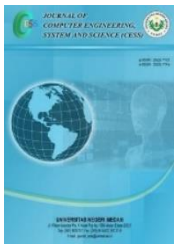
CESS
(Journal of Computing Engineering, System and Science) 7(2) July 2022 315-324

e-ISSN: 2502-714x
p-ISSN: 2502-7131

Contents list available at www.jurnal.unimed.ac.id

CESS
(Journal of Computing Engineering, System and Science)

journal homepage: https://jurnal.unimed.ac.id/2012/index.php/cess



Penerapan Algoritma Decision Tree dan K-Nearest Neighbor (K-NN) Berbasis Particle Swarm Optimization (PSO) Untuk Analisis Akurasi Prediksi Penyakit Diabetes Mellitus

Application of Decision Tree and K-Nearest Neighbor (K-NN) Algorithm Based on Particle Swarm Optimization (PSO) for Diabetes Mellitus Prediction Accuracy Analysis

Andi Nur Rachman[1*], Supratman[2], Euis Nur Fitriani Dewi[3]

[1,2,3] Universitas Siliwangi
Jl Siliwangi No 24 Tasikmalaya
email: [1]andy.rachman@unsil.ac.id, [2]supratman@unsil.ac.id, [3] euis.nurfitriani@unsil.ac.id

Submitted: 28 April 2022 | Revision: 15 Juni 2022 | Accepted: 24 Juni 2022

A B S T R A K

Penyakit *Diabetes Mellitus* merupakan penyakit tidak menular, tetapi penyakit ini merupakan salah satu penyakit yang mematikan bagi yang mengidapnya. Penyakit ini disebabkan oleh beberapa factor diantaranya pola makan hidup yang tidak teratur atau berlebihan. Apabila penyakit ini tidak dihentikan, maka penderita penyakit *Diabetes Mellitus* akan semakin memakan para pasien penderita penyakit ini. Menurut WHO atau *World Health Organization,* sekitar 425 juta orang menderita penyakit diabetes, kemudian 1,6 juta kematian setiap tahunnya di akibatkan oleh penyakit diabetes. Kemudian, pada tahun 2016 di Indonesia, kematian yang disebabkan oleh penyakit diabetes sekitar 99 ribu jiwa. Penyakit diabetes pada tahun ke tahun semakin meningkat, jadi perlu adanya sebuah sistem yang dapat membantu medis untuk melakukan klasifikasi terhadap diabetes berdasarkan data kesehatan pasien. Salah satu metode yang dapat digunakan untuk memprediksi penyakit *diabetes mellitus* adalah dengan menggunakan data mining. Data mining merupakan suatu proses yang interaktif untuk memprediksi penyakit *diabetes mellitus.* Prediksi untuk mendiagnosis penyakit ini menggunakan seleksi fitur berbasis *Particle Swarm Optimization* (PSO) pada dataset Kaggle.com dan metode klasifikasi yang digunakan yaitu metode *Decision Tree* dan *K-Nearest Neighbors* (K-NN). Hasil dari penelitian ini menghasilkan nilai akurasi tertinggi sebanyak 79.8% dengan AUC 0.71 dengan menggunakan metode *Decision Tree*, dan untuk menggunakan optimasi metode *K-Nearest Neighbors* (K-NN) menggunakan *Particle Swarm Optimization* (PSO) memiliki nilai akurasi tertinggi sebanyak 77.09%.

