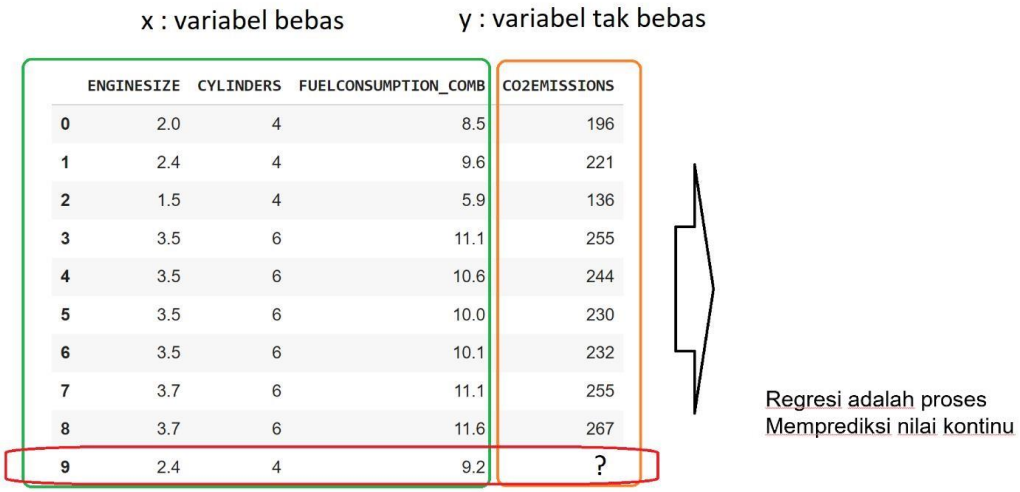
MATERI REGRESI

A. Pengertian Regresi

Regresi adalah proses Memprediksi nilai kontinu, sebagaimana contoh dalam Gambar 1.

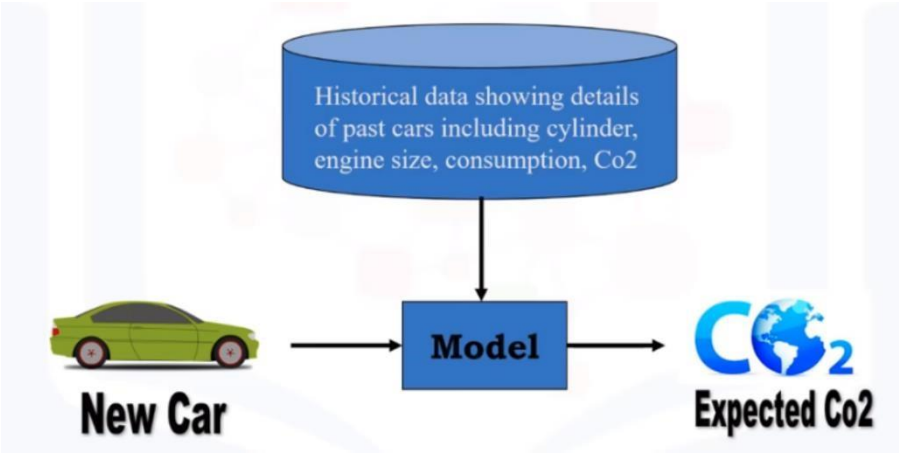


	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Gambar 1. Tabel Dataset Emisi C02

Regresi akan memprediksi nilai kontinyu CO2EMISSIONS sebagai variable tak bebas berdasarkan nilai-nilai dari variable bebas (ENGINESIZE, CYLINDERS, dst.).

B. Model Regresi



Gambar 2. Ilustrasi Penggunaan Model Regresi

Model regresi adalah hasil dari proses penentuan parameter regresi yang dapat dengan akurat memprediksi variabel tak bebas berdasar variabel bebas. Dalam contah pada Gambar 2, model yang dihasilkan digunakan untuk meprediksi emisi CO2 dari mobil baru yang belum pernah diketahui sebelumnya berdasaran variabel-variabel bebas antara lain cylinder, engine size, fuel consumption, dst.

C. Tipe-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.

D. Aplikasi Regresi

Hasil model regresi dapat digunakan dan diaplikasikan dalam berbagai bidang antara lain:

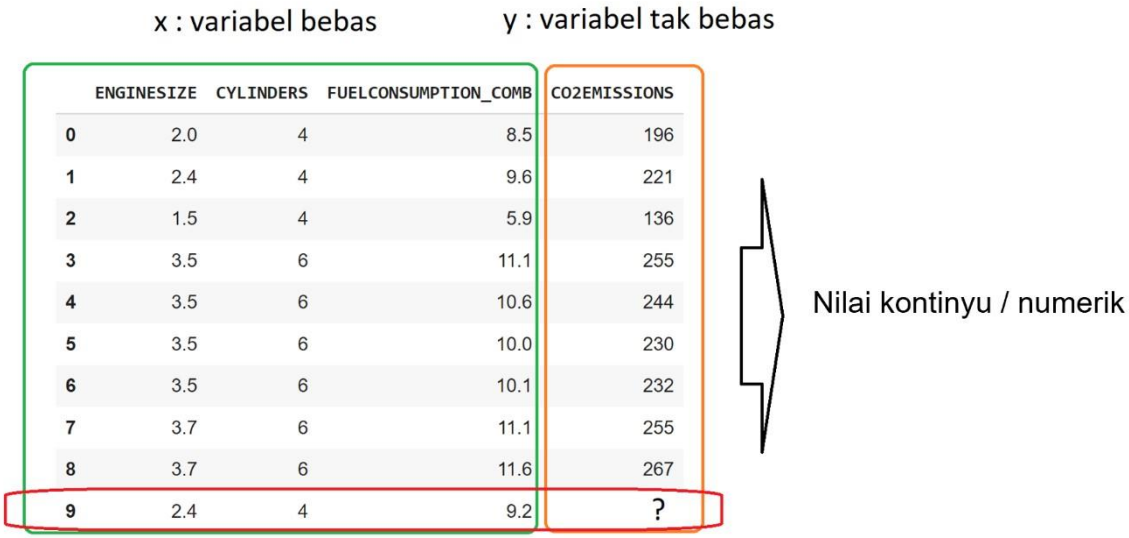
- Prakiraan penjualan produk
- Analisis kepuasan
- Estimasi harga • Pendapatan pekerjaan
- dst.

E. Algoritma-algoritma Regresi

Banyak algoritma yang telah dikembangkan untuk melakukan proses regresi dari sisi model dan penentuan parameternya antara lain:

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

F. Regresi Linier Sederhana 1. Regresi Linier untuk Memprediksi Nilai Kontinyu



	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	?

