



## KLASIFIKASI PENDERITA DIABETES MENGGUNAKAN ALGORITMA MACHINE LEARNING DAN Z-SCORE

Ichwanul Muslim Karo Karo<sup>1</sup>, Hendriyana<sup>2</sup>

<sup>1</sup> Ilmu Komputer, Universitas Negeri Medan

<sup>2</sup> Rekayasa Perangkat Lunak, UPI Kampus Daerah Cibiru, Universitas Pendidikan Indonesia

Medan, Sumatera Utara, Indonesia 20221

Bandung, Jawa Barat, Indonesia 40614

[ichwanul@unimed.ac.id](mailto:ichwanul@unimed.ac.id), [hendriyana@upi.edu](mailto:hendriyana@upi.edu)

### Abstract

*Diabetes is a deadly and chronic disease. It characterized by an increase in blood sugar. Many complications occur if diabetes does not treat and identified. The common identification process by visits to diagnostic centers and consulting physician. It makes bored patients. Machine learning approach can solve the problem of diabetic identification. However, the unbalanced range of diabetes variable values affects the quality of machine learning results. This study predicts the likelihood of diabetes in diabetic patients from 768 Indian women, using three machine learning classification algorithms and Z-Score normalization method. The machine learning algorithms used are Decision Tree, Support Vector Machine (SVM) and Naive Bayes. Experiments were run on the Pima Indians Diabetes Database (PIDD). Dataset retrieved from the UCI Machine Learning Repository. The performance of the three algorithms was evaluated using accuracy, precision, F1, and recall based on confusion matrix. SVM algorithm is an algorithm that has the highest performance that both algorithm the Naive Bayes and Decision Tree algorithms, the accuracy and F1 is 80.73% and 76%. The Z-Score method has positively contribution to increasing the accuracy of the classification model. Furthermore, this study also managed to get a higher accuracy than previous studies.*

**Keywords:** Diabetes, Decision Tree, Machine Learning, Naive Bayes, SVM, Z-Score

### Abstrak

Diabetes merupakan salah satu penyakit mematikan dan kronis ditandai dengan peningkatan gula darah. Banyak komplikasi terjadi jika diabetes tidak diobati dan tidak teridentifikasi. Proses identifikasi umumnya dilakukan dengan kunjungan ke pusat diagnostik dan dokter konsultasi membuat pasien bosan. Pendekatan *machine learning* dapat memecahkan masalah identifikasi penyakit diabetes. Namun, rentang nilai variabel penentu penyakit diabetes yang tidak seimbang mempengaruhi kualitas hasil *machine learning*. Penelitian ini memprediksi kemungkinan diabetes pada pasien penderita diabetes dari 768 wanita Indian, dengan tiga algoritma klasifikasi *machine learning* dan metode normalisasi *Z-Score*. Adapun algoritma *machine learning* yang digunakan adalah *Decision Tree*, *Support Vector Machine* (SVM) dan Naive Bayes. Eksperimen dilakukan pada *Pima Indians Diabetes Database* (PIDD) yang bersumber dari *UCI Machine Learning Repository*. Kinerja ketiga algoritma dievaluasi menggunakan dengan akurasi, *Precision*, *F1*, dan *Recall* berdasarkan *confusion matrix*. *SVM* merupakan algoritma yang memiliki performansi paling tinggi dibandingkan dengan algoritma Naive Bayes dan *Decision Tree*, dengan akurasi 80.73% dan *F1* 76 %. Metode *Z-Score* berkontribusi positif dalam meningkatkan akurasi model klasifikasi. Lebih lanjut, penelitian ini juga berhasil mendapat akurasi yang lebih tinggi dari penelitian sebelumnya.

**Kata kunci:** Diabetes, *Decision Tree*, *Machine Learning*, Naive Bayes, *SVM*, *Z-Score*

### 1. PENDAHULUAN

Diabetes merupakan penyakit kronis, umumnya ditandai dengan kadar gula darah yang tinggi [1]. Penumpukan glukosa dalam darah diakibatkan oleh sel tubuh tidak bekerja secara normal dalam menyerap glukosa. Sehingga,

hal tersebut dapat menimbulkan gangguan organ tubuh. Diabetes penting untuk diidentifikasi sejak dini karena bisa menyebabkan kematian. Terdapat beberapa faktor yang menandai seseorang terkena diabetes, hal tersebut penting

diketahui untuk segera antisipasi dan melakukan pemeriksaan lebih lanjut.

