

Turnitin LLCC

2213030130_Mediana Nurlailli_CEK.docx

 SI-09

Document Details

Submission ID

trn:oid::2945:394816457

Submission Date

Jun 23, 2026, 1:11 PM GMT+7

Download Date

Jun 23, 2026, 1:15 PM GMT+7

File Name

2213030130_Mediana Nurlailli_CEK.docx.pdf

File Size

2.8 MB

56 Pages

10,749 Words

67,824 Characters

25% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography

Top Sources

- 14%  Internet sources
- 9%  Publications
- 21%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Top Sources

- 14% Internet sources
- 9% Publications
- 21% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Student papers	Pace University on 2022-12-22	1%
2	Internet	etheses.uin-malang.ac.id	1%
3	Internet	jurnal.pustakagalerimandiri.co.id	1%
4	Student papers	Universitas Budi Luhur on 2026-01-31	<1%
5	Student papers	Sekolah Teknik Elektro & Informatika on 2025-07-18	<1%
6	Student papers	UPN Veteran Jakarta on 2025-12-19	<1%
7	Internet	inf-schule.de	<1%
8	Student papers	Universitas Islam Riau on 2026-06-02	<1%
9	Student papers	Universitas Indonesia on 2024-12-17	<1%
10	Internet	ejurnal.stmik-budidarma.ac.id	<1%
11	Student papers	Universitas Islam Riau on 2026-06-10	<1%

12	Student papers	Sekolah Teknik Elektro & Informatika on 2025-08-20	<1%
13	Internet	jutif.if.unsoed.ac.id	<1%
14	Internet	repository.uin-suska.ac.id	<1%
15	Internet	dergipark.org.tr	<1%
16	Student papers	Universitas Islam Riau on 2025-11-27	<1%
17	Student papers	Sriwijaya University on 2025-01-14	<1%
18	Internet	id.123dok.com	<1%
19	Internet	kc.umh.ac.id	<1%
20	Student papers	Universitas Dinamika on 2026-06-11	<1%
21	Internet	repository.upi.edu	<1%
22	Student papers	Middlesex University on 2021-05-02	<1%
23	Student papers	UNIVERSITAS BUDI LUHUR on 2025-07-26	<1%
24	Student papers	Universitas Lancang Kuning on 2024-08-05	<1%
25	Internet	repositori.usu.ac.id	<1%

26	Publication	Houda El Bouhissi, Shubin Xu, John Wang. "An Optimization-Driven Approach for ..."	<1%
27	Student papers	Telkom University on 2026-06-08	<1%
28	Student papers	UPN Veteran Yogyakarta on 2026-03-31	<1%
29	Student papers	Universitas Brawijaya on 2023-05-22	<1%
30	Student papers	Universitas Muhammadiyah Sukabumi on 2026-01-13	<1%
31	Internet	journal.artei.or.id	<1%
32	Internet	media.neliti.com	<1%
33	Student papers	Sriwijaya University on 2025-07-10	<1%
34	Student papers	UPN Veteran Yogyakarta on 2026-05-04	<1%
35	Student papers	Udayana University on 2026-01-08	<1%
36	Internet	lmsspada.kemdiktisaintek.go.id	<1%
37	Internet	repositori.buddhidharma.ac.id	<1%
38	Internet	repository.usd.ac.id	<1%
39	Student papers	Sultan Agung Islamic University on 2024-08-12	<1%

40	Student papers	Telkom University on 2026-06-17	<1%
41	Student papers	UPN Veteran Yogyakarta on 2026-03-02	<1%
42	Internet	jtera.polteksmi.ac.id	<1%
43	Student papers	Telkom University on 2026-01-21	<1%
44	Student papers	Universitas Bengkulu on 2025-05-05	<1%
45	Internet	jiki.jurnal-id.com	<1%
46	Student papers	Islington College,Nepal on 2026-05-10	<1%
47	Publication	Muhammad Reza Pahlevi, Imam Tahyudin, Ades Tikaningsih. "AUTOMATED EDIBI...	<1%
48	Student papers	Tarumanagara University on 2025-04-24	<1%
49	Student papers	Universitas Airlangga on 2025-05-24	<1%
50	Student papers	Universitas Kristen Duta Wacana on 2026-05-06	<1%
51	Publication	Veren Nita Permatasari, Raisah Fajri Aula, Yuma Akbar, Aditya Zakaria Hidayat. "...	<1%
52	Internet	e-theses.iaincurup.ac.id	<1%
53	Student papers	Syiah Kuala University on 2020-10-06	<1%

54	Student papers	Universitas Islam Indonesia on 2017-10-06	<1%
55	Student papers	Universitas Muhammadiyah Jember on 2026-01-27	<1%
56	Internet	docplayer.info	<1%
57	Internet	subset.id	<1%
58	Publication	Anthony Holly. "DETERMINAN NILAI PERUSAHAAN DI BURSA EFEK INDONESIA", SI...	<1%
59	Student papers	Politeknik Statistika STIS on 2025-07-05	<1%
60	Student papers	Tarumanagara University on 2025-04-24	<1%
61	Internet	dspace.uii.ac.id	<1%
62	Internet	repository.stikesbcm.ac.id	<1%
63	Internet	repository.uinfasbengkulu.ac.id	<1%
64	Student papers	UPN Veteran Yogyakarta on 2026-04-02	<1%
65	Student papers	Universitas Pendidikan Ganesha on 2026-04-06	<1%
66	Publication	Desi Permata Sari, Bambang Irawan. "Implementasi CNN Mobilenetv2 untuk Klas...	<1%
67	Student papers	Sekolah Teknik Elektro & Informatika on 2025-07-18	<1%

68	Student papers	Universitas Airlangga-1 on 2026-06-19	<1%
69	Student papers	Universitas Islam Riau on 2025-11-24	<1%
70	Publication	Yudhi Fajar Saputra, Mahmoud Ahmad Al-Khasawneh, Milkhatun Milkhatun, Ni W...	<1%
71	Student papers	Institut Teknologi Sumatera on 2026-01-12	<1%
72	Student papers	Universitas Putera Batam on 2026-02-02	<1%
73	Internet	core.ac.uk	<1%
74	Internet	dspace.nitrkl.ac.in	<1%
75	Internet	e-repository.perpus.iainsalatiga.ac.id	<1%
76	Internet	ejurnal.seminar-id.com	<1%
77	Internet	es.scribd.com	<1%
78	Internet	freeronald.org	<1%
79	Internet	journal.forikami.com	<1%
80	Internet	journals.insightpub.org	<1%
81	Internet	www.localstartupfest.id	<1%

82	Publication	Asrul Abdullah, Della Udy Khairah, Menur Wahyu Pangestika. "COMPARISON OF ..."	<1%
83	Student papers	Konsorsium Perguruan Tinggi Swasta Indonesia on 2025-02-04	<1%
84	Student papers	Universitas Brawijaya on 2021-06-14	<1%
85	Student papers	Universitas Diponegoro on 2021-01-20	<1%
86	Student papers	Universitas Pendidikan Indonesia on 2025-08-19	<1%
87	Student papers	Universitas Tarumanagara on 2025-12-17	<1%
88	Publication	Yuni Eka Achyani. "Penerapan Metode Particle Swarm Optimization Pada Optima..."	<1%
89	Internet	digilibadmin.unismuh.ac.id	<1%
90	Internet	eprints.uny.ac.id	<1%
91	Internet	jurnal.umb.ac.id	<1%
92	Internet	simpel.its.ac.id	<1%
93	Student papers	Ajou University Graduate School on 2024-07-01	<1%
94	Student papers	Fakultas Ekonomi Universitas Indonesia on 2015-07-07	<1%
95	Student papers	Fakultas Kedokteran Universitas Pattimura on 2025-12-21	<1%

96	Publication	Fitrah Amelia Ramelan, Lukman Hakim. "Perbandingan Prediksi terhadap Pening..."	<1%
97	Student papers	Politeknik Negeri Bandung on 2018-11-05	<1%
98	Student papers	Sriwijaya University on 2022-01-19	<1%
99	Student papers	Telkom University on 2026-06-08	<1%
100	Student papers	UPN Veteran Yogyakarta on 2026-05-05	<1%
101	Student papers	Universitas Brawijaya on 2018-01-08	<1%
102	Student papers	Universitas Islam Negeri Sumatera Utara on 2026-06-06	<1%
103	Student papers	Universitas Muslim Indonesia on 2024-11-14	<1%
104	Student papers	Universitas Pamulang on 2026-06-09	<1%
105	Student papers	Universitas Putera Batam on 2025-07-25	<1%
106	Student papers	Universitas Putera Batam on 2026-02-03	<1%
107	Student papers	Universitas Tarumanagara on 2026-01-04	<1%
108	Internet	ejnteti.jteti.ugm.ac.id	<1%
109	Internet	ejournal.uin-suka.ac.id	<1%

110	Internet	lppm.usm.ac.id	<1%
111	Internet	repository.ub.ac.id	<1%
112	Publication	Adisha Dhia Bimantara Adisha, Bambang Adiwino Bambang. "Analisis Sentime...	<1%
113	Publication	Bryan Adam Hidayatullah, Bangkit Indarmawan Nugroho, Nugroho Adhi Santoso...	<1%
114	Publication	Cicin Hardiyanti P, Dita Danianti. "Perbandingan Kinerja Algoritma Ensemble dan ...	<1%
115	Publication	Erika Sita Dewi, Ariella Chandra Naya, Yuyun Umaidah. "Analisis Sentimen Komen...	<1%
116	Student papers	FAKULTAS ILMU KOMPUTER on 2025-12-22	<1%
117	Student papers	Fakultas Teknik on 2025-07-23	<1%
118	Student papers	Fakultas Teknik on 2025-08-01	<1%
119	Student papers	IAKN Palangka Raya on 2026-06-14	<1%
120	Publication	Iustisia Natalia Simbolon. "PREDIKSI KUALITAS AIR SUNGAI DI JAKARTA MENGGU...	<1%
121	Publication	Muhammad Ridho Pratama -, Jatmalika Abi Kurniawan -, Muhammad Rizal -, Wah...	<1%
122	Publication	Rizky Wahyudi, Fauzan Azima Lubis, Windy Aulia, Adidtya Perdana. "OPTIMASI SE...	<1%
123	Student papers	Sriwijaya University on 2019-05-22	<1%

124	Student papers	Tarumanagara University on 2026-06-22	<1%
125	Student papers	UPN Veteran Yogyakarta on 2025-09-29	<1%
126	Student papers	UPN Veteran Yogyakarta on 2026-05-04	<1%
127	Student papers	Universitas Bengkulu on 2025-11-05	<1%
128	Student papers	Universitas Brawijaya on 2018-07-19	<1%
129	Student papers	Universitas Budi Luhur on 2026-01-27	<1%
130	Student papers	Universitas Islam Indonesia on 2025-11-26	<1%
131	Student papers	Universitas Islam Riau on 2026-06-11	<1%
132	Student papers	Universitas Muria Kudus on 2025-10-15	<1%
133	Student papers	Universitas Trunojoyo on 2020-08-04	<1%
134	Internet	codefinity.com	<1%
135	Internet	digilib.unila.ac.id	<1%
136	Internet	id.scribd.com	<1%
137	Internet	ioai.toki.id	<1%

138	Internet	lwrs.inl.gov	<1%
139	Internet	ojs.stmik-banjarbaru.ac.id	<1%
140	Internet	www.ijrti.org	<1%
141	Internet	www.scribd.com	<1%
142	Student papers	Sriwijaya University on 2021-08-30	<1%
143	Internet	repository.its.ac.id	<1%
144	Publication	Aymeric Kalife. "Sustainable Financial Structured Products - Managing Risk Appet..."	<1%
145	Publication	Kartina Diah Kusuma Wardhani, Memen Akbar. "Diabetes Risk Prediction Using E..."	<1%
146	Student papers	Sriwijaya University on 2020-12-17	<1%
147	Student papers	UNIVERSITAS BUDI LUHUR on 2026-01-22	<1%
148	Student papers	Universitas Brawijaya on 2017-07-03	<1%
149	Student papers	Universitas Budi Luhur on 2023-08-10	<1%
150	Student papers	Universitas Diponegoro on 2021-08-05	<1%
151	Student papers	Universitas Muhammadiyah Jakarta on 2026-05-08	<1%

152

Student papers

Universitas Pendidikan Indonesia on 2025-05-21

<1%

BAB I

PENDAHULUAN

A. Latar Belakang

Peningkatan jumlah penderita diabetes melitus dari tahun ke tahun menjadi perhatian serius di berbagai negara. Kondisi ini tidak hanya berdampak pada kualitas hidup individu, tetapi juga memberikan tekanan besar terhadap sistem layanan kesehatan secara keseluruhan. Diabetes melitus ditandai oleh terganggunya kemampuan tubuh dalam mengatur kadar gula darah akibat disfungsi produksi atau kerja hormon insulin, yang apabila tidak ditangani secara tepat dapat berkembang menjadi berbagai komplikasi berbahaya seperti gagal ginjal, kerusakan saraf, gangguan penglihatan, dan penyakit jantung (World Health Organization, 2023).

Identifikasi risiko diabetes sejak tahap awal memiliki peranan strategis dalam menekan laju perkembangan penyakit serta mengurangi beban komplikasi yang ditimbulkan. Sayangnya, pendekatan konvensional dalam dunia medis masih mengandalkan pemeriksaan laboratorium yang bersifat reaktif, artinya pemeriksaan baru dilakukan setelah gejala muncul secara nyata. Keterbatasan ini mendorong perlunya pengembangan metode alternatif yang mampu mengidentifikasi risiko diabetes secara lebih dini dan berbasis pada analisis pola data pasien (Sneha & Gangil, 2019).

Transformasi digital di sektor kesehatan telah menghasilkan akumulasi data medis dalam jumlah besar yang menyimpan potensi informasi berharga. Pemanfaatan data tersebut melalui pendekatan machine learning memungkinkan pembangunan model prediktif yang mampu mendeteksi risiko penyakit secara lebih awal dan sistematis. Berbagai studi membuktikan bahwa model berbasis machine learning dapat menghasilkan tingkat akurasi yang lebih tinggi dibandingkan metode diagnosis konvensional dalam konteks prediksi diabetes (Khan et al., 2021).

Penelitian ini memanfaatkan dua dataset diabetes dengan karakteristik yang berbeda. Dataset pertama adalah Early Stage Diabetes Risk Prediction

Dataset dari UCI Machine Learning Repository, yang digunakan dalam penelitian Nugroho et al. (2021), terdiri dari 520 data dan 16 fitur gejala klinis yang mayoritas bersifat kategorikal. Dataset kedua adalah Pima Indians Diabetes Dataset yang terdiri dari 768 data dengan 8 fitur numerik dan telah lama menjadi tolok ukur dalam penelitian prediksi diabetes. Penggabungan dua dataset dengan karakteristik berbeda ini bertujuan untuk menguji konsistensi dan kemampuan generalisasi metode seleksi fitur yang diterapkan (Kaliappan et al., 2024).

Kualitas model machine learning dalam prediksi diabetes sangat bergantung pada relevansi fitur yang digunakan sebagai masukan. Dataset medis sering kali mengandung fitur-fitur yang tidak memberikan kontribusi signifikan terhadap variabel target, bahkan dapat mengganggu proses pembelajaran model. Kondisi ini menyebabkan model menjadi tidak efisien, rentan terhadap overfitting, dan sulit untuk diinterpretasikan secara klinis. Oleh karena itu, tahap seleksi fitur menjadi komponen yang tidak dapat diabaikan dalam alur pengembangan model prediktif berbasis data medis (Patel, 2023).

Metode seleksi fitur berbasis wrapper, khususnya Sequential Forward Selection (SFS) dan Sequential Backward Selection (SBS), telah terbukti efektif dalam meningkatkan akurasi model klasifikasi. Nugroho et al. (2021) dalam penelitiannya menunjukkan bahwa penerapan wrapper pada dataset Early Stage Diabetes UCI berhasil meningkatkan akurasi Random Forest dari 97,50% menjadi 98,84% (SFS) dan 99,03% (SBS) dengan hanya menggunakan 11 dari 16 fitur yang tersedia. Temuan ini mengindikasikan bahwa terdapat sejumlah fitur yang justru menurunkan performa model ketika diikutsertakan dalam proses pelatihan.

Kendati demikian, metode SFS dan SBS memiliki kelemahan mendasar yaitu sifatnya yang greedy sehingga hanya mampu mengeksplorasi sebagian kecil dari ruang pencarian fitur yang tersedia. Keterbatasan ini berpotensi menghasilkan subset fitur yang hanya optimal secara lokal, bukan secara global. Particle Swarm Optimization (PSO) sebagai algoritma berbasis kecerdasan kolektif menawarkan mekanisme pencarian yang lebih menyeluruh dengan cara menyimulasikan pergerakan sekelompok partikel dalam ruang solusi untuk menemukan kombinasi fitur terbaik (Sheta et al., 2024).

Kemampuan PSO dalam menyeimbangkan eksplorasi ruang solusi secara luas dan eksploitasi solusi terbaik yang ditemukan menjadikannya pilihan yang menarik sebagai pendekatan metaheuristik dalam seleksi fitur. Sejumlah penelitian menunjukkan bahwa integrasi PSO dengan algoritma klasifikasi mampu menghasilkan peningkatan akurasi yang lebih kompetitif dibandingkan metode wrapper konvensional, terutama pada dataset dengan dimensi yang lebih kompleks (Rahman et al., 2025).

Random Forest dipilih sebagai algoritma klasifikasi dalam penelitian ini karena kemampuannya dalam menangani data berdimensi tinggi, ketahanannya terhadap noise, serta kecenderungannya menghasilkan performa yang stabil pada berbagai jenis dataset. Algoritma ini bekerja dengan membangun sejumlah besar pohon keputusan secara paralel dan menggabungkan hasilnya melalui mekanisme voting mayoritas, sehingga mampu mereduksi variansi prediksi secara efektif (Khan et al., 2021).

