Jurnal Krisnadana
Volume 3 Nomor 1, September 2023
e-ISSN 2808-3563 p-ISSN 2808-3520
https://ejournal.sidyanusa.org/index.php/jkdn

Deteksi Penyakit Diabetes Mellitus Menggunakan Algoritma Decision Tree Model Arsitektur C4.5

Achmad Afifuddin[1*], Lukman Hakim[2]
[1*,2]*Teknik Informatika, Teknik, Universitas Yudharta Pasuruan*
[1*]*uachmadafif@gmail.com* , [2]*lukman@yudharta.ac.id*

Abstract

***D**iabetes mellitus (DM) is a metabolic disorder characterized by an increase in blood sugar levels due to disruptions in insulin secretion, insulin function, or both. Diabetes affects individuals of all age groups, thus necessitating reliable diagnostic tools for prevention and management. One of the approaches is to utilize technological advancements in the field. This research focuses on diagnosing diabetes using an application employing the C4.5 decision tree algorithm. The Decision Tree C4.5 is utilized in the model to predict a tree-like hierarchical structure that transforms data into decision trees and decision rules. The dataset for this study consists of 2000 samples collected from Kaggle. The results of this research demonstrate that the predictive model using the Decision Tree C4.5 algorithm achieves a 96% accuracy rate by utilizing 5 variables. As a result of this accuracy, a diabetes mellitus detection application is developed to facilitate independent self-detection before consulting a medical professional.*

Keywords: *Diagnosis, Diabetes Mellitus, Decision Tree, C4.5, Application*

Abstrak

***D**iabetes mellitus (DM) merupakan penyakit metabolik yang ditandai dengan peningkatan kadar gula darah akibat gangguan pada sekresi insulin, kerja insulin atau keduanya. Penyakit diabetes menyerang dari segala kalangan usia, sehingga di perlukannya alat diagnosis baik untuk pencegahan, penanggulangan pada seseorang yang terdampak, salah satunya menggunakan bidang ilmu teknologi. Pada penelitian ini dilakukannya diagnosis pada penyakit diabetes menggunakan aplikasi dengan metode algoritma decision tree C4.5. Decision Tree C4.5 digunakan dalam model untuk memprediksi sebuah struktur pohon atau hirarki untuk mengubah data menjadi pohon keputusan dan aturan-aturan keputusan. Jenis pengumpulan dataset mengambil dari Kaggle sebanyak 2000. Hasil dari penelitian ini menunjukkan pada model prediksi algoritma Decision Tree C4.5 memiliki akurasi 96% dengan Menggunakan 5 variabel, maka dari hasil akurasi tersebut dibuatkan aplikasi deteksi penyakit diabetes mellitus guna untuk mendeteksi secara mandiri sebelum pergi kedokter.*

Received: **24/08/2023**; Revised: **25/09/2023**; Accepted: **27/09/2023**

Kata Kunci: *Diagnosis, Diabetes Mellitus, Decision Tree, C4.5, Aplikasi*

PENDAHULUAN

Perhatian terhadap kesehatan menjadi bagian terpenting dalam faktor kehidupan. Kesehatan menjadi bagian dalam kelancaran atau keberhasilan kegiatan seseorang. Penyakit menyerang tidak memperhatikan usia, salah satunya penyakit yang menyerang metabolik dengan gangguan pada sekresi insulin, kerja insulin atau keduanya sehingga terjadi peningkatan gula darah penyakit tersebut yakni Diabetes Melitus (DM) [1].

