e-ISBN 978-623-93343-4-5
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 15 September 2021

Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma *Naïve Bayes, Decision Tree,* dan *Support Vector Machine*

Fikri Adams[1], Realdy Agsar Dwi Anggoro[2], Muhammad Bayu Satria[3], Anggun Windari Oktavia[4], Nurul Chamidah[5]
Informatika / Fakultas Ilmu Komputer
Universitas Pembangunan Nasional Veteran Jakarta
Jakarta, Indonesia
fikriadams@upnvj.ac.id[1], realdyada@upnvj.ac.id[2], muhammadbs@upnvj.ac.id[3] , windarianggun@gmail.com[4], nurul.chamidah@upnvj.ac.id[5]

Abstrak. Dalam melakukan klasifikasi data menggunakan algoritma *machine learning*, keseimbangan rentang nilai pada suatu atribut dapat mempengaruhi kualitas hasil dari performa model klasifikasinya. Oleh karena itu, dataset dalam suatu penelitian diperlukan diterapkan tahap praproses data agar menghasilkan model klasifikasi dengan akurasi yang baik. Metode praproses data menggunakan tiga cara, yaitu normalisasi *min-max, z-score* dan *decimal scaling*. Setiap normalisasi data dikombinasikan dengan tiga algoritma *machine learning*, yaitu algoritma *Naïve Bayes, Decission Tree*, dan *Support Vector Machine* untuk mencari akurasi model yang terbaik dalam mengklasifikasikan dataset *wine*. Hasil perbandingan kombinasi ketiga metode klasifikasi dan ketiga metode normalisasi data menunjukkan bahwa akurasi terbaik terdapat pada algoritma *Support Vector Machine* dengan normalisasi *decimal scaling* dengan akurasi rata-rata yang diperoleh sebesar 57,1%. Hasil penelitian kali ini juga menunjukan bahwa suatu metode normalisasi data bisa saja mendapatkan hasil rata-rata akurasi tertinggi pada algoritma klasifikasi tertentu, akan tetapi belum tentu unggul apabila menggunakan metode klasifikasi yang lain.

Kata Kunci: Normalisasi, *Support Vector Machine*, *Decission Tree*, *Naïve Bayes.*

1 Pendahuluan

Wine adalah minuman yang berasal dari hasil fermentasi buah anggur menjadi minuman beralkohol. *Wine* merupakan salah satu minuman favorit serta banyak peminatnya terutama di luar negeri. Banyak penikmat *wine* yang berkembang menjadi seorang pakar *wine*, tugas dari seorang pakar *wine* adalah melabeli berbagai jenis *wine*. Dari hal tersebut klasifikasi data *wine* digunakan dalam mempermudah pakar *wine* dalam melakukan pelabelan awal pada minuman wine.

Pada klasifikasi dataset *wine*, dibutuhkan praproses data untuk meminimalisir *noise* pada dataset *wine* dengan cara menormalisasi nilai setiap atribut pada dataset *wine* tersebut. Tujuan dilakukannya praproses data dengan metode normalisasi tersebut adalah untuk menyeimbangkan rentang nilai pada tiap atribut agar hasil yang didapatkan bisa lebih akurat, serta dapat mengurangi waktu komputasi perhitungan pada model, dan nilai atribut pada dataset *wine* tersebut memiliki rentang yang seimbang tanpa merubah informasi pada dataset tersebut. Rentang nilai yang berbeda jauh pada setiap atribut menjadi salah satu permasalahan yang dapat diselesaikan pada tahap praproses. Perbedaan ruang nilai dari masing-masing atribut dapat menyebabkan atribut tersebut tidak memberikan pengaruh secara maksimal dalam menentukan hasil pengklasifikasian data karena memiliki nilai tidak seragam dibandingkan atribut lainnya. Praproses yang dilakukan adalah dengan cara transformasi data dengan normalisasi. Tujuan dilakukannya normalisasi pada data adalah untuk menyeimbangkan skala nilai pada masing-masing atribut dalam *range* tertentu. Terdapat 3 teknik untuk normalisasi dengan cara normalisasi *min-max, z-score* dan *decimal scaling.*

