PENERAPAN *NAÏVE BAYES* BERBASIS *PARTICLE SWARM OPTIMIZATION* UNTUK KLASIFIKASI DIABETES MELLITUS

SKRIPSI

Oleh :
ROUDLOTUL HANNAH
NIM. 210605110003

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2025

PENERAPAN *NAÏVE BAYES* BERBASIS *PARTICLE SWARM OPTIMIZATION* UNTUK KLASIFIKASI DIABETES MELLITUS

SKRIPSI

Diajukan kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)

Oleh:
ROUDLOTUL HANNAH
NIM. 210605110003

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2025

HALAMAN PERSETUJUAN

PENERAPAN *NAÏVE BAYES* BERBASIS *PARTICLE SWARM OPTIMIZATION* UNTUK KLASIFIKASI DIABETES MELLITUS

SKRIPSI

Oleh :
ROUDLOTUL HANNAH
NIM. 210605110003

Telah Diperiksa dan Disetujui untuk Diuji:
Tanggal: 4 Juni 2025

Pembimbing I,	Pembimbing II,
Prof. Dr. Suhartono, M.Kom	Ajib Hanani, M.T
NIP. 19680519 200312 1 001	NIP. 19840731 202321 1 013

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Ir. Fachrul Kurniawan, M.MT., IPU
NIP. 19771020 200912 1 001

HALAMAN PENGESAHAN

PENERAPAN *NAÏVE BAYES* BERBASIS *PARTICLE SWARM OPTIMIZATION* UNTUK KLASIFIKASI DIABETES MELLITUS

SKRIPSI

Oleh :
ROUDLOTUL HANNAH
NIM. 210605110003

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 13 Juni 2025

Susunan Dewan Penguji

Ketua Penguji	: Dr. M. Ainul Yaqin, M.Kom NIP. 19761013 200604 1 004	( )
Anggota Penguji I	: Supriyono, M.Kom NIP. 19841010 201903 1 012	( )
Anggota Penguji II	: Prof. Dr. Suhartono, M.Kom NIP. 19680519 200312 1 001	( )
Anggota Penguji III	: Ajib Hanani, M.T NIP. 19840731 202321 1 013	( )

Mengetahui dan Mengesahkan,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Ir. Fachrul Kurniawan, M.MT., IPU
NIP. 19771020 200912 1 001

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama	: Roudlotul Hannah
NIM	: 210605110003
Fakultas / Program Studi	: Sains dan Teknologi / Teknik Informatika
Judul Skripsi	: Penerapan *Naïve Bayes* Berbasis *Particle Swarm Optimization* untuk Klasifikasi Diabetes Mellitus

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambil alihan data, tulisan, atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan skripsi ini merupakan hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 13 Juni 2025
Yang membuat pernyataan,

Roudlotul Hannah
NIM. 210605110003

MOTTO

"Tidak ada kesuksesan tanpa kerja keras. Tidak ada keberhasilan tanpa kebersamaan. Tidak ada kemudahan tanpa do'a"

HALAMAN PERSEMBAHAN

Puji syukur atas kehadirat Allah Subhanahu wa ta'ala, karena berkat rahmat dan petunjuk-Nya, sehingga penulis dapat menyelesaikan skripsi ini dengan baik dan lancar. Shalawat dan sakan semoga tetap tercurahkan kepada Rasulullah Shallallahu 'alaihi wasallam, yang senantiasa membawa kita dari zaman jahiliyah menuju yang yang benar yakni addinul Islam. Skripsi ini tidak akan selesai tanpa adanya kontribusi dan dukungan dari berbagai pihak. Oleh karena itu, penulis mempersembahkan skripsi ini kepada seluruh pihak yang telah berjasa dalam pengerjaan penelitian ini. Karya ini penulis persembahkan kepada:

1. Keluarga tercinta Bapak Masyhudi dan Ibu Genduk Ifa'atun selaku orang tua penulis, serta Ismaudin Walidul Awwal A.Md.T, Shohihatun Nisa' Yulia Isnaini S.Tr.Keb selaku kakak penulis, serta M. Zam'ul Fawwaid Al-Ukhro sebagai adik dari penulis, yang sangat penulis cintai dan tidak pernah berhenti dalam memberikan semangat, dukungan maupun do'a kepada penulis hingga penulis dapat menyelesaikan skripsi ini dengan baik dan lancar.
2. Bapak Prof. Dr. Suhartono, M.Kom selaku dosen pembimbing I serta Bapak Ajib Hanani, M.T selaku dosen pembimbing II atas segala ilmu yang telah diberikan kepada penulis dan senantiasa membimbing, memberikan semangat, memberi arahan dan masukan, serta membantu penulis dalam mengerjakan dan menyelesaikan skripsi ini.

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh

Alhamdulillahirabbilalamin, segala puji dan rasa syukur yang senantiasa penulis panjatkan kepada baginda Rasulullah Shallallahu 'alaihi wasallam, atas berkat dan hidayah-Nya, sehingga penulis dapat menyelesaikan skripsi yang berjudul "Penerapan *Naïve Bayes* Berbasis *Particle Swarm Optimization* untuk Klasifikasi Diabetes Mellitus" dengan baik dan lancar. Skripsi ini diajukan untuk memenuhi salah satu syarat lulus sebagai sarjana komputer (S.Kom) pada Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang.

Penulisan skripsi ini tentunya tidak akan berjalan dengan baik tanpa adanya dukungan dan bantuan dari berbagai pihak, baik secara langsung maupun tidak langung. Oleh karena itu, dengan penuh rasa hormat, penulis menyampaikan rasa terima kasih yang sebesar-besarnya kepada:

1. Prof. Dr. M. Zainuddin, M.A., selaku rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Prof. Dr. Hj. Sri Harini, M.Si., selaku dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Dr. Ir. Fachrul Kurniawan, M.MT, IPU., selaku Ketua Program Studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.

4. Prof. Dr. Suhartono, M.Kom selaku pembimbing utama yang ddengan kesabaran dan ketulusan hati memberikan bimbingan, arahan, serta dorongan dalam setiap penyusunan skripsi ini.
5. Ajib Hanani, M.T selaku dosen pembimbing kedua penulis yang selalu memberikan bimbingan hingga terselesaikannya skripsi ini.
6. Dr. M. Ainul Yaqin, M.Kom selaku penguji utama dan Supriyono, M.Kom selaku pengji kedua yang telah berkenan menguji serta memberikan masukan sehingga skripsi ini dapat terselesaikan dengan baik.
7. Nia Faricha S, Si., selaku admin Program Studi Teknik Informatika yang selalu sabar memberikan informasi, membantu, dan memberikan arahan selama perkuliahan dan proses penulisan skripsi ini.
8. Segenap dosen, laboran, dan jajaran Staff Program Studi Teknik Informatika yang telah memberikan ilmu, pengetahuan, dan dukungan selama penulis menjalani studi hingga selesainya skripsi ini.
9. Kepada kedua orang tua tercinta, Abah Masyhudi dan Ibu Genduk Ifa'atun, terima kasih atas cinta yang tidak pernah habis, do'a yang selalu mengiringi, dan kesabaran tanpa batas yang menjadi cahaya di setiap langkah penulis. Beliau merupakan alasan utama penulis mampu bertahan dan menyelesaikan perjalanan ini. Terima kasih atas kekuatan dan ketulusan yang tak pernah pudar. Semoga Allah SWT membalasnya dengan limpahan rahmat dan keberkahan.
10. Kepada kakak-kakak tercinta Ismaudin Walidul Awwal, A.Md.T dan Shohihatun Nisa' Yulia Isnaini S.Tr.Keb., yang telah menjadi sumber

inspirasi, motivasi, dan dukungan selama perjalanan ini. Terima kasih atas semangat, nasihat, serta perhatian yang tak henti diberikan kepada penulis. Kepada adik tersayang M. Zam'ul Fawwaid Al-Ukhro yang kehadirannya turut memberi semangat dan ketulusan do'a dalam setiap langkah penulis. Dan kepada seluruh keluarga besar yang tiada henti memberikan do'a dan dukungan sehingga penulis mampu menyelesaikan skripsi ini.

11. Kepada teman beda umur penulis, Nur Shasmitta Zaen, yang telah menemani penulis, memberikan semangat, serta segala bantuan pada saat penulisan skripsi ini. Dukungan dan perhatianmu selalu menjadi penyemangat bagi penulis, terutama saat-saat sulit. Terima kasih atas kehadiranmu yang tidak hanya menjadi tempat berbagi cerita, tetapi juga selalu ada untuk mendengarkan dan menguatkan penulis.
12. Kepada teman-teman tersayang, Nabila Mahdiya Putri, Amalia Amriadi, Nurjihan Nabila Ramadhani, serta seluruh teman yang telah menjadi bagian penting dalam perjalanan studi ini. Terima kasih atas dukungan yang tulus, serta segala bantuan yang diberikan kepada penulis. Kehadiran kalian telah menjadi penguat dalam menjalani masa perkuliahan dan proses penulisan skripsi ini.
13. Teman-teman seperjuangan "Anak Prof", Umi Kunhayati, An Nisa' Puja Karimah Attamimi, dan Sita Maulida. Kehadiran kalian memberikan warna dalam perjalanan ini dengan segala bantuan, perhatian, dan dukungan yang tidak pernah kalian ragu berikan. Bersama kalian, penulis merasa tidak

pernah sendirian. Terima kasih telah menjadi teman sekaligus penyemangat bagi penulis.

14. Teman dekat penulis “Bismillah Kaya Raya”, Ria dan Khusnul, meskipun kini terpisah jarak, namun tetap menjadi bagian dalam perjalanan penulis. Terima kasih atas kebersamaan, perhatian, dan semangat yang selalu diberikan, meski tidak selalu dalam satu ruang yang sama. Hubungan jarak jauh tidak mengurangi rasa hangatnya persahabatan kita. Kehadiran, dukungan, dan pesan-pesan sederhana yang dikirimkan dari kejauhan telah menjadi penguat tersendiri bagi penulis dalam menyelesaikan skripsi ini. Semoga suatu saat kita bisa berkumpul kembali dalam keadaan yang lebih baik dan lebih bahagia tentunya.
15. Seluruh warga Teknik Informatika, khususnya angkatan 2021 “Aster”, yang telah memberikan kehangatan, semangat, dan motivasi dalam proses akademik mamupun non-akademik.
16. Kepada orang teristimewa yang tidak bisa penulis sebutkan namanya, yang telah menjadi sosok ayah untuk penulis, selalu menjadi pendukung terbaik dalam setiap langkah kehidupan. Terima kasih atas kasih sayang, perhatian, dan kekuatan yang senantiasa diberikan. Kehadiran dan do’a yang tulus telah menjadi sumber semangat yang luar biasa bagi penulis dalam menyelesaikan studi ini. Semoga segala kebaikan dibalas dengan limpahan rahmat dan keberkahan dari Allah SWT.
17. Seluruh pihak yang telah terlibat, baik secara langsung maupun tidak langsung dari awal perkuliahan hingga akhir penulisan skripsi ini.

Penulis menyadari bahwa penyusunan skripsi ini masih perlu disempurnakan. Maka dari itu penulis menerima saran, kritik, dan masukan yang bersifat membangun untuk meningkatkan kualitas dan pengembangan lebih lanjut terhadap penelitian ini. Penulis percaya bahwa proses pembelajaran tidak memiliki garis akhir dan setiap masukan dapat menjadi langkah berharga dalam perjalanan akademik dan pengembangan ilmu pengetahuan. Semoga skripsi ini dapat memberikan manfaat nyata bagi pembaca, peneliti selanjutnya, maupun pihak-pihak terkait. Serta semoga penelitian ini memberikan kontribusi bagi pengembangan ilmu pengetahuan dan dapat bermanfaat di masa mendatang.

Wassalamu'alaikum warahmatullahi wabarakatuh.

Malang, 13 Juni2025

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Hannah, Roudlotul. 2025. **Penerapan *Naïve Bayes* Berbasis *Particle Swarm Optimization* untuk Klasifikasi Diabetes Mellitus**. Skripsi. Jurusan Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Prof. Dr. Suhartono, M.Kom. Nama Pembimbing (II) Ajib Hanani, M,T.

.

Kata kunci: Diabetes Mellitus, *Klasifikasi*, Machine Learning, Naïve Bayes, Particle Swarm Optimization.

Diabetes Mellitus merupakan penyakit kronis yang prevalensinya terus meningkat setiap tahun. Deteksi dini menjadi sangat penting guna mencegah komplikasi lebih parah. Penelitian ini bertujuan untuk membangun model klasifikasi diabetes mellitus dengan menerapkan algoritma Naïve Bayes yang dioptimasi menggunakan Particle Swarm Optimization (PSO). Dataset yang digunakan adalah Pima Indian Diabetes Dataset yang terdiri dari 768 data dengan delapan fitur utama. Tahapan penelitian mencakup preprocessing data, penanganan missing value, normalisasi, teknik balancing menggunakan SMOTE, serta pengujian performa menggunakan 10-Fold Cross Validation. Evaluasi dilakukan dengan membandingkan hasil klasifikasi sebelum dan sesudah proses optimasi PSO. Hasil menunjukkan bahwa model Naïve Bayes yang dioptimasi mampu meningkatkan akurasi dari 72.73% menjadi 74.86% ketika menggunakan teknik optimasi PSO. Temuan ini membuktikan bahwa kombinasi Naïve Bayes dan PSO dapat menghasilkan klasifikasi yang lebih akurat dan efisien dalam deteksi dini Diabetes Mellitus. Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan sistem pendukung keputusan berbasis kecerdasan buatan di bidang kesehatan.

ABSTRACT

Hannah, Roudlotul. 2025. **Application of Naïve Bayes Based on Particle Swarm Optimization for Diabetes Mellitus Classification**. Thesis. Department of Informatics Engineering, Faculty of Science and Technology, Maulana Malik Ibrahim State Islamic University, Malang. Supervisor: (I) Prof. Dr. Suhartono, M.Kom. Supervisor Name (II) Ajib Hanani, M,T.

Diabetes mellitus is one of the chronic diseases whose incidence continues to rise year after year. Classifying diabetes patients is an important step in detecting the disease early. This study aims to build a diabetes mellitus classification model using the Naïve Bayes algorithm optimized with the Particle Swarm Optimization (PSO) method. The model was built using data from the Pima Indian Diabetes Dataset, which consists of 768 data points with eight relevant features, such as blood glucose levels, blood pressure, body mass index, age, and number of pregnancies. The research process included data preprocessing, feature selection using PSO, model training, and performance evaluation. Evaluation was conducted by comparing the results of the Naïve Bayes model before and after optimization using PSO. The results showed that the accuracy of the Naïve Bayes model increased from 71% to 74% after optimization with PSO. The optimized model was able to select the best features and produce more accurate and efficient classifications. This study is expected to contribute scientifically to the development of artificial intelligence-based classification systems for the early diagnosis of Diabetes Mellitus, as well as support rapid and accurate medical decision-making.

Keywords: Classification, Diabetes Mellitus, Machine Learning, Naïve Bayes, Particle Swarm Optimization

ملخص

هانا، رودلوتول. 2025. تطبيق نايف بايز القائم على تحسين سرب الجسيمات لتصنيف مرض السكري. أطروحة. قسم هندسة المعلوماتية، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم الإسلامية، مالانج. المشرف (I) :الأستاذ الدكتور سوهارتونو، ماجستير في علوم الكمبيوتر. اسم المشرف (II) أجيب حناني، ماجستير في التكنولوجيا.

الكلمات المفتاحية: نايف بايز، تحسين سرب الجسيمات، التصنيف، داء السكري، التعلم الآلي، التحقق المتقاطعK-Fold

مرض السكري هو أحد الأمراض المزمنة التي تزداد حالات الإصابة بها عامًا بعد عام. يعد تصنيف مرضى السكري خطوة مهمة في الكشف المبكر عن المرض. تهدف هذه الدراسة إلى بناء نموذج لتصنيف مرض السكري باستخدام خوارزمية Naïve Bayes المُحسّنة بطريقة Particle Swarm Optimization (PSO).تم بناء النموذج باستخدام بيانات من مجموعة بيانات مرض السكري لدى هنود بيما، والتي تتكون من 768 بيانات مع ثمانية ميزات ذات صلة، مثل مستوى الجلوكوز، وضغط الدم، ومؤشر كتلة الجسم، والعمر، وعدد الحمل. وتشمل مراحل البحث معالجة البيانات المسبقة، واختيار الميزات باستخدامPSO ، وتدريب النموذج، وتقييم الأداء. تم إجراء التقييم من خلال مقارنة نتائج نموذج Naïve Bayes قبل وبعد تحسينه باستخدام .PSO أظهرت نتائج البحث أن دقة نموذج Naïve Bayes زادت من 71٪ إلى 74٪ بعد تحسينه باستخدام .PSO تمكن النموذج المحسّن من اختيار أفضل الميزات وإنتاج تصنيف أكثر دقة وكفاءة. من المتوقع أن يساهم هذا البحث في تطوير نظام تصنيف قائم على الذكاء الاصطناعي لتشخيص مرض السكري في مرحلة مبكرة، فضلاً عن دعم اتخاذ القرارات الطبية بسرعة ودقة.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Diabetes mellitus (DM) merupakan salah satu penyakit kronis yang ditandai dengan meningkatnya kadar gula darah *(hiperglikemia)* yang terjadi karena kelainan gangguan sekresi insulin, kerja insulin, atau keduanya menurut (Ginting et al., 2022). Penyakit diabetes ini merupakan penyakit berbahaya yang mengalami peningkatan tiap tahunnya yang dipengaruhi berbagai faktor seperti tekanan darah tinggi, kadar gula berlebih, berat badan, riwayat keturunan diabetes, jumlah kadar insulin dalam tubuh, kurangnya aktivitas fisik dan pola gaya hidup, serta diet yang tidak sehat menurut (Aprillia et al., 2022). Diabetes dapat menyebabkan komplikasi yang berdampak buruk bagi fungsi mata, jantung, ginjal, kulit, saraf, sampai saluran pernapasan. Oleh karena itu, memprediksi atau mendiagnosa awal dapat menjadi langkah yang tepat untuk menghindari penyakit ini (Diana Dewi et al., 2023).

Diabetes merupakan salah satu penyakit yang disebabkan karena kadar gula darah di dalam tubuh manusia yang tinggi atau melampaui batas normal. Menurut *World Health Organizations* (WHO), di seluruh dunia, lebih dari 422 juta individu saat ini yang mengidap diabetes mellitus, dengan mayoritas populasi berada di negara-negara berpenghasilan menengah ke bawah. Setiap tahunnya, sekitar 1,5 juta kematian dapat secara langsung dikaitkan dengan diabetes. Kasus dan prevalensi diabetes terus mengalami peningkatan selama beberapa tahun terakhir.

