan Python dan Scikit-learn.

Tujuan Pembelajaran

Peserta mempelajari pengertian, cara menyiapkan, dan cara implementasi model regresi dengan algoritma:

A. Regresi Linier sederhana dan variabel jamak
B. Regresi Polinomial dan Non-linier
C. Support Vector Regression
D. Decision Tree Regression
E. Random Forest Regression

Beserta pengukuran performansinya menggunakan Python dan Scikit-learn.

Referensi

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, pp.2825-2830.
- Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F. and Mueller, A., 2015. Scikit-learn: Machine learning without learning the machinery. *GetMobile: Mobile Computing and Communications*, *19*(1), pp.29-33.
- https://scikit-learn.org/stable/index.html

Regresi

Pengertian Regresi

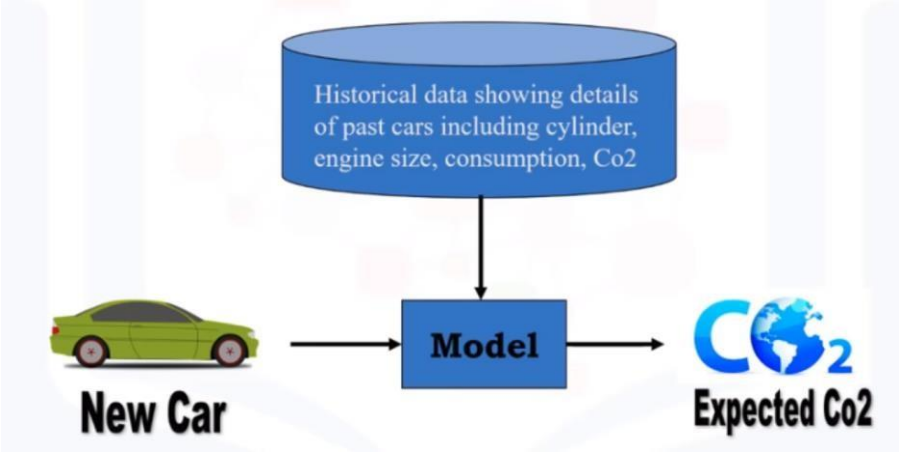
x : variabel bebas

y : variabel tak bebas

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Regresi adalah proses
Memprediksi nilai kontinu

Model Regresi



Tipe Model Regresi

Regresi Sederhana:

- Regresi sederhana linier
- Regresi sederhana non-linier
- Contoh: memprediksi co2emission vs EngineSize dari semua mobil.

Regresi Variabel Jamak:

- Regresi variabel jamak linier
- Regreasi variabel jamak non-linier
- Contoh: meprediksi co2emission vs EngineSize dan Cylinders dari semua mobil.

Aplikasi Regresi

- Prakiraan penjualan produk
- Analisis kepuasan
- Estimasi harga
- Pendapatan pekerjaan
- dst.

Algoritma Regresi

- Linier Regression
- Polynomial Regression
- Support Vector Regression
- Decision Tree Regression
- Random Forest Regression
- LASSO Regression
- ANN Regression
- K-NN Regression
- dst.

Regresi Linier Sederhana

Regresi Linier Untuk Memprediksi Nilai Kontinu

x : variabel bebas

y : variabel tak bebas

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Nilai kontinyu / numerik

Topologi Regresi Linier

Regresi Linier Sederhana:

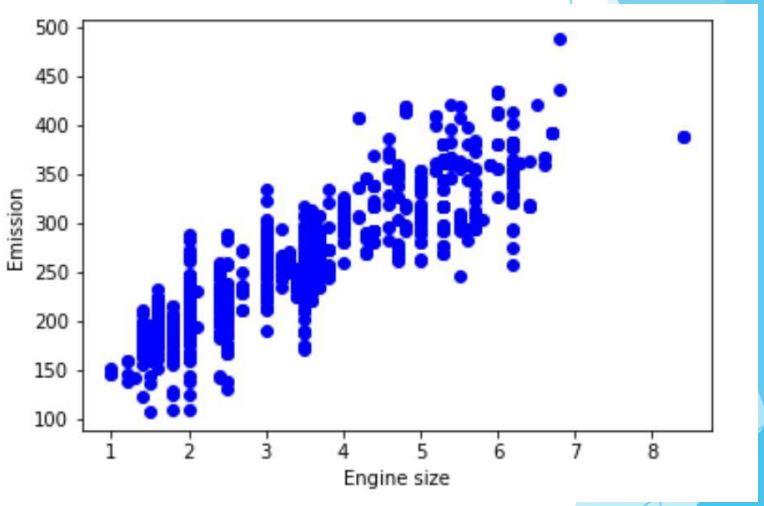
- Memprediksi co2emission vs EngineSize dari semua mobil
  - variabel bebas (x): EngineSize
  - variabel tak bebas (y): co2emission

Regresi Linier Variabel Jamak:

- Memprediksi co2emission vs EngineSize dan Cylinders dari semua mobil
  - variabel bebas (x): EngineSize, Cylinders, dst.
  - variabel tak bebas (y): co2emission

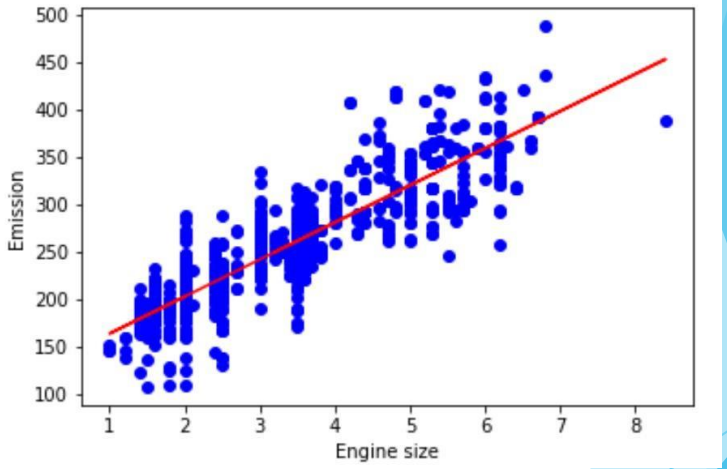
Cara Kerja Regresi Linier

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?



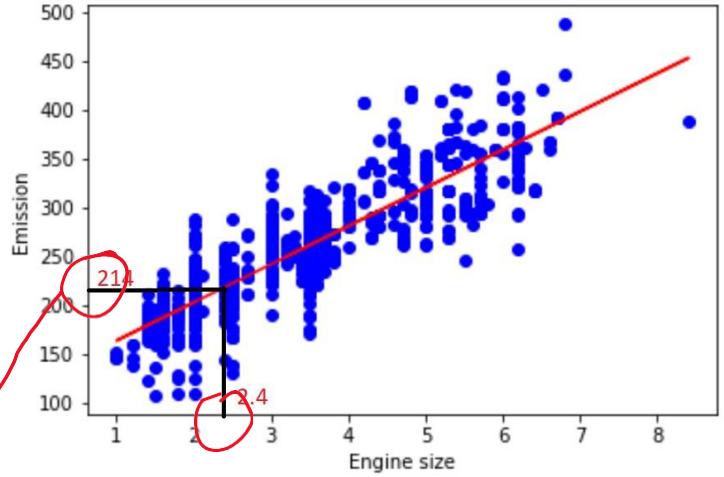
Cara Kerja Regresi Linier

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?



Cara Kerja Regresi Linier

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?



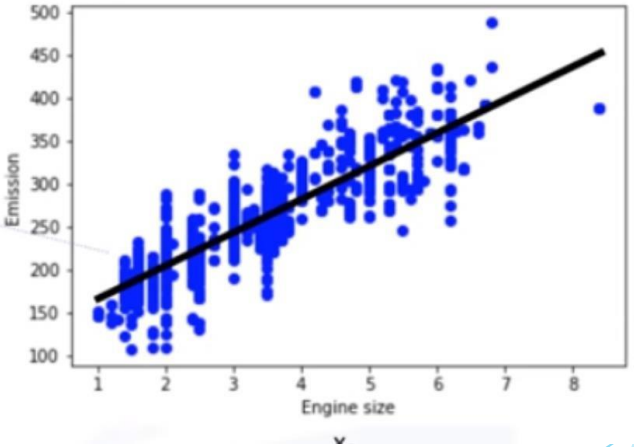
Cara Kerja Regresi Linier

$$\hat{y} = \theta_0 + \theta_1 x_1$$

response variable

a single predictor

y



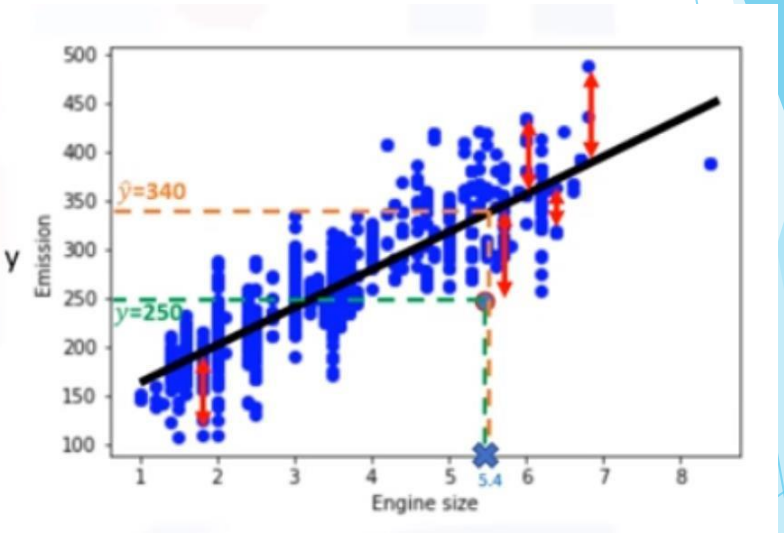
Cara Mencari Parameter Model Terbaik

x_1 = **5.4** independent variable
y= 250 actual Co2 emission of x1

$\hat{y} = \theta_0 + \theta_1 x_1$
$\hat{y}$ = 340 the predicted emission of x1

Error = y- $\hat{y}$
= 250 – 340
= -90

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



Estimasi Parameter

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

X_1 (ENGINESIZE) $\quad$ y (CO2EMISSIONS)

