1. Pendahuluan

# sorry berantakan. Pake libre. Bukan Tetex #

Abstrak : Feature Scaling adalah metode yang digunakan untuk normalisasi rentang dari variabel independen atau data dari suatu fitur. Pada pemrosesan data dikenal sebagai Normalisasi dan umumnya dilakukan pada pra-proses data. MinMax Scaling merupakan metode paling sederhana dalam melakukan scaling sebuah variabel menjadi rentang tertentu. Penggunaan Scaling pada algoritma Naive Bayes secara teori tidak memiliki dampak yang signifikan, namun pada pengujian didapatkan hasil bahwa Penggunaan Scaling menghasilkan waktu Training lebih cepat 45%, akurasi lebih tinggi sebesar 11% dan presisi lebih besar 25% dibanding penggunaan data tanpa Scaling pada algoritma yang sama pada saaat melakukan CrossValidation.

1.1 Latar Belakang

Dataset ini berasal dari National Institute of Diabetes and DIgestive and Kidney Diseases. PIMA INDIAN Diabetes adalah penduduk asli amerika hidup di Arizona. Kecenderungan genetik yang dialami memungkinkan kelompok ini untuk bertahan hidup secara normal untuk diet karbohidrat yang buruk selama bertahun-tahun. Beberapa tahun ini, dikarenakan transisi yang mendadak dari makanan pertanian tradisional menjadi makanan olahan, bersamaan dengan menurunnya aktivitas fisik, membuat mereka menjadi frekuensi tertinggi yang mengalami diabetes tipe 2 dan karena alasan ini mereka mejadi subjek dari berbagai macam studi.

Dataset berisi berbagai variabel prediksi medis dan satu variabel target "Outcome", jika output 1 maka diabetes, jika 0 maka tidak diabetes. Variabel prediksi medis berisi jumlah kehamilan yang dialami pasien, BMI mereka, level insulin, umur, dan seterusnya.

Dataset ini mengandung nilai yang sangat bervariasi pada fitur Glukosa, memiliki nilai minimum 0, nilai maksimum 199, dan variance 1022. Sedangkan metode Naive Bayes Gaussian secara teori tidak memiliki pengaruh signifikan pada feature scaling karena Naive Bayes secara desain dapat menangani algoritma seperti Linear Discriminant Analysis.

Scaling memiliki pengaruh yang signifikan pada algoritma yang sensitif pada jarak seperti KNN, karna itu data harus dilakukan scale agar tiap fitur memiliki bobot yang equal.

1.2 Tujuan

Untuk mencari dampak dari penggunaan "Scaling" pada waktu eksekusi, akurasi prediksi, preisisi dari model pada Gaussian Naive Bayes dan Support Vector Machine Support Vector Classification.