Indonesia menempati peringkat kelima dengan jumlah kasus diabetes tertinggi di dunia, dengan total 19,5 juta orang yang telah terdiagnosis.

Penelitian ini selaras dengan prinsip Islam yang mendorong umatnya untuk berikhtiar dalam mencari pengobatan atas setiap penyakit. Rasulullah SAW juga memotivasi umatnya untuk berikhtiar mencari pengobatan, sebagaimana dalam sebuah hadis yang diriwayatkan oleh Muslim:

مَا أَنْزَلَ اللهُ دَاءً إِلاَّ أَنْزَلَ لَهُ شِفَاء

"Sesungguhnya Allah tidak menurunkan suatu penyakit kecuali menurunkan pula obatnya."(HR. Bukhari No. 5678).

Hadis ini menjadi motivasi penting dalam pengembangan ilmu pengetahuan, khususnya dalam bidang kesehatan, bahwa setiap penyakit dapat ditangani apabila manusia berikhtiar untuk mencari pengobatan dan solusi yang tepat. Sejalan dengan semangat tersebut, penelitian ini bertujuan untuk mendukung proses klasifikasi penyakit diabetes mellitus secara dini melalui pendekatan teknologi kecerdasan buatan.

Dengan menerapkan algoritma *Naive Bayes* yang dioptimasi menggunakan *Particle Swarm Optimization* (PSO), penelitian ini berupaya mengklasifikasikan data pasien secara akurat untuk menentukan apakah seseorang berisiko menderita diabetes atau tidak. Upaya ini bertujuan agar penanganan terhadap penyakit dapat dilakukan lebih cepat dan tepat, sehingga dapat meminimalisir risiko komplikasi yang lebih parah. Dengan demikian, penelitian ini menjadi bentuk kontribusi dalam merealisasikan nilai-nilai Islam, khususnya dalam aspek ikhtiar dan upaya menemukan pengobatan bagi setiap penyakit yang Allah turunkan.

Klasifikasi penyakit diabetes mellitus menjadi langkah penting dalam proses penanganan dan pengendalian penyakit secara lebih awal. Karena sering kali penyakit ini tidak menunjukkan gejala pada awalnya, dan baru diketahui setelah muncul komplikasi yang serius. Menurut data dari WHO, lebih dari 50% kasus diabetes tidak terdeteksi. Untuk itu, metode klasifikasi berbasis data bisa menjadi solusi yang efektif karena kecepatannya. Penelitian ini memiliki kebaharuan pada penggunaan metode *Naïve Bayes* yang dioptimalkan dengan algoritma *Particle Swarm Optimization* (PSO). Sebagian besar penelitian sebelumnya hanya menggunakan *Naïve Bayes* tanpa proses optimasi. Dengan menggabungkan kedua metode tersebut, diharapkan dapat meningkatkan akurasi prediksi dan menghasilkan model klasifikasi yang lebih akurat untuk membantu proses deteksi dini terkait penyakit diabetes.

Dalam beberapa tahun terakhir, kemajuan dalam bidang teknologi informasi dan komputasi telah menciptakan peluang baru untuk memperbaiki proses diagnosis dan pengelolaan penyakit, termasuk diabetes. Salah satu cara untuk mendeteksi penyakit diabetes yaitu dengan memanfaatkan algoritma *machine learning* (Abdulhadi & Al-Mousa, 2021). *Naïve Bayes* merupakan algoritma pembelajaran mesin yang sering digunakan. Metode ini didasarkan pada konsep probabilitas dan klasifikasi statistik yang dikembangkan oleh Thomas Bayes seorang ilmuwan asal Inggris. Arti lain dari *Naïve Bayes* yaitu suatu metode klasifikasi berdasarkan probabilistik statistik yang memprediksi peluang dimasa depan berdasarkan pengalaman dari masa lalu (Gurning et al., 2024). Selain itu, klasifikasi dengan menggunakan algoritma *Naïve Bayes* melibatkan kumpulan data

pelatihan berdimensi tinggi. Dengan algoritma *Naïve Bayes* dapat membangun model dengan cepat dan menjadikannya *algoritma* prediksi yang paling cepat untuk dipelajari. Algoritma *Naïve Bayes* hanya mendukung pada atibut yang bertipe data *discrete* atau *discretized*.

Particle Swarm Optimization (PSO) merupakan salah satu teknik yang dapat digunakan untuk menyelesaikan masalah optimasi. *Particle Swarm Optimization* (PSO) sering digunakan dalam sebuah penelitian, karena *Particle Swarm Optimization* mudah diterapkan dan ada beberapa parameter untuk menyesuaikannya (Mutiara, 2020)

Beberapa penelitian telah dilakukan dalam upaya mendeteksi diabetes menggunakan metode machine learning. Menurut penelitian yang dilakukan oleh, menunjukkan bahwa algoritma *Naïve Bayes* mencapai akurasi sebesar 92% dalam memprediksi seorang menderita diabetes atau tidak dengan melalui tahapan pra-pemrosesan. Selain itu, penelitian juga telah dilakukan oleh (Maulidah et al., 2020) yang menerapkan *Particle Swarm Optimization* (PSO) pada *algoritma Naïve Bayes* dalam klasifikasi diabetes mellitus. Hasil penelitian menunjukkan metode *Naïve Bayes* mencapai akurasi sebesar 74,61%. Sedangkan dengan menerapkan *Particle Swarm Optimization* (PSO) untuk seleksi fitur akurasi meningkat menjadi 77,34%, mengalami peningkatan sebesar 2.73% dibandingkan dengan penggunaan *Naïve Bayes* tanpa optimasi. Hasil ini menunjukkan bahwa metode PSO mampu meningkatkan performa algoritma *Naïve Bayes* dalam mendiagnosis diabetes.

Penelitian kali ini membangun suatu sistem klasifikasi untuk memprediksi apakah seseorang sebagai penderita diabetes atau tidak berdasarkan dataset Pima

India Diabates Dataset dengan menggunakan metode *Naïve Bayes* berbasis *Particle Swarm Optimization* (PSO). Dengan demikian, penelitian ini diharapkan mampu memudahkan dalam melakukan pengklasifikasian penyakit diabetes yang efektif, sehingga seseorang dapat diketahui penyakit diabetes lebih awal agar dapat menekan angka kematian penyakit diabetes di Indonesia.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan sebelumnya, maka pernyataan masalah pada penelitian ini adalah bagaimana penerapan metode *Naïve Bayes* yang dioptimasi dengan algoritma *Particle Swarm Optimization* (PSO) dapat meningkatkan akurasi klasifikasi penyakit diabetes mellitus.

1.3 Batasan Masalah

a) Penelitian ini menggunakan Pima Indian Diabetes Dataset yang telah tersedia secara publik dengan jumlah data sebanyak 768 pasien yang terdiri dari 9 atribut sebagai data utama yang membangun model klasifikasi diabetes.
b) Penelitian ini hanya bertujuan untuk memprediksi apakah seseorang menderita diabetes atau tidak (klasifikasi biner: positif atau negatif), tanpa mengklasifikasikan jenis diabetes (seperti tipe 1, tipe 2, atau gestasional)**.** Hal ini disebabkan oleh keterbatasan informasi dalam dataset yang digunakan, yang hanya menyediakan label diagnosis secara biner.
c) Evaluasi performa model dibatasi pada metrik seperti *accuracy*, *precision*, *recall*, *f1-score*.

d) Penelitian ini tidak mencakup implementasi sistem di dunia nyata atau uji coba langsung pada pasien, tetapi hanya sebatas simulasi dan pengujian model menggunakan dataset yang telah tersedia.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini yaitu menerapkan metode *Naïve Bayes* berbasis *Particle Swarm Optimization* (PSO) untuk klasifikasi diabetes mellitus.

1.5 Manfaat Penelitian

Kegunaan yang dapat dihasilkan dari penelitian dalam tugas akhir ini adalah sebagai berikut:

a) Memberikan kontribusi pada pengembangan klasifikasi penyakit berbasis kecerdasan buatan, khususnya dengan menggunakan metode *Naïve Bayes* berbasis *Particle Swarm Optimization* (PSO).
b) Hasil penelitian ini dapat dijadikan sebagai acuan atau referensi bagi penelitian lanjutan yang ingin mengembangkan sistem prediksi kesehatan menggunakan metode serupa atau dataset yang berbeda.

BAB II

STUDI PUSTAKA

2.1 Penelitian Terkait

Dalam penelitian yang dilakukan oleh (Anisa & Jumanto, 2022) menggunakan dataset yang berasal dari *Kaggle* dengan jumlah 390 data, yang terdiri dari 60 pasien positif diabetes dan 330 pasien non- diabetes. Atribut dataset terdiri dari lima belas variabel input diantaranya cholosterol, glucose, hdl_chol, chol_hdl_ratio, age, gender, height, weight, bmi, systolic_bp, diastolic_bp, waist, hip, waist_hip_ratio, diabetes. Sebanyak 234 data digunakan sebagai data latih (*training*) dan 154 data digunakan sebagai data uji (*testing*). Dari dataset tersebut dilakukan tahapan *preprocessing*.

Preprocessing data dilakukan dengan menghilangkan beberapa atribut yang tidak terlalu dibutuhkan. Dari 15 atribut menjadi 9 atribut yang digunakan pada penelitian ini. Kemudian melakukan *scaling* terhadap data untuk normalisasi data. Hasil evaluasi menunjukkan bahwa model *Naïve Bayes* yang digunakan mampu mencapai akurasi sebesar 92,3% pada data latih dan 91,6% pada data uji. Penelitian ini menunjukkan bahwa metode *Naïve Bayes* memiliki potensi yang baik dalam prediksi diabetes. Perbedaan dengan penelitian ini terletak pada pendekatan yang digunakan, di mana dalam penelitian yang dilakukan oleh (Anisa & Jumanto, 2022) hanya menerapkan *Naïve Bayes* secara konvensional tanpa optimasi parameter. Sementara penelitian ini menggunakan kombinasi antara algoritma *Naïve Bayes* dan teknik optimasi *Particle Swarm Optimization* (PSO) untuk meningkatkan

akurasi klasifikasi. Selain itu, penelitian ini juga menggunakan dataset yang berbeda yaitu Pima Indian Diabetes Dataset yang memiliki jumlah 768 data dan menerapkan teknik SMOTE untuk menyeimbangkan distribusi kelas, sehingga diharapkan menghasilkan model yang lebih baik dan optimal.

Dalam penelitian yang dilakukan oleh (Pradhani et al., 2025) dataset yang digunakan berasal dari rumah sakit Sylhet Diabetes Hospital di Bangladesh yang tersedia di platform Kaggle. Dataset ini terdiri dari 520 data pasien dengan 16 atribut klinis seperti age, gender, polyuria, polydipsia, sudden_weight_loss, weakness, polyphagia, genital_thrush, visual_blurring, itching, irritability, delayed_healing, partial_paresis, muscle_stiffness, alopecia, dan obesity. Seluruh data melalui proses *preprocessing* yang mencakup penanganan missing value, normalisasi, dan seleksi fitur. Penelitian ini menerapkan algoritma *Gaussian Naïve Bayes* dalam pemodelan klasifikasi untuk memprediksi risiko diabetes mellitus.

Proses training dan evaluasi model dilakukan menggunakan bahasa pemrograman *Python* serta pustaka scikit-learn, pandas, dan matplotlib. Model kemudian diuji dengan pendekatan k-fold cross validation untuk mengurangi *overfitting* dan meningkatkan generalisasi model. Hasil evaluasi menunjukkan bahwa algoritma *Gaussian Naïve Bayes* mampu mencapai tingkat akurasi sebesar 91%, dengan klasifikasi pasien ke dalam kategori positif atau negatif diabetes berdasarkan probabilitas tertinggi. Penelitian ini menunjukkan bahwa metode *Gaussian Naïve Bayes* cukup andal dalam memproses data klinis untuk prediksi dini diabetes. Perbedaan dengan penelitian ini terletak pada pendekatan yang digunakan, di mana dalam penelitian (Pradhani et al., 2025) hanya menggunakan

metode *Gaussian Naïve Bayes* secara konvensional tanpa teknik optimasi atau penyeimbangan data. Sementara dalam penelitian ini, akan digunakan kombinasi antara algoritma *Naïve Bayes* dan teknik optimasi *Particle Swarm Optimization* (PSO) guna mengoptimalkan parameter dan meningkatkan akurasi *klasifikasi*. Selain itu, penelitian ini juga menggunakan dataset yang berbeda, yaitu Pima Indian Diabetes Dataset dengan 768 data, serta menerapkan teknik SMOTE untuk menangani ketidakseimbangan kelas, dengan tujuan menghasilkan model prediksi yang lebih presisi dan stabil.

Dalam penelitian yang dilakukan oleh (Muhammadiyah Jember & Tri Rahayu, 2022) menggunakan dataset Pima Indian Diabetes dari *Kaggle* yang berjumlah 768 data dengan 8 atribut yaitu kehamilan, glukosa, tekanan darah, ketebalan kulit, insulin, BMI, riwayat keturunan, dan umur. Penelitian ini membandingkan performa dua algoritma klasifikasi, yaitu *K-Nearest Neighbor* (KNN) dan *Gaussian Naïve Bayes* (GNB). Klasifikasi dilakukan menggunakan teknik k-fold cross validation (k=2 hingga k=10) dan evaluasi dilakukan dengan metrik akurasi, sensitivitas, spesifisitas, presisi, dan error rate. Hasil evaluasi menunjukkan bahwa algoritma *Gaussian Naïve Bayes* memiliki performa yang lebih baik dibandingkan KNN dengan akurasi sebesar 73,50%, sensitivitas 96,14%, spesifisitas 86,45%, presisi 78,98%, dan error rate 29,47%. Sebaliknya, algoritma KNN menghasilkan akurasi tertinggi sebesar 71,27%, sensitivitas 76,86%, spesifisitas 81,59%, presisi 76,15%, dan error rate 35,05%. Penelitian ini menunjukkan bahwa *Gaussian Naïve Bayes* lebih unggul dalam menangani data

klasifikasi diabetes, terutama pada dataset dengan jumlah atribut yang tidak terlalu banyak.

Perbedaan dengan penelitian ini terletak pada pendekatan yang digunakan, di mana dalam penelitian (Muhammadiyah Jember & Tri Rahayu, 2022) hanya membandingkan dua metode secara konvensional tanpa melibatkan teknik optimasi. Sementara penelitian ini mengusulkan penggunaan algoritma *Naïve Bayes* yang dikombinasikan dengan *Particle Swarm Optimization* (PSO) untuk mengoptimalkan parameter model serta menerapkan teknik SMOTE dalam menangani ketidakseimbangan kelas. Selain itu, fokus penelitian ini adalah pada peningkatan performa prediksi diabetes melalui integrasi metode klasifikasi dan optimasi, sehingga diharapkan menghasilkan akurasi yang lebih tinggi dan model yang lebih stabil.

Penelitian yang telah dilakukan oleh (Bangun & Rachmat, 2024) membahas dalam hal membandingkan kinerja algoritma *K-Nearest Neighbors* (KNN) dan *Naïve Bayes* dalam mendiagnosa penyakit diabetes. Dataset yang digunakan dalam penelitian ini berasal dari *Kaggle* dengan total 768 data, yang terdiri dari 9 atribut. Penelitian ini menggunakan skema pembagian data 80% untuk data latih (614 data) dan 20% untuk data uji (154 data) serta menerapkan model 5-fold-cross-validation untuk validasi model. Hasil penelitian menunjukkan bahwa algoritma *Naïve Bayes* memiliki akurasi lebih tinggi mencapai 74,7%, dibandingkan dengan algoritma KNN yang hanya mencapai akurasi sebesar 68,6%, dengan selisih akurasi sebesar 6,1%. Selain itu, perbandingan menggunakan kurva ROC (*Receiver Operating Characteristic*) menunjukkan bahwa algoritma *Naïve Bayes* memiliki nilai AUC

sebesar 0,823 yang lebih baik dibandingkan dengan algoritma KNN yang memiliki nilai AUC sebesar 0,710.

Berdasarkan hasil tersebut, penelitian ini menyimpulkan bahwa algoritma *Naïve Bayes* lebih unggul dalam memprediksi diabetes dibandingkan dengan KNN dan dapat digunakan sebagai alternatif dalam sistem diagnosa penyakit diabetes berbasis data mining. Perbedaan dengan penelitian ini terletak pada pendekatan dan metode optimasi yang digunakan. Penelitian (Bangun & Rachmat, 2024) hanya membandingkan dua algoritma klasifikasi tanpa melakukan proses optimasi parameter, sedangkan penelitian ini menerapkan metode *Particle Swarm Optimization* (PSO) untuk mengoptimalkan parameter dalam algoritma *Naïve Bayes* agar menghasilkan akurasi yang lebih tinggi. Selain itu, penelitian ini juga menerapkan teknik *Synthetic Minority Over-sampling Technique* (SMOTE) untuk menangani ketidakseimbangan kelas dalam data, sedangkan penelitian (Bangun & Rachmat, 2024) tidak mencantumkan penggunaan metode penyeimbangan data. Pendekatan ini membuat penelitian ini lebih fokus pada peningkatan performa model melalui optimasi.

Dalam penelitian yang dilakukan oleh (Susilowati et al., 2023) menggunakan metode K-Nearest Neighbor (KNN) yang dikombinasikan dengan algoritma *Particle Swarm Optimization* (PSO) sebagai teknik seleksi fitur untuk meningkatkan akurasi *klasifikasi* penyakit diabetes. Dataset yang digunakan adalah Pima Indian Diabetes Dataset dari *Kaggle*, yang terdiri dari 768 data dengan 8 atribut, antara lain: pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, dan age.

Proses penelitian dilakukan dalam beberapa tahapan, mulai dari *preprocessing* data (normalisasi min-max), pembagian data menggunakan 10-fold cross-validation, pencarian nilai k-optimal (diperoleh k=19), hingga proses klasifikasi dan evaluasi model. Hasil klasifikasi dengan algoritma KNN tanpa seleksi fitur menunjukkan akurasi sebesar 75%. Setelah dilakukan seleksi fitur menggunakan Binary *Particle Swarm Optimization* (BPSO), diperoleh peningkatan akurasi menjadi 77,213%, serta fitur-fitur yang paling berpengaruh terhadap klasifikasi adalah glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, dan age.

Penelitian ini membuktikan bahwa seleksi fitur menggunakan PSO mampu mengurangi dimensi data dan meningkatkan performa model klasifikasi. Perbedaan dengan penelitian ini terletak pada metode klasifikasi yang digunakan. Dalam penelitian (Susilowati et al., 2023), digunakan K-Nearest Neighbor sebagai algoritma utama untuk klasifikasi. Sedangkan dalam penelitian ini, digunakan algoritma *Naïve Bayes* yang dioptimalkan menggunakan *Particle Swarm Optimization* (PSO) untuk meningkatkan kinerja klasifikasi.