**Penulis Korespondensi:*
email: andy.rachman@unsil.ac.id

Kata Kunci: *Diabetes Mellitus; Data Mining; Particle Swarm Optimization.*

A B S T R A C T
Diabetes Mellitus is a non-communicable disease, but this disease is one of the deadly diseases for those who suffer from it. This disease is caused by several factors including an irregular or excessive diet. If this disease is not stopped, people with Diabetes Mellitus will eat more patients with this disease. According to WHO or the World Health Organization, about 425 million people suffer from diabetes, then 1.6 million deaths each year are caused by diabetes. Then, in 2016 in Indonesia, deaths caused by diabetes were around 99 thousand people. Diabetes is increasing year by year, so there is a need for a system that can help doctors to classify diabetes based on patient health data. One method that can be used to predict diabetes mellitus is to use data mining. Data mining is an interactive process to predict diabetes mellitus. Predictions for diagnosing this disease use feature selection based on Particle Swarm Optimization (PSO) on the Kaggle.com dataset. And the classification method used is the Decision Tree and K-Nearest Neighbors (K-NN) methods. The results of this study produce the highest accuracy value of 79.8% with AUC of 0.71 using the Decision Tree method, and to use the optimization of the K-Nearest Neighbors (K-NN) method using Particle Swarm Optimization (PSO) has the highest accuracy value of 77.09%.

Keywords: *Diabetes Mellitus; Data Mining; Particle Swarm Optimization.*

1. PENDAHULUAN

Diabetes adalah sebuah penyakit dimana kadar glukosa didalam tubuh terlalu tinggi. Penyakit ini biasanya diderita oleh orang sudah dewasa, terjadi ketika tubuh menjadi resisten terhadap insulin atau tidak menghasilkan cukup insulin. Faktor lingkungan yang dapat meningkatkan penyakit diabetes mellitus adalah perubahan gaya hidup seseorang, salah satunya kebiasaan makan yang tidak teratur. Diabetes mellitus salah satu penyakit tidak menular [1], tetapi penyakit diabetes mellitus ini penyakit yang berbahaya, yang dapat menyebabkan kematian. Penyakit diabetes mellitus juga merupakan penyakit dari segala penyakit. Penderita diabetes mellitus yang tidak bisa mengontrol kadar gula darahnya akan mengalami komplikasi hiperglikemi atau tingginya kadar gula didalam darah [2].

Menurut WHO atau World Health Organization, sekitar 425 juta orang menderita penyakit diabetes, kemudian 1.6 juta kematian setiap tahunnya di akibatkan oleh penyakit diabetes, kemudian pada tahun 2016 di Indonesia kematian yang disebabkan oleh penyakit diabetes sekitar 99 ribu jiwa. Penyakit diabetes ini dari tahun ke tahun semakin meningkat, sehingga diperlukan adalanya sebuah sistem yang dapat melakukan klasifikasi terhadap diabetes berdasarkan data kesehatan pasien.

Salah satu metode yang dapat digunakan untuk memprediksi sebuah penyakit bisa menggunakan data mining, data mining suatu proses yang interaktif untuk menemukan pola data [3].

Pada penelitian ini digunakan algoritma Decision Tree dan K-Nearest Neighbors (K-NN) dengan menggunakan optimasi Particle Swarm Optimization (PSO), metode optimasi PSO ini digunakan untuk melakukan tuning parameter terhadap data prediksi sehingga menghasilkan model yang optimal yang dapat menentukan klasifikasi penyakit diabetes mellitus.