2. Metode Scaling

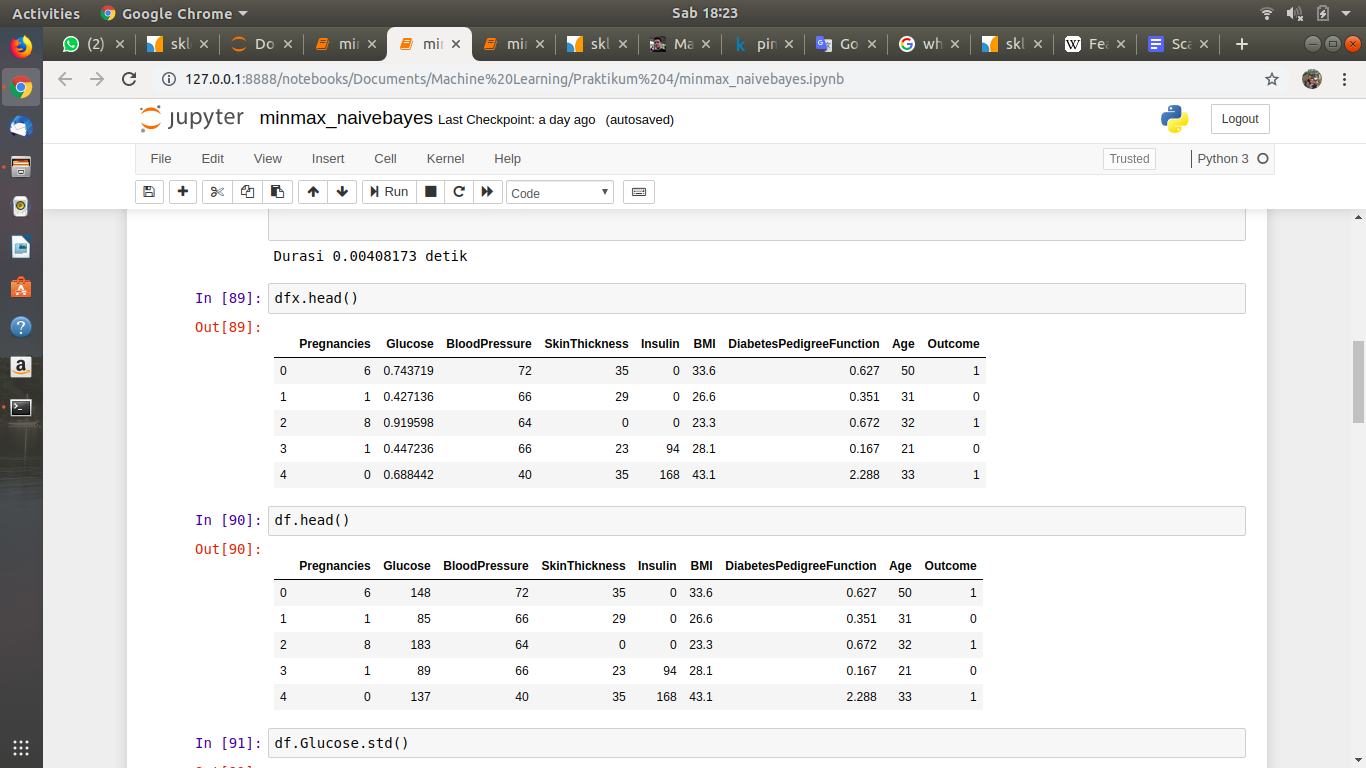
Juga dikenal sebagai min-max normalisasi, adalah metode paling sederhana untuk melakukan scaling. Mengubah features dengan melakukan scaling pada tiap instances menjadi range nilai