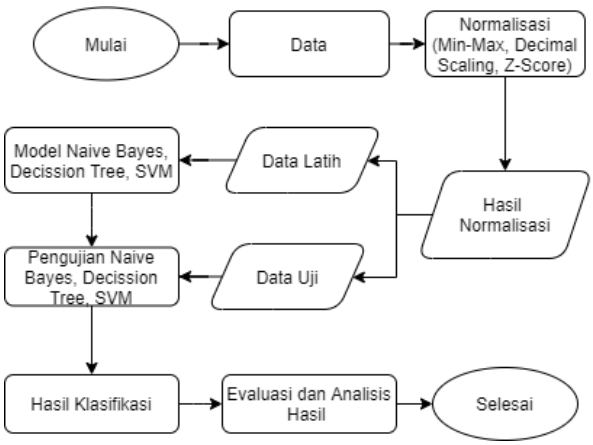
Klasifikasi dilakukan untuk mengelompokkan data ke suatu objek kelas (kategori) didasari variabel tertentu. Proses klasifikasi data dapat dilakukan dengan mengamati suatu kelompok dari data dengan variabel. Tujuan klasifikasi adalah untuk memprediksi suatu kelas yang belum diketahui. Tahap klasifikasi diantaranya, yaitu pembentukan model dan evaluasi model.

Pada penelitian terdahulu yang berkaitan dengan perbandingan metode normalisasi data pada dataset *wine* menggunakan alogirtma *K-Nearest Neighbor* (K-NN) pernah dilakukan sebelumnya dengan menghasilkan rata-rata akurasi sebesar 59,68% [1]. Kemudian penelitian terdahulu yang berkaitan dengan perbandingan metode normalisasi data juga pernah dilakukan pada dataset kasus kanker payudara dengan menggunakan algoritma JST *Backpropagasi Gradient Descent* dengan *Adaptive Gain* (BPGD/AG) dengan menghasilkan akurasi rata-rata hingga 96.86% [2].

Berdasarkan hal-hal diatas maka dilakukan pengembangan pada penelitian terdahulu dengan melakukan penelitian terhadap perbandingan kombinasi antara algoritma klasifikasi dan normalisasi data dengan tujuan mengamati hasil perfoma model dengan akurasi terbaik pada klasifikasi dataset *wine*.

2 Metodologi Penelitian

Ada beberapa tahapan untuk penelitian ini, bisa dilihat melalui diagram *flowchart* sebagai berikut:



Gambar. 1. Tahapan Penelitian

2.1 Dataset Penelitian

Dataset yang diperoleh merupakan dataset *wine* dari UCI *Machine Learning*. Jumlah *record* data sebanyak 1599. Data yang dimiliki sebesar 11 variabel atribut dengan satu output berupa kelas pada skla nilai 0-10.

2.2 *Wine*

Wine adalah larutan hidroalkohol dengan tingkat keasaman ph mencapai 3 sampai 4. Komponen utama *wine* adalah air dan etanol, biasanya terhitung sekitar 97% pada *weight-for-weight* (w/w). Senyawa yang tersisa bisanya sebagian besar rasa dan warna anggur yang kurang dari 10g/L. Senyawa yang ada dalam anggur juga bisa ditemukan pada kopi, bir, roti, rempah-rempah, sayuran, keju, dan bahan makanan lainnya [3].

2.3 Praproses Data

Dalam penelitian ini, praproses data dilakukan untuk mentranformasi data menggunakan teknik normalisasi. Normalisasi merupakan sebuah proses penskalaan kolom atribut menjadi nilai numerik pada rentang tertentu [4]. Berikut adalah tahapan normalisasi untuk penelitian ini:

a) Normalisasi *Min-Max*

Min-max merupakan teknik pengskalaan yang menggunakan *minimum* dan *maximum* dari fitur untuk mengubah skala nilai ke dalam suatu rentang, biasanya rentang yang dipakai 0 hingga 1 atau –1 hingga 1 [5]. Berikut ini adalah rumus yang digunakan pada persamaan 1 sebagai berikut [1]:

$$Data(x) = \frac{(x - minValue) * (maxRange - minRange)}{maxValue - minValue} + minRange \quad \textbf{(1)}$$

Keterangan :