Penderita penyakit Diabetes Melitus dapat terserang penyakit lainya sepeti hipertensi, jantung, strok, retinopati, kanker, ginjal dan beberapa penyakit lainnya. Umumnya penyakit Diabetes Melitus terbagi menjadi 2, tipe 1 penyakit tersebut dikarenakan kerusakan pada sel beta pankreas akibat faktor autoimun, genetik atau idiopatik, pada tipe 2 penyakit tersebut diakibatkan perubahan gaya hidup sehingga umumnya timbul akibat resistensi insulin. dengan demikian, perlunya menjaga kesehatan afar terhindar dari serangan penyakit Diabetes Melitus [2]

Word Health Organization (WHO) Indonesia mengalami lonjakan 21,3 juta penduduk pada tahun 2030 yang terjangkit penyakit Diabetes Melitus dimana pada tahun 2000 penderita penyakit hanya sebesar 8,4 juta. Word Diabetes Assocoation mengemukakan peningkatan prevelensi penyakit Diabetes Melitus di Indonesia sebesar 9,1 juta penderita pada tahun 2014 di mana nantinya meningkat 14,1 juta pada tahun 2035, Adapun grafik dari tahun ke tahun sebagai berikut [3]

Berdasarkan data dari *Word Health Organization* (WHO) menunjukkan peningkatan diabetes dari tahun 2000 terdapat 8,4 Juta yang terserang penyakit diabetes, ditahun 2014 penyakit diabetes meingkat menjadi 9,1 Juta, tahun 2030 indonesia mengalami lonjakan 21,3 Juta penduduk yang terserang, dan nantinya pada tahun 2035 meningkat menjadi 14,1 Juta.

Peningkatan diabetes dari tahun ketahun tidak hanya dikarenakan faktor umur, faktor keturunan namun juga pola hidup yang kurang baik. Penyakit diabetes dapat menyerang kapan saja sehingga perlunya pencegahan dan pengecekan untuk mengetahui apakah sesorang menderita penyakit diabetes dengan kriteria prediabetes ataupun normal sehingga di perlukannya diagnose secara berkala untuk dapat memprediksikan apakah orang tersebut terjangkit penyakit Diabetes melitus [4]. Dalam mendiagnosa penyakit tersebut dapat dilakukan dalam bidang keilmuan lainnya, namun dengan seiring berkembangnya zaman yang semakin canggih, fleksibel dan juga cepat diagnosa penyakit diabetes juga dapat dilakukan pada bidang teknologi seperti pada penggunaan aplikasi dimana salah satunya menggunakan Algoritma *Decision Tree* C4.5 [5]

Algoritma *Decision tree* C4.5 digunakan untuk memprediksi kelas atau objek ketika kelas data dari item baru tidak diketahui, algoritma pohon keputusan C4.5 adalah klasifikasi dari proses menemukan model atau fungsi yang menjelaskan dan membedakan kelas atau konsep data [4]. Decision tree mirip dengan grafik aliran, dengan setiap node berdiri untuk nilai atribut, setiap cabang mempresentasikan untuk hasil tes, dan setiap daun untuk mempresentasikan distribusi kelas atau kelas. Variasi dari metode ID3 yang dikenal sebagai Decision

Tree C4.5 menggunakan strategi *greedy* dengan pengambilan keputusan berdasarkan pohon yang dibuat menggunakan pendekatan rekursif *top down* dan sistem bagi serang [6]

Dengan hal ini dapat di ketahui bahwasannya penyakit diabetes meningkat di semua kalangan mulai dari 2013 sampai dengan 2018. Peningkatan diabetes dari tahunketahun tidak hanya dikarenakan faktor umur, faktor keturunan namun juga pola hidup yang kurang baik [7]. Penyakit diabetes dapat menyerang kapan saja sehingga perlunya pencegahan dan pengecekan untuk mengetahui apakah sesorang menderita penyakit diabetes dengan kriteria prediabetes ataupun normal sehingga di perlukannya media aplikasi untuk mengetahui tingkatan diabetes sesorang dengan lebih mudah serta memberikan informasi tentang pengetahuan terkait penyakit diabetes