Gambar 3. Dataset Emisi Co2

Model regresi akan ditentukan parameternya dengan data satu variabel bebas untuk memprediksi variabel tak bebas, sebagai contoh memprediksi nilai kontinyu CO2EMISSIONS dengan variabel ENGINESIZE berdasar data pembelajaran (No 0 sd No 8). Hasil pemodelan dapat digunakan memprediksi nilai numerik CO2EMISSIONS kasus baru yang belum pernah dihadapi yakni kasus No 9 dengan dasar ENGINESIZE.

2. Topologi Regresi Linier

Regresi Linier Sederhana:

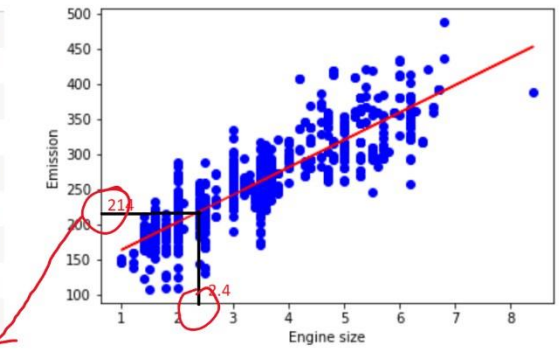
- Memprediksi co2emission vs EngineSize dari semua mobil
    a. variabel bebas (x): EngineSize
    b. variabel tak bebas (y): co2emission

Regresi Linier Variabel Jamak:

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

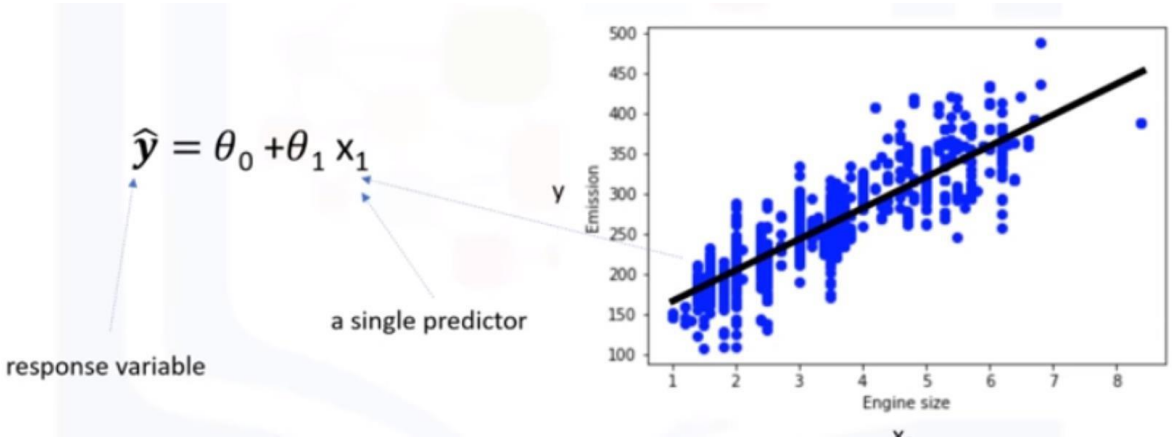
3. Cara Kerja Regresi Linier

	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	?

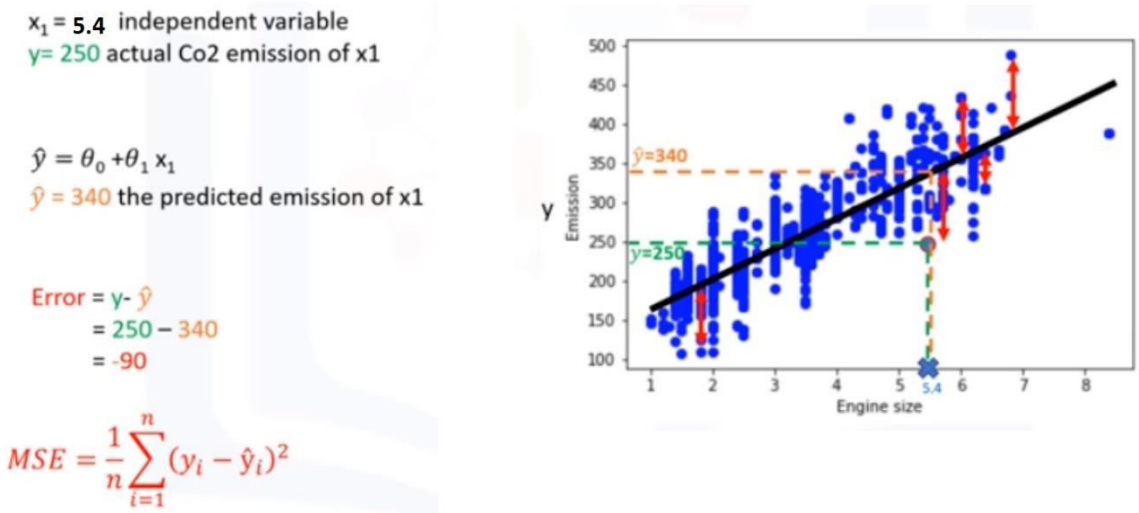


Gambar 4. Cara Kerja Regresi Linier

Garis merah pada Gambar 4 adalah model linier yang dihasilkan oleh algoritme regresi linier. Berdasarkan garis merah itu dapat dikethaui nilai CO2EMISSIONS sebagai variabel tak bebas dengan melihat nilai pada ENGINESIZE, dalam hal ini bernilai = 2,4 yang menghasilkan nilai CO2EMISSIONS = 214. Parameter regresi dari model persamaan linier yaitu θ_0 dan θ_1 sebagaimana dalam Gambar 5.



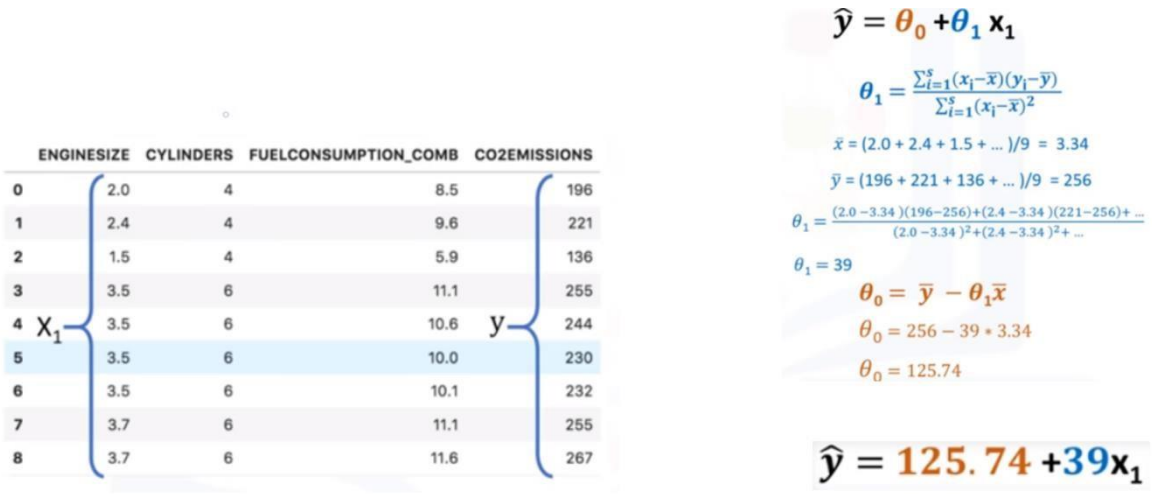
Gambar 5. Model linier dan parameternya



Gambar 6. Metrik Mean Squared Error untuk Mencari Parameter Terbaik (MSE)

