**Job Sheet 3: Regresi**

**Tugas 1: Multiple Linear Regression**

Mohammad Izamul Fikri Fahmi – 2141720171

19/TI-3H

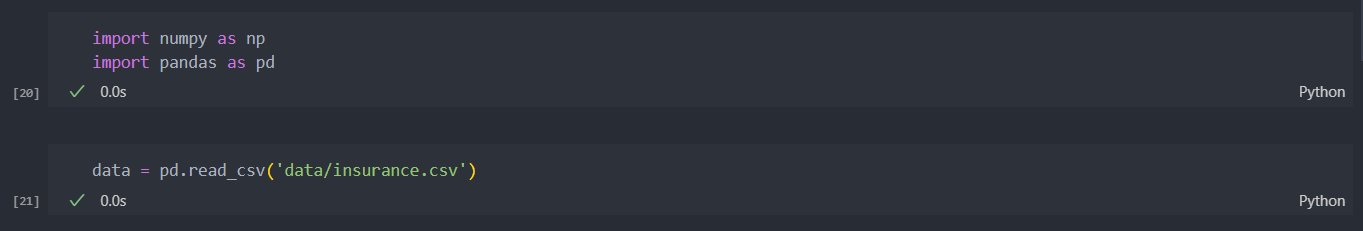
**Intruksi:**

1. Identifikasi variabel-variabel yang akan digunakan sebagai variabel bebas (fitur) dan variabel target (biaya medis personal).
2. Bagi dataset menjadi data latih (train) dan data uji (test) dengan proporsi yang sesuai.
3. Lakukan feature scaling jika diperlukan.
4. Buat model multiple linear regression menggunakan Scikit-Learn.
5. Latih model pada data latih dan lakukan prediksi pada data uji.
6. Evaluasi model dengan menghitung metrik seperti R-squared, MSE, dan MAE. Tampilkan hasil evaluasi.