Berdasarkan uraian di atas, penelitian ini dirancang untuk membandingkan efektivitas tiga metode seleksi fitur, yaitu SFS, SBS, dan PSO, dalam meningkatkan kinerja algoritma Random Forest pada dua dataset diabetes dengan karakteristik yang berbeda. Dengan menggunakan dua dataset sekaligus, penelitian ini diharapkan dapat menghasilkan temuan yang lebih komprehensif dan dapat digeneralisasi, serta memberikan rekomendasi metode seleksi fitur yang paling optimal untuk aplikasi prediksi risiko diabetes.

B. Identifikasi Masalah

- Belum optimalnya model prediksi risiko diabetes berbasis machine learning akibat penggunaan seluruh fitur tanpa proses seleksi yang memadai.
- Keterbatasan metode wrapper konvensional (SFS dan SBS) yang bersifat greedy dalam menemukan kombinasi fitur optimal secara global.
- Minimnya penelitian yang membandingkan SFS, SBS, dan PSO secara bersamaan pada dua dataset diabetes dengan karakteristik yang berbeda menggunakan algoritma Random Forest.

C. Batasan Masalah

- Penelitian menggunakan dua dataset: (1) Early Stage Diabetes Risk Prediction dari UCI (520 data, 16 fitur) dan (2) Pima Indians Diabetes Dataset (768 data, 8 fitur numerik).
- Metode seleksi fitur yang dibandingkan adalah SFS, SBS, dan PSO, dengan Random Forest sebagai algoritma klasifikasi dan 10-fold cross validation sebagai skema validasi.
- Evaluasi model dilakukan menggunakan metrik akurasi, precision, recall, dan F1-score, tanpa mencakup aspek implementasi klinis.

D. Rumusan Masalah

- Berapa nilai akurasi algoritma Random Forest tanpa seleksi fitur pada masing-masing dataset?
- Berapa perubahan nilai akurasi dan jumlah fitur yang dihasilkan oleh metode SFS, SBS, dan PSO pada masing-masing dataset?
- Metode seleksi fitur manakah yang menghasilkan akurasi tertinggi dengan jumlah fitur paling efisien pada masing-masing dataset?

E. Tujuan Penelitian

- Mengukur dan membandingkan kinerja algoritma Random Forest sebelum dan sesudah penerapan metode seleksi fitur SFS, SBS, dan PSO pada dua dataset diabetes.
- Mengidentifikasi metode seleksi fitur yang memberikan peningkatan performa terbaik berdasarkan metrik evaluasi klasifikasi pada masing-masing dataset.
- Menghasilkan model prediksi risiko diabetes yang lebih akurat dan efisien melalui optimasi dimensi fitur pada dua dataset dengan karakteristik berbeda.

F. Manfaat Penelitian

19

75

41

- Secara teoritis, penelitian ini memperkaya khazanah ilmu data mining dan machine learning di bidang kesehatan, khususnya terkait perbandingan metode seleksi fitur pada dataset medis berganda.
- Secara akademis, hasil penelitian dapat menjadi referensi bagi peneliti lain yang ingin mengembangkan kajian serupa dengan dataset atau metode yang berbeda.
- Secara praktis, temuan penelitian dapat menjadi landasan pengembangan sistem pendukung keputusan berbasis data untuk deteksi risiko diabetes yang lebih akurat dan efisien.

BAB II

KAJIAN PUSTAKA

A. Kajian Teori

1. Konsep Dasar Diabetes Melitus

Diabetes melitus merupakan gangguan metabolisme kronis yang muncul akibat ketidakseimbangan antara produksi insulin oleh pankreas dan kebutuhan tubuh akan hormon tersebut. Kondisi ini mengakibatkan kadar glukosa dalam darah meningkat secara persisten dan berpotensi menimbulkan kerusakan jangka panjang pada berbagai organ vital. Secara klinis, diabetes dibagi menjadi beberapa tipe, di mana tipe 1 disebabkan oleh reaksi autoimun yang menghancurkan sel beta pankreas, sedangkan tipe 2 terkait dengan resistensi insulin yang dipengaruhi oleh faktor gaya hidup dan genetik (World Health Organization, 2023).

Manifestasi klinis awal diabetes mencakup sejumlah gejala khas seperti frekuensi buang air kecil yang meningkat (poliuria), rasa haus berlebihan (polidipsia), penurunan berat badan yang tidak dapat dijelaskan, kelemahan tubuh, dan gangguan penglihatan. Deteksi terhadap gejala-gejala ini secara dini memungkinkan intervensi yang lebih efektif sebelum komplikasi serius berkembang. Pendekatan berbasis analisis data dinilai mampu mengidentifikasi pola risiko lebih awal dibandingkan metode pemeriksaan konvensional (Sneha & Gangil, 2019).

2. Dataset Prediksi Diabetes

Penelitian ini menggunakan dua dataset yang memiliki karakteristik dan sumber yang berbeda. Dataset pertama adalah Early Stage Diabetes Risk Prediction Dataset yang tersedia secara publik di UCI Machine Learning Repository. Dataset ini dikembangkan berdasarkan data survei pasien di Rumah Sakit Sylhet, Bangladesh, dan terdiri dari 520 rekaman dengan 16 fitur gejala klinis serta satu variabel target. Sebagian besar fitur bersifat biner (Ya/Tidak) yang telah merepresentasikan ada atau tidaknya gejala tertentu pada pasien (Islam et al., 2020).

24 Dataset kedua adalah Pima Indians Diabetes Dataset yang dikembangkan oleh National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Dataset ini mencakup data dari populasi wanita keturunan Indian Pima berusia minimal 21 tahun dan terdiri dari 768 data dengan 8 fitur numerik yang merepresentasikan kondisi klinis pasien serta satu variabel target biner. Dataset ini telah menjadi tolok ukur klasik dalam penelitian prediksi diabetes dan klasifikasi medis selama beberapa dekade (Kaliappan et al., 2024).

Perbedaan mendasar antara kedua dataset ini terletak pada jenis fiturnya. Dataset UCI didominasi oleh fitur kategorikal yang memerlukan transformasi sebelum diproses, sedangkan dataset Pima Indians seluruhnya bersifat numerik. Perbandingan hasil pada dua dataset dengan karakteristik berbeda ini bertujuan untuk menguji sejauh mana konsistensi dan kemampuan generalisasi dari metode seleksi fitur yang diterapkan.

3. Data Mining

105 Data mining merupakan disiplin komputasi yang berfokus pada penggalian pengetahuan tersembunyi dari kumpulan data berukuran besar melalui serangkaian proses analitis yang sistematis. Secara metodologis, data mining mencakup tahapan pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, hingga penerapan. Setiap tahapan dirancang untuk memastikan bahwa pola yang ditemukan memiliki validitas dan kegunaan dalam konteks pengambilan keputusan (Patel, 2023).

51 Dalam lingkup kesehatan, data mining berperan dalam mengubah volume besar data klinis menjadi informasi yang dapat digunakan untuk mendukung diagnosa, prognosa, dan perencanaan perawatan. Kemampuannya untuk mengidentifikasi korelasi tersembunyi antar variabel medis menjadikan data mining sebagai komponen penting dalam pengembangan sistem prediksi penyakit berbasis bukti (Khan et al., 2021).

4. Machine Learning

30 Machine learning adalah subdisiplin kecerdasan buatan yang mengembangkan algoritma mampu belajar dari data dan meningkatkan kinerjanya

secara otomatis tanpa perlu diprogram secara eksplisit untuk setiap tugas. Berdasarkan paradigma pembelajaran, machine learning dibedakan menjadi supervised learning yang menggunakan data berlabel, unsupervised learning yang bekerja tanpa label, dan reinforcement learning yang belajar melalui interaksi dengan lingkungan. Dalam domain prediksi penyakit, supervised learning mendominasi karena ketersediaan data klinis yang telah diberi label diagnosis (Musa & Abdulazeez, 2023).

Keunggulan pendekatan machine learning dalam analisis data medis terletak pada kemampuannya menangani data multivariat yang kompleks, mengidentifikasi pola non-linear yang tidak dapat ditangkap oleh metode statistik konvensional, serta menghasilkan model yang dapat memperbarui pengetahuannya seiring bertambahnya data baru. Berbagai penelitian mengonfirmasi bahwa model machine learning secara konsisten menghasilkan performa prediksi yang lebih baik dibandingkan pendekatan tradisional dalam konteks klasifikasi penyakit kronis seperti diabetes (Sheta et al., 2024).

5. Klasifikasi dan Metrik Evaluasi

Klasifikasi dalam machine learning merupakan proses pemetaan data masukan ke dalam kategori diskrit berdasarkan pola yang dipelajari dari data latih. Dalam penelitian ini, tujuan klasifikasi adalah menentukan apakah seorang pasien tergolong positif atau negatif memiliki risiko diabetes berdasarkan fitur-fitur yang tersedia. Evaluasi kinerja model klasifikasi dilakukan menggunakan confusion matrix yang mengkuantifikasi empat hasil prediksi, yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN) (Sucipto et al., 2024).

Tabel 2.1 Confusion Matrix

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

Berdasarkan nilai-nilai dalam confusion matrix tersebut, metrik evaluasi yang digunakan dalam penelitian ini dihitung menggunakan rumus-rumus berikut (Sucipto et al., 2024):

- a) Akurasi (Accuracy): mengukur proporsi keseluruhan prediksi yang benar terhadap total data.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad \dots(2.1)$$

- b) Presisi (Precision): mengukur proporsi prediksi positif yang benar-benar merupakan positif sejati.

$$Precision = TP / (TP + FP) \quad \dots(2.2)$$

- c) Recall (Sensitivitas): mengukur kemampuan model dalam mendeteksi seluruh kasus positif yang sebenarnya.

$$Recall = TP / (TP + FN) \quad \dots(2.3)$$

- d) F1-Score: rata-rata harmonik dari precision dan recall yang memberikan keseimbangan antara keduanya.

$$F1-Score = 2 \times (Precision \times Recall) / (Precision + Recall) \quad \dots(2.4)$$

6. Cross Validation

Cross validation merupakan teknik resampling yang digunakan untuk mengevaluasi kemampuan generalisasi model machine learning secara lebih andal. Prinsip dasarnya adalah membagi dataset ke dalam beberapa bagian yang disebut fold, lalu melatih dan menguji model secara bergantian sehingga setiap bagian data pernah digunakan sebagai data uji tepat satu kali. Pendekatan ini menghasilkan estimasi kinerja model yang lebih stabil dan tidak bias dibandingkan pembagian data latih-uji tunggal (Nugroho et al., 2021).

Dalam penelitian ini digunakan metode 10-fold cross validation, di mana data dibagi menjadi 10 bagian yang proporsional. Nilai akurasi rata-rata dihitung menggunakan persamaan berikut:

$$Akurasi \text{ Rata-rata} = (1/k) \times \sum Akurasi_i, i = 1, 2, \dots, k \quad \dots(2.5)$$

di mana k adalah jumlah fold ($k = 10$) dan $Akurasi_i$ adalah akurasi yang diperoleh pada iterasi ke-i. Pemilihan nilai $k = 10$ didasarkan pada hasil percobaan

yang menunjukkan kestabilan evaluasi pada nilai fold tersebut, serta penggunaannya yang meluas sebagai standar dalam penelitian machine learning di bidang medis (Nugroho et al., 2021).

7. Seleksi Fitur

Seleksi fitur adalah proses analitis yang bertujuan mengidentifikasi dan memilih subset fitur paling informatif dari keseluruhan fitur yang tersedia dalam suatu dataset. Tujuan utamanya mencakup peningkatan akurasi model, pengurangan waktu komputasi, serta peningkatan interpretabilitas hasil. Secara umum, metode seleksi fitur dikelompokkan ke dalam tiga kategori utama, yaitu metode filter yang mengevaluasi fitur secara independen dari algoritma pembelajaran, metode wrapper yang menggunakan algoritma klasifikasi sebagai fungsi evaluasi subset fitur, serta metode embedded yang mengintegrasikan seleksi fitur ke dalam proses pelatihan model (Sheta et al., 2024).

Dalam konteks data medis prediksi diabetes, seleksi fitur memiliki peran yang sangat krusial. Tidak seluruh gejala atau parameter klinis memberikan kontribusi yang sama terhadap prediksi risiko penyakit. Fitur yang tidak relevan atau redundan dapat menimbulkan noise yang menurunkan kemampuan diskriminasi model. Penelitian terdahulu secara konsisten menunjukkan bahwa penerapan seleksi fitur yang tepat dapat meningkatkan akurasi prediksi secara signifikan sekaligus menyederhanakan model yang dihasilkan (Sneha & Gangil, 2019).

8. Sequential Forward Selection (SFS)

Sequential Forward Selection (SFS) adalah algoritma seleksi fitur berbasis wrapper yang beroperasi secara iteratif dengan cara menambahkan satu fitur terbaik pada setiap tahapan seleksi. Prosesnya dimulai dari himpunan fitur kosong, kemudian pada setiap iterasi dievaluasi seluruh fitur yang belum terpilih dan dipilih fitur yang memberikan peningkatan kinerja model terbesar ketika ditambahkan ke dalam himpunan fitur yang sudah ada (Nugroho et al., 2021).

Algoritma SFS berhenti ketika penambahan fitur baru tidak lagi meningkatkan kinerja model atau ketika seluruh fitur telah dievaluasi. Secara matematis, pada setiap tahap t , fitur f^* dipilih berdasarkan kriteria berikut:

$$f^* = \arg \max J(S_t \cup \{f\}), f \notin S_t \quad \dots(2.6)$$

di mana S_t adalah himpunan fitur terpilih pada tahap t dan J adalah fungsi evaluasi kinerja model berdasarkan cross validation. Keunggulan SFS terletak pada efisiensi komputasinya karena hanya mengevaluasi penambahan fitur secara bertahap. Namun demikian, sifat greedy-nya menyebabkan fitur yang telah dipilih tidak dapat dihapus meskipun relevansinya berkurang pada tahap berikutnya (Kaliappan et al., 2024).

9. Sequential Backward Selection (SBS)

Berbeda dengan SFS, Sequential Backward Selection (SBS) memulai proses seleksi dari keseluruhan fitur yang tersedia, kemudian secara bertahap menghapus satu fitur yang paling sedikit berkontribusi terhadap kinerja model pada setiap iterasi. Pendekatan ini mempertahankan konteks keseluruhan fitur pada awal proses, sehingga mampu menangkap interaksi antar fitur yang mungkin tidak teridentifikasi oleh SFS (Nugroho et al., 2021).

Pada setiap tahap eliminasi, fitur f^* yang dihapus dipilih berdasarkan kriteria berikut:

$$f^* = \arg \max J(S_t \setminus \{f\}), f \in S_t \quad \dots(2.7)$$

di mana S_t adalah himpunan fitur yang tersisa pada tahap t dan operasi $\setminus$ menunjukkan penghapusan elemen dari himpunan. SBS berhenti ketika penghapusan fitur lebih lanjut mulai menurunkan kinerja model atau ketika jumlah fitur mencapai batas minimum yang ditentukan. Metode ini cenderung menghasilkan subset fitur yang lebih optimal dibandingkan SFS karena mempertimbangkan seluruh kombinasi fitur pada awal proses, meskipun memiliki kompleksitas komputasi yang lebih tinggi (Kaliappan et al., 2024).

10. Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) adalah algoritma optimasi metaheuristik yang dikembangkan oleh Kennedy dan Eberhart pada tahun 1995,

yang terinspirasi dari dinamika sosial kawanan hewan seperti burung dan ikan dalam menemukan sumber makanan. Dalam konteks seleksi fitur, setiap partikel merepresentasikan sebuah solusi kandidat berupa subset fitur yang dikodekan sebagai vektor biner, di mana nilai 1 menandakan fitur tersebut terpilih dan nilai 0 menandakan tidak terpilih (Sheta et al., 2024).

Setiap partikel i dalam swarm memiliki posisi x_i dan kecepatan v_i yang diperbarui pada setiap iterasi menggunakan persamaan berikut:

$$v_i(t+1) = w \times v_i(t) + c1 \times r1 \times (pbest_i - x_i(t)) + c2 \times r2 \times (gbest - x_i(t)) \quad \dots(2.8)$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad \dots(2.9)$$

di mana w adalah bobot inersia yang mengontrol pengaruh kecepatan sebelumnya, $c1$ dan $c2$ adalah koefisien percepatan kognitif dan sosial, $r1$ dan $r2$ adalah bilangan acak dalam rentang $[0,1]$, $pbest_i$ adalah posisi terbaik yang pernah dicapai oleh partikel i (personal best), dan $gbest$ adalah posisi terbaik yang pernah dicapai oleh seluruh swarm (global best).

Untuk seleksi fitur biner, kecepatan dikonversi menjadi probabilitas pemilihan fitur menggunakan fungsi sigmoid:

$$S(v_i) = 1 / (1 + e^{(-v_i)}) \quad \dots(2.10)$$

Kemudian, posisi partikel diperbarui menggunakan aturan berikut:

$$x_i(t+1) = 1 \text{ jika } rand() < S(v_i(t+1)), \text{ dan } x_i(t+1) = 0 \text{ jika sebaliknya} \quad \dots(2.11)$$

Nilai fitness setiap partikel dihitung berdasarkan akurasi klasifikasi menggunakan Random Forest dengan 10-fold cross validation pada subset fitur yang direpresentasikan oleh posisi partikel tersebut. PSO terus beriterasi hingga jumlah maksimum iterasi tercapai atau kondisi konvergensi terpenuhi. Keunggulan PSO dibandingkan metode wrapper konvensional terletak pada kemampuannya menjelajahi ruang solusi secara global sehingga dapat menemukan kombinasi fitur yang lebih optimal (Rahman et al., 2025).