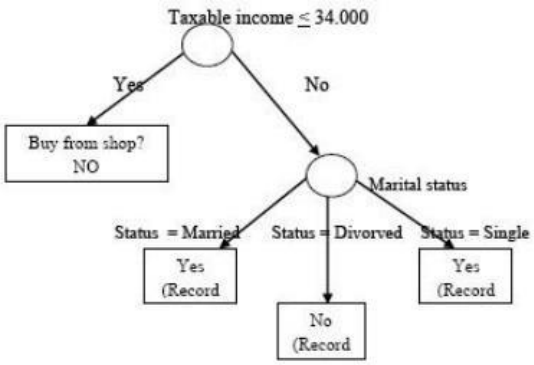
Penelitian sebelumnua, analisis dan perbandingan teknik penambangan data seperti "Method Decision Tree, Naive Bayes, Smo dan Part" untuk mengidentifikasi diabetes mellitus. Tujuan dari penelitian ini adalah untuk memilih klasifikasi data mining untuk mendiagnosis Diabetes Mellitus (DM) [8]. Diagnosis dilakukan pada basis komputer menggunakan teknik pemilihan fitur dan klasifikasi terhadap Dataset Diabetes Pima Indians. Metode pemilihan fitur/seleksi fitur yang digunakan disebut Correlation-based Featured Selection (CFS). Dibandingkan dengan klasifikator lain, SMO memiliki nilai akurasi terbesar, menurut hasil klasifikasi data mining dalam penelitian ini dibandingkan Classifiers lainnya [9]

Pada penelitian terdahulu penggunaan aplikasi pendeteksi penyakit diabetes ditujukan hanya sebagai alat bantu pada dokter saja, sedangkan pada penelitian ini aplikasi deteksi penyakit diabetes mellitus digunakan untuk mendiagnosis secara mandiri sebelum pergi ke dokter dengan menggunakan 5 variabel diantaranya adalah Pregnancies, Glucose, Boold Pressure, BMI, Age, dan tanpa harus mengetahui tentang Skin Tehickness, Insulin, dan Diabetes Pedigree Fungsi.

Berdasarkan pemaparan diatas maka terdapat tujuan penelitian yaitu untuk merancang aplikasi deteksi penyakit diabetes mellitus guna dapat mempermudah pasien dalam mendiagnosis penyakit diabetes milletus dengan mandiri sebelum pergi kedokter.

METODE PENELITIAN

Penelitian ini menggunakan Decision Tree C4.5 untuk membangun sebuah pohon keputusan yang lebih mudah untuk dimengertu, fleksibel, dan menarik. *Decision Tree* atau pohon keputusan adalah model untuk memprediksi sebuah struktur pohon atau hirarki untuk mengubah data menjadi pohon keputusan dan aturan-aturan keputusan [10].



Gambar 2. Contoh konsep pohon keputusan sederhana

Ada Beberapa tahap untuk membuatpohon keputusan dengan algoritma *Decision Tree* adalah sebagai berikut:

1. Mempersiapkan data training, bisa diambil dari data *history* yang sudah pernah terjadi sebelumnya dan sudahdikelompokkan dalam bentuk kelas- kelas tertentu.
2. Menentukan akar pohon dengan menghitung nilai *gain* yang tertinggi untuk masing-masing atribut atauberdasarkan nilai *index entropy*terendah. Sebelumnya telah dihitung terlebih dahulu nilai index entropy, dengan rumus:

$$\text{Entropy (s)} = \sum_{i=1}^{n} - pi * \log_2 pi \quad (1)$$

3. Hitung nilai *gain* dengan rumus:

$$\text{Gain (S, A)} = \text{Entropy (S)} - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si) \quad (2)$$

4. Untuk menghitung *gain ratio* diperlukan suatu term baru yang disebut *Split Information* dengan rumus:

$$SplitInfo\ (S, A) = \sum_{n=1}^{c} - \frac{Si}{S} \log_2 \frac{Si}{S} \quad (3)$$

5. elanjutnya menghitung *ratio:*

$$Gain\ ratio(S.A) = \frac{Gain(S.A)}{Split\ Information(S.A)} \quad (4)$$

6. Mengulangi Langkah ke-2 hingga semua *record* terpartisi

HASIL DAN PEMBAHASAN

Hasil penelitian ini menggunakan Algoritma C4.5 dengan dataset berjumlah 2000 dan menggunakan 5 variabel, dataset diabetes ini akan dibuatkan aplikasi pendeteksi penyakit diabetes untuk mendiagnosis penyakit tersebut. Sebelum dibuat aplikasi maka dataset tesebut dicari tingkat akurasinya.

