**Tutorial 2**

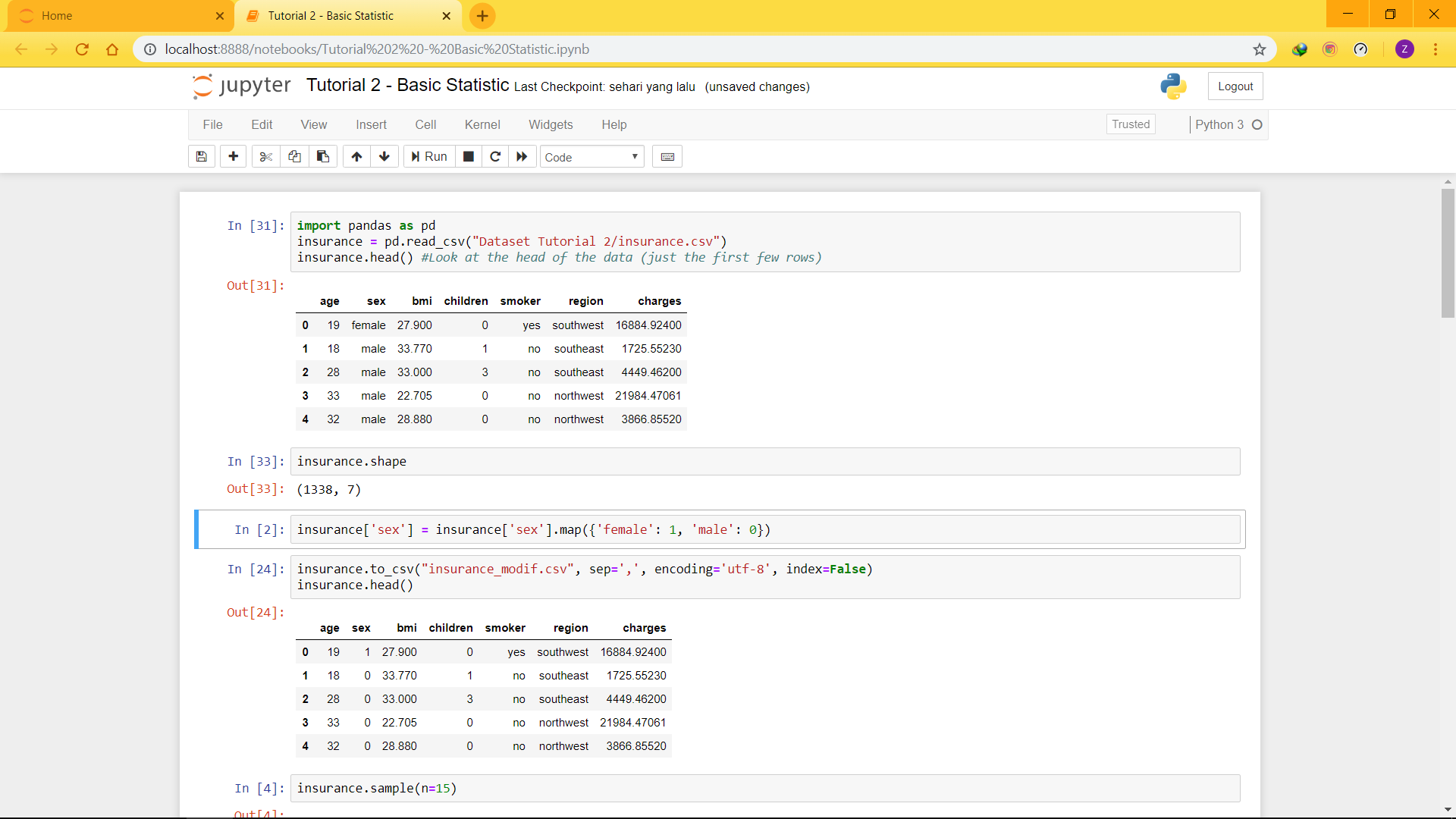
***Machine Learning***

Nama : Zaki Raihan

NPM : 1606878505

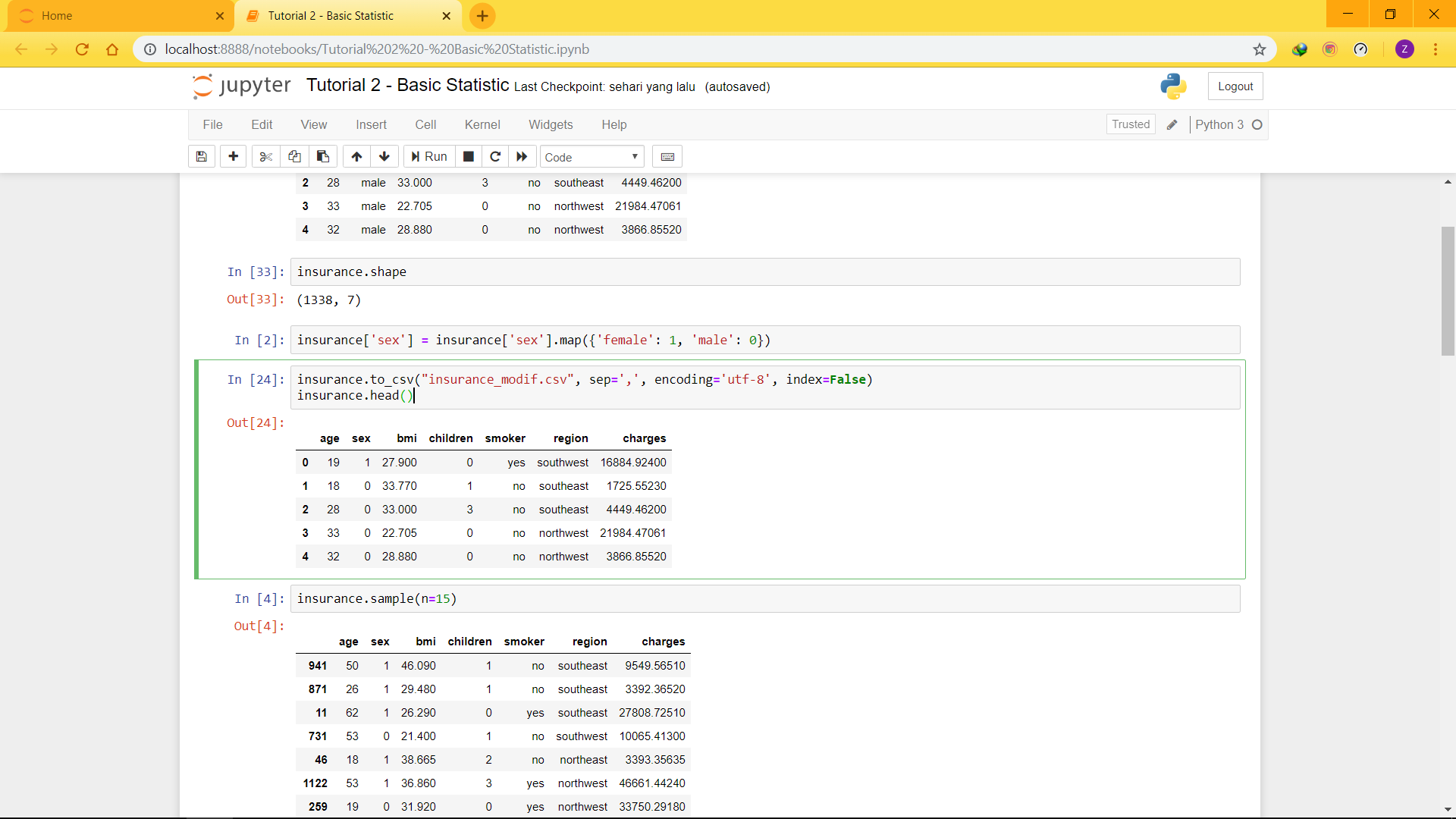
**Basic Statistics**

1. Import data insurance.csv dan hitung dimensi:

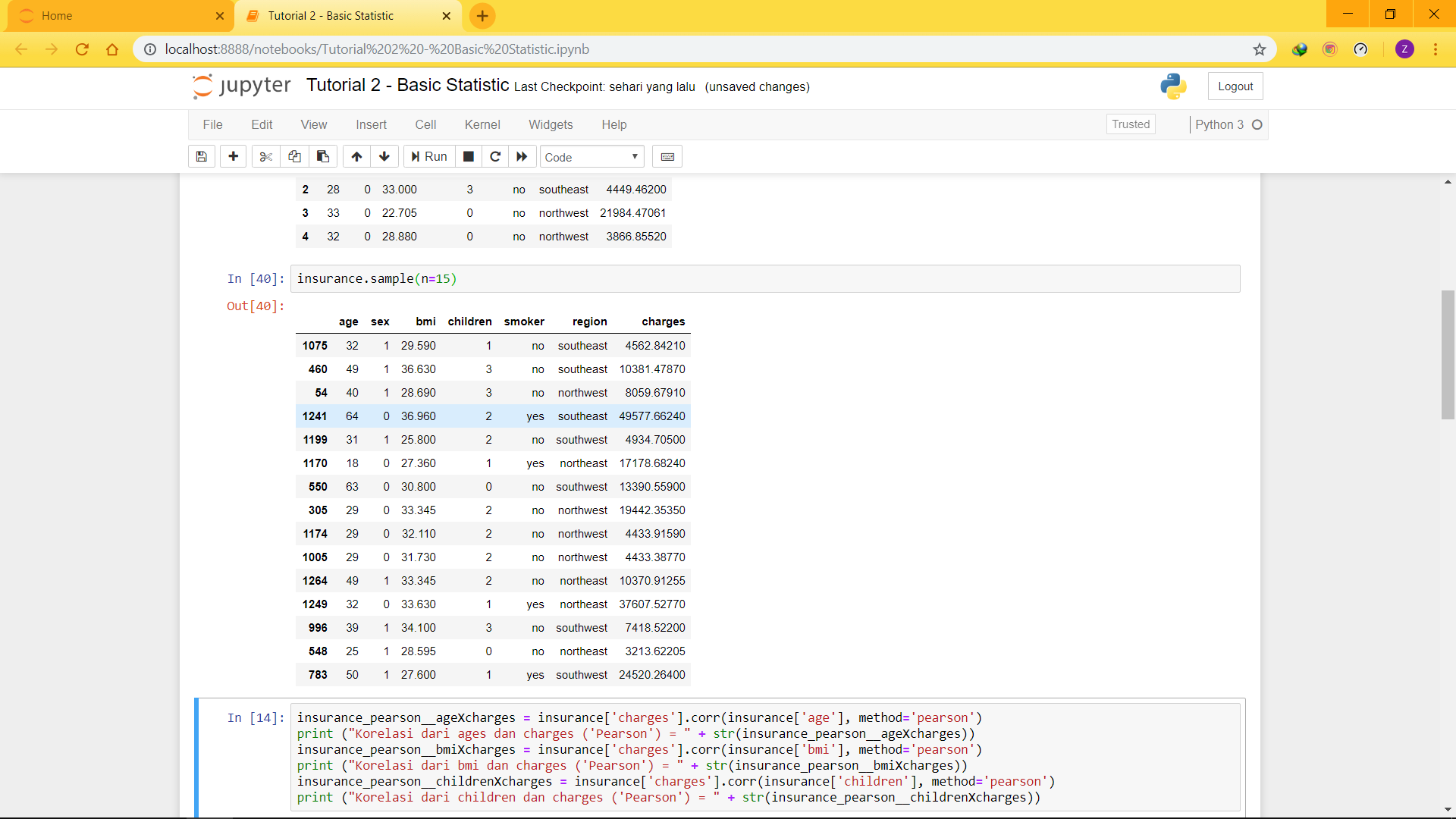


Dengan menggunakan command insurance.shape, saya mendapatkan bahwa terdapat 1338 Row dan 7 Column pada data insurance.csv

1. Mengubah semua kolom yang berisi kategorikal data menjadi numerik. Misal “male” menjadi 0, “female” menjadi 1. Simpan data hasil perubahan ini kedalam insurance\_modif.csv