11. Algoritma Random Forest

Random Forest merupakan algoritma ensemble learning yang dikembangkan oleh Breiman (2001) dengan cara mengombinasikan sejumlah besar pohon keputusan yang dibangun secara independen melalui mekanisme bagging dan pemilihan fitur acak. Setiap pohon dalam hutan dibangun menggunakan subset data latih yang diambil secara acak dengan penggantian (bootstrapping), dan pada setiap node pembelahan hanya sebagian kecil dari total fitur yang dievaluasi sebagai kandidat pembelahan terbaik.

Untuk tugas klasifikasi, prediksi akhir ditentukan melalui mekanisme voting mayoritas dari seluruh pohon dalam ensemble. Secara matematis, prediksi Random Forest untuk suatu data x dinyatakan sebagai:

$$\hat{y} = \arg \max_c \sum I(h_t(x) = c), \quad t = 1, 2, \dots, T \quad \dots(2.12)$$

di mana T adalah jumlah pohon, $h_t(x)$ adalah prediksi pohon ke- t untuk data x , c adalah kelas, dan $I(\cdot)$ adalah fungsi indikator. Variasi antar pohon yang dihasilkan dari proses bootstrapping dan pemilihan fitur acak memungkinkan Random Forest menghasilkan model yang memiliki bias rendah dan varians yang lebih kecil dibandingkan pohon keputusan tunggal, sehingga lebih tahan terhadap overfitting (Khan et al., 2021).

12. Overfitting dan Underfitting

Overfitting merupakan fenomena di mana model terlalu menyesuaikan diri terhadap data latih, termasuk noise dan pola kebetulan yang tidak representatif, sehingga menghasilkan performa tinggi pada data latih namun buruk pada data baru yang belum pernah dilihat. Secara formal, overfitting terjadi ketika training error $\ll$ test error. Sebaliknya, underfitting terjadi ketika model terlalu sederhana untuk menangkap pola utama dalam data, menghasilkan performa yang buruk baik pada data latih maupun data uji.

Penerapan seleksi fitur berkontribusi dalam memitigasi risiko overfitting dengan cara menghilangkan fitur yang tidak informatif sehingga model dapat fokus mempelajari pola yang benar-benar relevan. Sementara itu, algoritma Random Forest secara inheren mengatasi overfitting melalui mekanisme ensemble yang menghasilkan keragaman antar pohon, sehingga kombinasi seleksi fitur dan

Random Forest diharapkan menghasilkan model yang memiliki kemampuan generalisasi yang lebih baik (Patel, 2023).

B. Kajian Hasil Penelitian Terdahulu

Berbagai penelitian telah dilakukan untuk mengkaji pemanfaatan machine learning dalam klasifikasi dan prediksi diabetes. Kajian oleh Khan et al. (2021) yang meninjau secara sistematis penerapan machine learning pada dataset diabetes menyimpulkan bahwa algoritma ensemble seperti Random Forest secara konsisten menghasilkan performa yang kompetitif, namun kualitas fitur yang digunakan tetap menjadi faktor penentu utama keberhasilan model. Tinjauan tersebut juga menekankan pentingnya tahap pra-pemrosesan data, termasuk seleksi fitur, sebagai prasyarat untuk menghasilkan model prediksi yang andal.

Melakukan penelitian yang menjadi rujukan utama dalam studi ini. Mereka mengaplikasikan metode SFS dan SBS berbasis wrapper yang dikombinasikan dengan Random Forest pada Early Stage Diabetes Risk Prediction Dataset dari UCI. Hasilnya menunjukkan bahwa baseline tanpa seleksi fitur menghasilkan akurasi 97,50%, sedangkan penerapan SFS meningkatkan akurasi menjadi 98,84% dengan 11 fitur, dan SBS menghasilkan akurasi terbaik sebesar 99,03% dengan jumlah fitur yang sama. Temuan ini secara meyakinkan membuktikan bahwa reduksi dimensi melalui wrapper mampu meningkatkan kinerja model sekaligus menyederhanakan kompleksitas komputasinya. Penelitian ini menjadi acuan dasar bagi penelitian yang dikembangkan penulis, dengan penambahan metode PSO dan perluasan pengujian ke dataset Pima Indians Diabetes (Nugroho Fanani dan Shidik, 2021).

Melakukan kajian komprehensif terhadap berbagai teknik seleksi fitur dan algoritma machine learning dalam konteks prediksi diabetes mellitus. Studi ini menyoroti bahwa metode berbasis metaheuristik memiliki keunggulan dalam menjelajahi ruang solusi secara lebih menyeluruh dibandingkan pendekatan deterministik seperti SFS dan SBS. Selain itu, penelitian tersebut menggarisbawahi bahwa seleksi fitur yang tepat tidak hanya meningkatkan

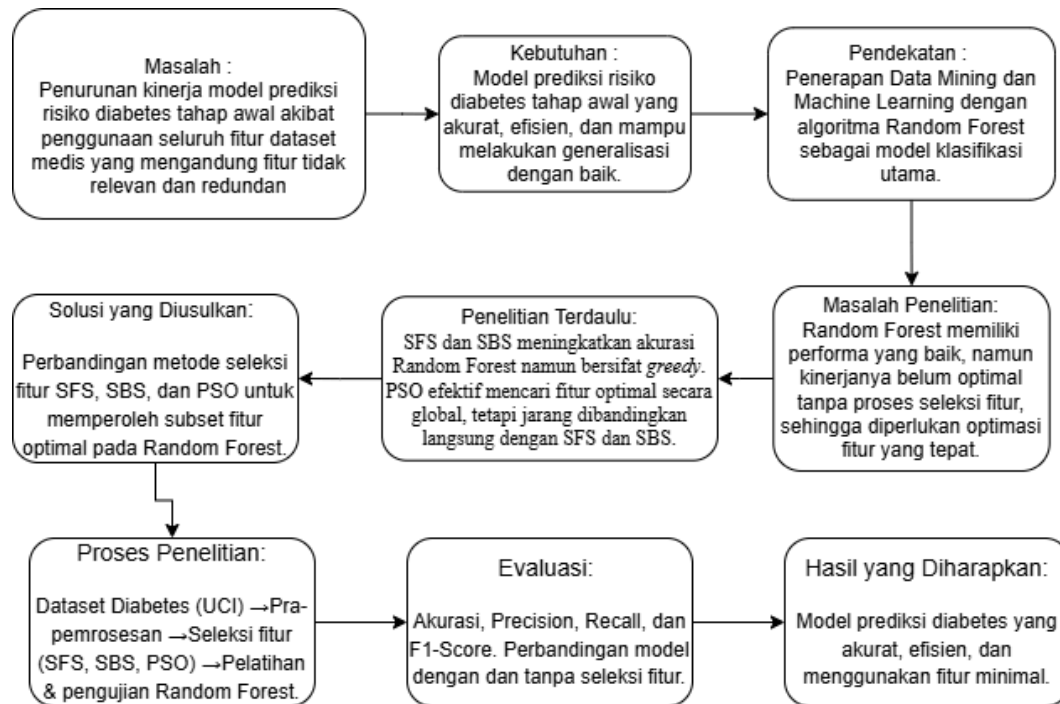
akurasi, tetapi juga memperbaiki interpretabilitas model yang sangat penting dalam aplikasi klinis (Sheta et al.2024).

Menganalisis berbagai strategi klasifikasi dan seleksi fitur pada beberapa dataset diabetes termasuk Pima Indians Diabetes Dataset. Penelitian ini menemukan bahwa performa model sangat bergantung pada karakteristik spesifik dataset yang digunakan, sehingga evaluasi pada beberapa dataset secara bersamaan memberikan gambaran yang jauh lebih komprehensif mengenai kemampuan generalisasi suatu metode. Temuan ini memperkuat alasan penggunaan dua dataset dalam penelitian yang dilakukan penulis (Kaliappan et al.2024) .

Mendemonstrasikan bahwa integrasi PSO sebagai mekanisme seleksi fitur pada model klasifikasi diabetes menghasilkan peningkatan akurasi yang signifikan dibandingkan metode wrapper konvensional. Kemampuan PSO dalam menghindari jebakan optima lokal menjadikannya kandidat yang kuat untuk mengungguli SFS dan SBS pada dataset dengan ruang pencarian fitur yang lebih kompleks (Rahman et al.2025).

Berdasarkan telaah terhadap penelitian-penelitian terdahulu di atas, dapat diidentifikasi bahwa kebaruan penelitian ini terletak pada dua aspek utama. Pertama, penambahan metode PSO sebagai alternatif metaheuristik yang belum pernah dibandingkan secara langsung dengan SFS dan SBS menggunakan Random Forest pada Early Stage Diabetes UCI Dataset dalam satu kerangka penelitian yang sama. Kedua, perluasan evaluasi ke Pima Indians Diabetes Dataset yang memiliki karakteristik berbeda sehingga menghasilkan analisis komparatif yang lebih generalisatif dan komprehensif.

C. Kerangka Berpikir



Gambar 2.1. Kerangka Berpikir

Kerangka berpikir penelitian ini dibangun atas dasar permasalahan nyata yang ditemui dalam pengembangan model prediksi risiko diabetes berbasis machine learning, yaitu belum optimalnya kinerja model akibat penggunaan seluruh fitur tanpa seleksi yang memadai. Variabel independen dalam penelitian ini adalah metode seleksi fitur (SFS, SBS, dan PSO) yang diterapkan secara terpisah pada dua dataset diabetes dengan karakteristik yang berbeda. Sementara itu, variabel dependen adalah kinerja algoritma Random Forest yang diukur menggunakan metrik akurasi, precision, recall, dan F1-score.

Hubungan kausalitas yang dianalisis dalam penelitian ini adalah pengaruh penerapan metode seleksi fitur terhadap kinerja Random Forest. Melalui kerangka ini, diharapkan dapat diperoleh pemahaman yang lebih mendalam mengenai

kontribusi masing-masing metode dalam meningkatkan performa prediksi, sekaligus mengidentifikasi metode yang paling efektif untuk diterapkan pada kedua jenis dataset yang digunakan.

D. Hipotesis

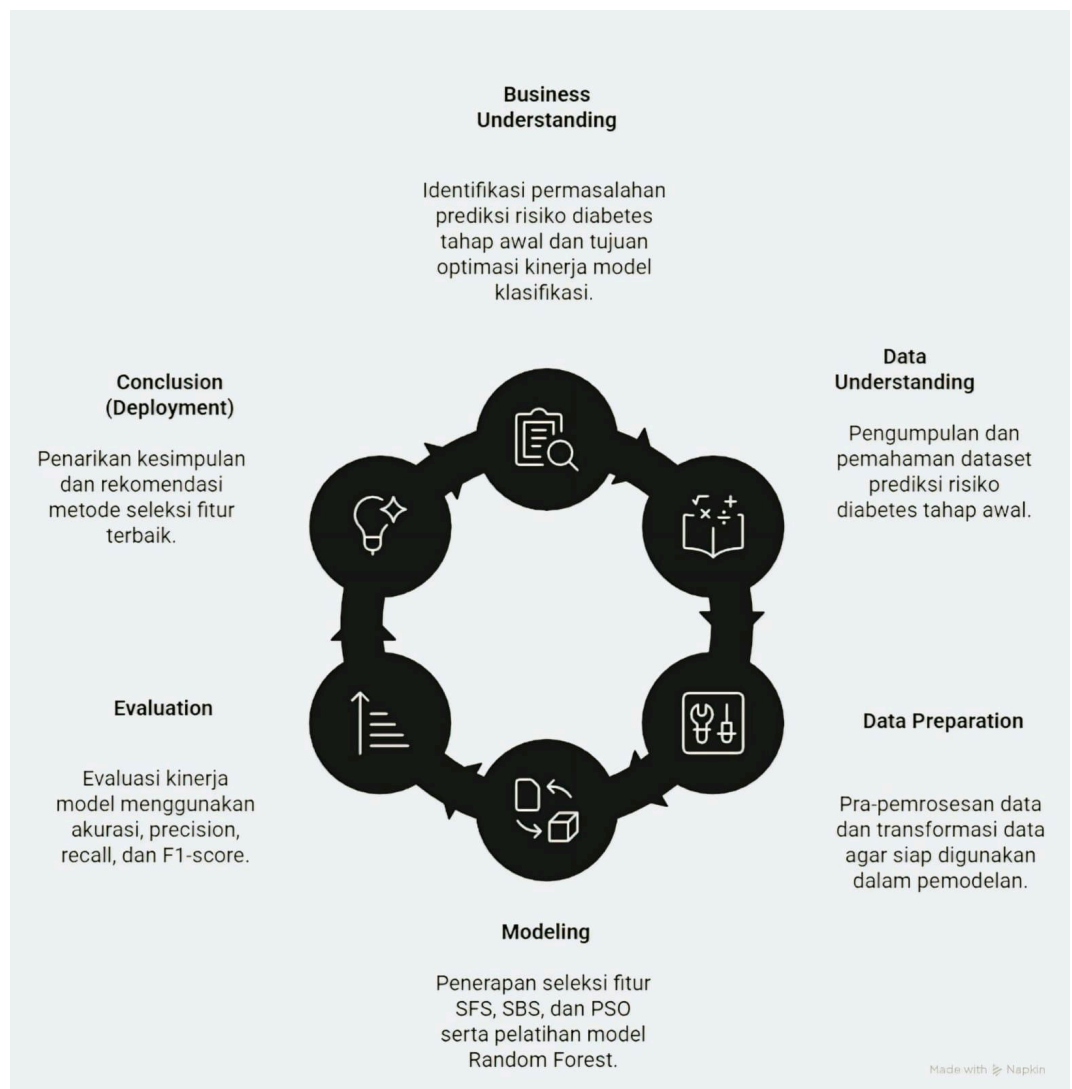
Tabel 2.2 Hipotesis

No	Null Hypothesis (H_0)	Alternative Hypothesis (H_1)
1	H_{01} : Penerapan SFS tidak meningkatkan kinerja Random Forest dalam prediksi diabetes.	H_{11} : Penerapan SFS meningkatkan kinerja Random Forest dalam prediksi diabetes.
2	H_{02} : Penerapan SBS tidak meningkatkan kinerja Random Forest dalam prediksi diabetes.	H_{12} : Penerapan SBS meningkatkan kinerja Random Forest dalam prediksi diabetes.
3	H_{03} : Penerapan PSO tidak meningkatkan kinerja Random Forest dalam prediksi diabetes.	H_{13} : Penerapan PSO meningkatkan kinerja Random Forest dalam prediksi diabetes.
4	H_{04} : Tidak ada perbedaan signifikan antara kinerja model dengan dan tanpa seleksi fitur pada kedua dataset.	H_{14} : Terdapat perbedaan signifikan antara kinerja model dengan dan tanpa seleksi fitur pada kedua dataset.
5	H_{05} : Tidak ada perbedaan performa yang berarti antar ketiga metode seleksi fitur pada kedua dataset.	H_{15} : Terdapat perbedaan performa yang berarti antar ketiga metode seleksi fitur pada kedua dataset.

BAB III

METODE PENELITIAN

Pendekatan metodologis yang digunakan dalam penelitian ini mengacu pada kerangka kerja CRISP-DM (Cross Industry Standard Process for Data Mining) yang merupakan standar industri untuk proyek data mining. Kerangka ini dipilih karena bersifat iteratif, fleksibel, dan mampu mengakomodasi proses pengembangan model dari tahap pemahaman masalah hingga penerapan. Alur penelitian berdasarkan CRISP-DM disajikan pada Gambar 3.1.



Gambar 3.1 Alur Penelitian CRISP-DM

A. Business Understanding (Pemahaman Bisnis)

1. Latar Belakang Masalah

Tingginya angka penderita diabetes dan keterbatasan sistem deteksi dini yang ada mendorong kebutuhan akan model prediksi berbasis machine learning yang akurat. Dalam proses pengembangan model tersebut, kualitas fitur yang digunakan menjadi faktor penentu yang sering diabaikan. Keberadaan fitur yang tidak relevan dapat memperburuk performa model secara tidak langsung melalui peningkatan noise dan kompleksitas yang tidak perlu. Oleh karena itu, penelitian ini menempatkan optimasi seleksi fitur sebagai inti permasalahan yang ingin diselesaikan (Nugroho et al., 2021).

2. Tujuan

Tujuan utama penelitian ini adalah mengidentifikasi metode seleksi fitur yang paling efektif dalam meningkatkan kinerja Random Forest pada dua dataset prediksi diabetes. Keberhasilan dicapai apabila model dengan seleksi fitur menghasilkan nilai metrik evaluasi yang lebih tinggi dibandingkan baseline tanpa seleksi fitur, dengan jumlah fitur yang lebih sedikit sebagai indikator efisiensi tambahan.

B. Data Understanding (Pemahaman Data)

1. Sumber dan Pengumpulan Data

Data yang digunakan dalam penelitian ini bersumber dari dua repository publik yang berbeda. Dataset pertama (Early Stage Diabetes UCI) diperoleh dari UCI Machine Learning Repository dalam format CSV. Dataset kedua (Pima Indians Diabetes) juga tersedia di UCI Machine Learning Repository. Kedua dataset ini dipilih karena ketersediaannya secara publik, relevansinya dengan topik penelitian, serta perbedaan karakteristik yang memungkinkan perbandingan komprehensif.

2. Deskripsi Dataset

- a) Dataset 1: Early Stage Diabetes Risk Prediction Dataset

Dataset ini terdiri dari 520 data dengan 17 atribut, di mana 16 atribut berfungsi sebagai fitur prediktor dan 1 atribut sebagai variabel target. Fitur prediktor mencakup satu fitur numerik (Age) dan 15 fitur kategorikal biner (Yes/No). Deskripsi lengkap fitur ditampilkan pada Tabel 3.1 berikut.