Data(x) : data baru dari hasil normalisasi
x : data yang akan dinormalisasi
minValue : nilai terkecil dari satu kolom baris
maxValue : nilai terbesar dari satu kolom baris
minRange : batas nilai terkecil dari normalisasi
maxRange : batas nilai terbesar dari normalisasi

b) Normalisasi *Z-score*

Z-Score adalah penskalaan fitur agar mendekati standar yang didistribusikan secara normal. Standardisasi untuk mengubah data yang sedemikian rupa sehingga memiliki mean x̄, dari 0 dan standar deviasi, σ, dari 1 [5]. Pada tahap ini menggunakan persamaan 2 sebagai berikut:

$$x_i' = \frac{x_i - \bar{x}}{\sigma} \quad \textbf{(2)}$$

Keterangan :
$x_i' x_i'$: data baru dari hasil normalisasi
$x_i x_i$: data yang akan di normalisasi
$\bar{x}\bar{x}$: rata-rata dari setiap kolom
$\sigma\sigma$: standar deviasi dari beberapa kolom

c) *Decimal Scaling*

Decimal Scaling merupakan sebuah teknik normalisasi data yang digunakan untuk pengskalaan desimal dengan menggeser nilai variabel atribut pada titik desimal. Besar titik desimal tergantung dalam nilai absolut atribut [6]. Dalam tahap ini, rumus yang digunakan pada persamaan 3 sebagai berikut:

$$v_i' = \frac{v_i}{10^n} \quad \textbf{(3)}$$

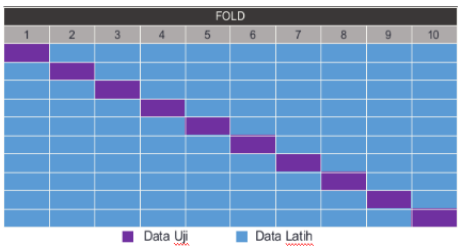
Keterangan :
$v_i' v_i'$: data baru dari hasil normalisasi
$v_i v_i$: data yang akan dinormalisasi
n : pangkat untuk membagi

2.4 Klasifikasi

Klasifikasi merupakan tahapan untuk mencari model dalam memprediksi objek kelas biasanya disebut dengan target (kategori). Model yang digunakan biasanya kumpulan dari turunan data pelatihan yang labelnya sudah diketahui [6]. Dalam penelitian ini, menggunakan tahapan untuk pengklasifikasian sebagai berikut:

a) *K-fold Cross Validation*

Cross validation pada dasarnya sebuah teknik pembagian data dalam membuat model dengan cara sampel asli masuk ke partisi *training* set sehingga digunakan untuk melatih model, sedangkan pada *testing* set digunakan untuk mengevaluasi model [7]. Pada *k-fold cross-validation*, pembagi dataset dilakukan secara acak menjadi *k-fold* tanpa penggantian. Di mana k-1 *fold* digunakan untuk *training* model dan 1 *fold* digunakan untuk pengujian [8]. Dalam penelitian ini proses pembagian data pada klasifikasi adalah k-*fold cross validation* dengan range nilai k = 1-10.



Gambar. 2. Ilustrasi 10-*fold cross validation*

b) *Naïve Bayes*

Naïve Bayes merupakan salah satu teknik untuk klasifikasi, di mana proses perthitungan menggunakan probabilitas dan statistik. Metode *Naïve Bayes* digunakan untuk memahami probabilitas beberapa peristiwa, P (A | B), dengan beberapa informasi baru, P (B | A), dan kejadian sebelumnya pada probabilitas peristiwa, P (A) [5]. Persamaan rumus dapat dilihat pada persamaan 4 sebagai berikut:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \quad \textbf{(4)}$$

Keterangan :