2. TINJAUAN TEORI

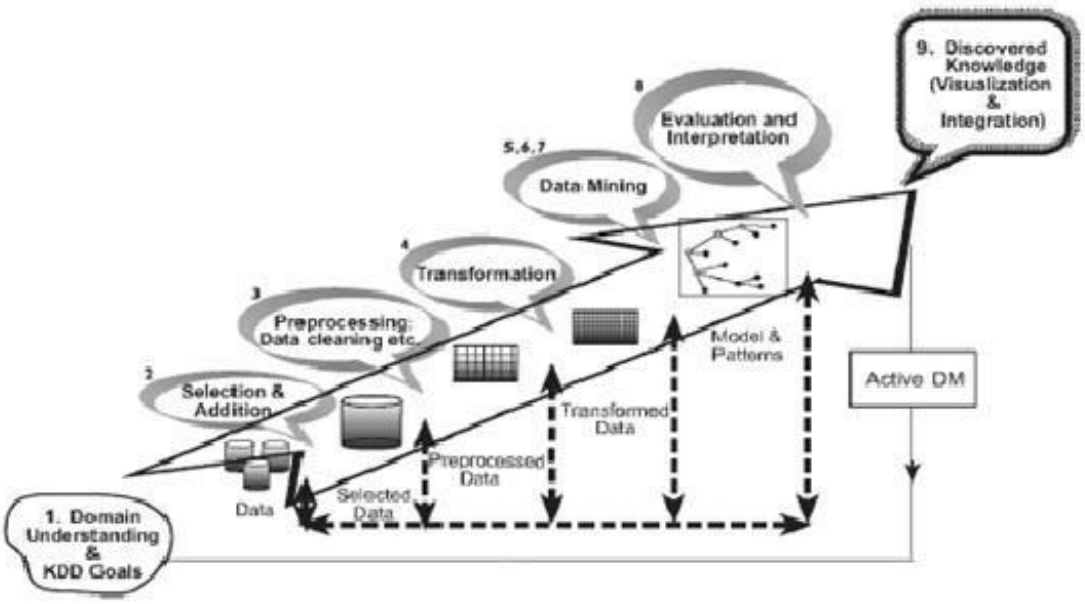
2.1. Diabetes Mellitus

Penyakit *diabetes mellitus* merupakan penyakit metabolic yang ditandai dengan hiperglikemi, hiperglikemi ini kondisi dimana penderita *diabetes mellitus* akan mengalami komplikasi, dimana komplikasi ini selalu diikuti dengan komplikasi 3 penyempitan vaskuler [4].

2.2. Data Mining

Data mining merupakan sekelompok tahapan untuk nenemukan informasi baru dari suatu kumpulan data yang belum diketahui dalam bentuk pengetahuan [5].

Istilah data mining dan *knowledge discovery in database* (KDD) seringkali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Salah satu tahapan dalam keseluruhan proses KDD adalah data mining. Proses KDD secara garis besar dapat dijelaskan sebagai berikut:



Gambar 1. Proses KDD

1. *Data Selection* : pada tahapan pertama ini, dilakukan penyeleksian, pembuatan kelompok data informasi, dan target. Kemudian hasil dari penyeleksian data tersebut disimpan terpisah dari basisdata operasional didalam berkas.
2. *Preprocessing* dan *Cleaning* Data : Meliputi pemilihan data, tugas KDD yaitu untuk mengambil data yang relevan dari *database*. Data *cleaning* berfungsi untuk menghilangkan *noise* dan data *double*, untuk menangani data yang hilang. Dan data *integration* berfungsi untuk menyatukan data dari berbagai sumber
3. *Transformation* : Untuk mengubah data menjadi bentuk yang sesuai, yaitu untuk menemukan fitur yang berguna untuk contoh data.
4. *Data Mining* : Ini merupakan proses paling penting digunakan untuk mengekstraksi pola.

5. *Interpretation/*Evaluasi : Operasi dasar yang termasuk pengidentifikasian pola yang benar-benar menarik yang mewakili pengetahuan, dan menyajikan pengetahuan yang digali dengan cara yang mudah untuk dipahami.

2.3 Decision Tree

Decision Tree atau pohon keputusan merupakan salah satu metode klasifikasi terhadap data [6], D*ecision tree* merupakan metode yang digunakan untuk mengubah data menjadi sebuah pohon keputusan dengan aturan keputusan. Dalam pengembangan klasifikasi dengan *Decision Tree* terdapat beberapa tahapan, yaitu sebagai berikut:

a. Mempersiapkan data training.
b. Menghitung akar dari pohon, untuk menghitung entropy digunakan rumus sebagai berikut:

$$Entropy(S) = \Sigma_{i=1}^{n} - pi * \log 2\, pi$$

Kemudian, hitung nilai gain menggunakan rumus sebagai berikut:

$$Gain(S, A) = Entropy\ (S)\Sigma_{i=1}^{n} \frac{Si}{S} * Entropy(Si)$$

c. Ulangi langkah b dan c hingga semua record terpartisi.
d. Dan partisi akan berhenti.