1. Melakukan random sampling sederhana pada data Insurance dimana k = 15!

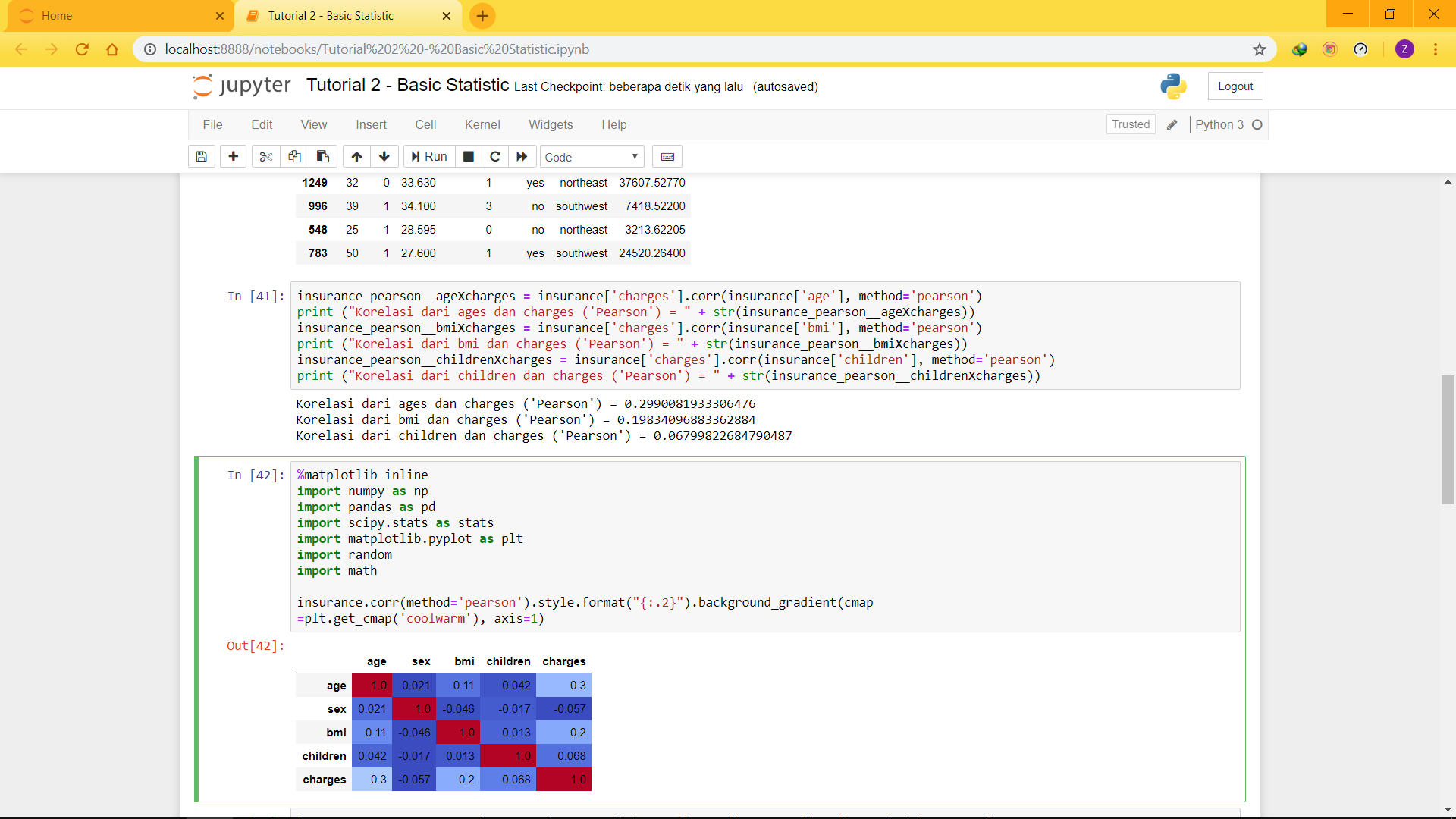


1. Melakukan uji korelasi spearman dan pearson

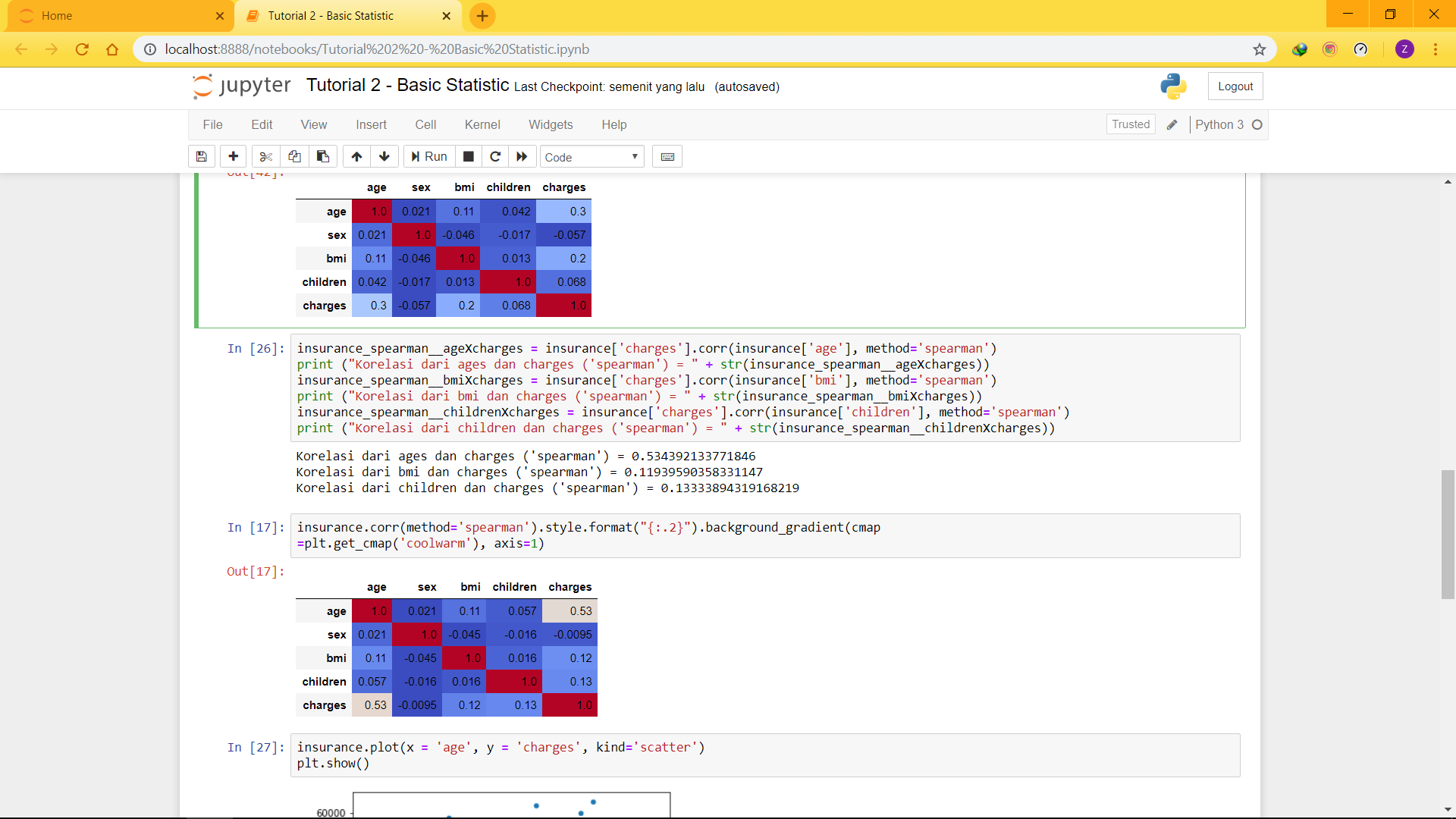
Variable yang saya uji korelasinya adalah pengaruh ‘*age’* terhadap ‘*charges’*, pengaruh ‘*bmi’* terhadap ‘*charges’* dan pengaruh ‘*children’* terhadap ‘*charges’*.

Saya memilih *age*, *bmi*, dan *children* karena saya merasa ketiga variable tersebut dapat mempengaruhi besaran biaya asuransi untuk seseorang (*charges*)

Uji korelasi dengan pearson:



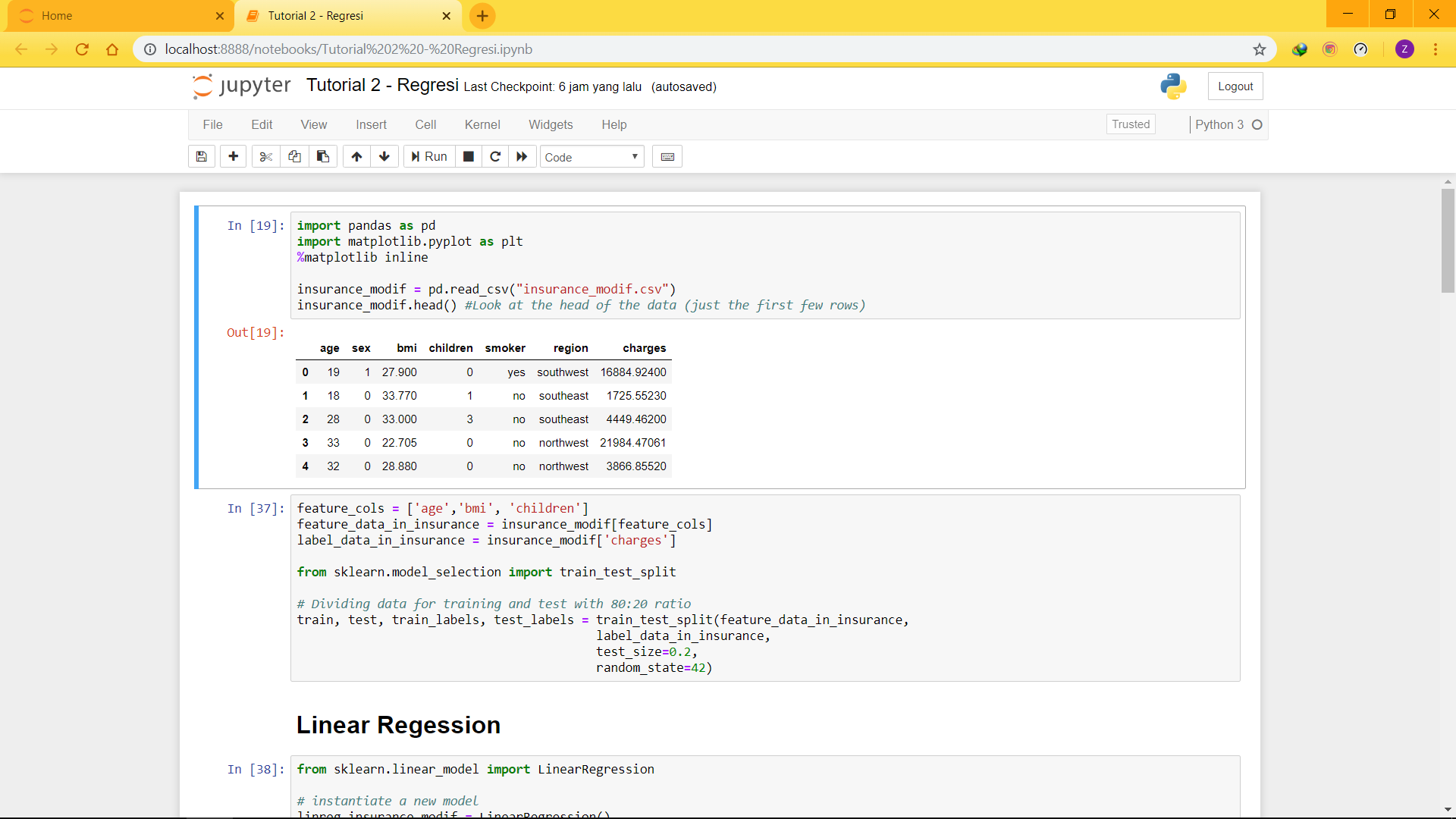
Uji korelasi dengan spearman:



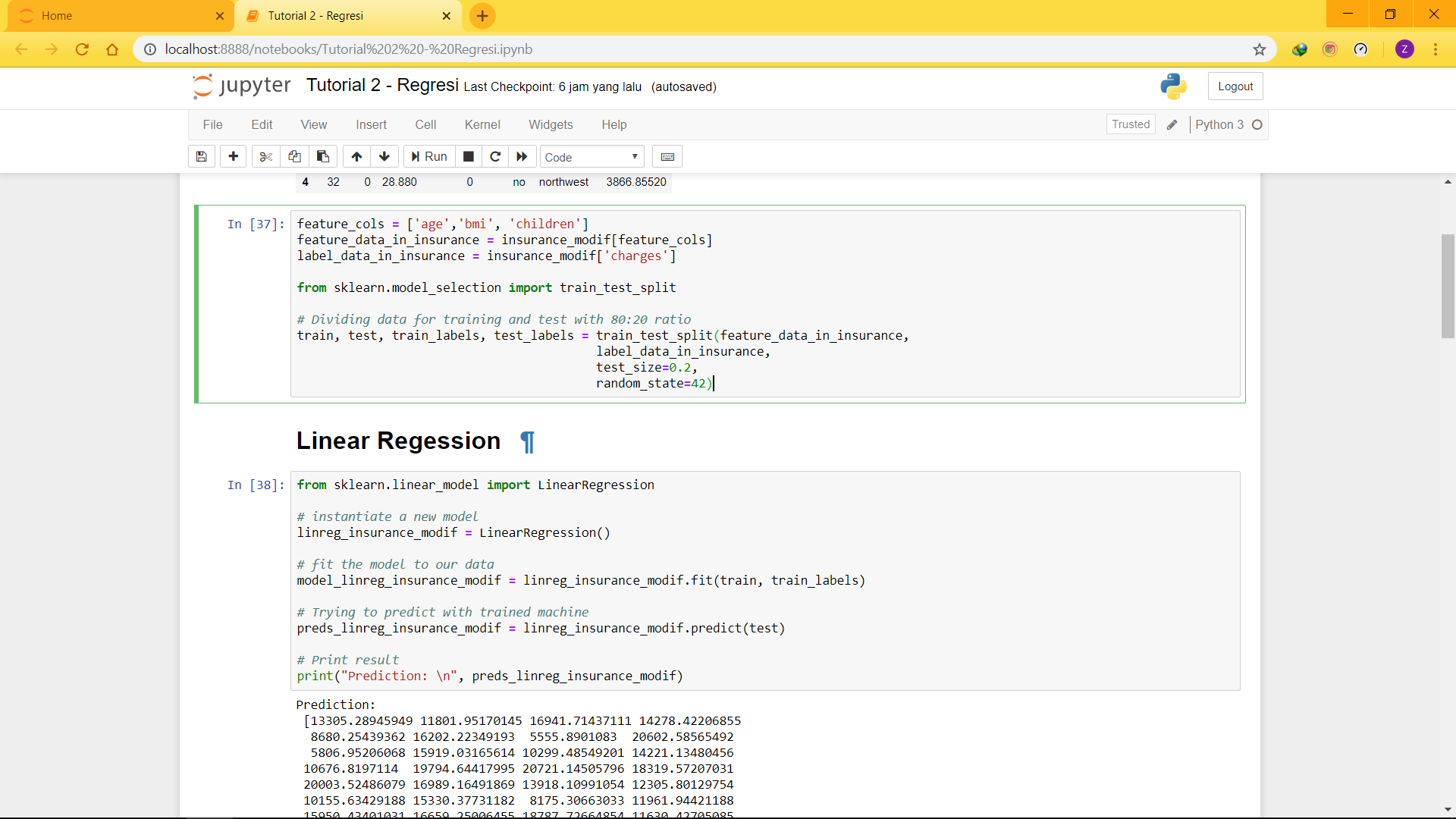
1. Dari hasil pengujian korelasi terhadap variabel *age*, *bmi*, dan *children*, saya dapat melihat bahwa ketiga variabel tersebut memiliki korelasi positif terhadap besaran biaya asuransi yang dikenakan kepada seseorang (*charges*). Itu artinya jika ada peningkatan umur seseorang maka kemungkinan besaran biaya asuransi juga akan naik. Hal itu juga berlaku untuk *bmi* dan jumlah anak (*children*), walaupun kenaikannya tidak signifikan.