Untuk mendapatkan parameter model regresi terbaik maka dicari parameter yang membuat selisih terkecil antara prediksi dengan nilai aktual yang disebut error. Dalam hal ini metrik yang paling sering digunakan adalah Mean Squared Error (MSE) sebagaimana ditunjukkan dalam Gambar 6 untuk menghindari saling menegasikan antara error positif dan error negatif.

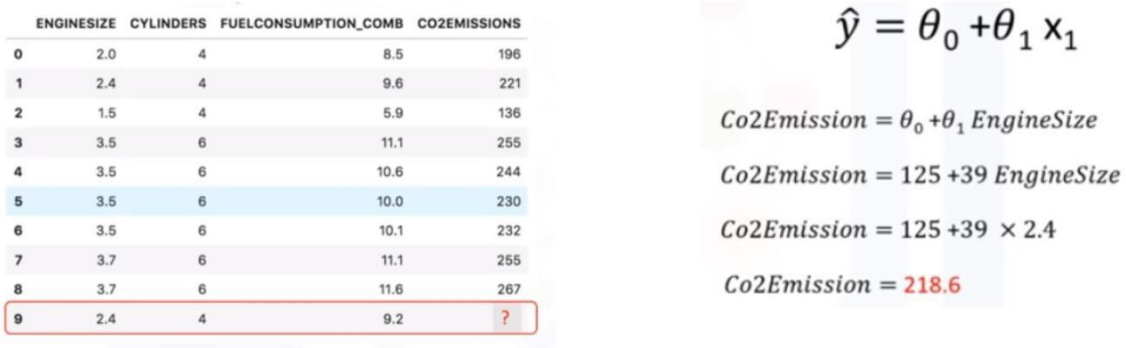
4. Estimasi Parameter



Gambar 7. Estimasi Parameter Regresi Linier

Paramater yang terbaik dicari dengan menggunakna metode Least Square seperti pada Gambar 7 sehingga menghasilkan parameter terbaik dilihat dari MSE.

5. Prediksi dengan Model Regresi Linier



Gambar 8. Prediksi dengan Model Regresi Linier

Prediksi nilai kontinyu variabel tak bebas dilakukan dengan memasukkan nilai variabel bebas ke dalam model yang sudah ditemukan. Dalam contoh Gambar 8, variabel EngineSize = 2,4 memberikan hasil Co2Emission = 218,6.

6. Kelebihan Regresi Linier

- Ringan – komputasi sederhana
- Tidak perlu tuning parameter – parameter langsung dapat dihitung
- Mudah dipahami dan diinterpretasikan – pengaruh variabel tak bebas tampak jelas

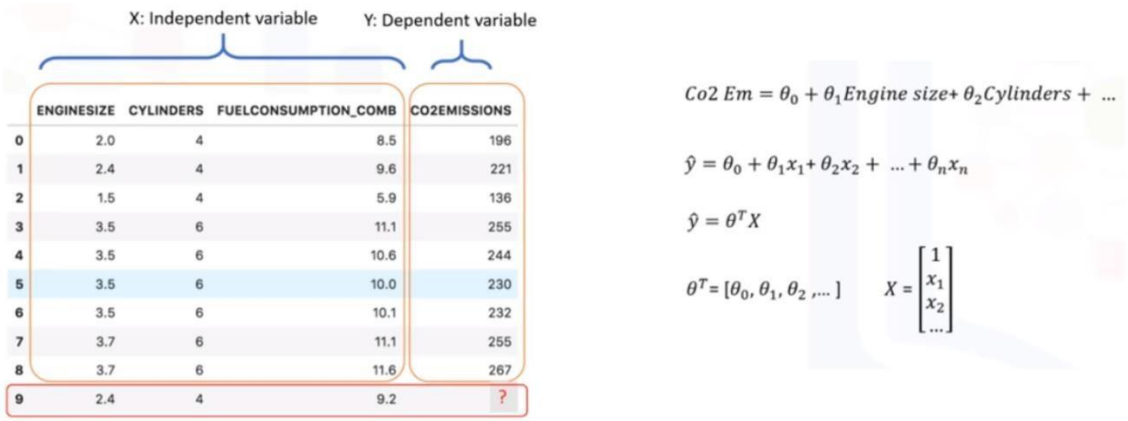
Lab: silakan jalankan notebook Regresi Linear Sederhana.ipynb

G. Regresi Linier Variabel Jamak 1. 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?

2. Prediksi Nilai Kontinyu pada Regresi Linier Variabel Jamak



Gambar 9. Prediksi pada Regresi Linier Variabel Jamak

Regresi linier variabel jamak menggunakan lebih dari satu variabel bebas antara lain ENGINESIZE, CYLINDERS, FUELCONSUMPTION_COMB untuk memprediksi nilai kontinyu variabel tak bebas dalam hal ini CO2EMISSION sebagaimana ditunjukkan pada Gambar 9.

3. Mean Squared Error (MSE) Sebagai Metrik Kesalahan Pada Model

	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

Gambar 10. MSE pada Regresi Linier Variabel Jamak

Regresi linier variabel jamak menggunakan MSE sebagai metrik kesalahan atau selisih antara hasil prediksi dengan nilai aktual variabel tak bebas sebagaimana ditunjukkan pada Gambar 10.