2.4 K-Nearest Neighbor

K-NN merupakan metode pengklasifikasian yang menggunakan fungsi jarak dari data baru ke data training [7]. Berikut langkah-langkah dari metode K-NN:

a. Menentukan parameter k (jumlah tetangga paling dekat).
b. Menghitung kuadrat jarak Euclid.
c. Mengurutkan objek-objek ke dalam kelompok yang mempunyai jarak Euclid terkecil.
d. Mengumpulkan kategori y (kelas tetangga terdekat).
e. Tentukan label yang frekuensinya paling banyak.

Untuk prediksi penyakit DM dengan menggunakan metode K-NN digunakan model sebagai berikut:

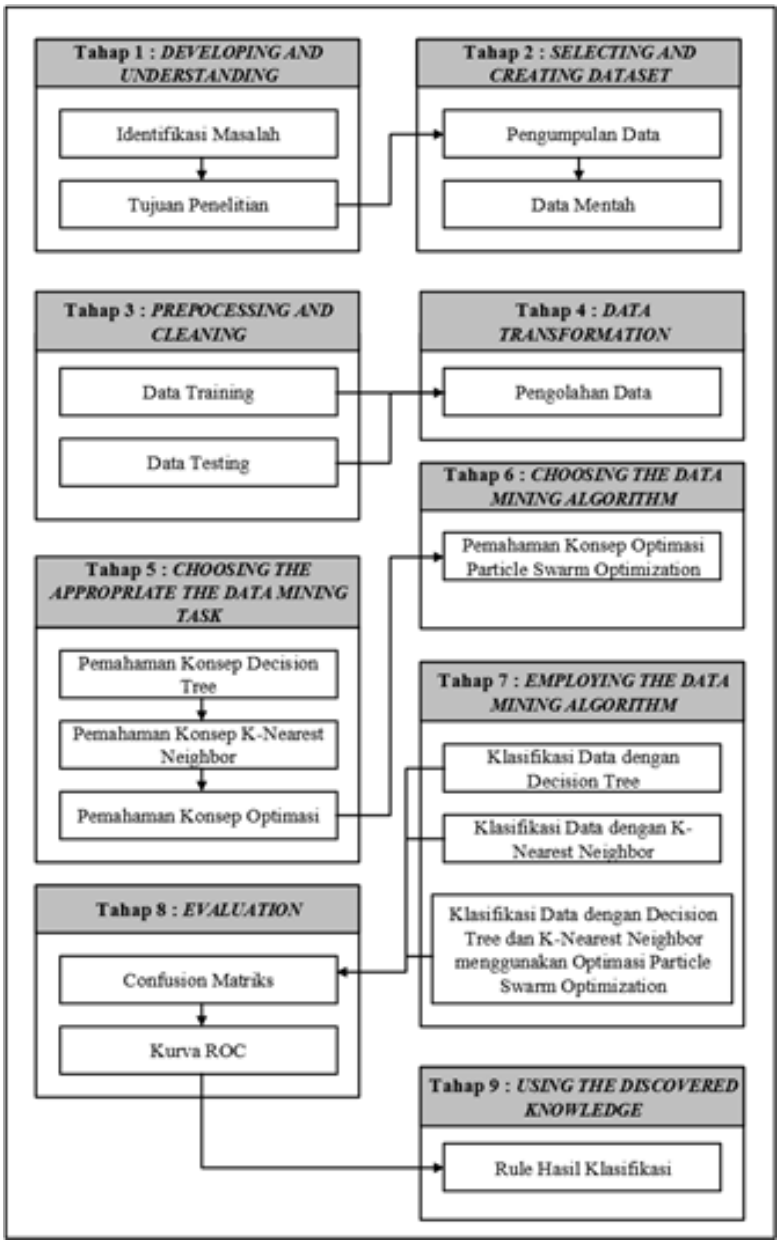
$$dxy = \sqrt{\sum_{i=1}^{n} (x_1 - y_2)2}$$

$$dxfk_{nn}(x) = \frac{1}{k} \Sigma_{iEN\ k(x)} yi$$

2.5 Particle Swarm Optimization

PSO merupakan metode optimasi *heuristic* global yang dikenalkan oleh Dokter Kenneldy dan Eberhart pada tahun 1995 [8]. PSO bisa digunakan sebagai pengambilan keputusan. PSO banyak digunakan untuk memecahkan masalah optimasi bobot dan seleksi fitur.