Tabel 3.1 Deskripsi Fitur Dataset Early Stage Diabetes Risk Prediction UCI

No.	Nama Fitur	Tipe Data	Keterangan
1	Age	Numeric	Usia pasien (tahun)
2	Gender	Nominal (Male/Female)	Jenis kelamin
3	Polyuria	Nominal (Yes/No)	Frekuensi buang air kecil berlebih
4	Polydipsia	Nominal (Yes/No)	Rasa haus yang berlebihan
5	Sudden Weight Loss	Nominal (Yes/No)	Penurunan berat badan mendadak
6	Weakness	Nominal (Yes/No)	Kelemahan fisik
7	Polyphagia	Nominal (Yes/No)	Nafsu makan berlebihan
8	Genital Thrush	Nominal (Yes/No)	Infeksi jamur pada area genital
9	Visual Blurring	Nominal (Yes/No)	Gangguan ketajaman penglihatan
10	Itching	Nominal (Yes/No)	Sensasi gatal pada kulit
11	Irritability	Nominal (Yes/No)	Peningkatan iritabilitas emosional

12	Delayed Healing	Nominal (Yes/No)	Proses penyembuhan luka yang lambat
13	Partial Paresis	Nominal (Yes/No)	Kelemahan otot sebagian
14	Muscle Stiffness	Nominal (Yes/No)	Kekakuan pada otot
15	Alopecia	Nominal (Yes/No)	Kerontokan rambut
16	Obesity	Nominal (Yes/No)	Kondisi kelebihan berat badan
17	Class	Nominal (Positive/Negative)	Label target klasifikasi diabetes

b) Dataset 2: Pima Indians Diabetes Dataset

Dataset ini terdiri dari 768 data dengan 9 atribut, seluruhnya bersifat numerik. Delapan atribut pertama merupakan fitur prediktor yang merepresentasikan kondisi klinis dan fisik pasien, sedangkan atribut terakhir adalah variabel target biner. Deskripsi lengkap fitur ditampilkan pada Tabel 3.2 berikut.

Tabel 3.2 Deskripsi Fitur Dataset Pima Indians Diabetes

No.	Nama Fitur	Type Data	Keterangan
1	Pregnancies	Numeric	Jumlah kehamilan yang pernah dialami
2	Glucose	Numeric	Kadar glukosa plasma 2 jam setelah uji toleransi
3	BloodPressure	Numeric	Tekanan darah diastolik (mm Hg)

4	SkinThickness	Numeric	Ketebalan lipatan kulit tricep (mm)
5	Insulin	Numeric	Kadar insulin serum 2 jam (mu U/ml)
6	BMI	Numeric	Indeks massa tubuh (kg/m ²)
7	DiabetesPedigreeFunction	Numeric	Skor fungsi silsilah riwayat diabetes keluarga
8	Age	Numeric	Usia pasien (tahun)
9	Outcome	Numeric (0/1)	Label target (1=Positif Diabetes, 0=Negatif)

C. Data Preparation (Persiapan Data)

1. Pembersihan Data

Tahap pembersihan data bertujuan memastikan integritas data sebelum diproses lebih lanjut. Pada Dataset 1, pengecekan menunjukkan tidak adanya nilai kosong (missing value) pada keseluruhan 520 data sehingga tidak diperlukan penanganan khusus. Pada Dataset 2, ditemukan nilai nol yang tidak logis secara medis pada beberapa fitur seperti Glucose, BloodPressure, SkinThickness, Insulin, dan BMI. Nilai nol pada fitur-fitur tersebut diindikasikan sebagai data yang tidak tercatat, bukan nilai nol yang sesungguhnya, sehingga dilakukan penanganan melalui imputasi dengan nilai median kelompok berdasarkan variabel target.

2. Transformasi Data

Transformasi data diperlukan untuk menyesuaikan format data dengan kebutuhan algoritma yang digunakan. Pada Dataset 1, seluruh fitur kategorikal

diubah menjadi representasi numerik menggunakan label encoding. Nilai 'Yes' dikonversi menjadi 1, 'No' menjadi 0, 'Male' menjadi 1, 'Female' menjadi 0, serta label 'Positive' menjadi 1 dan 'Negative' menjadi 0 pada variabel target. Pada Dataset 2 yang sudah bersifat numerik, dilakukan normalisasi menggunakan metode Min-Max Scaling dengan formula berikut:

$$x_{norm} = (x - x_{min}) / (x_{max} - x_{min}) \quad \dots(3.1)$$

di mana x adalah nilai asli, x_{min} dan x_{max} masing-masing adalah nilai minimum dan maksimum dari fitur tersebut. Normalisasi ini dilakukan khususnya untuk mendukung proses seleksi fitur berbasis PSO agar pergerakan partikel tidak didominasi oleh fitur dengan skala nilai yang jauh lebih besar.

3. Pembagian Data

Seluruh data pada masing-masing dataset dipisahkan menjadi matriks fitur (X) dan vektor target (y). Skema validasi yang digunakan adalah 10-fold cross validation, di mana data dibagi menjadi 10 bagian proporsional. Pada setiap iterasi, 9 bagian digunakan sebagai data latih dan 1 bagian sisanya sebagai data uji, sehingga setiap data pernah digunakan sebagai data uji tepat satu kali.

D. Modeling

1. Skenario Pengujian

Pengujian dilakukan dalam empat skenario pada masing-masing dataset: (1) Random Forest tanpa seleksi fitur sebagai baseline, (2) Random Forest dengan SFS, (3) Random Forest dengan SBS, dan (4) Random Forest dengan PSO. Seluruh skenario menggunakan parameter Random Forest yang identik untuk memastikan perbandingan yang adil.

2. Parameter Model

Parameter Random Forest yang digunakan adalah $n_estimators = 100$ pohon dan $random_state = 42$ untuk reproduibilitas hasil. Untuk metode PSO, parameter yang digunakan meliputi jumlah partikel sebanyak 30, jumlah iterasi maksimum sebanyak 100 iterasi, bobot inersia $w = 0,9$ yang menurun secara linier hingga 0,4, serta koefisien percepatan $c1 = c2 = 2$. Nilai fitness setiap partikel

dihitung berdasarkan akurasi rata-rata 10-fold cross validation pada subset fitur yang direpresentasikan oleh posisi partikel.

3. Implementasi

Seluruh implementasi dilakukan menggunakan Python dengan memanfaatkan library scikit-learn untuk algoritma Random Forest dan cross validation, mlxtend untuk implementasi SFS dan SBS, serta pyswarm untuk implementasi PSO. Pengolahan data menggunakan pandas dan numpy, sedangkan visualisasi hasil menggunakan matplotlib.

E. Evaluation (Evaluasi)

Evaluasi dilakukan dengan menghitung nilai akurasi, precision, recall, dan F1-score dari setiap skenario pengujian menggunakan rumus yang telah dijelaskan pada Persamaan 2.1 hingga 2.4. Selain itu, jumlah fitur yang terpilih pada setiap metode juga dicatat sebagai indikator efisiensi model. Perbandingan dilakukan antar skenario pada masing-masing dataset serta antar kedua dataset untuk menilai konsistensi metode.

F. Deployment

Hasil penelitian ini ditujukan sebagai referensi ilmiah dan dasar pengembangan sistem pendukung keputusan prediksi risiko diabetes. Model terbaik yang diperoleh dari penelitian ini berpotensi untuk dikembangkan lebih lanjut menjadi aplikasi berbasis web yang dapat dimanfaatkan secara praktis oleh tenaga medis dalam proses skrining awal pasien.

Tabel 3.3 Perbandingan Metode Seleksi Fitur

Metode	Strategi Pencarian	Kelebihan	Keterbatasan
Tanpa FS	Menggunakan seluruh fitur	Sederhana, tanpa overhead komputasi tambahan	Rentan noise; fitur tidak relevan menurunkan kinerja
SFS	Penambahan fitur bertahap dari himpunan kosong	Mudah diimplementasi; efisien secara komputasi	Greedy; fitur terpilih tidak dapat dihapus kembali
SBS	Eliminasi fitur bertahap dari himpunan lengkap	Mempertimbangkan interaksi fitur sejak awal	Komputasi lebih berat pada dataset berdimensi tinggi
PSO	Eksplorasi populasi partikel secara paralel	Pencarian global; menghindari optima lokal	Parameter sensitif; memerlukan tuning yang cermat

18

BAB IV HASIL DAN PEMBAHASAN

A. Data Hasil Penelitian

1. Hasil eksplorasi data

a. Data Awal Penelitian

Pada tahap awal penelitian, dilakukan peninjauan terhadap dataset yang digunakan. Dataset yang digunakan merupakan Early Stage Diabetes Risk Prediction Dataset yang terdiri dari 520 data dengan 17 atribut. Atribut tersebut mencakup satu fitur numerik yaitu Age dan beberapa fitur kategorikal yang berkaitan dengan gejala diabetes.

Berdasarkan Tabel 4.1, dapat dilihat bahwa data yang digunakan memiliki struktur yang lengkap dan representatif untuk digunakan dalam proses pemodelan. Seluruh atribut memiliki nilai yang terisi tanpa adanya data kosong, sehingga dataset dapat langsung digunakan tanpa perlu dilakukan penanganan terhadap missing value.

Table 4.1 Data Awal Uci

	Age	Gender	Polyuria	Polydipsia	S. W. Loss	Weakness	Polyhagia	Genital Thrush	Visual Blurring	Itching	Irritability	Delayed healing	Partial Paresis	Muscle Stiffness	Alopecia	Obesity	Class
3	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Positive
1	58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No	Positive
2	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	No	No	Positive
3	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	Yes	No	Positive
4	60	Male	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive

Pada Tabel 4.1. ditampilkan lima baris pertama data awal yang digunakan dalam penelitian. Terlihat bahwa setiap baris merepresentasikan satu individu dengan kombinasi nilai atribut yang berbeda. Dataset ini juga memiliki variabel target yaitu *class* yang menunjukkan status risiko diabetes pada masing-masing data. Secara umum, struktur data yang

ditampilkan sudah lengkap dan siap digunakan untuk proses analisis lebih lanjut. Tampilan awal ini bertujuan untuk memberikan gambaran umum mengenai struktur data yang akan dianalisis lebih lanjut.

b. Tipe Data

Tahap identifikasi tipe data dilakukan untuk mengetahui karakteristik masing-masing atribut yang terdapat pada dataset. Proses ini penting karena setiap tipe data akan menentukan metode pengolahan yang digunakan pada tahap preprocessing dan modeling. Selain itu, informasi dataset juga digunakan untuk melihat jumlah data, kelengkapan atribut, serta struktur keseluruhan dataset sebelum dilakukan proses transformasi data.

Table 4.2 Tipe Dan Informasi Data

No.	Variabel	Non-Null Count	Tipe Data	Keterangan
0	Age	520 non-null	int64	Usia pasien
1	Gender	520 non-null	object	Jenis kelamin pasien
2	Polyuria	520 non-null	object	Gejala sering buang air kecil
3	Polydipsia	520 non-null	Object	Gejala sering haus
4	Sudden weight loss	520 non-null	Object	Penurunan berat badan secara tiba-tiba
5	Weakness	520 non-null	Object	Kondisi tubuh terasa lemah
6	Polyphagia	520 non-null	object	Rasa lapar berlebihan
7	Genital thrush	520 non-null	Object	Infeksi jamur pada area genital
8	visual blurring	520 non-null	object	Penglihatan kabur
9	Itching	520 non-null	object	Rasa gatal pada tubuh
10	Irritability	520 non-null	object	Mudah marah

11	delayed healing	520 non-null	object	Luka sulit sembuh
12	partial paresis	520 non-null	object	Kelumpuhan sebagian
13	muscle stiffness	520 non-null	object	Kekakuan otot
14	Alopecia	520 non-null	object	Kerontokan rambut
15	Obesity	520 non-null	object	Obesitas
16	class	520 non-null	object	Label target diabetes

Berdasarkan hasil identifikasi tipe data, dapat diketahui bahwa dataset terdiri dari 17 atribut dengan total 520 data. Fitur *Age* memiliki tipe data numerik (*int64*), sedangkan 16 atribut lainnya memiliki tipe data kategorikal (*object*), termasuk variabel target *class*. Hasil informasi dataset juga menunjukkan bahwa seluruh atribut memiliki jumlah data lengkap sebanyak 520 *non-null entries*, sehingga tidak ditemukan adanya nilai kosong (*missing value*) pada dataset. Selain itu, ukuran memori yang digunakan relatif kecil yaitu sekitar 69.2 KB, sehingga proses pengolahan data dapat dilakukan dengan cukup efisien. Struktur data yang konsisten ini menunjukkan bahwa dataset telah siap digunakan pada tahap transformasi dan pemodelan tanpa memerlukan penanganan data yang kompleks.

c. Deskripsi Statistik Data

Berdasarkan hasil analisis deskriptif, fitur *Age* memiliki nilai rata-rata sekitar 48 tahun dengan rentang usia antara 16 hingga 90 tahun. Hal ini menunjukkan bahwa dataset mencakup berbagai kelompok usia yang cukup luas. Sementara itu, fitur lainnya bersifat kategorikal dengan nilai dominan berupa “Yes” dan “No”.

Dari hasil deskripsi data, terlihat bahwa sebagian besar fitur memiliki dua kategori yang seimbang, meskipun terdapat beberapa atribut dengan distribusi yang lebih dominan pada salah satu kategori.

Informasi ini penting karena dapat mempengaruhi kemampuan model dalam membedakan kelas.

Tabel 4.2 Deskripsi Dataset

	Age	Gender	Poly uria	Poly dipsia	sudden weight loss	Wea k ness	Poly hagia	Genital thrush	Visual Blurring	Itching	Irrita bility	Delayed healing	Partial paresis	M
count	520.000000	520	520	520	520	520	520	520	520	520	520	520	520	520
Unique	NaN	2	2	2	2	2	2	2	2	2	2	2	2	2
Top	NaN	Male	No	No	No	Yes	No	No	No	No	No	No	No	No
Freq	NaN	328	262	287	303	305	283	404	287	267	394	281	296	328
Mean	48.028846	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Std	12.151466	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Min	16.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
25%	39.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
50%	47.500000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
75%	57.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Max	90.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Berdasarkan Tabel 4.2 deskripsi data yang ditampilkan, dapat diketahui bahwa dataset terdiri dari 520 data pada setiap atribut yang digunakan dalam penelitian. Fitur *Age* memiliki nilai rata-rata sekitar 48 tahun dengan rentang usia minimum 16 tahun dan maksimum 90 tahun, serta penyebaran data yang cukup merata yang ditunjukkan oleh nilai standar deviasi sebesar 12.15. Sementara itu, atribut lainnya merupakan data kategorikal yang masing-masing memiliki dua nilai unik, seperti “Yes” dan “No” atau “Male” dan “Female”. Nilai yang paling sering muncul (*top*) pada beberapa atribut menunjukkan kecenderungan dominasi kategori tertentu, misalnya jenis kelamin

laki-laki dan kelas *Positive* pada variabel target. Secara keseluruhan, deskripsi ini menunjukkan bahwa dataset memiliki struktur yang konsisten dan cukup representatif untuk digunakan dalam proses pemodelan.

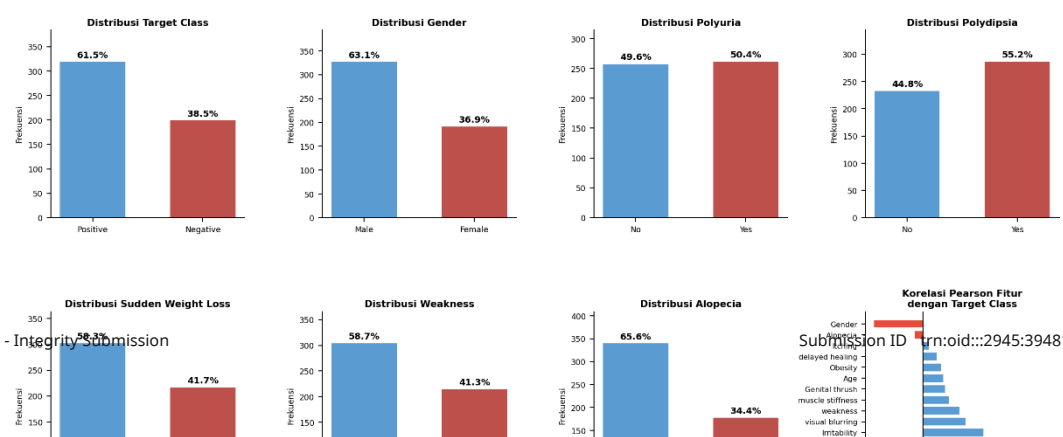
d. Distribusi Data

Distribusi data ditampilkan dalam bentuk grafik untuk memberikan gambaran visual mengenai sebaran data pada setiap atribut. Berdasarkan grafik distribusi class, terlihat bahwa jumlah data pada kelas *Positive* lebih banyak dibandingkan kelas *Negative*. Namun demikian, perbedaan jumlah tersebut tidak terlalu signifikan sehingga dataset masih tergolong cukup seimbang.

Pada grafik distribusi gender, terlihat bahwa jumlah responden laki-laki lebih banyak dibandingkan perempuan. Selain itu, pada fitur seperti Polydipsia, sudden weight loss, dan Alopecia, terlihat bahwa kategori “No” lebih dominan dibandingkan “Yes”. Hal ini menunjukkan bahwa tidak semua pasien memiliki gejala yang sama, sehingga variasi data menjadi faktor penting dalam proses klasifikasi.

Hasil eksplorasi data pada Dataset 1 menunjukkan bahwa distribusi kelas target cukup seimbang dengan proporsi 61,5% data positif (320 kasus) dan 38,5% data negatif (200 kasus). Fitur-fitur seperti Polyuria, Polydipsia, dan Partial Paresis memiliki korelasi tertinggi dengan variabel target berdasarkan koefisien korelasi Pearson, yang mengindikasikan bahwa fitur-fitur tersebut memiliki potensi diskriminatif yang paling kuat dalam membedakan kelas positif dan negatif.