4. Estimasi Parameter Regresi Linier Variabel Jamak

Cara-cara mengestimasi parameter θ yaitu dengan mengguanakan metode:

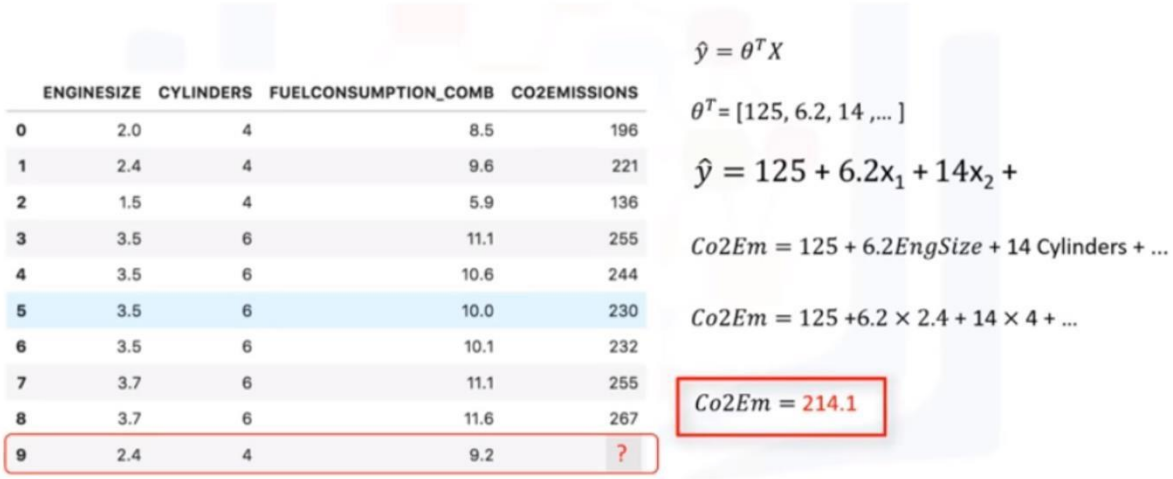
- Least Squares dengan aspek-aspek:
    a. Memanfaatkan operasi aljabar linier
    b. Memerlukan waktu yang lama untuk dataset yang besar

  (lebih dari 10000 baris)
- Algoritma optimisasi dengan aspek-aspek:

a. Menggunakan metode optimisasi berdasarkan Gradient Descent

b. Metode ini sesuai apabila dataset sangat besar

5. 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	?

Gambar 11. Contoh Prediksi Menggunakan Regresi Linier Variabel Jamak

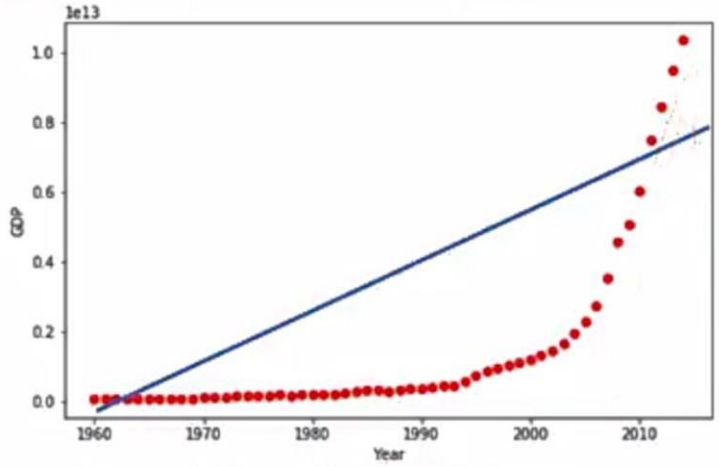
Gambar 11 menunjukkan bagaimana prediksi nilai numerik CO2EMISSIONS berdasarkan variabel bebas jamak yaitu ENGINESIZE = 2,4 serta CYLINDERS = 4 dan FUELCONSUMPTION = 9,2 dengan hasil prediksi = 214,1 menggunakna parameter terbaik yang sudah didapatkan dari data latih.

Lab : silakan jalankan notebook Regresi Linear Variabel Jamak.ipynb

H. Regresi Non-linier

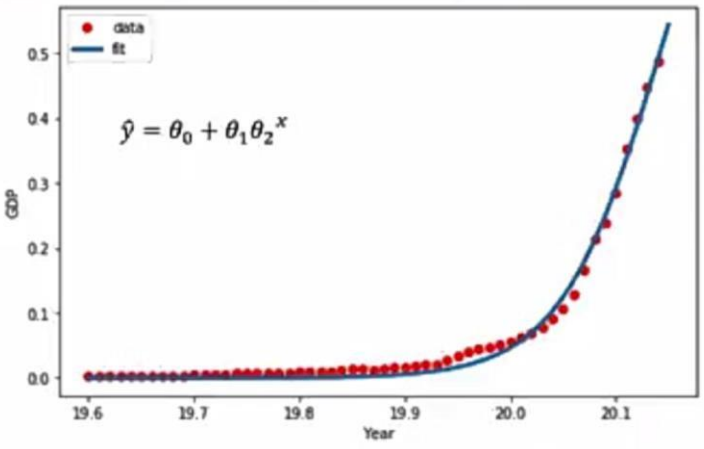
1. Mengapa Regresi Non-linier Diperlukan?

	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
...	.....	.....



Gambar 12. Pemodelan Linier Terhadap Data Non-linier

	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
...	.....	.....

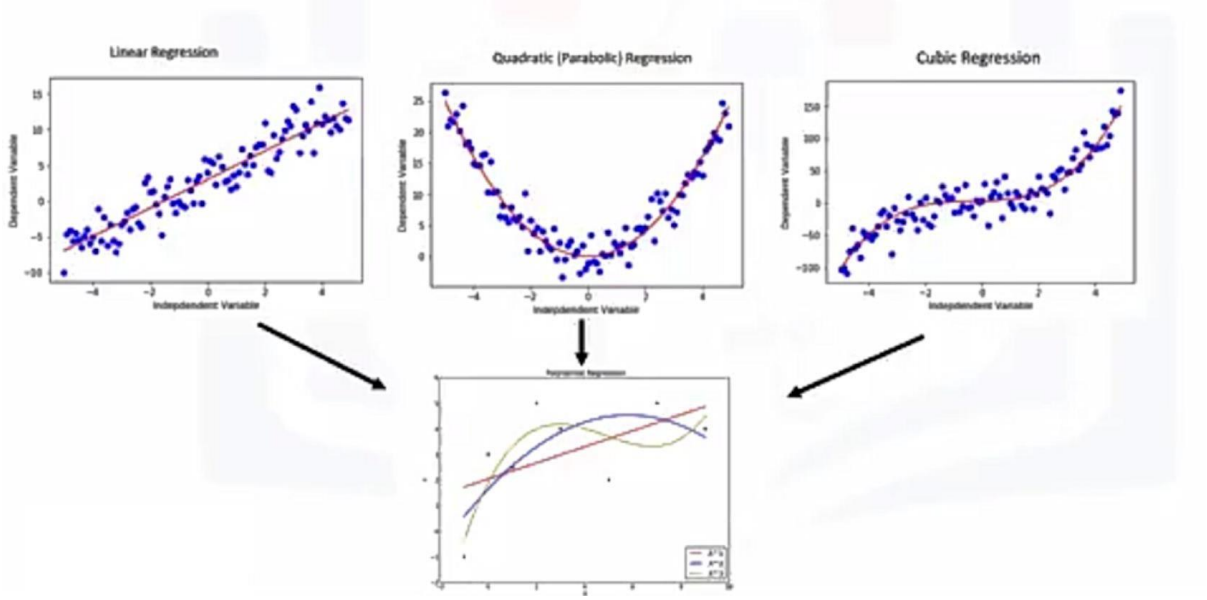


Gambar 13. Pemodelan Non-linier Terhadap Data Non-linier

Regresi non-linear diperlukan karena tidak setiap data menunjukkan hubungan linier sehingga error akan besar ketika dipaksakan menggunakan model linier sebagaimana

Gambar 12. Penggunaan model non-linier tampak mempunyai error yang lebih kecil sebagaimana Gambar 13.