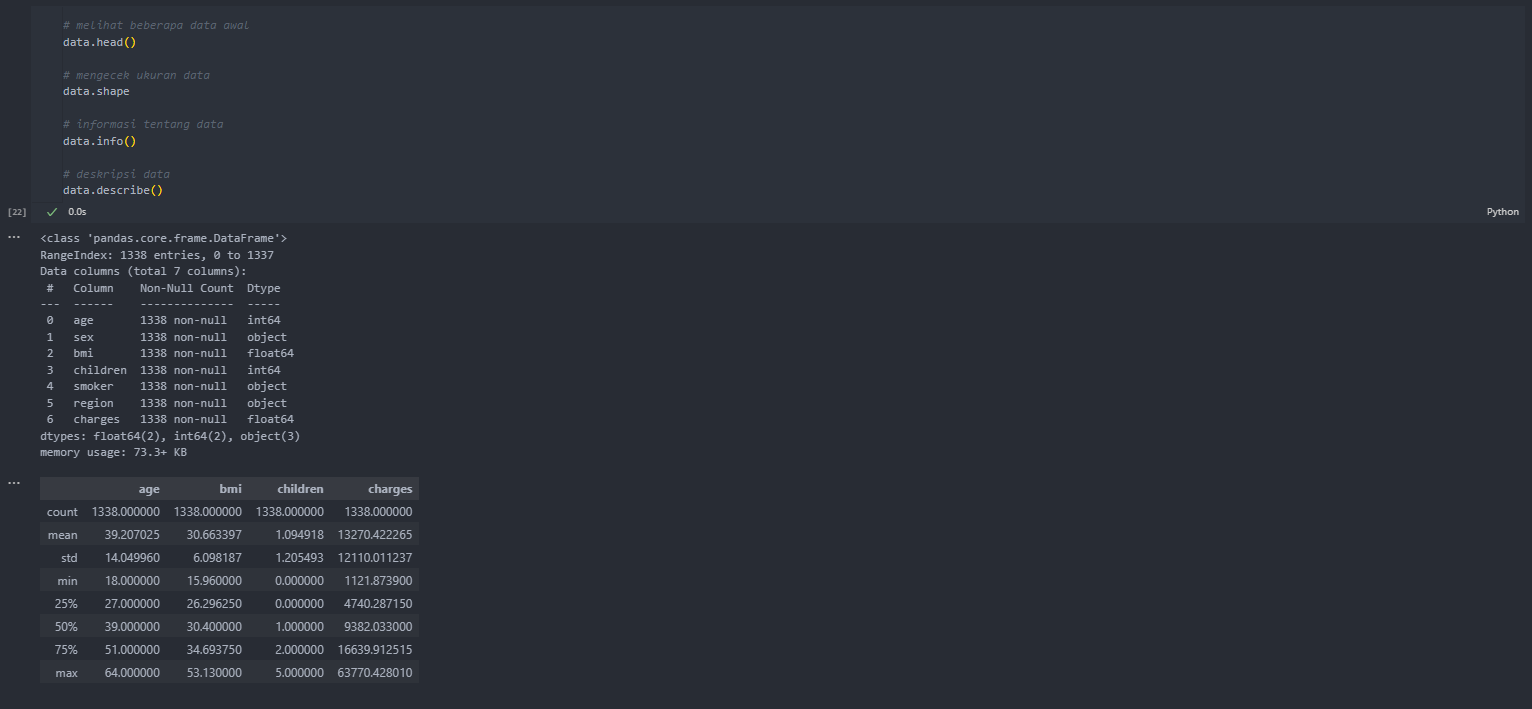
**Pengerjaan:**

Sebelum kita mengerjakan intruksi ini, kita perlu memahami apa itu Multiple Linear Regression, jadi MLR adalah teknik statistik yang digunakan untuk memodelkan hubungan antara satu variabel dependen (output) dan dua atau lebih variabel independen (input) dengan asumsi bahwa hubungan tersebut adalah linier. Dengan modal pengetahuan dasar tentang MLR, mari kita kerjakan sesuai intruksi.

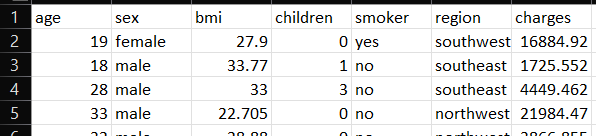
kita import library numpy dan pandas juga data yang akan kita lakukan analisis yaitu data insurance



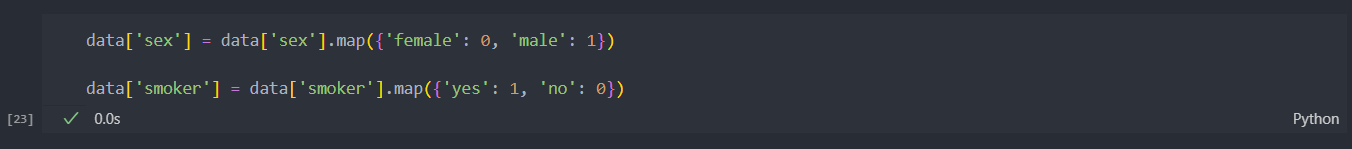
Setelah itu, kita akan melihat beberapa informasi tentang data tersebut



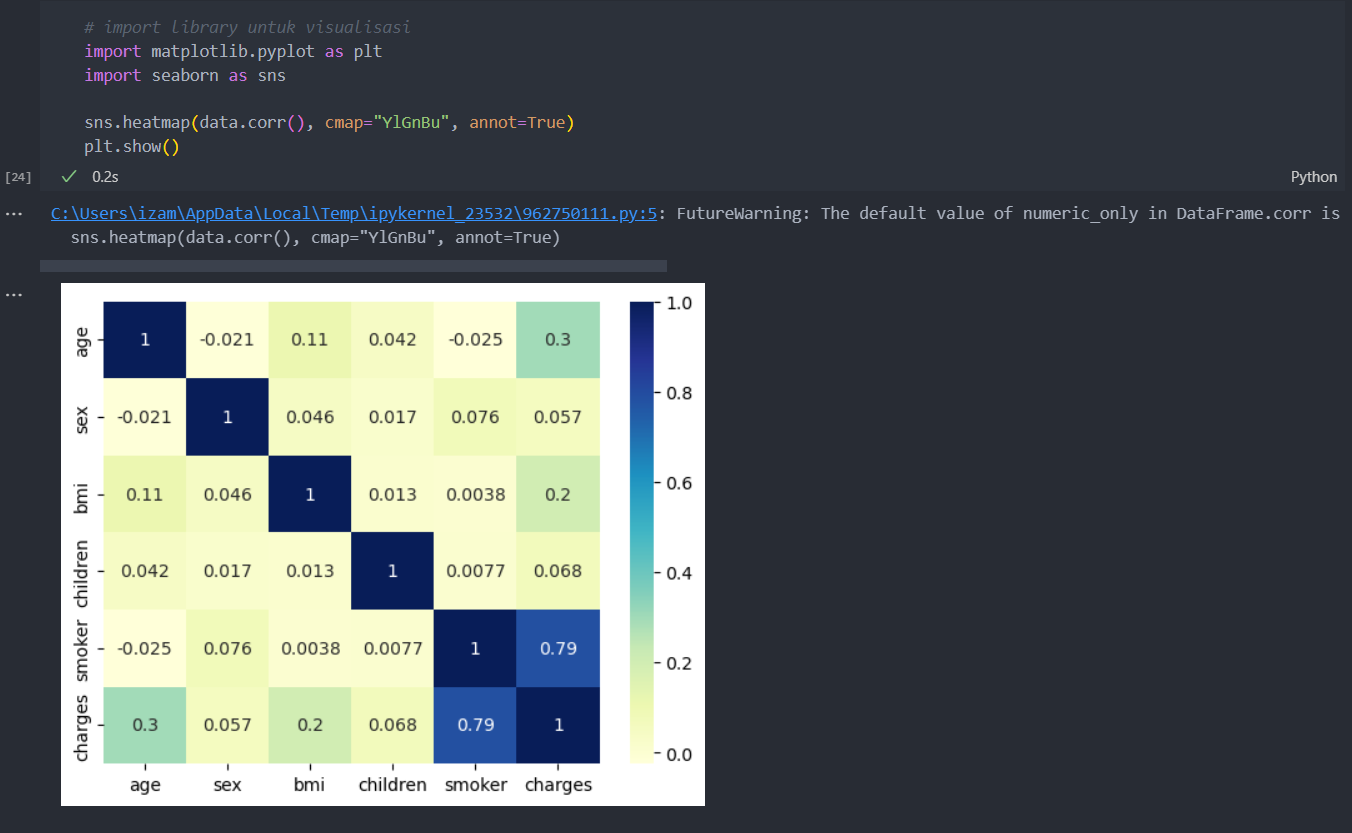
Kita juga bisa langsung mengecek dengan membuka file csv tersebut