**Regresi**

1. Menggunakan data insurance\_modiv.csv yang dihasilkan dari nomor 1

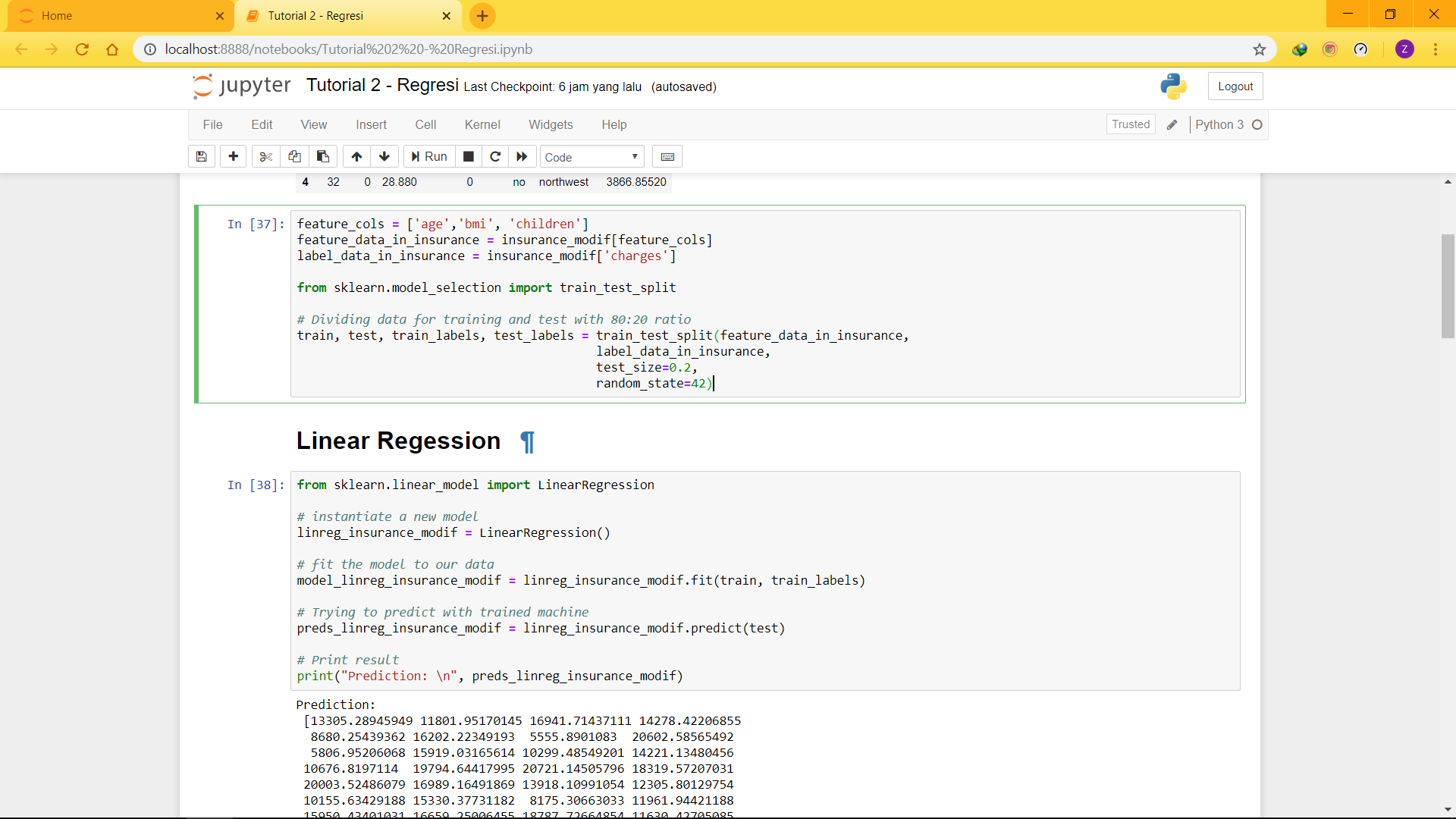


1. Memilih fitur yang digunakan sebagai model

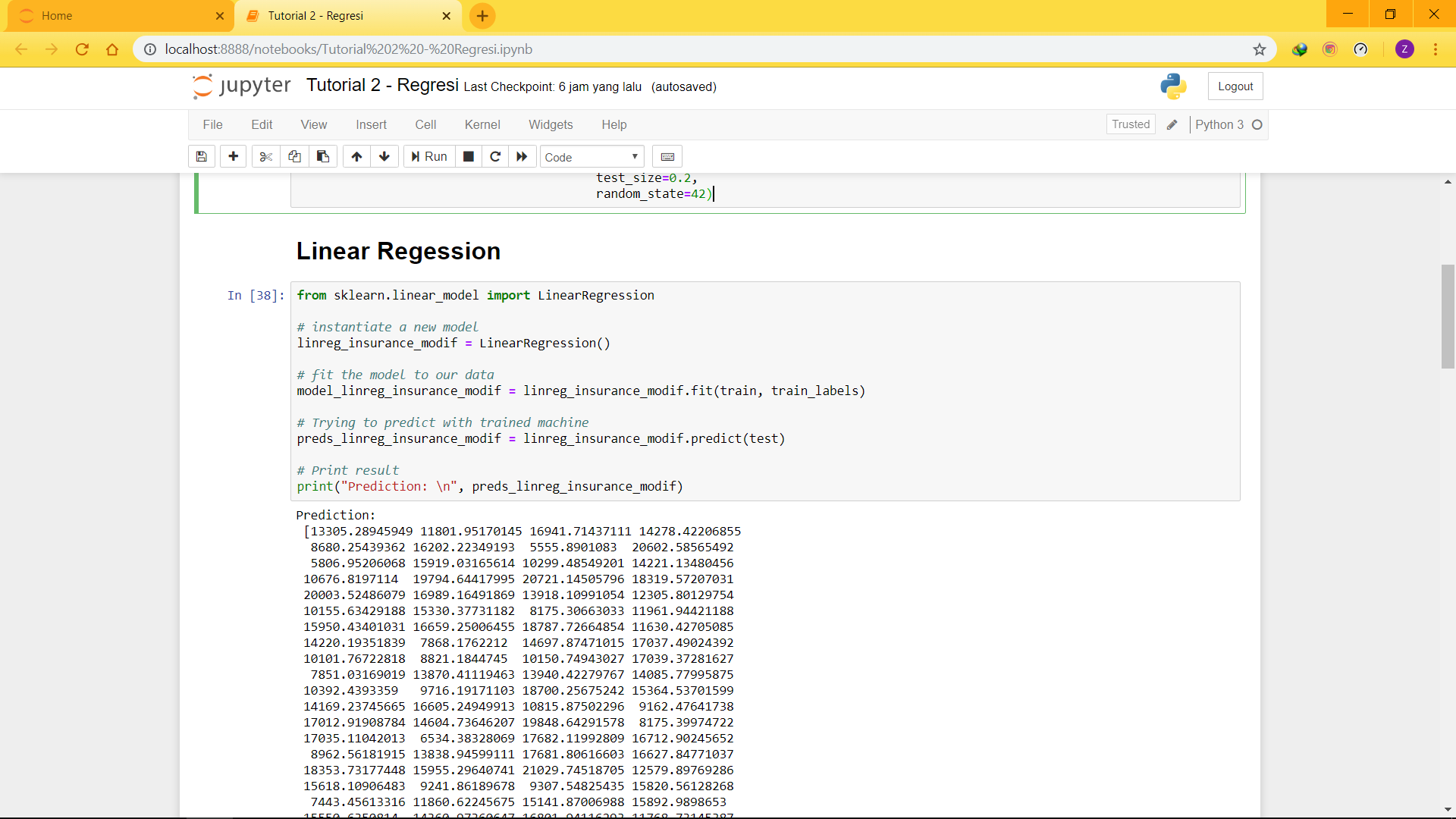


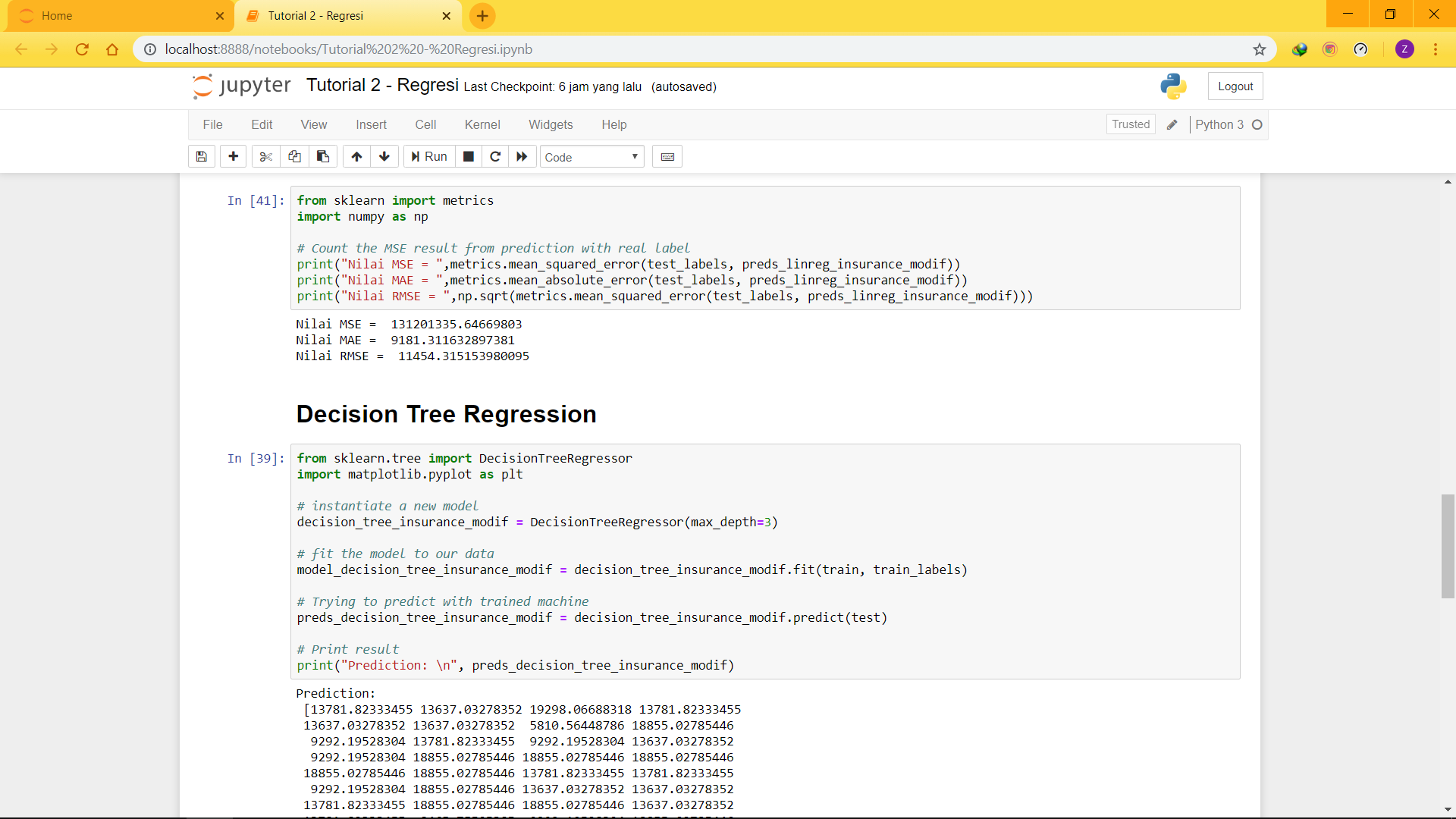
Saya memilih variabel *age*, *bmi*, dan *children*, karena nilai korelasinya yang cukup baik dengan variabel *charge*.

1. Menggunakan proporsi 80:20 untuk pembagian data training dan testing

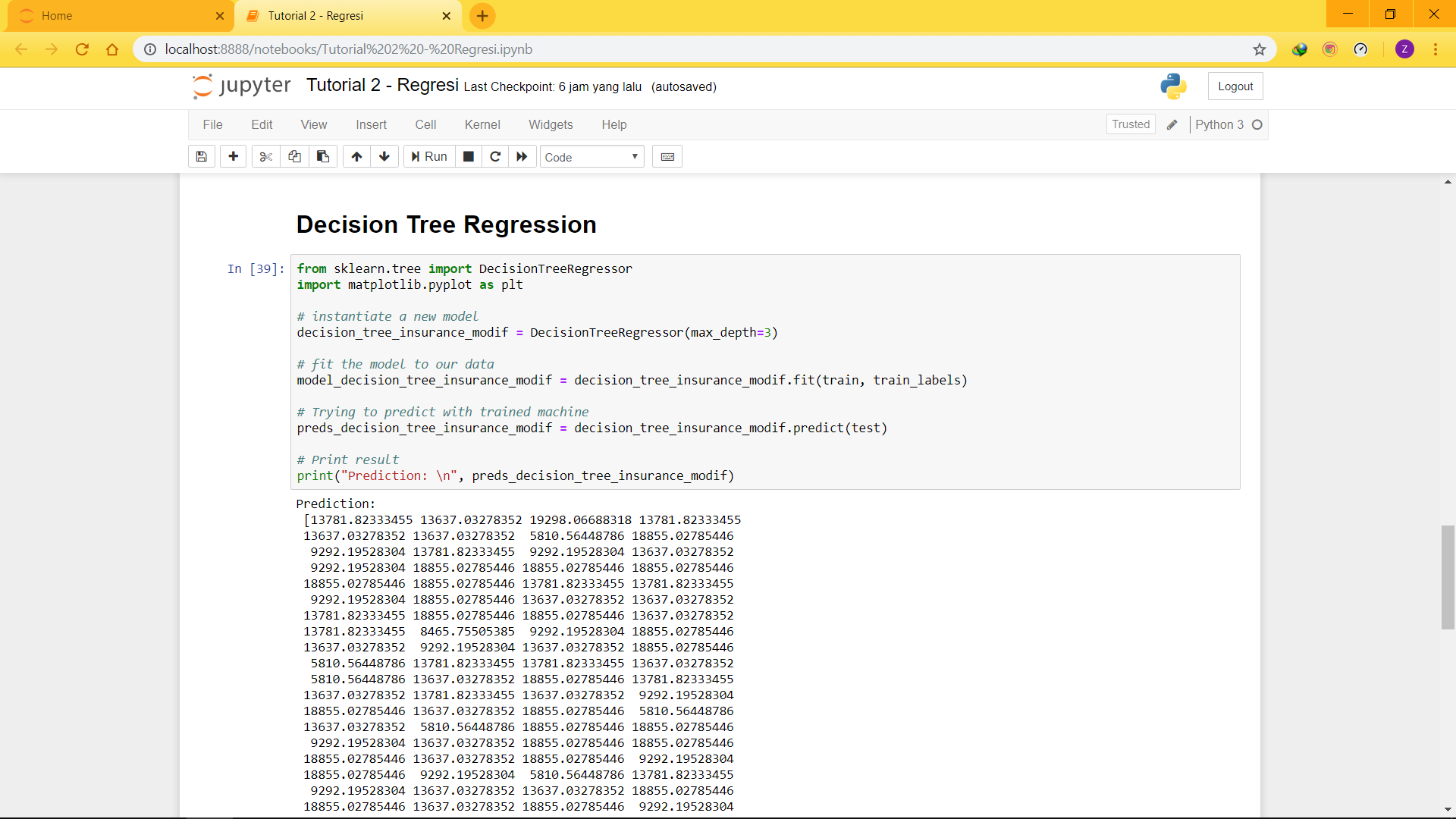


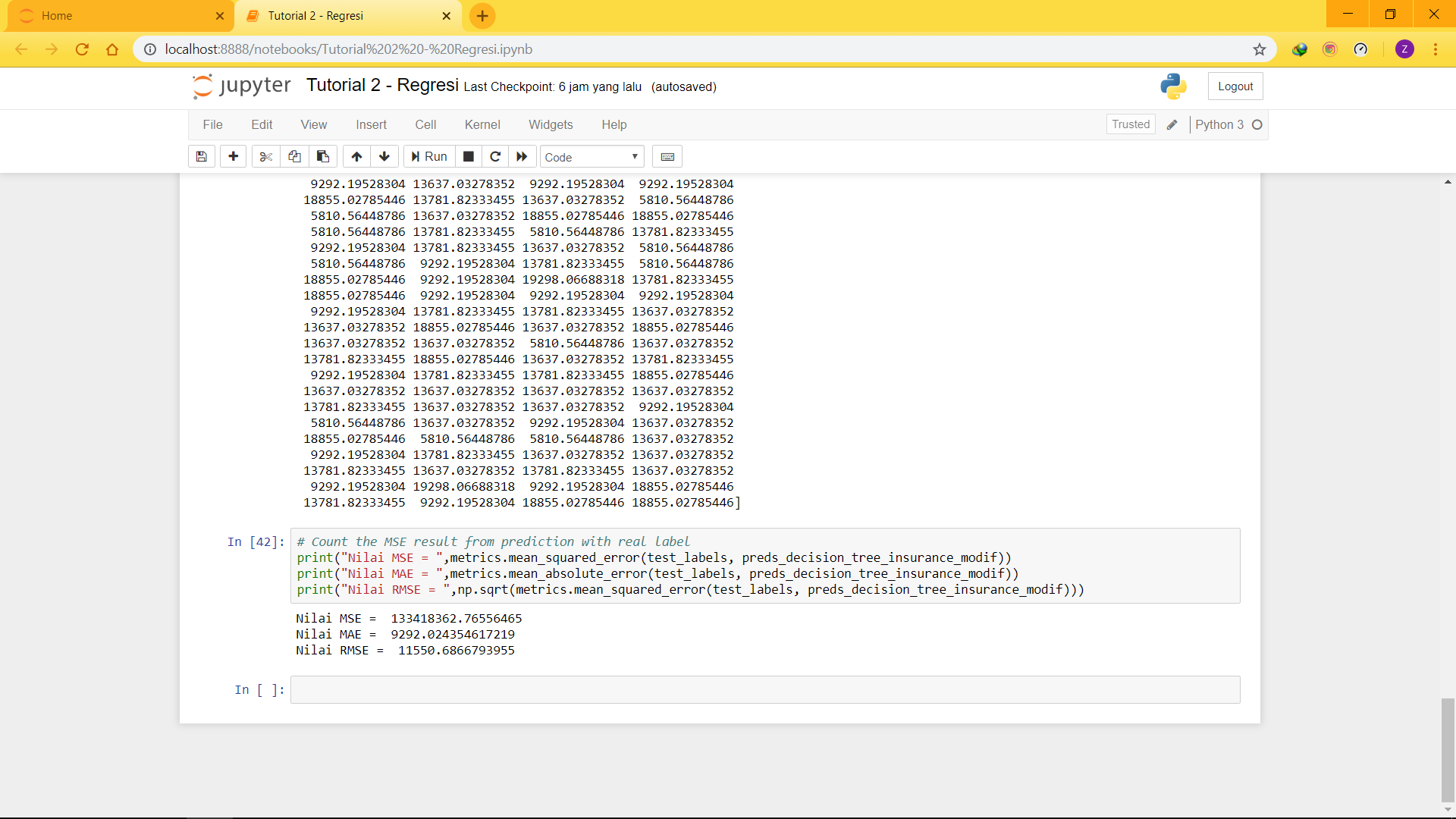
Menggunakan Linear Regression untuk memprediksi nilai *charges* (data hasil prediksi tidak saya tampilkan semua):





Menggunakan DecisionTree Regressor untuk memprediksi nilai *charges* (data hasil prediksi tidak saya tampilkan semua):