A. Dataset Diabetes

Tabel 1 Pada proses ini terdapat dataset penyakit diabetes yang berjumlah 2000 data dan menggunakan 5 variabel. Data tersebut akan digunakan mencari nilai entropy, gain, Split Info, Gain Ratio. Dapat dilihat pada tabel 1.

Table 1: Dataset Dabetes
[Sumber: Kaggle.com]

No	Pregnancies	Glucose	Blood Pressure	BMI	Age	Outcome
1	6	148	72	33.06.00	50	1
2	1	85	66	26.06.00	31	0
3	8	183	64	23.03	32	1
4	1	89	66	28.01.00	21	0
5	0	137	40	43.01.00	33	1
6	5	116	74	25.06.00	30	0
7	3	78	50	31	26	1
8	10	115	0	35.03.00	29	0
9	2	197	70	30.05.00	53	1
10	8	125	96	0	54	1
11	4	110	92	37.06.00	30	0
12	10	168	74	38	34	1
13	10	139	80	27.01.00	57	0
14	1	189	60	30.01.00	59	1
15	5	166	72	25.08.00	51	1
16	7	100	0	30	32	1
17	0	118	84	45.08.00	31	1
18	7	107	74	29.06.00	31	1
19	1	103	30	43.03.00	33	0
20	1	115	70	34.06.00	32	1

Berdasarkan tabel 1 dapat diketahui jumlah dataset penyakit diabetes yang digunakan sebagai dataset penelitian.

B. Perhitungan Dataset Diabetes

			Jumlah (S)	YA (Si)	Tidak (Si)	Entropy	Gain	Split Info	Gain Ratio
Total			2000	684	1316	0,926719721			
Pregnancies							5,44613E-05	0,611090142	8,91216E-05
	Belum Pernah	0	301	100	201	0,917184431			
	Pernah	>1	1699	584	1115	0,928344913			
Glucose							0,13102029	1,57267878	0,083310268
	Normal	100	569	58	511	0,475093959			
	Prediabetes	100-125	650	179	471	0,849085547			
	Diabetes	>126	781	447	334	0,984846087			
Boold Pressure							0	0	0
	Normal	<140	2000	684	1316	0,926719721			
	Diatas Normal	>140	0	0	0	0			
BMI							0,067394032	0,685515678	0,098311437
	Normal	18,5 - 24,9	365	24	341	0,349872589			
	Berlebihan	>25	1635	660	975	0,97305681			
Age							0,06167649	1,525246666	0,04043706
	Remaja	17-25 Tahun	717	122	595	0,658055612			
	Dewasa	26-45 Tahun	989	420	569	0,983564635			
	Lansia	46-65 Tahun	253	132	121	0,998635964			
	Manula	65 Ke Atas	41	10	31	0,801469893			

Gambar 3. Perhitungan dataset diabetes

Gambar 3 adalah cara perhitungan entropy total dan gain dari dataset diabetes pengerjaan dataset sebagai berikut:

a. Baris total kolom Entropy pada Tabel dihitung dengan rumus sebagai berikut:

   Entropy =

((-Jumlah Data Training Ya/Jumlah Data Tarining)*IMLOG2(Jumlah Data Training Ya/Jumlah Data Training) + (-Jumlah Data Traning Tidak/Jumlah Data Training)*IMLOG2(Jumlah Data Training Tidak/Jumlah Data Training))

Entropy Jumlah Data =

$$\left(\left(-\frac{684}{2000}\right)*\text{IMLOG2}\left(\frac{684}{2000}\right)+\left(-\frac{1316}{2000}\right)*\text{IMLOG2}\left(\frac{1316}{2000}\right)\right)=0{,}926719721$$