Selain itu, penelitian ini juga menerapkan teknik *Synthetic Minority Over-sampling Technique* (SMOTE) untuk menangani ketidakseimbangan data, yang tidak diterapkan dalam penelitian oleh (Susilowati et al., 2023). Dengan pendekatan yang berbeda ini, diharapkan penelitian ini dapat menghasilkan tingkat akurasi klasifikasi yang lebih tinggi dan model yang lebih andal dalam mendeteksi risiko diabetes.

Tabel 2.1 Penelitian Terkait

No	Peneliti	Metode	Input	Output	Hasil	Perbedaan
1.	(Anisa & Jumanto, 2022)	*Naïve Bayes*	Cholesterol, Glucose, Age, BMI, dll. (15→9 fitur)	Diabetes (Ya/Tidak)	Akurasi sebesar 92,3% pada data latih dan 91,6% pada data uji	Tidak menggunakan optimasi fitur atau parameter
2.	(Pradhani et al., 2025)	*Naïve Bayes*	Gender, Age, Cholesterol, Heart Rate, BMI, dll.	Diabetes (Ya/Tidak)	Hasil akurasi sebesar 91%	Menggunakan dataset yang berbeda dan juga tidak menggunakan teknik optimasi PSO
3.	(Muhammadiyah Jember & Tri Rahayu, 2022)	K-Nearest Neighbor dan Gaussian *Naïve Bayes*	15 atribut dari UCI dataset	Diabetes (Ya/Tidak)	Gaussian Naïve Bayes menghasilkan akurasi 73,50%, sensitivitas 96,14%, dan error rate 29,47%, sedangkan KNN mencatat akurasi tertinggi 71,27%, sensitivitas 76,86%, dan error rate 35,05%.	Membandingkan dua algoritma untuk klasifikasi diabetes yaitu K-Nearest Neighbor dan Gaussian *Naïve Bayes* tanpa melakukan teknik optimasi parameter menggunakan PSO
4.	(Bangun & Rachmat, 2024)	K-Nearest Neighbor dan *Naïve Bayes*	9 fitur dari dataset PIMA	Diabetes (Ya/Tidak)	Hasil akurasi model *Naïve Bayes* yaitu 74,7%, hasil akurasi model KNN yaitu 68,6%	Membandingkan performa algoritma *K-Nearest Neighbor* dan *Gaussian Naïve Bayes* dalam klasifikasi diabetes tanpa menggunakan teknik optimasi parameter berbasis PSO.
5.	(Susilowati et al., 2023)	KNN, *Particle Swarm Optimization*	20 atribut	Diabetes (Ya/Tidak)	Hasil akurasi tanpa PSO yaitu 75%, dengan PSO 77,213%	Menggunakan algoritma KNN

Berdasarkan Tabel 2.1 menyajikan terkait ringkasan dari penelitian terdahulu yang menggunakan beberapa metode dalam memprediksi penyakit diabetes. Tabel ini terdiri dari nama peneliti, dan metode penelitian, input, output, hasil, perbedaan. Penelitian ini menghadirkan kebaruan (*novelty*) yang terletak pada pengembangan pendekatan klasifikasi penyakit diabetes dengan menggabungkan model probabilistik sederhana yang telah dioptimalkan menggunakan teknik optimasi PSO serta didukung dengan teknik balancing data dan evaluasi menggunakan *10-fold cross validation.* Informasi yang terdapat dalam Tabel tersebut memberikan gambaran terkait pengaruh penggunaan metode yang digunakan dalam konteks kesehatan.

Selain metode yang digunakan, terdapat pula perbedaan pada dataset yang digunakan oleh masing-masing penelitian. Pada penelitian yang dilakukan (Anisa & Jumanto, 2022) menggunakan dataset dari *Kaggle* yang terdiri dari 390 data dan memiliki atribut berjumlah 15 yang direduksi menjadi 9 atribut. Sementara penelitian yang dilakukan oleh (Bangun & Rachmat, 2024) menggunakan dataset yang sama dengan penelitian ini, namun hanya fokus pada perbandingan dua *algoritma* tanpa adanya teknik optimasi atau balancing data. Perbedaan dengan penelitian ini yaitu menggunakan teknik SMOTE untuk mengatasi ketidakseimbangan kelas serta mengoptimalkan parameter dari algoritma *Naïve Bayes* menggunakan *Particle Swarm Optimization* (PSO).

2.2 Penyakit Diabetes

Penyakit diabetes merupakan salah satu masalah kesehatan yang semakin mengkhawatirkan di seluruh dunia. Diabetes atau dikenal sebagai diabetes mellitus,

adalah kondisi kronis yang ditandai dengan peningkatan kadar gula darah yang signifikan, disertai dengan gejala utama berupa urine yang mengandung gula melebihi batas normalnya. Kondisi ini bisa menyebabkan komplikasi seperti penyakit jantung, gangguan ginjal, dan juga kerusakan saraf (Salissa et al., 2023). Diabetes adalah penyakit tidak menular dan tidak dapat disembuhkan, tetapi dapat dikendalikan dengan melalui pengaturan pola makan, aktivitas fisik, serta pengobatan medis yang teratur. Penyakit ini dapat menyerang siapa saja, baik pria maupun wanita.

Secara umum diabetes terbagi menjadi 3 jenis, yaitu diabetes tipe 1 yaitu kondisi di mana pancreas tidak mampu memproduksi insulin sama sekali. Diabetes tipe ini biasanya terjadi pada usia anak-anak atau remaja. Kemudian diabetes tipe 2 yaitu kondisi di mana tubuh tidak dapat menggunakan insulin secara efektif atau produksi insulin tidak mencukupi, umumnya terjadi pada usia dewasa dan sering dikaitkan dengan gaya hidup tidak sehat. Diabetes gestasional yaitu jenis diabetes yang muncul setelah melahirkan, namun berisiko tinggi yang akan berkembang menjadi diabetes tipe 2 di kemudian hari (Kasandra et al., 2022).

Penyebab utama diabetes adalah kurangnya hormon insulin yang merupakan satu-satunya hormon yang berperan dalam menurunkan kadar gula dalam darah. Kekurangan hormone ini menyebabkan glukosa tidak dapat masuk ke dalam sel untuk digunakan sebagai sumber energi, sehingga menumpuk di dalam darah. Faktor-faktor yang dapat menyebabkan terjadinya diabetes antara lain adalah faktor genetik atau keturunan, obesitas, pola makan tidak seimbang (tinggi gula dan

karbohidrat sederhana), kurang aktivitas fisik, hingga penyakit infeksi tertentu (Tarigan, 2022).

Gejala awal dari diabetes mellitus sering kali tidak disadari oleh penderitanya. Gejala klasik yang umum dijumpai dikenal dengan istilah yang disebut dengan 3P, yaitu Poliuri (sering buang air kencing), polidipsi (sering merasa haus), dan polifagi (sering merasa lapar). Keluhan umum lainnya yaitu seperti penurunan berat badan secara tiba-tiba, penglihatan kabur, sering kesemutan, mudah lelah, serta terdapat luka yang sulit untuk sembuh, dan lain-lain (Ghozali et al., 2023).

2.3 Klasifikasi

Klasifikasi adalah kata serapan dari bahasa Belanda ‘Classificatie’ lalu kata ‘Classificatie’ tersebut berasal dari bahasa Prancis yakni ‘Classification’. Menurut Kamus Besar Bahasa Indonesia (KBBI) menjelaskan pengertian klasifikasi adalah penyusunan bersistem dalam kelompok atau golongan menurut kaidah atau standar yang ditetapkan. Secara istilah klasifikasi merupakan cara pengelompokan benda berdasarkan ciri-ciri yang dimiliki oleh objek klasifikasi.

Dalam prosesnya, klasifikasi dapat dilakukan dengan berbagai metode, baik secara manual maupun dengan memanfaatkan teknologi. Klasifikasi secara manual dilakukan secara langsung oleh manusia tanpa mengunakan algoritma kecerdasan buatan. Sementara itu, klasifikasi berbasis teknologi memanfaatkan berbagai algoritma, seperti *Naïve Bayes, Support Vector Machine* (SVM), *Decision Tree, Fuzzy Logic*, serta Jaringan Saraf Tiruan (Wibawa et al., 2018). *Klasifikasi*

merupakan salah satu tugas yang penting dalam data mining. Sebuah peng*klasifikasi*an dibuat berdasarkan pada pola analisis kumpulan data latih.

2.4 *Machine Learning*

Machine Learning merupakan bagian dari ilmu kecerdasan buatan, yang banyak digunakan untuk menggantikan atau menirukan perilaku manusi dalam menyelesaikan masalah atau melakukan otomatisasi. Sesuai dengan namanya yaitu *machine learning* mencoba untuk menirukan bagaiman proses manusia atau makhluk cerdas belajar dan megeneralisasi. Terdapat beberapa macam dari *machine learning* diantaranya ada *supervised learning, unsupervised learning, dan reinforcement learning*. *Machine learning* dapat digunakan untuk melakukan pendeteksian, melakukan klasifikasi, dan melakukan prediksi (Purnomo & Yuhana, 2016).

Dalam penggunaan *machine earning* adanya proses pelatihan, pembelajaran, atau training. Oleh karenanya, *Machine Learning* membutuhkan data untuk dipelajari atau bisa disebut dengan data training. Klasifikasi adalah metode dalam *machine learning* yang digunakan oleh mesin untuk memilah atau mengklasifikasikan objek berdasarkan ciri tertentu sebagaimana manusia mencoba untuk membedakan benda satu dengan benda yang lain. Sedangkan prediksi digunakan oleh mesin untuk menerka keluaran dari suatu data input berdasarkan data yang sudah dipelajari dalam proses training.

Machine Learning sering diterapkan dalam berbagai bidang salah satunya bidang kesehatan, terutama dalam mendeteksi dan memprediksi suatu penyakit. Dengan memanfaatkan algoritma pembelajaran mesin, sistem dapat menganalisis

data medis secara otomatis dan memberikan hasil yang akurat. Sebagai contoh, dalam diagnosis diabetes mellitus, *Machine Learning* dapat membantu mengidentifikasi pola dari riwayat kesehatan pasien. Teknologi ini tidak hanya mempercepat proses diagnosa akan tetapi juga membantu tenaga medis dalam mengambil keputusan yang lebih tepat berdasarkan hasil prediksi yang diberikan.

Dalam penelitian ini menerapkan salah satu dari algoritma yang terdapat dalam *supervised learning* yaitu *Naïve Bayes* yang digunakan untuk menyeleksi suatu objek dengan karakteristik tertentu untuk membedakan objek satu dengan yang lain berdasarkan dataset yang diperoleh dari dataset publik melalui situs web *Kaggle.*

2.5 Balancing Data

Balancing data merupakan proses untuk mengatasi ketidakseimbangan distribusi kelas (*class imbalance*) dalam dataset, terutama dalam konteks klasifikasi. Ketidakseimbangan data terjadi ketika jumlah sampel dari satu kelas jauh lebih banyak dibandingkan kelas lainnya. Ketidakseimbangan data dapat menyebabkan model machine learning menjadi bias terhadap kelas mayoritas, sehingga kurang mampu mengenali pola dari kelas minoritas yang justru seringkali merupakan kelas penting. Misalnya, dalam kasus prediksi penyakit, sering kali data penderita diabetes jauh lebih sedikit dibandingkan dengan data non-diabetes. Oleh karena itu, balancing data bertujuan untuk menciptakan distribusi kelas yang seimbang agar model dapat memberikan prediksi yang lebih akurat terhadap semua kelas.

Teknik yang digunakan dalam balancing data terdiri dari tiga pendekatan utama, yaitu *oversampling, undersampling*, dan metode *hybrid. Oversampling* merupakan teknik menyeimbangkan data dengan cara menambah jumlah data pada kelas minoritas, misalnya dengan menduplikasikan data atau menggunakan metode sintetik seperti SMOTE. Undersampling merupakan teknik menyeimbangkan data dengan cara mengurangi jumlah data pada kelas mayoritas, biasanya dengan enghapus sebagian data secara acak. Sementara itu, metode hybrid yaitu mengombinasikan kedua pendekatan tersebut. Dengan melakukan balancing data, model dapat menghasilkan performa yang lebih baik dalam hal recall, F1-score, akurasi dan presisi.

2.6 *Naïve Bayes*

Algoritma *Naïve Bayes* merupakan metode klasifikasi yang merepresentasikan setiap kelas objeks berdasarkan perhitungan probabilitas dan menentukan kelas yang paling mungkin sesuai untuk setiap objek yang diuji. Penentuan kelas dilakukan dengan mempertimbangkan atribut atau variabel yang telah diketahui nilainya. Algoritma ini dapat digunakan untuk memprediksi peluang suatu objek termasuk dalam kategori tertentu. Selain itu, *Naïve Bayes* sangat efektif dalam mengklasifikasikan dataset yang memiliki tipe data nominal. Keunggulan utama *Naïve Bayes* adalah mudah digunakan dan efektis dalam berbagai situasi. Teorema Bayes memiliki bentuk persamaan umun sebagai berikut:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (2.1)$$

Dengan keterangan sebagai berikut:
X adalah data dengan class yang belum dikeahui,
H adalah hipotesis data,

X merupakan suatu class spesifik,
P(H|X) adalah probabilitas hipotesis H berdasar dengan kondisi X (posterior probability,
P(H) adalah probabilitas hipotesi H (prior probability),
P(H|X) adalah Probabilitas X berdasar kondisi pada hipotesis H,
P(X) adalah Probabilitas dari X.

2.7 Gaussian *Naïve Bayes*

Gaussian Naïve Bayes merupakan salah satu varian dari algoritma *Naïve Bayes* yang dirancang untuk menangani data numerik. Algoritma ini bekerja dengan mengasumsikan bahwa setiap fitur input memiliki distribusi normal (*Gaussian*) di dalam setiap kelas target. *Gaussian Naïve Bayes* sangat sesuai untuk dataset seperti diabetes yang memiliki atribut bertipe numerik, seperti kadar glukosa, tekanan darah, dan indeks massa tubuh. Dalam konteks klasifikasi diabetes mellitus, fitur-fitur seperti kadar glukosa, tekanan darah, BMI, usia bersifat numerik atau kontinu. Oleh karena itu, Gaussian *Naïve Bayes* menjadi pilihan yang tepat dibandingkan varian lain seperti Multinomial (untuk data frekuensi) atau Bernoulli (untuk data biner).

Dengan distribusi tersebut, *Gaussian Naïve Bayes* menghitung probabilitas setiap fitur terhadap masing-masing kelas, dan mengalikannya untuk mendapatkan total probabilitas data terhadap kelas tertentu. Kelas dengan probabilitas tertinggi yang kemudian dipilih sebagai nilai hasil prediksi.

2.8 *Particle Swarm Optimization* (PSO)

Particle Swarm Optimization (PSO) adalah sebuah teknik optimasi yang terinspirasi dari pola pergerakan dan perilaku hewan, seperti ikan dan burung dalam mencari makanan atau mangsa. Metode ini pertama kali diperkenalkan oleh James Kennedy dan Russel C. Eberhart pada tahun 1995. PSO bekerja dengan

menggunakan sejumlah partikel yang bergerak untuk menemukan posisi terbaik dalam menyelesaikan masalah optimasi. Salah satu keunggulam utama dari teknik ini adalah konsepnya yang sederhana, mudah di implementasikan, serta memiliki efisiensi komputasi yang lebih tinggi dibandingkan dengan algoritma matematika dan teknik optimasi lainnya (Arifin & Ariesta, 2019).

2.9 K-Fold Cross Validation

K-fold cross validation adalah metode validasi silang yang digunakan untuk meng*Evaluasi* performa suatu algoritma dengan cara membagi dataset secara acak menjadi kelompok atau partisi. Dalam prosesnya, setiap iterasi dari *k-fold cross validation* akan menggunakan satu partisi sebagai data validasi, sementara sisanya digunakan sebagai data latih. Proses ini terus berulang hingga semua partisi telah digunakan sebagai data uji secara bergantian.

Teknik ini memungkinkan pengujian yang lebih akurat terhadap model karena setiap data memiliki kesempatan untuk menjadi data uji, sehingga mengurangi bias dan meningkatkan generalisasi model. Dalam penelitian ini, proses *k-fold cross validation* dilakukan sebanyak 10 kali (*10-fold cross validation*), dimana setiap iterasi menggunakan data latih dan data uji yang berbeda. Dengan melakukan pengujian secaraberulang pada berbagai posisi data, metode ini bertujuan untuk mendapatkan hasil yang lebih optimal. Pengujian sebanyak *10-fold cross validation* dapat dilihat pada tabel 2.1 berikut.

Tabel 2.2 10-Fold Cross Validation

K-Fold	**Cross Validation**									
1	Test	Train	Train	Train	Train	Train	Train	Train	Train	Train
2	Train	Test	Train	Train	Train	Train	Train	Train	Train	Train
3	Train	Train	Test	Train	Train	Train	Train	Train	Train	Train
4	Train	Train	Train	Test	Train	Train	Train	Train	Train	Train

K-Fold	Cross Validation									
5	Train	Train	Train	Train	Test	Train	Train	Train	Train	Train
6	Train	Train	Train	Train	Train	Test	Train	Train	Train	Train
7	Train	Train	Train	Train	Train	Train	Test	Train	Train	Train
8	Train	Train	Train	Train	Train	Train	Train	Test	Train	Train
9	Train	Train	Train	Train	Train	Train	Train	Train	Test	Train
10	Train	Train	Train	Train	Train	Train	Train	Train	Train	Test

Terlihat dari tabel 2.2 diatas, proses validasi silang menggunakan nilai k = 10, yang berarti data dibagi menjadi 10 subset (*fold*). Data kemudian dijalankan sebanyak 10 kali (*10-fold*) untuk memastikan setiap subset data mendapatkan kesempatan sebagai data latih dan data uji. Pada setiap iterasi, satu *fold* digunakan sebagai data uji, sementara 9 fold lainnya digunakan sebagai data latih. Posisi data uji berubah secara bergantian pada setiap iterasi. Misalnya, pada iterasi pertama, fold pertama digunakan sebagai data uji, sedangkan fold lainnya sebagai data latih. Pada iterasi kedua, fold kedua menjadi data uji, dan seterusnya hingga iterasi kesepuluh, dimana fold terakhir digunakan sebagi data uji.