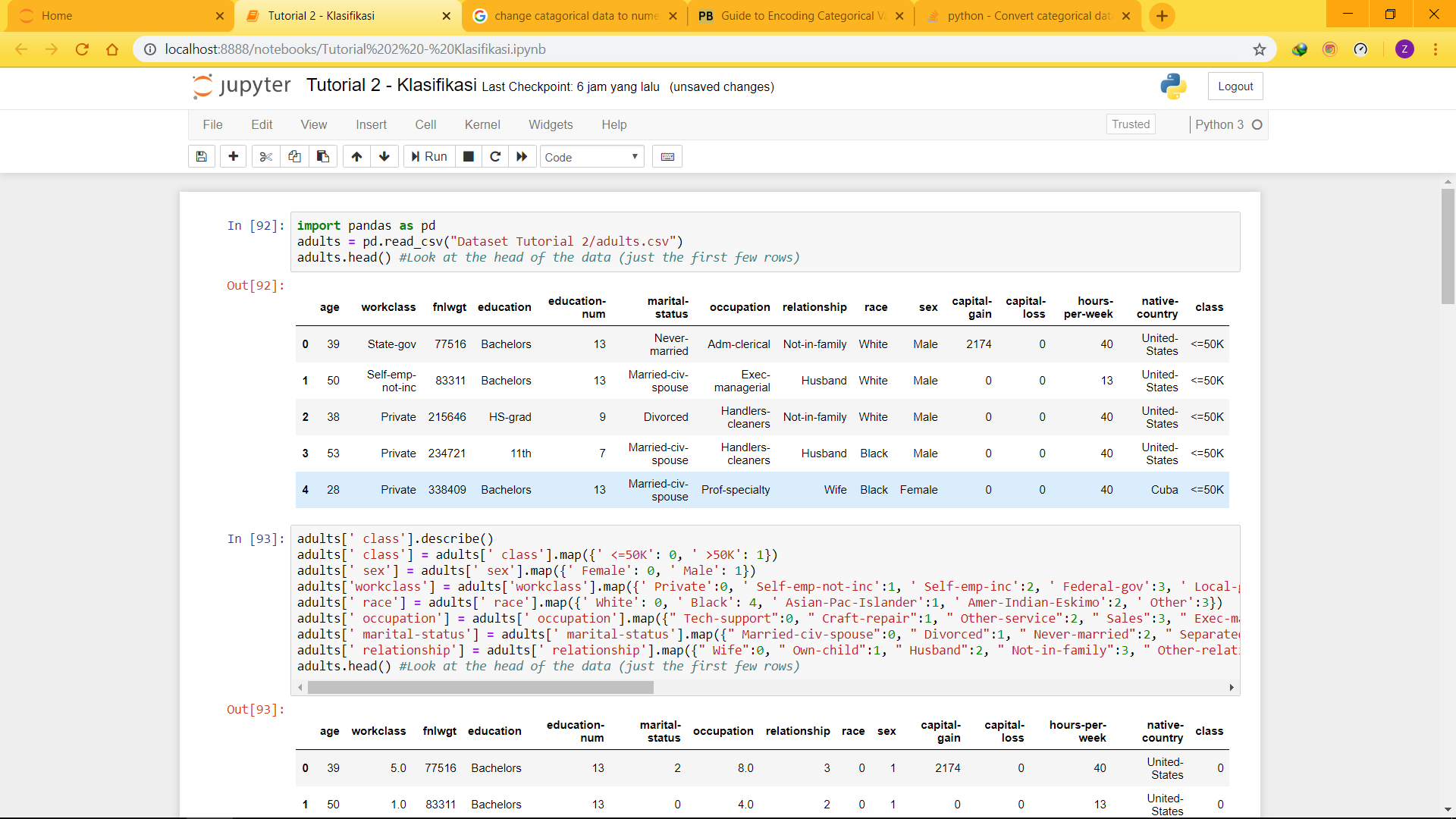


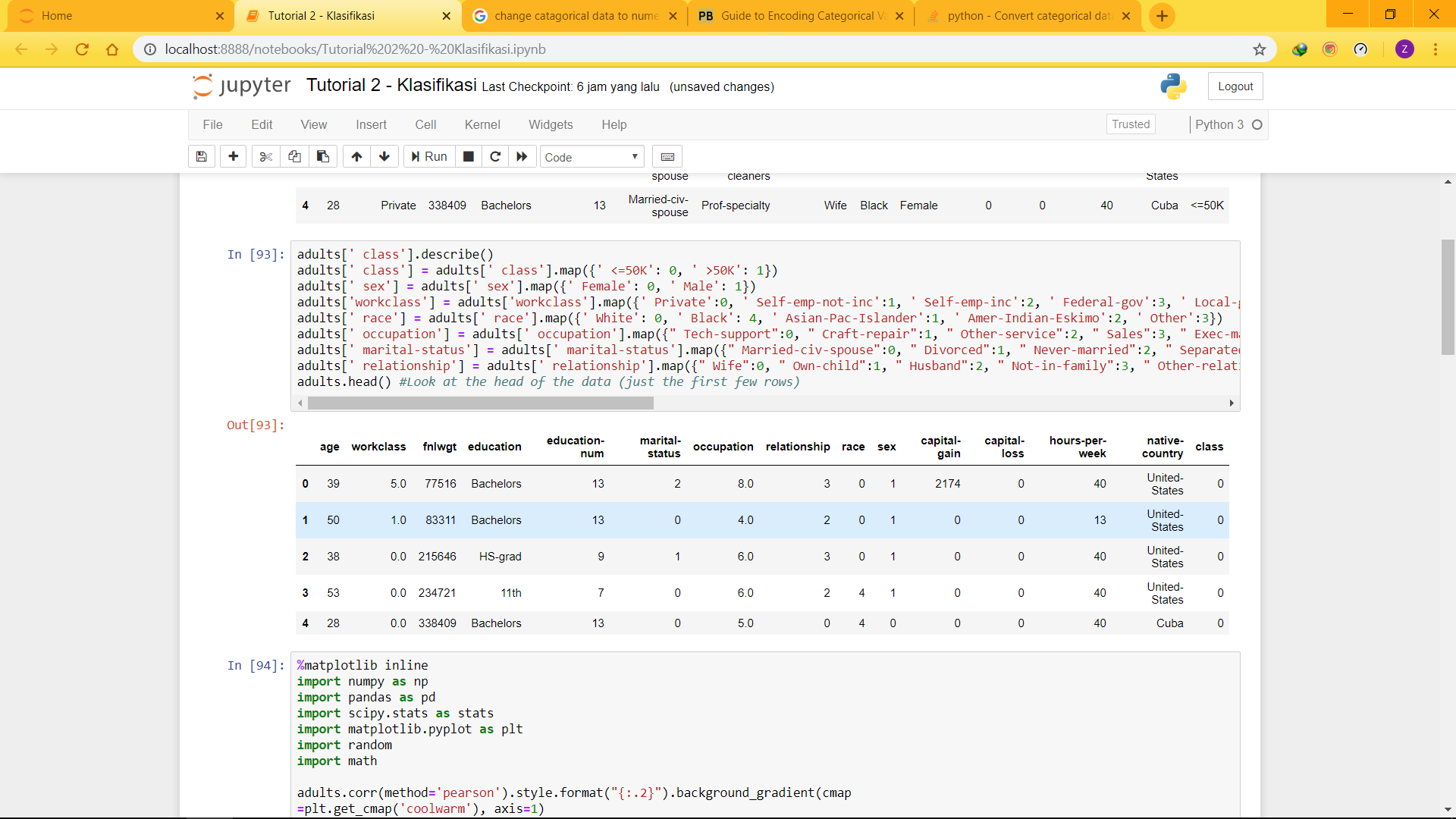


1. Dari hasil perhitungan menggunakan linear regression dan decision tree regressor, maka terlihat bahwa hasil MAE dari linear regression lebih kecil ketimbang menggunakan decision tree regressor.

**Klasifikasi**

1. menggunakan data adults.csv dan melakukan perubahan kolom yang berisi kategorikal data ke numerik





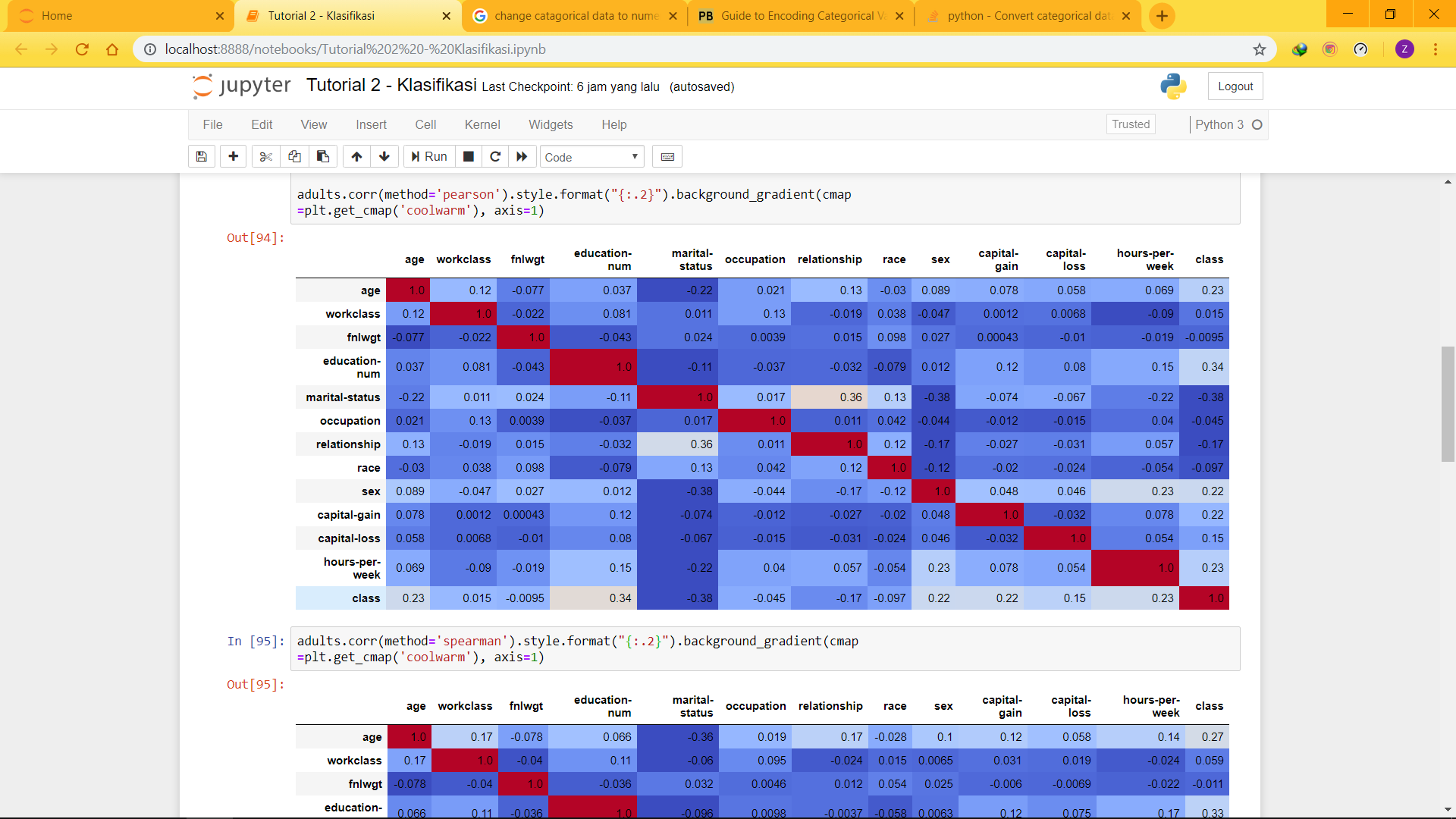
\*note: jika gambar kurang jelas dapat melihat Tutorial 2 - Klasifikasi.ipynb yang telah disertakan

1. mamiilih fitur yang akan digunakan sebagai model

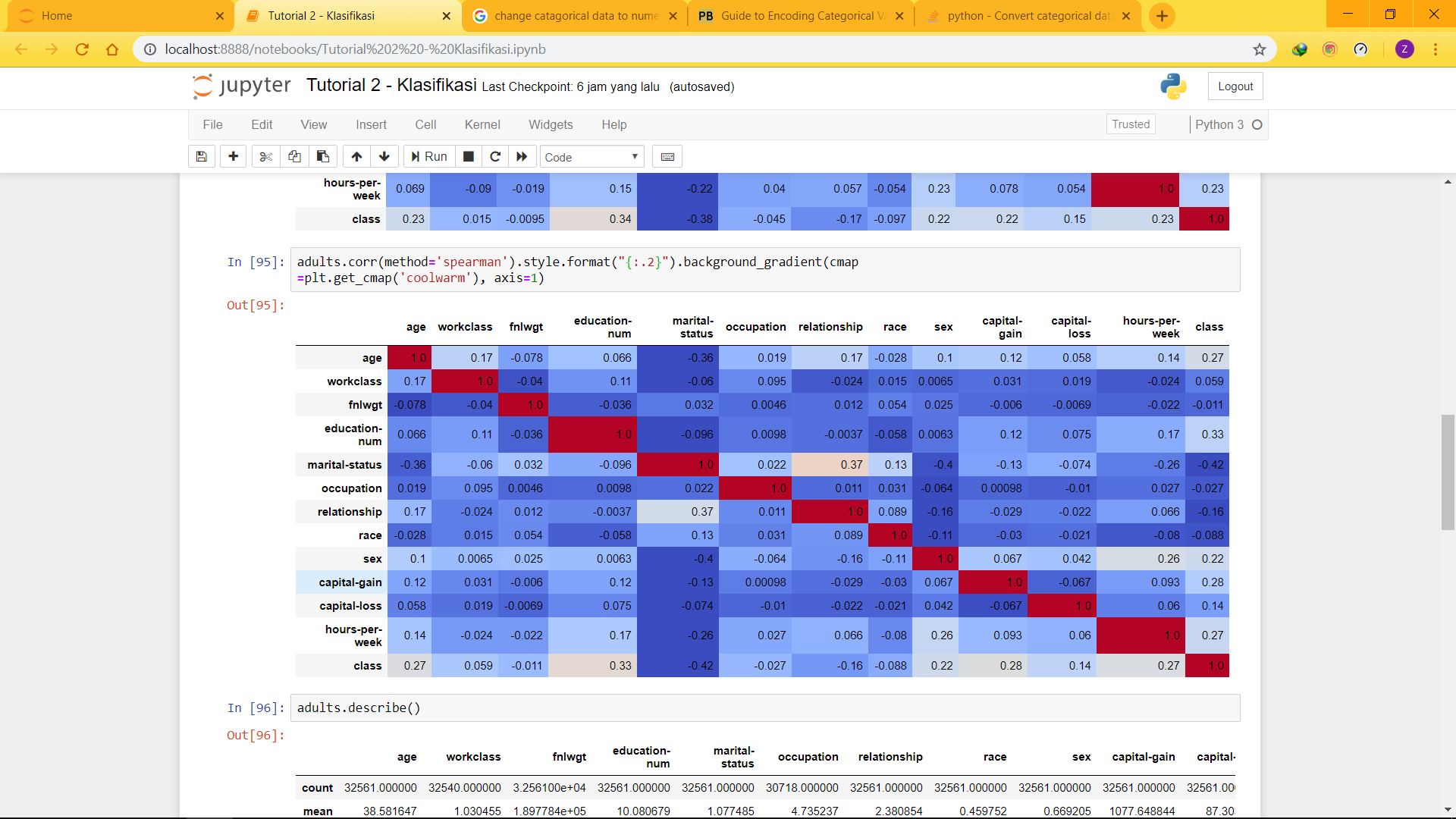
Saya memilih *age, education-num* (representasi dari *education*), *capital-gain,* dan *hours-per-week* sebagai vitur yang saya gunakan untuk mengklasifikasikan tingkatan pendapatan seseorang atau class(<=50K ataupun >50K)