tertentu sepert misalnyai 0 hingga 1.



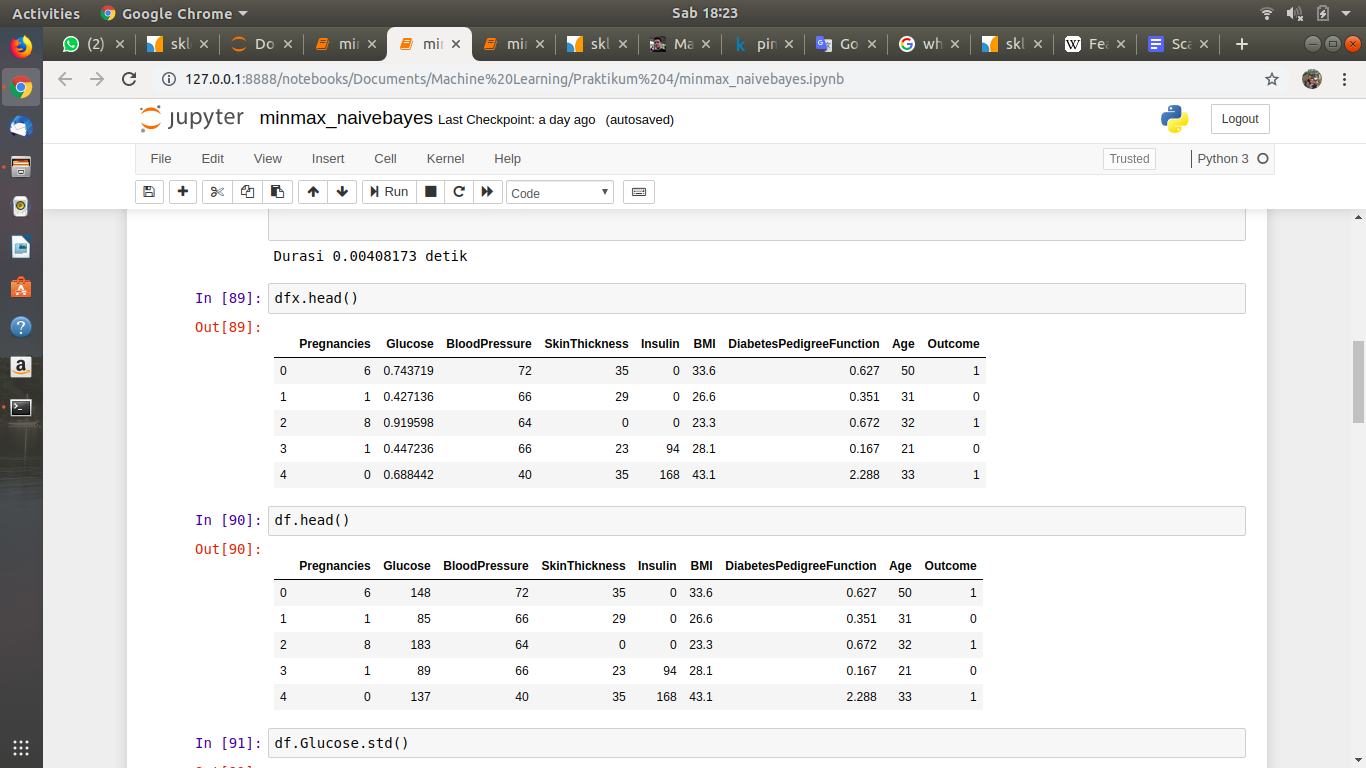
Dimana  
a : range bawah   
b : range atas  
X : value dari suatu features.   
min(X) : nilai minimum dari features X  
max(X) : nilai maksimum dari features X