B	: Data kelas yang belum diketahui
A	: Hipotesis data yang merupakan suatu kelas
P(A|B)	: Probabilitas hipotesis dari suatu kondisi (posteriori p*robability*)
P(A)	: Probabilitas hipotesis (*prior probability*)
P(B|A)	: Probabilitas dari kondisi berdasarkan hipotesis
P(B)	: Probabilitas dari A

c) *Decision Tree*

Decision tree merupakan sebuah pohon keputusan di mana setiap simpul mewakili fitur (atribut), setiap tautan (cabang) mewakili keputusan (aturan), dan setiap daun mewakili hasil (nilai kategorikal atau kontinu) dari variabel kelas. Pohon keputusan (*decision tree*) memiliki struktur diagram alir sehingga setiap node internal (node *non-leaf*) adalah tes pada atribut, sedangkan setiap pada cabang mewakili percobaan dan hasil node terminal adalah label (kelas). Node paling atas dalam sebuah pohon adalah simpul akar [6]. Dalam tahapan ini, menggunakan persamaan rumus 5 dan 6 sebagai berikut [9]:

$$Gain(S,A) = Entropy(S) - \left(\sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(|S_i|)\right) \quad \textbf{(5)}$$

Keterangan :

S	: Himpunan dari kasus dataset
A	: Atribut

$|S_i||S_i|$: Jumlah dari beberapa sampel pada nilai ke-i.
$|S||S|$: Jumlah dari seluruh sampel data

Dimana terdapat rumus persamaan untuk menghitung nilai Entropy:

$$Gain(S) = \sum_{i=1}^{n} -p_i * \log_2 p_i \quad \textbf{(6)}$$

Keterangan :

S : Himpunan dari kasus dataset
n : Jumlah partisi dari S
pi : Probabilitas dari *Si* ke S

d) *Support Vector Machine*

Support Vector Machine adalah kategori *supervised lerning*, biasanya digunakan untuk sebuah proses klasifikasi dan regresi. Tujuannya digunakan untuk memaksimalkan margin pada *hyperplane*. Di mana margin adalah jarak antara *hyperplane* pemisah (batas keputusan) dan sampel titik pelatihan yang paling dekat dengan *hyperplane* [10]. Pada tahapan ini digunakan persamaan rumus 7,8 dan 9 sebagai berikut [11]:

$$d(x) = w.x + b \quad \textbf{(7)}$$

atau

$$d(x) = \sum_{i=1}^{l} d_j y_i K(X, X_i) + b_0 \quad \textbf{(8)}$$

Keterangan :
w : Parameter *hyperplane* dengan mencari garis lurus antara garis *hyperplane* pada titik data.
x : Titik data.
$d_i d_i$: Bobot pada nilai disetiap titik data.
$K(X,X_i)K(X,X_i)$: Fungsi kernel.
$b_0 b_0$: Nilai bias.
Yang mana dalam tahap ini menggunakan kernel linier. Dengan rumus sebagai berikut:

$$K(x, y) = K(x_1^T.x_2) \quad \textbf{(9)}$$

Keterangan :
K(*x*,*y*) : Nilai kernel dari data x dan y.
$x_i^T x_i^T$: Nilai fitur data 1.
$x_2 x_2$: Nilai fitur data 2.

2.5 Evaluasi

Evaluasi digunakan untuk mengukur performa dari model pada permasalahan klasifikasi [12]. Berikut adalah persamaan rumus 10 untuk evaluasi dengan menganalisa akurasi sebagai berikut [13] :

$$K(x, y) = K(x_1^T.x_2) \quad \textbf{(10)}$$

Keterangan :
True Positive (TP), data yang diprediksi benar dan nyatanya benar.
True Negative (TN), data yang diprediksi salah dan nyatanya salah.
False Positive (FP), data yang diprediksi benar dan nyatanya salah.
False Negative (FN), data yang diprediksi salah dan nyatanya benar.

3 Hasil dan Pembahasan

Pada tahapan yang pertama dilakukan adalah praproses data dengan cara normalisasi pada dataset *wine*. Proses normalisasi dilakukan dengan 3 cara yaitu normalisasi *min-max, z-score* dan *decimal scaling*. Berikut adalah dataset sebelum dipraproses yang dapat dirangkum pada Tabel 1:

Tabel 1. Dataset *Wine*

No	Fixed acidity	Volatile acidity	Citric acid	…..	quality
1	7.4	0.7	0	….	5
2	7.8	0.88	0	….	5
3	7.8	0.76	0.04	….	5
4	11.2	0.28	0.56	….	6
….	….	….	….	….	5
1599	6	0.31	0.47	….	6

Pada tabel 1 merupakan dataset *wine* asli yang masih belum di praproses sehingga skala nilai masih berbeda maka dari hal tersebut perlu dilakukannya normalisasi pada dataset *wine*.