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$\theta_1 = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$$

$\bar{x}$ = (2.0 + 2.4 + 1.5 + ...)/9 = 3.34

$\bar{y}$ = (196 + 221 + 136 + ...)/9 = 256

$$\theta_1 = \frac{(2.0 - 3.34)(196 - 256) + (2.4 - 3.34)(221 - 256) + \dots}{(2.0 - 3.34)^2 + (2.4 - 3.34)^2 + \dots}$$

$\theta_1 = 39$

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

$$\theta_0 = 256 - 39 * 3.34$$

$$\theta_0 = 125.74$$

$$\hat{y} = 125.74 + 39 x_1$$

Prediksi dengan Model Regresi Linier

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$Co2Emission = \theta_0 + \theta_1 \, EngineSize$$

$$Co2Emission = 125 + 39 \, EngineSize$$

$$Co2Emission = 125 + 39 \times 2.4$$

$$Co2Emission = 218.6$$

Kelebihan Regresi Linier

- Ringan
- Tidak perlu tuning parameter
- Mudah dipahami dan diinterpretasikan

Regresi Linier Variabel Jamak

Contoh Regresi Linier Variabel Jamak

Efektivitas variabel-variabel bebas terhadap prediksi

- Apakah kegelisahan, kehadiran dosen, dan jenis kelamin mempunyai efek pada kinerja ujian mahasiswa?

Prediksi dampak perubahan

- Seberapa besar kenaikan/penurunan tekanan darah terhadap kenaikan/penurunan BMI dari pasien?

Prediksi Nilai Kontinu pada Regresi Linier Variabel Jamak

X: Independent variable

Y: Dependent variable

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

$$Co2\ Em = \theta_0 + \theta_1 Engine\ size + \theta_2 Cylinders + \ ...$$

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ ... + \theta_n x_n$$

$$\hat{y} = \theta^T X$$

$$\theta^T = [\theta_0, \theta_1, \theta_2, ...] \qquad X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ ... \end{bmatrix}$$

MSE Untuk Menunjukkan Error Pada Model

$\hat{y} = \theta^T X$

$\hat{y}_i = 140$ the predicted emission of x_i

$y_i = 196$ actual value of x_i

$y_i - \hat{y}_i = 196 - 140 = 56$ residual error

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

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

Estimasi Parameter Regresi Linier Variabel Jamak

Cara-cara mengestimasi parameter θ

Least Squares
- Operasi aljabar linier
- Perlu waktu yang lama untuk dataset yang besar (lebih dari 10000 baris)

Algoritma optimisasi
- Gradient Descent
- Metode yang sesuai apabila dataset sangat besar

Prediksi Menggunakan Regresi Linier Variabel Jamak

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

$$\hat{y} = \theta^T X$$

$\theta^T = [125, 6.2, 14, \dots]$

$$\hat{y} = 125 + 6.2x_1 + 14x_2 +$$

$$Co2Em = 125 + 6.2EngSize + 14\ \text{Cylinders} + \dots$$

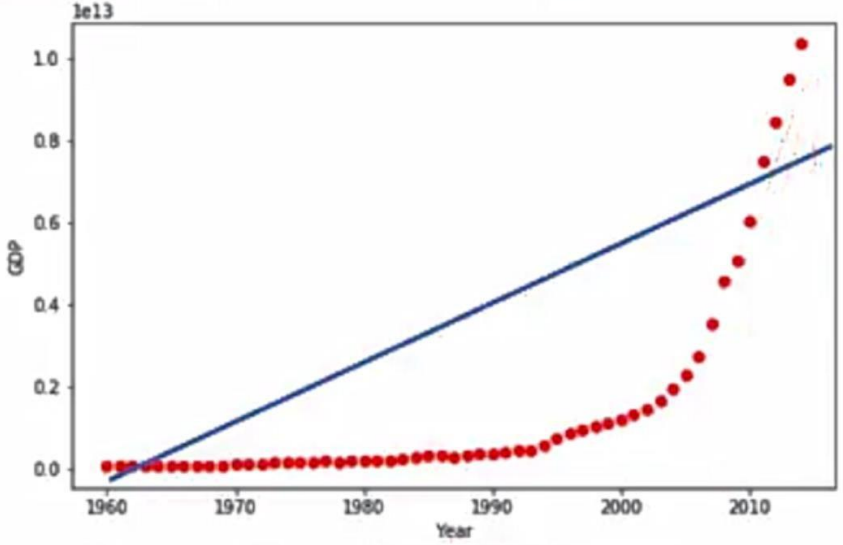
$$Co2Em = 125 + 6.2 \times 2.4 + 14 \times 4 + \dots$$

$$Co2Em = 214.1$$

Regresi Non Linier

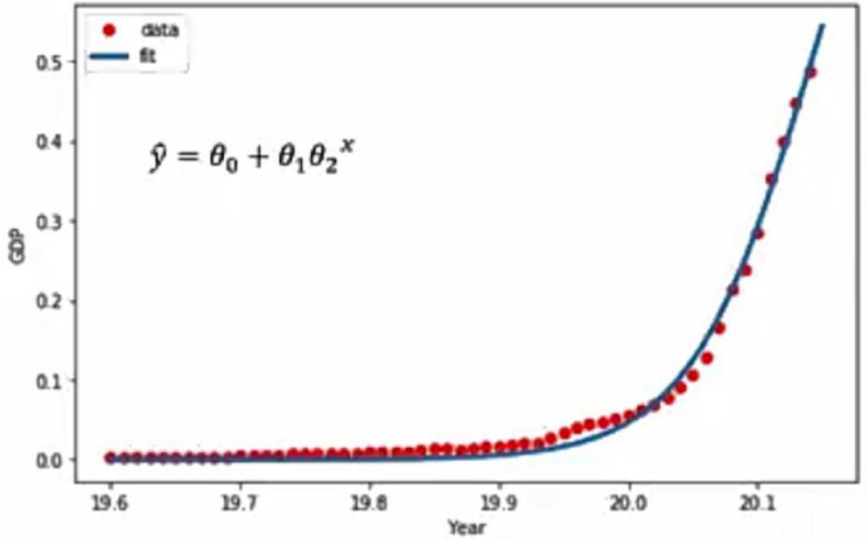
Mengapa Regresi Non-Linier Diperlukan?

	Year	Value
0	1960	5.918412e+10
1	1961	4.955705e+10
2	1962	4.668518e+10
3	1963	5.009730e+10
4	1964	5.906225e+10
5	1965	6.970915e+10
6	1966	7.587943e+10
7	1967	7.205703e+10
8	1968	6.999350e+10
9	1969	7.871882e+10
...	.....	.....