3. METODE



Gambar 2. Kerangka Konsep Penelitian

1. Tahap *discovering and understanding* bertujuan untuk memahami apa yang akan dilakukan dengan mengidentifikasi dan menetapkan tujuan penelitian.
2. Tahap *selecting and creating dataset,* yang dilakukan dengan cara mengumpulkan data yang akan digunakan dalam penelitian yaitu data penderita diabetes dari repository Kaggle.com.
3. Tahap *prepocessing and cleansing* yaitu memproses data dengan cara memilih data yang akan digunakan dan membuang data yang *missing value* (bernilai kosong atau tidak

lengkap) dan *noise* (tidak tepat). Data klasifikasi dibagi menjadi dua bagian untuk *testing* dan training, dengan komposisi 80% dan 20%.

4. Tahap *data transformation* yang ditujukan untuk mengembangkan data menjadi lebih baik sesuai dengan pola informasi yang dibutuhkan.
5. Tahap *choosing the appropriate data mining task* yaitu proses pemilihan teknik data mining.
6. Tahap *choosing data mining algorithm* yaitu memilih algoritma optimasi, algoritma yang digunakan adalah *Particle Swarm* Optimization (PSO).
7. Tahap *employing data mining algorithm* yaitu membuat klasifikasi pada dataset dengan menggunakan klasifikasi *Decision Tree, K-Nearest Neighbor* menggunakan optimasi algoritma *Particle Swarm* Optimization (PSO).
8. Tahap *evaluation,* mengevaluasi hasil klasifikasi untuk mengukur tingkat akurasi dalam membuat prediksi penyakit diabetes. Evaluasi digunakan dengan *confusion matrix* dan kurva ROC.
9. Tahap *using the discovered knowledge* yaitu hasil dari rule klasifikasi.

4. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan 6 fitur yang diambil dari dataset yang diambil yaitu fitur *pregnancies* (kehamilan), *glucose* (glukosa), *blood pressure* (tekanan darah), BMI, *age* (umur), dan fitur *outcome*. Setelah menentukan ekstrasi fitur dilakukan pengolahan data yang dibagi menjadi data training sebanyak 80% dan data testing 20%. Data penyakit diabetes terdiri menjadi dua kategori, yaitu jika kondisi hasil perhitungan *outcome* menghasilkan data "TIDAK" yang menjadi indikator pengolahan data tersebut tidak mengidap penyakit diabetes, dan jika kondisi *outcome* menghasilkan data "YA" maka pengolahan data tersebut mengidap penyakit diabetes. Tabel 1 merupakan data mentah penentuan penyakit diabetes:

Tabel 1. Data Mentah

Pregnancies	Gluscose	Blood Pressure	BMI	Age	Outcome
6	148	72	33,6	50	Ya
1	85	66	26,6	31	Tidak
8	183	64	23,3	32	Ya
1	89	66	28,1	21	Tidak
0	137	40	43,1	33	Ya
5	116	74	25,6	30	Tidak
3	78	50	31	26	Ya
10	115	0	35,3	29	Tidak
2	197	70	30,5	53	Ya
8	125	96	0	54	Ya
4	110	92	37,6	30	Tidak
10	168	74	38	34	Ya