2.10 *Confusion matrix*

Confusion matrix merupakan alat ukur berbentuk matriks yang digunakan untuk mengevaluasi kinerja suatu model klasifikasi. Matriks ini menggambarkan jumlah ketetapan *klasifikasi* terhadap kelas yang ditentukan oleh algoritma yang digunakan. *Confusion matrix* biasanya digunakan untuk menghitung metrik evaluasi seperti akurasi, presisi, dan recall, yang penting dalam menilai performa model prediksi. Akurasi adalah mengukur tingkat kesesuaian antara nilai prediksi dan nilai aktual, atau dengan kata lain, seberapa baik model dapat mengklasifikasikan data dengan benar. Sedangkan presisi atau positif predictive value adalah mengukur seberapa banyak data yang diprediksi sebagai kelas positif

benar-benar merupakan kelas positif. Sedangkan *recall* menunjukkan sejauh mana model dapat mengenali kelas positif dari keseluruhan data yang benar-benar termasuk dalam kelas tersebut.

Dalam *Confusion matrix*, setiap baris mewakili kelas sebenarnya, sedangkan setiap kolom menunjukkan kelas yang di prediksi oleh model. Elemen-elemen dalam *Confusion matrix* meliputi, True Positive (TP) yaitu data yang benar-benar positif dan di prediksi sebagai positif. True Negatif (TN) yaitu data yang benar-benar negatif dan di prediksi sebagai negatif. False Positif (FP) yaitu data yang sebenarnya negative tapi di prediksi sebagai positif. False Negatif (FN) yaitu data yang sebenarnya positif tetapi di prediksi sebagai negative (Suryadewiansyah et al., 2020).

Tabel 2.3 Contoh *Confusion matrix*

Nilai Prediksi	**Nilai Aktual**	
	Positive	***Negative***
Positive	***TP***	***FN***
Negative	***FP***	***TN***

Untuk mengevaluasi performa suatu model, dapat menggunakan beberapa metric dari *Confusion matrix* seperti akurasi, presisi, *recall*, dan *f-1 score*. Nilai-nilai tersebut memberikan gambaran bahwa semakin tinggi nilainya, maka semakin baik pula kualitas model yang dihasilkan. Agar lebih mudah dipahami, hasil evaluasi dikonversi ke dalam bentuk persentase dengan mengalikannya sebesar 100%. Akurasi menunjukkan tingkat kedekatan antara hasil prediksi dan nilai actual setelah dilakukan pengujian. Semakin tinggi nilai akurasi yang diperoleh, maka semakin baik hasil evaluasi model tersebut. Adapun akurasi, presisi, *recall*, dan *f1-score* masing-masing memiliki rumus yang dijabarkan sebagai berikut:

$$Accuracy = \frac{TP + FN}{TP + FP + TN + FN} \quad (2.2)$$

$$Precision = \frac{TP}{TP + FP} \quad (2.3)$$

$$Recall = \frac{TP}{TP+FN} \quad (2.4)$$

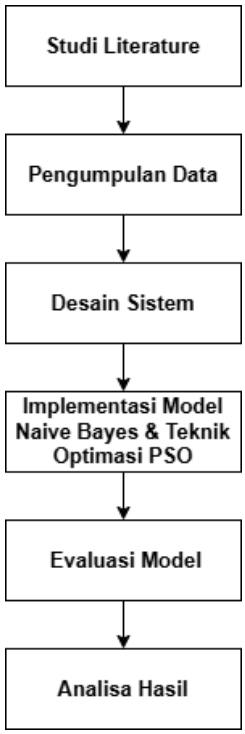
$$F\ 1\ Score = \frac{Precision + Recall}{Precision + Recall} \quad (2.5)$$

BAB III

METODE PENELITIAN

3.1 Tahapan Penelitian

Tahapan penelitian ini merupakan alur utama yang menjelaskan proses dari penelitian ini. Gambar 3.1 adalah metode dan langkah-langkah yang akan digunakan agar penulisan terstruktur dengan baik.



Gambar 2. 1 Tahapan Penelitian

3.2 Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan dataset yang bisa diakses secara publik. Dataset yang digunakan dalam penelitian ini adalah *Pima Indian Diabetes Dataset* yang dikumpulkan oleh *National Institute of Diabetes and Digestive and Kidney Diseases*. Dataset diperoleh dari situs resmi *Kaggle* yang

bisa diakses oleh siapapun. Dataset berjumlah 768 data dan terdiri dari 9 fitur. Pada penelitian ini digunakan 8 fitur dimana 1 atributnya yaitu *outcome* yang merupakan variable output. Dalam dataset Pima Indian Diabetes Dataset terdapat 268 pasien yang menderita diabetes dan 500 pasien yang tidak menderita diabetes sebelum dilakukan *cleaning* data. Data yang mengandung missing value dilakukan proses imputasi data menggunakan median. Berikut merupakan tabel penjelasan terkait fitur dataset dan keterangannya.

Tabel 3.1 Fitur Dataset

No	Fitur	Keterangan
1	Pregnancies	Jumlah kehamilan/kelahiran pada perempuan
2	Glucose	Kadar glukosa dalam darah 2 jam setelah makan
3	BloodPressure	Tekanan darah
4	SkinThickness	Ketebalan lipatan kulit
5	Insulin	Kadar insulin dalam 2 jam setelah makan
6	BMI	Tingkat ideal tubuh
7	DiabetesPedigreeFunction	Riwayat keturunan diabetes keluarga
8	Age	Umur pasien
9	Outcome	Kelas target (0 = Tidak memiliki diabetes, 1 = Memiliki diabetes)

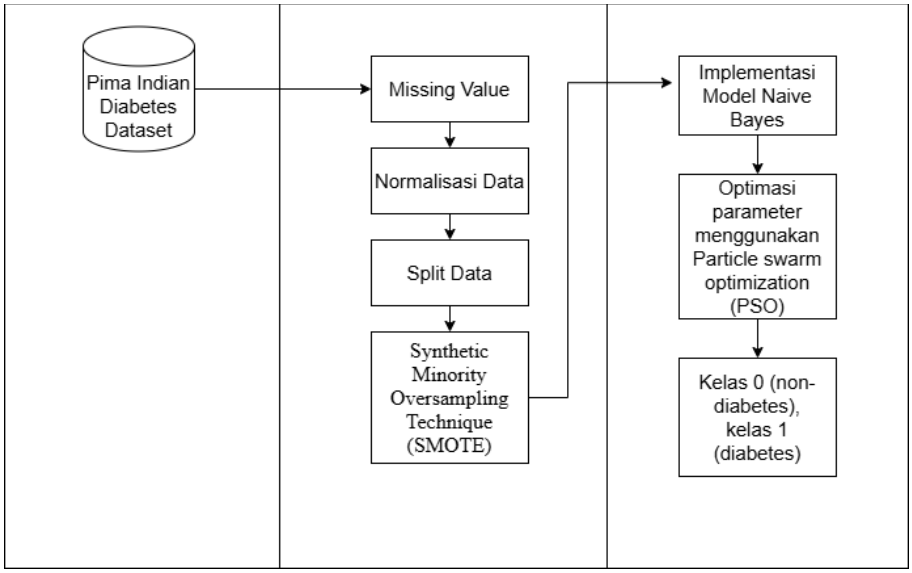
Tabel 3.2 Contoh Dataset Diabetes

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigree Function	Age	Outcome
6	148	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	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1

Dalam dataset ini, dari total 9 atribut, 8 fitur digunakan untuk model klasifikasi. Atribut yang berfungsi sebagai variabel output atau target adalah *outcome***,** yang menunjukkan diagnosis diabetes (1 untuk positif diabetes, 0 untuk negatif diabetes). Penggunaan delapan fitur ini bertujuan untuk membangun model yang akurat dalam memprediksi kemungkinan seseorang menderita diabetes berdasarkan karakteristik yang terdapat dalam dataset.

3.3 Desain Sistem

Desain sistem ini berisi tahapan-tahapan yang dilakukan mulai dari input dataset hingga menghasilkan prediksi yang menunjukkan diabetes atau tidak. Dalam penelitian ini dibuat sebuah alur desain sistem prediksi penyakit diabetes yang dapat dilihat pada Gambar 3.2.



Gambar 2.2 Desain Sistem

Penelitian ini menggunakan bahasa pemrograman *Python* dengan dataset *Pima Indian Diabetes Dataset* dari situs web *Kaggle*. Dalam tahapan *preprocessing* dimulai dengan melakukan eliminasi data, normalisasi data, split data, balancing

data. Setelah tahap *preprocessing*, Setelah itu, menerapkan model *Naïve Bayes* untuk melakukan proses prediksi data pasien dengan menggunakan 8 atribut untuk memprediksi seseorang terkena diabetes. Kemudian setelah proses prediksi, menerapkan algoritma *Particle Swarm Optimization* (PSO) untuk mengoptimasi parameter dalam klasifikasi.

3.4 Preprocessing

Preprocessing merupakan tahap penting dalam pengolahan data sebelum data tersebut digunakan untuk analisis atau pelatihan model pembelajaran mesin. Tahap ini bertujuan untuk meningkatkan kualitas data sehingga model dapat bekerja secara optimal.

3.4.1 Missing Value

Kehilangan data atau *missing value* dapat berdampak pada performa evaluasi model dalam klasifikasi. Untuk mengatasi masalah ini, diperlukan penanganan khusus agar hasil penelitian tetap optimal. Banyaknya *missing value* dalam dataset dapat disebabkan oleh berbagai fakor yang berhubungan dengan informasi dari sampel pasien, seperti ketidaksediaan pasien dalam memberikan data atau adanya kesalahan teknik saat proses pengumpulan informasi.

Dalam Pima Indian Diabetes Dataset, jumlah data yang mengandung *missing value* mencapai ratusan, dengan beberapa pasien memiliki satu hingga lima atribut tidak lengkap. Oleh karena itu, penelitian ini menerapkan teknik imputasi menggunakan nilai median untuk menangnai permasalahan *missing value*. Nilai 0 pada atribut seperti Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin,

dan BMI dianggap sebagai data yang tidak valid atau hilang. Untuk mengatasinya, nilai-nilai 0 tersebut diganti dengan nilai tengah (median) dari nilai valid pada masing-masing atribut guna menjaga kualitas data dan mencegah bias dalam proses pelatihan model. contoh data yang mengandung *missing value* ditampilkan pada tabel 3.3 sebagai berikut.

Tabel 3.3 Data Mengandung *Missing Value*

No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
…	…	…	…	…	…	…	…	…	…
764	10	101	76	48	180	32.9	0.171	63	0
765	2	122	70	27	0	36.8	0.34	27	0
766	5	121	72	23	112	26.2	0.245	30	0
767	1	126	60	0	0	30.1	0.349	47	1
768	1	93	70	31	0	30.4	0.315	23	0

3.4.2 Normalisasi Data

Data yang terdapat pada dataset biasanya memiliki nilai dengan rentang yang tidak sama. Tentunya hal ini dapat mempengaruhi hasil pengukuran analisis data, sehingga diperlukan suatu metode normalisasi data. Normalisasi data merupakan proses membuat skala nilai artibut ke dalam rentang yang lebih kecil dengan bobot yang sama. Skala nilai atribut yang baru bisa membantu kinerja *klasifikasi* karena dapat menghapus fitur noise yang tinggi dan relevansi yang rendah. Terdapat beberapa metode untuk normalisasi data seperti *Min-max Normalization*, *Z-score Normalization*, dan *Decimal scaling*. Dalam penelitian ini menggunakan metode normalisasi, sebagai berikut.

a. Min-max Normalization

Min-max Normalization suatu metode yang untuk melakukan transformasi linear dengan menggunakan nilai maksimum dan juga nilai minimum yang bisa menghasilkan nilai keseimbangan antara data satu dengan data yang lain pada rentang yang sama. Dalam mencapai konvergensi, metode ini membutuhkan waktu yang paling cepat dibandingkan dengan metode yang lain. *Min-max Normalization* dapat dihitung dengan menggunakan rumus berikut:

$$Xnew = \frac{X - \min(X)}{max\ (X) - \min(X)} \quad (3.1)$$

Dengan Keterangan:
X_new : Nilai baru dari hasil normalisasi
X : Nilai lama
Max (X) : Nilai maksimum dalam dataset
Min (X) : Nilai minimum dalam dataset

3.4.3 Split Data

Setelah tahap preprocessing dan eksplorasi data, langkah berikutnya adalah membagi dataset menjadi dua segmen yaitu data pelatihan (training set) dan data pengujian (testing set). Data pelatihan digunakan untuk membangun dan melatih model machine learning, sedangkan data pengujian digunakan untuk meng*Evaluasi* performa model pada data baru yang belum pernah diakses sebelumnya. Dalam penelitian ini, data tersegmentasi menggunakan rasio 8:2, dimana 80% dialokasikan menjadi data latih dan 20% dialokasikan menjadi data uji dan juga 9:1, dimana 90% dialokasikan menjadi data latih dan 10% dialokasikan menjadi data uji.

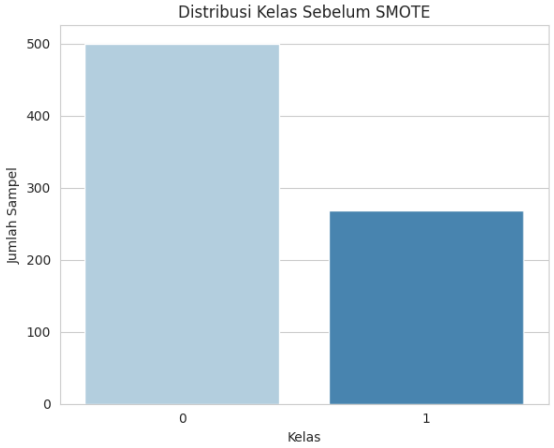
3.4.4 *Synthetic Minority Oversampling Technique* (SMOTE)

Dalam penelitian ini untuk mengatasi ketidakseimbangan data dengan

menggunakan teknik *oversampling* yaitu *Synthetic Minority Oversampling Technique* (SMOTE). SMOTE bekerja dengan membangkitkan sampel baru secara interpolative antara data miknoritas yang ada dan tetangga terdekatnya. Secara teknik, SMOTE memilih sebuah sampel dari kelas minoritas, kemudian mencari beberapa tetangga terdekatnya menggunakan metrik jarak. Selanjutnya, algortima akan membuat data baru dengan menggabungkan informasi dari sampel asli dan salah satu tetangganya secara acak pada garis lurus di antara keduanya. Proses ini membantu memperluas representasi kelas minoritas di ruang fitur tanpa menimbulkan duplikasi data. Keuntungan utama dari SMOTE adalah mengurangi risiko overfitting yang biasa terjadi pada oversampling dan meningkatkan kemampuan model dalam mengenali pola pada kelas minoritas. Proses pembuatan data sintetis dalam SMOTE secara matematis dirumuskan sebagai berikut:

$$X_{new} = X_i + \alpha.(X_{zi} - X_i) \quad (3.1)$$

Keterangan:
X_{new} adalah data sintetis baru yang dihasilkan.
X_{zi} adalah tetangga terdekat dari xi yang juga berasal dari kelas minoritas.
α adalah bilangan acak antara 0 dan 1
X_i adalah vektor fitur dari salah satu data kelas minoritas.



Gambar 2.3 Distribusi Kelas Data

Gambar 2.3 memperlihatkan bahwa persebaran kelas pada Diabetes Dataset tidak seimbang. Kelas *healthy* memiliki 500 data, sedangkan kelas *diabetic* hanya memiliki 268 data. Penggunaan teknik SMOTE dilakukan setelah tahap split data, sehingga pada masing-masing rasio pembagian data diterapkan teknik SMOTE untuk mengatasi persebaran kelas datanya.

3.5 *Gaussian Naïve Bayes*

Algoritma klasifikasi yang digunakan dalam penelitian ini adalah *Gaussian Naïve Bayes* yaitu salah satu varian dari metode *Naïve Bayes* yang digunakan untuk data numerik dengan asumsi distribusi Gaussian pada setia fitur. Algoritma ini bekerja dengan asumsi bahwa setiap fitur dalam dataset bersifat independen atau tidak saling bergantung (*naïve assumption*). Meskipun asumsi ini jarang sepenuhnya benar dalam dunia nyata, algoritma ini tetap menjadi metode yang efektif dan sering digunakan dalam berbagai kasus klasifikasi, temasuk deteksi penyakit diabetes mellitus.

Dalam kasus klasifikasi diabetes mellitus, algoritma digunakan untuk memprediksi apakah seorang pasien menderita diabetes atau tidak dengan didasarkan pada sejumlah fitur, seperti Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, dan Age. Algoritma ini menghitung probabilitas suatu kelas (diabetes atau tidak diabetes) berdasarkan fitur-fitur yang ada. *Gaussian Naïve Bayes* menghitung probabilitas suatu kelas (misalnya: diabetes atau non-diabetes) berdasarkan distribusi normal untuk masing-

masing fitur. Setiap nilai fitur pada data uji akan dievaluasi terhadap distribusi Gaussian yang dihasilkan dari data latih.

Proses langkah-langkah implementasi *Gaussian Naïve Bayes* dalam konteks klasifikasi diabetes mellitus sebagai berikut.

1. Input data uji yang akan diklasifikasi, yang berisi berbagai fitur seperti Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, dan Age.
2. Hitung probabilitas awal kelas (*Prior Probability*)

   Probabilitas awal untuk setiap kelas dihitung dari proporsi jumlah data pada masing-masing kelas dalam data latih:

$$P(C_k) = \frac{Jumlah\ data\ dengan\ kelas\ C_k}{Total\ data\ latih} \tag{3.2}$$

3. Menghitung parameter distribusi Gaussian tiap fitur

   Untuk setiap fitur dan setiap kelas, hitung nilai *mean* (μ) dan *standar deviasi* (σ). Nilai ini digunakan untuk membentuk fungsi distribusi Gaussian:

$$P(X|C) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2a}} \tag{3.3}$$

   Di mana:
   X adalah nilai fitur dalam data uji
   μ adalah rata-rata fitur dari data latih
   σ *standar deviasi* fitur dari data latih
   P(X|C) adalah probabilitas fitur X diberikan kelas C.

4. Menghitung probabilitas gabungan (*Posterior Probability*)

Probabilitas suatu kelas C_k diberikan fitur-fitur dari data uji dihitung sebagai hasil perkalian semua probabilitas fitur yang bersesuaian, kemudian dikalikan dengan prior.

$$P(C_k \mid X) \propto P(C_k) \cdot \prod_{\{i=1\}}^{n} P(x_i|C_k) \quad (3.4)$$

Dimana X = (x_1,x_2, …., x_n) adalah fitur dalam dataset.

5. Klasifikasi berdasarkan probabilitas tertinggi

   Nilai probabilitas posterior dari setiap kelas dibandingkan. Kemudian, kelas dengan nilai probabilitas tertinggi dipilih sebagai hasil prediksi:

   a) Jika *P* (Diabetes | *X*) > *P* (NonDiabetes | *X*), maka hasilnya yaitu diabetes.
   b) Jika sebaliknya, maka hasilnya yaitu non-diabetes.

