Scaling Data Menggunakan Metode MinMax Scaler

1st Aditya Rahman
Teknik Elektro
Universitas Negeri Malang
Malang, Indonesia
aditya@klim.or.id

2nd Galih Hadi Wibowo
Teknik Elektro
Universitas Negeri Malang
Malang, Indonesia
hadigalih@gmail.com

***Abstract*—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.**

***Index Terms*—Scaling, MinMax Scaler**

I. Pendahuluan

A. 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 algoritma KMeans, karna itu data harus dilakukan scale agar tiap fitur memiliki bobot yang equal.

Pengujian menggunakan MinMax Scale karna metode ini mudah diterapkan untuk data yang tidak memiliki begitu banyak outlier sehingga hanya untuk melakukan normalisasi MinMax cukup untuk dataset ini.

B. 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.

II. 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. Pada artikel ini menerapkan scaling dengan range 0 hingga 1. MinMaxScale juga dapat memiliki output berupa range tertentu misal 1 hingga 3 dan seterusnya.

$$X' = a + \frac{(X_1 - min(X))(b - a)}{max(X) - min(X)} \tag{1}$$

Dengan:
a : range bawah
b : range atas
X_i : value dari suatu features
$min(X)$: nilai minimum dari features X
$max(X)$: nilai maksimum dari features X

III. Penerapan

Data PAMA INDIAN sebelum diterapkan MinMax Scaler jika ditampilkan seperti berikut ini

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Fig. 1. Data sebelum menggunakan MinMax Scaler.

Setelah diterapkan MinMax Scaler pada features Glucose

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	0.743719	72	35	0	33.6	0.627	50	1
1	1	0.427136	66	29	0	26.6	0.351	31	0
2	8	0.919598	64	0	0	23.3	0.672	32	1
3	1	0.447236	66	23	94	28.1	0.167	21	0
4	0	0.688442	40	35	168	43.1	2.288	33	1

Fig. 2. Data sesudah menggunakan MinMax Scaler.

Penerapan menggunakan modul MinMaxScaler dari library sklearn.preprocessing

Visualisasi data setelah diberlakukan Scaling pada features Glucose dan Pregnancies. Dilihat dari plot sebelah kiri bahwa kedua feaures tersebut memiliki rentang nilai yang berbeda, namun pada plot sebelah kanan data tersebut telah diterapkan MinMaxScale memiliki plot yang serupa karna memiliki rentang nilai yang serupa.

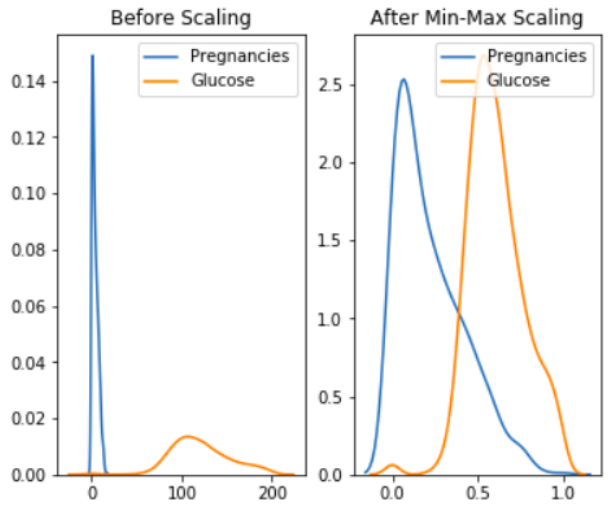

Fig. 3. Menggunakan module sklearn.preprocessing.

```
start_time = time.time()

scaler = MinMaxScaler(feature_range=(0, 1))
scaler.fit(df.iloc[:,1:2])
dfx['Glucose'] = scaler.transform(df.iloc[:,1:2])

end_time = time.time()
print("Durasi %g detik" % (end_time - start_time))
```

Fig. 4. Menggunakan module sklearn.preprocessing.

Durasi yang diperlukan untuk transformasi features Glucose

```
Durasi 0.00408173 detik
```

Fig. 5. Durasi menggunakan module sklearn.preprocessing.

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

IV. 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

	fit_time	score_time	test_accuracy	train_accuracy	test_precision	train_precision	test_recall	train_recall
0	0.001494	0.005153	0.625000	0.786885	0.454545	0.766667	0.454545	0.547619
1	0.001117	0.003874	0.741935	0.731707	0.636364	0.645161	0.636364	0.476190
2	0.001022	0.004326	0.612903	0.796748	0.444444	0.707317	0.363636	0.690476
3	0.002311	0.002193	0.766667	0.766129	0.714286	0.718750	0.500000	0.534884
4	0.000564	0.004489	0.700000	0.782258	0.571429	0.750000	0.400000	0.558140

Fig. 6. Hasil Crossvalidation GaussianNB.

Rata-rata dari nilai TEST_ACCURACY yaitu sebesar 0.6893010752688171

Rata-rata dari nilai TEST_PRECISION yaitu sebesar 0.5642135642135642

	fit_time	score_time	test_accuracy	train_accuracy	test_precision	train_precision	test_recall	train_recall
0	0.000593	0.001945	0.718750	0.803279	0.750000	0.735294	0.272727	0.625000
1	0.000630	0.004177	0.677419	0.796748	0.500000	0.735294	0.500000	0.609756
2	0.000588	0.001906	0.806452	0.764228	0.750000	0.676471	0.600000	0.560976
3	0.000556	0.002048	0.800000	0.782258	0.750000	0.718750	0.600000	0.560976
4	0.000568	0.001924	0.833333	0.782258	0.777778	0.705882	0.700000	0.585366

Fig. 7. Hasil Crossvalidation GaussianNB.

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

V. 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.

References

[1] Wikipedia. *"Feature scaling,"*. (Online). (https://en.m.wikipedia.org/wiki/Feature_scaling). Diakses pada 27 September 2019