

**KLASIFIKASI DATA PENYAKIT DIABETES MENGGUNAKAN
ALGORITMA *DECISION TREE* DAN *PARTICLE SWARM*
*OPTIMIZATION***

Diajukan Sebagai Syarat Untuk Menyelesaikan
Pendidikan Program Strata-1 Pada
Jurusan Teknik Informatika



Oleh:

Putri Patricia
NIM : 09021182025033

**Jurusan Teknik Informatika
FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA
2024**

LEMBAR PENGESAHAN SKRIPSI

KLASIFIKASI DATA PENYAKIT DIABETES MENGGUNAKAN ALGORITMA *DECISION TREE* DAN *PARTICLE SWARM* *OPTIMIZATION*

Oleh:

Putri Patricia

NIM : 09021182025033

Palembang, 10 Agustus 2024

Mengetahui,
Ketua Jurusan Teknik Informatika,

Pembimbing,



Dr. M. Fachrurrozi, S.Si., M.T.
NIP. 198005222008121002



Dian Palupi Rini, M.Kom., Ph.D.
NIP. 197802232006042002

TANDA LULUS UJIAN KOMPREHENSIF SKRIPSI

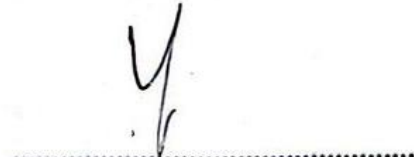
Pada hari Jum'at tanggal 26 Juli 2024 telah dilaksanakan ujian komprehensif skripsi oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : Putri Patricia
NIM : 09021182025033
Judul : *KLASIFIKASI DATA PENYAKIT DIABETES MENGGUNAKAN ALGORITMA DECISION TREE DAN PARTICLE SWARM OPTIMIZATION*

dan dinyatakan **LULUS**.

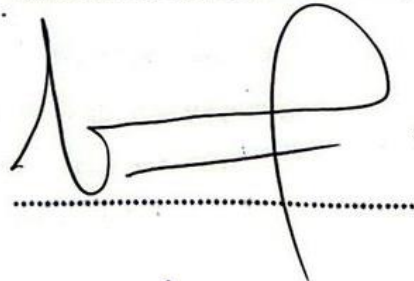
1. Ketua Penguji

Yunita, M.Cs.
NIP. 198306062015042002



2. Penguji

Alfarissi, M.Comp.Sc.
NIP. 198512152014041001



3. Pembimbing

Dian Palupi Rini, M.Kom., Ph.D.
NIP. 197802232006042002



Mengetahui,

Ketua Jurusan Teknik Informatika



Dr. M. Fachrurrozi, S.Si., M.T.
NIP. 198005222008121002

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini:

Nama : Putri Patricia

NIM : 09021182025033

Prodi : Teknik Informatika

Judul : KLASIFIKASI DATA PENYAKIT DIABETES MENGGUNAKAN
ALGORITMA *DECISION TREE* DAN *PARTICLE SWARM*
OPTIMIZATION

Hasil Pengecekan *Software iThenticate/Turnitin* : 14%

Menyatakan bahwa laporan skripsi saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan/plagiat dalam laporan skripsi ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya sesuai dengan ketentuan yang berlaku.

Demikian, pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan oleh siapapun.

Palembang, 18 Juli 2024



Putri Patricia
NIM. 09021182025033

People come and go, but this thesis will be eternal. —ppatriciasy

Bismillah,

“Keep smiling, Keep learning, Keep fighting!”
—you have to keep going.

Kupersembahkan karya tulis ini kepada
teman seperjuangan:

Alm. Muhammad Rizky Agung Elfatih

CLASSIFICATION OF DIABETES DISEASE DATA USING ALGORITHM DECISION TREE AND PARTICLE SWARM OPTIMIZATION

Oleh:

Putri Patricia
09021182025033

ABSTRACT

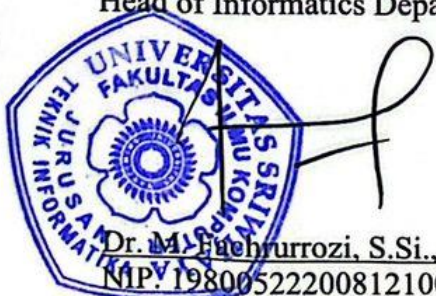
The prevalence of diabetes is very high in Indonesia so diabetes classification is crucial for early diagnosis in detecting this disease. The classification of diabetes data carried out in this study used an algorithm Decision Tree C4.5. However, the algorithm decision tree has weaknesses when handling large datasets, where not all features are relevant for the classification process, which can reduce the level of accuracy. One of the algorithms that can be integrated with the C4.5 decision tree algorithm to select relevant features is Particle Swarm Optimization (PSO). This research also uses a data balancing method to obtain more accurate results. The data balancing method used is Synthetic Minority Over-sampling Technique (SMOTE) day Random Undersampling (RUS). The research results show that the combination of the C4.5 algorithm with PSO in feature selection and the data balancing method significantly increases the accuracy of diabetes classification, producing the highest accuracy of 83.33% and an increase in accuracy of 16.33% compared to using C4.5 without PSO. The optimal PSO parameters are number of particles=10, C_1 value=2, C_2 value=2, and maximum iteration=10.

Keywords: Diabetes Classification, C4.5, Feature Selection, PSO, SMOTE, RUS

Palembang, 13 August 2024

Approved,
Head of Informatics Department,

Supervisor,



Dr. M. Fachrurrozi, S.Si., M.T.
NIP. 198005222008121002



Dian Falupi Rini, M.Kom., Ph.D.
NIP. 197802232006042002

KLASIFIKASI DATA PENYAKIT DIABETES MENGGUNAKAN ALGORITMA *DECISION TREE* DAN *PARTICLE SWARM OPTIMIZATION*

Oleh:

Putri Patricia
09021182025033

ABSTRAK

Prevalensi diabetes sangat tinggi di Indonesia sehingga klasifikasi diabetes menjadi krusial untuk diagnosis awal dalam mendeteksi penyakit ini. Klasifikasi data penyakit diabetes yang dilakukan pada penelitian ini menggunakan algoritma *Decision Tree C4.5*. Namun, algoritma *Decision Tree* memiliki kelemahan ketika menangani dataset yang besar, dimana tidak semua fitur relevan untuk proses klasifikasi, dapat menurunkan tingkat akurasi. Salah satu algoritma yang dapat diintegrasikan dengan algoritma *Decision Tree C4.5* untuk menyeleksi fitur-fitur yang relevan adalah *Particle Swarm Optimization (PSO)*. Penelitian ini juga menggunakan metode penyeimbangan data untuk mendapatkan hasil yang lebih akurat. Metode penyeimbangan data yang digunakan adalah *Synthetic Minority Over-sampling Technique (SMOTE)* dan *Random Undersampling (RUS)*. Hasil penelitian menunjukkan bahwa kombinasi algoritma *C4.5* dengan *PSO* dalam seleksi fitur serta metode penyeimbangan data secara signifikan meningkatkan akurasi klasifikasi diabetes, dengan menghasilkan akurasi tertinggi sebesar 83.33% dan peningkatan akurasi sebesar 16.33% dibandingkan penggunaan *C4.5* tanpa *PSO*. Parameter *PSO* yang optimal adalah jumlah partikel=10, nilai $C_1=2$, nilai $C_2=2$, dan iterasi maksimum=10.