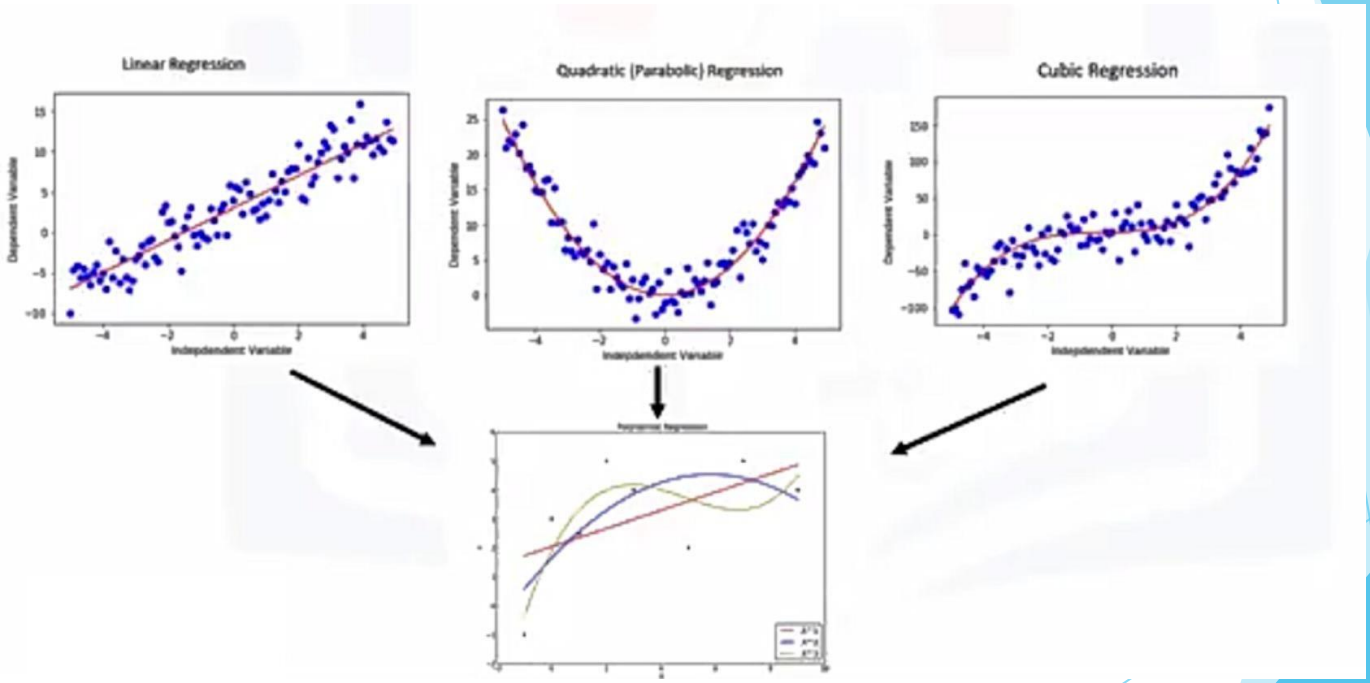


Mengapa Regresi Non-Linier Diperlukan?

	Year	Value
0	1960	5.918412e+10
1	1961	4.955705e+10
2	1962	4.668518e+10
3	1963	5.009730e+10
4	1964	5.906225e+10
5	1965	6.970915e+10
6	1966	7.587943e+10
7	1967	7.205703e+10
8	1968	6.999350e+10
9	1969	7.871882e+10
...	.....	.....



Tipe Regresi



Regresi Polinomial

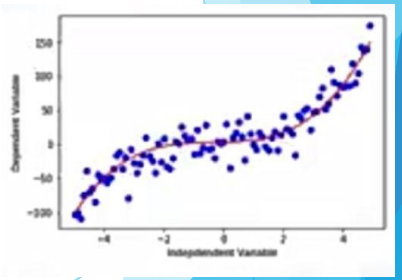
- Beberapa data yang berbentuk kurva dapat dimodelkan dengan regresi linier
- Contoh:

$$\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

- Model regresi polinomial dapat ditransformasikan menjadi model regresi linier.

$$x_1 = x$$
$$x_2 = x^2$$
$$x_3 = x^3$$

$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$ -> dapat diselesaikan dengan least squares Regresi linier variabel jamak



Regresi Non-Linier

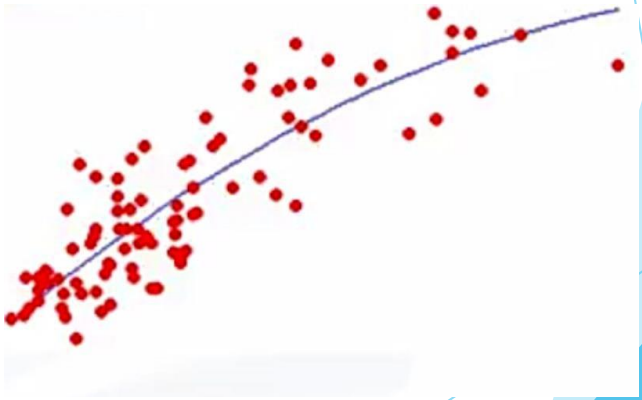
- Memodelkan hubungan tidak linier antara variabel tak bebas dengan himpunan variabel bebas
- $\hat{y}$ berupa fungsi non-linier dari parameter θ dan fitur x.

$$\hat{y} = \theta_0 + \theta_2{}^2 x$$

$$\hat{y} = \theta_0 + \theta_1 \theta_2{}^x$$

$$\hat{y} = \log(\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3)$$

$$\hat{y} = \frac{\theta_0}{1 + \theta_1{}^{(x-\theta_2)}}$$



Regresi Linier atau Non-Linier?

Cara untuk mengetahui apakah permasalahan cocok diselesaikan dengan regresi linier atau non linier

- Pengamatan visual atas data (visualisasi)
- Pengamatan akurasi hasil pemodelan

Cara untuk memodelkan data apabila visualisasi mengindikasikan non-linier

- Regresi polinomial
- Regresi non-linier
- Transformasi data non-linier menjadi linier

Support Vector Regression

Support Vector Regression

- SVR memberi fleksibilitas untuk menentukan seberapa besar kesalahan yang dapat diterima dalam model dan akan menemukan garis yang sesuai (atau hyperplane dalam dimensi yang lebih tinggi) agar sesuai dengan data.
- Berbeda dengan Least Square biasa, fungsi tujuan SVR adalah untuk meminimalkan koefisien — lebih khusus lagi, *l2*-norm vektor koefisien — bukan *squared error*.