Entropy Jumlah Data Belum Pernah Hamil =

$$\left(\left(-\frac{100}{301}\right)*\text{IMLOG2}\left(\frac{100}{301}\right)+\left(-\frac{201}{301}\right)*\text{IMLOG2}\left(\frac{201}{301}\right)\right)=0{,}917184431$$

Entropy Jumlah Data Pernah Hamil =

$$\left(\left(-\frac{584}{1699}\right)*\text{IMLOG2}\left(\frac{584}{1699}\right)+\left(-\frac{1115}{1699}\right)*\text{IMLOG2}\left(\frac{1115}{1699}\right)\right)=0{,}928344913$$

b. Sementara itu untuk nilai Gain dihitung dengan menggunakan rumus Gain sebagai berikut:

Gain =

(Entropy Jumlah Data Training)-((Jumlah Total Belum Pernah Hamil/Jumlah Data Training)*Entropy Jumlah Data Belum Pernah Hamil)-((Jumlah Data Pernah Hamil/Jumlah Data Training)*Entropy Jumlah Data Pernah Hamil)

Gain =

$$(0{,}926719721)-\frac{301}{2000}*0{,}917184431-\frac{1699}{2000}*0{,}928344913=5{,}44613E-05$$

c. Untuk menghitung *Split Info* dengan menggunakan rumus sebagai berikut:

Split Info =

-(Jumlah Total Belum Pernah Hamil/Jumlah Data Training)*IMLOG2(Jumlah Total Belum Pernah Hamil/Jumlah Data Training)-(Jumlah Total Pernah Hamil/Jumlah Data Training)*IMLOG2(Jumlah Total Pernah Hamil/Jumlah Data Training)

Split Info =

$$-\left(\frac{301}{2000}\right)*\text{IMLOG2}\left(\frac{301}{2000}\right)-\left(\frac{1699}{2000}\right)*\text{IMLOG2}\left(\frac{1699}{2000}\right)=0.611090142$$

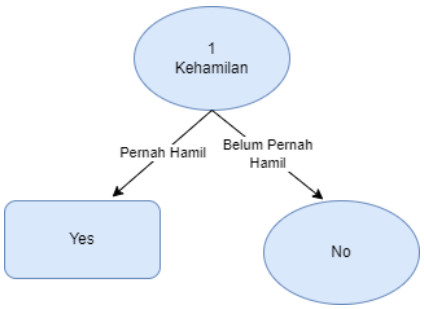
d. Sedangkan untuk menghitung *Gain Ratio* dengan menggunakan rumus sebagai berikut:

Hasil Gain/Split Info

Gain Ratio =

$$\frac{5{,}44613E-05}{0{,}611090142}=8{,}91216E-05$$

C. Decision Tree Pohon Keputusan



Gambar 4.

```
X_validation = validation_data.drop("outcome", axis=1)
y_validation = validation_data["outcome"]
y_pred_validation = model.predict(X_validation)
accuracy_validation = accuracy_score(y_validation, y_pred_validation)
print("Validation Accuracy: %d persen" %(accuracy_validation*100))
```

```
Validation Accuracy: 90 persen
```

hon keputusan yang sudah terbentuk:

- ❖ Jika Kel
- ❖ Jika Kehamilan = Pernah Hamil maka Yes

Berdasarkan gambar 4 dapat dijelaskan dari pohon keputusan diatas adalah rata-rata yang terkena penyakit diabetes adalah dari variable kehamilan, apabila seseorang tersebut sering hamil maka terserang, apabila seseorang tersebut tidak pernah maka tidak terserang.