Kata Kunci: Klasifikasi Diabetes, *C4.5*, Seleksi Fitur, *PSO*, *SMOTE*, *RUS*

Palembang, 1 Agustus 2024

Dosen Pembimbing,

Mengetahui,
Ketua Jurusan Teknik Informatika,



Dr. M. Fachrurrozi, S.Si., M.T.
NIP. 198005222008121002


Dian Ratupi Rini, M.Kom., Ph.D.
NIP. 197802232006042002

KATA PENGANTAR

Alhamdulillah, segala puji syukur kepada Allah SWT atas nikmat-Nya, sehingga penulis dapat menyelesaikan skripsi berjudul “KLASIFIKASI DATA PENYAKIT DIABETES MENGGUNAKAN ALGORITMA *DECISION TREE* DAN *PARTICLE SWARM OPTIMIZATION*” yang disusun untuk memenuhi salah satu syarat kelulusan Sarjana (Strata-1) di Jurusan Teknik Informatika, Fakultas Ilmu Komputer, Universitas Sriwijaya.

Pada kesempatan ini, penulis ingin mengucapkan terima kasih kepada pihak-pihak yang telah memberikan dukungan, bimbingan, dan motivasi dalam penyelesaian skripsi ini, yaitu:

1. Kepada orang tua penulis, Melianti (Umak) dan Herlansyah (Bapak), serta adik-adik penulis, Vino dan Fattan, terima kasih atas semua doa, cinta, dan dukungan yang telah kalian berikan sepanjang hidup penulis, terutama dalam proses penyelesaian penelitian ini.
2. Bapak Prof. Dr. Erwin, S.Si., M.Si. selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya.
3. Terima kasih kepada Ibu Alvi Syahrini Utami, M.Kom., dan Dr. M. Fachrurrozi, S.Si., M.T., yang masing-masing telah menjabat sebagai Ketua Jurusan Teknik Informatika selama penelitian ini, atas dukungan dan kebijakan yang telah memfasilitasi kelancaran proses studi penulis.
4. Penulis mengucapkan terima kasih yang sebesar-besarnya kepada Ibu Dian Palupi Rini, M.Kom., Ph.D., selaku Dosen Pembimbing Skripsi, atas bimbingan, ilmu, dan arahan yang telah diberikan dengan penuh kesabaran dan dedikasi, sehingga penulis dapat menyelesaikan skripsi ini dengan baik.
5. Penulis juga menyampaikan terima kasih yang sebesar-besarnya kepada Ibu Yunita, M.Cs. selaku Ketua Penguji, dan Bapak Alfarissi, M.Comp.Sc. selaku Dosen Penguji Skripsi, atas masukan, saran, dan evaluasi yang sangat berharga dalam penyusunan dan penyelesaian skripsi ini.

6. Ucapan terima kasih yang tulus penulis sampaikan kepada Ibu Mastura Diana Marieska, M.T., selaku Dosen Pembimbing Akademik, atas bimbingan dan nasihat yang telah membantu penulis selama masa studi.
7. Seluruh Dosen dan Staff Fakultas Ilmu Komputer Universitas Sriwijaya.
8. Teman-teman seperjuangan di Teknik Informatika Reguler A Angkatan 2020.
9. Deens, Aryaaa, Aura, Mipta, Fairy S.Kom (Titos, Akbar, Pian, Erlangga dan Putra), Karin, Haris, Mba Ayu, Mba Santi, Mba Desti, Mba Lilis, Dini, Wawan, Hafiz, Syafiq, Fathir, Aul, Adinda, Rosa, Roomate 305 (Nope dan Olip), Fiaa, Buni, Atika, Summon mamen (Ragils, Fanii, Ameng, Juli, Nabila dan Vina).
10. Rekan-rekan di Medinfo Fasilkom Science Community (FASCO), Kominfo Himpunan Mahasiswa Informatika (HMIF), dan Bisnis Kemitraan LDK Nadwah Universitas Sriwijaya.
11. Serta teman-teman di Bangkit Batch 1 2023 Grup CC-51 dan Capstone C23-PS111, Volunteer TS CDC UNSRI, Aslab Multimedia dan Game Fasilkom Unsri, ECFSAW Grup 429 dan Kelompok Mata Kuliah Metode Penelitian.
12. Beserta semua pihak yang telah membantu dalam penyusunan yang tidak dapat disebutkan satu persatu.

Penulis berharap bahwa skripsi ini dapat memberikan kontribusi yang bermanfaat bagi pengembangan ilmu pengetahuan dan sebagai referensi bagi penelitian-penelitian selanjutnya. Akhir kata, penulis menyadari bahwa skripsi ini masih jauh dari kesempurnaan, oleh karena itu penulis sangat terbuka terhadap saran dan kritik yang membangun.

Lahat, 26 Juli 2024

Penulis,

Putri Patricia

DAFTAR ISI

LEMBAR PENGESAHAN SKRIPSI.....	ii
TANDA LULUS UJIAN KOMPREHENSIF SKRIPSI.....	iii
HALAMAN PERNYATAAN BEBAS PLAGIAT	iv
MOTTO DAN PERSEMBAHAN	v
ABSTRACT	vi
ABSTRAK	vii
KATA PENGANTAR	viii
DAFTAR ISI.....	x
DAFTAR GAMBAR	xii
DAFTAR TABEL.....	xiii
BAB I PENDAHULUAN	I-1
1.1 Pendahuluan	I-1
1.2 Latar Belakang	I-1
1.3 Rumusan Masalah	I-4
1.4 Tujuan Penelitian.....	I-4
1.5 Manfaat Penelitian.....	I-5
1.6 Batasan Penelitian	I-5
1.7 Sistematika Penulisan.....	I-6
1.8 Kesimpulan.....	I-7
BAB II KAJIAN LITERATUR	II-1
2.1 Pendahuluan	II-1
2.2 Landasan Teori	II-1
2.2.1 Diabetes Melitus	II-1
2.2.2 Klasifikasi	II-2
2.2.3 C4.5	II-3
2.2.4 <i>Particle Swarm Optimization</i>	II-5
2.2.5 <i>Confusion Matrix</i>	II-7
2.2.6 <i>Cross Validation</i>	II-8
2.2.7 SMOTE	II-9
2.2.8 RUS	II-9
2.3 <i>Rational Unified Process (RUP)</i>	II-10
2.4 Penelitian lain yang relevan	II-12
2.5 Kesimpulan.....	II-14
BAB III METODOLOGI PENELITIAN.....	III-1
3.1 Pendahuluan	III-1
3.2 Pengumpulan Data	III-1
3.3 Tahapan Penelitian	III-2
3.3.1 Menetapkan Kerangka Kerja.....	III-3
3.3.2 Menetapkan Kriteria Pengujian.....	III-7
3.3.3 Menentukan Format Data Pengujian.....	III-8