Selajutnya masuk pada tahapan praproses dengan normalisasi *min-max*, di mana data yang diolah menjadi nilai minimum dan maksimum di setiap variabel atribut. Skala nilai pada metode normalisasi *min-max* adalah 0 sampai 1. Pada rumus persamaan 1 digunakan untuk normalisasi *min-max.* Hasil dari normalisasi dapat dilihat pada Tabel 2 bahwa nilai yang diperoleh memiliki *range* yang seimbang.

Tabel 2. Hasil normalisasi pada *min-max*

No	Fixed acidity	Volatile acidity	Citric acid	…..	quality
1	0.247788	0.39726	0	….	5
2	0.283186	0.520548	0	….	5
3	0.283186	0.438356	0.04	….	5
4	0.584071	0.109589	0.56	….	6
….	….	….	….	….	5
1599	0.123894	0.130137	0.47	….	6

Kemudian masuk pada praproses dengan normalisasi *z-score*. Data yang digunakan dataset *wine* asli bisa dilihat pada Tabel 1. Metode normalisasi *z-score* konsep data yang diolah diambil dari nilai atribut menggunakan parameter rata-rata (mean) dan standar deviasi. Dari metode tersebut metode *z-score* dapat di aplikasikan ke persamaan rumus 2. Hasil normalisasi *z-score* dapat dilihat pada Tabel 3.

Tabel 3. Hasil normalisasi setelah menggunakan *z-score*

No	Fixed acidity	Volatile acidity	Citric acid	…..	quality
1	-0.52836	0.961877	-1.39147	….	5
2	-0.298547	1.96744	-1.39147	….	5
3	-0.298547	1.29707	-1.18607	….	5

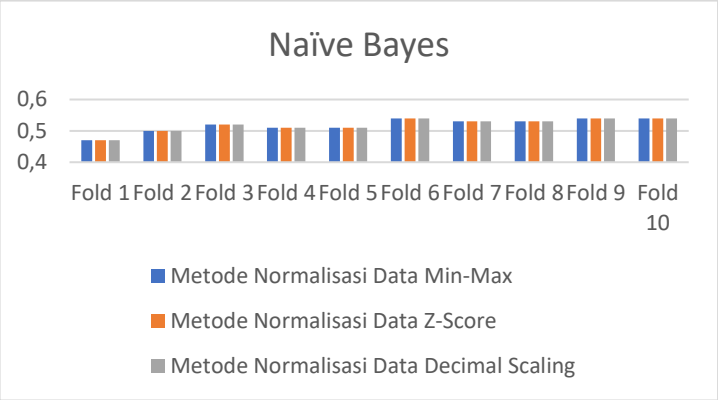
4	1.65486	-1.38444	1.48415	….	6
….	….	….	….	….	5
1599	-1.3327	-1.21685	1.022	….	6

Tahap terakhir pada praproses ini adalah *decimal scaling*, di mana dataset *wine* pada Tabel 1 di praproses kembali menggunakan normalisasi *decimal scaling*. Pada proses *decimal scaling* menggunakan persamaan rumus 3. Berikut adalah hasil normalisasi *decimal scaling* menggunakan dataset *wine*, dapat dilihat pada Tabel 4.

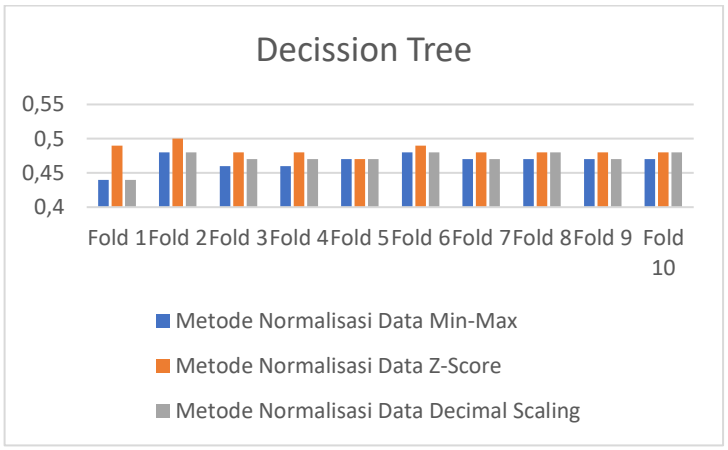
Tabel 4. Hasil normalisasi menggunakan *decimal scaling*

