**Nama :Muhammad Dzaka Murran Rusid**

**NIM :2141720076**

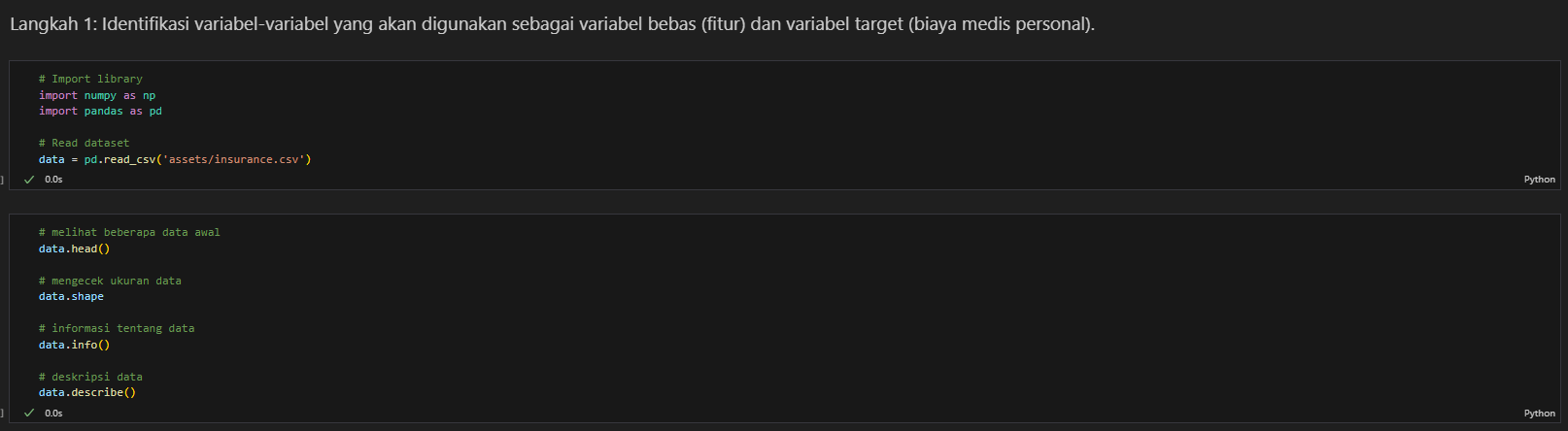
**Kelas/Abs :TI-3H/21**

**Intruksi:**

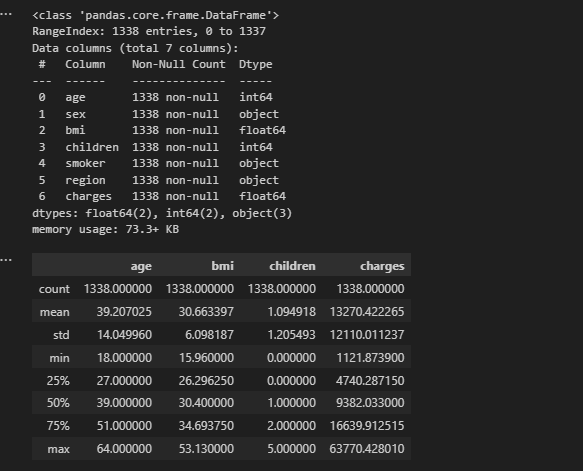
1. Identifikasi variabel-variabel yang akan digunakan sebagai variabel bebas (fitur) dan variabel target (biaya medis personal).
2. Bagi dataset menjadi data latih (train) dan data uji (test) dengan proporsi yang sesuai.
3. Lakukan feature scaling jika diperlukan.
4. Buat model multiple linear regression 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:**

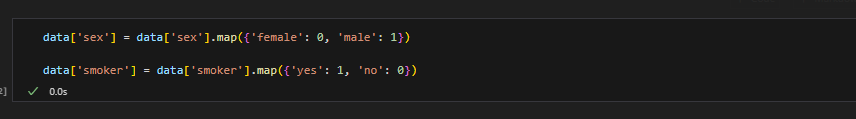
Langkah pertama yang harus dilakukan adalah mengimpor library numpy dan pandas serta membaca dataset insurance.csv yang telah diberikan



Kemudian setelah me-run data.head() dan lain sebagainya, akan terlihat informasi berkaitan dataset tersebut. Alternatif lainnya bisa juga langsung melihatnya di aplikasi excel.



Data **Sex** dan **Smoker masih** bukan berbentuk numerik, jadi harus dilakukan proses mapping terlebih dahulu.