Saya memilih keempat fitur tersebut karena memiliki korelasi yang cukup kuat dengan variabel class, hal itu terlihat dari perhitungan korelasi dengan menggunakan pearson dan spearman di bawah:

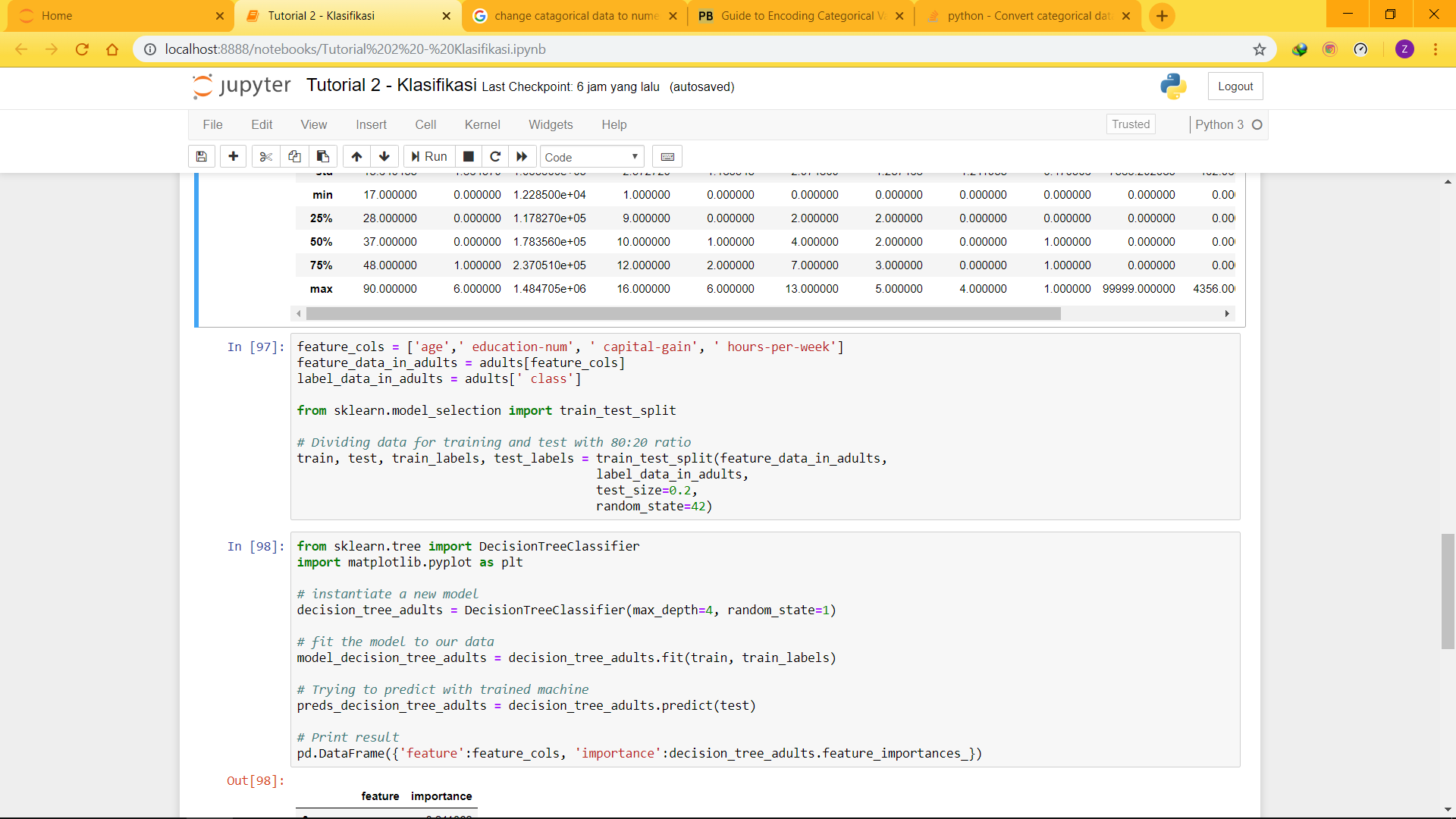
Pearson:



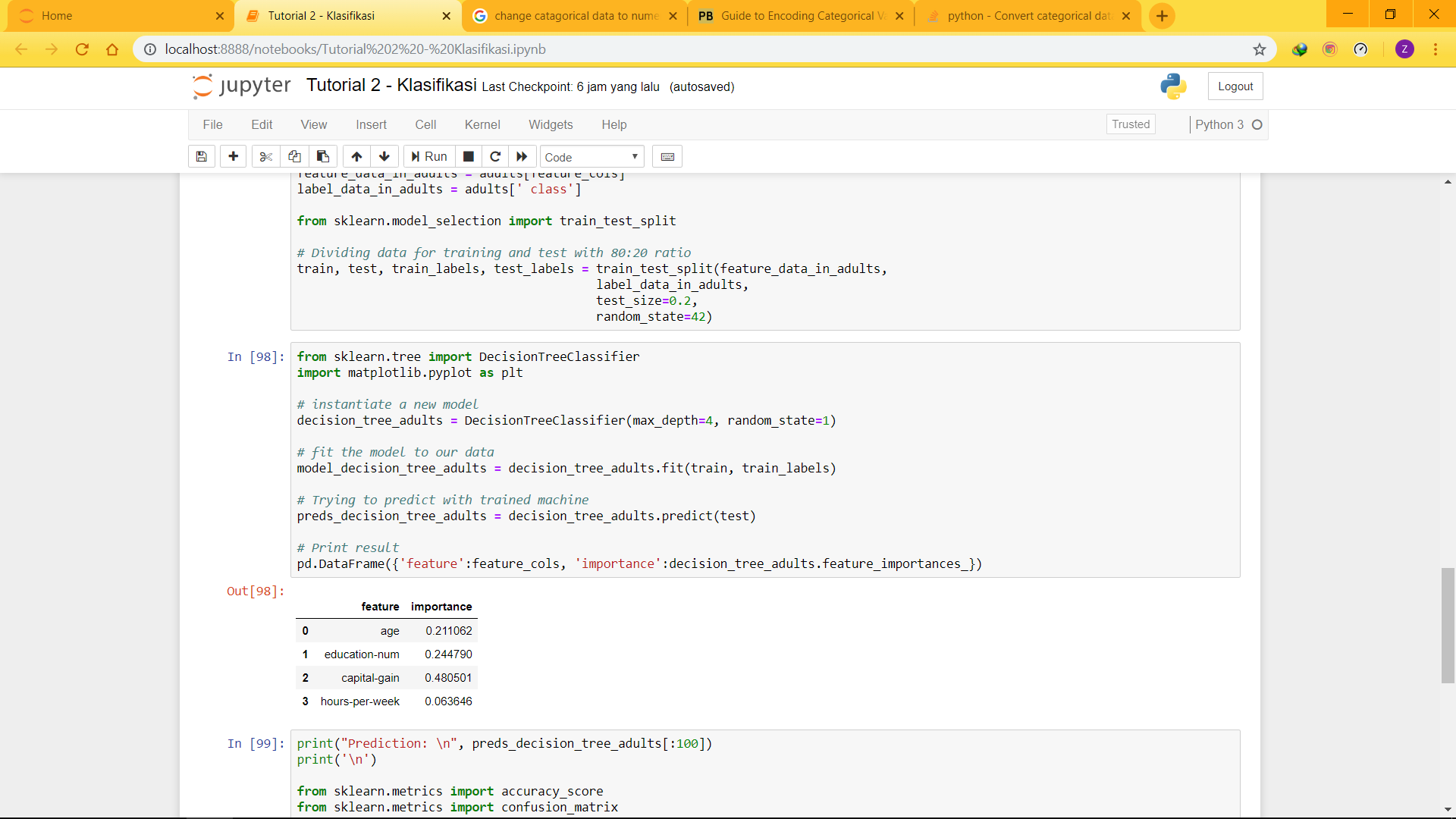
Spearman:



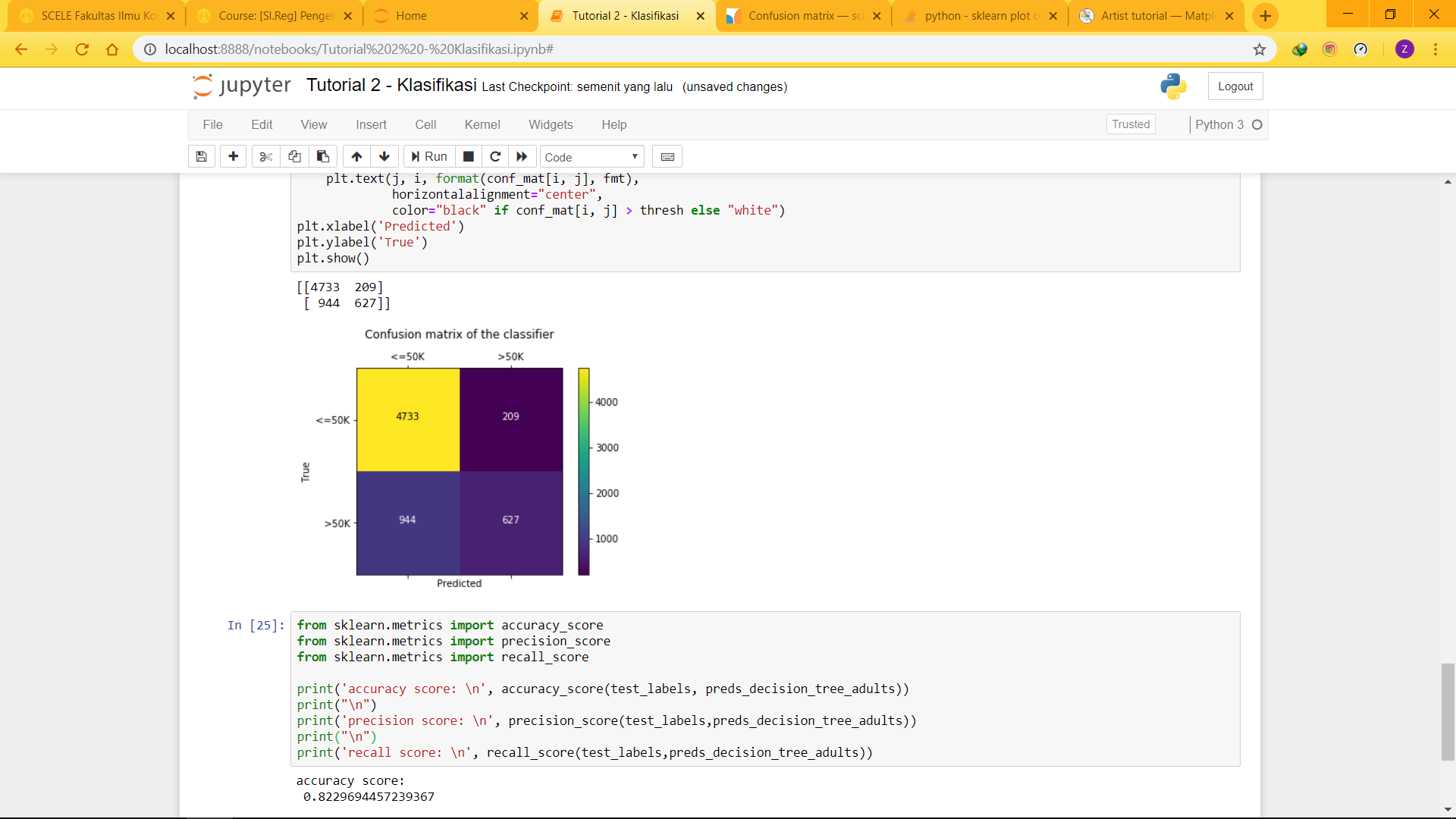
1. Menggunakan proporsi 80:20 untuk melakukan pembagian data training dan testing.