Dengan perkembangan teknologi, diabetes dapat diidentifikasi sejak dini dengan menggunakan pendekatan *Data mining*[2]. *Data mining* adalah proses pengumpulan dan pengolahan data yang bertujuan untuk mengekstrak informasi penting atau sebuah pola pada data[3]. Proses ekstraksi dapat dilakukan dengan algoritma *machine learning*.

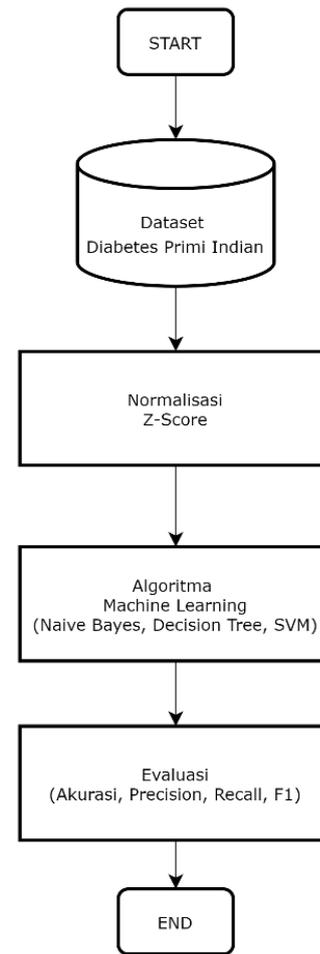
Algoritma *machine learning* telah banyak diimplementasikan untuk mengidentifikasi penderita diabetes. Sebuah penelitian oleh[2], menganalisis hasil klasifikasi diabetes dari tiga buah algoritma *machine learning*, yakni Naive Bayes, *Decision Tree*, dan *Support Vector Machine (SVM)*. *Dataset* yang digunakan adalah Pima Indian yang berasal dari *National Institute of Diabetes and Digestive and Kidney Diseases*. Algoritma Naive Bayes memiliki akurasi tertinggi dibanding dengan algoritma lainnya. Dengan *dataset* yang sama, peneliti [4] mengklasifikasikan penderita diabetes menggunakan algoritma *SVM*. Adapun akurasi terbaik yang diperoleh adalah 77.92%.

Penelitian sebelumnya belum mampu menyajikan akurasi di atas 80%. Dengan kata lain, penelitian yang sudah dilakukan belum mencapai performansi maksimal. Selain itu, metode normalisasi atas rentang nilai variabel penentu penyakit diabetes yang tidak seimbang ditengarai menjadi salah satu faktor rendahnya performansi model klasifikasi [5], [6].

Penelitian ini mengklasifikasikan penderita diabetes dengan algoritma *machine learning* serta pelibatan normalisasi data pada tahap *pre-processing*. Adapun algoritma *machine learning* yang digunakan meliputi Naive Bayes, *Decision Tree* dan *SVM*. Penelitian ini juga merupakan upaya perbaikan atas penelitian sebelumnya[2], [4].

## 2. METODE PENELITIAN

Penelitian ini memiliki rancangan sistem seperti yang terlihat pada Gambar 1. Secara garis besar, penelitian ini melawati tiga tahapan, normalisasi data, implementasi algoritma dan proses evaluasi.



Gambar 1. Flowchart Penelitian

### 2.1 Dataset

*Dataset* yang digunakan adalah *dataset* Pima Indian yang berasal dari *National Institute of Diabetes and Digestive and Kidney Diseases*. *Dataset* diperoleh dari *UCI Machine Learning Repository*. *Dataset* ini juga telah digunakan pada penelitian sebelumnya [2], [4]. *Dataset* terdiri dari 768 *record*, delapan atribut numerik dan dua kelas. Lebih lanjut, semua pasien di sini adalah wanita dari suku Pima Indians dengan usia di atas 21 tahun. Tabel 1 merupakan contoh *dataset* sedangkan penjelasan atribut dapat dilihat pada Tabel 2.