No	Fixed acidity	Volatile acidity	Citric acid	…..	quality
1	0.074	0.07	0	….	5
2	0.078	0.088	0	….	5
3	0.078	0.076	0.004	….	5
4	0.112	0.028	0.056	….	6
….	….	….	….	….	5
1599	0.06	0.031	0.047	….	6

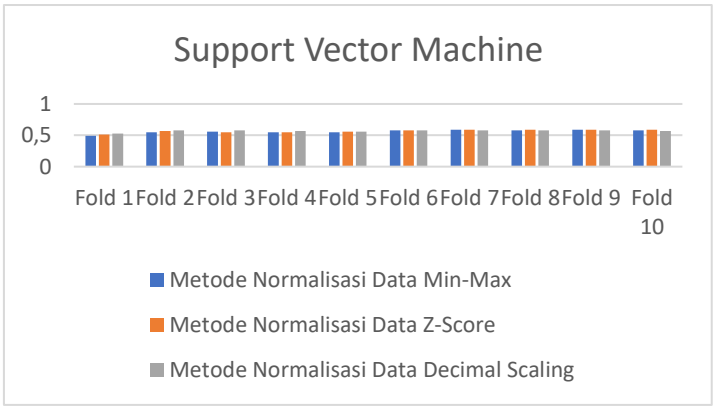
Dari hasil normalisasi *decimal scaling* perubahan dari masing-masing nilai variable atribut memiliki rentang nilai tidak terlalu jauh. Setelah melakukan praproses data, langkah selanjutnya pembagian data menggunakan metode *k-fold cross validation*. Proses pembagian data dilakukan dengan k-*fold cross validation* dengan nilai k = 10. Dari masing-masing data yang telah normalisasi seperti *min-max*, *z-score* dan *decimal scaling* kemudian klasfikasikan menggunakan algoritma *naïve bayes*, *decision tree* dan *support vector machine*. Kemudian hasil klasifikasi tersebut dievaluasi ketiga modelnya dengan indicator evaluasinya yaitu akurasi. Berikut adalah hasil evaluasi model menggunakan algoritma *naïve bayes*, *decision tree* dan *support vector machine*, dapat dilihat pada Gambar 3, 4 dan 5:



Gambar 3. Hasil Evaluasi Model *Naïve Bayes*

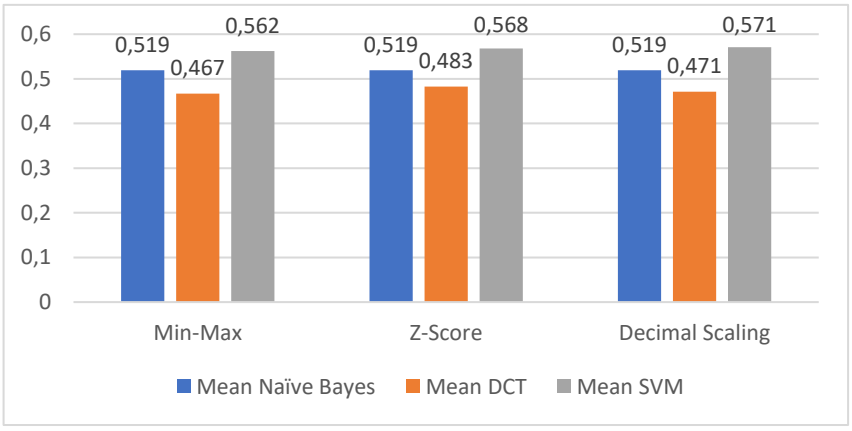


Gambar 4. Hasil Evaluasi Model *Decision Tree*



Gambar 5. Hasil Evaluasi Model *Support Vector Machine*

Pada Gambar 3, 5 dan 6 dilakukan evaluasi model klasifikasi *naïve bayes*, *decision tree* dan *support vector machine*, dengan hasil rata-rata akurasi pada percobaann data yang telah di normalisasi dengan *min-max*, *z-score*, dan *decimal scaling*. Berikut adalah perbandingan hasil rata-rata akurasi pada model yang dibentuk dengan algoritma *naïve bayes*, *decision tree* dan *support vector machine* dan dataset yang telah normalisasi menggunakan *min-max*, *z-score*, dan *decimal scaling* dapat dilhat pada Gambar 6 :