3.6 *Particle Swarm Optimization (PSO)*

Particle Swarm Optimization (PSO) adalah algoritma optimasi berbasis populasi yang terinspirasi dari perilaku kawanan burung atau ikan dalam mencari makanan. Algoritma ini digunakan untuk menemukan solusi yang optimal dengan memanfaatkan interaksi antar partikel dalam suatu ruang pencarian. Dalam konteks klasifikasi penyakit diabetes mellitus menggunakan *Naïve Bayes*, PSO digunakan untuk optimasi parameter *Naïve Bayes*, seperti probabilitas prior atau parameter distribusi Gaussian.

Langkah-langkah *Particle Swarm Optimization* (PSO) dalam klasifikasi diabetes sebagai berikut.

1. Inisialisasi Partikel

Setiap partikel populasi merepresentasikan kemungkinan solusi, misalnya subset fitur yang dipilih atau parameter *Naïve Bayes* dan juga menentukan kecepatan awal dan posisi awal partikel dalam ruang pencarian.

2. Evaluasi Fungsi objektif

   Untuk setiap partikel, hitung nilai fungsi objektif menggunakan model *Naïve Bayes*. fungsi objektif yang digunakan bisa berupa akurasi klasifikasi pada dataset validasi.

3. Perbarui Posisi dan Kecepatan Partikel

   Dengan menggunakan rumus berikut untuk memperbarui kecepatan (v) dan posisi (x) setiap partikel:

$$V_i^{(t+1)} = W.V_i^{(t)} + c_1 \,.\, r_1 \left(pbest_i - X_i^{(t)}\right) + c_2 \,.\, r_2 \left(gbest_i - X_i^{(t)}\right) X_i^{(t+1)} = X_i^{(t)} + V_i^{(t+1)} \quad (3.5)$$

Di mana:
w adalah inertia weight untuk menjaga keseimbangan eksplorasi dan eksploitasi.
c1 dan c2 adalah konstanta pembelajaran untuk mengontrol pengaruh pengalaman individu dan kawanan.
r1 dan r2 adalah angka acak antara 0 dan 1.
pbesti adalah posisi terbaik partikel individu.
gbesti adalah posisi terbaik dari seluruh populasi.

4. Pengecekan Konvergensi

   Jika nilai fungsi objektif tidak mengalami perubahan signifikan atau sudah mencapai jumlah iterasi maksimum, maka proses dihentikan.

5. Gunakan Hasil Optimasi untuk Klasifikasi

   Partikel terbaik yang ditemukan digunakan sebagai konfigurasi optimal dalam model *Naïve Bayes* untuk klasifikasi diabetes.

Dengan PSO, model *Naïve Bayes* dapat menghasilkan klasifikasi yang lebih akurat karena parameter distribusi probabilitasnya telah dioptimasi dengan PSO. Integrase PSO dalam rumus *Naïve Bayes*:

Rumus utama *Naïve Bayes*:

$$P(C|X) = \frac{P(X|H)P(H)}{P(X)} \tag{3.6}$$

Dengan:

$$P(C|X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2a}} \tag{3.7}$$

PSO berperan dalam mengoptimasi parameter μ dan σ^2 yang terdapat rumus di atas. Artinya, nilai-nilai tersebut diatur oleh PSO agar menghasilkan probabilitas $P(C|X)$ yang lebih akurat sehingga meningkatkan performa klasifikasi akhir $P(C|X)$. Dengan kata lain, PSO tidak mengubah rumus *Naïve Bayes*, melainkan menyesuaikan parameter dalam rumus tersebut agar hasil probabilitas yang maksimal.

3.7 Skema Pengujian

Skenario pengujian bertujuan untuk mengevaluasi performa sistem klasifikasi diabetes mellitus yang dibangun dengan menggunakan varian dari metode *Naïve Bayes* yaitu *Gaussian Naïve Bayes* serta hasil optimasinya menggunakan teknik optimasi *Particle Swarm Optimization* (PSO). Pengujian ini dilakukan untuk mengetahui sejauh mana model dapat memprediksi kelas target (diabetes atau tidak diabetes) secara akurat.

1. Skema pembagian data

Data dibagi menggunakan dua skenario rasio yaitu 80:20 dan 90:10, di mana data latih digunakan untuk melatih model dan data uji digunakan untuk mengevaluasi performa model pada data baru. Selain itu, digunakan juga teknik validasi silang menggunakan *10-Fold Cross Validation* untuk memastikan performa model yang lebih stabil dan menghindari *overfitting*.

2. Pengujian model *baseline* (tanpa optimasi)

   Pengujian pertama dibangun dengan model *Gaussian Naïve Bayes* tanpa tuning atau optimasi parameter. Evaluasi dilakukan menggunakan data uji dan validasi silang, dengan metrik evaluasi berupa akurasi, presisi, *recall*, dan *F1-score*.

3. Pengujian model dengan PSO (*Tuning Hyperparameter*)

   Pengujian kedua dibangun dengan model *Gaussian Naïve Bayes* dengan melakukan optimasi menggunakan *Particle Swarm Optimization* (PSO) untuk menemukan nilai optimal dari parameter distribusi Gaussian. PSO dijalankan dengan:

   a) Fungsi objektif: rata-rata akurasi hasil *10-Fold Cross Validation*
   b) Parameter yang dioptimasi: nilai *mean* dan *standar deviasi* dari fitur distribusi Gaussian dalam model *Naïve Bayes*
   c) Jumlah partikel
   d) Maksimal iterasi
   e) Konstanta percepatan
   f) Inertia weight

4. Evaluasi kinerja model

   Evaluasi dilakukan dengan menggunakan:

a) *Confusion matrix* untuk melihat prediksi benar dan salah

b) Akurasi, presisi, *recall*, dan *f1-score* sebagai metrik utama

c) Perbandingan performa model sebelum dan sesudah tuning PSO

Tabel 3.4 Skenario Pengujian

Uji Coba	Metode	Rasio Data	SMOTE	Validasi	PSO	Jumlah Partikel	Maksimal Iterasi	Keterangan
1	Naïve Bayes	90:10, 80:20, 70:30, 60:40	Ya	Tanpa K-Fold	Tidak	-	-	Baseline tanpa optimasi dan validadi silang
2	Naïve Bayes	60:40	SMOTE, Tanpa SMOTE	Ya (10-Fold CV)	Tidak	-	-	Baseline tanpa optimasi dan tanpa SMOTE
3	Naïve Bayes + PSO	60:40	Ya	Ya (10-Fold CV)	Ya	30, 60, 90	100	Optimasi parameter dengan PSO
4	Naïve Bayes + PSO	60:40	Ya	Ya (10-Fold CV)	Ya	30	100, 200, 300	Uji sensivitas jumlah partikel PSO
5	Naïve Bayes + PSO	60:40	Tidak	Ya (10-Fold CV)	Ya	30, 60, 90	100, 200, 300	Analisis dampak tanpa balancing (SMOTE)
6	Naïve Bayes Standar	60:40	Tidak	Ya (10-Fold CV)	Tidak	-	-	Baseline tanpa optimasi

BAB IV

HASIL DAN PEMBAHASAN

4.1 Langkah-Langkah Pengujian

Pada bagian ini dijelaskan secara rinci langkah-langkah yang dilakukan untuk melakukan pengujian sistem klasifikasi diabetes mellitus menggunakan metode *Naïve Bayes* yang dikombinasikan dengan teknik optimasi *Particle Swarm Optimization* (PSO). Tujuan dari penelitian ini adalah untuk mengevaluasi performa metode dalam mengklasifikasi penyakit diabetes berdasarkan data riwayat kesehatan pasien. Langkah-langkah pengujian sistem mengacu pada skenario-skenario yang telah disebutkan dalan bab 3 pada poin 3.3. Adapun urutan pengujian sistem sebagai berikut:

a) Dataset yang digunakan merupakan dataset publik yang tersedia di Kaggle dengan jumlah data sebanyak 768 pasien dan terdiri 9 fitur diantaranya, Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, Outcome.
b) Sebelum model diimplementasikan, dilakukan proses pengecekan nilai yang tidak valid, khususnya nilai 0 pada atribut Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI. Nilai 0 dianggap sebagai nilai yang tidak valid dan dilakukan imputasi menggunakan nilai median dari masing-masing kolom.

c) Selanjutnya, data dinormalisasi mengunakan *Min-Max Normalization* untuk memastikan seluruh fitur berada dalam rentang yang sama, sehingga model dapat mempelajari pola antar fitur secara adil.

d) Pembagian dataset dilakukan menjadi dua bagian, yaitu data latih dan data uji. Proses pembagian ini menggunakan metode train_test_split() dengan dua skenario pembagian data yaitu 80:20 dan 70:30.

e) Karena data target pada dataset bersifat *imbalance* (tidak seimbang) yaitu lebih banyak kelas 0 dibanding kelas 1, maka dilakukan proses balancing data mengunakan teknik SMOTE (*Synthetic Minority Over-sampling Technique*) khusus pada data latih.

f) Pengujian dilakukan menggunakan dua model yaitu *Naïve Bayes* tanpa optimasi dan *Naïve Bayes* dengan optimasi PSO. Dengan melakukan proses optimasi parameter *var_smothing* serta distribusi *mean* dan *standar deviasi* dari fitur. Proses optimasi PSO dijalankan dengan konfigurasi jumlah partikel dan iterasi yang berbeda (contohnya: 40 partikel/100 iterasi dan 80 partikel/200 iterasi).

g) Untuk memastikan performa model tidak bias terhadap pembagian data tertentu, digunakan teknik *K-Fold Cross Validation* sebanyak 10 lipatan pada setiap pengujian. Hal ini bertujuan untuk menghasilkan estimasi performa yang lebih stabil.

h) Hasil evaluasi ditampilkan menggunakan metrik *accuracy*, *precision*, *recall*, *F1-score*.

4.2 Hasil Uji Coba

Pada sub bab ini, dijelaskan dengan rinci terkait analisis hasil dari pengujian berdasarkan dengan scenario pengujian yang telah disusun, dengan tujuan untuk menilai performa metode *Naïve Bayes* dengan teknik optimasi *Particle Swarm Optimization* dalam klasifikasi diabetes mellitus yang menggunakan 768 data pasien dengan 9 atribut. Dengan melalui tahapan preprocessing yaitu pengecekan missing value kemudian dilakukan imputasi data menggunakan *mean*, normalisasi data menggunakan min-max normalization, pembagian data menjadi 2 rasio (80:20 dan 90:10), kemudian dilakukan balancing data dengan menggunakan teknik SMOTE agar distribusi jumlah data seimbang. Kemudian dilakukan pengujian model menggunakan teknik silang yaitu 10-fold cross validation, lalu di evaluasi menggunakan *Confusion matrix*. Dengan demikian, evaluasi sistem yang terbangun menjadi krusial untuk memahami secara mendalam bagaimana kinerja metode *Naïve Bayes* dengan optimasi PSO dalam konteks klasifikasi diabetes.

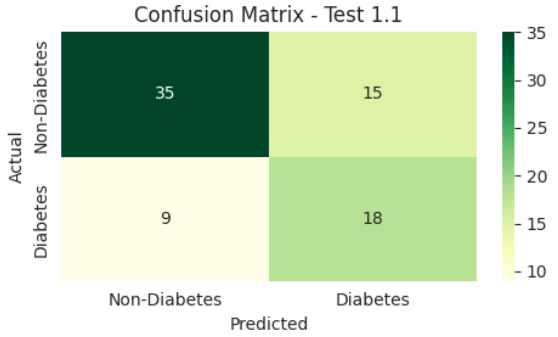
4.2.1 Uji Coba Pertama

Pengujian pertama dilakukan dengan menggunakan algoritma *Naïve Bayes* pada data yang telah dinormalisasi. Model ini diuji dengan pendekatan baseline tanpa optimasi menggunakan algoritma PSO. Validasi dilakukan menggunakan metode hold-out, dan untuk mengatasi ketidakseimbangan kelas antara data penderita Diabetes dan Non-Diabetes, digunakan teknik SMOTE (*Synthetic Minority Oversampling Technique*) pada data pelatihan. Pengujian dilakukan dengan membandingkan performa model pada empat skenario rasio data latih dan

uji, yaitu 90:10, 80:20, 70:30, dan 60:40. Rasio ini menunjukkan persentase data yang digunakan untuk pelatihan dibandingkan data untuk pengujian.

Sebelum dilakukan SMOTE, data pelatihan memiliki distribusi kelas yang tidak seimbang. Misalnya, pada rasio 80:20, data pelatihan terdiri dari 400 data Non-Diabetes dan 214 data Diabetes. Setelah dilakukan proses SMOTE, distribusi kelas menjadi seimbang, yaitu masing-masing sebanyak 400 data per kelas. Setelah pelatihan, model diuji terhadap data uji untuk masing-masing rasio. Hasil evaluasi pengujian ditampilkan dalam bentuk *Confusion matrix* dan dicatat metrik performa seperti akurasi, presisi, recall, dan f1-score. Selain itu, akurasi dari masing-masing rasio juga dikumpulkan dan divisualisasikan dalam bentuk grafik agar dapat dianalisis tren performa model.

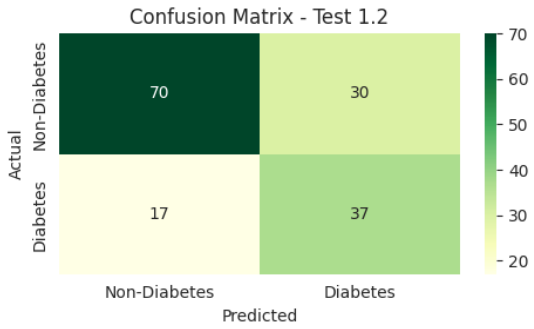
Berdasarkan hasil pengujian pertama, diperoleh *Confusion matrix* seperti ditunjukkan pada Gambar 4.1 berikut:



Gambar 4.1 Confusion Matrik (90:10)

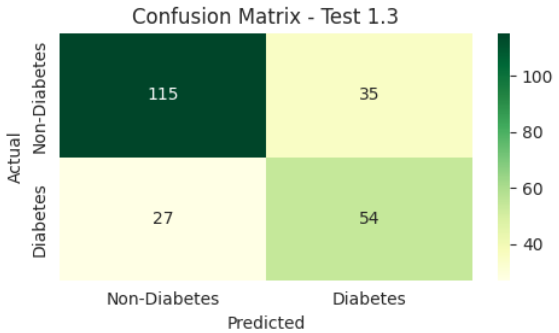
Berdasarkan Gambar 4.1 pada skenario pembagian data latih dan uji dengan rasio 90:10, menghasilkan confusion matrix dari total 77 data uji, model berhasil mengklasifikasikan 35 data non-diabetes dan 18 data diabetes dengan benar, sedangkan terjadi kesalahan klasifikasi sebanyak 15 data non-diabetes

diklasifikasikan sebagai diabetes dan 9 data diabetes diklasifikasikan sebagai non-diabetes. Hasil ini menunjukkan bahwa model memiliki performa yang cukup baik dalam mengidentifikasi kedua kelas, meskipun terdapat ketidakseimbangan dalam klasifikasi non-diabetes.



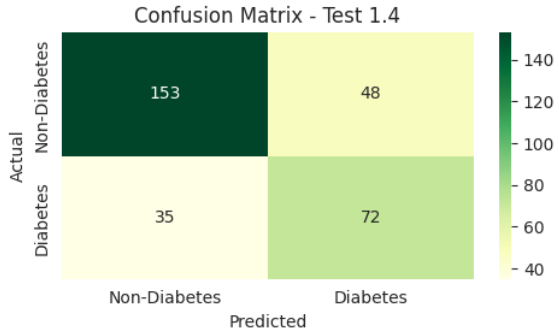
Gambar 4.2 Confusion Matrik (80:20)

Gambar 4.2 menunjukkan hasil confusion matrix pada scenario pembagian data 80:20. Pada skenario ini dari total 154 data uji, model berhasil mengklasifikasikan 70 data non-diabetes dan 37 data diabetes dengan benar, sementara 30 data non-diabetes salah diklasifikasikan sebagai diabetes dan 17 data diabetes salah diklasifikasikan sebagai non-diabetes. Meskipun jumlah data uji lebih banyak pada rasio ini, performa model tetap stabil, dengan kecenderungan kesalahan klasifikasi yang sedikit lebih tinggi dibandingkan skenario 90:10.



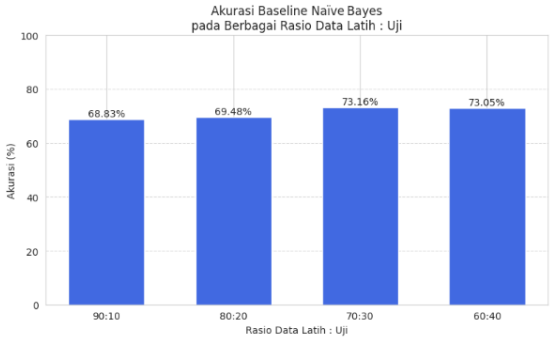
Gambar 4.3 Confusion Matrik (70:30)

Berdasarkan Gambar 4.3 pada skenario pembagian data dengan rasio 70:30, model menghasilkan 115 prediksi benar untuk kelas non-diabetes dan 54 prediksi benar untuk kelas diabetes. Namun, terdapat 35 kesalahan klasifikasi pada kelas non-diabetes yang diprediksi sebagai non-diabetes. Hasil ini menunjukkan bahwa model mampu mempertahankan akurasi yang baik meskipun proporsi data uji lebih besar.



Gambar 4.4 *Confusion matrix* (60:40)

Gambar 4.4 menunjukkan hasil confusion matrix pada scenario pembagian data 60:40. Model berhasil mengklasifikasikan 153 data non-diabetes dan 72 data diabetes dengan benar. Sementara itu, terdapat 48 data non-diabetes yang salah diklasifikasikan sebagai diabetes dan 35 data diabetes yang salah diklasifikasikan sebagai non-diabetes. Meskipun terjadi peningkatan jumlah kesalahan klasifikasi dibandingkan skenario 70:30, jumlah prediksi benar juga meningkat yang menunjukkan bahwa model tetap bekerja secara optimal meskipun dengan beban data uji yang lebih besar.



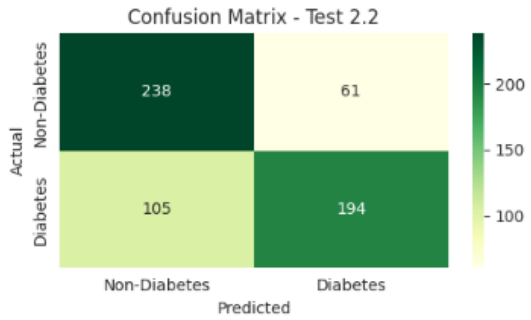
Gambar 4.5 Perbandingan Hasil Akurasi

Gambar 4.5 menunjukkan bahwa hasil akurasi model cenderung lebih tinggi pada rasio data 60:40 dan 70:30 dibandingkan rasio 90:10. Hal ini mungkin disebabkan oleh data latih yang lebih banyak memberikan informasi lebih kaya untuk model, sementara data uji yang relatif kecil dapat menyebabkan evaluasi yang kurang stabil. Sebaliknya, rasio seperti 90:10 memiliki data uji lebih sedikit, sehingga metrik evaluasi lebih rentan terhadap fluktuasi.