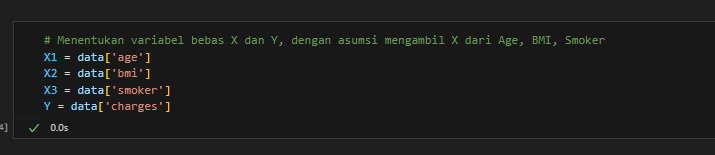
Setelah di-mapping, barulah bisa dicek korelasi antar data dan charges.



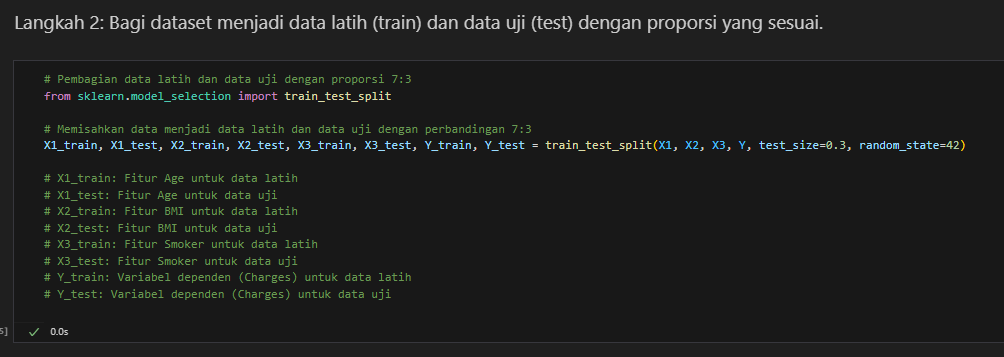
Dengan melihat pada matriks, terlihat ada setidaknya 3 variabel independent yang memiliki korelasi dekat dengan charges dengan urutan; **smoker(0.79), age(0.3), bmi(0.2)**, berangkat dari data heat map ini maka kita melakukan dua uji coba, yaitu dengan 3 variable tersebut dan dengan 2 variable saja yaitu;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

**Menggunakan variabel Age, BMI, dan Smoker**

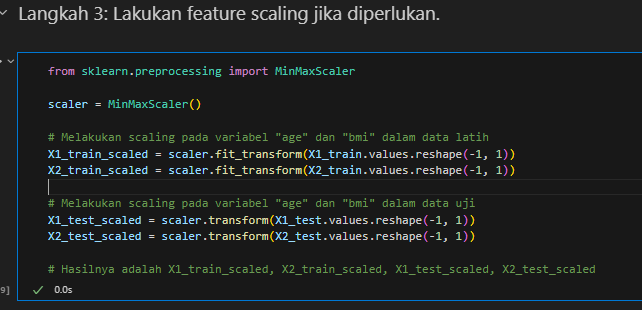
Membuat variable independen dari masing-masing data:



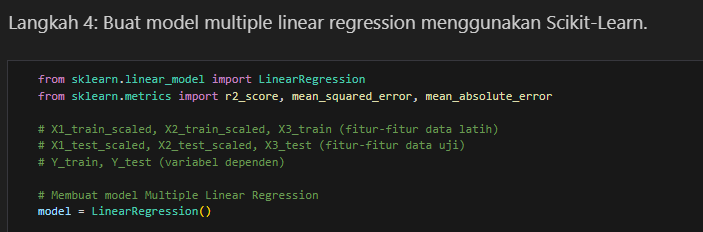
Membagi dataset menjadi data latih dan data uji, proporsi 70:30



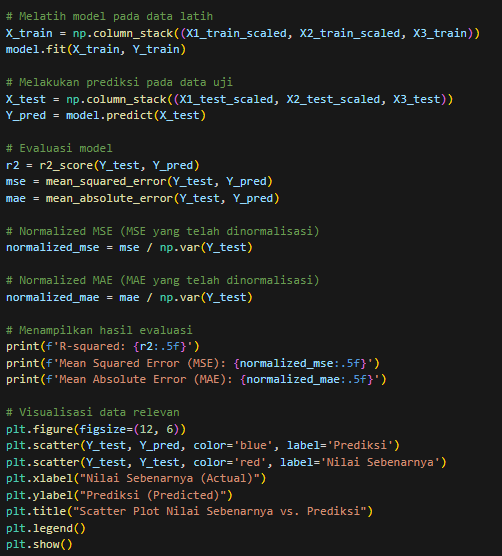
Karena smoker sudah dalam rentang 0-1, maka kita lakukan scaling variable untuk “age” dan “bmi”