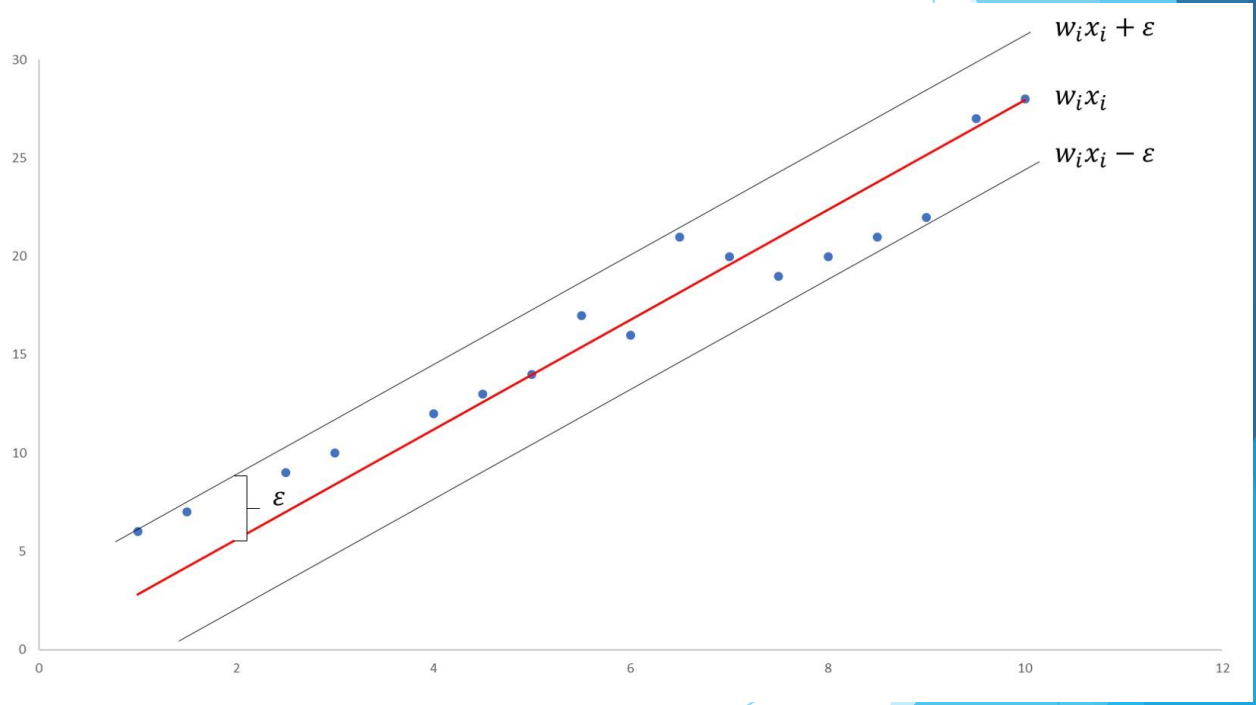
Support Vector Regression

Minimize

$$MIN\ \frac{1}{2}||\boldsymbol{w}||^2$$

Constraint

$$|y_i - w_i x_i| \leq \varepsilon$$



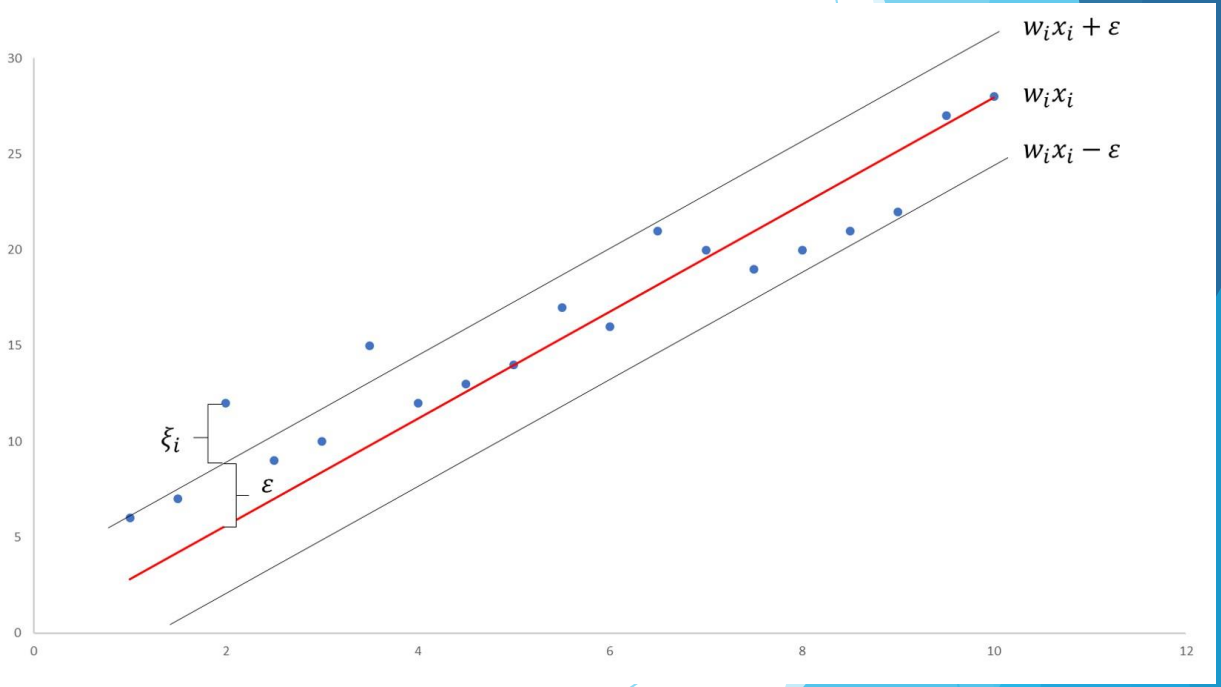
Support Vector Regression with Slack Variable

Minimize

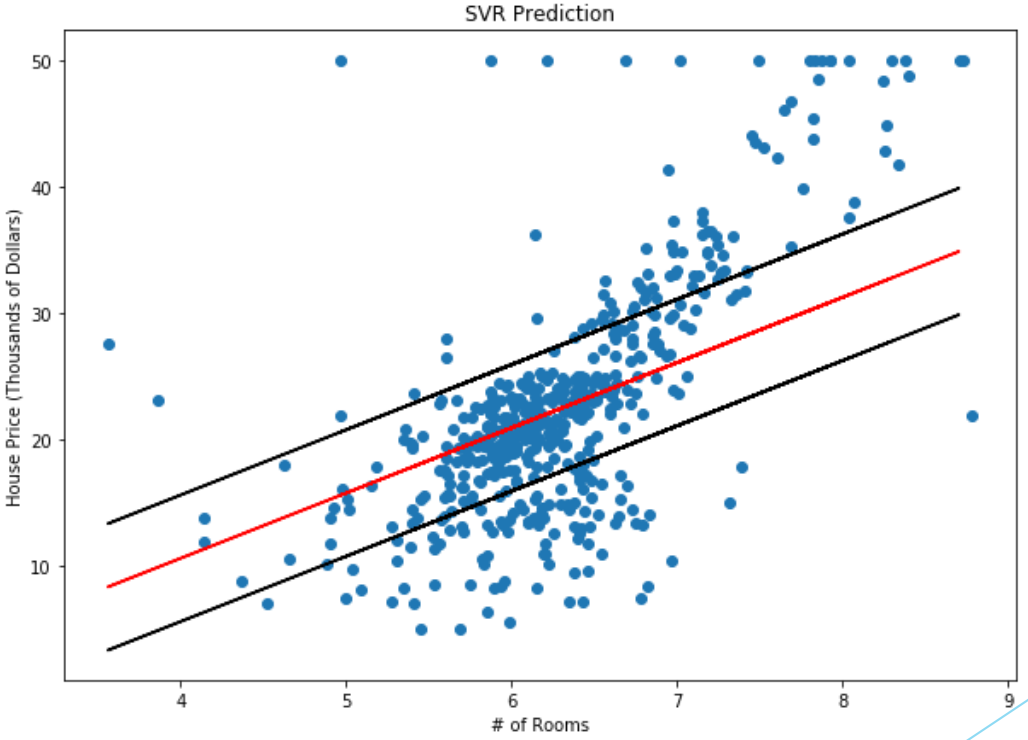
$$MIN\ \frac{1}{2}\left|\left|\boldsymbol{w}\right|\right|^2 + C\sum_{i=1}^{n}|\xi_i|$$

Constraint

$$|y_i - w_i x_i| \leq \varepsilon + |\xi_i|$$

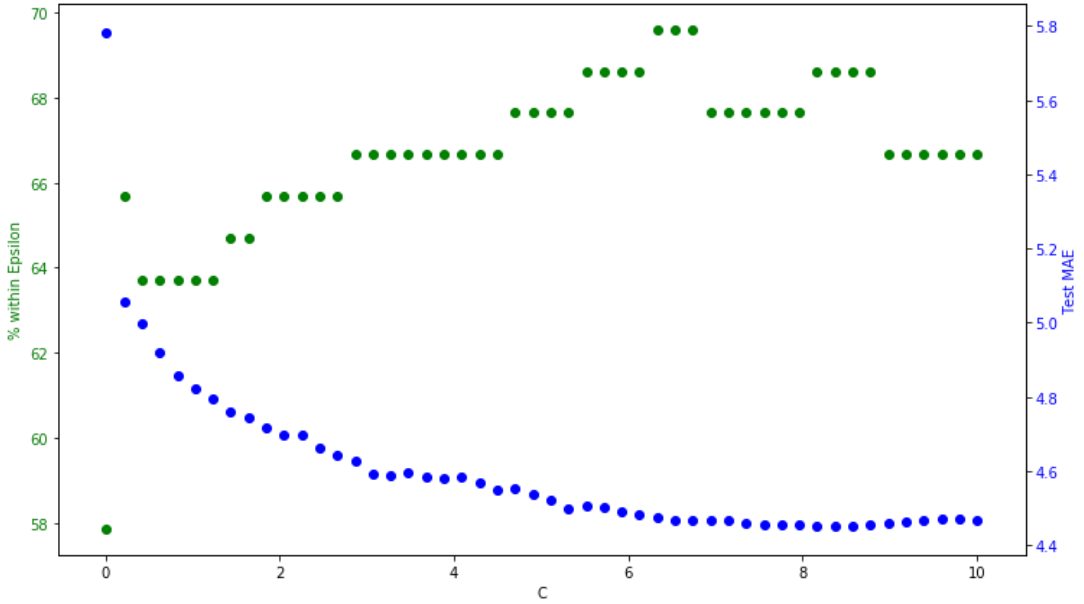


Efek C



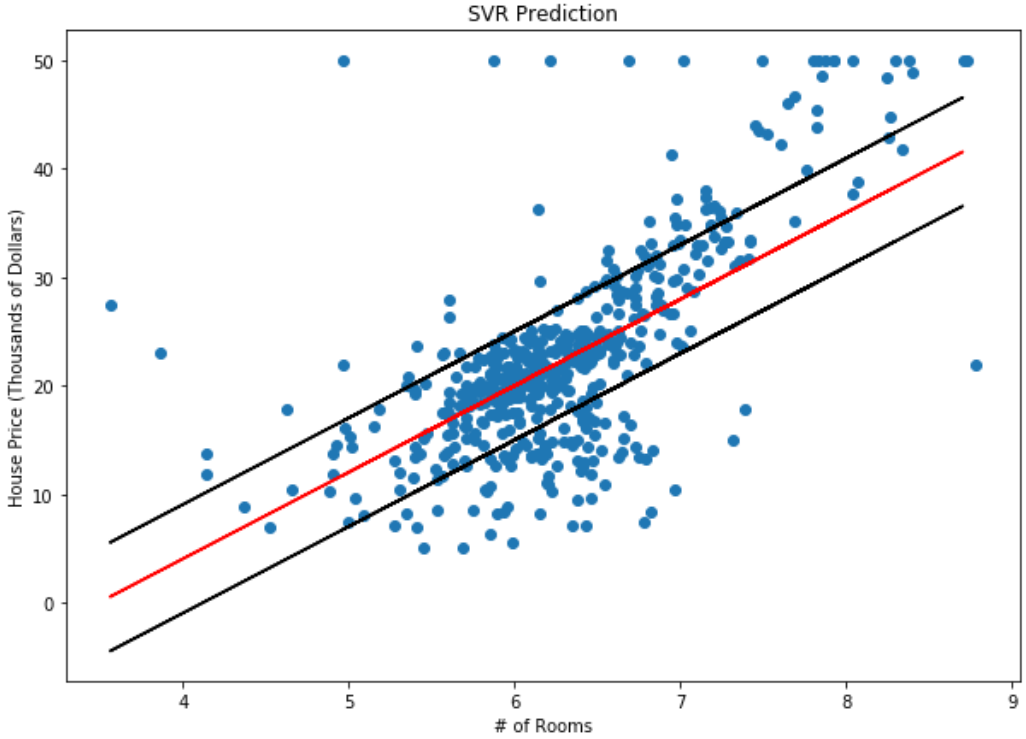
SVR Prediction of Boston Housing Prices with ϵ=5, C=1.0