Karena data **sex** dan **smoker** masih berupa abjad, mari kita ubah ke numerik sesuai dengan nilainya



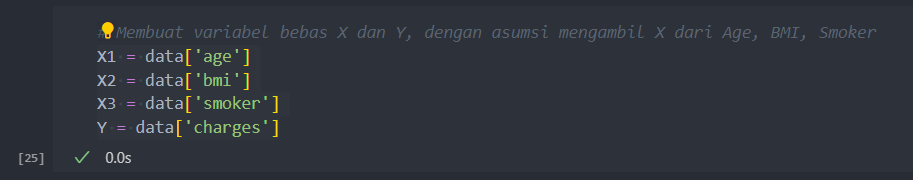
Dengan, demikian kita bisa melihat heat map untuk mengecek korelasi data terhadap charges(sebagai dependen variable)



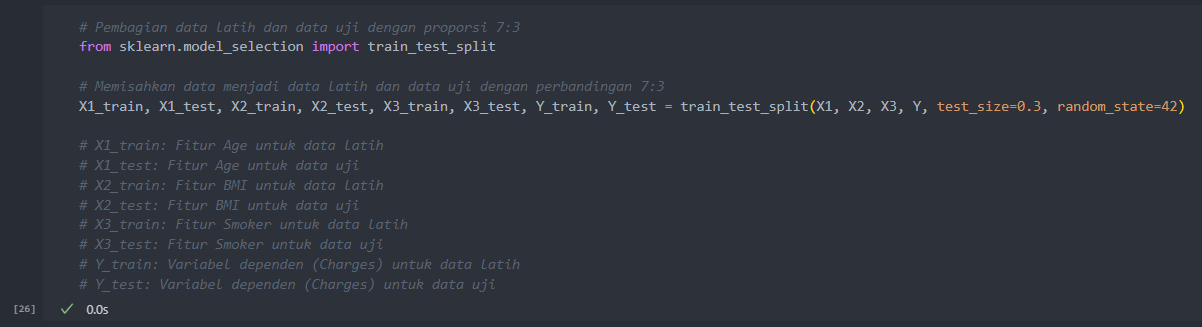
Dengan melihat pada matriks, terlihat ada setidaknya 3 variabel independent yang memiliki korelasi dekat dengan charges dengan urutan; **smoker(0.79), age(0.3), bmi(0.2)**, berangkat dari data heat map ini maka kita melakukan dua uji coba, yaitu dengan 3 variable tersebut dan dengan 2 variable saja yaitu; **smoker** dan **age**.