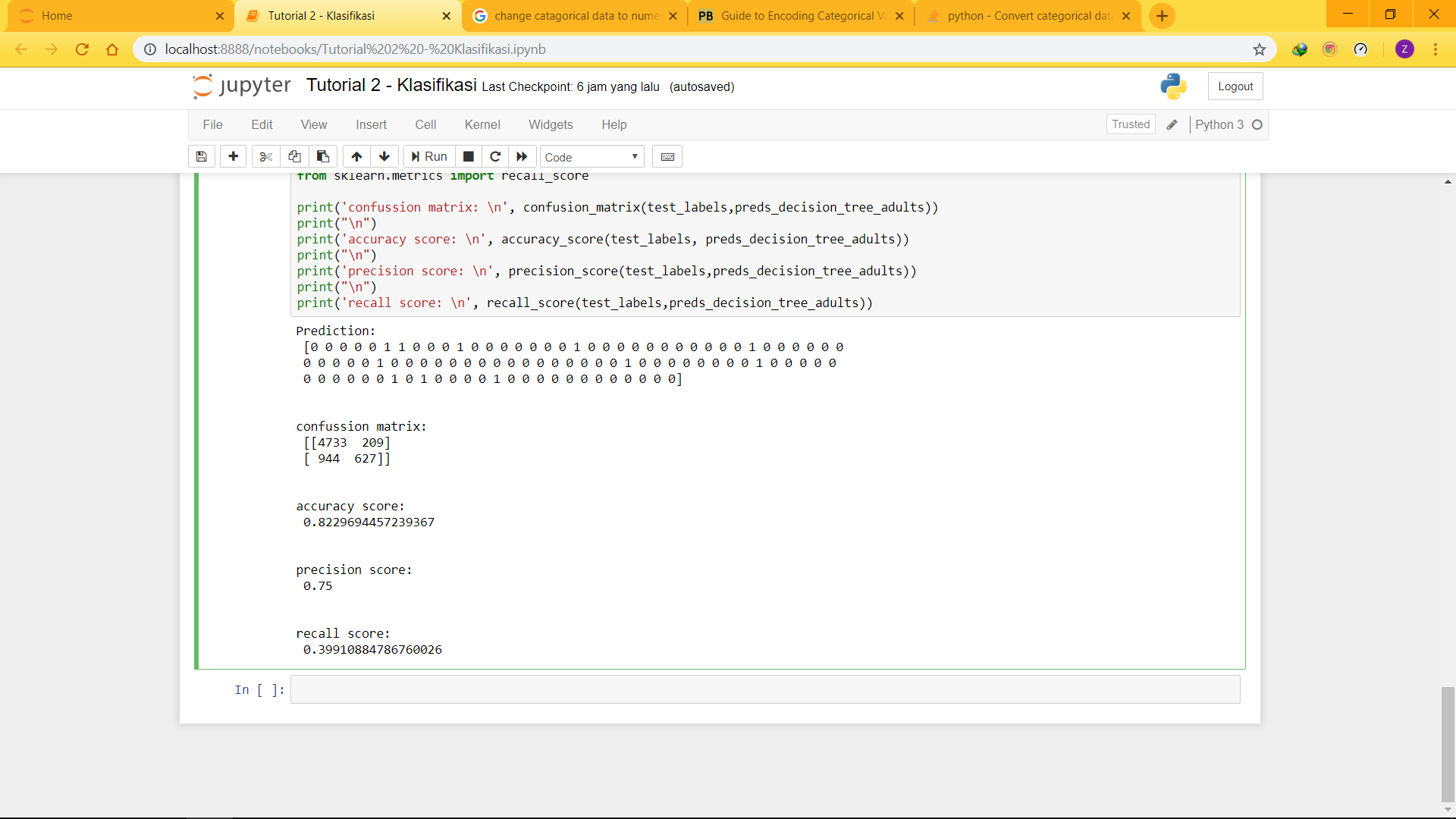
Menggunakan algoritma Decision Tree untuk melakukan pengelompokkan class di data tersebut.



1. Visualisasi dari confussion matrix hasil klasifikasi:

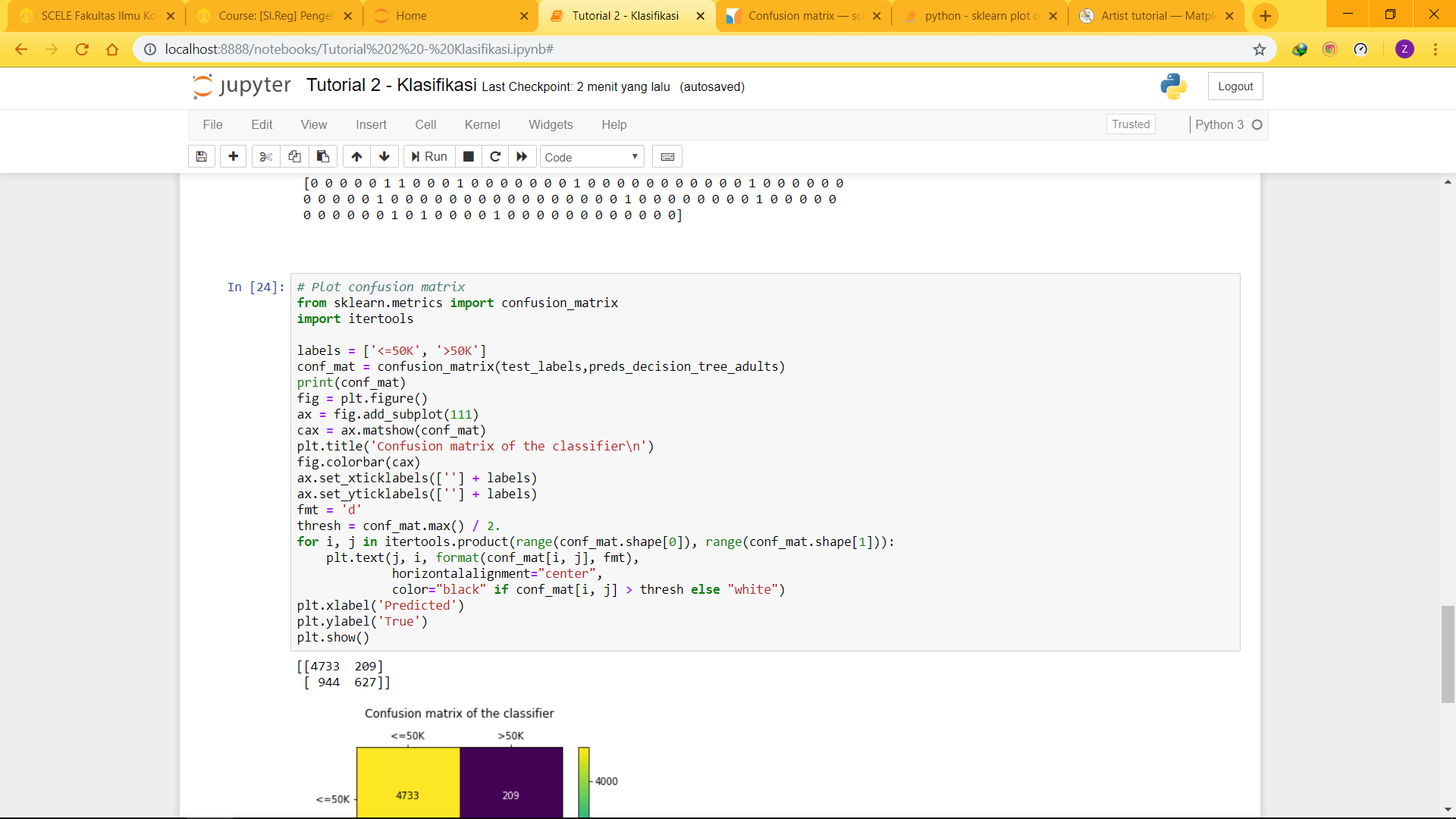


1. Akurasi, precision dan recall pada hasil klasifikasi model:

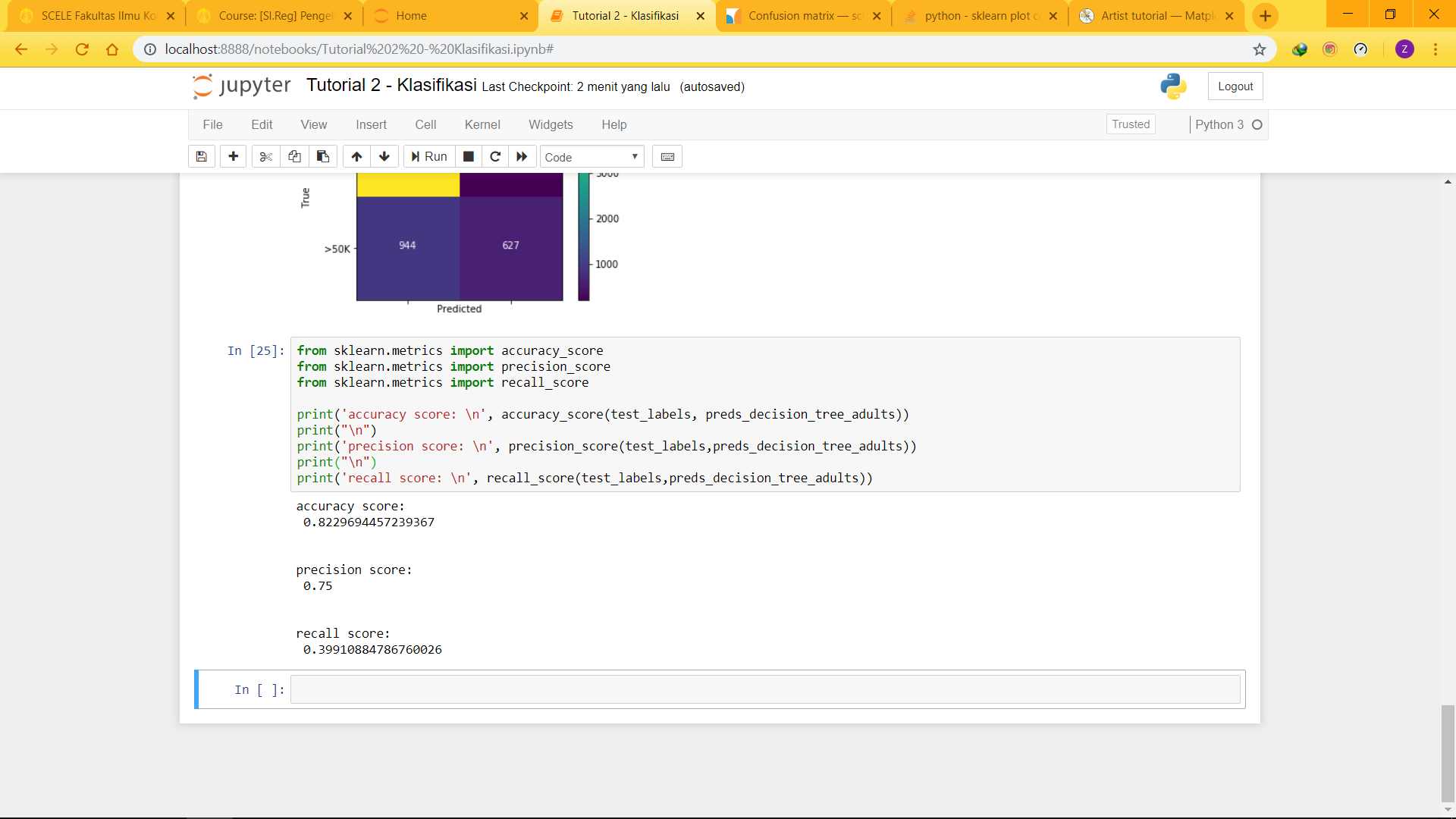


Code untuk menampilkan confussion matrix, akurasi, precision dan recall:

Confussion Matrix:



Akurasi, precision dan recall:



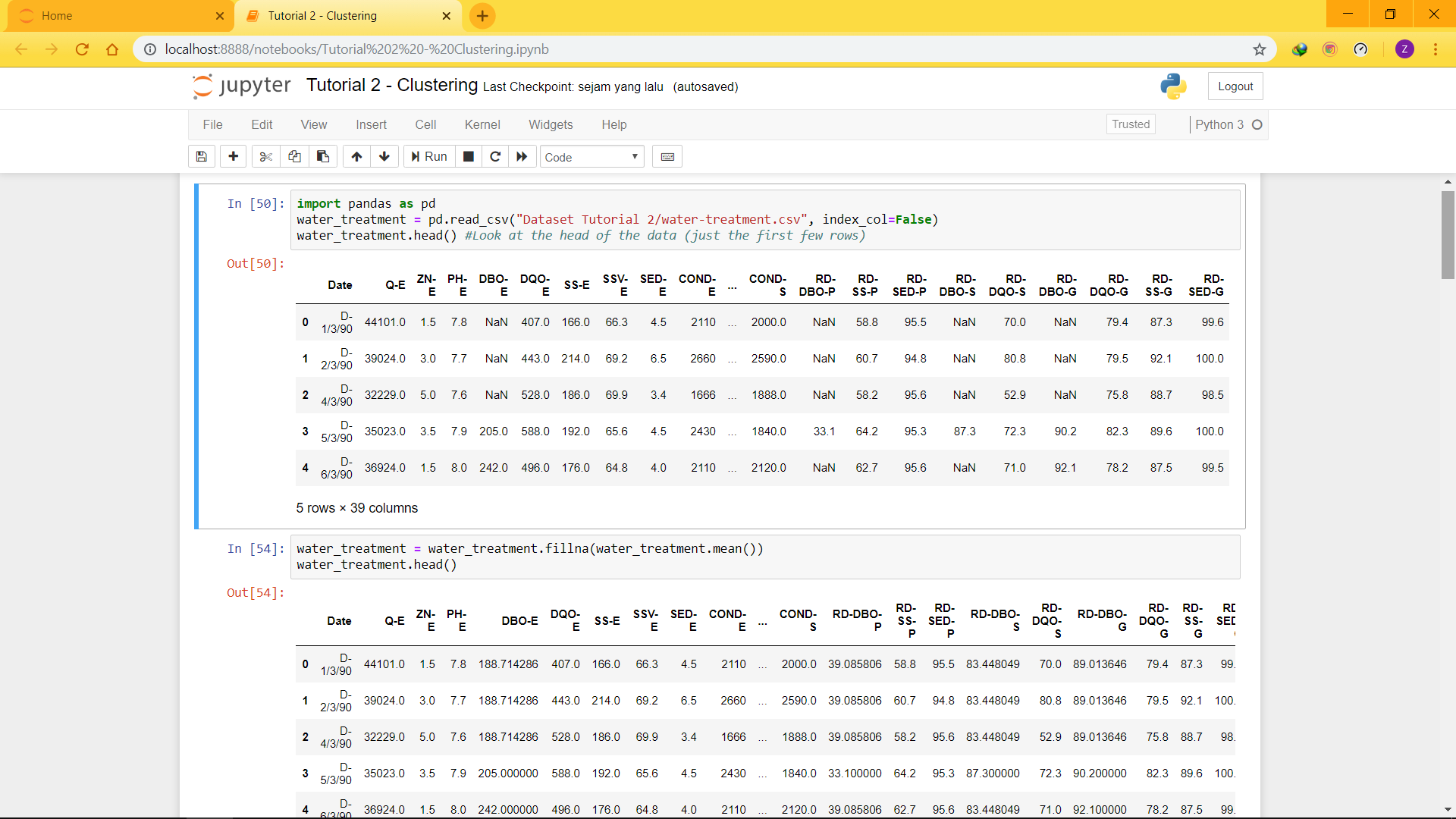
1. Penjelasan mengenai interpretasi hasil evaluasi dan error analysis.

Berdasarkan confussion matrix terlohat bahwa:

* Banyaknya data diprediksi termasuk dalam kelas <=50K dan kenyataannya benar termasuk kelas <=50K adalah sebanyak: 4733 data
* Banyaknya data diprediksi termasuk dalam kelas <=50K dan kenyataannya termasuk kelas >50K adalah sebanyak: 944 data
* Banyaknya data diprediksi termasuk dalam kelas >50K dan kenyataannya benar termasuk kelas >50K adalah sebanyak: 627 data
* Banyaknya data diprediksi termasuk dalam kelas >50K dan kenyataannya termasuk kelas <=50K adalah sebanyak: 209 data
* Dari sini dapat dilihat bahwa peluang perkiraan benar sekitar: (4733+627)/6513 = 0.82 atau 82%, sesuai dengan accuracy score

**Clustering**

1. Menggunakan data water-treatment.csv sebagai input data untuk clustering.

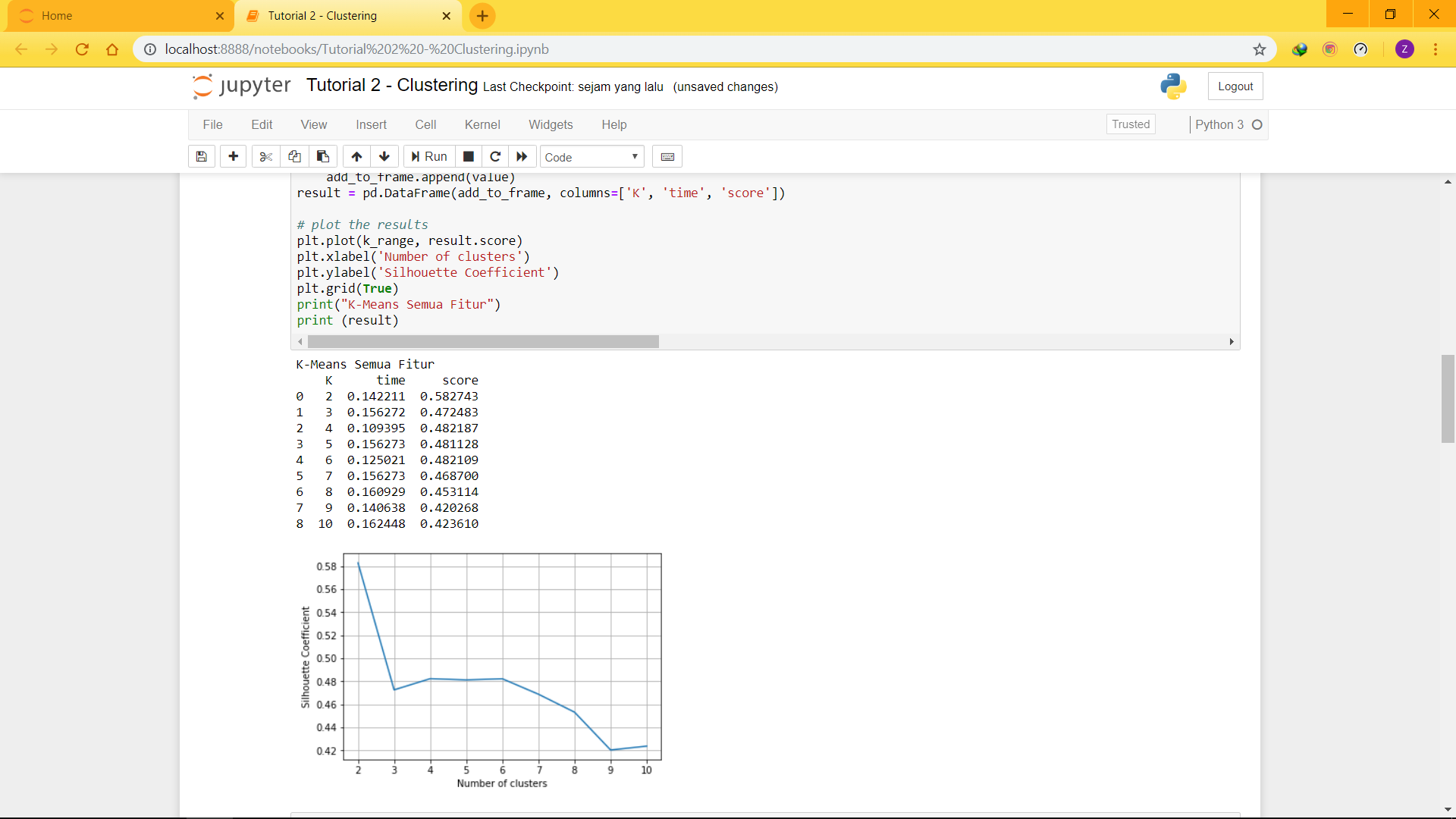


\*note: saya melakukan persiapan data (data preparation) dengan asumsi bahwa data dengan ‘?’ merupakan data yang tidak diketahui nilainya dan saya ganti dengan menggunakan rata2 dari kolom yang bersangkutan (jia ‘?’ berada kolom Q-E maka nilainya saya ganti dengan rata-rata nilai Q-E)

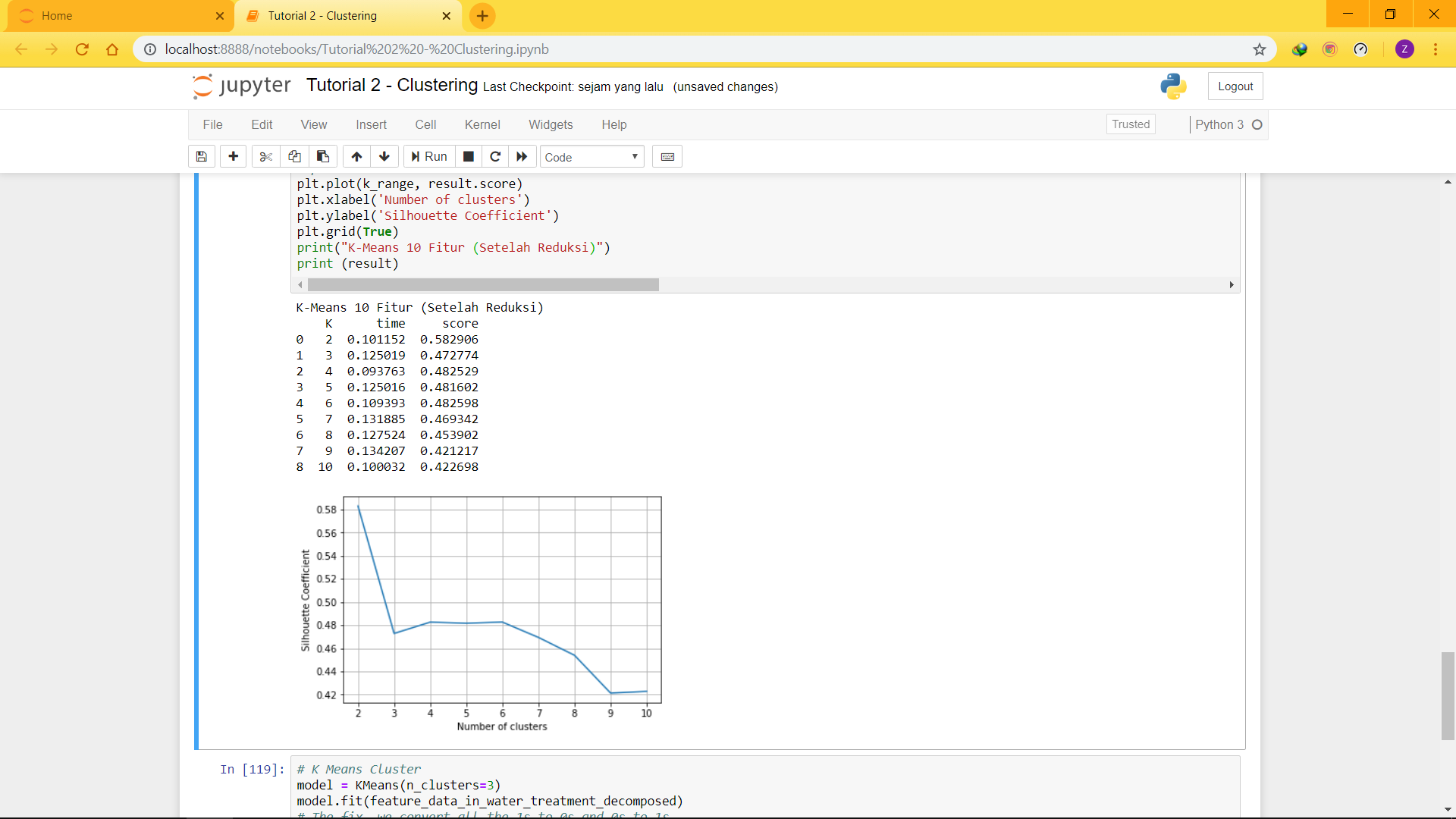
1. Apakah anda memerlukan reduksi dimensi pada data tersebut?, Jika Iya berapa jumlah fitur yang anda gunakan dan berapa jumlah fitur yang anda hilangkan?

Saya melakukan reduksi dengan menggunakan feature extraction dengan menggunakan modul PCA, dengan begitu maka tidak ada fitur yang dihilangkan melainkan saya menggabungkannya sehingga total hanya ada 10 fitur dari 38 fitur awal

1. Hasil K-Means semua fitur:



Hasil K-Means 10 fitur



1. Dari perbandingan kedua tabel hasil perhitungan K-Means untuk model diatas terlihat bahwa jika kita melakukan feature extraction maka waktu yang kita butuhkan akan lebih cepat ketimbang menggunakan semua fitur yang ada namun tidak begitu memiliki pengaruh yang berarti pada nilai Sihoulette Coeficientnya, sehingga dapat disimpulkan bahwa menggunakan feature extraction lebih efisien ketimbang menggunakan seluruh fitur yang ada.