Mencari C Terbaik



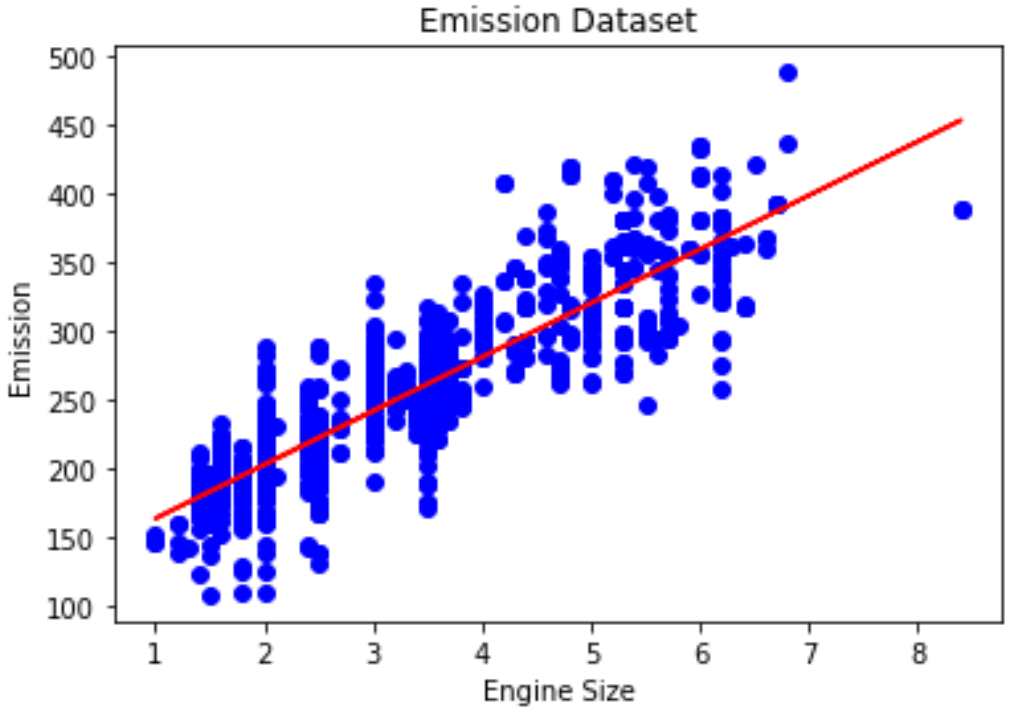
GridSearch for C

Efek C



SVR Prediction of Boston Housing Prices with ϵ=5, C=6.13

Aplikasi SVR pada Regresi Sederhana Emission Dataset

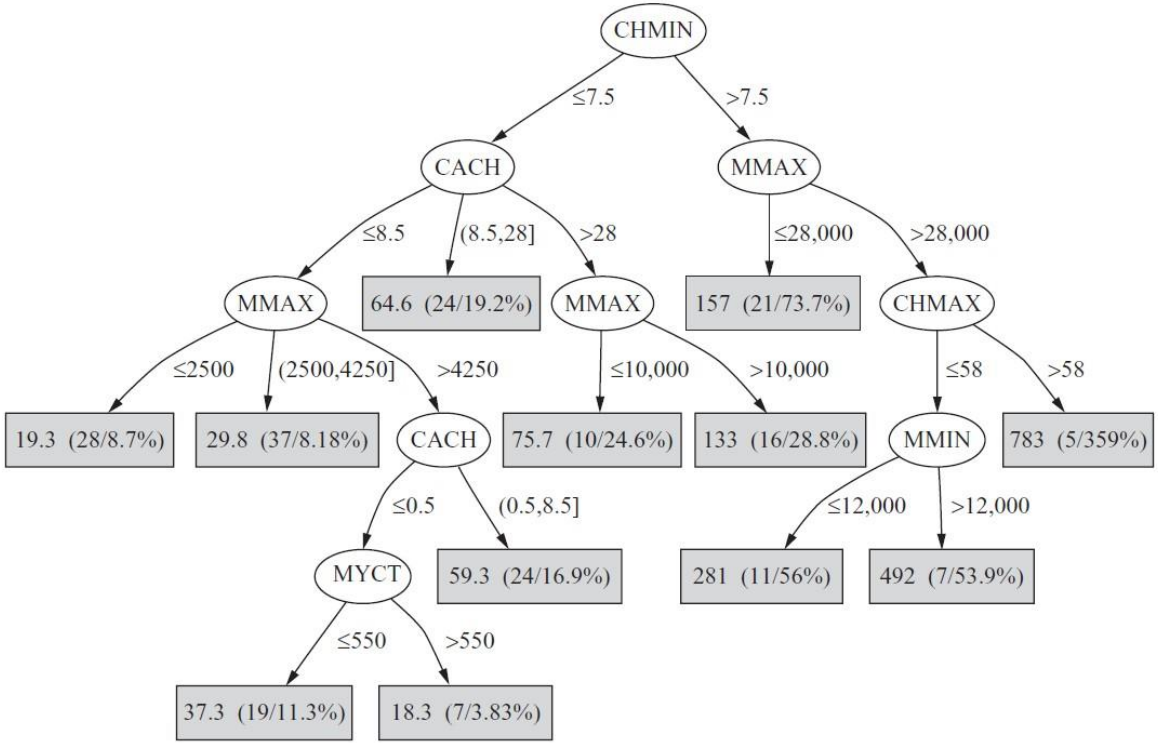


Decision Tree Regression

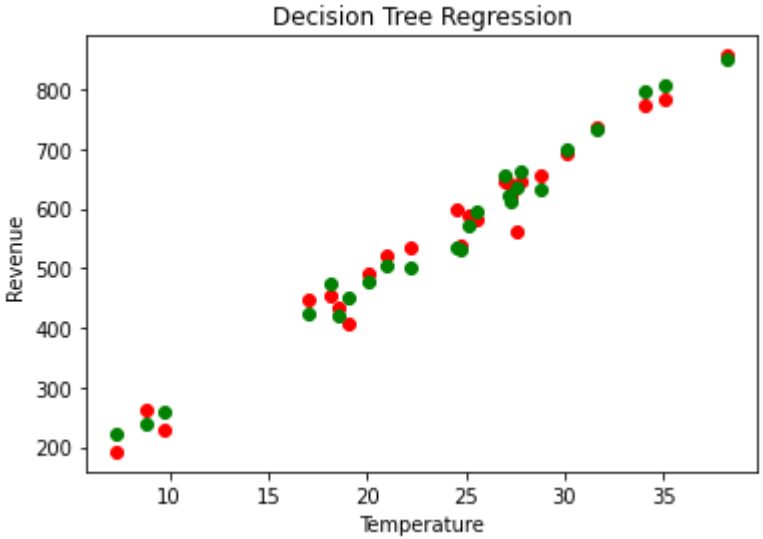
Decision Tree Regression

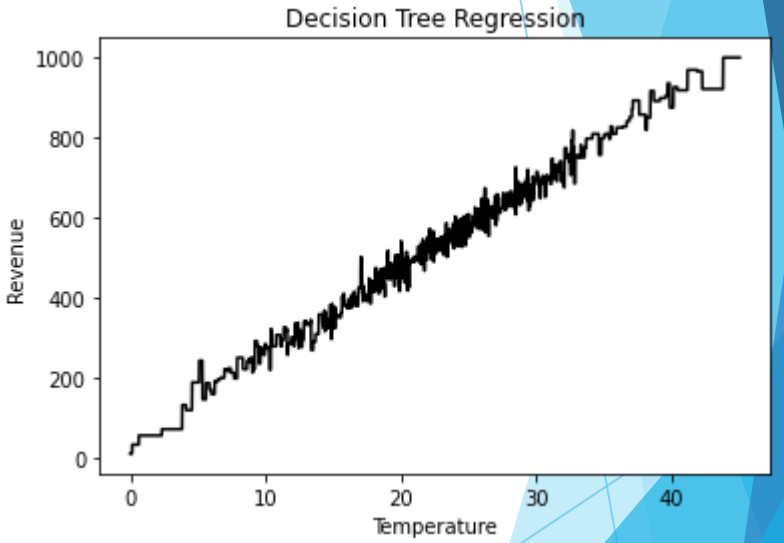
- Decision Tree Regression (DTR) membangun model regresi dalam bentuk struktur pohon. DTR memecah dataset menjadi subset yang lebih kecil dan lebih kecil sementara pada saat yang sama pohon keputusan terkait dikembangkan secara bertahap. Hasil akhirnya adalah pohon dengan simpul keputusan dan simpul daun.
- Dengan titik data tertentu, DTR dijalankan sepenuhnya melalui seluruh pohon dengan menjawab pertanyaan Benar/Salah hingga mencapai simpul daun
- Prediksi terakhir adalah rata-rata dari nilai variabel dependen dalam simpul daun tertentu. Melalui beberapa iterasi, Pohon mampu memprediksi nilai yang tepat untuk titik data.