# Percobaan 1: Smoker, Age, BMI

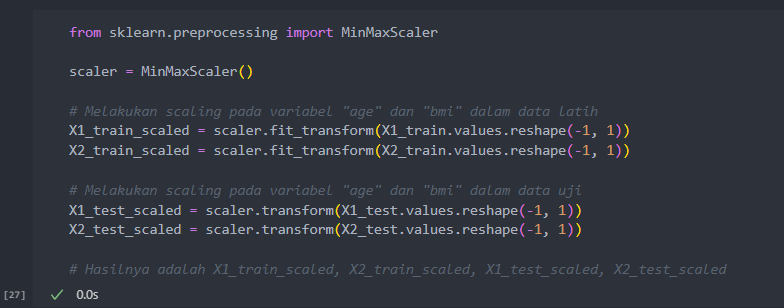
Membuat variable bebas dari masing-masing data:



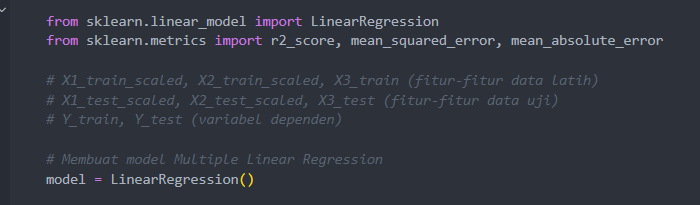
Membagi dataset menjadi data latih dan data uji, dengan proporsi 70:30



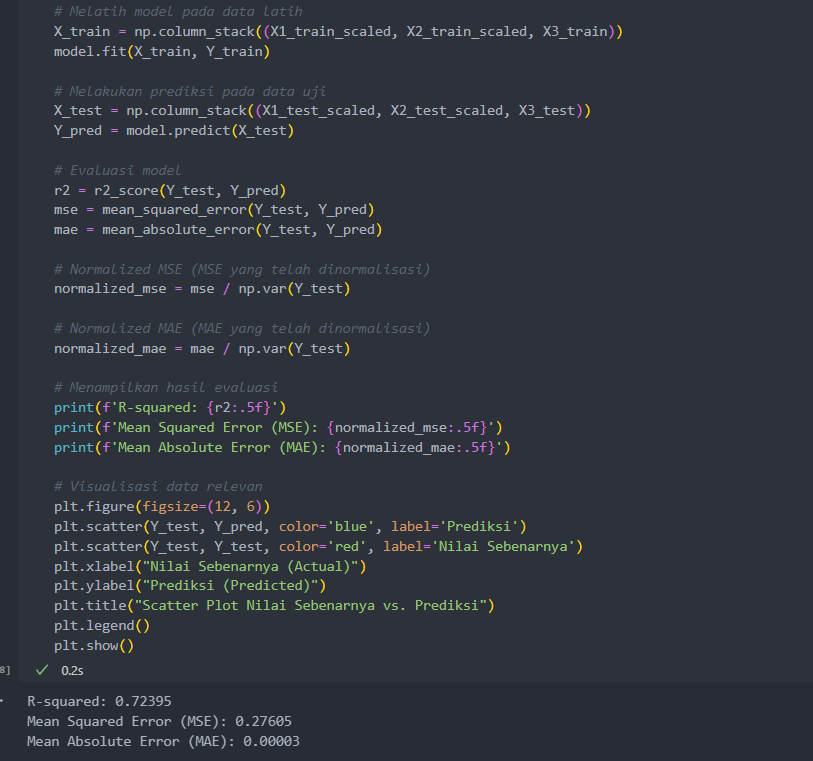
Karena smoker sudah dalam rentang 0-1, maka kita lakukan scaling variable untuk “age” dan “bmi”

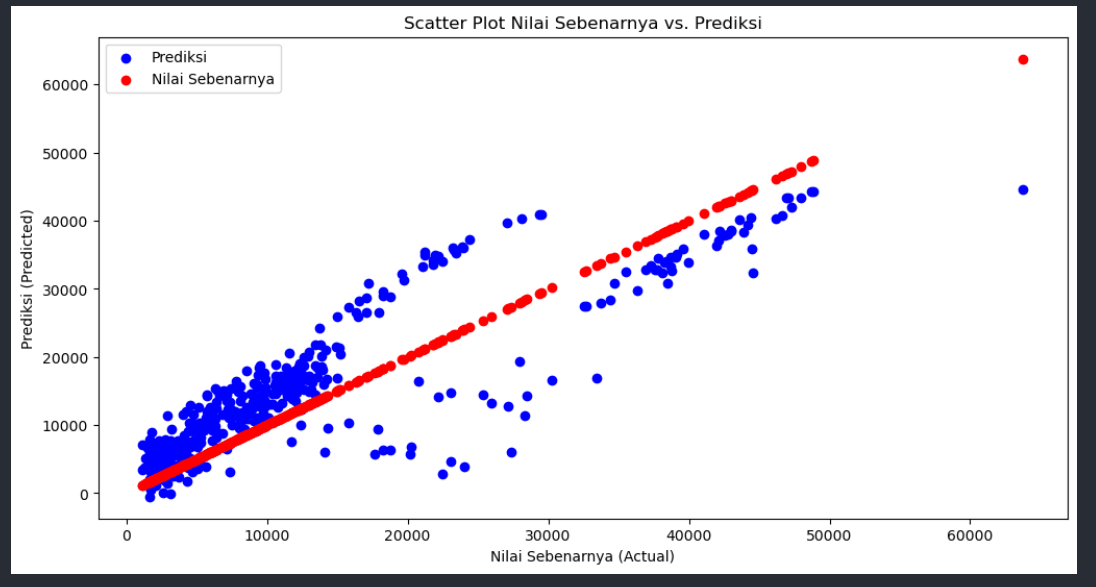


Setelah dilakukan scalling, kita bisa memuat model multiple linear regression menggunakan Scikit-Learn