4.1. Decision Tree dan K-Nearest Neighbor (K-NN)

Pada penelitian ini tahap pertama melakukan pengujian tanpa penerapan metode *Particle Swarm Optimization* (PSO) untuk melihat nilai akurasi pembanding dari pengujian menggunakan metode tanpa seleksi fitur dengan metode PSO. Hasil pengolahan data mentah dengan menggunakan metode *Decision Tree* yaitu menghasilkan nilai akurasi sebesar 79.87%. Sedangkan untuk akurasi menggunakan metode *K-Nearest Neighbor* menghasilkan nilai 77.27%. Berikut Gambar hasil proses pengolahan data memiliki nilai akurasi sebagai berikut:

```
In [9]: #mengevaluasi model yang telah dibangun
        print("Akurasi:",metrics.accuracy_score(y_test, y_pred))
```

```
Akurasi: 0.7987012987012987
```

Gambar 3. Hasil pengolahan data menggunakan metode *Decision Tree*

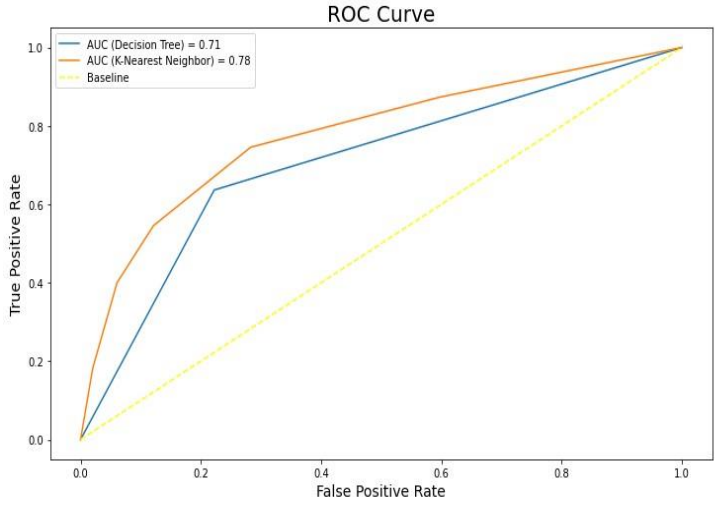
Metode *Decision Tree* menghasilkan, bisa dilihat pada gambar dibawah ini:

```
In [17]: #accuracy
         print(accuracy_score(y_test,y_pred)* 100, '%')
         plt.show()
```

```
77.27272727272727 %
```

Gambar 4. Hasil pengolahan data menggunakan metode *K-Nearest Neighbor*

Setelah mendapatkan nilai akurasi tahapan selanjutnya yaitu memetakan terhadap kurva *Receiver Operating Characteristic* (ROC) untuk mengetahui nilai Area under the ROC Curve (AUC) dari metode *Decision Tree* dan *K-Nearest Neighbor*. Dari hasil pengujian nilai AUC dengan menggunakan metode *Decision Tree* menghasilkan nilai sebesar 0.71, sedangkan menggunakan metode *K-Nearest Neighbor* menghasilkan nilai AUC sebesar 0.78. Berikut gambar kurva ROC dari hasil ekstrasi aplikasi:

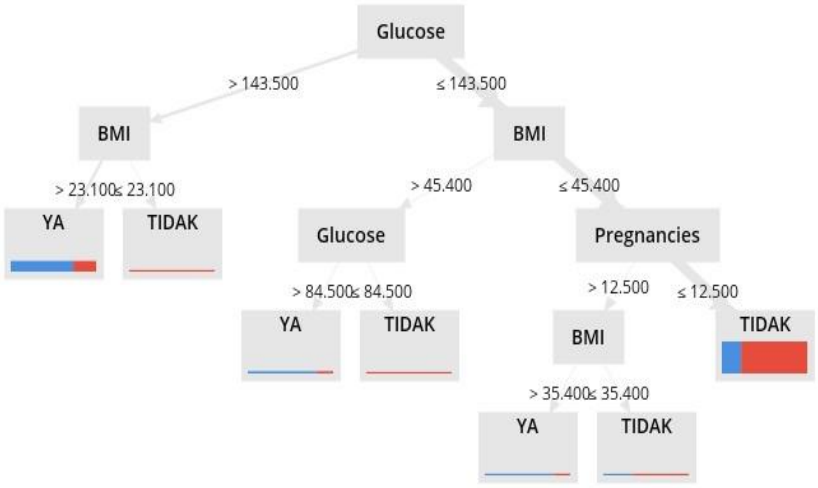


Gambar 5. Kurva ROC metode *Decision Tree* dan *K-Nearest Neighbor*

Dari hasil pengujian pengolahan data menggunakan metode *Decision Tree* dan *K-Nearest Neighbor* tanpa menggunakan fitur seleksi PSO menunjukkan metode *Decision Tree* memperoleh nilai akurasi lebih baik, hal ini menunjukkan metode *Decision Tree* lebih terstruktur dalam menentukan kecepatan mengambil keputusan. Sedangkan untuk pengolahan data metode *K-Nearest Neighbor* menunjukkan nilai ketepatan keputusan lebih baik dikarenakan banyak fitur yang menjadi indikator penentuan hasil *outcome*.

4.2. Decision Tree dan K-Nearest Neighbor (K-NN) dengan Particle Swarm Optimization

Tahapan penelitian kedua yaitu pengujian data menggunakan metode *Decision Tree* dan metode *Decision Tree* dan *K-Nearest Neighbor* menggunakan fitur seleksi PSO. Dari hasil pengujian yang dilakukan pada *confusion matrix* mendapatkan hasil dari 769 data dengan 135 data diklasifikasikan "YA" menujukan *outcome* memiliki penyakit *diabetes mellitus.* Hasil ini menujukan nilai akurasi 75.66% sesuai prediksi pada pengujian sebelumnya dengan tidak menggunakan PSO kecepatan mencari hasil keputusan sangat cepat, Seperti digambarkan dalam pohon keputusan pada gambar 6:



Gambar 6. Pohon Keputusan *Decision Tree* dengan PSO.

Dari *margin error* pengolahan sebanyak 52 data yang diprediksi nilai *outcame* "YA" tetapi ternyata pada kenyataannya tidak memiliki penyakit *diabetes mellitus*. Selanjutnya pengujian berikutnya dari 135 data diprediksi nilai outcome "TIDAK' ternyata hasilnya memiliki penyakit *diabetes mellitus* dan pada pengujian terakhir sebanyak 445 data nilai outcome dinyatakan "TIDAK" mendapatkan hasilnya memiliki penyakit *diabetes mellitus* sesuai prediksinya dari pengujian metode *Decision Tree* tanpa fitur seleksi PSO, penerapan fitur PSO tidak menjamin nilai akurasinya menjadi lebih baik. Berikut tabel pengujian akurasi *Decision Tree* dengan PSO:

Tabel 2. Hasil Akurasi *Decision Tree* dengan PSO

	True "YA"	True "TIDAK"	*Class Precision*
Pres "YA"	136	52	72,34%
Pres "TIDAK"	135	445	76,72%
Class recall	50,18%	89,54%	

Hasil dari akurasi metode *K-Nearest Neighbors* menggunakan PSO, data *confusion matrix* mendapatkan hasil dari 769 data dengan pengujian sebanyak 174 data menghasilkan klasifikasi "YA" artinya menunjukkan nilai memiliki penyakit *diabetes mellitus,* hal ini sesuai dengan prediksi yang dilakukan dengan metode *K-Nearest Neighbor* tanpa menggunakan fitur seleksi PSO, Selanjutnya pengujian dari 67 data diprediksi "YA" memiliki penyakit *diabetes mellitus* tetapi ternyata "TIDAK" memiliki penyakit *diabetes mellitus*. Pengujian berikutnya sebanyak 97 data diprediksi "TIDAK" memiliki penyakit *diabetes mellitus* ternyata hasilnya "YA" memiliki penyakit *diabetes mellitus* dan terakhir pengujian sebanyak 430 data dinyatakan "TIDAK" memiliki penyakit *diabetes mellitus* hasil *outcome* sesuai prediksi dengan nilai akurasi yang dihasilkan metode K-NN menggunakan PSO sebesar 78.65%. Berikut tabel hasil nilai akurasi metode K-NN menggunakan PSO:

Tabel 3. Hasil Akurasi KNN dengan PSO

	True "YA"	True "TIDAK"	*Class Precision*
Pres "YA"	174	67	72,20%
Pres "TIDAK"	97	430	81,59%
Class recall	64,21%	86,52%	

5. KESIMPULAN

Hasil penelitian yang dilakukan dengan menggunakan metode *Decision Tree* mendapatkan hasil akurasi 79.8% dan AUC 0.71, sedangkan penelitian yang dilakukan dengan menggunakan metode *K-Nearest Neighbors* (K-NN) mendapatkan hasil akurasi 77.27% dan AUC sebesar 0.78 hasil pengujian yang pertama tanpa menggunakan fitur seleksi *Particle Swarm Optimization* (PSO). Pengujian yang kedua dengan penerapan fitur seleksi *Particle Swarm Optimization* (PSO) untuk penelitian menggunakan metode *Decision Tree* menggunakan (PSO) mendapatkan hasil akurasi sebesar 75.66% dengan nilai AUC sebesar 0.93, dan yang terakhir penelitian menggunakan metode *K-Nearest Neighbors* (K-NN) menggunakan (PSO) mendapatkan hasil akurasi 78.65% dengan nilai AUC sebesar 0.82.

Berdasarkan penelitian yang sudah dilakukan metode *Decision Tree* menggunakan *Particle Swarm Optimization* (PSO) mengalami penurunan sebesar 4.14%, sedangkan pada metode *K-Nearest Neighbors* (K-NN) menggunakan *Particle Swarm Optimization* (PSO) mengalami kenaikan akurasi sebesar 1.38%.

UCAPAN TERIMA KASIH

Penulis ucapkan terima kasih kepada Universitas Siliwangi dan penelitian hibah mandiri dukungan keuangan dalam proses penelitian ini.

REFERENSI

[1] R. Jundinatra, K. R. and H. Hermansyah, "Analisa Bakteriuria Asimtomatik Pada Penderita Diabetes Mellitus Tipe 2 Di Rumah Sakit Bhayangkara Palembang," *Mitra Kesehatan,* vol. 3, pp. 1-5, 2020.

[2] M. A. Maharini and E. G. Zulfa Nugroho, "Pengaruh Senam Diabetes Mellitus Terhadap Penurunan Kadar Gula Darah Pada Penderita Diabetes Mellitus Tipe 2 Di RSI NU Demak," Profesi Keperawatan, vol. 8, pp. 105-117, 2021.

[3] R. R. Mahmudah and E. Aribowo, "Penggunaan Algoritma FP-Growth Untuk Menemukan Aturan Asosiasi Pada Data Transaksi Penjualan Obat Di Apotek (Studi Kasus: Apotek UAD)," *Sarjana Teknik Informatika,* vol. 2, pp. 130-139, 2014.
[4] N. Azwanti and E. Elisa, "Analisis Pola Penyakit Hipertensi Menggunakan Algoritma C4.5," Jurnal Nasional Informatika dan Teknologi Jaringan, vol. 3, pp. 116-123, 2019.
[5] M. Y. Kurniawan and M. E. Rosadi, "Optimasi Decision Tree Menggunakan Particle Swarm Optimization Pada Data Siswa Putus Sekolah," JTIULM, vol. 2, pp. 15-22, 2017.
[6] L. Andiani, S. and D. P. Rini, "Analisis Penyakit Jantung Menggunakan Metode KNN Dan Random Forest," Computer Science and ICT, vol. 5, pp. 165-169, 2019.
[7] Y. E. Achyani, "Penerapan Metode Particle Swarm Optimization Pada Optimasi Prediksi Pemasaran Langsung," Informatika, vol. 5, pp. 1-11, 2018.