3.3.4	Menentukan Alat yang Digunakan dalam Penelitian.....	III-9
3.3.5	Melakukan Pengujian Penelitian.....	III-9
3.3.6	Melakukan Analisis Hasil Pengujian	III-10
3.4	Metode Pengembangan Perangkat Lunak	III-11
3.4.1	Fase Insepsi	III-11
3.4.2	Fase Elaborasi	III-11
3.4.3	Fase Konstruksi.....	III-12
3.4.4	Fase Transisi	III-12
3.5	Manajemen Proyek Penelitian.....	III-12
3.6	Kesimpulan.....	III-15
BAB IV	PENGEMBANGAN PERANGKAT LUNAK	IV-1
4.1	Pendahuluan	IV-1
4.2	Fase Insepsi	IV-1
4.2.1	Analisis Perangkat Lunak	IV-1
4.2.2	Desain Perangkat Lunak	IV-2
4.3	Fase Elaborasi	IV-8
4.3.1	Perancangan Data.....	IV-8
4.3.2	Perancangan Antarmuka	IV-8
4.3.3	Perancangan <i>Sequence Diagram</i>	IV-9
4.3.4	Perancangan <i>Class Diagram</i>	IV-11
4.4	Fase Konstruksi	IV-11
4.4.1	Implementasi Kelas.....	IV-11
4.4.2	Implementasi Antarmuka	IV-12
4.5	Fase Transisi.....	IV-14
4.5.1	Lingkungan Pengujian	IV-14
4.5.2	Rencana Pengujian	IV-15
4.5.3	Pengujian.....	IV-16
4.6	Kesimpulan.....	IV-18
BAB V	HASIL DAN ANALISIS PENELITIAN.....	V-1
5.1	Pendahuluan	V-1
5.2	Analisis Hasil Penelitian dan Pembahasan.....	V-1
5.2.1	Percobaan Konfigurasi	V-1
5.2.2	Hasil Konfigurasi I (Jumlah Partikel)	V-4
5.2.3	Hasil Konfigurasi II (Nilai c1 dan c2).....	V-5
5.2.4	Hasil Konfigurasi III (Maksimum Iterasi)	V-6
5.2.5	Hasil Konfigurasi IV (Metode Penyeimbangan Data)	V-7
5.3	Kesimpulan.....	V-8
BAB VI	KESIMPULAN DAN SARAN	VI-1
6.1	Pendahuluan	VI-1
6.2	Kesimpulan.....	VI-1
6.3	Saran.....	VI-1
DAFTAR PUSTAKA		xv

DAFTAR GAMBAR

Gambar II-1 <i>Workflow</i> RUP (Kruchten, 2000).....	II-10
Gambar III-2 Tahapan Penelitian.....	III-3
Gambar III-3 Kerangka Kerja	III-4
Gambar III-4 <i>Flowchart</i> Klasifikasi C4.5-PSO	III-6
Gambar III-5. Diagram Tahapan Pengujian Penelitian.....	III-10
Gambar IV-1. <i>Use Case Diagram</i>	IV-3
Gambar IV-2. <i>Activity Diagram</i> Memasukkan File Data	IV-7
Gambar IV-3. <i>Activity Diagram</i> Melakukan Klasifikasi	IV-7
Gambar IV-4. Rancangan Antarmuka Perangkat Lunak	IV-9
Gambar IV-5. <i>Sequence Diagram</i> Masukkan File Data	IV-10
Gambar IV-6. <i>Sequence Diagram</i> Melakukan Klasifikasi	IV-10
Gambar IV-7. <i>Class Diagram</i>	IV-11
Gambar IV-8. Antarmuka Perangkat Lunak	IV-14

DAFTAR TABEL

Tabel I-1. Ringkasan Penelitian Terkait.....	I-3
Tabel II-1. Model <i>Confusion Matrix</i>	II-7
Tabel III-1. Deskripsi Atribut Dataset	III-2
Tabel III-2. <i>Confusion Matrix</i>	III-8
Tabel III-3. Rancangan Tabel <i>Confusion Matrix</i>	III-8
Tabel III-4. Rancangan Tabel Perbandingan Akurasi.....	III-8
Tabel III-5. <i>Work Breakdown Structure</i> (WBS)	III-13
Tabel IV-1. Kebutuhan Fungsional.....	IV-2
Tabel IV-2. Kebutuhan Non-Fungsional.....	IV-2
Tabel IV-3. Definisi Aktor <i>Use Case</i>	IV-3
Tabel IV-4. Definisi <i>Use Case</i>	IV-4
Tabel IV-5. Skenario <i>Use Case</i> Memasukkan File Data	IV-4
Tabel IV-6. Skenario <i>Use Case</i> Melakukan Klasifikasi	IV-6
Tabel IV-7. Implementasi Kelas	IV-11
Tabel IV-8. Rencana Pengujian <i>Use Case</i> Memasukkan File Data.....	IV-15
Tabel IV-9. Rencana Pengujian <i>Use Case</i> Melakukan Klasifikasi.....	IV-16
Tabel IV-10. Pengujian Masukkan File Data.....	IV-16
Tabel IV-11. Pengujian Melakukan Klasifikasi.....	IV-17
Tabel V-1. Konfigurasi I (Jumlah Partikel)	V-2
Tabel V-2. Konfigurasi II (Nilai c_1 dan c_2).....	V-3
Tabel V-3. Konfigurasi III (Iterasi Maksimum)	V-3
Tabel V-4. Konfigurasi IV (Metode Penyeimbangan Data)	V-4
Tabel V-5. Hasil Konfigurasi I (Jumlah Partikel).....	V-4
Tabel V-6. Hasil Konfigurasi II (Nilai C_1 dan C_2).....	V-5
Tabel V-7. Hasil Konfigurasi III (Maksimum Iterasi).....	V-6
Tabel V-8. Hasil Konfigurasi IV (Metode Penyeimbangan Data).....	V-7

BAB I

PENDAHULUAN

1.1 Pendahuluan

Bab ini menyajikan latar belakang, perumusan masalah, tujuan penelitian, manfaat penelitian, batasan penelitian, serta sistematika penulisan skripsi.

1.2 Latar Belakang

International Diabetes Federation (IDF) 2021 menyatakan bahwa sekitar 537 juta orang dewasa dalam rentang usia 20-79 tahun hidup dengan diabetes. Proyeksi yang diberikan oleh IDF juga menunjukkan bahwa jumlah penderita diabetes diperkirakan akan meningkat menjadi 643 juta pada tahun 2030, dan mencapai angka yang lebih tinggi, yaitu 783 juta pada tahun 2045¹⁾. IDF juga mengungkapkan bahwa prevalensi diabetes di Indonesia telah meningkat sejak tahun 2000, mencapai angka 19 ribu pada tahun 2021, dengan sekitar 73.7% penderita yang belum menerima diagnosis resmi²⁾.

Karena meningkatnya angka kejadian diabetes, diagnosis dini menjadi krusial untuk mencegah potensi komplikasi dan mengurangi risiko masalah kesehatan. Data mining dapat digunakan sebagai landasan untuk memprediksi dan mendiagnosis karakteristik penyakit dengan menggunakan metode yang sesuai. (Rahayu et al., 2023). Pemanfaatan data mining sangat umum digunakan dalam

¹⁾ <https://idf.org/about-diabetes/diabetes-facts-figures/>

²⁾ <https://www.diabetesatlas.org/data/en/country/94/id.html>

klasifikasi data medis. Klasifikasi memiliki beragam aplikasi untuk mengelompokkan data dari berbagai aspek kehidupan, termasuk pengklasifikasian penyakit (Saritas and Yasar, 2019). Salah satu metode klasifikasi yang umum digunakan adalah algoritma C4.5, dan telah terbukti memiliki performa yang baik dalam klasifikasi (Ningsih, 2020).