Kemudian kita lakukan uji data dan evaluasi hasil





Analisis hasil evaluasi:

*R-squared (R^2):*

* R-squared adalah metrik yang mengukur seberapa baik model regresi sesuai dengan data. Nilai R-squared berkisar antara 0 hingga 1, dan semakin tinggi nilainya, semakin baik modelnya sesuai dengan data.
* Nilai R-squared adalah 0.72395, yang berarti sekitar 72,4% dari variabilitas dalam data dapat dijelaskan oleh model. Ini adalah nilai yang relatif tinggi, yang menunjukkan bahwa model cukup baik dalam menjelaskan variasi dalam data.

*Mean Squared Error (MSE):*

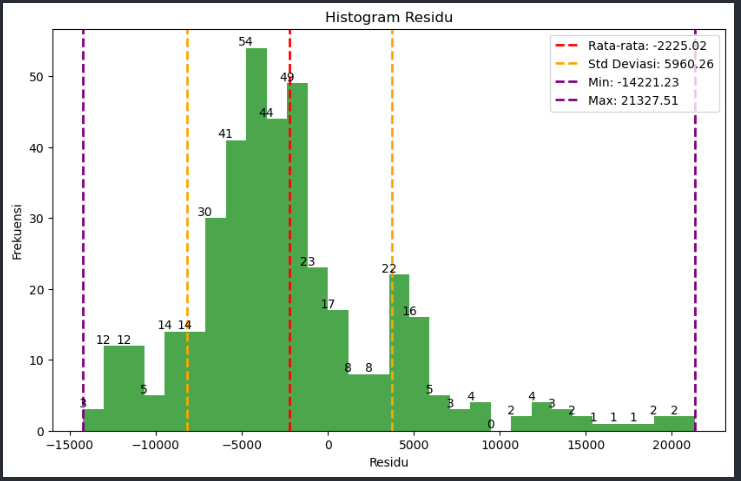
* MSE mengukur seberapa besar kesalahan rata-rata kuadrat antara nilai sebenarnya dan nilai prediksi model. Semakin rendah MSE, semakin baik modelnya.
* Nilai MSEadalah 0.27605, yang menunjukkan bahwa kesalahan rata-rata kuadrat antara nilai sebenarnya dan prediksi model adalah 0.27605. Ini adalah nilai yang relatif kecil, yang mengindikasikan bahwa model memiliki kinerja yang baik dalam meminimalkan kesalahan.

*Mean Absolute Error (MAE):*

* MAE mengukur seberapa besar kesalahan rata-rata mutlak antara nilai sebenarnya dan nilai prediksi model. Semakin rendah MAE, semakin baik modelnya.
* Nilai MAE adalah 0.00003, yang menunjukkan bahwa kesalahan rata-rata mutlak antara nilai sebenarnya dan prediksi model adalah sangat kecil, yaitu 0.00003. Ini adalah nilai yang sangat rendah, yang mengindikasikan bahwa modelmemiliki kinerja yang sangat baik dalam meminimalkan kesalahan.