Decision Tree Regression

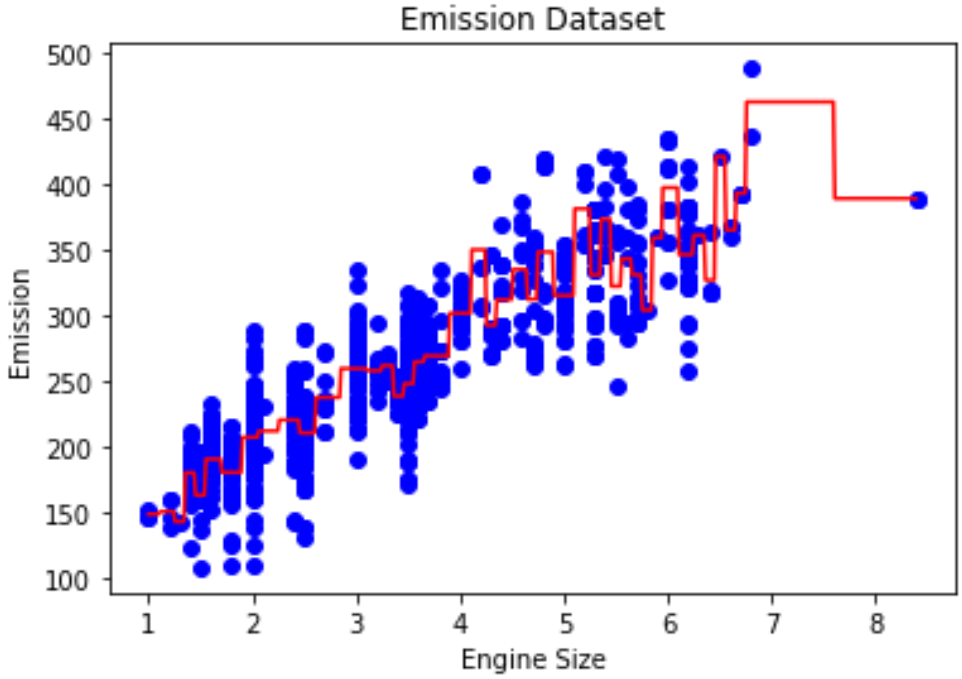


Decision Tree Regression - Contoh

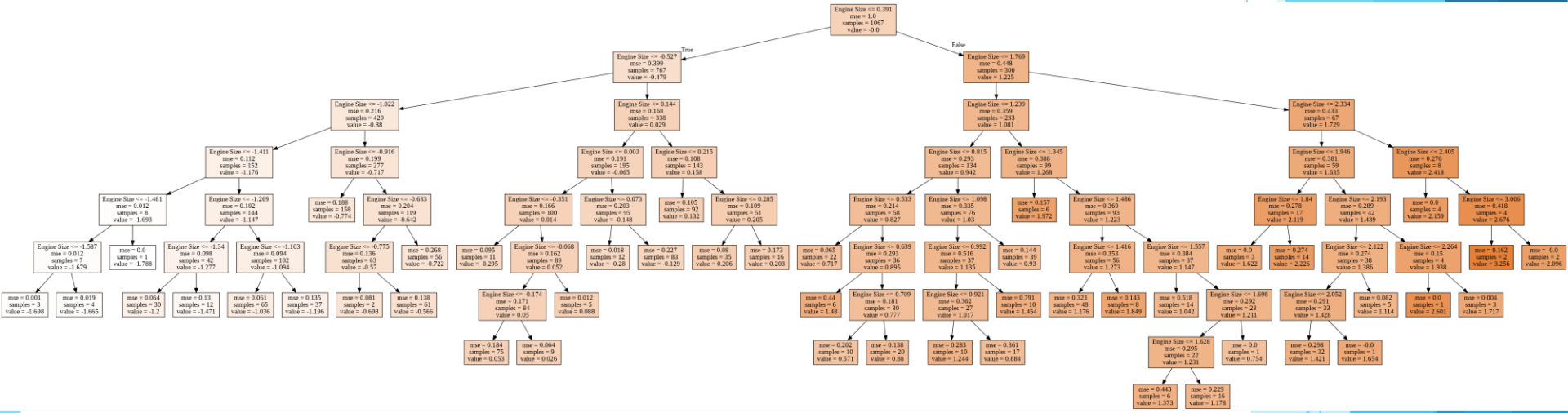




Aplikasi DTR pada Regresi Sederhana Emission Dataset



Visualisasi Struktur Tree DTR – Emission Dataset

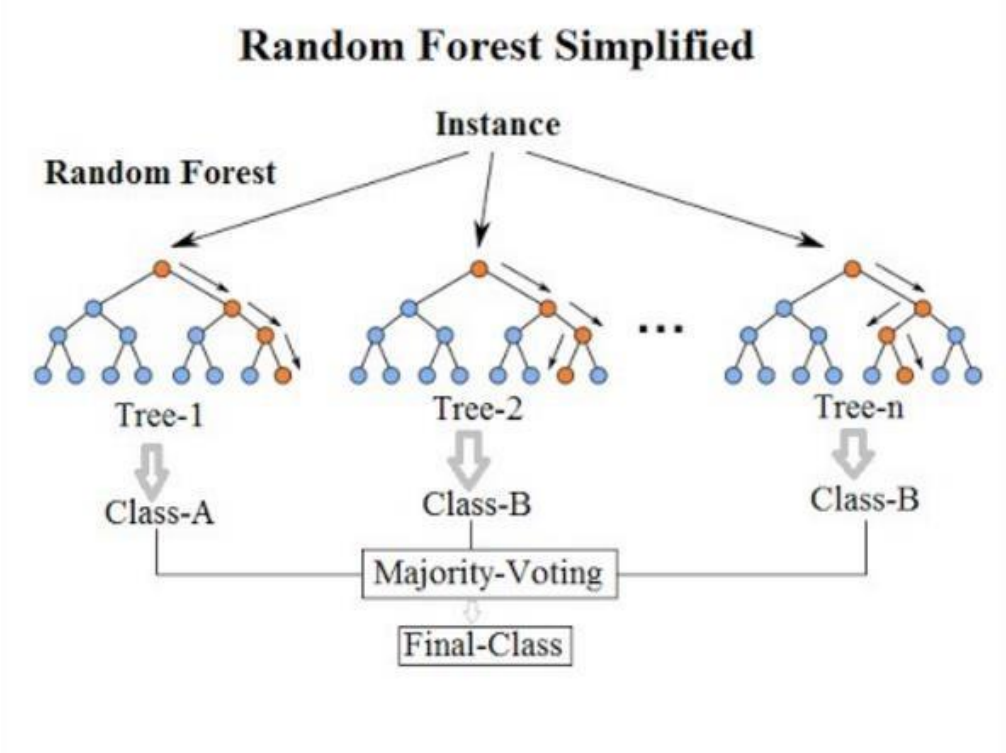


Random Forest Regression

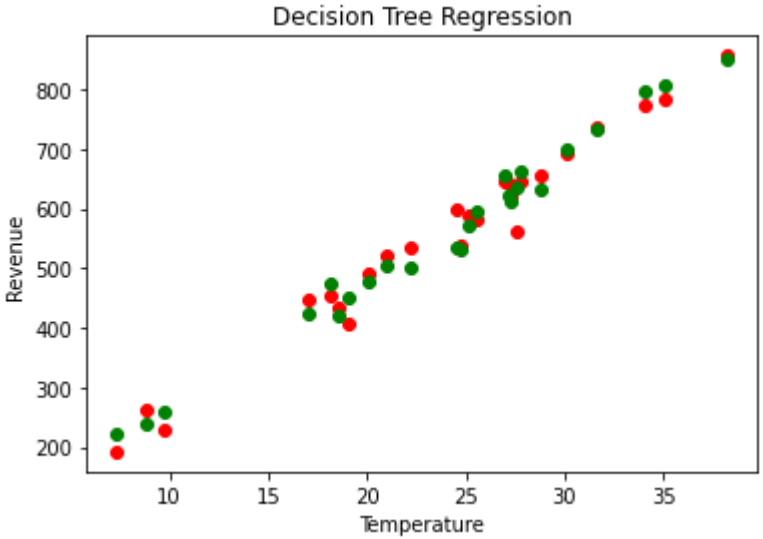
Random Forest Regression (RFR)