3. Penerapan

Data PIMA INDIAN sebelum diterapkan MINMAX SCALER

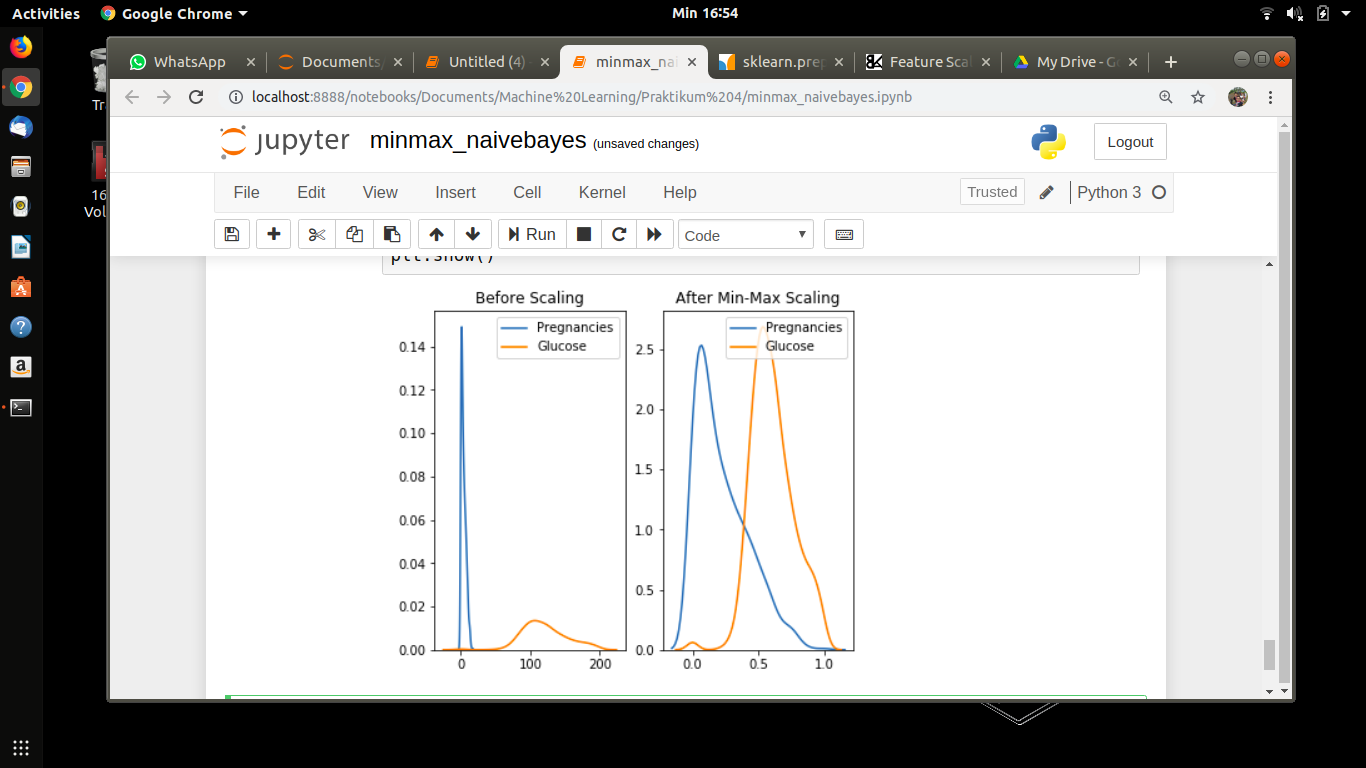


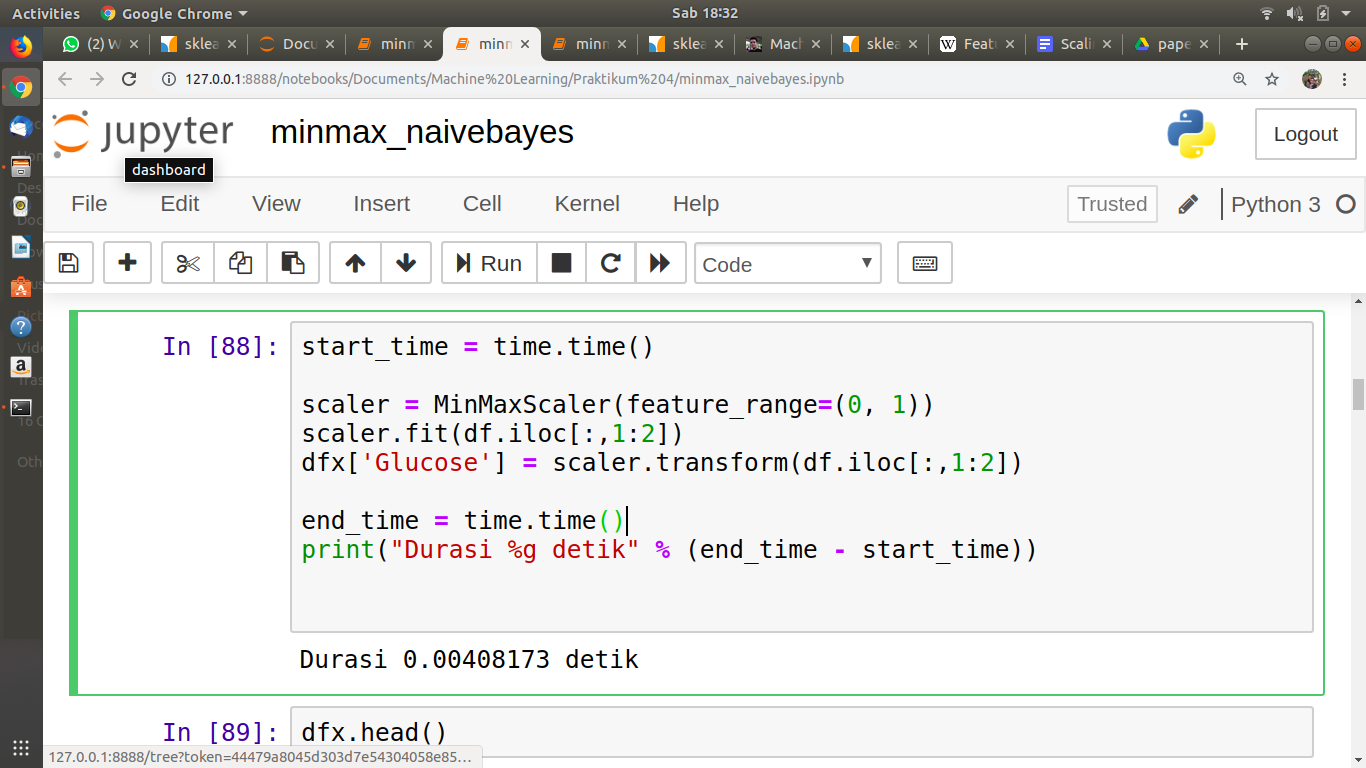
Setelah diterapkan MINMAX SCALER pada features Glucose