Sebelumnya telah dilakukan penelitian mengklasifikasikan penyakit diabetes melitus algoritma *decision tree* C4.5. Dataset yang digunakan pada penelitian adalah *Pima Indians Diabetes Dataset*. Hasil Penelitian menunjukkan bahwa *decision tree* C4.5 menghasilkan akurasi yang baik mencapai 80.72% dengan *k-fold cross validation(k=10)* (Junior, Saedudin, dan Widharta 2021). C4.5 juga telah terbukti efektif dalam mengklasifikasikan data teks, yang dalam penelitian ini menggunakan data cuaca (Novandya dan Oktria, 2017). Meskipun demikian, algoritma *decision tree* ini memiliki kelemahan yaitu munculnya masalah ketika berhadapan dengan dataset berukuran besar, di mana tidak semua fitur pada dataset memiliki relevansi dalam proses klasifikasi. Adanya fitur yang tidak relevan tersebut dapat berdampak negatif pada akurasi (Wiantara, 2021). Kelemahan tersebut dapat diatasi dengan melakukan seleksi fitur dan telah terbukti mampu meningkatkan akurasi (Wiantara, 2021). Dalam meningkatkan kinerja algoritma C4.5, berbagai teknik optimasi dapat digunakan, salah satunya *Particle Swarm Optimization* (PSO) yang dalam penelitian ini telah berhasil meningkatkan akurasi C4.5 dalam mengklasifikasikan harapan hidup pasien hepatitis setelah menerapkan metode PSO (Arizaldi, 2020). Penelitian terkait lainnya meneliti tentang *Decision Tree* dan PSO untuk klasifikasi sel pap smear (Arifin dan

Rachman, 2020), klasifikasi penyakit demensia dengan menggunakan metode C4.5 dan PSO menunjukkan peningkatan akurasi setelah optimasi menggunakan PSO (Wiantara, 2021). Berikut ringkasan penelitian terkait:

Tabel I- 1. Ringkasan Penelitian Terkait

No.	Peneliti	Judul	Hasil
1.	Joshua Bonardo Junior, Rd. Rohmat Saedudin, Vandha Pradiwiyasma Widharta	Perbandingan Akurasi Algoritma <i>Decision Tree</i> dan Algoritma <i>Support Vector Machine</i> pada Penyakit Diabetes. Dataset: <i>Pima Indian Diabetes Dataset</i>	Hasil penelitian menunjukkan bahwa <i>decision tree</i> C4.5 menghasilkan akurasi yang baik mencapai 80.72% dengan <i>k-fold cross validation</i> ($k=10$).
2.	Adhika Novandya, Isni Oktria	Penerapan Algoritma Klasifikasi Data Mining C4.5 Pada Dataset Cuaca Wilayah Bekasi.	Menghasilkan akurasi sebesar 88.89% dengan <i>10-fold cross validation</i> .
3.	Tubagus Wiantara	Optimasi Metode C4.5 Menggunakan <i>Particle Swarm Optimization</i> untuk Klasifikasi Penyakit Demensia.	<ul style="list-style-type: none"> - PSO meningkatkan akurasi C4.5 untuk klasifikasi penyakit demensia - Nilai <i>hyperparameter</i> hasil konfigurasi yaitu jumlah partikel = 10 dan jumlah iterasi = 75. - Nilai akurasi klasifikasi C4.5-PSO mencapai 91.93% dengan peningkatan akurasi metode C4.5 adalah 2.68%.
4.	Arizaldi	Pengaruh <i>Particle Swarm Optimization</i> Pada Algoritma C4.5 Untuk Klasifikasi Harapan Hidup Pasien Hepatitis	Algoritma C4.5 menghasilkan nilai rata-rata akurasi sebesar 56 % dan mengalami peningkatan nilai rata-rata akurasi sebesar 70,2 % setelah seleksi atribut menggunakan metode PSO.
5.	Toni Arifin, Rizal Rachman	Optimasi <i>Decision Tree</i> Menggunakan <i>Particle Swarm Optimization</i> Untuk Klasifikasi Sel Pap Smear.	Akurasi <i>decision tree</i> sebesar 91.39% dengan AUC 0.858, meningkat menjadi 96.76 % dengan AUC 0.926 setelah dioptimasi dengan algoritma PSO.

6.	Putri Patricia	Klasifikasi Data Penyakit Diabetes Menggunakan Algoritma <i>Decision Tree</i> dan <i>Particle Swarm Optimization</i> .	
----	----------------	--	--

Berdasarkan latar belakang diatas, maka pada penelitian ini akan dilakukan klasifikasi data penyakit diabetes menggunakan algoritma *Decision Tree* dan *Particle Swarm Optimization* (PSO).

1.3 Rumusan Masalah

Dengan latar belakang tersebut, rumusan masalah penelitian ini adalah mengevaluasi efektivitas penggunaan *Particle Swarm Optimization* (PSO) dalam meningkatkan kinerja algoritma C4.5 untuk mengklasifikasikan data penyakit diabetes. Berdasarkan rumusan masalah ini, berikut pertanyaan penelitian yang diajukan:

- 1) Bagaimana seleksi fitur pada C4.5 menggunakan algoritma PSO dalam mengklasifikasikan data penyakit diabetes?
- 2) Bagaimana hasil klasifikasi data penyakit diabetes menggunakan C4.5 yang dilakukan seleksi fitur dengan PSO?

1.4 Tujuan Penelitian

Tujuan Penelitian ini adalah:

- 1) Mengintegrasikan algoritma C4.5 dan PSO agar dapat digunakan dalam klasifikasi data penyakit diabetes.

- 2) Membandingkan akurasi hasil klasifikasi antara penggunaan algoritma C4.5 yang diseleksi fitur dengan algoritma PSO dan algoritma C4.5 tanpa seleksi fitur PSO pada klasifikasi data penyakit diabetes.

1.5 Manfaat Penelitian

Manfaat yang dapat diperoleh dari penelitian ini adalah sebagai berikut:

- 1) Hasil penelitian ini dapat menjadi pedoman untuk penerapan algoritma C4.5 dan algoritma PSO dalam klasifikasi data penyakit diabetes.
- 2) Penelitian ini diharapkan dapat memberikan kontribusi dalam meningkatkan efisiensi dan akurasi pengelompokan dataset penyakit diabetes.
- 3) Hasil penelitian dapat dikembangkan dan digunakan sebagai referensi untuk penelitian masa depan.

1.6 Batasan Penelitian

Berikut merupakan batasan pada penelitian ini:

- 1) Dataset yang dipakai dalam penelitian ini merupakan dataset sumber terbuka yang diunduh dari situs *Kaggle*³⁾.
- 2) Dilakukan *pre-processing data* sebelum masuk ke sistem.
- 3) Evaluasi performa klasifikasi dilakukan dengan melakukan pengujian menggunakan *Confusion Matrix*.

³⁾ <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/>

1.7 Sistematika Penulisan

BAB I PENDAHULUAN

Bab ini menyajikan latar belakang, perumusan masalah, tujuan penelitian, manfaat penelitian, batasan penelitian, serta sistematika penulisan skripsi.

BAB II KAJIAN LITERATUR

Bab ini menguraikan teori-teori dasar yang digunakan dalam penelitian, termasuk penelitian-penelitian terkait yang relevan dengan topik yang akan diteliti.

BAB III METODOLOGI PENELITIAN

Bab ini menguraikan tentang metode pengumpulan data, tahapan penelitian, metode pengembangan perangkat lunak, dan manajemen proyek penelitian. Tahapan penelitian dijadikan acuan dalam setiap fase pengembangan perangkat lunak yang memberikan sebuah solusi untuk rumusan masalah dan mencapai tujuan dari penelitian.

BAB IV PENGEMBANGAN PERANGKAT LUNAK

Bab ini membahas rancangan dan implementasi perangkat lunak menggunakan metode *Object-Oriented Programming* (OOP), dengan pedoman dari metode pengembangan perangkat lunak UML, yaitu *Rational Unified Process* (RUP) yang meliputi empat tahap insepri, elaborasi, konstruksi, dan transisi.

BAB V HASIL DAN ANALISIS PENELITIAN

Bab ini membahas hasil klasifikasi data penyakit diabetes menggunakan algoritma C4.5 dan optimasi dengan PSO dari pengujian perangkat lunak yang dikembangkan dalam penelitian. Di akhir bab, disajikan analisis dari hasil yang diperoleh selama proses pengujian.