2. Tipe-tipe Regresi



Gambar 14. Tipe-tipe Regresi

Tipe-tipe regresi antara lain regresi linier, regresi quadratic atau parabolic, regresi cubic.

3. 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

4. Regresi Non-linier

Regresi Non-linier adalah pemodelan hubungan tidak linier antara variabel tak bebas dengan himpunan variabel bebas ŷ berupa fungsi non-linier dari parameter θ dan fitur x.

$$\hat{y} = \theta_0 + {\theta_2}^2 \mathrm{x}$$
$$\hat{y} = \theta_0 + \theta_1 {\theta_2}^x$$
$$\hat{y} = \log(\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3)$$
$$\hat{y} = \frac{\theta_0}{1 + {\theta_1}^{(x-\theta_2)}}$$

5. Regresi Linier atau Non-Linier?

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

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

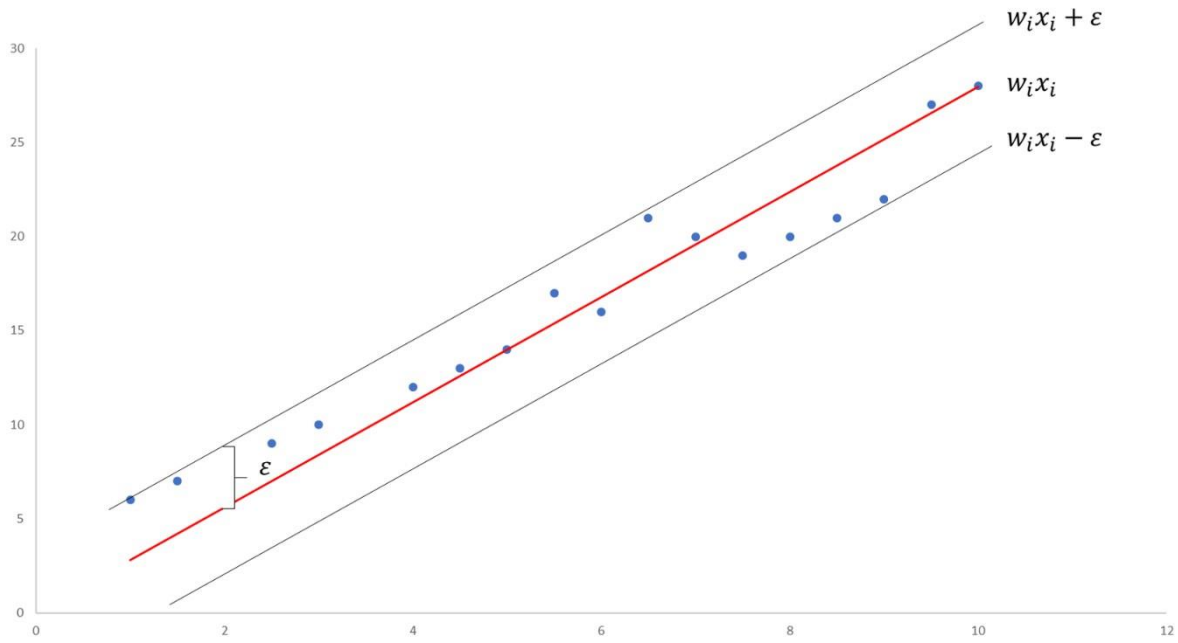
Cara untuk memodelkan data apabila visualisasi mengindikasikan nonlinier

- Regresi polynomial
- Regresi non-linier
- Transformasi data non-linier menjadi linier

Lab : silakan jalankan notebook Regresi Nonlinear.ipynb

6. 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.



Gambar 15. Support Vector Regression

Konsep SVR adalah dengan meminimalkan fungsi objektif yakni koefisien dengan konstrain margin sebagaimana tampak secara visual pada Gambar 15.