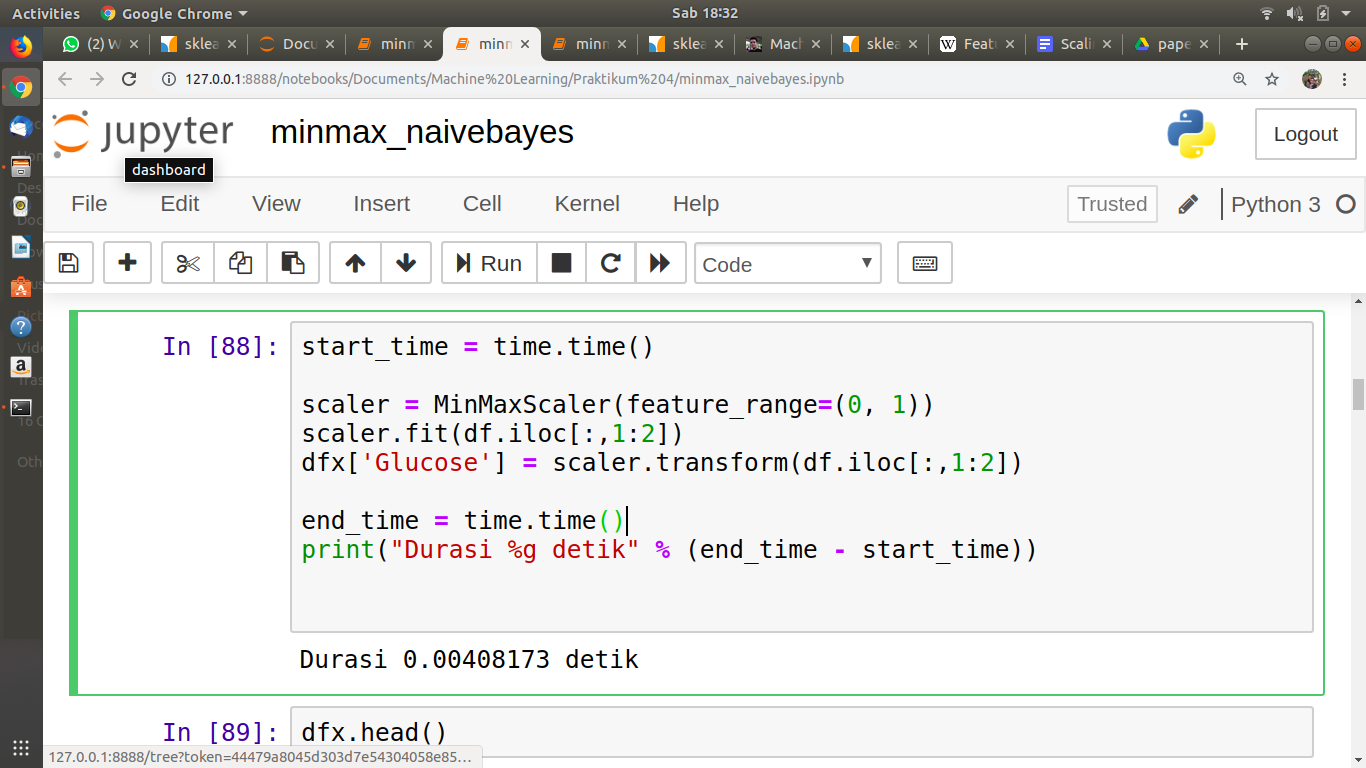
Penerapan menggunakan modul MinMaxScaler dari library sklearn.preprocessing.

Visualisasi data setelah diberlakukan Scaling pada features Glucose dan Pregnancies.