**Tambahan**

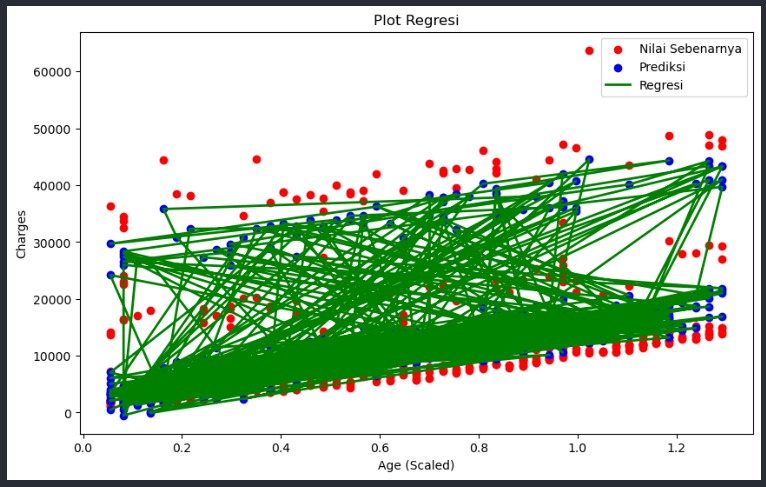
*Histogram Residu*

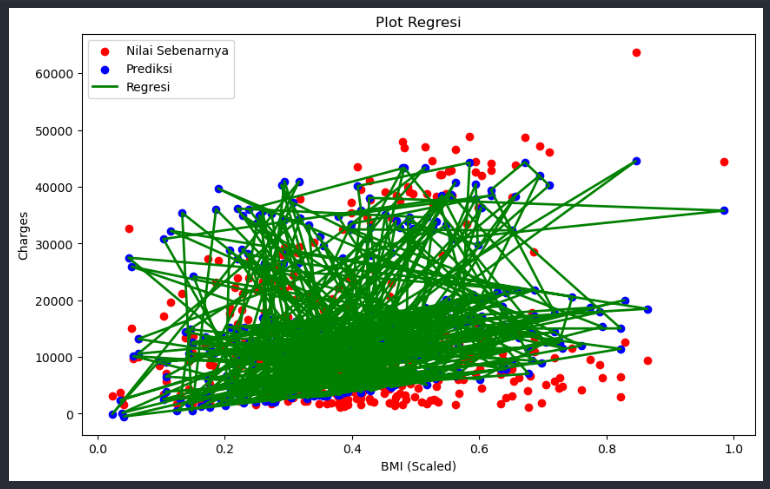


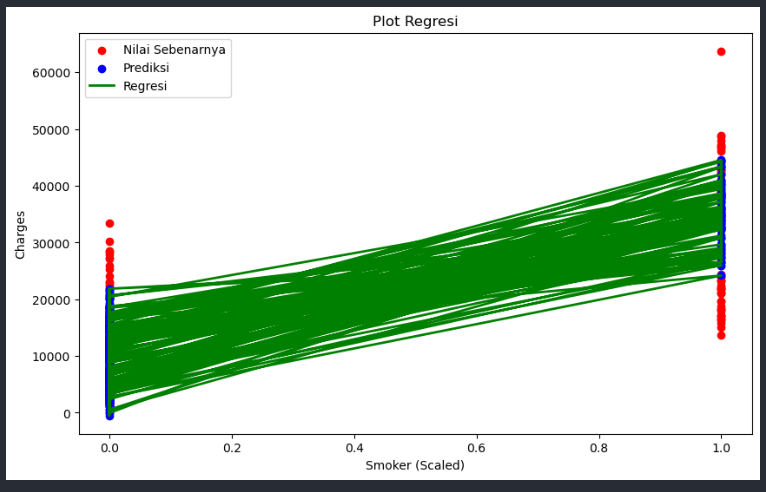
Dengan data histogram:  
Rata-rata : -2225.02  
STD Deviasi : 5960.26  
Min : -14221.23  
Max : 21327.51

Menunjukan rata-rata residu masih cukup besar, berkisar -2225.0 dari data sebenarnya denagn data prediksi.

*Plot Regresi: Age, BMI, Smoker*

****

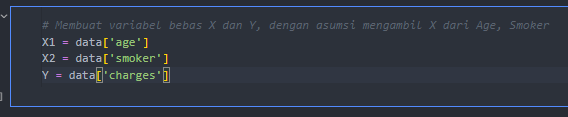
****

****

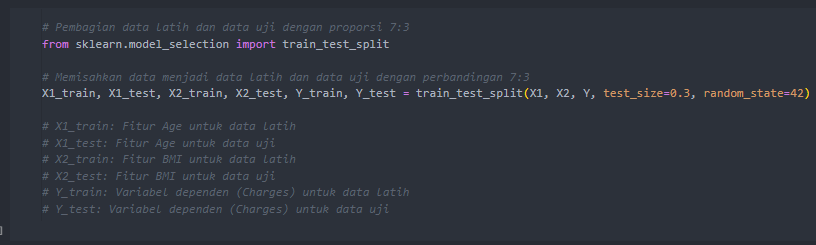
Menunjukan Smoker jauh lebih baik dalam memprediksi pengaruh X terhadap Y(Charges)

# Percobaan 2: Smoker, Age

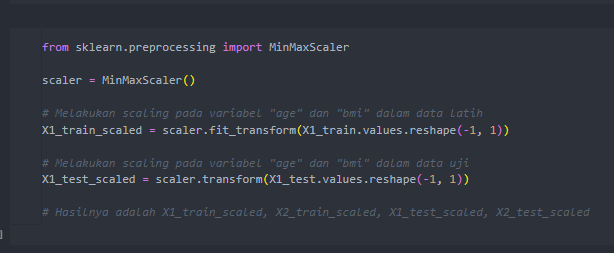
Membuat variable bebas dari masing-masing data:



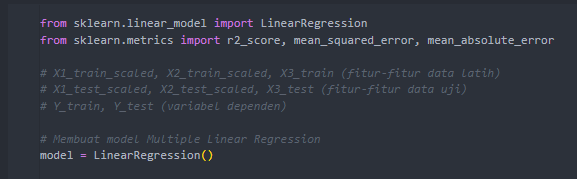
Membagi dataset menjadi data latih dan data uji, dengan proporsi 70:30



Karena smoker sudah dalam rentang 0-1, maka kita lakukan scaling variable untuk “age”

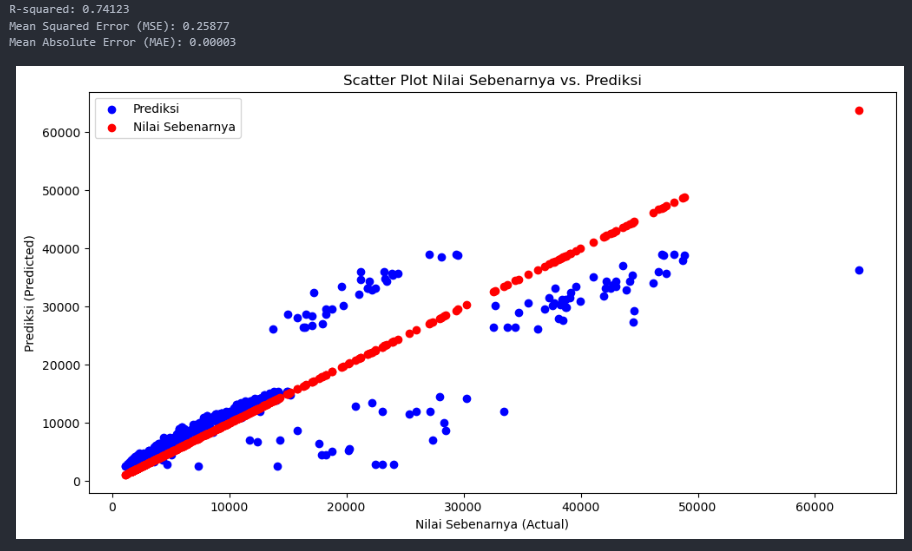


Setelah dilakukan scalling, kita bisa memuat model multiple linear regression menggunakan Scikit-Learn



Kemudian kita lakukan uji data dan evaluasi hasil





Analisis hasil evaluasi:

*R-squared (R^2):*

* R-squared adalah metrik yang mengukur seberapa baik model regresi sesuai dengan data. Nilai R-squared berkisar antara 0 hingga 1, dan semakin tinggi nilainya, semakin baik modelnya sesuai dengan data.
* Nilai R-squared adalah 0.74123, yang berarti sekitar 74,1% dari variabilitas dalam data dapat dijelaskan oleh model. Ini adalah nilai yang relatif tinggi, yang menunjukkan bahwa model cukup baik dalam menjelaskan variasi dalam data.

*Mean Squared Error (MSE):*

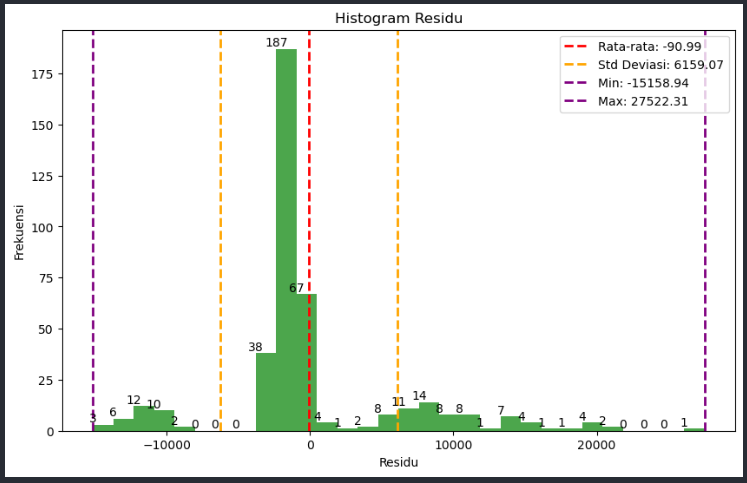
* MSE mengukur seberapa besar kesalahan rata-rata kuadrat antara nilai sebenarnya dan nilai prediksi model. Semakin rendah MSE, semakin baik modelnya.
* Nilai MSEadalah 0.25877, yang menunjukkan bahwa kesalahan rata-rata kuadrat antara nilai sebenarnya dan prediksi model adalah 0.27605. Ini adalah nilai yang relatif kecil, yang mengindikasikan bahwa model memiliki kinerja yang baik dalam meminimalkan kesalahan.