Tabel 1. Contoh *Dataset*

Preg	Glu	Blood	Skin	Ins	BMI	Ped.	Age	Class
6	14	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.221	21	0
0	137	40	35	168	43.1	2.288	33	1

Tabel 2. Deskripsi Atribut

Atribut	Deskripsi
<i>Preg</i>	Jumlah berapa kali hamil
<i>Glu</i>	konsentrasi glukosa setelah 2 jam test
<i>Blo</i>	Tekanan darah diastolik
<i>Skin</i>	Ketebalan kulit bagian triceps
<i>Ins</i>	2 jam setelah serum insulin
<i>BMI</i>	Indeks massa tubuh (berat dalam kg / (tinggi dalam m)^2
<i>Ped.</i>	Fungsi Pedigree Diabetes
<i>Age</i>	Umur
<i>Class</i>	Kelas untuk klasifikasi, 1 artinya terkena diabetes dan 0 berarti tidak.

### 2.2 Data Pre-processing

Data *pre-processing* memiliki peranan penting dalam meningkatkan kinerja model algoritma *machine learning*[7]. Harapan dilakukannya data *pre-processing* data adalah diperolehnya kondisi data yang ideal untuk diproses[8]. Terdapat banyak teknik data *pre-processing*, salah satunya adalah normalisasi data. Penelitian ini menerapkan data normalisasi sebagai teknik data *pre-processing*. Terdapat banyak metode data normalisasi, yang umum digunakan adalah metode *Z-Score*[9]. Salah satu tujuannya adalah untuk men-standarisasi *dataset*.

Nilai *Z-Score* merupakan ukuran penyimpangan data dari nilai rata-ratanya ( $\mu$ ) yang diukur dalam satuan standar deviasinya ( $\sigma$ ) [6]. *Z-Score* ini juga disebut dengan Nilai Standar atau Nilai baku. Persamaan (1) merupakan formula untuk menghitung nilai *Z-Score* dari setiap data yang diamati ( $x$ ).

$$Z = \frac{x - \mu}{\sigma} \tag{1}$$

Pada tahap normalisasi data, semua nilai  $x$  akan bertransformasi ke nilai  $Z$ . Sehingga Tabel 1 (*Dataset* yang belum dinormalisasi), akan bertransformasi nilainya ke Tabel 3.

Tabel 3. Contoh *Dataset* Setelah Proses Normalisasi

Preg	Glu	Blood	Skin	Ins	BMI	Ped.	Age	Class
0.64	0.85	0.15	0.91	-0.7	0.2	0.47	1.43	1
-0.	-1.	-0.16	0.53	-0.7	-0.	-0.2	-0.	0
1.23	1.94	-0.26	-1.	-0.7	-1.1	-0.60	-0.	1
-0.	-1	-0.16	0.15	0.12	-0.5	-0.92	-1.	0
-1.	0.5	-1.5	0.91	0.77	1.41	5.48	-0.	1

### 2.2 Algoritma Decision Tree

Algoritma *Decision Tree* merupakan salah satu algoritma *machine learning* dengan teknik pemodelan paling sederhana untuk mengklasifikasikan data[10]. Algoritma *Decision Tree* menggunakan *graph* seperti pohon yang mewakili struktur *root* atau akar dan *leaf* atau daun. Setiap *root* pohon mewakili sebuah atribut. Setiap cabang dari simpul mewakili hasil tes, dan simpul terakhir adalah "daun" yang mewakili label atau kelas. Untuk menentukan *root*, algoritma *Decision Tree* umumnya menggunakan *Information Gain* atau Gini Index. Pada penelitian ini, proses membangun model klasifikasi dengan algoritma *Decision Tree* mengikuti panduan penelitian [2].