Durasi yang diperlukan untuk transformasi features Glucose

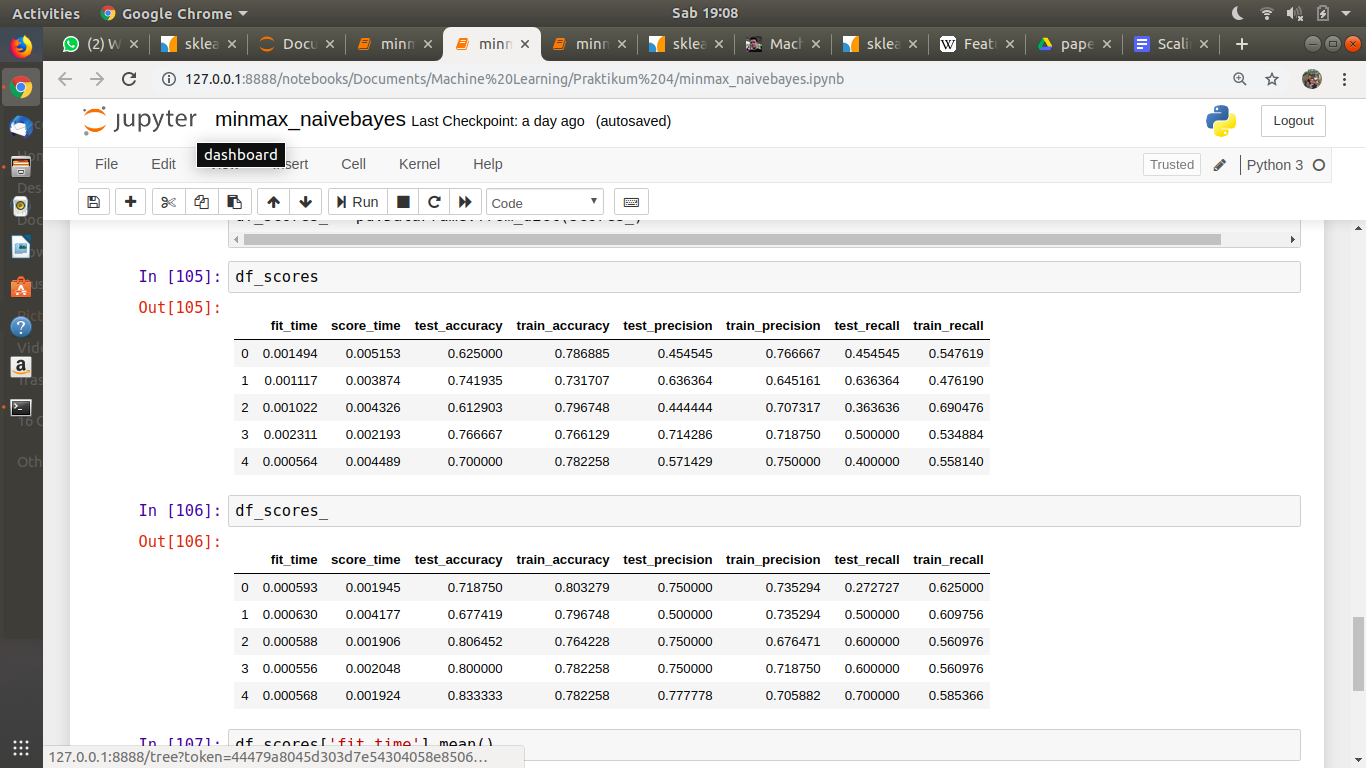


Perangkat yang digunakan adalah Thinkpad X240, processor i5-4300u 2c/4t, ram 4gb, sistem operasi Ubuntu 18.04

4. Hasil

Penerapan pada Gaussian Naive Bayes menggunakan Cross Validation.