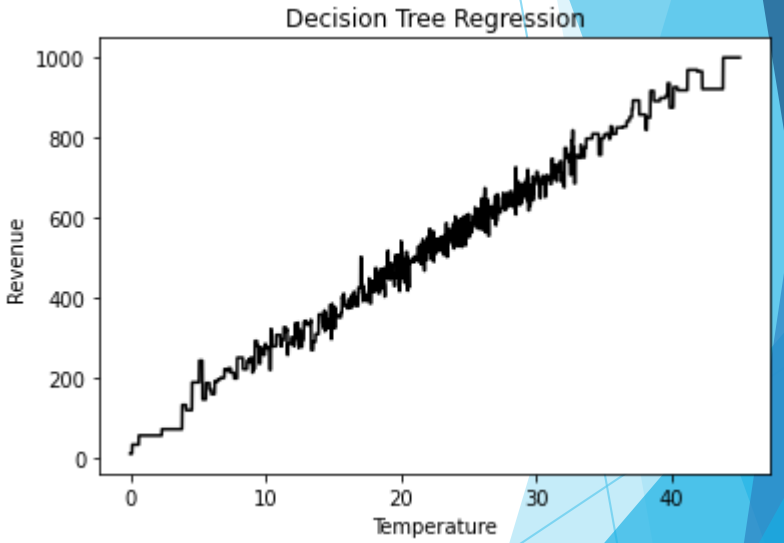
- Pohon Keputusan (Decision Tree) adalah algoritma yang mudah dipahami dan diinterpretasikan dan karenanya satu pohon mungkin tidak cukup bagi model untuk mempelajari fitur-fiturnya. Di sisi lain, Random Forest juga merupakan algoritma berbasis “Pohon” yang menggunakan fitur kualitas dari beberapa Pohon Keputusan untuk membuat keputusan.
- Oleh karena itu, dapat disebut sebagai ‘Forest’ atau ‘Hutan’ dari pohon-pohon dan karenanya disebut “Random Forest". Istilah ‘Random’ atau ‘Acak' disebabkan oleh fakta bahwa algoritma ini adalah hutan dari 'Pohon Keputusan atau Decision Tree yang dibuat secara acak atatu random'.
- Algoritma Decision Tree memiliki kelemahan utama yaitu menyebabkan over-fitting. Masalah ini dapat diatasi dengan menerapkan Regresi Random Forest (Random Forest Regression) sebagai pengganti DTR. Selain itu, algoritma Random Forest juga sangat cepat dan kuat dibandingkan model regresi lainnya.