Gambar 6. Hasil Perbandingan Rata-Rata Akurasi Dengan Metode Normalisasi

4 Kesimpulan

Dari hasil penelitian yang sudah dibuat, maka didapatkan kesimpulan, yaitu:

a) Hasil pengembangan dari penelitian sebelumnya [1] menemukan bahwa suatu metode normalisasi data dapat mendapatkan hasil rata-rata akurasi tertinggi pada algoritma klasifikasi tertentu tetapi belum tentu unggul apabila menggunakan metode klasifikasi yang lain.

b) Akurasi dengan rata-rata tertinggi adalah 57,1% pada algoritma SVM dengan menggunakan metode *decimal scaling*.

c) Akurasi dengan rata-rata terendah terdapat pada dataset dengan normalisasi *min-max* pada algoritma Decission Tree sebesar 46,7%.

d) Metode praproses pada normalisasi dapat mempengaruhi hasil perfoma dari model akurasi.

e) Akurasi menggunakan dengan metode normalisasi pada penelitian ini tidak lebih baik dari akurasi penelitian sebelumnya [1] sebesar 59,68%.

Referensi

[1] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN," *Comput. Eng. Sci. Syst. J.*, vol. 4, no. 1, p. 78, 2019, doi: 10.24114/cess.v4i1.11458.

[2] N. Chamidah, . W., and U. Salamah, "Pengaruh Normalisasi Data pada Jaringan Syaraf Tiruan Backpropagasi Gradient Descent Adaptive Gain (BPGDAG) untuk Klasifikasi," *J. Teknol. Inf. ITSmart*, vol. 1, no. 1, p. 28, 2012.

[3] A. L. Waterhouse, G. L. Sacks, and D. W. Jeffery, *Understanding Wine Chemistry*. O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472., 2016.

[4] M. Bowles, *Machine Learning in Python: Essential Techniques for Predictive Analysis*. 2015.

[5] Chris Albon, *Machine Learning with Python Cookbook Practical Solutions from Preprocessing to Deep Learning*. O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472., 2018.

[6] J. Han and M. Kamber, *Data Mining: Concepts and Techniques Second Edition*. Diane Cerra, 2007.

[7] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann; 3rd edition (January 20, 2011), 2016.

[8] S. Raschka, *Python Machine Learning: Unlock deeper insights into Machine Leaning with this vital guide to cutting-edge predictive analytics*. Packt Publishing Ltd, 2015.

[9] H. Widayu, S. Darma, N. Silalahi, and Mesran, "Data Mining Untuk Memprediksi Jenis Transaksi Nasabah Pada Koperasi Simpan Pinjam Dengan Algoritma C4.5," *Issn 2548-8368*, vol. Vol 1, No, no. June, p. 7, 2017, [Online]. Available: https://ejurnal.stmik-budidarma.ac.id/index.php/mib/article/view/323.

[10] S. Raschka and V. Mirjalili, *Python Machine Learning Third edition*, vol. 53. Packt Publishing Ltd, 2019.

[11] A. M. Pravina, I. Cholissodin, and P. P. Adikara, "Analisis Sentimen Tentang Opini Maskapai Penerbangan pada Dokumen Twitter Menggunakan Algoritme Support Vector Machine (SVM)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 3, pp. 2789–2797, 2019, [Online]. Available: http://j-ptiik.ub.ac.id.

[12] A. Novantirani, M. K. Sabariah, and V. Effendy, "Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode Support Vector Machine," *e-Proceeeding Eng.*, vol. 2, no. 1, pp. 1–7, 2015.

[13] J. LING, I. P. E. N. KENCANA, and T. B. OKA, "Analisis Sentimen Menggunakan Metode Naïve Bayes Classifier Dengan Seleksi Fitur Chi Square," *E-Jurnal Mat.*, vol. 3, no. 3, p. 92, 2014, doi: 10.24843/mtk.2014.v03.i03.p070.