D. Hasil Akurasi Dataset

1. Hasil Akurasi Data Testing

Gambar 5 merupakan hasil dataset *testing* dengan akurasi 90%. Dengan menggunakan 200 dataset.

```
X_test = test_data.drop("outcome", axis=1)
y_test = test_data["outcome"]
y_pred_test = model.predict(X_test)
accuracy_test = accuracy_score(y_test, y_pred_test)
print("Testing Accuracy: %d persen" %(accuracy_test*100))
```

```
Testing Accuracy: 90 persen
```

Gambar 5. Hasil akurasi data testing

2. Hasil Akurasi Data Validation

Gambar 6 merupakan hasil dataset *validation* dengan akurasi 90%. Dengan menggunakan 200 dataset.

```
X_validation = validation_data.drop("outcome", axis=1)
y_validation = validation_data["outcome"]
y_pred_validation = model.predict(X_validation)
accuracy_validation = accuracy_score(y_validation, y_pred_validation)
print("Validation Accuracy: %d persen" %(accuracy_validation*100))
```

```
Validation Accuracy: 90 persen
```

Gambar 6. Hasil akurasi data validation

3. Hasil Akurasi Dataset

Gambar 7 merupakan hasil akurasi dataset diabetes. Untuk hasil akurasi precision mencapai 0.99, recall 0.95, f1-score 0.97, dan untuk hasil akurasi untuk pembuatan aplikasi ini mencapai 96%.

```
Confusion Marix
[[255  14]
 [  2 129]]
Tingkat Akurasi Algoritma C4.5
Akurasi :              precision    recall  f1-score   support

           0       0.99      0.95      0.97       269
           1       0.90      0.98      0.94       131

    accuracy                           0.96       400
   macro avg       0.95      0.97      0.96       400
weighted avg       0.96      0.96      0.96       400

Tingkat Akurasi: 96 persen
```

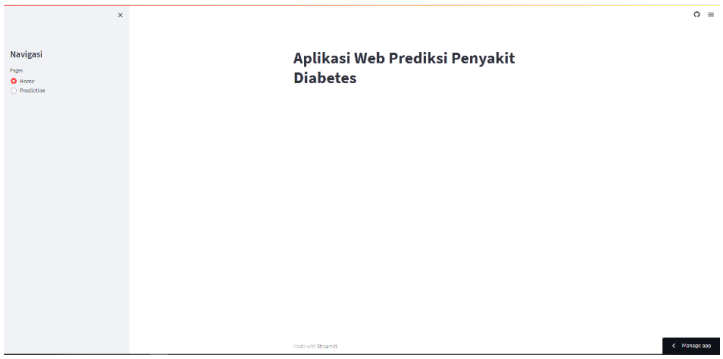
Gambar 7. Hasil akurasi dataset

E. Implementasi Aplikasi

Implementasi aplikasi adalah sebuah aplikasi yang dibuat oleh Bahasa pemograman python menggunakan framework streamlit untuk mempermudah atau mempercepat pembuatan aplikasi

1. Halaman Dashboard

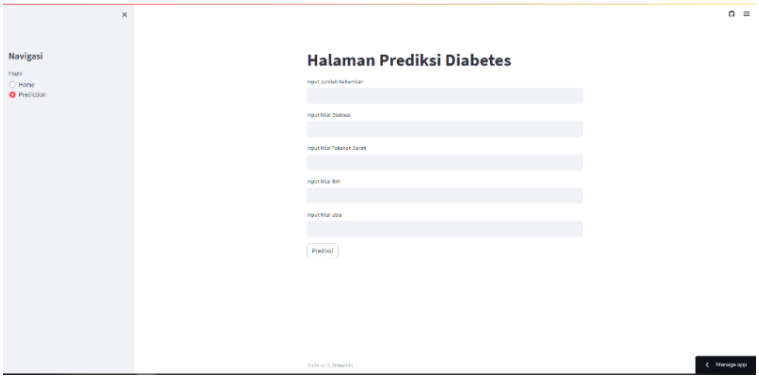
Halaman dashboard ini menampilkan sebuah halaman utama pada aplikasi web prediksi penyakit diabetes.