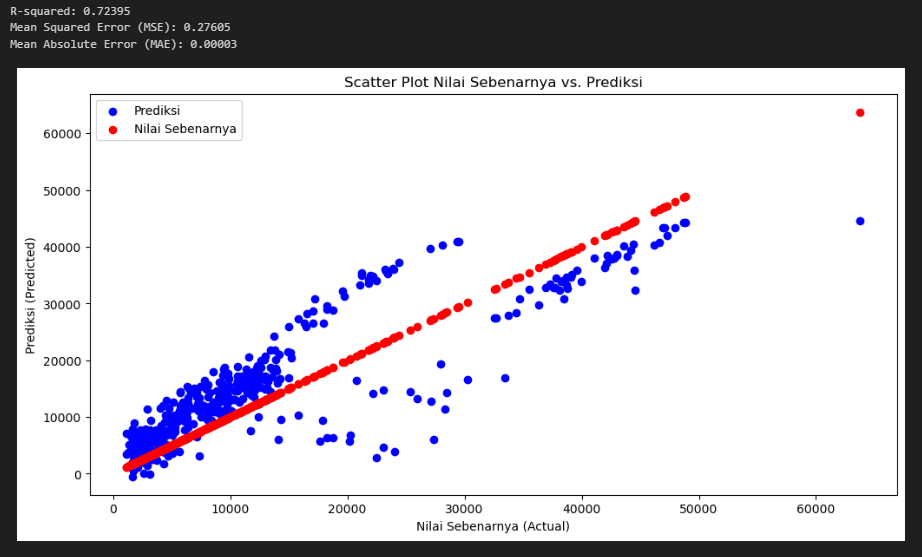


Setelah dilakukan scalling, kita bisa memuat model multiple linear regression menggunakan Scikit-Learn



Kemudian dilakukan uji data dan evaluasi hasil





Analisis hasil evaluasi:

Koefisien determinasi (R-squared):

* R-squared adalah ukuran seberapa tepat model regresi cocok dengan data. Nilai R-squared berkisar dari 0 hingga 1, dengan semakin tinggi menunjukkan tingkat kesesuaian yang lebih baik antara model dan data.
* R-squared memiliki nilai sebesar 0,72395, yang berarti sekitar 72,4% dari variasi dalam data dapat dijelaskan oleh model. Ini adalah nilai yang tinggi, menandakan model mampu dengan baik dalam menjelaskan variasi dalam data.

Error Kuadrat Rata-rata (MSE):

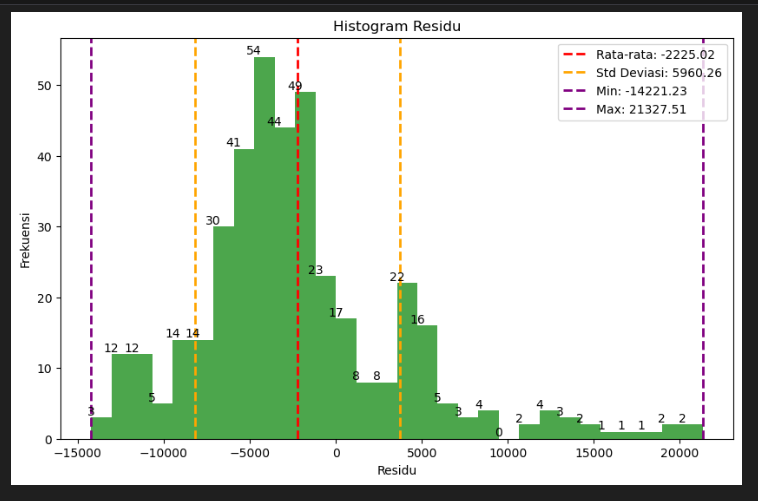
* MSE mengukur seberapa besar rata-rata dari perbedaan kuadrat antara nilai sebenarnya dan nilai yang diprediksi oleh model. Semakin rendah nilai MSE, semakin baik kinerja model.
* Nilai MSE adalah 0,27605, menandakan bahwa kesalahan rata-rata kuadrat antara nilai aktual dan prediksi model adalah 0,27605. Ini adalah nilai yang rendah, yang mengindikasikan bahwa model memiliki kinerja yang baik dalam mengurangi kesalahan.