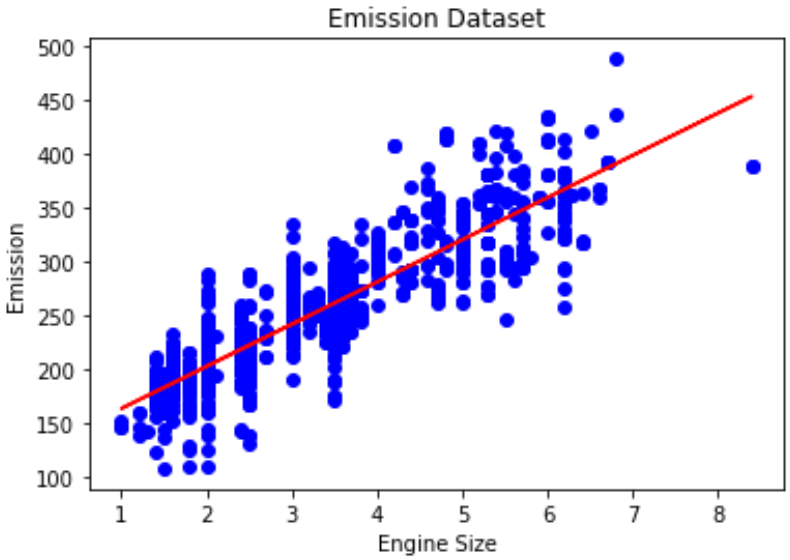
Minimize

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

Constraint

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

7. Aplikasi SVR pada Regresi Sederhana Emission Dataset



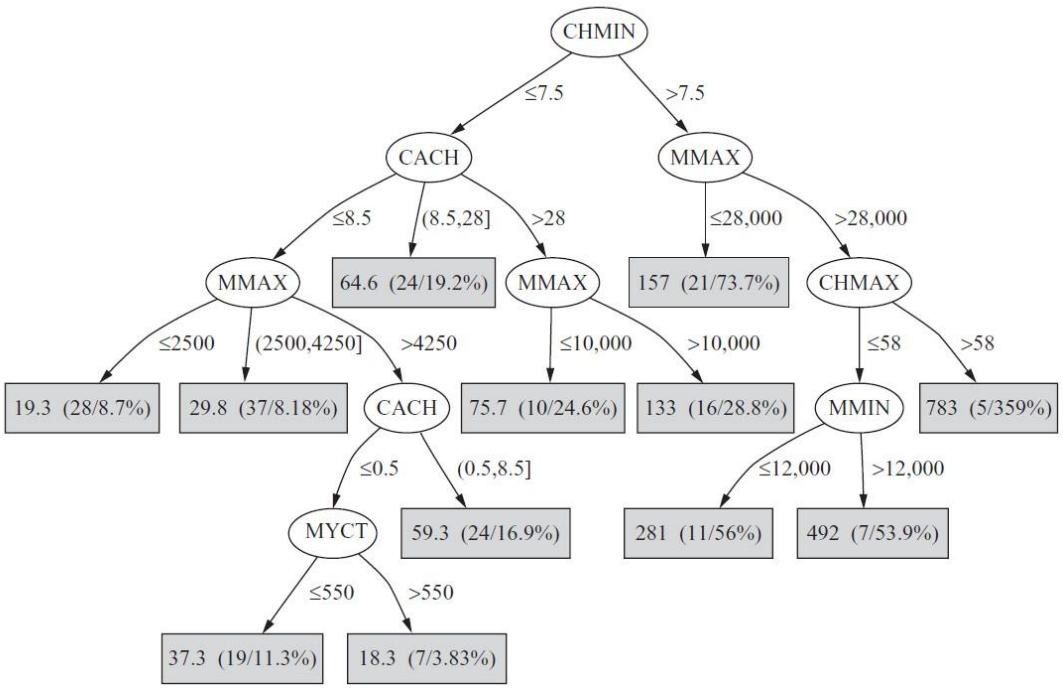
Gambar 16. SVR untuk Emission Dataset

Gambar 16 adalah hasil model regresi dengan SVR untuk Emission Dataset.

Lab : silakan jalankan notebook Support Vector Regression.ipynb

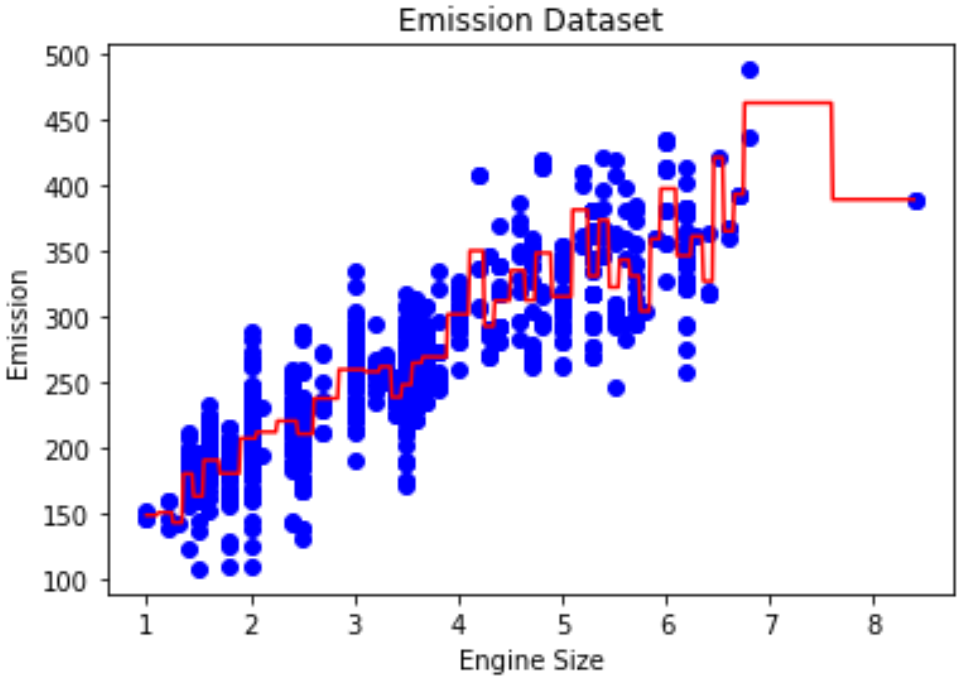
I. Decision Tree Regression

- 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.



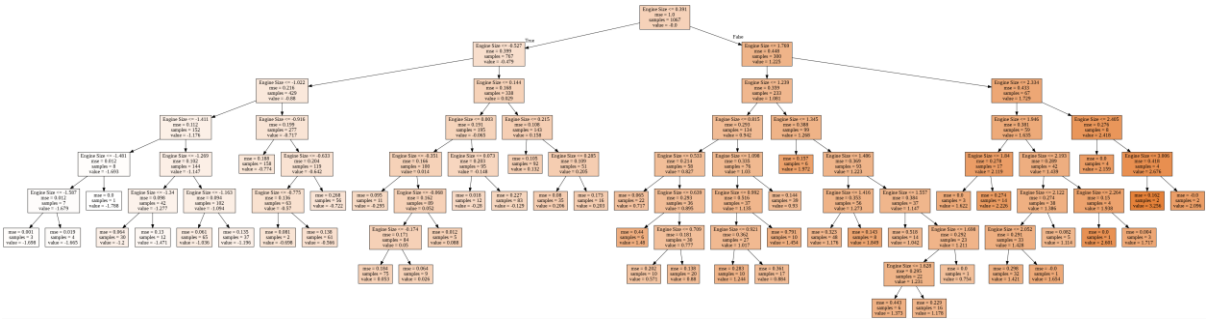
Gambar 17. Decision Tree Regression

Hasil dari pemodelan regresi berupa Decision Tree dengan leaves berupa nilai rata-rata dari data yang ada di leaf tersebut dalam bentuk numerik sebagaiman tampak pada Gambar 17.



Gambar 18. Aplikasi DTR pada Emission Dataset.

Hasil pemodelan DTR pada Emission Dataset tampak pada Gambar 18 berupa garis patah-patah warna merah, sedangkan hasil tree-nya dapa dilihat pada Gambar 19.