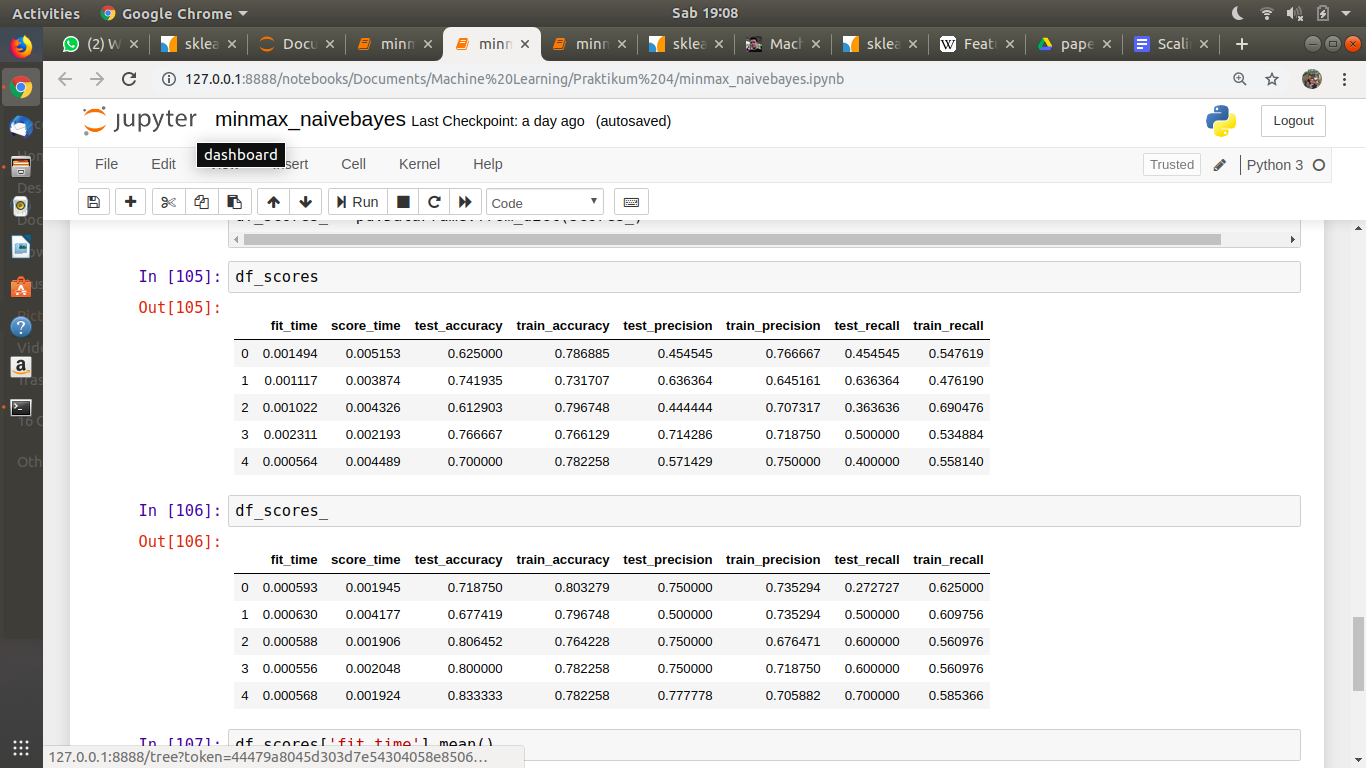
Hasil Crossvalidation GaussianNB menggunakan dataset yang tidak menerapkan feature scaling.



Rata-rata dari nilai FIT\_TIME yaitu 0.0013017654418945312 detik

Rata-rata dari nilai TEST\_ACCURACY yaitu sebesar 0.6893010752688171

Rata-rata dari nilai TEST\_PRECISION yaitu sebesar 0.5642135642135642



Rata-rata dari nilai FIT\_TIME yaitu 0.0005871295928955078 detik

Rata-rata dari nilai TEST\_ACCURACY yaitu sebesar 0.7671908602150537

Rata-rata dari nilai TEST\_PRECISION yaitu sebesar 0.7055555555555555

5. Kesimpulan

Berdasarkan pengujian tersebut, dapat disimpulkan bahwa penggunaan MinMaxScaler untuk feature Glucose pada dataset PIMA INDIAN menghasilkan rata-rata waktu training lebih cepat 45%, akurasi 11% lebih tinggi, dan precision 25% lebih tinggi dibanding tiidak menggunakan MinMaxScaling pada pengujian cross validation dengan menggunakan model Gaussian Naive Bayes.