*Mean Absolute Error (MAE):*

* MAE mengukur seberapa besar kesalahan rata-rata mutlak antara nilai sebenarnya dan nilai prediksi model. Semakin rendah MAE, semakin baik modelnya.
* Nilai MAE adalah 0.00003, yang menunjukkan bahwa kesalahan rata-rata mutlak antara nilai sebenarnya dan prediksi model adalah sangat kecil, yaitu 0.00003. Ini adalah nilai yang sangat rendah, yang mengindikasikan bahwa modelmemiliki kinerja yang sangat baik dalam meminimalkan kesalahan.

**Tambahan**

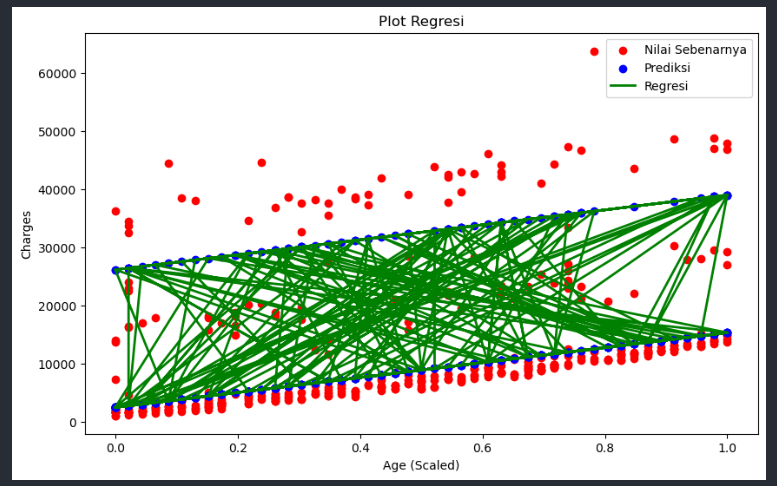
*Histogram Residu*

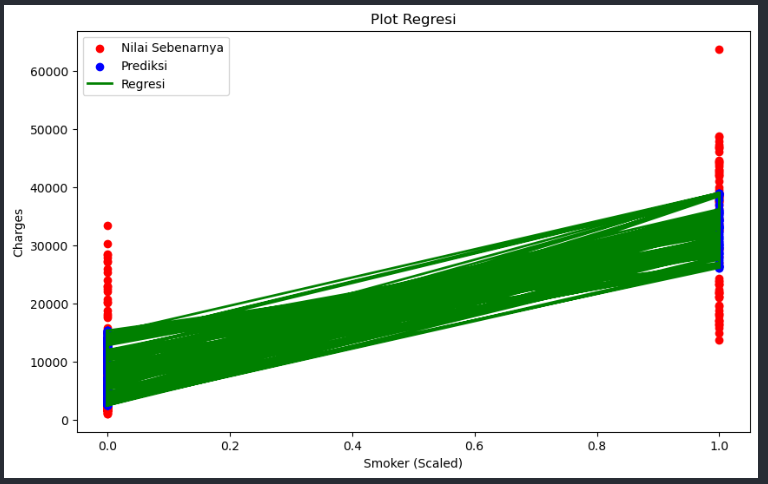


Dengan data histogram:  
Rata-rata : -90.99  
STD Deviasi : 6159.07  
Min : -15158.94  
Max : 27522.31

Menunjukan rata-rata residu sangat rendah di angka -90.99

*Plot Regresi: Age, Smoker*

****

****

Menunjukan Smoker jauh lebih baik dalam memprediksi pengaruh X terhadap Y(Charges) daripada Age

# Kesimpulan: Percobaan 1 dengan Percobaan 2

Dalam menentukan fitur independent untuk dependent, sangat penting untuk melihat korelasi paling mempengaruhi dependent.

Dengan melakukan dua percobaan dengan variable independent yang berbeda dapat dinyatakan beberapa perbandingan hasil dalam tabel berikut:

|  |  |  |
| --- | --- | --- |
|  | Percobaan 1 | Percobaan 2 |
| R-Squared | 0.72395 | 0.74123 |
| Mean Squared Error | 0.27605 | 0.25877 |
| Mean Absolute Error | 0.00003 | 0.00003 |
| Rata-rata residu | -2225.02 | -90.99 |

Dapat disimpulkan bahwa percobaan 2 lebih baik dalam melakukan prediksi dengan menggunakan variable Independent **Smoker** dan **Age** untuk menentukan pengaruh terhadap **Charges**.