### 2.3 Algoritma Naïve Bayes

Algoritma Naïve Bayes menjadi salah satu dari sepuluh algoritma klasifikasi terpopuler dan mudah untuk diimplementasikan pada berbagai kasus[11]. Naive Bayes merupakan algoritma klasifikasi sederhana dengan pendekatan probabilitas[12]. Idanya adalah menentukan nilai probabilitas dari setiap variabel dengan menjumlahkan frekuensi dan kombinasi nilai dari kumpulan data. Proses pembangunan model klasifikasi, menggunakan teorema Bayes (persamaan (2)). Sehingga untuk memprediksi kelas sebuah data yang diamati ( $X$ ), ditentukan dari probabilitas tertinggi dari keanggotaannya. Algoritma Naïve Bayes mengasumsikan bahwa semua variabel ( $Y$ ) haruslah independen satu dengan yang lainnya. Lebih lanjut, antar variabel dengan kelas juga tidak terdapat hubungan satu dengan yang lainnya. Karena *dataset* bertipe numerik, maka rumus yang digunakan yaitu Gaussian Naïve Bayes mengikuti panduan pada penelitian [11].

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{2}$$

### 2.4 Algoritma Support Vector Machine

*Support Vector Machine (SVM)* merupakan salah satu algoritma pembelajaran mesin yang efektif untuk mengakomodasi banyak variabel dan banyak kelas[13]. Ide algoritma *SVM* adalah menciptakan *hyperplane* yang mampu memisahkan *dataset*. *Hyperplane* berupa sebuah fungsi dengan  $N_{in} - 1$  variabel independen. Adapun algoritma *SVM* yang digunakan mengikuti panduan sebagaimana pada Tabel 4[13].

Tabel 4. Algoritma *SVM*

Algoritma
<b>Masukan:</b>
$N_{in}$ (Jumlah vektor)
$N_{sv}$ (Jumlah <i>support vector</i> )
$N_{ft}$ (Jumlah atribut di <i>support</i> )
$SV[N_{sv}]$ (Array <i>support vector</i> )
$IN[N_{in}]$ (masukan array vektor)
$b *$ (bias)
<b>Luaran:</b>
$F$ (Fungsi <i>hyperplane</i> )

```

Algoritma
for  $i \leftarrow 1$  to  $N_{in}$  by 1 do
   $F = 0$ 
  for  $j \leftarrow 1$  to  $N_{sv}$  by 1 do
     $dist = 0$ 
    for  $k \leftarrow 1$  to  $N_{ft}$  by 1 do
       $dist += (SV[j].variabel[k] - IN[i].variabel[k])^2$ 
    end
     $k = \exp(-\gamma \times dist)$ 
     $F += SV[j].a * k$ 
  end
   $F = F + b *$ 
end

```

## 2.5 Proses Evaluasi

Untuk memperoleh model klasifikasi terbaik, proses *training* oleh setiap algoritma *machine learning* dilakukan dengan 5 skema, dengan mengganti komposisi data latih dan data uji (Tabel 5).

Tabel 5. Skema Percobaan

Skema	Komposisi Data latih: Data Uji
I	90:10
II	80:20
III	70:30
IV	60:40
V	50:50

Setiap skema dari setiap algoritma *machine learning* dievaluasi menggunakan akurasi (persamaan (3)), *precision* (persamaan (4)), *recall* (persamaan (5)) dan *F1* (persamaan (6)). Pengukuran tersebut berdasarkan *confusion matrix* (Tabel 6). Nantinya pengukuran performansi setiap skema dan algoritma berdasarkan nilai *confusion matrix*.

Tabel 6. Confusion Matrix

		Prediksi			
		Positif		Negatif	
Aktual	Positif	True (TP)	Positive (FP)	False (FN)	Negative (TN)
	Negatif	False (FP)	Positive (TP)	True (TN)	Negative (FN)