BAB VI KESIMPULAN DAN SARAN

Bab ini menyimpulkan hasil analisis selama pengujian klasifikasi data penyakit diabetes menggunakan algoritma *decision tree* dan *particle swarm optimization*, serta memberikan saran berdasarkan hasil penelitian.

1.8 Kesimpulan

Berdasarkan uraian diatas, maka akan dikembangkan sebuah perangkat lunak untuk mengklasifikasi data penyakit diabetes dengan menerapkan teknik optimasi *particle swarm optimization* pada algoritma *decision tree*. Untuk lebih lanjut, teori-teori yang berhubungan erat dengan penelitian akan dijelaskan pada bab II.

BAB II

KAJIAN LITERATUR

2.1 Pendahuluan

Bab ini menguraikan teori-teori dasar yang digunakan dalam penelitian, termasuk penelitian-penelitian terkait yang relevan dengan topik yang akan diteliti.

2.2 Landasan Teori

2.2.1 Diabetes Melitus

Diabetes Mellitus (DM), yang sering disebut sebagai kencing manis merupakan penyakit kronis yang dapat berlangsung seumur hidup (Lestari, Zulkarnain dan Siji 2021). Diabetes ditandai dengan peningkatan kadar glukosa yang seiring waktu dapat menyebabkan kerusakan serius pada jantung, pembuluh darah, mata, ginjal, dan saraf⁴⁾. Terdapat beberapa jenis diabetes, di antaranya diabetes tipe 1 yang dapat berkembang pada usia berapa pun dan memerlukan pengobatan insulin untuk bertahan hidup. Diabetes tipe 2, yang mencakup sekitar 90% dari seluruh kasus diabetes, lebih sering didiagnosis pada orang dewasa dan melibatkan penggunaan insulin yang tidak efektif atau produksi insulin tidak mencukupi. Diabetes Gestasional terjadi selama kehamilan, ditandai dengan peningkatan kadar glukosa darah dan potensi komplikasi bagi ibu dan anak.

⁴⁾ <https://www.who.int/health-topics/diabetes>

Memahami berbagai jenis diabetes ini sangat penting untuk perawatan yang sesuai dan manajemen yang efektif guna menjaga kesejahteraan umum individu dengan diabetes. Deteksi dini juga memiliki peranan krusial dalam menjalani hidup dengan baik bagi penderita diabetes. Semakin lama seseorang hidup dengan diabetes yang tidak terdiagnosis dan tidak diobati, kemungkinan besar dampak kesehatannya akan semakin buruk. Diabetes melitus dapat diklasifikasikan dengan mengidentifikasi parameter tertentu, termasuk *Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, dan Age*⁵⁾

2.2.2 Klasifikasi

Han, Pei, and Kamber (2022) menyatakan bahwa proses klasifikasi data melibatkan dua tahap utama. Pertama, dalam tahap pembelajaran (*learning*), model klasifikasi dikembangkan melalui analisis data pelatihan untuk mengidentifikasi atribut label kelas. Model tersebut dibangun dengan merinci aturan klasifikasi yang mencerminkan korelasi antara atribut-atribut tertentu dan keputusan kelas. Langkah kedua, klasifikasi (*classification*), melibatkan penerapan model yang telah dibangun pada *data testing* untuk memprediksi label kelas. *Data testing* digunakan sebagai sarana pengujian untuk mengevaluasi sejauh mana model dapat mengaplikasikan informasi dari data pelatihan ke data baru. Jika hasil prediksi model memenuhi standar

⁵⁾ <https://idf.org/about-diabetes/what-is-diabetes/>

keakuratan yang diinginkan, model dapat dianggap efektif dan dapat diterapkan pada klasifikasi tupel data baru untuk memprediksi label kelasnya. Dengan demikian, esensi dari proses klasifikasi terletak pada pembangunan model melalui pembelajaran dan penerapan model tersebut untuk mengklasifikasikan label kelas pada data yang belum pernah diidentifikasi sebelumnya.

2.2.3 C4.5

J. Ross Quinlan memperkenalkan Algoritma C4.5 sebagai penerus ID3 pada tahun 1993 (Han, Pei, and Kamber, 2022). Algoritma C4.5 banyak digunakan dalam penelitian klasifikasi data mining karena mudah untuk diinterpretasikan (Muttaqien, Galih, and Pramuntadi, 2021). Algoritma C4.5 dapat mengelola data nominal, kategorikal, dan numerik (Chanmee and Kesorn, 2023), serta bekerja dengan baik pada dataset berukuran besar (Roy and Garg, 2017).

Berikut langkah-langkah membangun pohon keputusan dengan algoritma C4.5 (Bulolo, 2020):

- a. Menentukan nilai $Entropy(S)$ untuk setiap nilai kriteria

Pada langkah ini, dilakukan pencarian nilai $entropy(S)$, di mana $entropy(S)$ berfungsi sebagai ukuran untuk menilai variasi setiap nilai atribut kriteria terhadap atribut keputusan dalam suatu *dataset*. Semakin rendah nilai $entropy(S)$, maka tingkat variasi dalam *dataset* semakin rendah,

sebaliknya, semakin tinggi nilai $entropy(S)$, maka tingkat variasinya menjadi lebih tinggi. Rumus matematika $entropy(S)$, adalah seperti berikut:

(2.1)

$$Entropy(S) = \sum_{i=1}^n - p_i * \log_2(p_i)$$

Dimana:

S = Jumlah kasus (*Sampling*)

N = Jumlah partisi S

Pi = Proporsi dari S_i terhadap S

Dalam proses pencarian nilai jika semua nilai $entropy(S)$ kriteria terhadap atribut keputusan sama, maka nilai $entropy = 1$, dan jika hanya satu nilai kriteria yang tidak sama dengan 0, maka nilai $Entropy = 0$.

b. Menentukan nilai $Gain(S,A)$ untuk setiap atribut

$Gain(S,A)$ adalah hasil dari pengurangan total nilai $entropy$ dari nilai setiap atribut kriteria dikalikan dengan proporsi nilai atribut terhadap jumlah kasus. $Gain(S,A)$ berperan sebagai ukuran efektivitas setiap atribut kriteria dalam proses klasifikasi data. Untuk mencari nilai $Gain(S,A)$ digunakan rumus matematika sebagai berikut:

(2.2)

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{S} Entropy(S_i)$$

Dimana:

S = Jumlah kasus (*Sampling*)

A = atribut

n = Jumlah partisi S

$|S_i|$ = Jumlah kasus pada partisi ke- i

$|S|$ = Jumlah kasus dalam S

- c. Membentuk node dengan akar dan cabang berdasarkan *Gain* tertinggi
- d. Mengulangi proses untuk masing-masing cabang

2.2.4 Particle Swarm Optimization

Particle Swarm Optimization (PSO) merupakan algoritma optimasi yang diusulkan pertama kali oleh Kennedy dan Eberhart pada tahun 1995 (Mercangöz, 2022) sebagai cara untuk menyelesaikan masalah optimasi dengan menggunakan kelompok partikel yang bergerak secara acak. PSO adalah algoritma metaheuristik yang terinspirasi dari alam dan bekerja dengan sekumpulan partikel seperti kelompok hewan yang mencari makan bersama-sama. Setiap partikel dianggap solusi untuk masalah dalam ruang keputusan dan memiliki dua karakteristik yaitu kecepatan dan posisi (Samala, 2022).