Random Forest

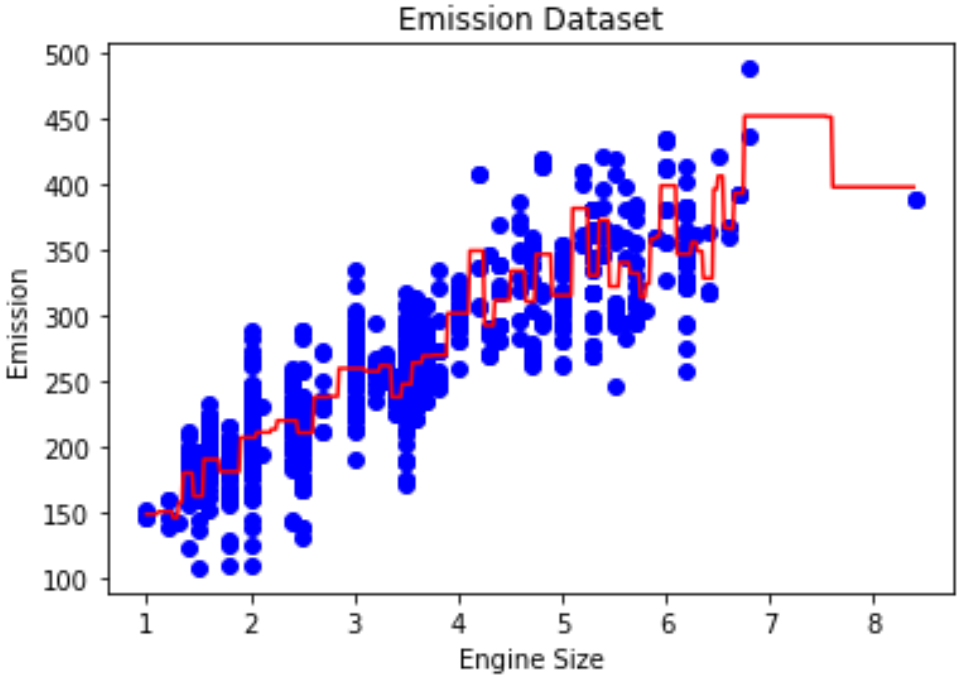


Random Forest Regression - Contoh

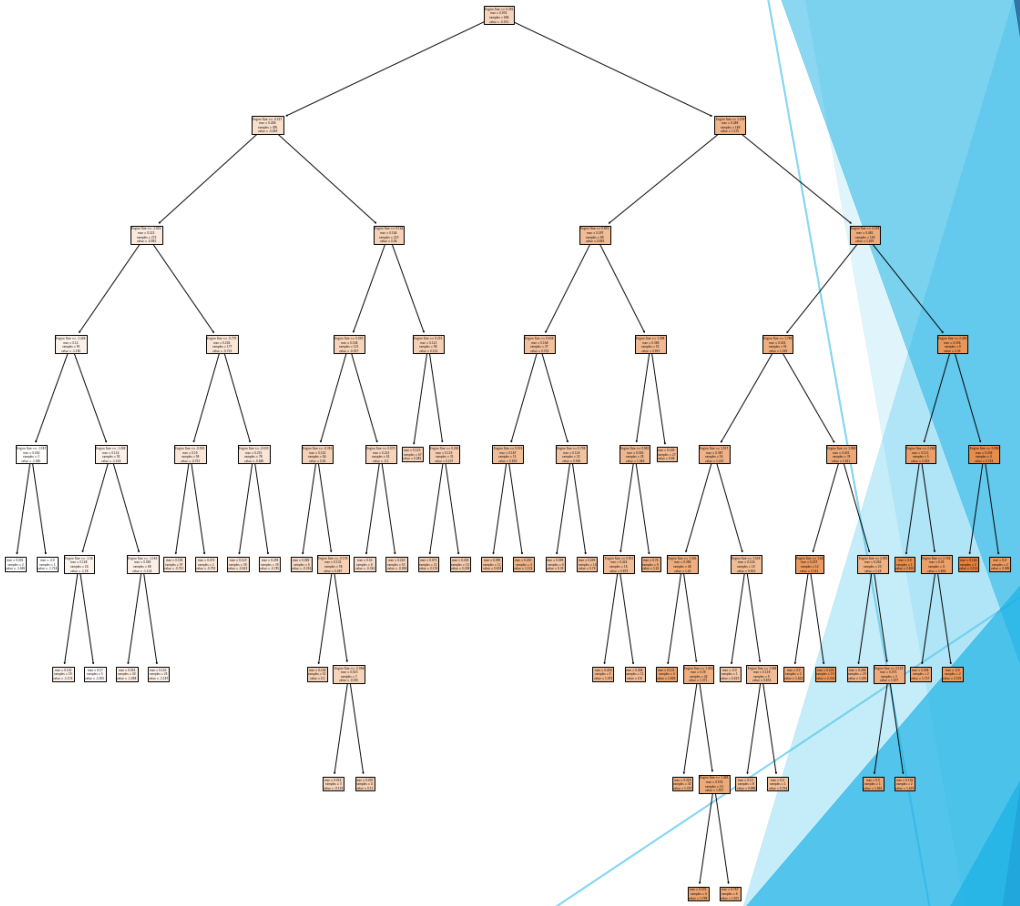




Aplikasi RFR pada Regresi Sederhana Emission Dataset



Visualisasi Struktur Tree RFR – Emission Dataset

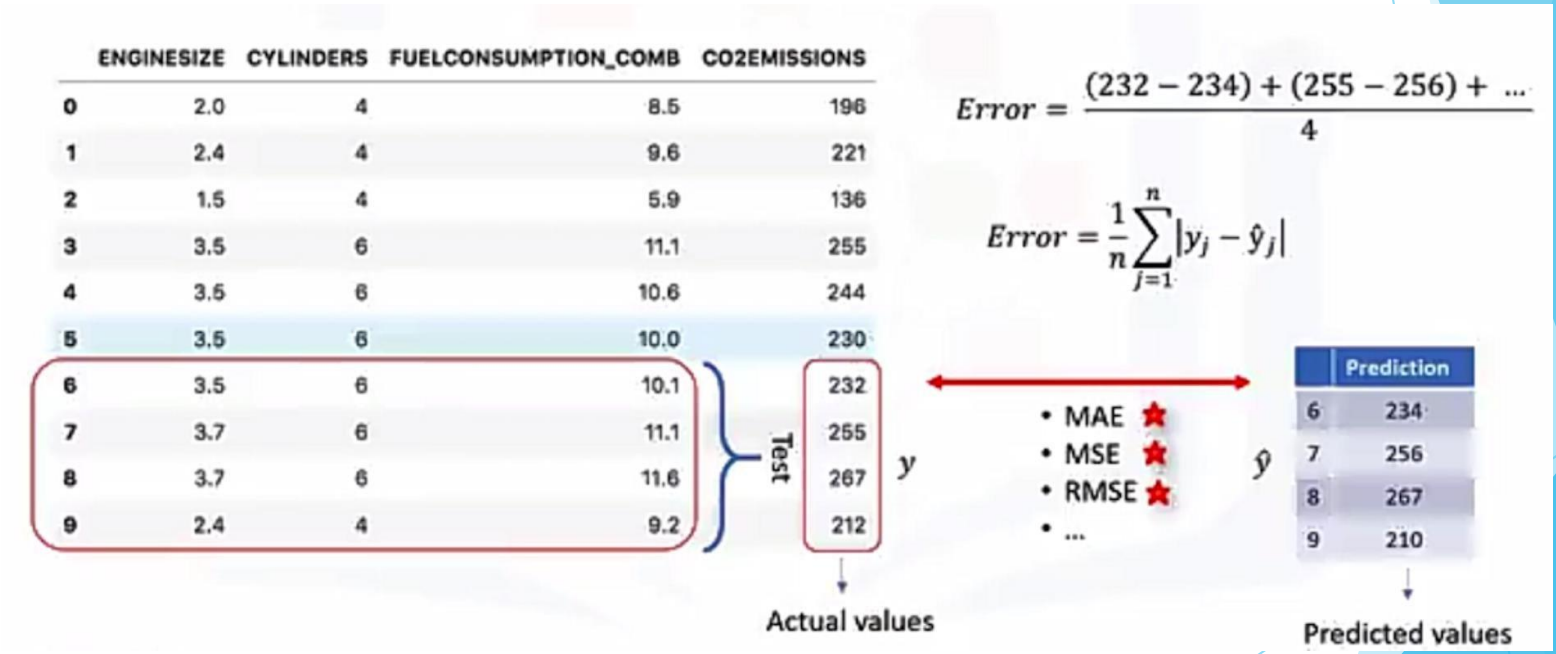


Lab

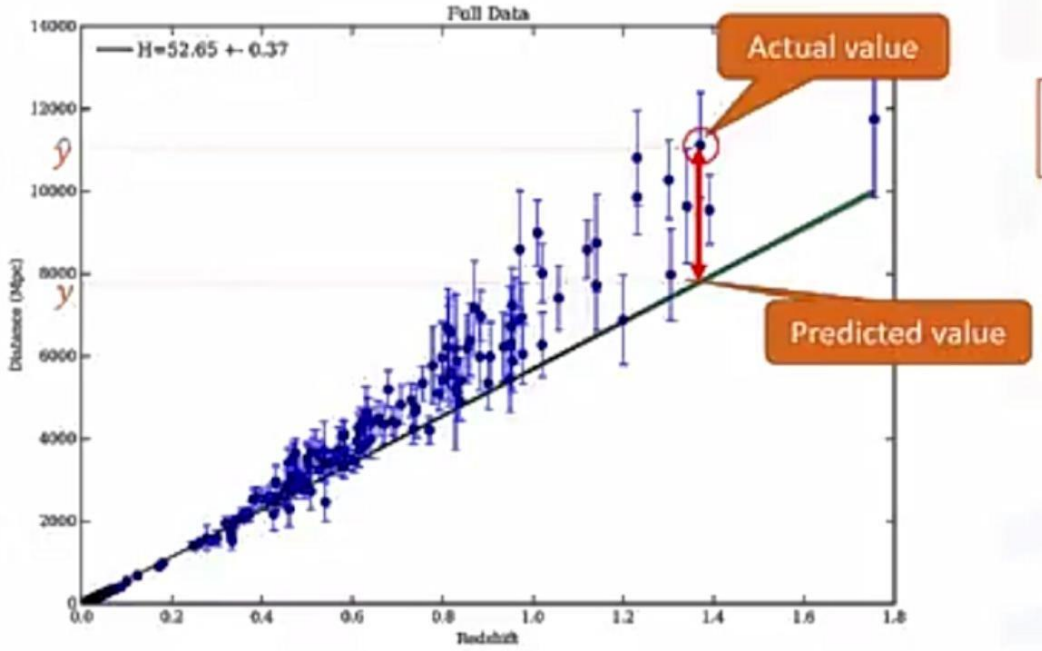
- Jalankan file Jupyter Notebook untuk Random Forest Regression

Metrik Evaluasi

Error

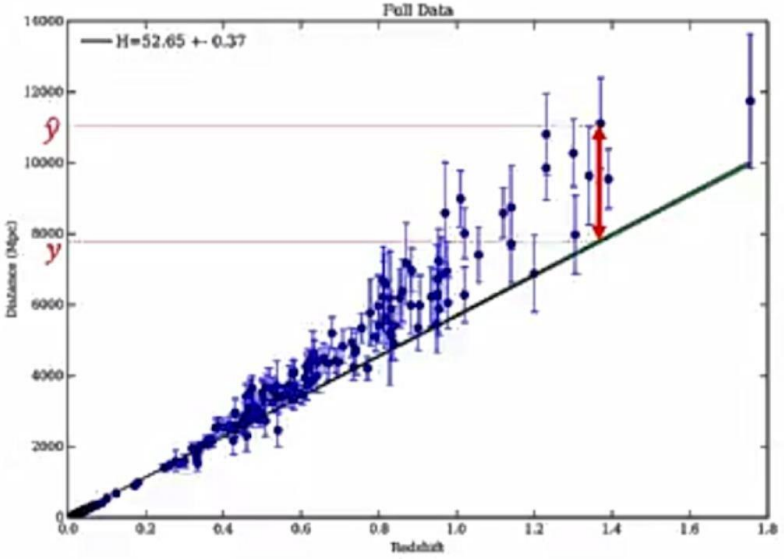

ENGINESIZE CYLINDERS FUELCONSUMPTION_COMB CO2EMISSIONS
0 2.0 4 8.5 196
1 2.4 4 9.6 221
2 1.5 4 5.9 136
3 3.5 6 11.1 255
4 3.5 6 10.6 244
5 3.5 6 10.0 230
6 3.5 6 10.1 232
7 3.7 6 11.1 255
8 3.7 6 11.6 267
9 2.4 4 9.2 212
Test
Actual values
y
$Error = \frac{(232 - 234) + (255 - 256) + \ldots}{4}$
$Error = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$
• MAE
• MSE
• RMSE
• ...
ŷ
Prediction
6 234
7 256
8 267
9 210
Predicted values

Error dari Model



Error: measure of how far the data is from the fitted regression line.

Error dari Model



$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

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

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

$$RAE = \frac{\sum_{j=1}^{n}|y_j - \hat{y}_j|}{\sum_{j=1}^{n}|y_j - \bar{y}|}$$

$$RSE = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

$$R^2 = 1 - RSE$$

Perbandingan

Regresi Linier (LR) vs DTR

- DTR mendukung non linearitas, di mana RL hanya mendukung solusi linier.
- Ketika ada sejumlah besar fitur dengan lebih sedikit kumpulan data (dengan noise rendah), regresi linier dapat mengungguli DTR/Random Forest Regression (RFR). Dalam kasus umum, DTR akan memiliki akurasi rata-rata yang lebih baik.
- Untuk variabel bebas kategorikal, DTR lebih baik daripada regresi linier.
- DTR menangani kolinearitas lebih baik daripada LR.

RL vs SVR

- SVR mendukung solusi linier dan non-linier menggunakan trik kernel.
- SVR menangani outlier lebih baik daripada RL.
- Keduanya berkinerja baik ketika data pelatihan lebih sedikit, dan ada banyak fitur.

DTR vs RFR

- RFR adalah kumpulan DT, suara (vote) mayoritas atau rata-rata dari forest dipilih sebagai keluaran yang diprediksi.
- RFR akan kurang rentan terhadap overfitting daripada DTR, dan memberikan solusi yang lebih general.
- RFR lebih robust dan akurat daripada pohon keputusan.

Tools / Lab Online

- Scikit-Learn
- Jupyter Notebook
- Conda / Anaconda
- Google Colaboratory
- PyPI (pip)

Summary

- Regresi adalah prediksi nilai kontinyu atau numeris dari variabel tak bebas berdasarkan variabel bebas atau predictor.
- Regresi ada dua tipe sederhana atau satu variabel dan variabel jamak.
- Masing-masing regresi tersebut terdapat dua pendekatan terhadap data: linier dan non-linier.
- Evaluasi pemodelan regresi menggunakan metrik berdasar error seperti MAE, MSE, maupun kecocokan model terhadap data seperti R^2
- Terdapat banyak algoritma regresi yang dapat digunakan dengan kelebihan dan kekurangan masing-masing.

TERIMA KASIH