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \cdot 100\% \quad (3)$$

$$P = \frac{TP}{TP + FP} \cdot 100\% \quad (4)$$

$$Recall(R) = \frac{TP}{TP + FN} \cdot 100\% \quad (5)$$

$$F1 \text{ score} = \frac{2PR}{P + R} \quad (6)$$

## 3. HASIL DAN PEMBAHASAN

### 3.1 Hasil Algoritma Naïve Bayes

Percobaan pertama merupakan implementasi algoritma Naïve Bayes terhadap *dataset* yang telah dilakukan normalisasi. Naïve Bayes merupakan algoritma berdasarkan

probabilitas, bukan berdasarkan parameter tertentu. Performansi model klasifikasi dengan beberapa skema ditunjukkan pada Tabel 7. Nilai pada tabel tersebut berdasarkan *confusion matrix* dan formula setiap parameter performansi yang telah dijelaskan di bagian sebelumnya. Berdasarkan tabel tersebut, model klasifikasi terbaik diperoleh dari skema IV, dengan jumlah data latih 60% dari keseluruhan data dan sisanya sebagai data uji. Adapun nilai akurasi, *precision*, *recall* dan *F1* secara berturut turut ialah 78.9, 75, 76 dan 75.5 persen. Dengan kata lain, performansi algoritma Naïve Bayes dalam mengidentifikasi penderita diabetes adalah nilai tersebut pula.

Tabel 7. Performansi Algoritma Naive Bayes (%)

Skema	Akurasi	Precision	Recall	F1
I	75.32	75	72.5	73
II	74.68	72.5	70.5	71
III	76.62	74	73	73.5
IV	<b>78.9</b>	<b>75</b>	<b>76</b>	<b>75.5</b>
V	77.08	74	73.5	73.5

### 3.2 Hasil Algoritma Decision Tree

Percobaan kedua merupakan implementasi algoritma *Decision Tree* terhadap *dataset* yang telah dinormalisasi. Performansi model klasifikasi dengan beberapa skema ditunjukkan pada Tabel 8. Nilai pada tabel tersebut berdasarkan *confusion matrix* dan formula setiap parameter performansi yang telah dijelaskan di bagian sebelumnya. Berdasarkan tabel tersebut, model klasifikasi terbaik diperoleh dari skema II, dengan jumlah data latih 80 % dari keseluruhan data dan sisanya sebagai data uji. Adapun nilai akurasi sebesar 75.32%, sedangkan nilai *precision*, *recall* dan *F1* ialah 73%. Dengan kata lain, performansi algoritma *Decision Tree* dalam mengidentifikasi penderita diabetes adalah nilai tersebut pula.

Tabel 8. Performansi Algoritma Decision Tree (%)

Skema	Akurasi	Precision	Recall	F1-score
I	68.83	67.5	66	66.5
II	<b>75.32</b>	<b>73</b>	<b>73</b>	<b>73</b>
III	74.03	71	72	71.5
IV	70.45	67	69.5	67
V	71.35	68	70.5	69

### 3.3 Hasil Algoritma Support Vector Machine (SVM)

Percobaan ketiga merupakan implementasi algoritma *SVM* terhadap *dataset* yang telah dilakukan normalisasi. Adapun parameter yang digunakan pada algoritma *SVM* antara lain *regular parameter* (*C*) adalah 1.0, kernel function berupa *linear* dan nilai *auto gamma*. Performansi model klasifikasi dengan beberapa skema ditunjukkan pada Tabel 9. Nilai pada tabel tersebut berdasarkan *confusion matrix* dan

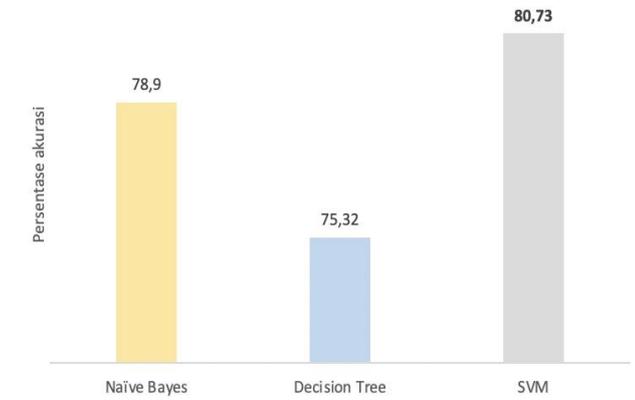
formula setiap parameter performansi yang telah dijelaskan di bagian sebelumnya. Berdasarkan tabel tersebut, model klasifikasi terbaik diperoleh dari setengah *dataset*. Adapun nilai akurasi, *precision*, *recall* dan *F1* secara berturut turut ialah 80.73, 79.5, 74.5 dan 76 persen. Dengan kata lain, performansi algoritma *SVM* dalam mengidentifikasi penderita diabetes adalah nilai tersebut pula.