Berikut rumus matematika yang menentukan posisi dan kecepatan partikel di suatu ruang keputusan (Wiantara, 2021):

(2.3)

$$x_i(k) = x_{i_1}(k), x_{i_2}(k), \dots, x_{i_N}(k)$$

$$v_i(k) = v_{i_1}(k), v_{i_2}(k), \dots, v_{i_N}(k)$$

Dimana:

x = posisi partikel

v = kecepatan partikel

i = indeks partikel

k = iterasi ke- k

N = ukuran dimensi ruang

Berikut rumus matematika yang digunakan dalam PSO untuk memperbarui kecepatan dan posisi setiap partikel (Samala, 2022):

(2.4)

$$v_i(k + 1) = wv_i(k) + \rho_1 c_1 (p_i^{best} - X_i(k)) + \rho_2 c_2 (GX_i(k))$$

$$X_i(k + 1) = X_i(k) + v_i(k + 1)$$

Dimana, i adalah indeks partikel, k adalah nomor iterasi, $v_i(k)$ adalah kecepatan, $X_i(k)$ adalah posisi partikel i pada iterasi k , p_i^{best} adalah posisi terbaik yang ditemukan oleh partikel i (*personal best*), G adalah posisi terbaik secara keseluruhan oleh *swarm* atau seluruh kelompok partikel (*global best, best of personal best*), w adalah koefisien inersia, $(\rho_1 \rho_2)$ adalah bilangan acak dalam interval $[0,1]$, c_1, c_2 adalah konstanta positif yang merepresentasikan faktor daya tarik partikel terhadap posisi terbaiknya sendiri atau terhadap posisi terbaik *swarm* (Samala, 2022).

Dalam algoritma PSO, proses pemodelan dilakukan dengan menempatkan solusi ke dalam suatu ruang keputusan selama beberapa iterasi. Pada setiap iterasi, posisi dan partikel akan diperbarui untuk mendekati tujuan yang ingin dicapai yang dapat berupa fungsi minimal atau maksimum. Proses ini berlanjut hingga syarat atau kondisi maksimum iterasi terpenuhi, dan juga dapat melibatkan penggunaan parameter kriteria terminasi atau metode pemberhentian lainnya (Wiantara, 2021).

2.2.5 *Confusion Matrix*

Confusion Matrix merupakan tabel *error matrix* yang mengindikasikan hasil dari algoritma klasifikasi yang digunakan untuk mengevaluasi hasil model yang diusulkan (Karimi, 2021).

Tabel II-1. Model *Confusion Matrix*

<i>Correct Classification</i>	<i>Classified (+)</i>	<i>Classified (-)</i>
<i>Actual (+)</i>	FP	TN
<i>Actual (-)</i>	TP	FN

Berdasarkan Tabel II-1 Model *Confusion Matrix*, *True Positive* (TP) mengindikasikan banyaknya data dari kelas positif yang berhasil diprediksi dengan benar sebagai positif. *True Negative* (TN) merujuk pada banyaknya data dari kelas negatif yang berhasil diprediksi dengan benar sebagai negatif. *False Positive* (FP) mengacu pada banyaknya data dari kelas negatif yang salah diprediksi sebagai positif. *False Negative* (FN) menunjukkan banyaknya data dari kelas positif yang salah diprediksi sebagai negatif.

Confusion matrix dari suatu model akan menghasilkan nilai *accuracy*, *precision*, *recall*, dan *F1-score* yang menjadi indikator performa model. Berikut adalah definisi dan rumus untuk masing-masing *matrix*:

- a. *Accuracy* mengukur perbandingan prediksi benar dengan total data.

(2.5)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- b. *Precision* menunjukkan rasio prediksi positif yang sebenarnya positif dari total prediksi positif.

(2.6)

$$Precision = \frac{TP}{TP + FP}$$

- c. *Recall* mengindikasikan rasio positif aktual yang diprediksi dengan benar dari total positif aktual.

(2.7)

$$Recall = \frac{TP}{TP + FN}$$

- d. *F1-score* adalah nilai mean harmonis dari *precision* dan *recall*

(2.8)

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad \square$$

2.2.6 Cross Validation

Cross Validation atau validasi silang adalah metode yang umumnya digunakan untuk mengevaluasi kesalahan prediksi model yang disesuaikan dengan data pelatihan (Bates, Hastie, and Tibshirani, 2022). Salah satu

metode validasi silang yaitu *k-fold cross validation*. Dalam *k-fold cross validation*, dataset awal diacak dan dibagi menjadi k subset. Proses pelatihan dan pengujian diulang sebanyak k kali, di mana setiap sampel digunakan dengan jumlah yang sama untuk pelatihan dan yang lainnya untuk pengujian. Untuk klasifikasi, akurasi dihitung dari total klasifikasi benar dibagi total tuple pada data awal, dan nilai k yang umumnya direkomendasikan adalah 10 (Han, Pei, and Kamber, 2022). Dengan contoh $k=10$, data 100 dibagi menjadi 10 bagian, setiap bagian 10 data. Bagian tersebut bergantian sebagai *data testing*, sementara sisanya sebagai *data training*. Jika tidak habis dibagi 10, data yang tersisa dikurangi sehingga jumlah data habis dibagi oleh k (Brownlee, 2018).

2.2.7 SMOTE

Synthetic Minority Oversampling TEchnique (SMOTE) adalah metode *oversampling* yang dikembangkan oleh Chawla (IEEE, 1976). Salah satu teknik *oversampling* yang mengatasi masalah ketidakseimbangan kelas untuk meningkatkan kinerja model klasifikasi dengan memperkaya batas kelas minoritas dengan menciptakan sampel sintesis dalam di kelas minoritas daripada hanya menggandakan sampel yang ada untuk menghindari masalah *overfitting* (At et al., 2016).

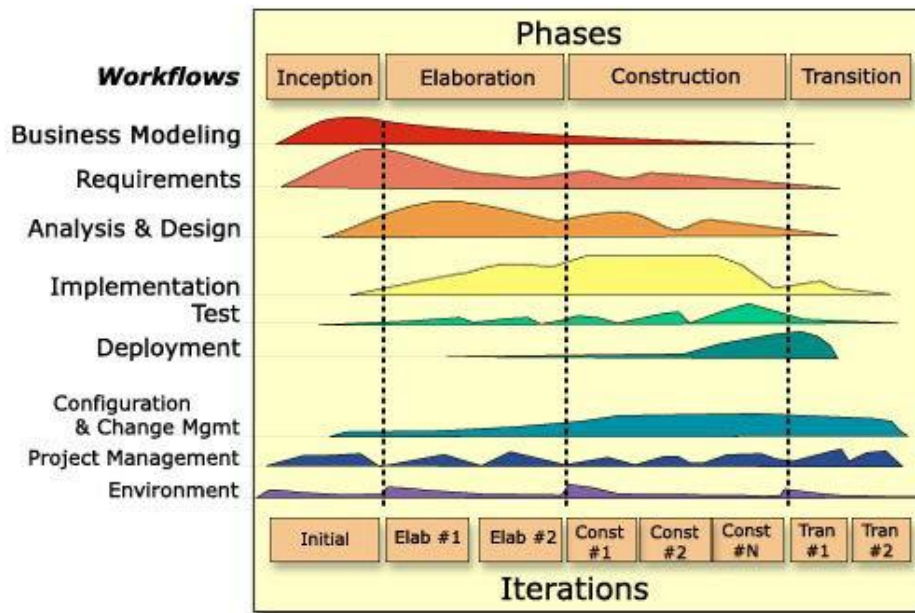
2.2.8 RUS

Random Undersampling (RUS) adalah salah satu metode *resampling* yang dapat menangani ketidakseimbangan kelas yang sering terjadi dalam dataset. Metode ini melibatkan penghapusan sampel secara

acak dari kelas mayoritas, dengan tujuan untuk mengurangi dominasi jumlah contoh dari kelas tersebut. Penggunaan RUS tidak hanya untuk menyeimbangkan proporsi antara kelas mayoritas dan minoritas, tetapi juga sebagai fondasi utama dalam pembentukan ensemble. Dalam konteks ini, *ensemble* menggabungkan berbagai model atau *classifier* untuk mengatasi tantangan klasifikasi yang disebabkan oleh distribusi tidak merata dari kelas-kelas yang berbeda dalam dataset dan meningkatkan akurasi dalam mengklasifikasikan data yang tidak seimbang (Chen et al., 2004).