Gambar 19.Visualisasi DTR hasi pemodelan dengan Emission Dataset

Lab : silakan jalankan notebook Decision Tree Regression.ipynb

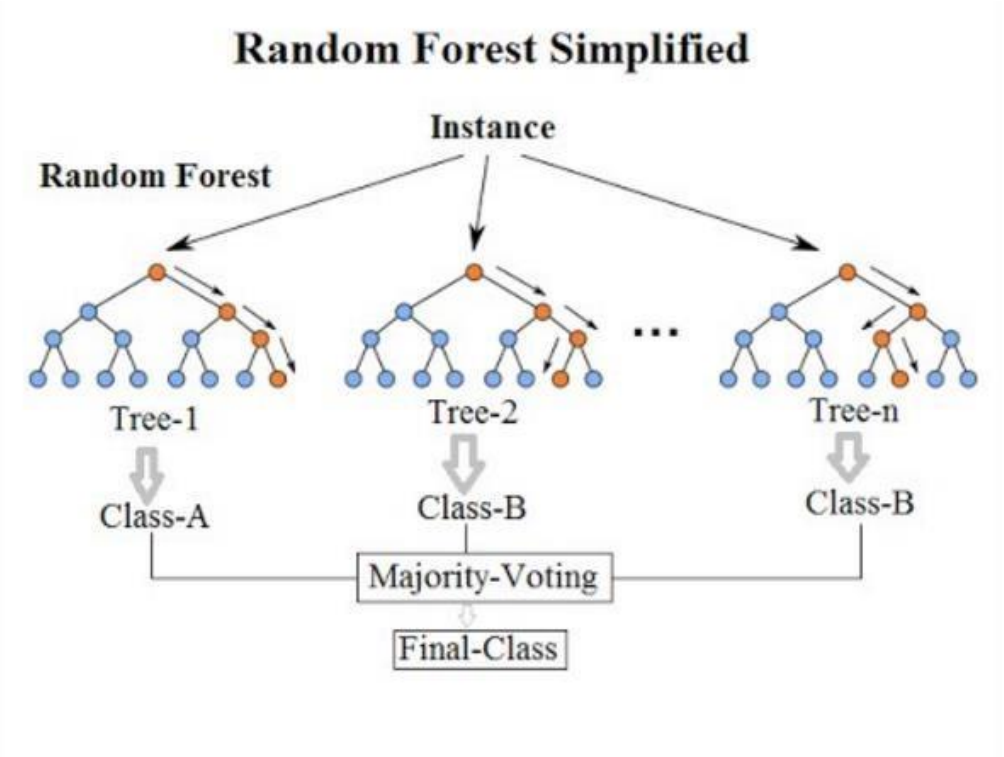
J. Random Forest Regression (RFR)

- Pohon Keputusan (Decision Tree) adalah algoritma yang mudah dipahami dan diinterpretasikan dan karenanya satu pohon mungkin tidak cukup bagi model untuk mempelajari

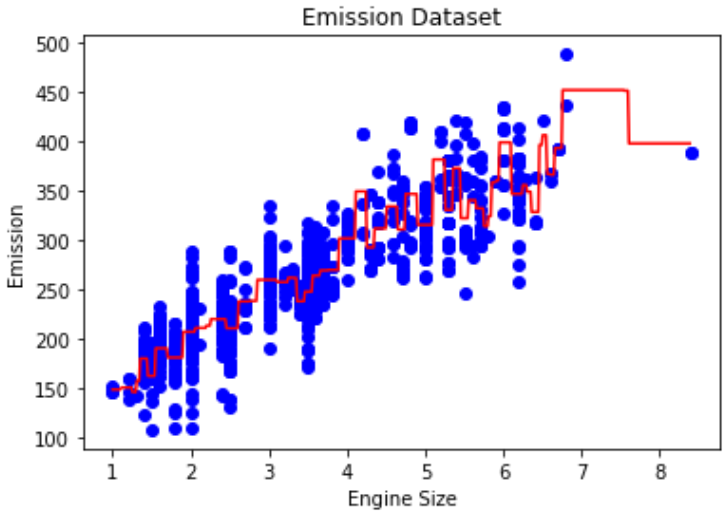
fitur-fiturnya. Di sisi lain, Random Forest juga merupakan algoritma berbasis “Pohon” yang menggunakan fitur kualitas dari beberapa Pohon Keputusan untuk membuat keputusan.

- Oleh karena itu, dapat disebut sebagai ‘Forest’ atau ‘Hutan’ dari pohonpohon 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.

Ilustrasi pembentukan random trees dapat dilihat pada Gambar 20.