**Tabel 9.** Performansi Algoritma *SVM* (%)

Skema	Akurasi	<i>Precision</i>	<i>Recall</i>	<i>F1-score</i>
I	77.92	79	74.5	75.5
II	75.32	74.5	70	71
III	77.92	77	72	73
IV	79.55	76.5	73	74.5
V	80.73	79.5	74.5	76

### 3.4 Analisis Perbandingan Algoritma

Bagian ini menyajikan analisis perbandingan performansi ketiga algoritma. Berdasarkan tiga percobaan yang telah dilakukan, kita tidak dapat menentukan skema terbaik dalam menghasilkan model klasifikasi terbaik. Dengan kata lain, model klasifikasi terbaik dapat berasal dari komposisi data latih mana pun. Namun, penelitian oleh [13] tidak menyarankan menggunakan komposisi data latih yang terlalu sedikit.



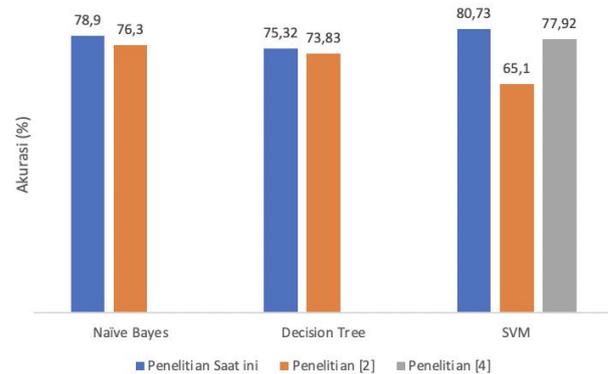
**Gambar 2.** Perbandingan Performansi Algoritma

Selanjutnya, Gambar 2 menyajikan perbandingan performansi antar algoritma *machine learning* dalam mengidentifikasi penderita diabetes. Berdasarkan gambar tersebut, algoritma *SVM* lebih unggul dibandingkan dengan algoritma Naive Bayes dan *Decision Tree*, dengan akurasi 80.73%.

### 3.5 Analisis Perbandingan dengan Penelitian Sebelumnya

Bagian ini menyajikan analisis hasil penelitian yang dilakukan dengan penelitian sebelumnya. Dengan menggunakan *dataset* yang sama, penelitian[2] mengidentifikasi penderita diabetes dengan menggunakan algoritma Naive Bayes, *Decision Tree* dan *SVM* serta tidak

menerapkan proses normalisasi. Sedangkan penelitian [4] menggunakan algoritma *SVM* serta metode *Min-Max* untuk mengidentifikasi penderita diabetes.



**Gambar 3.** Perbandingan dengan Penelitian Sebelumnya

Gambar 3 menyajikan perbandingan hasil penelitian saat ini dengan penelitian sebelumnya. Penelitian saat ini memiliki akurasi sedikit lebih unggul dibandingkan dengan hasil penelitian sebelumnya. Metode *Z-Score* pada semua algoritma *machine learning* yang diujicoba memiliki akurasi model klasifikasi yang lebih baik dibandingkan dengan tanpa metode *Z-Score*. Lebih lanjut, metode normalisasi *Z-Score* pada algoritma *SVM* memiliki akurasi hasil lebih besar (80.73%) dibandingkan dengan metode *Min-Max* (77.92%) untuk mengidentifikasi penderita diabetes (*PIMA Indian Dataset*).