Error Mutlak Rata-rata (MAE):

* MAE mengukur seberapa besar rata-rata kesalahan absolut antara nilai aktual dan prediksi model. Semakin rendah nilai MAE, semakin baik kinerja model.
* Nilai MAE adalah 0,00003, menunjukkan bahwa kesalahan rata-rata mutlak antara nilai aktual dan prediksi model sangat kecil, yaitu 0,00003. Ini menunjukkan bahwa model memiliki kinerja yang sangat baik dalam mengurangi kesalahan.

**Tambahan**

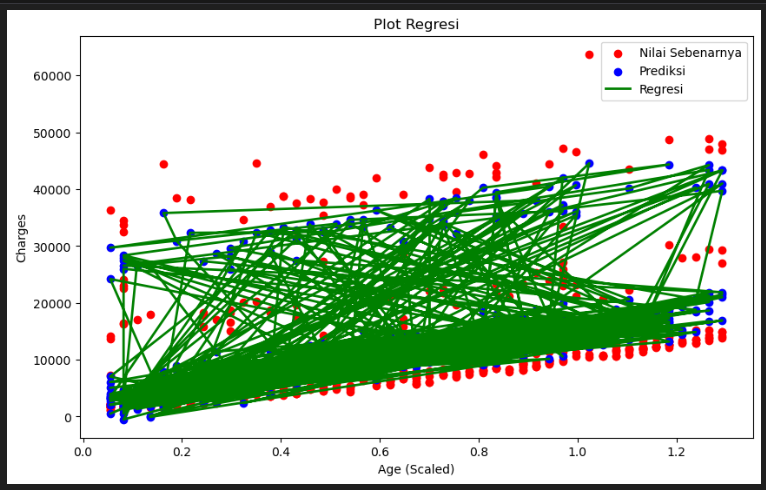
*Histogram Residu*

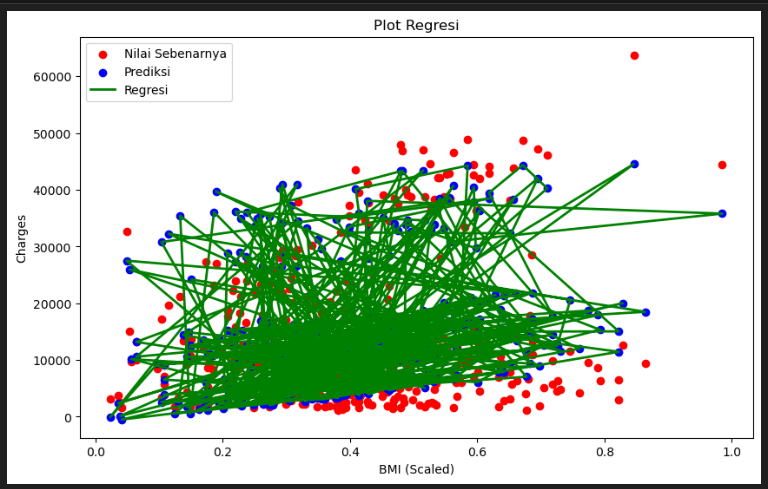


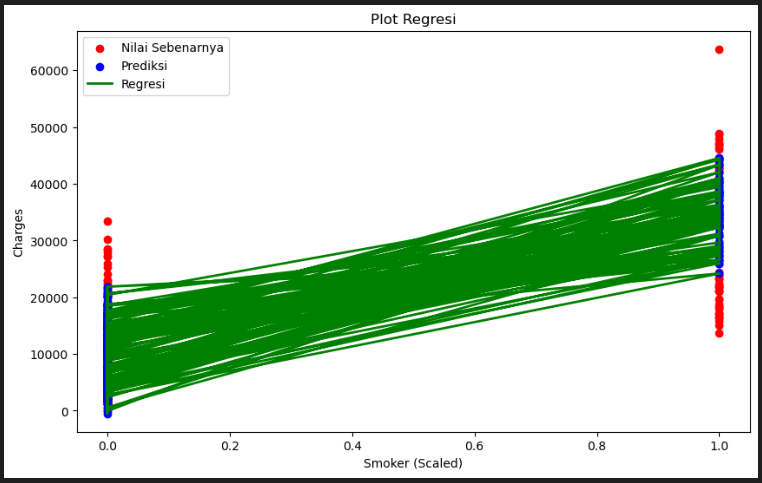
Dengan data histogram:  
Rata-rata : -2225.02  
STD Deviasi : 5960.26  
Min : -14221.23  
Max : 21327.51

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

**Plot Regresi: Age, BMI, Smoker**

****

****

****

Dari ketiga variabel tersebut, ditemukan bahwa variabel Smoker yang paling baik dalam memprediksi pengaruh X terhadap Y (Charges).