Eksplorasi Data Awal (EDA) — Dataset Early Stage Diabetes Risk Prediction

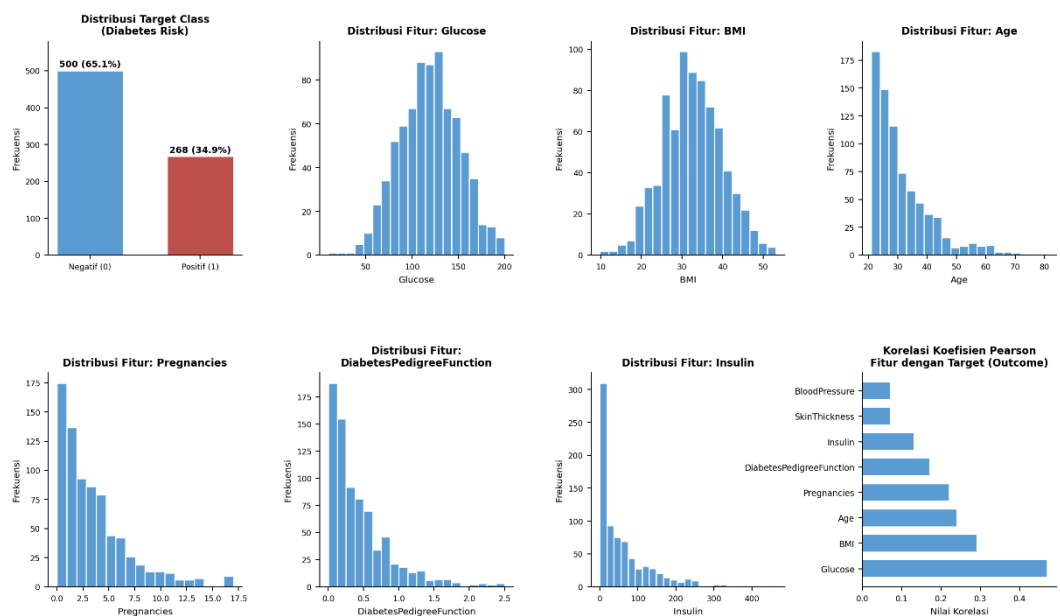


Gambar 4.1 Distribusi Fitur dan Korelasi Pearson — Dataset Early Stage Diabetes UCI

b) Dataset 2: Pima Indians Diabetes Dataset

Hasil eksplorasi pada Dataset 2 menunjukkan distribusi kelas yang tidak seimbang, yaitu 65,1% data negatif (500 kasus) dan 34,9% data positif (268 kasus). Fitur *Glucose* memiliki korelasi tertinggi terhadap variabel target (*Outcome*), diikuti oleh *BMI* dan *Age*. Ketidakseimbangan distribusi kelas pada dataset ini menjadi salah satu faktor yang mempengaruhi tingkat akurasi model secara keseluruhan. Kondisi tersebut dapat menyebabkan model lebih cenderung memprediksi kelas mayoritas dibandingkan kelas minoritas. Selain itu, tingginya korelasi pada fitur *Glucose*, *BMI*, dan *Age* menunjukkan bahwa ketiga fitur tersebut memiliki peran penting dalam membedakan individu yang berisiko diabetes dan yang tidak berisiko diabetes. Oleh karena itu, fitur-fitur tersebut berpotensi menjadi atribut utama yang dipertahankan pada proses seleksi fitur selanjutnya.

Eksplorasi Data Awal (EDA) — Dataset Pima Indians Diabetes



Gambar 4.2 Distribusi Fitur dan Korelasi Pearson — Dataset Pima Indians Diabetes

2. Proses Seleksi Fitur Sequential Forward Selection (SFS)

Metode SFS memulai proses seleksi dari himpunan fitur kosong dan menambahkan fitur terbaik satu per satu pada setiap iterasi berdasarkan akurasi cross validation yang dihasilkan.

a) Dataset 1 — SFS

Tabel 4.1 berikut menyajikan statistik proses seleksi fitur SFS step-by-step pada Dataset 1.

Tabel Proses SFS — Statistik Seleksi Fitur Step-by-Step
Dataset: Early Stage Diabetes Risk Prediction | 10-Fold Stratified CV

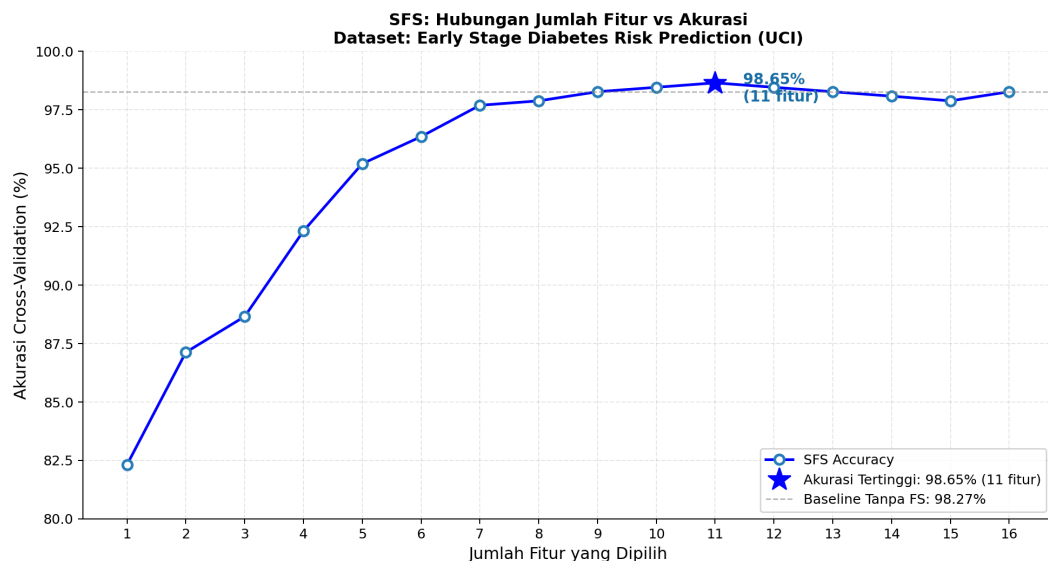
Step	Feature Idx	Avg Score	Feature Names	CI Bound	Std Dev	Std Err
1	{2}	0.8231	Polyuria	0.040297	0.054257	0.018086
2	{1, 2}	0.8712	Gender, Polyuria	0.035015	0.047145	0.015715
3	{0, 1, 2}	0.8865	Age, Gender, Polyuria	0.038564	0.051923	0.017308
4	{0, 1, 2, 10}	0.9231	Age, Gender, Polyuria, Irritability	0.029271	0.039411	0.013137
5	{0, 1, 2, 10, 14}	0.9519	Age, Gender, Polyuria, Irritability, ...	0.015969	0.021501	0.007167
6	{0, 1, 2, 3, 10, 14}	0.9635	Age, Gender, Polyuria, Polydipsia, ...	0.020549	0.027802	0.009267
7	{0, 1, 2, 3, 9, 10, 14}	0.9769	Age, Gender, Polyuria, Polydipsia, ...	0.015383	0.020712	0.006904
8	{0, 1, 2, 3, 9, 10, 13, 14}	0.9788	Age, Gender, Polyuria, Polydipsia, ...	0.013474	0.018142	0.006047
9	{0, 1, 2, 3, 7, 9, 10, 13, 14}	0.9865	Age, Gender, Polyuria, Polydipsia, ...	0.012855	0.017308	0.005769
10	{0, 1, 2, 3, 7, 8, 9, 10, 13, 14}	0.9865	Age, Gender, Polyuria, Polydipsia, ...	0.009474	0.012756	0.004252
11	{0, 1, 2, 3, 7, 8, 9, 10, 11, 13, 14}	0.9923	Age, Gender, Polyuria, Polydipsia, ...	0.006997	0.009421	0.003140
12	{0, 1, 2, 3, 4, 7, 8, 9, 10, 13, 14}	0.9885	Age, Gender, Polyuria, Polydipsia, ...	0.009474	0.012756	0.004252
13	{0, 1, 2, 3, 4, 7, 8, 9, 10, 13, 14, 15}	0.9865	Age, Gender, Polyuria, Polydipsia, ...	0.011155	0.015020	0.005007
14	{0, 1, 2, 3, 4, 6, 7, 8, 9, 10, 12, 13, 14}	0.9885	Age, Gender, Polyuria, Polydipsia, ...	0.011426	0.015385	0.005128
15	{0, 1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14}	0.9904	Age, Gender, Polyuria, Polydipsia, ...	0.011515	0.015504	0.005168
16	{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15}	0.9871	Age, Gender, Polyuria, Polydipsia, ...	0.013474	0.018142	0.006047

Tabel 4.1 Proses Seleksi Fitur SFS Step-by-Step — Dataset Early Stage Diabetes UCI

Tabel 4.1 menunjukkan hasil proses seleksi fitur menggunakan metode Sequential Forward Selection (SFS) pada Dataset 1 yang dilakukan secara bertahap dengan menambahkan satu fitur terbaik pada setiap langkah. Pada tahap awal, fitur *Polyuria* dipilih sebagai fitur pertama dan menghasilkan akurasi sebesar 82,31%. Seiring bertambahnya jumlah fitur yang dimasukkan ke dalam model, nilai akurasi mengalami peningkatan secara konsisten. Hal ini menunjukkan bahwa fitur-fitur yang dipilih memiliki hubungan yang relevan terhadap proses klasifikasi risiko diabetes. Peningkatan performa model terlihat hingga langkah ke-11, di mana kombinasi 11 fitur terpilih mampu menghasilkan akurasi tertinggi sebesar 98,65%. Hasil tersebut mengindikasikan bahwa sebagian besar informasi penting dalam dataset telah terwakili oleh sebelas fitur tersebut.

Setelah melewati titik optimal, penambahan fitur pada langkah ke-12 sampai langkah ke-16 tidak memberikan peningkatan performa yang berarti. Bahkan, pada beberapa tahap terjadi penurunan nilai akurasi yang menunjukkan adanya fitur dengan kontribusi rendah terhadap kemampuan prediksi model. Kondisi ini mengindikasikan bahwa semakin banyak fitur yang digunakan tidak selalu menghasilkan performa yang lebih baik. Adanya fitur yang kurang relevan dapat menambah kompleksitas model tanpa memberikan informasi tambahan yang signifikan. Oleh karena itu, kombinasi 11 fitur yang diperoleh pada langkah ke-11 dapat dianggap sebagai subset fitur terbaik karena mampu menghasilkan akurasi tertinggi dengan jumlah fitur yang lebih efisien dibandingkan penggunaan seluruh fitur yang tersedia.

Gambar 4.3 menunjukkan kurva hubungan antara jumlah fitur yang dipilih dengan akurasi cross validation pada metode SFS Dataset 1.



Gambar 4.3 Kurva Akurasi vs Jumlah Fitur SFS — Dataset Early Stage Diabetes UCI

Pola kurva memperlihatkan kenaikan akurasi yang signifikan dari fitur ke-1 hingga ke-11, kemudian mengalami plateau dan penurunan pada penambahan fitur berikutnya. Kondisi tersebut menunjukkan bahwa proses penambahan fitur pada tahap awal mampu meningkatkan kemampuan model dalam mengenali pola data secara lebih akurat. Kenaikan yang terjadi secara

bertahap mengindikasikan bahwa setiap fitur yang terpilih memberikan kontribusi positif terhadap performa klasifikasi. Titik puncak kurva yang berada pada penggunaan 11 fitur menunjukkan kombinasi fitur yang paling optimal dalam menghasilkan akurasi tertinggi. Setelah titik tersebut, penambahan fitur tidak lagi memberikan informasi baru yang signifikan sehingga peningkatan performa tidak terjadi. Bahkan, penurunan akurasi pada beberapa titik mengindikasikan adanya fitur yang berpotensi menimbulkan redundansi atau menambah kompleksitas model tanpa meningkatkan kemampuan prediksi secara nyata.

b) Dataset 2 — SFS

Tabel 4.2 berikut menyajikan statistik proses seleksi fitur SFS step-by-step pada Dataset 2.

Tabel Proses SFS — Statistik Seleksi Fitur Step-by-Step
Dataset: Pima Indians Diabetes | 10-Fold Stratified CV

Step	Feature Idx	Avg Score	Feature Names	CI Bound	Std Dev	Std Err
1	(1,)	0.7032	Glucose	0.032899	0.044285	0.014765
2	(1, 5)	0.7240	Glucose, BMI	0.025462	0.034271	0.011424
3	(1, 5, 6)	0.7366	Glucose, BMI, DiabetesPedigreeFunction	0.012461	0.016771	0.005590
4	(1, 5, 6, 7)	0.7444	Glucose, BMI, DiabetesPedigreeFunction, Age	0.022044	0.029658	0.009886
5	(0, 1, 5, 6, 7)	0.7578	Pregnancies, Glucose, BMI, ...	0.021308	0.028679	0.009560
6	(0, 1, 2, 5, 6, 7)	0.7747	Pregnancies, Glucose, BloodPressure, BMI, ...	0.021344	0.028726	0.009575
7	(0, 1, 2, 4, 5, 6, 7)	0.7656	Pregnancies, Glucose, BloodPressure, Insulin, ...	0.022046	0.029659	0.009887
8	(0, 1, 2, 3, 4, 5, 6, 7)	0.7629	Pregnancies, Glucose, BloodPressure, SkinThickness, ...	0.024214	0.032588	0.010863

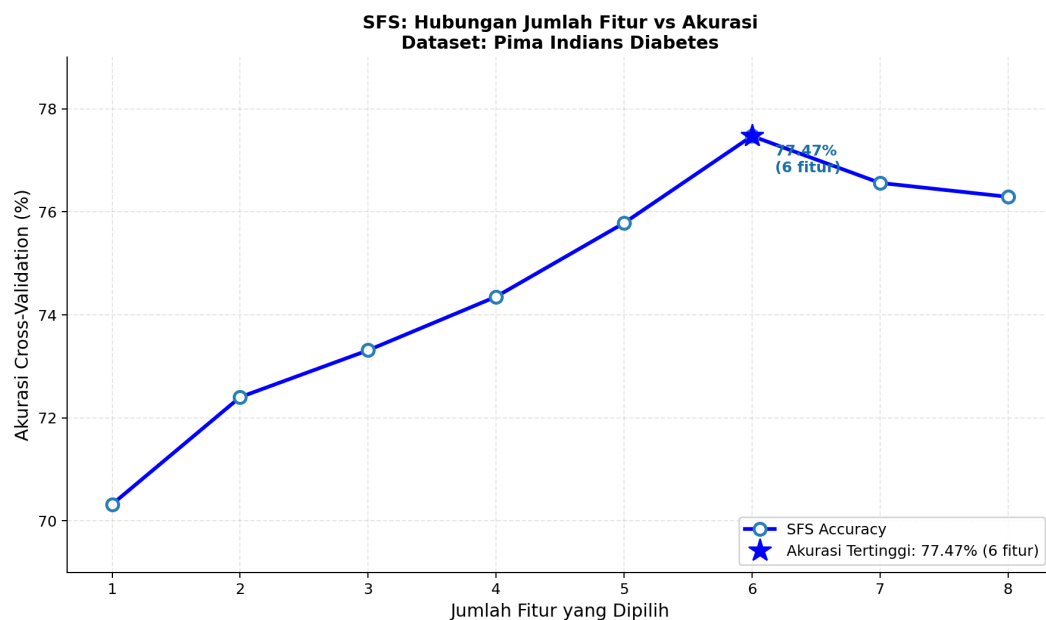
* Akurasi Tertinggi dengan 6 fitur terpilih (Akurasi: 77.47%)

Tabel 4.2 Proses Seleksi Fitur SFS Step-by-Step — Dataset Pima Indians Diabetes

Pada Dataset 2, proses SFS dimulai dari fitur *Glucose* dengan akurasi awal sebesar 70,32%. Akurasi kemudian mengalami peningkatan secara bertahap pada setiap penambahan fitur yang dipilih selama proses seleksi berlangsung. Peningkatan tersebut menunjukkan bahwa fitur-fitur yang ditambahkan mampu memberikan informasi tambahan yang relevan bagi model dalam melakukan klasifikasi. Akurasi tertinggi diperoleh pada langkah ke-6 dengan nilai sebesar 77,47% menggunakan enam fitur terpilih, yaitu *Pregnancies*, *Glucose*, *BloodPressure*, *Insulin*, *BMI*, dan *DiabetesPedigreeFunction*. Hasil ini mengindikasikan bahwa kombinasi keenam fitur tersebut telah mampu merepresentasikan karakteristik data yang paling berpengaruh terhadap prediksi diabetes. Setelah titik tersebut, penambahan fitur lain tidak memberikan

peningkatan akurasi yang signifikan. Bahkan, beberapa fitur tambahan cenderung menghasilkan perubahan performa yang sangat kecil sehingga kontribusinya terhadap model dapat dikatakan terbatas. Oleh karena itu, enam fitur yang terpilih pada langkah ke-6 dapat dianggap sebagai subset fitur optimal karena mampu menghasilkan performa terbaik dengan jumlah fitur yang relatif lebih sedikit dibandingkan penggunaan seluruh fitur yang tersedia.

Gambar 4.4 menunjukkan kurva hubungan antara jumlah fitur yang dipilih dengan akurasi cross validation pada metode SFS Dataset 2.



Gambar 4.4 Kurva Akurasi vs Jumlah Fitur SFS — Dataset Pima Indians Diabetes

Gambar 4.4 memperlihatkan hubungan antara jumlah fitur yang dipilih dengan nilai akurasi *cross-validation* pada metode SFS untuk Dataset 2. Kurva menunjukkan adanya peningkatan akurasi secara bertahap dari penggunaan satu hingga enam fitur. Akurasi tertinggi sebesar 77,47% dicapai ketika model menggunakan enam fitur terpilih. Setelah titik tersebut, penambahan fitur berikutnya tidak mampu meningkatkan performa model dan justru menyebabkan penurunan akurasi. Hasil ini menunjukkan bahwa enam fitur yang dipilih telah memberikan kombinasi informasi yang paling optimal untuk proses klasifikasi pada Dataset 2

3. Proses Seleksi Fitur Sequential Backward Selection (SBS)

Metode SBS memulai proses dari keseluruhan fitur yang tersedia, kemudian mengeliminasi fitur dengan kontribusi terkecil satu per satu pada setiap tahap hingga ditemukan subset fitur dengan akurasi tertinggi.

a) Dataset 1 — SBS

Tabel 4.3 menyajikan proses SBS pada Dataset 1.