4.2.2 Uji Coba Kedua

Pada uji coba kedua, algoritma *Naïve Bayes* diterapkan tanpa penggunaan teknik penyeimbangan data (*balancing*), yakni tanpa penerapan SMOTE (*Synthetic Minority Oversampling Technique*). Data dibagi dengan rasio 60:40, di mana 60% data digunakan sebagai data latih, sedangkan 40% sisanya digunakan sebagai data uji. Validasi model dilakukan dengan menggunakan metode *10-Fold Cross Validation* guna mengurangi kemungkinan *overfitting* dan meningkatkan kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya. Distribusi kelas dalam data latih sebelum proses pelatihan menunjukkan kondisi yang tidak seimbang, yaitu sebanyak 350 data untuk kelas Non-Diabetes dan 187 data untuk kelas Diabetes, dengan total data latih sebanyak

537 data, sementara data uji terdiri dari 231 data. Hasil pengujian menghasilkan *confusion matrix* sebagaimana ditampilkan pada Gambar 4.6 sebagai berikut:



Gambar 4.6 *Confusion matrix* Tanpa SMOTE

Gambar 4.6 model diuji tanpa menggunakan teknik *oversampling* SMOTE. Dari hasil pengujian tersebut, model berhasil mengklasifikasikan dengan benar sebanyak 238 data non-diabetes dan 194 data diabetes. Namun masih terdapat 61 data non-diabetes yang salah diklasifikasikan sebagai diabetes, serta 105 data diabetes yang salah diklasifikasikan sebagai non-diabetes. Tingginya jumlah kesalahan pada kelas diabetes mengindikasikan adanya ketidakseimbangan dalam performa model, yang kemungkinan disebabkan oleh ketidakseimbangan jumlah data antar kelas. Hal ini menunjukkan bahwa tanpa penerapan teknik SMOTE, model cenderung lebih bias terhadap kelas mayoritas dan kurang optimal dalam mengenali kelas minoritas.

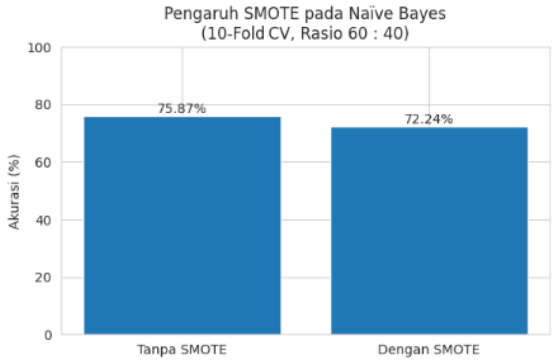
Sebagai bentuk pengembangan pengujian, dilakukan perbandingan terhadap performa model *Naïve Bayes* dengan dua konfigurasi, yakni tanpa SMOTE dan dengan SMOTE. Tujuan dari pengujian ini adalah untuk mengetahui sejauh mana pengaruh teknik *balancing* data terhadap peningkatan performa klasifikasi. Kedua konfigurasi tersebut diuji menggunakan metode validasi silang

10-Fold Cross Validation dan rasio pembagian data 60:40, dengan hasil akurasi disajikan pada Tabel 4.1 dan visualisasinya ditampilkan pada Gambar 4.7 berikut:

Tabel 4.1 Akurasi SMOTE dan Tanpa SMOTE

Konfigurasi	**Akurasi**
Tanpa SMOTE	75.87%
Dengan SMOTE	72.24%

Tabel 4.1 menyajikan perbandingan hasil akurasi algoritma Naïve Bayes dengan dan tanpa penerapan SMOTE. Dari hasil pengujian menggunakan skenario pembagian data 60:40 dan validasi silang 10-fold, terlihat bahwa akurasi model tanpa SMOTE mencapai 75.87%, sementara akurasi model dengan SMOTE sedikit lebih rendah yaitu sebesar 74.24%.



Gambar 4.7 Visualisasi Hasil Akurasi SMOTE dan Tanpa SMOTE

Visualisasi pada Gambar 4.7 menunjukkan bahwa akurasi tanpa SMOTE lebih tinggi. Hal ini menunjukkan bahwa penerapan teknik SMOTE dalam skenario ini tidak memberikan peningkatan terhadap performa model, bahkan cenderung sedikit menurunkan akurasi. Kemungkinan hal ini terjadi karena model *Naïve Bayes* memiliki sensitivitas terhadap distribusi data sintetis yang dihasilkan oleh SMOTE, sehingga tidak secara konsisten memperbaiki performa klasifikasi. Oleh

karena itu, *evaluasi* terhadap metrik lain seperti *presisi* dan *recall* juga perlu dipertimbangkan untuk menilai efektivitas SMOTE secara komprehensif.

4.2.3 Uji Coba Ketiga

Pada uji coba ketiga, algoritma Naïve Bayes dikombinasikan dengan skema optimasi *hyperparameter* menggunakan Particle Swarm Optimization (PSO). Tujuan utama dari pengujian ini adalah untuk mengevaluasi pengaruh variasi jumlah partikel, jumlah iterasi maksimum, serta kombinasi keduanya terhadap performa klasifikasi. Dataset dibagi dengan rasio 60 : 40 (60 % data latih dan 40 % data uji). Pada pengujian ini data tidak dilakukan proses balancing data. Sebagaimana tercantum pada skenario pengujian (Tabel 3.4), dilakukan analisis sensitivitas terhadap jumlah partikel (swarmsize) dengan tiga konfigurasi linear, yaitu 30, 60, dan 90 partikel sementara maksimal iterasi pada 100, 200, 300. Seluruh pengujian menggunakan validasi 10-Fold Cross Validation. Hasil akurasi tiap konfigurasi diringkas pada Tabel 4.2 dan divisualisasikan pada Gambar 4.8 sebagai berikut.

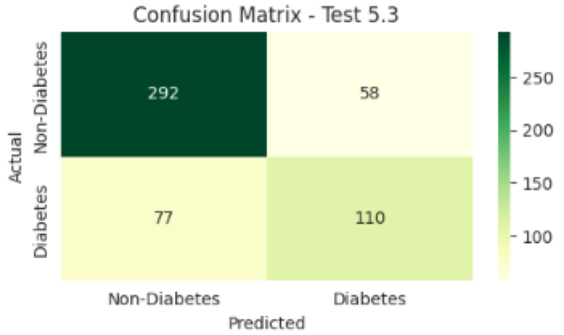
Tabel 4.2 Akurasi Naive Bayes+PSO dengan jumlah partikel

Jumlah Partikel	Maksimal Iterasi	Rasio Data	Balancing Data	Akurasi
30, 60, 90	100	60:40	Tanpa SMOTE	74,86%
30	100, 200, 300	60:40	Tanpa SMOTE	74,86%
30, 60, 90	100, 200, 300	60:40	Tanpa SMOTE	74,86%

Tabel 4.2 menunjukkan bahwa variaso jumlah partikel (30, 60, 90) dan iterasi maksimum (100, 200, 300) pada algoritma PSO tidak mempengaruhi hasil akurasi model Naïve Bayes. Semua konfigurasi menghasilkan akurasi yang sama

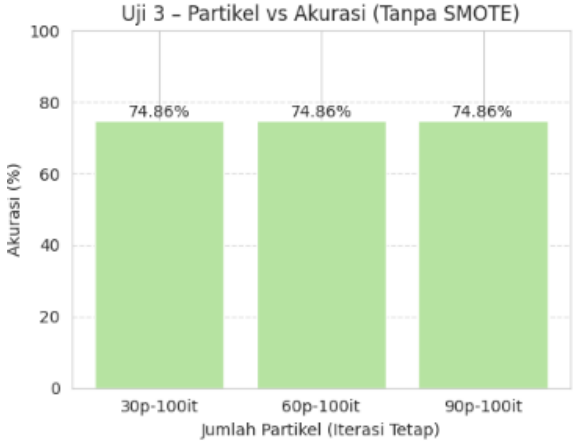
yaitu 74.86%. Hal ini menunjukkan bahwa dalam skenario ini, peningkatan jumlah partikel dan iterasi belum memberikan peningkatan performa model.

Konfigurasi dasar kombinasi (30 partikel dan 100 iterasi, 60 partikel dan 100 iterasi, 90 partikel dan 100 iterasi) menghasilkan *confusion matrix* sebagaimana ditunjukkan pada Gambar 4.8 sebagai berikut.



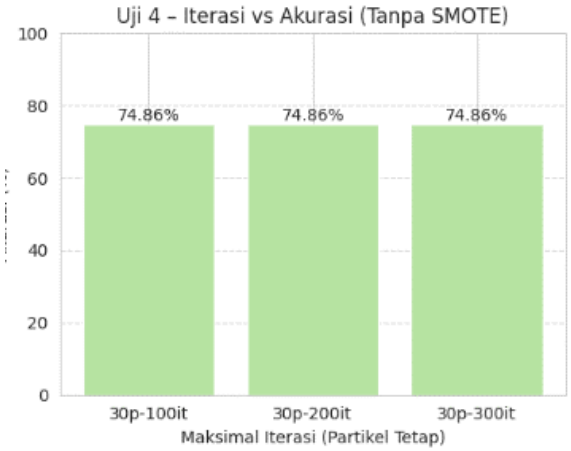
Gambar 4.8 *Confusion matrix* Uji Coba Ketiga

Gambar 4.8 menjukkan hasil confusion matrix, di mana model berhasil mengklasifikasikan 292 data non-diabetes dan 110 data diabetes dengan benar. Namun, terdapat 58 data non-diabetes yang salah diklasifikasikan sebagai diabetes, serta 77 data diabetes yang diklasifikasikan sebagai non-diabetes.



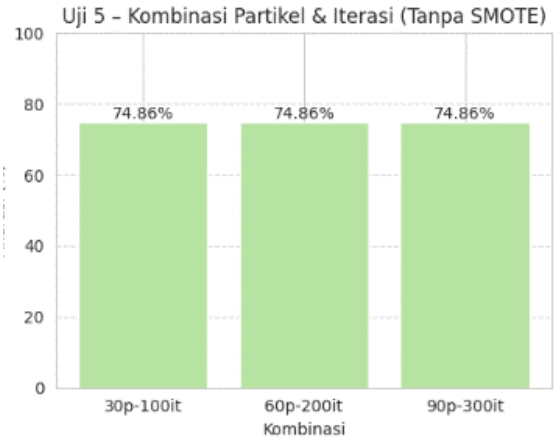
Gambar 4.9 Visualisasi Jumlah Partikel

Gambar 4.9 menunjukkan bahwa perubahan jumlah partikel dengan iterasi tetap tidak mempengaruhi akurasi model, di mana hasil akurasi menunjukkan 74.86%. Hal ini menunjukkan bahwa variasi jumlah partikel dalam scenario tidak memperngaruhi terdapat performa klasifikasi.



Gambar 4.10 Visualisasi Jumlah Iterasi

Gambar 4.10 menunjukkan hubungan antara jumlah iterasi maksimal terhadap akurasi model tanpa menggunakan SMOTE, terlihat bahwa perubahan jumlah iterasi tidak memberikan pengaruh terhadap hasil akurasi. Tiga konfigurasi iterasi yang diuji yaitu 100, 200, dan 300 iterasi dengan jumlah partikel tetap sebanyak 30 partikel menghasilkan akurasi yang sama, yaitu sebesar 74.86%. Hal ini menunjukkan bahwa dalam skenario ini, peningkatan jumlah iterasi tidak secara langsung meningkatkan performa model Naïve Bayes yang dioptimasi dengan PSO.



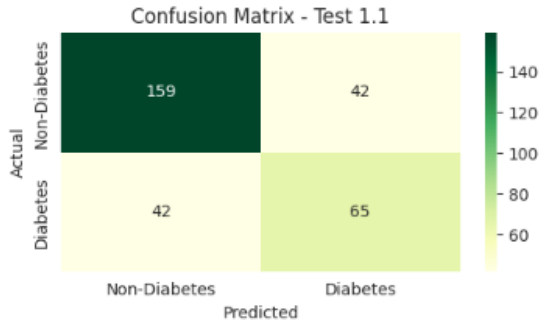
Gambar 4.11 Visualisasi Jumlah Partikel dan Maksimal Iterasi

Pada Gambar 4.11 menampilkan visualisasi pengujian dari kombinasi antara jumlah partikel dan jumlah iterasi tanpa penerapan SMOTE. Tiga kombinasi yang diuji yaitu 30 partikel dan 100 iterasi, 60 partikel dan 200 iterasi, 90 partikel dan 300 iterasi juga menghasilkan akurasi yang sama yaitu 74.8%. ini menunjukkan bahwa baik penambahan jumlah partikel maupun iterasi tidak memberikan dampat terhadap peningkatan akurasi model.

Temuan ini menunjukkan bahwa peningkatan parameter PSO, baik dalam jumlah partikel maupun iterasi, tidak memberikan pengaruh berarti terhadap akurasi model dalam kondisi data yang tidak seimbang. Hal ini mengindikasikan bahwa keterbatasan akurasi lebih disebabkan oleh distribusi data yang tidak seimbang, bukan oleh ketidaksempurnaan parameter PSO. Oleh karena itu, dapat disimpulkan bahwa untuk memperoleh peningkatan performa yang lebih signifikan, perhatian sebaiknya difokuskan pada penanganan ketidakseimbangan data atau eksplorasi metode klasifikasi yang lebih kompleks daripada sekadar meningkatkan parameter optimasi.

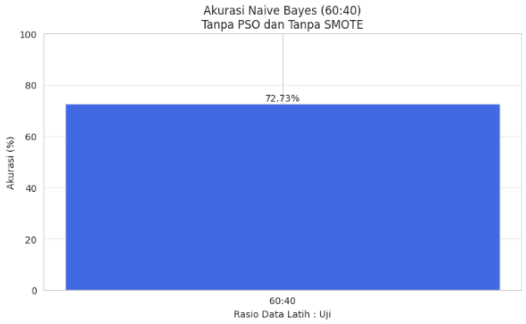
4.2.4 Hasil Uji Coba Keempat

Pada pengujian keempat ini, dilakukan evaluasi awal terhadap kinerja algoritma Naïve Bayes tanpa menggunakan metode optimasi (seperti PSO) maupun teknik penyeimbangan data (seperti SMOTE). Dataset dibagi dengan rasio 60:40, artinya 60% data digunakan untuk pelatihan dan 40% untuk pengujian, menggunakan teknik validasi hold-out. Konfigurasi ini bertujuan untuk memberikan baseline atau tolok ukur performa dasar dari model sebelum dilakukan penerapan metode optimasi lebih lanjut.



Gambar 4.12 Confusion Matrix Uji Coba Keempat

Gambar 4.12 menunjukkan hasil confusion matrix bahwa model berhasil mengklasifikasikan 159 data non-diabetes dan 65 data diabetes dengan benar, sementara 42 data non-diabetes dan 42 data diabetes salah diklasifikasikan. Hal ini mencerminkan adanya keseimbangan kesalahan antara kedua kelas.



Gambar 4.13 Visualisasi Naive Bayes Tanpa SMOTE dan PSO

Gambar 4.13 menunjukkan bahwa akurasi model Naïve Bayes tanpa penggunaan SMOTE dan PSO menghasilkan akurasi sebesar 72.73% yang menunjukkan hasil performa model tanpa penggunakan teknik optimasi.

Berdasarkan hasil pengujian, model menghasilkan akurasi sebesar 72,73%. Dari confusion matrix yang dihasilkan, untuk kelas non-diabetes diperoleh nilai presisi dan recall sebesar 79%, dengan f1-score sebesar 0,79. Sementara untuk kelas diabetes, diperoleh nilai presisi 61%, recall 61%, dan f1-score sebesar 0,61. Nilai-nilai ini menunjukkan bahwa model cukup baik dalam mengenali kelas mayoritas, namun memiliki performa yang lebih rendah dalam mengenali kelas minoritas. Ketidakseimbangan data tampak cukup berpengaruh, mengingat model lebih condong untuk mengklasifikasikan data ke kelas mayoritas. Oleh karena itu, hasil dari pengujian ini menguatkan pentingnya penggunaan teknik penyeimbangan data atau optimasi parameter untuk meningkatkan performa klasifikasi, terutama terhadap kelas minoritas. Pengujian ini berfungsi sebagai referensi dasar sebelum implementasi pendekatan yang lebih kompleks seperti PSO atau SMOTE.

4.3 Pembahasan

Penelitian ini bertujuan untuk membangun model *klasifikasi* penyakit Diabetes Mellitus menggunakan *algoritma* Naïve Bayes yang dioptimasi dengan *algoritma* Particle Swarm Optimization (PSO). *Evaluasi* dilakukan dengan membandingkan performa antara model Naïve Bayes sebelum dan sesudah optimasi menggunakan berbagai skenario pembagian data dan parameter PSO.

4.3.1 Pengaruh Rasio Data Latih dan Data Uji

Berdasarkan hasil pengujian pertama yang terdiri dari empat variasi rasio data latih dan uji, yaitu 90:10 (Uji 1.1), 80:20 (Uji 1.2), 70:30 (Uji 1.3), dan 60:40 (Uji 1.4), dapat disimpulkan bahwa penerapan algoritma *Naïve Bayes* dengan metode validasi *Hold-out* dan penyeimbangan data menggunakan SMOTE menunjukkan tren peningkatan akurasi seiring meningkatnya proporsi data uji, sampai batas tertentu.

Pada rasio 90:10 (Uji 1.1), akurasi yang diperoleh sebesar 68,83%, dengan ketepatan (*precision*) sebesar 0,80 untuk kelas *non-diabetes* dan hanya 0,55 untuk kelas *diabetes*, menunjukkan bahwa model cenderung bias terhadap kelas mayoritas. Ketika rasio diubah menjadi 80:20 (Uji 1.2), akurasi meningkat sedikit menjadi 69,48%, tetapi masih menunjukkan ketidakseimbangan dalam performa antar kelas. Pada rasio 70:30 (Uji 1.3), akurasi meningkat signifikan menjadi 73,16%, di mana performa pada kedua kelas menjadi lebih seimbang, dengan f1-score sebesar 0,79 untuk *non-diabetes* dan 0,64 untuk *diabetes*. Hal ini menunjukkan bahwa model mulai mendapatkan informasi pembelajaran yang lebih beragam dan cukup data uji untuk evaluasi yang lebih representatif. Rasio 60:40

(Uji 1.4) menghasilkan akurasi 73,05%, sedikit lebih rendah dibanding uji 1.3, tetapi masih menunjukkan stabilitas performa model.