2.3 Rational Unified Process (RUP)

Rational Unified Process (RUP) is a software engineering process (Kruchten, 2004).



Gambar II- 1 *Workflow* RUP (Kruchten, 2000)

Gambar diatas merupakan arsitektur RUP. Adapun tahapan yang ada pada RUP adalah sebagai berikut (Rohmat dan Nuriyah, 2023):

a. Insepsi

Tahap pertama difokuskan pada penggambaran model proses bisnis dengan menetapkan ruang lingkup proyek. Selain itu, tahap ini melibatkan analisis kebutuhan sistem dan penentuan batasan-batasan sistem yang akan dibuat.

b. Elaborasi

Tahap kedua ditekankan pada proses desain. Arsitektur sistem dirancang berdasarkan hasil tahap sebelumnya, dengan menghasilkan model kebutuhan sistem melalui berbagai representasi seperti *activity*, *sequence*, dan *class diagram*. Selanjutnya, desain tersebut diimplementasikan melalui proses *coding*.

c. Konstruksi

Tahap *construction*, melibatkan aktivitas pemrograman (*coding*) untuk mengimplementasikan rancangan yang telah disusun sebelumnya. Selanjutnya, sistem diuji menggunakan metode *blackbox* dan *whitebox testing* untuk memastikan aspek fungsional dan non-fungsionalnya. Hasil tahap ini adalah sistem perangkat lunak yang siap digunakan dan didokumentasikan sebelum di *deploy* kepada pengguna.

d. Transisi

Pada tahap ini akan dilakukan pengujian kesesuaian sistem dengan harapan pengguna. Ini mencakup pemindahan sistem dari lingkungan pengembangan ke lingkungan publik atau target pasar, di mana sistem perangkat lunak diharapkan berfungsi secara optimal. Tahap ini juga

melibatkan dokumentasi menyeluruh tentang sistem perangkat lunak yang beroperasi dengan baik dalam lingkungannya.

2.4 Penelitian lain yang relevan

Beberapa peneliti telah melakukan penelitian tentang algoritma C4.5. (Septiani dan Marlina, 2021) melakukan penelitian yang membandingkan algoritma *Decision Tree*, *Naïve Bayes*, dan *Neural Network* untuk deteksi dini penyakit diabetes. Tujuan utama penelitian ini adalah melakukan klasifikasi prediksi guna mendapatkan hasil yang akurat untuk pengambilan keputusan terkait deteksi dini penyakit diabetes. Data yang digunakan berasal dari riwayat pasien rumah sakit di Sylhet, Bangladesh. Hasil pengukuran menunjukkan tingkat akurasi masing-masing 95,96% dengan *Decision Tree*, 87,69% dengan *Naive Bayes*, dan 61,54% dengan *Neural Network*. Oleh karena itu, hasil penelitian menyimpulkan bahwa algoritma terbaik untuk prediksi deteksi dini penyakit diabetes adalah *Decision Tree*.

Penelitian selanjutnya tentang klasifikasi penderita penyakit diabetes menggunakan algoritma *Decision Tree* C4.5 dilakukan oleh (Hana, 2020). Dataset yang digunakan dalam penelitian adalah *Early Stage Diabetes Risk Prediction dataset* yang diambil dari sumber dataset *UCI Machine Learning* yang terdiri dari 17 variabel dan 520 data. Dataset ini dibagi menjadi *data training* (80%) dan *data testing* (20%). Hasil pengujian menunjukkan tingkat akurasi sebesar 97,12%, *precision* 93,02%, dan *recall* 100,00%. Analisis menggunakan Kurva ROC menegaskan bahwa algoritma C4.5 memiliki nilai AUC sebesar 0,994 yang artinya *Excellent Classification*, ini menunjukkan bahwa menggunakan Algoritma C4.5 untuk klasifikasi penyakit diabetes menghasilkan akurasi yang tinggi.

Selain penelitian tentang algoritma C4.5, beberapa peneliti juga telah melakukan penelitian tentang algoritma C4.5 yang di optimasi menggunakan PSO, yaitu: (Arifin dan Herliana, 2020), (Arifin dan Rachman, 2020), (Pratiwi dan Arifin, 2020), (Arizaldi, 2020), dan (Wiantara, 2021).

(Arifin dan Herliana, 2020) melakukan penelitian tentang *optimasi decision tree* dengan menggunakan *particle swarm optimization* (PSO) untuk mengidentifikasi penyakit mata berdasarkan analisis tekstur. Dengan menggunakan 311 citra mata, termasuk 233 citra mata normal dan 78 citra mata berpenyakit (28 glaukoma, 28 katarak, dan 22 uveitis), yang diperoleh dari *Center For Biometrics and Security Research* serta *Eye Health Center*, penelitian ini membagi data menjadi 249 *data training* dan 62 *data testing*. PSO digunakan dalam tahap *learning scheme* untuk mengoptimalkan data dan fitur setiap citra. Metode *decision tree* diterapkan untuk mencari pola klasifikasi dari hasil analisis tekstur menggunakan *Gray-Level Co-occurrence Matrix* (GLCM). Hasilnya, implementasi ekstraksi fitur citra mata dengan metode GLCM dan klasifikasi DT yang dioptimasi oleh PSO mampu meningkatkan akurasi sistem identifikasi penyakit mata hingga mencapai 88,09%, menunjukkan peningkatan kinerja yang signifikan dalam deteksi penyakit mata.

(Arifin dan Rachman, 2020) melanjutkan penelitian sebelumnya dengan fokus pada optimasi *Decision Tree* menggunakan *Particle Swarm Optimization* untuk klasifikasi sel Pap Smear. Dataset yang digunakan berasal dari Herlev dengan total 917 data. Penelitian ini bertujuan untuk mengevaluasi apakah penerapan algoritma optimasi *Particle Swarm Optimization* dapat meningkatkan kinerja

Decision Tree dalam mengklasifikasikan data sel pap smear. Tahap penelitian melibatkan *preprocessing, feature optimization, knowledge rules, evaluation, dan performance report*. Hasil penelitian menunjukkan bahwa *Decision Tree* pada data Sel Pap Smear mencapai akurasi sebesar 91,39% dengan AUC 0,858. Namun, ketika diterapkan *Particle Swarm Optimization* pada *Decision Tree*, akurasi meningkat menjadi 96,76% dengan AUC 0,926, yang termasuk dalam kategori *excellent classification* untuk identifikasi sel pap smear normal dan abnormal.

Penelitian lain yang berhasil meningkatkan akurasi dengan menggunakan algoritma C4.5 yang dioptimasi menggunakan PSO untuk klasifikasi kesuburan pada pria (Pratiwi dan Arifin, 2021), klasifikasi harapan hidup pasien hepatitis (Arizaldi, 2020), dan klasifikasi penyakit demensia (Wiantara, 2021).

2.5 Kesimpulan

Bab ini telah membahas tentang literatur terkait dengan penelitian, teori-teori yang relevan dengan metode yang akan digunakan selama penelitian, dan juga merinci beberapa penelitian lain yang mendukung penelitian ini. Penjelasan dalam bab ini akan menjadi dasar teoritis untuk penelitian yang akan dilakukan.