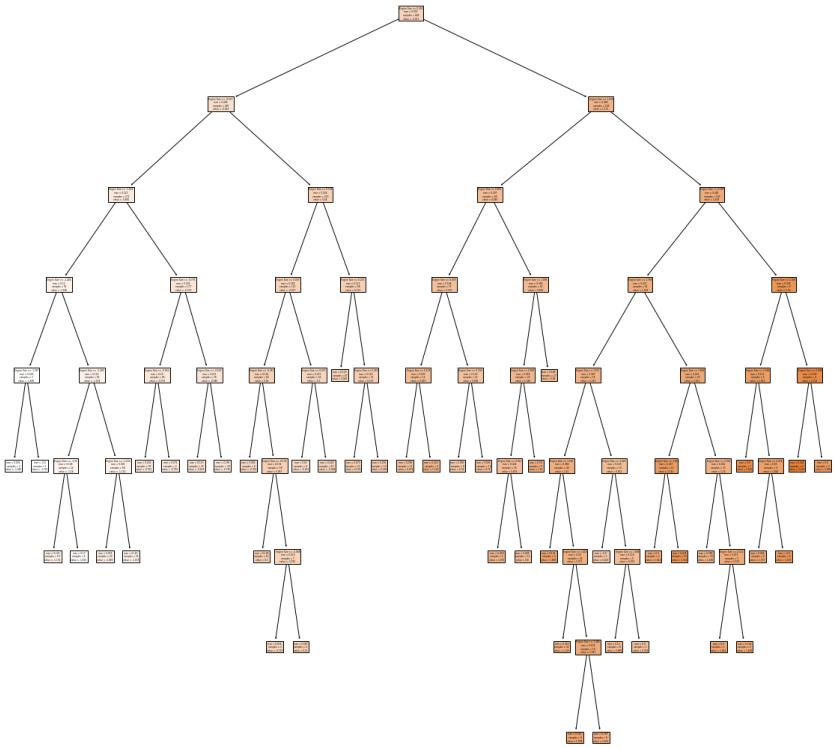


Gambar 20. Ilustrasi RFR



Gambar 21. Aplikasi RFR pada Emission Dataset

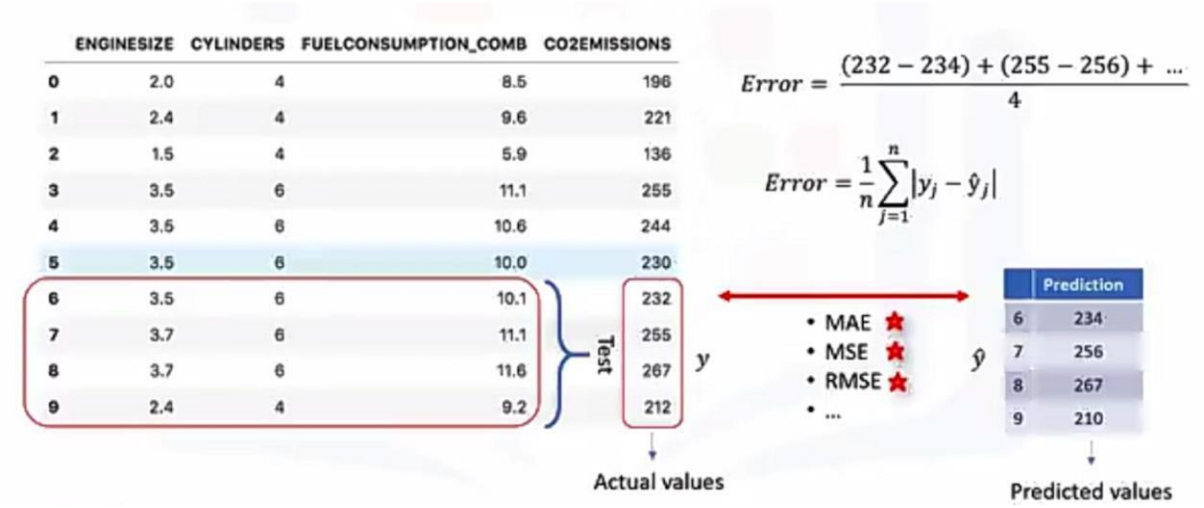
Hasil model dari aplikasi RFR pada Emission Dataset tampak pada Gambar 21, dengan garis putus-putus berwarna merah.
Visualisasi model yang dihasilakn oleh RFR pada Emission Dataset tampak paa Gambar 22.



Gambar 22. Visualisasi Struktur Tree RFR – Emission Dataset

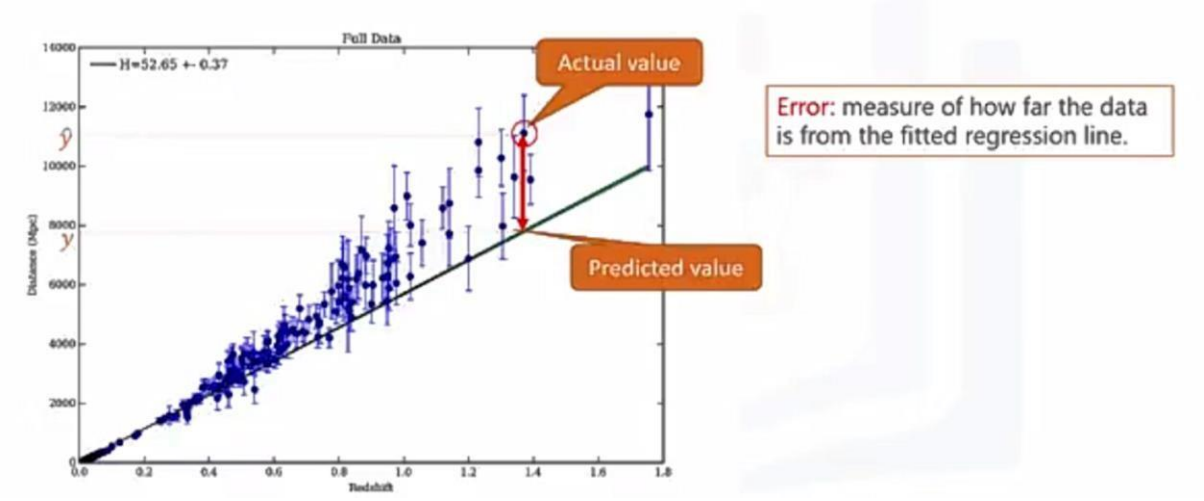
Lab : silakan jalankan notebook Random Forest Regression.ipynb

K. Metrik Evaluasi 1. Error



Gambar 23. Formula Error (MAE)

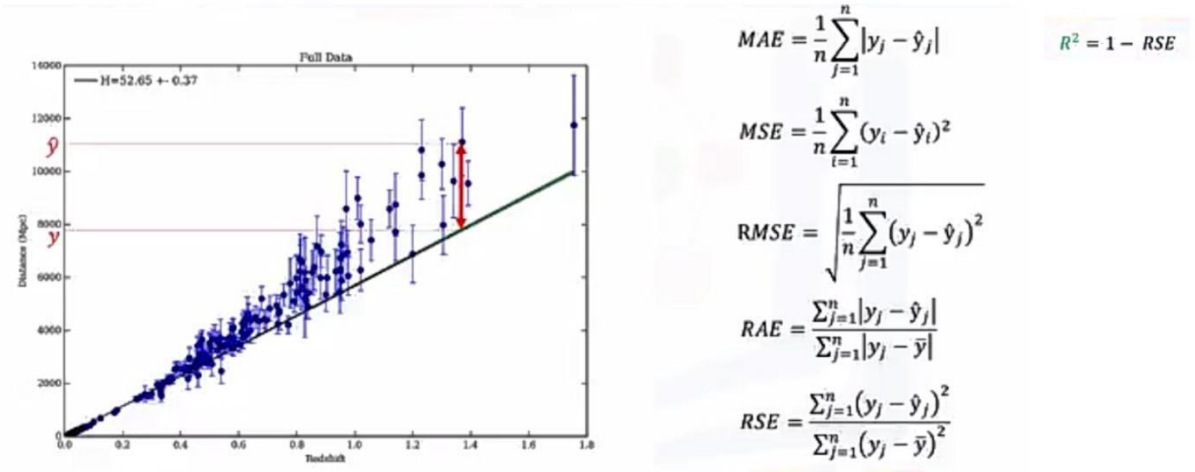
Metrik evaluasi yang paling umum digunakan adalah Error yang terdiri dari berbagai formula antara lain Mean Absoute Error sebagaimana pada Gambar 23. Selain itu terdapat Mean Squared Error, Root Mean Squared Error, dan lainnya.



Gambar 24. Ilustrasi Error

Error didefinisikan sebagai ukuran sejauh mana data aktual terhadap garis regresi yang sudah dicocokan sebagaimana

diilustrasikan pada Gambar 24. Garis regresi tersebut memuat nilai prediksi sedangkan titik-titik biru yang ada adalah nilai aktualnya.



Gambar 25. Berbagai Tipe Error (MAE, MSE, RMSE, RAE, RSE) dan metrik R2

Metrik kinerja model regresi yang dihasilkan diukur dengan metrik berbagai macam seperti pada Gambar 25.

L. Perbandingan Berbagai Algoritma Regresi 1. Regresi Linier (RL) 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.

2. 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.

3. 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.

M. Ringkasan

- 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 R2
- Terdapat banyak algoritma regresi yang dapat digunakan dengan kelebihan dan kekurangan masing-masing.