Distribusi data latih yang semakin sedikit dari rasio 90:10 ke 60:40 menyebabkan perbedaan hasil, namun penerapan SMOTE berhasil menjaga keseimbangan antar kelas setelah pembangkitan data sintetis. Secara umum, pengujian ini menunjukkan bahwa rasio data latih dan uji yang lebih seimbang seperti 70:30 atau 60:40 mampu memberikan evaluasi yang lebih adil terhadap kinerja model, dengan akurasi yang relatif tinggi dan distribusi f1-score yang lebih merata antar kelas. Hal ini juga menandakan bahwa model Naïve Bayes cukup sensitif terhadap jumlah data latih yang tersedia, dan penerapan SMOTE sangat membantu dalam menstabilkan *performa* klasifikasi, terutama untuk mendeteksi kelas minoritas seperti diabetes.

4.3.2 Pengaruh Penggunaan SMOTE

Pada skenario pengujian kedua, dilakukan evaluasi terhadap pengaruh penerapan metode SMOTE terhadap performa algoritma *Naïve Bayes* dengan validasi *10-Fold Cross Validation* dan rasio data latih dan uji sebesar 60:40. Dua kondisi diuji: tanpa SMOTE (Uji 2.1) dan dengan SMOTE (Uji 2.2). Pada pengujian 2.1, tanpa dilakukan penyeimbangan data, model menghasilkan akurasi sebesar 75,87%, dengan f1-score sebesar 0,81 untuk kelas *non-diabetes* dan 0,62 untuk kelas *diabetes*. Nilai *recall* untuk kelas *diabetes* hanya sebesar 0,59, menunjukkan bahwa model memiliki kecenderungan untuk tidak cukup baik dalam mendeteksi kasus diabetes, yang merupakan kelas minoritas. Hal ini dapat

disebabkan oleh distribusi data yang tidak seimbang, yakni jumlah data *non-diabetes* lebih banyak dibandingkan *diabetes*.

Ketika SMOTE diterapkan pada uji 2.2, distribusi data menjadi seimbang antara kelas 0 dan 1 (masing-masing 350 data). Hasil akurasi yang diperoleh sedikit menurun menjadi 72,24%, tetapi performa antar kelas menjadi lebih seimbang. Hal ini terlihat dari meningkatnya f1-score untuk kelas *diabetes* menjadi 0,73 dan *recall*-nya meningkat menjadi 0,70, menandakan bahwa model lebih sensitif terhadap mendeteksi kasus diabetes. Meskipun akurasi secara keseluruhan tidak meningkat signifikan, penerapan SMOTE terbukti membantu dalam mengurangi bias terhadap kelas mayoritas dan meningkatkan kemampuan model dalam mengidentifikasi kelas minoritas. Oleh karena itu, penggunaan SMOTE dalam konteks klasifikasi diabetes dengan *Naïve Bayes* dapat dipandang efektif untuk memperbaiki ketimpangan klasifikasi antar kelas, khususnya dalam konteks data yang tidak seimbang.

4.3.3 Pengaruh Optimasi PSO Terhadap Performa Model

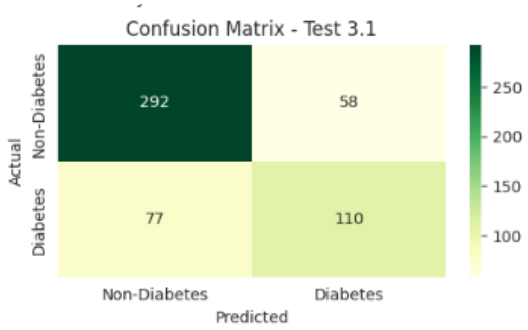
Pada pengujian ketiga hingga kelima, fokus utama adalah mengevaluasi efektivitas PSO (Particle Swarm Optimization) dalam meningkatkan performa model Naïve Bayes. Seluruh skenario menggunakan rasio data latih dan uji 60:40 serta telah dilakukan penyeimbangan data dengan SMOTE, sehingga PSO dijalankan dalam kondisi data yang sudah siap.

Pengujian ke-3 mencoba mengetahui pengaruh jumlah partikel terhadap hasil optimasi, dengan iterasi tetap 100. Partikel yang digunakan adalah 30, 60, dan 90. Pengujian ke-4 fokus pada pengaruh jumlah iterasi, dengan partikel tetap 30, dan

iterasi dinaikkan dari 100 ke 200 hingga 300. Pengujian ke-5 menggabungkan peningkatan jumlah partikel dan iterasi sekaligus, dari konfigurasi 30 partikel-100 iterasi hingga 90 partikel-300 iterasi. Namun, hasil dari semua pengujian ini menunjukkan akurasi yang identik, yaitu 74,86%. Tidak ada perubahan pada confusion matrix, nilai precision, recall, maupun f1-score. Ini menandakan bahwa proses optimasi PSO sudah mencapai konvergensi sangat awal bahkan pada parameter paling dasar (30 partikel, 100 iterasi). Dengan kata lain, menambah jumlah partikel maupun iterasi tidak membuat hasil klasifikasi menjadi lebih baik.

Dari sini dapat disimpulkan bahwa dalam konteks data yang telah diseimbangkan, penggunaan PSO tetap efektif, tetapi peningkatan parameternya tidak selalu berdampak langsung terhadap akurasi model. Justru, penggunaan konfigurasi minimum pada PSO sudah cukup untuk mendapatkan performa terbaik, tanpa perlu menambah beban komputasi dari parameter yang lebih besar.

Seluruh pengujian juga dilengkapi dengan confusion matrix untuk mengevaluasi kinerja klasifikasi pada masing-masing kelas (diabetes dan non-diabetes). Confusion matrix berfungsi untuk menunjukkan rincian hasil prediksi model, dan terdiri dari empat elemen utama: True Positive (TP), False Negative (FN), False Positive (FP), dan True Negative (TN). Misalnya, pada salah satu konfigurasi dengan 30 partikel dan 100 iterasi, confusion matrix yang dihasilkan dapat dilihat pada Gambar 4.13 sebagai berikut.



Gambar 4.14 Penjelasan Confusion Matrix

Berdasarkan Gambar 4.14, angka tersebut menunjukkan bahwa dari total 537 data uji, sebanyak 292 data non-diabetes berhasil diprediksi dengan benar (True Negative). Sebanyak 110 data diabetes juga berhasil diklasifikasikan dengan benar (True Positive). Sebanyak 58 data non-diabetes salahh diklasifikasikan sebagai diabetes (False Positive). Dan sebanyak 77 data diabetes salah diklasifikasin sebagai non-diabetes (False Negative).

Berdasarkan confusion matrix tersebut, diperoleh akurasi sebesar 74,86%, dengan f1-score kelas non-diabetes sebesar 0,81, dan f1-score kelas diabetes sebesar 0,62. Hasil ini mengindikasikan bahwa model memiliki kecenderungan untuk lebih akurat dalam mengenali kelas mayoritas, namun performanya menurun dalam mendeteksi kelas minoritas, terutama ketika teknik balancing data seperti SMOTE tidak digunakan. Oleh karena itu, evaluasi confusion matrix sangat penting dalam menilai keseimbangan performa klasifikasi antar kelas, serta sebagai dasar dalam menganalisis apakah peningkatan akurasi mencerminkan peningkatan yang adil pada kedua kelas atau hanya disebabkan oleh bias terhadap kelas mayoritas.

Dalam beberapa skenario pengujian, validasi model dilakukan menggunakan metode 10-Fold Cross Validation untuk memastikan bahwa model

Naïve Bayes yang dibangun memiliki kemampuan generalisasi yang baik terhadap data baru. Teknik ini membagi data latih menjadi 10 bagian (fold) yang kurang lebih sama besar. Dalam setiap iterasi, 9 fold digunakan untuk melatih model dan 1 fold sisanya digunakan untuk validasi. Proses ini diulang sebanyak 10 kali, sehingga setiap fold berperan sebagai data validasi satu kali. Hasil evaluasi seperti akurasi, precision, *recall*, dan *f1-score* dihitung berdasarkan prediksi gabungan dari seluruh iterasi validasi tersebut. Evaluasi ini dilakukan pada data latih, sedangkan data uji tidak dilibatkan dalam proses K-Fold.

Dengan demikian, metode ini memungkinkan pengujian performa model secara menyeluruh tanpa mengorbankan sebagian besar data hanya untuk validasi. Penggunaan K-Fold terbukti efektif karena memberikan hasil evaluasi yang lebih stabil dan bebas dari fluktuasi yang biasa terjadi pada validasi hold-out. Hasil akurasi sebesar 74,86% yang diperoleh pada pengujian ketiga merupakan hasil dari validasi silang menggunakan 10-Fold Cross Validation terhadap data latih sebanyak 537 data. Dengan skema validasi ini, setiap data diuji secara adil, sehingga memberikan gambaran lebih representatif terhadap performa sebenarnya dari model yang dibangun.

Tabel 4.3 Hasil Skenario Pengujian

Pengujian	Pembagian Data	SMOTE	Validasi	Jumlah Partikel	Maksimal Iterasi	Akurasi
Naïve Bayes Standar	90:10, 80:20, 70:30, 60:40	Ya	Tanpa K-Fold	-	-	68,83% (90:10), 69,48% (80:20), 73,16% (70:30), 73,05 (60:40)
Naïve Bayes Standar	60:40	Ya Tidak	Ya (10-Fold CV)	-	-	75,87% (Tanpa SMOTE), 72,24% (Dengan SMOTE)
Naïve Bayes + PSO	60:40	Tidak	Ya (10-Fold CV)	30, 60, 90	100	74,86% (30 partikel), 74,86% (60 partikel), 74,86% (90 partikel)
Naïve Bayes + PSO	60:40	Tidak	Ya (10-Fold CV)	30	100, 200, 300	74,86% (100 iterasi), 74,86% (200 partikel), 74,86% (300 partikel)
Naïve Bayes + PSO	60:40	Tidak	Ya (10-Fold CV)	30, 60, 90	100, 200, 300	74,86% (100 iterasi), 74,86% (200 iterasi), 74,86% (300 iterasi)
Naïve Bayes Standar	60:40	Tidak	Ya (10-Fold CV)	-	-	72,73%

Tabel 4.3 menyajikan hasil pengujian dari beberapa skenario berbeda untuk mengevaluasi performa algoritma *Naïve Bayes,* baik dalam bentuk standar maupun versi yang telah dioptimasi menggunakan *Particle Swarm Optimization* (PSO). Pengujian dilakukan dengan berbagai rasio pembagian data, yakni 90:10, 80:20, 70:30, dan 60:40, serta melibatkan validasi menggunakan metode hold-out dan 10-Fold Cross Validation (CV). Selain itu, teknik SMOTE juga diterapkan pada beberapa skenario untuk menangani masalah ketidakseimbangan data.

Pada pengujian awal menggunakan *Naïve Bayes* standar tanpa validasi silang dan dengan SMOTE, diperoleh akurasi yang bervariasi tergantung rasio data. Akurasi terendah terjadi pada rasio 90:10 sebesar 63,83%, sementara akurasi tertinggi diperoleh pada rasio 70:30 sebesar 73,16%. Rata-rata akurasi di empat rasio tersebut berkisar antara 69–73%, yang menunjukkan adanya fluktuasi performa akibat perbedaan jumlah data latih dan uji. Ketika validasi silang diterapkan dengan rasio 60:40, akurasi meningkat secara signifikan. Pada skenario ini, penggunaan *Naïve Bayes* standar tanpa SMOTE menghasilkan akurasi sebesar 75,87%, sedangkan dengan SMOTE justru sedikit menurun menjadi 72,24%. Hal ini menunjukkan bahwa pada dataset tertentu, penerapan SMOTE tidak selalu memberikan dampak positif terhadap akurasi model.

Selanjutnya, dilakukan pengujian dengan menerapkan algoritma PSO untuk mengoptimasi parameter model Naïve Bayes. Pada pembagian data 60:40 tanpa SMOTE dan dengan validasi 10-Fold CV, dilakukan variasi jumlah partikel (30, 60, 90) dan iterasi maksimum (100). Hasilnya, seluruh kombinasi tersebut memberikan akurasi yang seragam sebesar 74,86%. Ini mengindikasikan bahwa

penambahan jumlah partikel dalam konfigurasi tersebut tidak memberikan pengaruh signifikan terhadap akurasi. Uji sensitivitas terhadap jumlah iterasi juga dilakukan dengan konfigurasi partikel tetap (30) dan iterasi bervariasi (100, 200, 300). Akurasi yang diperoleh juga konstan pada angka 74,86%. Hal ini memperkuat kesimpulan bahwa proses optimasi telah mencapai titik konvergensi, sehingga penambahan iterasi tidak lagi berpengaruh secara signifikan. Terakhir, pada pengujian Naïve Bayes standar tanpa SMOTE dan menggunakan validasi silang 10-Fold, akurasi yang diperoleh adalah sebesar 72,12%. Hal ini menunjukkan bahwa meskipun PSO tidak diterapkan, model masih dapat memberikan hasil yang kompetitif dengan konfigurasi tertentu.

Secara keseluruhan, hasil pengujian menunjukkan bahwa penggunaan PSO dapat membantu menstabilkan dan meningkatkan akurasi model *Naïve Bayes*. Namun, variasi jumlah partikel dan iterasi maksimum tidak selalu memberikan dampak signifikan terhadap performa. Di sisi lain, penerapan SMOTE juga perlu dievaluasi secara kontekstual karena dalam beberapa kasus justru menyebabkan penurunan akurasi. Akurasi tertinggi sebesar 75,87% diperoleh pada *Naïve Bayes* standar tanpa SMOTE dengan validasi 10-Fold CV, menegaskan bahwa pemilihan strategi balancing dan optimasi yang tepat sangat krusial dalam membangun model klasifikasi yang efektif untuk deteksi dini diabetes mellitus.

4.4 Integrasi Islam

Ilmu pengetahuan dalam Islam tidak hanya dipandang sebagai sarana untuk memahami fenomena alam, tetapi juga sebagai bentuk ibadah apabila digunakan untuk kemaslahatan umat manusia. Dalam konteks ini, penelitian mengenai

penerapan metode *Naïve Bayes* yang dioptimasi menggunakan *Particle Swarm Optimization* (PSO) untuk klasifikasi diabetes mellitus merupakan bagian dari kontribusi ilmiah yang sejalan dengan nilai-nilai Islam. Upaya ini menunjukkan komitmen untuk memanfaatkan teknologi dalam meningkatkan kualitas hidup, khususnya di bidang kesehatan.

Diabetes mellitus merupakan penyakit metabolik kronis yang ditandai dengan kadar gula darah tinggi. Penyakit ini dapat menimbulkan komplikasi serius jika tidak segera ditangani, seperti kerusakan saraf, gagal ginjal, dan penyakit jantung. Islam memandang menjaga kesehatan sebagai bentuk tanggung jawab manusia atas amanah tubuh yang diberikan oleh Allah SWT. Upaya preventif dan deteksi dini terhadap penyakit, termasuk melalui sistem klasifikasi berbasis kecerdasan buatan, merupakan bentuk nyata dari ikhtiar menjaga keselamatan jiwa. Semangat kolaborasi dalam menjaga kesehatan ini juga sejalan dengan firman Allah dalam Surah Al-Mā'idah ayat 2:

وَتَعَاوَنُوْا عَلَى الْبِرِّ وَالتَّقْوٰىۖ وَلَا تَعَاوَنُوْا عَلَى الْاِثْمِ وَالْعُدْوَانِۚ وَاتَّقُوا اللّٰهَۗ اِنَّ اللّٰهَ شَدِيْدُ الْعِقَابِ ٢

"*Tolong-menolonglah kamu dalam (mengerjakan) kebajikan dan takwa, dan jangan tolong-menolong dalam berbuat dosa dan permusuhan. Bertakwalah kepada Allah, sesungguhnya Allah sangat berat siksaan-Nya."*(*QS. Al-Maidah:2*).

Menurut tafsir yang disusun oleh tim penyempurnaan Tafsir Kementerian Agama (Kemenag), 2011), menjelaskan tentang perintah "tolong-menolonglah kamu dalam kebajikan dan takwa" mengandung anjuran agar setiap individu muslim terlibat aktif dalam kolaborasi yang berlandaskan nilai kebaikan (*al-birr*) dan ketakwaan (*at-taqwa*). Al-birr dalam konteks ini tidak hanya terbatas pada amal

ritual seperti salat atau zakat, tetapi juga mencakup segala bentuk perbuatan yang memberikan manfaat sosial termasuk dalam bidang pendidikan, kesehatan, dan teknologi. Sementara itu, at-taqwa dimaknai sebagai sikap menjaga diri dari segala bentuk pelanggaran terhadap ketentuan Allah SWT, baik dalam tatanan individu maupun sistemik.

Sebaliknya, larangan untuk saling membantu dalam dosa (*al-itsm*) dan permusuhan (*al-'udwan*) merupakan bentuk pengingat agar kerja sama tidak dilakukan dalam konteks yang menyimpang dari nilai keadilan dan kebenaran. Ini meliputi kolaborasi dalam perbuatan zalim, manipulatif, atau yang merugikan pihak lain, baik secara langsung maupun tidak langsung. Ayat ini menuntun umat Islam untuk memiliki standar etik dalam bersinergi, yakni bahwa kerja sama hanya dibenarkan bila tujuannya membawa maslahat dan tidak menimbulkan kerusakan.

Dalam Penelitian ini, yang mengembangkan sistem klasifikasi Diabetes Mellitus dengan menggunakan metode *Naïve Bayes* berbasis PSO, merupakan aplikasi nyata dari prinsip ini. Teknologi ini memfasilitasi deteksi dini dan mempercepat interaksi antara pasien, tenaga medis, dan masyarakat luas, sehingga meningkatkan kualitas pelayanan kesehatan sebagai bentuk "*ta'awun*" dalam kebaikan. Selain itu, implementasi teknologi ini harus dilandasi nilai-nilai integritas, keadilan, dan transparansi, menghindarkan diri dari potensi penyalahgunaan data atau diskriminasi sejalan dengan larangan ayat untuk tidak menolong dalam hal yang merusak nilai dan martabat manusia. Dengan demikian, penelitian ini bukan hanya inovasi teknis, tetapi juga representasi spiritual dan etis

yang mendalam dari nilai Islam, yakni menolong dan menjaga sesama dalam bingkai ketakwaan.