Gambar 8. Halaman dashboard

2. Halaman Prediksi

Halaman prediksi digunakan untuk memprediksi penyakit diabetes



Gambar 9. Halaman prediksi

KESIMPULAN

Dari pembuatan aplikasi deteksi penyakit diabetes dengan penerapan algoritma klasifikasi C4.5 dapat digunakan untuk membatu seseorang dalam mendiagnosis pertama sebelum pergi ke dokter, dapat disipulkan bahwa 1) Algoritma C4.5 dapat digunakan untuk memudahkan dalam pengambilan keputusan dengan memproyeksikan data-data yang ada ke dalam bentuk pohon keputusan, berdasarkan nilai entropy dan gain yang dimiliki masing-masing atribut data. 2) Dalam hasil prediksi yang lebih akurat dibutuhkan data dalam jumlah besar, artinya semakin besar jumlah data yang digunakan maka semakin akurat hasil prediksi yang

dihasilkan. 3) Aplikasi deteksi penyakit diabetes mellitus dapat di prediksi dengan memanfaatkan Algoritma C4.5 dengan hasil presentase memiliki tingkat akurasi sebesar 96%.

DAFTAR PUSTAKA

[1] F. M. Hana, "Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4 . 5," 2020.

[2] M. T. Informatika and U. G. Jakarta, "Perbandingan Hasil Analisis Teknik Data Mining ' Metode Decision Tree , Naive Bayes , Smo Dan Part ' Untuk Mendiagnosa Penyakit Diabetes Mellitus," vol. 4, no. 1, pp. 38–44, 2019, doi: 10.25139/inform.v3i2.1010.

[3] H. Y. Resti, W. H. Cahyati, and I. Artikel, "Kejadian Diabetes Melitus pada Usia Produktif di Puskesmas Kecamatan Pasar Rebo," *Higeia J. Public Heal. Res. Dev.*, vol. 6, no. 3, pp. 350–361, 2022.

[4] P. Arsi and O. Somantri, "Deteksi Dini Penyakit Diabetes Menggunakan Algoritma Neural Network Berbasiskan Algoritma Genetika," vol. 03, no. 03, pp. 290–294, 2018, doi: 10.30591/jpit.v3i3.1008.

[5] Z. Fadilah and Murnawan, "Performance Comparison of K-Nearest Neighbor and Decision Tree C4.5 by Utilizing Particle Swarm Optimization for Prediction of Liver Disease," *Int. J. Open Inf. Technol.*, vol. 9, no. 10, pp. 9–15, 2021.

[6] S. Supangat, A. R. Amna, and T. Rahmawati, "Implementasi Decision Tree C4.5 Untuk Menentukan Status Berat Badan dan Kebutuhan Energi Pada Anak Usia 7-12 Tahun," *Teknika*, vol. 7, no. 2, pp. 73–78, 2018, doi: 10.34148/teknika.v7i2.90.

[7] N. Nurdiana and A. Algifari, "Studi Komparasi Algoritma Id3 Dan Algoritma Naive Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus," *INFOTECHjournal*, vol. 6, no. 2, pp. 18–23, 2020.

[8] U. Pujianto, A. L. Setiawan, H. A. Rosyid, and A. M. M. Salah, "Comparison of Naïve Bayes Algorithm and Decision Tree C4.5 for Hospital Readmission Diabetes Patients using HbA1c Measurement," *Knowl. Eng. Data Sci.*, vol. 2, no. 2, p. 58, 2019, doi: 10.17977/um018v2i22019p58-71.

[9] H. Sitorus, V. Yasin, and A. B. Yulianto, "JurnalSainsdanTeknologiWidyaloka Perancangan sistem pakar diagnosis penyakit diabetes berbasis web menggunakan algoritma naive bayes JurnalSainsdanTeknologiWidyaloka," vol. 1, pp. 135–144, 2022.

[10] A. Rohman and A. Rufiyanto, "Implementasi Data Mining Dengan Algoritma Decision Tree C4 . 5 Untuk Prediksi Kelulusan Mahasiswa Di Universitas Pandaran," *Proceeding SINTAK 2019*, pp. 134–139, 2019.