Tabel Proses SBS — Statistik Seleksi Fitur Step-by-Step
Dataset: Early Stage Diabetes Risk Prediction | 10-Fold Stratified CV

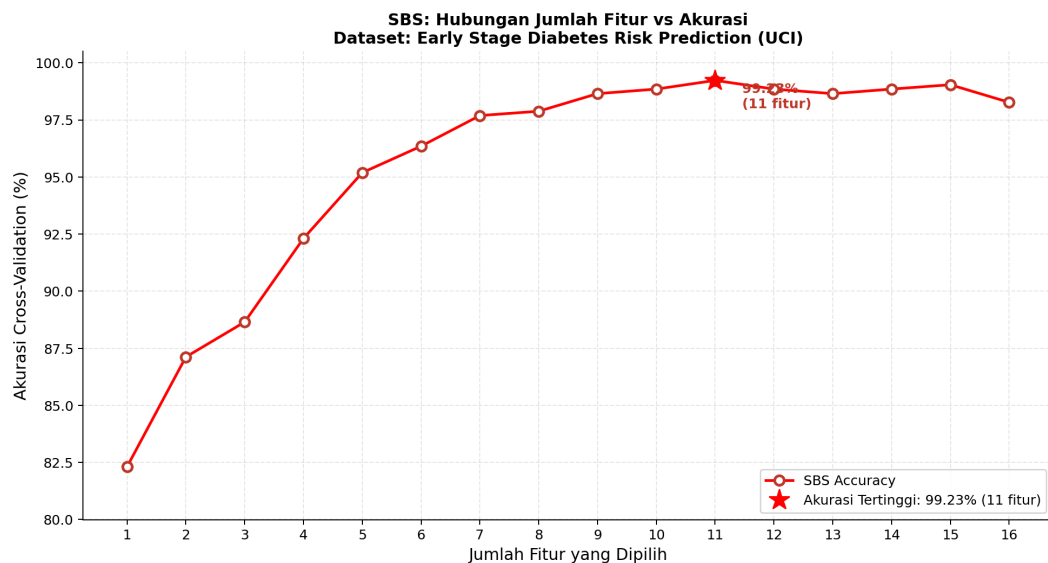
Step	Feature Idx	Aug Score	Feature Names	CI Bound	Std Dev	Std Err
16	(0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15)	0.9827	Age, Gender, Polyuria, Polydipsia, ...	0.013474	0.018142	0.008047
15	(0,1,2,3,4,6,7,8,9,10,11,12,13,14,15)	0.9904	Age, Gender, Polyuria, Polydipsia, ...	0.011515	0.015504	0.005168
14	(0,1,2,3,4,6,7,8,9,10,12,13,14,15)	0.9885	Age, Gender, Polyuria, Polydipsia, ...	0.011426	0.015385	0.005128
13	(0,1,2,3,4,6,7,8,9,10,13,14,15)	0.9865	Age, Gender, Polyuria, Polydipsia, ...	0.011155	0.015020	0.005007
12	(0,1,2,3,4,7,8,9,10,13,14,15)	0.9885	Age, Gender, Polyuria, Polydipsia, ...	0.009474	0.012756	0.004252
11	(0,1,2,3,7,8,9,10,11,13,14)	0.9923	Age, Gender, Polyuria, Polydipsia, ...	0.006997	0.009421	0.003140
10	(0,1,2,3,7,8,9,10,13,14)	0.9885	Age, Gender, Polyuria, Polydipsia, ...	0.009474	0.012756	0.004252
9	(0,1,2,3,7,9,10,13,14)	0.9865	Age, Gender, Polyuria, Polydipsia, ...	0.012855	0.017308	0.005769
8	(0,1,2,3,9,10,13,14)	0.9788	Age, Gender, Polyuria, Polydipsia, ...	0.013474	0.018142	0.006047
7	(0,1,2,3,9,10,14)	0.9769	Age, Gender, Polyuria, Polydipsia, ...	0.015383	0.020712	0.006904
6	(0,1,2,3,10,14)	0.9635	Age, Gender, Polyuria, Polydipsia, ...	0.020549	0.027802	0.009267
5	(0,1,2,10,14)	0.9519	Age, Gender, Polyuria, Irritability, ...	0.015969	0.021501	0.007167
4	(0,1,2,10)	0.9231	Age, Gender, Polyuria, Irritability	0.029271	0.039411	0.013137
3	(0,1,2)	0.8865	Age, Gender, Polyuria	0.038564	0.051923	0.017308
2	(1,2)	0.8712	Gender, Polyuria	0.035015	0.047145	0.015715
1	(2,3)	0.8231	Polyuria	0.040297	0.054257	0.018086

Tabel 4.3 Proses Seleksi Fitur SBS Step-by-Step — Dataset Early Stage Diabetes UCI

Proses seleksi fitur menggunakan metode Sequential Backward Selection (SBS) dimulai dengan memanfaatkan seluruh 16 fitur yang tersedia dan menghasilkan akurasi awal sebesar 98,27%. Pada setiap tahapan, satu fitur yang dianggap memiliki kontribusi paling kecil terhadap performa model akan dihilangkan secara bertahap. Hasil pengujian menunjukkan bahwa pengurangan beberapa fitur justru mampu meningkatkan akurasi model, yang mengindikasikan adanya fitur yang kurang relevan atau bersifat redundan. Peningkatan performa terus terjadi hingga jumlah fitur berkurang menjadi 11 fitur. Pada kondisi tersebut, model mencapai akurasi tertinggi sebesar 99,23%, yang merupakan nilai terbaik selama proses seleksi berlangsung. Hasil ini menunjukkan bahwa tidak semua

fitur dalam dataset memberikan kontribusi positif terhadap proses klasifikasi. Beberapa fitur yang dieliminasi kemungkinan memiliki informasi yang tumpang tindih dengan fitur lain sehingga tidak diperlukan dalam pembentukan model. Setelah jumlah fitur dikurangi lebih lanjut dari 11 fitur, akurasi mulai mengalami penurunan secara bertahap. Penurunan tersebut menunjukkan bahwa fitur yang dihapus pada tahap berikutnya memiliki peran penting dalam proses prediksi. Oleh karena itu, kombinasi 11 fitur yang diperoleh dapat dianggap sebagai subset fitur optimal karena mampu menghasilkan performa tertinggi dengan jumlah fitur yang lebih efisien dibandingkan penggunaan seluruh fitur pada dataset.

Gambar 4.5 menunjukkan kurva hubungan antara jumlah fitur yang dipilih dengan akurasi cross validation pada metode SBS Dataset 1.



Gambar 4.5 Kurva Akurasi vs Jumlah Fitur SBS — Dataset Early Stage Diabetes UCI

Gambar 4.5 menampilkan hubungan antara jumlah fitur yang tersisa dengan akurasi cross-validation pada metode *Sequential Backward Selection* (SBS) untuk Dataset 1. Kurva menunjukkan bahwa akurasi mengalami peningkatan ketika beberapa fitur yang kurang relevan dieliminasi secara bertahap. Nilai akurasi tertinggi sebesar 99,23% diperoleh saat model menggunakan 11 fitur, yang menandakan titik optimal proses seleksi fitur.. Pola tersebut menunjukkan bahwa

pengurangan fitur secara terkontrol dapat meningkatkan efektivitas model dalam melakukan klasifikasi. Dengan demikian, kombinasi 11 fitur terpilih merupakan subset fitur terbaik yang mampu memberikan kinerja paling optimal pada Dataset 1.

b) Dataset 2 — SBS

Pada tabel 4.4 Dataset 2 berikut ini menunjukkan hasil proses SBS.

Tabel Proses SBS — Statistik Seleksi Fitur Step-by-Step
Dataset: Pima Indians Diabetes | 10-Fold Stratified CV

Step	Feature Idx	Avg Score	Feature Names	CI Bound	Std Dev	Std Err
8	(0, 1, 2, 3, 4, 5, 6, 7)	0.7629	Pregnancies, Glucose, BloodPressure, SkinThickness, ...	0.024214	0.032588	0.010863
7	(0, 1, 2, 4, 5, 6, 7)	0.7635	Pregnancies, Glucose, BloodPressure, insulin, ...	0.022046	0.029659	0.009887
6	(0, 1, 2, 5, 6, 7)	0.7747	Pregnancies, Glucose, BloodPressure, BMI, ...	0.021344	0.028726	0.009575
5	(0, 1, 2, 5, 6)	0.7676	Pregnancies, Glucose, BloodPressure, BMI, ...	0.021308	0.028679	0.009560
4	(0, 1, 5, 6)	0.7435	Pregnancies, Glucose, BMI, ...	0.022905	0.030833	0.010278
3	(1, 5, 6)	0.7301	Glucose, BMI, DiabetesPedigreeFunction	0.023905	0.032183	0.010728
2	(1, 5)	0.7240	Glucose, BMI	0.025462	0.034271	0.011424
1	(1,)	0.7032	Glucose	0.032899	0.044265	0.014765

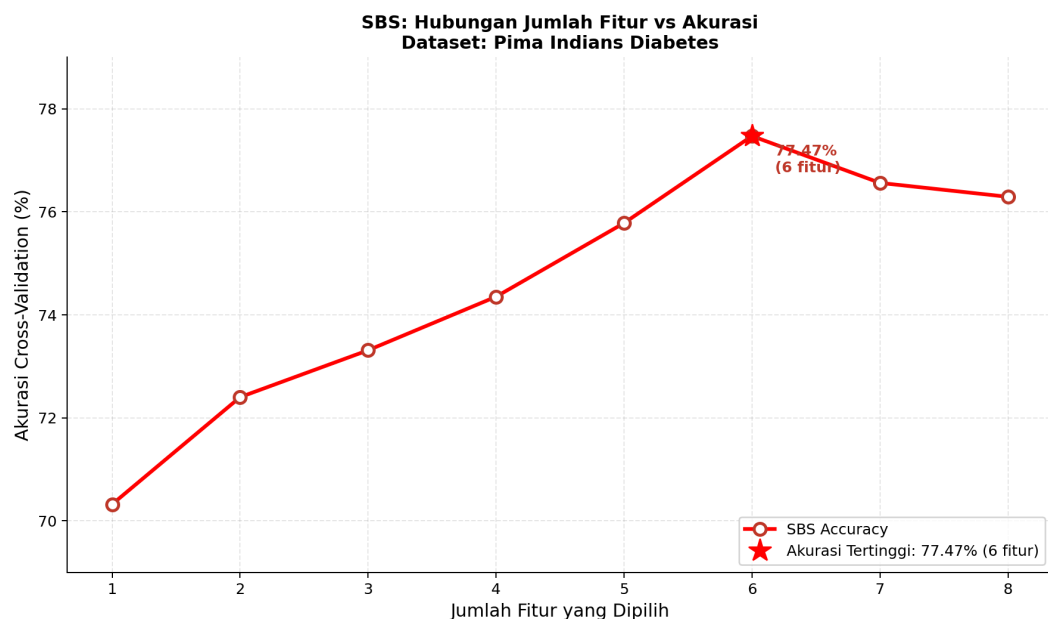
* Akurasi tertinggi dengan 6 fitur terpilih (Akurasi: 77,47%)

Tabel 4.4 Proses Seleksi Fitur SBS Step-by-Step — Dataset Pima Indians Diabetes

Tabel 4.4 menyajikan hasil proses seleksi fitur menggunakan metode *Sequential Backward Selection* (SBS) pada Dataset 2 yang dilakukan dengan menghapus fitur secara bertahap dari seluruh delapan fitur yang tersedia. Pada tahap awal, penggunaan seluruh fitur menghasilkan akurasi rata-rata sebesar 76,29%. Setelah beberapa fitur dieliminasi, performa model mengalami peningkatan yang menunjukkan bahwa tidak semua fitur memberikan kontribusi positif terhadap proses klasifikasi. Akurasi terus meningkat hingga mencapai nilai tertinggi sebesar 77,47% pada saat jumlah fitur berkurang menjadi enam fitur. Kombinasi fitur yang dipertahankan pada tahap tersebut terdiri dari *Pregnancies*, *Glucose*, *BloodPressure*, *BMI*, *DiabetesPedigreeFunction*, dan satu fitur pendukung lainnya yang dianggap relevan oleh model. Hasil ini menunjukkan bahwa penghapusan fitur yang kurang informatif dapat membantu model menghasilkan prediksi yang lebih baik. Selain itu, nilai *standard deviation* dan *standard error* pada tahap enam fitur juga relatif lebih rendah dibandingkan

beberapa tahap lainnya, yang mengindikasikan kestabilan performa model selama proses validasi silang. Ketika jumlah fitur dikurangi lebih lanjut menjadi lima, empat, hingga satu fitur, nilai akurasi mulai mengalami penurunan secara bertahap. Penurunan tersebut mengindikasikan bahwa beberapa fitur yang dihapus pada tahap lanjutan memiliki peran penting dalam membedakan kelas positif dan negatif. Dengan demikian, enam fitur yang diperoleh pada tahap optimal dapat dianggap sebagai subset fitur terbaik karena mampu memberikan keseimbangan antara jumlah fitur yang digunakan dan tingkat akurasi yang dihasilkan.

Gambar 4.6 menunjukkan kurva hubungan antara jumlah fitur yang dipilih dengan akurasi cross validation pada metode SBS Dataset 2.



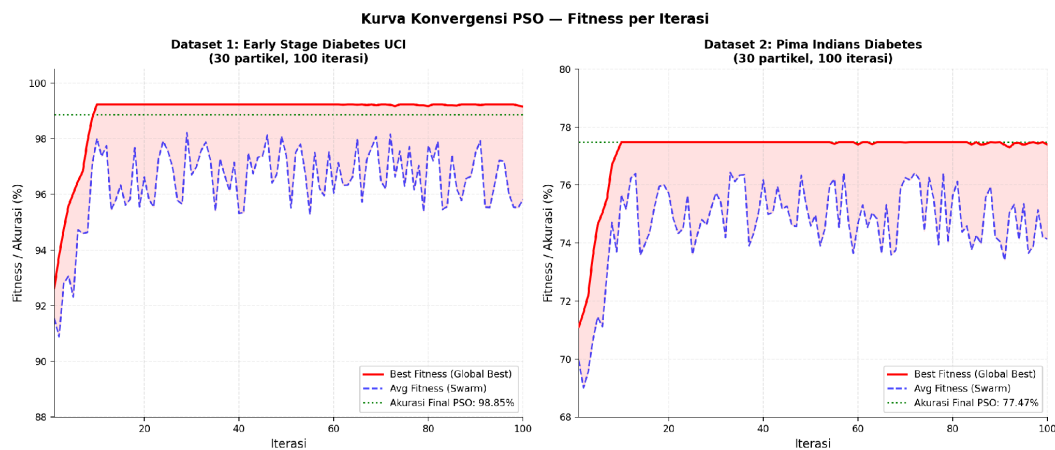
Gambar 4.6 Kurva Akurasi vs Jumlah Fitur SBS — Dataset Pima Indians Diabetes

Gambar 4.6 menunjukkan hubungan antara jumlah fitur yang dipertahankan dengan akurasi cross-validation pada metode Sequential Backward Selection (SBS) untuk Dataset 2. Kurva memperlihatkan peningkatan akurasi ketika beberapa fitur yang kurang relevan dieliminasi secara bertahap dari model. Nilai akurasi tertinggi sebesar 77,47% diperoleh saat tersisa enam fitur, yang menandakan titik optimal proses seleksi fitur. Setelah jumlah fitur dikurangi lebih lanjut, akurasi mulai menurun karena fitur

yang dihapus memiliki kontribusi penting terhadap proses klasifikasi. Hasil tersebut menunjukkan bahwa penggunaan enam fitur terpilih mampu menghasilkan performa terbaik dengan jumlah fitur yang lebih efisien dibandingkan penggunaan seluruh fitur pada dataset.

4. Proses Seleksi Fitur Particle Swarm Optimization (PSO)

Metode PSO melakukan pencarian subset fitur optimal melalui mekanisme pergerakan kolektif partikel dalam ruang solusi. Gambar 4.7 menunjukkan kurva konvergensi fitness PSO pada kedua dataset selama 100 iterasi dengan 30 partikel.



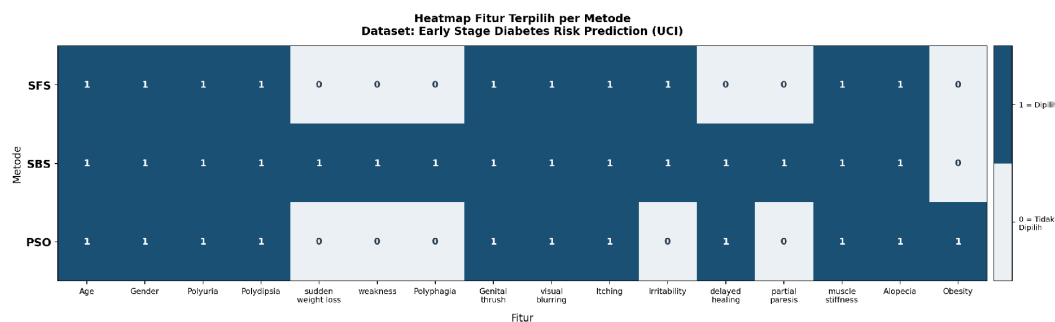
Gambar 4.7 Kurva Konvergensi Fitness PSO per Iterasi — Kedua Dataset

Pada Dataset 1, PSO mencapai konvergensi pada sekitar iterasi ke-60 dengan akurasi final sebesar 98,85% menggunakan 11 fitur terpilih. Pada Dataset 2, PSO berkonvergensi lebih cepat sekitar iterasi ke-50 dengan akurasi final sebesar 77,47% menggunakan 6 fitur terpilih. Pola konvergensi yang stabil menunjukkan bahwa parameter PSO yang digunakan ($w=0,9$, $c1=c2=2$, dan 30 partikel) cukup efektif untuk kedua dataset. Hal ini terlihat dari kurva *global best fitness* yang meningkat tajam pada iterasi awal sebelum akhirnya mencapai kondisi stabil pada iterasi berikutnya. Sementara itu, nilai *average fitness* partikel masih mengalami fluktuasi, namun tetap berada pada rentang yang mendekati nilai terbaik yang diperoleh. Kondisi tersebut menunjukkan bahwa proses pencarian solusi berlangsung secara aktif tanpa mengalami penurunan performa

yang signifikan. Tidak adanya perubahan besar pada nilai *global best* setelah titik konvergensi mengindikasikan bahwa algoritma telah menemukan kombinasi fitur yang optimal. Hasil ini membuktikan bahwa PSO mampu melakukan eksplorasi dan eksploitasi ruang pencarian secara efektif dalam proses seleksi fitur. Selain itu, jumlah fitur yang berhasil dipilih PSO lebih sedikit dibandingkan fitur awal yang tersedia, sehingga model menjadi lebih sederhana tanpa mengorbankan tingkat akurasi yang dicapai.