Ayat ini juga bersifat universal, berlaku untuk semua orang tanpa memandang status sosial maupun kedudukan. Dalam konteks penelitian ini, pengembangan sistem klasifikasi untuk deteksi diabetes dapat dilihat sebagai bentuk *ta'āwun* (tolong-menolong) antara ilmu pengetahuan, teknologi, dan kepedulian sosial. Misalnya, seorang peneliti membantu masyarakat melalui solusi berbasis data dan kecerdasan buatan untuk mempercepat diagnosis, sementara tenaga medis memanfaatkannya untuk mendukung proses pengambilan keputusan dalam pelayanan kesehatan. Semua ini merupakan implementasi dari perintah Allah untuk saling membantu dalam kebaikan dan takwa.

Dengan demikian, penerapan teknologi dalam penelitian ini tidak hanya mencerminkan ikhtiar ilmiah, tetapi juga menjadi bentuk pengamalan nilai-nilai Islam dalam ranah sosial dan kemanusiaan, yang didasarkan pada prinsip kolaborasi, tanggung jawab, dan kemanfaatan.

BAB V

KESIMPULAN

5.1 Kesimpulan

Penelitian ini membahas penerapan algoritma *Naïve Bayes* yang dioptimasi menggunakan metode *Particle Swarm Optimization* (PSO) dalam membangun model klasifikasi untuk mendeteksi penyakit diabetes mellitus berdasarkan data medis. Permasalahan utama yang diangkat adalah bagaimana meningkatkan akurasi klasifikasi pada data yang tidak seimbang (*imbalanced*), serta bagaimana memanfaatkan teknik optimasi dan data preprocessing secara tepat untuk menghasilkan model yang efektif dan efisien.

Proses penelitian mencakup beberapa tahapan penting, seperti pengolahan missing value, normalisasi data, pembagian data dengan rasio tertentu, serta penerapan teknik SMOTE untuk mengatasi ketidakseimbangan kelas. Algoritma *Naïve Bayes* digunakan sebagai dasar model klasifikasi, sedangkan metode PSO diterapkan untuk mengoptimalkan parameter distribusi Gaussian yang digunakan oleh model tersebut. Evaluasi dilakukan menggunakan metrik accuracy, precision, recall, dan F1-score, serta validasi silang 10-Fold Cross Validation guna menguji stabilitas model.

Hasil pengujian menunjukkan bahwa model *Naïve Bayes* yang telah dioptimasi dengan PSO menunjukkan peningkatan performa klasifikasi, dengan peningkatan akurasi dari 72,73% menjadi 74,86%. Selain itu, penggunaan SMOTE dalam skenario tertentu memberikan pengaruh terhadap distribusi data latih,

meskipun tidak selalu meningkatkan akurasi secara signifikan. Pengujian dengan berbagai skenario menunjukkan bahwa kombinasi antara *balancing* data, optimasi parameter, dan validasi silang mampu menghasilkan model yang lebih akurat dan andal.

Secara keseluruhan, penelitian ini telah membuktikan bahwa integrasi antara *Naïve Bayes* dan *Particle Swarm Optimization* mampu meningkatkan kinerja *klasifikasi* data medis, khususnya dalam konteks diagnosis dini penyakit diabetes. Penelitian ini memberikan kontribusi terhadap pengembangan sistem klasifikasi berbasis kecerdasan buatan di bidang kesehatan dan dapat dijadikan referensi untuk penelitian lanjutan dalam pengembangan sistem prediktif berbasis machine learning lainnya.

5.2 Saran

Untuk pengembangan lebih lanjut pada penelitian ini, beberapa saran yang dapat diberikan antara lain.

1. Penelitian ini hanya menggunakan satu metode optimasi yaitu *Particle Swarm Optimization* (PSO). Untuk penelitian selanjutnya, disarankan agar dilakukan perbandingan dengan *algoritma* optimasi lainnya seperti Genetic *Algorithm* (GA), *Simulated Annealing*, atau *Grid Search* guna mengetahui *algoritma* mana yang paling optimal dalam meningkatkan performa klasifikasi.
2. Dataset yang digunakan masih terbatas dari sisi jumlah data dan keberagaman fitur. Oleh karena itu, disarankan untuk menggunakan dataset yang lebih besar dan lebih bervariasi agar model yang dihasilkan lebih general dan memiliki kemampuan prediksi yang lebih tinggi.

3. Sistem klasifikasi ini masih berbentuk prototipe dalam lingkungan pengujian. Untuk implementasi yang lebih luas, disarankan agar hasil penelitian ini dapat dikembangkan menjadi sebuah aplikasi berbasis web atau mobile yang dapat digunakan oleh tenaga medis maupun masyarakat umum untuk membantu diagnosis awal diabetes mellitus.

DAFTAR PUSTAKA

Abdulhadi, N., & Al-Mousa, A. (2021). Diabetes Detection Using Machine Learning Classification Methods. *2021 International Conference on Information Technology, ICIT 2021 - Proceedings, July*, 350–354. https://doi.org/10.1109/ICIT52682.2021.9491788

Anisa, D. N., & Jumanto, J. (2022). *Klasifikasi* Penyakit Diabetes Menggunakan *Algoritma* Naive Bayes. *Jurnal Dinamika Informatika*, *14*(1), 33–42. https://doi.org/10.35315/informatika.v14i1.9135

Aprillia, N., Hadi, Z., & Ishak, N. I. (2022). Hubungan Aktivitas Fisik Dan Obesitas Terhadap Kejadian DiabetesMelitus Di Wilayah Kerja Puskesmas Birayang Kabupaten HuluSungai Tengah Tahun 2022. *EPrints Uniska*, 1–9. http://eprints.uniska-bjm.ac.id/12139/

Arifin, T., & Ariesta, D. (2019). Prediksi Penyakit Ginjal Kronis Menggunakan *Algoritma* Naive Bayes Classifier Berbasis Particle Swarm Optimization. *Jurnal Tekno Insentif*, *13*(1), 26–30. https://doi.org/10.36787/jti.v13i1.97

Bangun, A., & Rachmat, E. (2024). *Analisis Perbandingan Algoritma KNN dan Naïve Bayes dalam Mendiagnosis Penyakit Diabetes Mellitus Pendahuluan*. *23*(September), 387–396.

Diana Dewi, D., Qisthi, N., Lestari, S. S. S., & Putri, Z. H. S. (2023). Perbandingan Metode Neural Network Dan Support Vector Machine Dalam *Klasifikasi* Diagnosa Penyakit Diabetes. *Cerdika: Jurnal Ilmiah Indonesia*, *3*(09), 828–839. https://doi.org/10.59141/cerdika.v3i09.662

Ghozali, A., Pratiwi, H., & Handajani, S. S. (2023). Implementasi Data Mining Menggunakan Metode Random Forest Dan Support Vector Machine Dalam *Klasifikasi* Penyakit Diabetes. *Delta: Jurnal Ilmiah Pendidikan Matematika*, *11*(2), 147. https://doi.org/10.31941/delta.v11i2.2686

Ginting, R. G., Girsang, E., Ginting, J. B., & Hartono, H. (2022). Analisis Determinan Dan Prediksi Penyakit Diabetes Melitus Tipe 2 Menggunakan Metode Machine Learning: Scoping Review. *Jurnal Maternitas Kebidanan*, *7*(1), 58–72. https://doi.org/10.34012/jumkep.v7i1.2538

Gurning, U. R., Octavia, S. F., Andriyani, D. R., Nurainun, N., & Permana, I. (2024). Prediksi Risiko Stunting pada Keluarga Menggunakan Naïve Bayes Classifier dan Chi-Square. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(1), 172–180. https://doi.org/10.57152/malcom.v4i1.1074

Kasandra, T. A., Kurniasih, E., & Ekayamti, E. (2022). *Media Publikasi Penelitian ; 2022 ; Volume 9 ; No 1 Website : http://jurnal.akperngawi.ac.id Hubungan Dukungan Keluarga Terhadap Kualitas Hidup pada Penderita*

Diabetes Melitus di Dusun Cung Belud Kecamatan Paron Kabupaten Ngawi The Relationship of Family. *9*(1), 74–82.

Kemenag), D. A. R. (atau: T. P. T. A.-Q. (2011). *Al-Qur'an dan Tafsirnya (Edisi yang Disempurnakan), Jilid II*. http://scioteca.caf.com/bitstream/handle/123456789/1091/RED2017-Eng-8ene.pdf?sequence=12&isAllowed=y%0Ahttp://dx.doi.org/10.1016/j.regsciurbeco.2008.06.005%0Ahttps://www.researchgate.net/publication/305320484_SISTEM_PEMBETUNGAN_TERPUSAT_STRATEGI_MELESTARI

Maulidah, N., Abdilah, A., Nurlelah, E., Gata, W., Nur Hasan, F., Ilmu Komputer, J., & Nusa Mandiri Jalan Margonda Raya No, S. (2020). Seleksi Fitur *Klasifikasi* Penyakit Diabetes Menggunakan Particle Swarm Optimization (PSO) Pada *Algoritma* Naive Bayes. *Daerah Khusus Ibukota Jakarta*, *13*(2), 21231170. http://journal.stekom.ac.id/index.php/elkom■page40

Muhammadiyah Jember, U., & Tri Rahayu, P. (2022). Perbandingan *Algoritma* K-Nearest Neighbor Dan Gaussian Naïve Bayes Pada Klsifikai Penyakit Diabetes Melitus Comparison Of K-Nears Neighbor And Gaussian Naïve Bayes Algorithm On The Classification Of Diabetes Mellitus. *Jurnal Smart Teknologi*, *3*(4), 2774–1702. http://jurnal.unmuhjember.ac.id/index.php/JST

Mutiara, E.-. (2020). *Algoritma Klasifikasi* Naive Bayes Berbasis Particle Swarm Optimization Untuk Prediksi Penyakit Tuberculosis (Tb). *Swabumi*, *8*(1), 46–58. https://doi.org/10.31294/swabumi.v8i1.7668

Pradhani, P. C., Indrayani, A. S., Azzahra, N., Aflikha, E., Zahra, F., & Kalifia, A. D. (2025). *342 | P a g e*. *3*, 342–353.

Purnomo, M. H., & Yuhana, U. L. (2016). Implementasi IOT dan Machine Learning Dalam Bidang Pendidikan Pembelajaran Matematika Tingkat SD melalui Serious Game. *National Conference of Applied Sciences, Engineering, Business and Information Technology*, 250–257.

Salissa, I., N, W. T., & Ningsih, W. T. (2023). Gambaran Pola Makan , Pola Istirahat , Pola Aktivitas Dan. *Jurnal Multidisiplin Indonesia*, *2*(September), 2435–2444.

Suryadewiansyah, M. K., Endra, T., & Tju, E. (2020). Naïve Bayes dan *Confusion matrix* untuk Efisiensi Analisa Intrusion Detection System Alert. *Jurnal Nasional Teknologi Dan Sistem Informasi*, *8*(2), 81–88.

Susilowati, D., Sutrisno, S., & Yunus, M. (2023). Penerapan Particle Swarm Optimization Untuk Meningkatkan Kinerja *Algoritma* K-Nearest Neighbor Dalam *Klasifikasi* Penyakit Diabetes. *J-REMI : Jurnal Rekam Medik Dan Informasi Kesehatan*, *4*(3), 176–184. https://doi.org/10.25047/j-remi.v4i3.3980

Tarigan, R. (2022). HUBUNGAN GAYA HIDUP DENGAN TERJADINYA PENYAKIT DIABETES MELITUS DI RSU DAERAH Dr R.M

DJOELHAM. *Jurnal Keperawatan Priority*, *5*(1), 94–102. https://doi.org/10.34012/jukep.v5i1.2105

Wibawa, A. P., Guntur, M., Purnama, A., Fathony Akbar, M., & Dwiyanto, F. A. (2018). Metode-metode *Klasifikasi*. *Prosiding Seminar Ilmu Komputer Dan Teknologi Informasi*, *3*(1), 134–138.

Abdulhadi, N., & Al-Mousa, A. (2021). Diabetes Detection Using Machine Learning Classification Methods. *2021 International Conference on Information Technology, ICIT 2021 - Proceedings*, *July*, 350–354. https://doi.org/10.1109/ICIT52682.2021.9491788

Anisa, D. N., & Jumanto, J. (2022). *Klasifikasi* Penyakit Diabetes Menggunakan *Algoritma* Naive Bayes. *Jurnal Dinamika Informatika*, *14*(1), 33–42. https://doi.org/10.35315/informatika.v14i1.9135

Aprillia, N., Hadi, Z., & Ishak, N. I. (2022). Hubungan Aktivitas Fisik Dan Obesitas Terhadap Kejadian DiabetesMelitus Di Wilayah Kerja Puskesmas Birayang Kabupaten HuluSungai Tengah Tahun 2022. *EPrints Uniska*, 1–9. http://eprints.uniska-bjm.ac.id/12139/

Arifin, T., & Ariesta, D. (2019). Prediksi Penyakit Ginjal Kronis Menggunakan *Algoritma* Naive Bayes Classifier Berbasis Particle Swarm Optimization. *Jurnal Tekno Insentif*, *13*(1), 26–30. https://doi.org/10.36787/jti.v13i1.97

Bangun, A., & Rachmat, E. (2024). *Analisis Perbandingan Algoritma KNN dan Naïve Bayes dalam Mendiagnosis Penyakit Diabetes Mellitus Pendahuluan*. *23*(September), 387–396.

Diana Dewi, D., Qisthi, N., Lestari, S. S. S., & Putri, Z. H. S. (2023). Perbandingan Metode Neural Network Dan Support Vector Machine Dalam *Klasifikasi* Diagnosa Penyakit Diabetes. *Cerdika: Jurnal Ilmiah Indonesia*, *3*(09), 828–839. https://doi.org/10.59141/cerdika.v3i09.662

Ghozali, A., Pratiwi, H., & Handajani, S. S. (2023). Implementasi Data Mining Menggunakan Metode Random Forest Dan Support Vector Machine Dalam *Klasifikasi* Penyakit Diabetes. *Delta: Jurnal Ilmiah Pendidikan Matematika*, *11*(2), 147. https://doi.org/10.31941/delta.v11i2.2686

Ginting, R. G., Girsang, E., Ginting, J. B., & Hartono, H. (2022). Analisis Determinan Dan Prediksi Penyakit Diabetes Melitus Tipe 2 Menggunakan Metode Machine Learning: Scoping Review. *Jurnal Maternitas Kebidanan*, *7*(1), 58–72. https://doi.org/10.34012/jumkep.v7i1.2538

Gurning, U. R., Octavia, S. F., Andriyani, D. R., Nurainun, N., & Permana, I. (2024). Prediksi Risiko Stunting pada Keluarga Menggunakan Naïve Bayes Classifier dan Chi-Square. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(1), 172–180. https://doi.org/10.57152/malcom.v4i1.1074

Kasandra, T. A., Kurniasih, E., & Ekayamti, E. (2022). *Media Publikasi*

Penelitian ; 2022 ; Volume 9 ; No 1 Website : http://jurnal.akperngawi.ac.id Hubungan Dukungan Keluarga Terhadap Kualitas Hidup pada Penderita Diabetes Melitus di Dusun Cung Belud Kecamatan Paron Kabupaten Ngawi The Relationship of Family. *9*(1), 74–82.

Kemenag), D. A. R. (atau: T. P. T. A.-Q. (2011). *Al-Qur'an dan Tafsirnya (Edisi yang Disempurnakan), Jilid II*. http://scioteca.caf.com/bitstream/handle/123456789/1091/RED2017-Eng-8ene.pdf?sequence=12&isAllowed=y%0Ahttp://dx.doi.org/10.1016/j.regsciurbeco.2008.06.005%0Ahttps://www.researchgate.net/publication/305320484_SISTEM_PEMBETUNGAN_TERPUSAT_STRATEGI_MELESTARI

Maulidah, N., Abdilah, A., Nurlelah, E., Gata, W., Nur Hasan, F., Ilmu Komputer, J., & Nusa Mandiri Jalan Margonda Raya No, S. (2020). Seleksi Fitur *Klasifikasi* Penyakit Diabetes Menggunakan Particle Swarm Optimization (PSO) Pada *Algoritma* Naive Bayes. *Daerah Khusus Ibukota Jakarta*, *13*(2), 21231170. http://journal.stekom.ac.id/index.php/elkom■page40

Muhammadiyah Jember, U., & Tri Rahayu, P. (2022). Perbandingan *Algoritma* K-Nearest Neighbor Dan Gaussian Naïve Bayes Pada Klsifikai Penyakit Diabetes Melitus Comparison Of K-Nears Neighbor And Gaussian Naïve Bayes Algorithm On The Classification Of Diabetes Mellitus. *Jurnal Smart Teknologi*, *3*(4), 2774–1702. http://jurnal.unmuhjember.ac.id/index.php/JST

Mutiara, E.-. (2020). *Algoritma Klasifikasi* Naive Bayes Berbasis Particle Swarm Optimization Untuk Prediksi Penyakit Tuberculosis (Tb). *Swabumi*, *8*(1), 46–58. https://doi.org/10.31294/swabumi.v8i1.7668

Pradhani, P. C., Indrayani, A. S., Azzahra, N., Aflikha, E., Zahra, F., & Kalifia, A. D. (2025). *342 | P a g e*. *3*, 342–353.

Purnomo, M. H., & Yuhana, U. L. (2016). Implementasi IOT dan Machine Learning Dalam Bidang Pendidikan Pembelajaran Matematika Tingkat SD melalui Serious Game. *National Conference of Applied Sciences, Engineering, Business and Information Technology*, 250–257.

Salissa, I., N, W. T., & Ningsih, W. T. (2023). Gambaran Pola Makan , Pola Istirahat , Pola Aktivitas Dan. *Jurnal Multidisiplin Indonesia*, *2*(September), 2435–2444.

Suryadewiansyah, M. K., Endra, T., & Tju, E. (2020). Naïve Bayes dan *Confusion matrix* untuk Efisiensi Analisa Intrusion Detection System Alert. *Jurnal Nasional Teknologi Dan Sistem Informasi*, *8*(2), 81–88.

Susilowati, D., Sutrisno, S., & Yunus, M. (2023). Penerapan Particle Swarm Optimization Untuk Meningkatkan Kinerja *Algoritma* K-Nearest Neighbor Dalam *Klasifikasi* Penyakit Diabetes. *J-REMI : Jurnal Rekam Medik Dan Informasi Kesehatan*, *4*(3), 176–184. https://doi.org/10.25047/j-remi.v4i3.3980

Tarigan, R. (2022). HUBUNGAN GAYA HIDUP DENGAN TERJADINYA PENYAKIT DIABETES MELITUS DI RSU DAERAH Dr R.M DJOELHAM. *Jurnal Keperawatan Priority*, *5*(1), 94–102. https://doi.org/10.34012/jukep.v5i1.2105

Wibawa, A. P., Guntur, M., Purnama, A., Fathony Akbar, M., & Dwiyanto, F. A. (2018). Metode-metode *Klasifikasi*. *Prosiding Seminar Ilmu Komputer Dan Teknologi Informasi*, *3*(1), 134–138.