## 4. KESIMPULAN

Berdasarkan hasil percobaan yang telah dilakukan, dapat disimpulkan bahwa algoritma *machine learning* dapat digunakan pada bidang Kesehatan, yakni dalam mengidentifikasi penderita diabetes. Setiap algoritma klasifikasi memiliki karakteristik yang berbeda-beda sehingga menghasilkan performansi yang berbeda-beda pula. Setelah melakukan pengujian terhadap tiga algoritma klasifikasi dengan lima skema, ditemukan bahwa kombinasi algoritma *SVM* dan *Z-Score* merupakan kombinasi terbaik sehingga memiliki performansi terbaik dalam mengidentifikasi penderita diabetes dibandingkan dengan algoritma Naive Bayes dan *Decision Tree*, dengan akurasi 80.73% dan *F1* 76 %. Lebih lanjut, *Z-Score* berkontribusi positif untuk meningkatkan akurasi, hal ini diperlihatkan dengan adanya kombinasi algoritma *SVM* dan *Z-Score* lebih unggul dibandingkan dengan hasil penelitian sebelumnya.

## DAFTAR PUSTAKA

- [1] A. M. Argina, "Penerapan Metode Klasifikasi K-Nearest Neighbor pada Dataset Penderita Penyakit Diabetes," *Indonesian Journal of Data and Science*, vol. 1, no. 2, 2020, doi: 10.33096/ijodas.v1i2.11.
- [2] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," in *Procedia Computer Science*, 2018, vol. 132. doi: 10.1016/j.procs.2018.05.122.

- [3] D. Tomar and S. Agarwal, "A survey on data mining approaches for healthcare," *International Journal of Bio-Science and Bio-Technology*, vol. 5, no. 5, 2013, doi: 10.14257/ijbsbt.2013.5.5.25.
- [4] D. A. Agatsa, R. Rismala, and U. N. Wisesty, "Klasifikasi Pasien Pengidap Diabetes Metode Support Vector Machine," *e-Proceeding of Engineering*, vol. 7, no. 1, 2020.
- [5] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN," *Computer Engineering, Science and System Journal*, vol. 4, no. 1, 2019, doi: 10.24114/cess.v4i1.11458.
- [6] H. Henderi, "Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer," *IJIS: International Journal of Informatics and Information Systems*, vol. 4, no. 1, 2021, doi: 10.47738/ijis.v4i1.73.
- [7] I. M. K. Karo, A. Khosuri, and R. Setiawan, "Effects of Distance Measurement Methods in K-Nearest Neighbor Algorithm to Select Indonesia Smart Card Recipient," 2021. doi: 10.1109/ICoDSA53588.2021.9617476.
- [8] B. Muhamad Akbar, A. T. Akbar, and R. Husaini, "Analisis Sentimen dan Emosi Vaksin Sinovac pada Twitter menggunakan Naïve Bayes dan Valence Shifter," *Jurnal Teknologi Terpadu*, vol. 7, no. 2, pp. 83–92, 2021.
- [9] F. W. Townes, S. C. Hicks, M. J. Aryee, and R. A. Irizarry, "Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model," *Genome Biol*, vol. 20, no. 1, 2019, doi: 10.1186/s13059-019-1861-6.
- [10] A. Rana and R. Pandey, "A review of popular decision tree algorithms in data mining," *Asian Journal of Multidimensional Research*, vol. 10, no. 10, 2021, doi: 10.5958/2278-4853.2021.00837.5.
- [11] I. M. K. Karo, M. Y. Fajari, N. U. Fadhilah, and W. Y. Wardani, "Benchmarking Naïve Bayes and ID3 Algorithm for Prediction Student Scholarship," *IOP Conf Ser Mater Sci Eng*, vol. 1232, no. 1, p. 012002, Mar. 2022, doi: 10.1088/1757-899X/1232/1/012002.
- [12] I. M. Karo Karo, S. Nadia Amalia, dan Dian Septiana, P. Ilmu Komputer, and P. Matematika, "Klasifikasi Kebakaran Hutan Menggunakan Feature Selection dengan Algoritma K-NN, Naive Bayes dan ID3," *Journal of Software Engineering, Information and Communication Technology*, vol. 3, no. 1, pp. 121–126, 2022.
- [13] I. M. Karo Karo, M. F. M. Fudzee, S. Kasim, and A. A. Ramli, "Sentiment Analysis in Karonese Tweet using Machine Learning," *Indonesian Journal of Electrical Engineering and Informatics*, vol. 10, no. 1, pp. 219–231, Mar. 2022, doi: 10.52549/ijeei.v10i1.3565.