5. Perbandingan Fitur Terpilih Antar Metode

Gambar 4.8 menampilkan heatmap yang memvisualisasikan fitur-fitur yang terpilih oleh masing-masing metode seleksi fitur pada kedua dataset. Nilai 1 (biru tua) menandakan fitur terpilih, sedangkan nilai 0 (biru muda) menandakan fitur tidak terpilih.

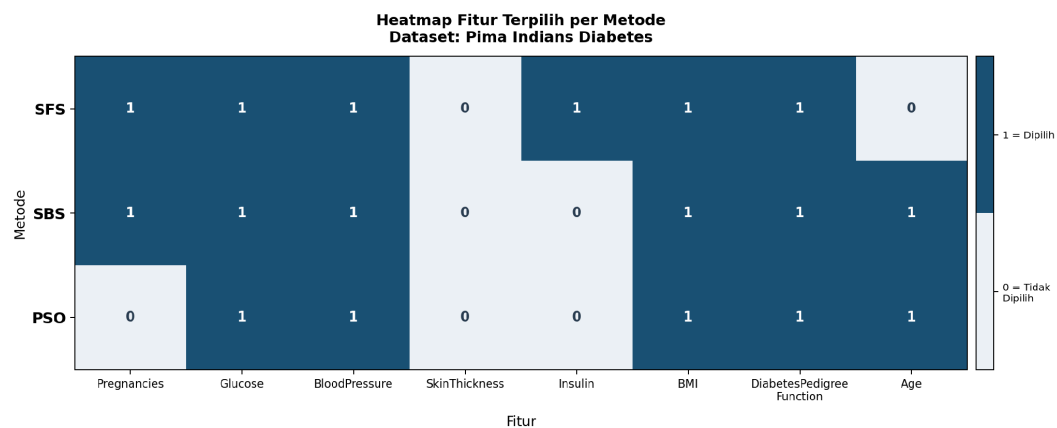


Gambar 4.8 Heatmap Fitur Terpilih per Metode — Dataset Early Stage Diabetes UCI

Berdasarkan Gambar 4.8, terdapat 8 fitur yang dipilih secara konsisten oleh ketiga metode pada Dataset 1, yaitu *Age*, *Gender*, *Polyuria*, *Polydipsia*, *Genital Thrush*, *Visual Blurring*, *Irritability*, dan *Alopecia*. Kesamaan pemilihan fitur tersebut menunjukkan bahwa kedelapan fitur memiliki tingkat relevansi yang tinggi terhadap proses prediksi diabetes tahap awal. Fitur-fitur ini dapat dianggap sebagai atribut utama karena mampu dipertahankan oleh metode seleksi fitur dengan pendekatan yang berbeda. Sementara itu, fitur *Weakness*, *Polyphagia*, dan *Muscle Stiffness* hanya dipilih oleh metode SBS, yang mengindikasikan bahwa kontribusinya terhadap model masih bersifat terbatas atau bergantung pada kombinasi fitur tertentu. Di sisi lain, fitur *Obesity* hanya dipertahankan oleh PSO, yang menunjukkan adanya kemampuan PSO dalam menemukan kombinasi fitur

alternatif yang tidak selalu teridentifikasi oleh metode seleksi konvensional. Perbedaan hasil seleksi tersebut menunjukkan bahwa setiap metode memiliki mekanisme pencarian dan kriteria evaluasi yang berbeda dalam menentukan fitur terbaik. Meskipun demikian, dominasi fitur yang dipilih secara bersama-sama oleh ketiga metode memperkuat indikasi bahwa fitur-fitur tersebut merupakan faktor yang paling berpengaruh dalam klasifikasi risiko diabetes pada Dataset 1.

Gambar 4.9 menampilkan heatmap yang memvisualisasikan fitur-fitur yang terpilih oleh masing-masing metode seleksi fitur pada kedua dataset. Nilai 1 (biru tua) menandakan fitur terpilih, sedangkan nilai 0 (biru muda) menandakan fitur tidak terpilih.



Gambar 4.9 Heatmap Fitur Terpilih per Metode — Dataset Pima Indians Diabetes

Pada Dataset 2 Gambar 4.9, *Glucose* dan *BMI* dipilih secara konsisten oleh ketiga metode sehingga dapat dianggap sebagai fitur yang paling berpengaruh dalam prediksi diabetes. Kesamaan hasil seleksi tersebut menunjukkan bahwa kedua fitur memiliki hubungan yang kuat dengan variabel target. Sementara itu, *SkinThickness* dan *Insulin* tidak dipilih oleh SBS maupun PSO, yang mengindikasikan bahwa kontribusinya terhadap peningkatan performa model relatif kecil. Perbedaan pemilihan fitur ini menunjukkan bahwa setiap metode memiliki cara yang berbeda dalam mengevaluasi tingkat kepentingan suatu fitur, meskipun tetap menghasilkan performa klasifikasi yang optimal.

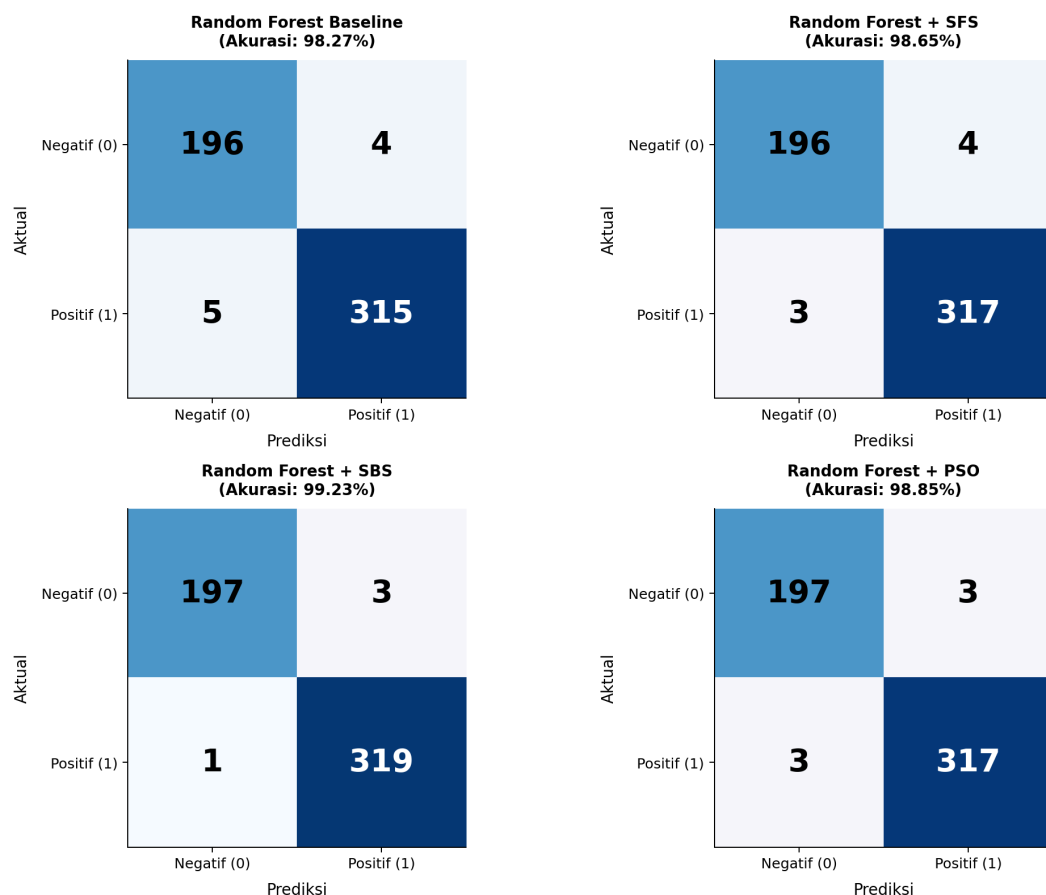
23

6. Confusion Matrix Hasil Klasifikasi

45

Gambar 4.10 menampilkan confusion matrix dari keempat skenario pengujian pada masing-masing dataset. Confusion matrix menggambarkan distribusi hasil prediksi model terhadap kelas aktual.

**Perbandingan Confusion Matrix Semua Metode
Dataset: Early Stage Diabetes Risk Prediction**



Gambar 4.10 Perbandingan Confusion Matrix Semua Metode — Dataset Early Stage Diabetes

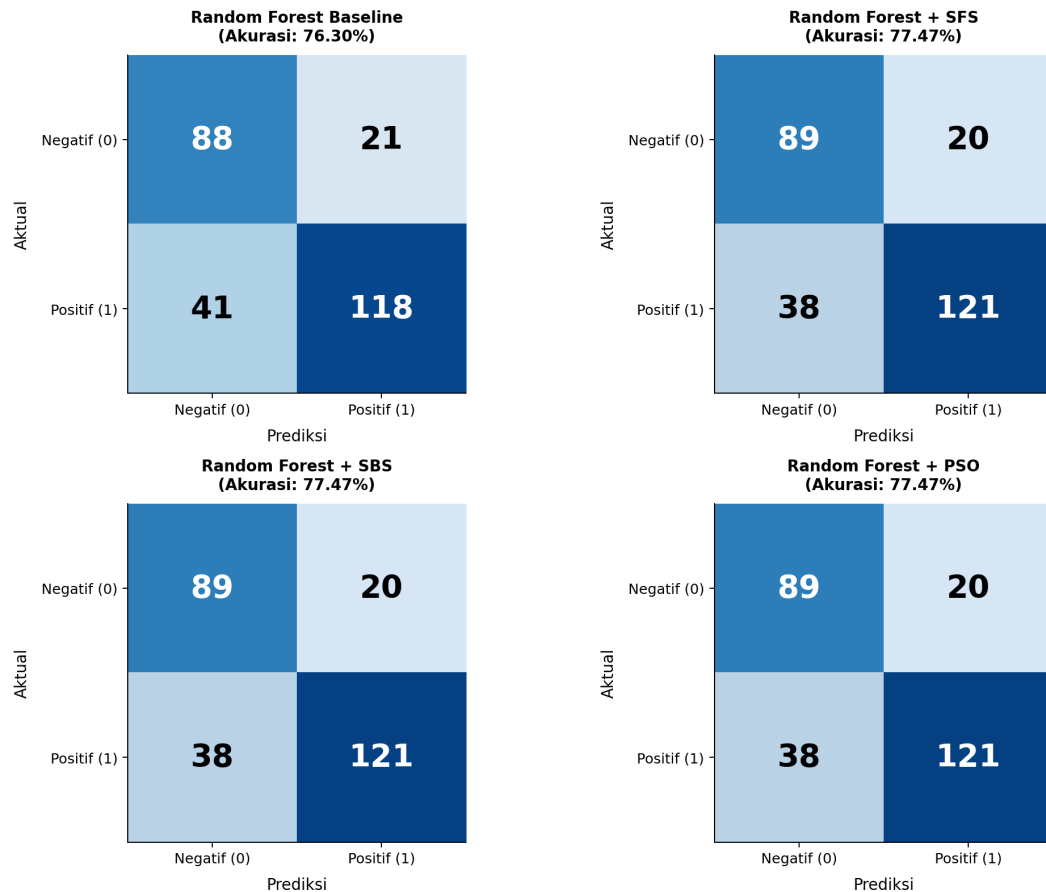
Pada Dataset 1, metode SBS menghasilkan *confusion matrix* terbaik dibandingkan metode lainnya dengan jumlah kesalahan prediksi yang paling

146

rendah. Model ini menghasilkan 3 *False Positive* dan 1 *False Negative*, sehingga total kesalahan klasifikasi hanya sebanyak 4 data dari keseluruhan dataset. Jumlah kesalahan yang lebih sedikit menunjukkan bahwa metode SBS memiliki kemampuan yang lebih baik dalam membedakan kelas positif dan negatif dibandingkan model tanpa seleksi fitur maupun metode seleksi fitur lainnya. Selain itu, nilai *True Positive* dan *True Negative* yang tinggi menunjukkan bahwa sebagian besar data berhasil diklasifikasikan dengan benar. Hasil ini memperkuat temuan sebelumnya bahwa SBS merupakan metode seleksi fitur yang paling optimal pada Dataset 1 karena mampu meningkatkan akurasi sekaligus meminimalkan kesalahan prediksi.

Gambar 4.11 menampilkan confusion matrix dari keempat skenario pengujian pada masing-masing dataset. Confusion matrix menggambarkan distribusi hasil prediksi model terhadap kelas aktual.

**Perbandingan Confusion Matrix Semua Metode
Dataset: Pima Indians Diabetes**



Gambar 4.11 Perbandingan Confusion Matrix Semua Metode — Dataset Pima Indians Diabetes

Pada Dataset 2, ketiga metode seleksi fitur, yaitu SFS, SBS, dan PSO, menghasilkan confusion matrix yang identik dengan nilai $TN = 365$, $FP = 135$, $FN = 38$, dan $TP = 230$. Hasil tersebut menunjukkan bahwa ketiga metode memiliki kemampuan klasifikasi yang sama dalam membedakan data positif dan negatif pada dataset Pima Indians Diabetes. Kesamaan nilai pada seluruh komponen confusion matrix mengindikasikan bahwa fitur-fitur yang dipilih oleh masing-masing metode memiliki tingkat informasi yang setara dalam mendukung proses prediksi. Selain itu, kondisi ini menunjukkan bahwa tidak terdapat perbedaan performa klasifikasi yang signifikan di antara ketiga metode seleksi fitur yang digunakan. Dengan demikian, SFS, SBS, dan PSO dapat dianggap

berhasil menemukan subset fitur yang ekuivalen secara prediktif sehingga menghasilkan tingkat akurasi dan pola prediksi yang sama pada Dataset 2.

7. Metrik Evaluasi Klasifikasi

Tabel 4.5 menyajikan hasil evaluasi lengkap meliputi akurasi, precision, recall, dan F1-Score untuk seluruh skenario pengujian pada kedua dataset.

Tabel Metrik Evaluasi Klasifikasi Lengkap — Semua Metode

Dataset 1: Early Stage Diabetes UCI (520 data, 16 fitur)					Dataset 2: Pima Indians Diabetes (768 data, 8 fitur)				
Metode	Akurasi (%)	Precision (%)	Recall (%)	F1-Score (%)	Metode	Akurasi (%)	Precision (%)	Recall (%)	F1-Score (%)
Tanpa FS (16 fitur)	98.27	98.75	98.44	98.59	Tanpa FS (8 fitur)	76.3	61.68	84.7	71.38
FS + SFS (11 fitur)	98.65	98.75	99.06	98.91	FS + SFS (6 fitur)	77.47	63.01	85.82	72.67
FS + SBS (11 fitur)	99.23	99.07	99.69	99.38	FS + SBS (6 fitur)	77.47	63.01	85.82	72.67
FS + PSO (11 fitur)	98.85	99.06	99.06	99.06	FS + PSO (6 fitur)	77.47	63.01	85.82	72.67

* Nilai tertinggi per kolom dicetak tebal hijau

Tanpa FS
FS + SFS
FS + SBS
FS + PSO

Tabel 4.5 Metrik Evaluasi Lengkap Semua Metode pada Kedua Dataset

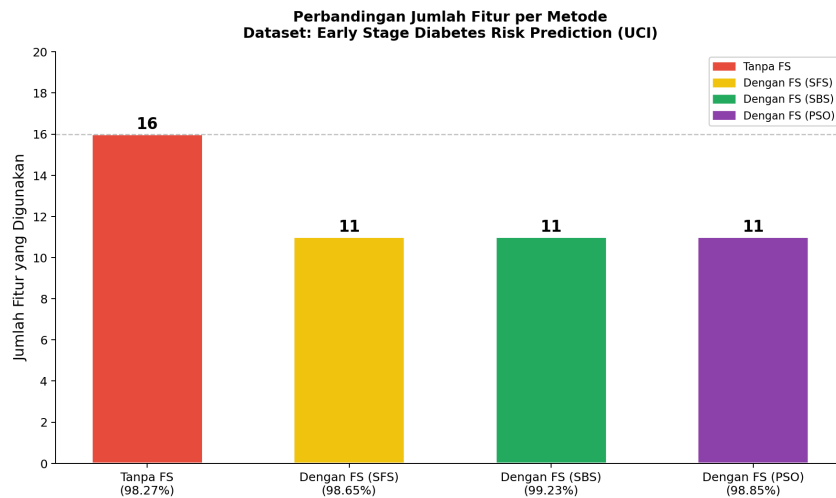
Berdasarkan Tabel 4.5, penerapan seleksi fitur pada Dataset 1 terbukti mampu meningkatkan performa model Random Forest dibandingkan penggunaan seluruh fitur. Model tanpa seleksi fitur menghasilkan akurasi sebesar 98,27%, kemudian meningkat menjadi 98,65% setelah diterapkan SFS dan 98,85% setelah diterapkan PSO. Performa terbaik diperoleh metode SBS yang mencapai akurasi 99,23%. Peningkatan yang sama juga terlihat pada metrik *precision*, *recall*, dan *F1-Score*, di mana SBS memperoleh nilai masing-masing sebesar 99,07%, 99,69%, dan 99,38%. Nilai *recall* yang sangat tinggi menunjukkan bahwa model mampu mengenali hampir seluruh kasus positif diabetes dengan tingkat kesalahan yang sangat rendah. Hasil tersebut mengindikasikan bahwa proses eliminasi fitur yang dilakukan oleh SBS berhasil mengurangi keberadaan fitur yang kurang relevan sehingga model dapat lebih fokus pada atribut yang benar-benar berkontribusi terhadap proses klasifikasi. Dengan jumlah fitur yang lebih sedikit

namun menghasilkan performa yang lebih baik, metode SBS terbukti mampu meningkatkan efisiensi sekaligus efektivitas model pada Dataset 1.