DAFTAR PUSTAKA

- Arifin, T., & Herliana, A. 2020. *Optimizing Decision Tree Using Particle Swarm Optimization to Identify Eye Diseases Based on Texture Analysis*. Jurnal Teknologi dan Sistem Komputer, 8(1):59–63.
<https://doi.org/10.14710/jtsiskom.8.1.2020.59-63>
- Arifin, T., & R. Rachman. 2020. *Optimasi Decision Tree menggunakan Particle Swarm Optimization untuk Klasifikasi Sel Pap Smear*. JATISI (Jurnal Teknik Informatika dan Sistem Informasi), 7(3):572–579.
<https://doi.org/10.35957/jatisi.v7i3.361>
- Arizaldi. 2020. *Pengaruh Particle Swarm Optimization Pada Algoritma C4.5 Untuk Klasifikasi Harapan Hidup Pasien Hepatitis*. Skripsi Program Teknik Informatika Fakultas Ilmu Komputer Palembang.
- At, E., M, A., F, A., & M, S. 2016. *Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method*. Global Journal of Technology and Optimization, 1, 1-11. <https://doi.org/10.4172/2229-8711.S1111>
- Bates, S., Hastie, T., & Tibshirani, R. 2022. *Cross-validation: What does it estimate and how well does it do it?* (arXiv:2104.00673). arXiv.
<http://arxiv.org/abs/2104.00673>
- Buulolo, E. 2020. *Data Mining Untuk Perguruan Tinggi*. CV Budi Utama, Yogyakarta, Indonesia.
- Chanmee, S., & Kesorn, K. 2023. *Semantic decision Trees: A new learning system for the ID3-Based algorithm using a knowledge base*. *Advanced Engineering Informatics*, 58, 102156.
<https://doi.org/10.1016/j.aei.2023.102156>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2002. *SMOTE: Synthetic minority over-sampling technique*. *Journal of Artificial Intelligence Research*, 16, 321–357.
<https://arxiv.org/abs/1106.1813>
- Chen, CH., Chiu, CC., Hung, SC. et al. BER Performance of Wireless BPSK Communication System in Tunnels With and Without Traffic. *Wireless Pers Commun* 30, 1–12 (2004). <https://doi.org/10.1007/s11277-004-1675-y>

- Han, J., Kamber, M., & Pei, J. 2022. *Data Mining: Concepts and Techniques (4th Edition)*. Katey Birtcher.USA.
- Hana, F. M. 2020. *Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5*. Jurnal SISKOM-KB (Sistem Komputer dan Kecerdasan Buatan), 4(1), 32–39. <https://doi.org/10.47970/siskom-kb.v4i1.173>
- Junior, J. B., Saedudin, R., & Widharta, V. P. 2021. *Perbandingan Akurasi Algoritma Decision Tree dan Algoritma Support Vector Machine Pada Penyakit Diabetes*. *e-Proceeding of Engineering*. 8(5).
- Karim, Z. (2021). *Confusion Matrix*. *ResearchGate*
<https://www.researchgate.net/publication/355096788>
- Kruchten, P. 2004. *The Rational Unified Process: An Introduction (3rd Edition)*. Canada.
- Lestari, Zulkarnain & Siji, S. A. 2021. *Diabetes Melitus: Review Etiologi, Patofisiologi, Gejala, Penyebab, Cara Pemeriksaan, Cara Pengobatan dan Cara Pencegahan*. Jurnal UIN Alauddin Makassar.
- Mercangöz, B. A. (Ed.). 2021. *Applying Particle Swarm Optimization: New Solutions and Cases for Optimized Portfolios (Vol. 306)*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-70281-6>
- Muttaqien, R., Pradana, M. G., & Pramuntadi, A. 2021. *Implementation of Data Mining Using C4.5 Algorithm for Predicting Customer Loyalty of PT. Pegadaian (Persero) Pati Area Office*. *International Journal of Computer and Information System (IJCIS)*, 2(3), 64–68. <https://doi.org/10.29040/ijcis.v2i3.36>
- Ningsih, W. R. 2020. *Penerapan Data Mining Dalam Menentukan Masa Studi Mahasiswa Menggunakan Algoritma C4.5*. Skripsi Program Ilmu Komputer Fakultas Sains Dan Teknologi Medan.
- Novandya, A., & Oktaria. I. 2017. *Penerapan Algoritma Klasifikasi Data Mining C4.5 Pada Dataset Cuaca Wilayah Bekasi*. 6(2):368–372.
- Pratiwi, T. W., & Arifin, T. 2021. *Optimasi Decision Tree Menggunakan Particle Swarm Optimization untuk Klasifikasi Kesuburan pada Pria*. *SISTEMASI*, 10(1), 13. <https://doi.org/10.32520/stmsi.v10i1.967>
- Rahayu, D. S., Afifah, J., & Intan, S. 2023. *Classification of Diabetes Mellitus Using C4.5 Algorithm, Support Vector Machine (SVM) and Linear Regression*. IRPI (Institut Riset dan Publikasi Indonesia).

- Rohmat, L. C. & Nuriyah, R. 2023. *Implementasi Human Resource Information System Berbasis Website Pada Pt Litedex Digital Indonesia*. JATI (Jurnal Mahasiswa Teknik Informatika), 7(1), 720–726. <https://doi.org/10.36040/jati.v7i1.6506>
- Roy, S., & Garg, A. 2017. *Analyzing performance of students by using data mining techniques a literature survey*. 2017 4th IEEE UPCON (Uttar Pradesh Section International Conference on Electrical, Computer and Electronics), 130–133. <https://doi.org/10.1109/UPCON.2017.8251035>
- Samala, R. K. 2023. *Particle Swarm Optimization*. *Artificial Intelligence* (Vol. 15). IntechOpen.London. <https://doi.org/10.5772/intechopen.107156>
- Saritas, M. M., & Yasar, A. 2019. *Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification*. *International Journal of Intelligent Systems and Applications in Engineering*.
- Septiani, W. D., & Marlina, M. 2021. *Comparison Of Decision Tree, Naïve Bayes, And Neural Network Algorithm For Early Detection Of Diabetes*. *Jurnal Pilar Nusa Mandiri*, 17(1), 73–78. <https://doi.org/10.33480/pilar.v17i1.2213>
- Taher, R., Basha, S.H., Abdalla, A. (2023). Improving Machine Learning Techniques with Imbalanced Data Treatment for Predicting Diabetes. In: Hassanien, A., Rizk, R.Y., Pamucar, D., Darwish, A., Chang, KC. (eds). *Proceedings of the 9th International Conference on Advanced Intelligent Systems and Informatics 2023*. AISI 2023. Lecture Notes on Data Engineering and Communications Technologies, vol 184. Springer, Cham. https://doi.org/10.1007/978-3-031-43247-7_34
- Wang, K. J., Makond, B., & Chen, K. H. 2014. *A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients*. *Applied Soft Computing*, 20, 15–24. <https://doi.org/10.1016/j.asoc.2014.03.014>
- Wiantara, T. 2021. *Optimasi Metode C4.5 Menggunakan Particle Swarm Optimization untuk Klasifikasi Penyakit Demensia*. Skripsi Program Teknik Informatika Fakultas Ilmu Komputer Palembang.
- Zhang, T., Chen, J., Li, F., Zhang, K., Lv, H., He, S., & Xu, E. 2021. *Intelligent fault diagnosis of machines with small & imbalanced data: a state-of-the-art review and possible extensions*. *ISA Transactions*, 119, 152–171. <https://doi.org/10.1016/j.isatra.2021.04.025>