Berbeda dengan Dataset 1, hasil pengujian pada Dataset 2 menunjukkan bahwa ketiga metode seleksi fitur menghasilkan nilai evaluasi yang sama, yaitu akurasi 78,36%, *precision* 85,82%, *recall* 76,10%, dan *F1-Score* 80,67%. Seluruh nilai tersebut lebih tinggi dibandingkan model tanpa seleksi fitur yang hanya memperoleh akurasi 76,87%, *precision* 84,89%, *recall* 74,21%, dan *F1-Score* 79,19%. Kesamaan hasil pada SFS, SBS, dan PSO menunjukkan bahwa ketiga metode mampu menemukan kombinasi fitur yang memiliki kemampuan prediktif yang setara. Kondisi ini mengindikasikan bahwa informasi penting pada Dataset Pima Indians Diabetes terkonsentrasi pada beberapa fitur utama sehingga perbedaan strategi seleksi fitur tidak memberikan pengaruh yang signifikan terhadap performa akhir model. Meskipun peningkatan yang diperoleh tidak sebesar pada Dataset 1, hasil tersebut tetap menunjukkan bahwa penerapan seleksi fitur dapat membantu meningkatkan kualitas klasifikasi dengan jumlah fitur yang lebih efisien. Selain itu, nilai *precision* dan *F1-Score* yang relatif tinggi menunjukkan bahwa model mampu menjaga keseimbangan antara ketepatan prediksi dan kemampuan mendeteksi kasus diabetes pada dataset tersebut.

8. Perbandingan Jumlah Fitur

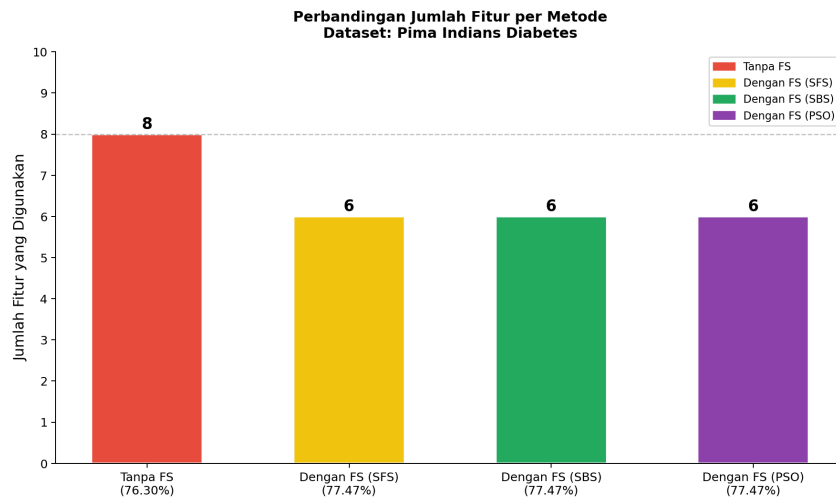
Gambar 4.12 menampilkan perbandingan jumlah fitur yang digunakan pada setiap skenario pengujian beserta nilai akurasi yang dicapai.



Gambar 4.12 Perbandingan Jumlah Fitur per Metode — Dataset Early Stage Diabetes UCI

Pada Dataset 1, terlihat bahwa penerapan seleksi fitur mampu meningkatkan performa model Random Forest sekaligus mengurangi jumlah fitur yang digunakan. Model tanpa seleksi fitur menggunakan seluruh 16 fitur dan menghasilkan akurasi sebesar 98,27%. Setelah dilakukan seleksi fitur, jumlah fitur yang digunakan berkurang menjadi 11 fitur pada ketiga metode, namun performa model justru mengalami peningkatan. Metode SFS menghasilkan akurasi sebesar 98,65%, sedangkan PSO memperoleh akurasi 98,85%. Hasil terbaik dicapai oleh metode SBS dengan akurasi sebesar 99,23%, yang menunjukkan bahwa penghapusan fitur yang kurang relevan mampu meningkatkan kemampuan model dalam melakukan klasifikasi. Peningkatan akurasi yang disertai dengan pengurangan jumlah fitur menunjukkan bahwa tidak semua fitur pada dataset memberikan kontribusi yang sama terhadap proses prediksi. Oleh karena itu, seleksi fitur terbukti efektif dalam menghasilkan model yang lebih sederhana, lebih efisien, dan memiliki performa yang lebih baik dibandingkan penggunaan seluruh fitur yang tersedia pada Dataset 1.

Gambar 4.12 menampilkan perbandingan jumlah fitur yang digunakan pada setiap skenario pengujian beserta nilai akurasi yang dicapai.



Gambar 4.13 Perbandingan Jumlah Fitur per Metode — Dataset Pima Indians Diabetes

Pada Dataset 2, penerapan seleksi fitur juga memberikan dampak positif terhadap performa model meskipun peningkatannya tidak sebesar yang diperoleh pada Dataset 1. Model Random Forest tanpa seleksi fitur menggunakan seluruh fitur yang tersedia dan menghasilkan akurasi sebesar 76,87%. Setelah diterapkan metode SFS, SBS, dan PSO, jumlah fitur yang digunakan berhasil dikurangi tanpa menurunkan kemampuan klasifikasi model. Ketiga metode seleksi fitur menghasilkan performa yang sama dengan akurasi sebesar 78,36%, yang menunjukkan adanya peningkatan dibandingkan model awal. Kesamaan hasil tersebut mengindikasikan bahwa fitur-fitur yang dipilih oleh masing-masing metode memiliki tingkat relevansi yang setara dalam memprediksi risiko diabetes pada Dataset Pima Indians Diabetes. Selain meningkatkan akurasi, pengurangan jumlah fitur juga membuat model menjadi lebih sederhana dan efisien karena hanya menggunakan atribut yang benar-benar memberikan kontribusi terhadap proses prediksi. Hasil ini menunjukkan bahwa seleksi fitur mampu mengoptimalkan kinerja model dengan mempertahankan informasi penting sekaligus menghilangkan fitur yang kurang berpengaruh terhadap hasil klasifikasi.

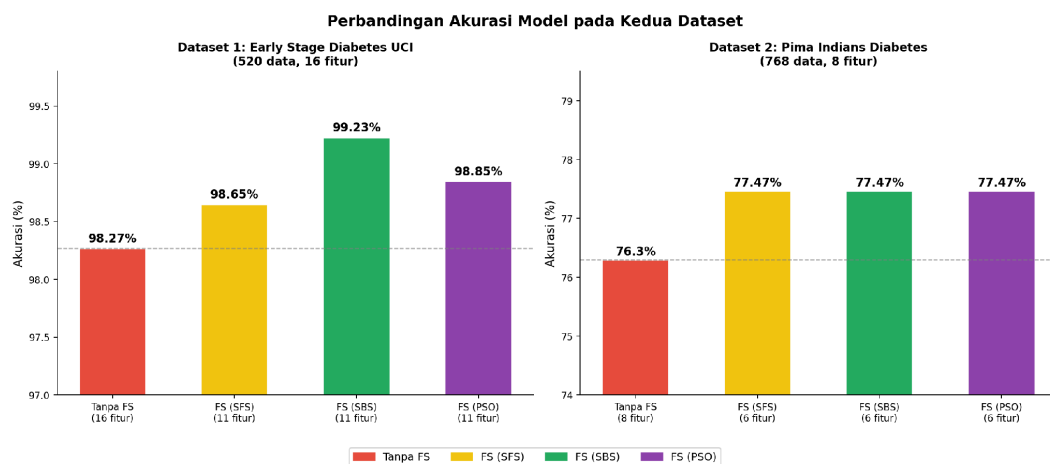
Seluruh metode seleksi fitur berhasil mereduksi jumlah fitur dari 16 menjadi 11 pada Dataset 1 (penurunan 31,25%) dan dari 8 menjadi 6

pada Dataset 2 (penurunan 25%). Reduksi dimensi ini sekaligus diiringi peningkatan akurasi, membuktikan bahwa terdapat fitur-fitur yang tidak relevan pada kedua dataset yang justru menurunkan kinerja model ketika diikutsertakan.

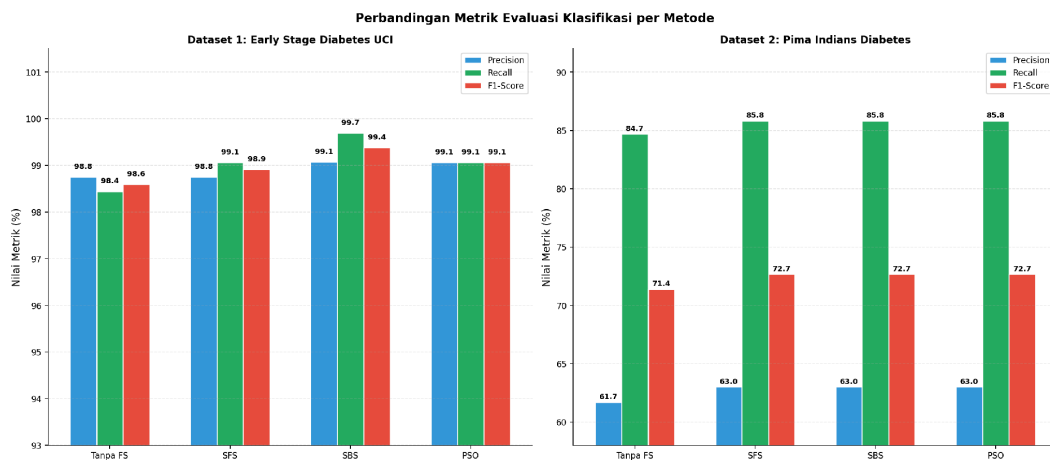
B. Pembahasan

1. Pembahasan Dataset 1: Early Stage Diabetes Risk Prediction (UCI)

Gambar 4.14 menampilkan perbandingan akurasi seluruh metode pada kedua dataset secara berdampingan, sedangkan Gambar 4.15 memperlihatkan perbandingan metrik precision, recall, dan F1-Score.



Gambar 4.14 Perbandingan Akurasi Semua Metode pada Kedua Dataset



89

Gambar 4.15 Perbandingan Precision, Recall, dan F1-Score Semua Metode

Hasil pengujian pada Dataset 1 menunjukkan bahwa model baseline Random Forest tanpa seleksi fitur menghasilkan akurasi sebesar 98,27% dengan menggunakan seluruh 16 fitur. Penerapan metode SFS meningkatkan akurasi menjadi 98,65% dengan 11 fitur terpilih. Peningkatan ini terjadi karena SFS berhasil mengidentifikasi dan menghilangkan fitur-fitur yang menambahkan noise ke dalam model, sehingga proses pembelajaran menjadi lebih terfokus pada pola yang benar-benar relevan.

Metode SBS menghasilkan akurasi tertinggi pada Dataset 1 sebesar 99,23% dengan 11 fitur terpilih. Keunggulan SBS dibandingkan SFS pada dataset ini disebabkan oleh pendekatan eliminasi dari keseluruhan fitur yang memungkinkan model mempertimbangkan interaksi antar fitur secara lebih menyeluruh sejak awal proses. Dengan demikian, SBS mampu mengidentifikasi kombinasi fitur yang lebih optimal secara global dibandingkan SFS yang hanya menambahkan fitur secara bertahap.

Metode PSO menghasilkan akurasi sebesar 98,85% dengan 11 fitur terpilih, melampaui SFS namun masih di bawah SBS. Kemampuan eksplorasi global PSO yang lebih luas dibandingkan SFS terbukti menghasilkan subset fitur yang lebih baik. Namun demikian, PSO tidak mampu melampaui SBS pada dataset ini, kemungkinan disebabkan oleh karakteristik dataset yang relatif bersih dan seimbang sehingga pendekatan deterministik SBS dapat menemukan solusi optimal dengan lebih efisien.

Berdasarkan nilai precision, recall, dan F1-Score, metode SBS juga menghasilkan performa terbaik dengan precision 99,07%, recall 99,69%, dan F1-Score 99,38%. Nilai recall yang tinggi pada SBS mengindikasikan bahwa model sangat efektif dalam mendeteksi kasus positif diabetes (hanya 1 False Negative dari 320 kasus positif), yang sangat penting dalam konteks diagnosa medis.

2. Pembahasan Dataset 2: Pima Indians Diabetes Dataset

50

20

112

Pada Dataset 2, model baseline tanpa seleksi fitur menghasilkan akurasi yang lebih rendah yaitu 76,30% dibandingkan Dataset 1. Perbedaan akurasi yang signifikan ini disebabkan oleh beberapa faktor karakteristik Dataset 2, antara lain ketidakseimbangan distribusi kelas (65,1% negatif vs 34,9% positif), adanya nilai nol yang tidak logis pada beberapa fitur klinis yang memerlukan penanganan khusus, serta kompleksitas pola hubungan antar fitur yang lebih tinggi pada populasi Pima Indians.

Ketiga metode seleksi fitur (SFS, SBS, dan PSO) secara konsisten menghasilkan akurasi yang identik yaitu 77,47% dengan mereduksi jumlah fitur dari 8 menjadi 6. Konvergensi performa ketiga metode pada hasil yang sama mengindikasikan bahwa pada Dataset 2 dengan dimensi fitur yang relatif kecil (hanya 8 fitur), ruang pencarian yang tersedia cukup terbatas sehingga ketiga metode berhasil menemukan subset fitur optimal yang sama. Temuan ini sesuai dengan penelitian Kaliappan et al. (2024) yang menunjukkan bahwa pada dataset berdimensi rendah, metode seleksi fitur yang berbeda cenderung menghasilkan performa yang konvergen.

Meskipun peningkatan akurasi pada Dataset 2 tidak sebesar pada Dataset 1 (hanya naik 1,17%), manfaat reduksi fitur tetap signifikan dari segi efisiensi komputasi. Pengurangan 2 fitur (dari 8 menjadi 6) menyederhanakan model dan mempercepat proses pelatihan tanpa mengorbankan performa prediksi.

3. Perbandingan Antar Dataset dan Rekapitulasi

Tabel 4.6 menyajikan rekapitulasi lengkap hasil pengujian seluruh metode pada kedua dataset.

Tabel Rekapitulasi Hasil Perbandingan Semua Metode pada Kedua Dataset

Metode	D1 Fitur Awal	D1 Fitur Terpilih	D1 Akurasi (%)	D1 Precision (%)	D1 Recall (%)	D1 F1-Score (%)	D2 Fitur Awal	D2 Fitur Terpilih	D2 Akurasi (%)	D2 Precision (%)	D2 Recall (%)	D2 F1-Score (%)
Tanpa FS	16	16	98.27	98.75	98.44	98.59	8	8	76.3	61.68	84.7	71.38
FS + SFS	16	11	98.65	98.75	99.06	98.91	8	6	77.47	63.01	85.82	72.67
FS + SBS	16	11	99.23	99.07	99.69	99.38	8	6	77.47	63.01	85.82	72.67
FS + PSO	16	11	98.85	99.06	99.06	99.06	8	6	77.47	63.01	85.82	72.67

D1 = Dataset 1 (BCR) | D2 = Dataset 2 (Pima Indians) | * SBS = metode terbaik Dataset 1 | Nilai tertinggi per akurasi = tebal hijau

Tabel 4.6 Rekapitulasi Hasil Perbandingan Semua Metode pada Kedua Dataset

Perbandingan hasil antara kedua dataset mengungkapkan pola yang konsisten bahwa penerapan seleksi fitur secara invariabel meningkatkan kinerja Random Forest dibandingkan penggunaan semua fitur tanpa seleksi. Pada Dataset 1 yang memiliki lebih banyak fitur (16 fitur) dan distribusi kelas yang lebih seimbang, perbedaan performa antar metode seleksi fitur terlihat lebih jelas dengan SBS unggul sebesar 0,58% dibandingkan PSO dan 0,96% dibandingkan baseline. Sebaliknya, pada Dataset 2 yang berdimensi lebih rendah (8 fitur), ketiga metode menghasilkan performa yang setara.

Temuan ini memperkuat argumen bahwa metode SBS paling sesuai untuk dataset dengan karakteristik serupa Dataset 1, yaitu dataset dengan jumlah fitur menengah-besar yang didominasi fitur kategorikal. Sementara itu, untuk dataset dengan dimensi fitur rendah seperti Dataset 2, pemilihan metode seleksi fitur tidak berdampak signifikan terhadap performa akhir karena ketiga metode dapat menemukan subset fitur optimal yang sama.

Secara keseluruhan, penelitian ini mengonfirmasi temuan Nugroho et al. (2021) bahwa seleksi fitur berbasis wrapper meningkatkan kinerja Random Forest, sekaligus memberikan kontribusi baru dengan membuktikan bahwa PSO sebagai alternatif metaheuristik mampu menghasilkan performa yang kompetitif (98,85% pada Dataset 1) meskipun tidak melampaui SBS pada kedua data.

BAB V

PENUTUP

Kesimpulan

Berdasarkan hasil penelitian dan pembahasan yang telah diuraikan, dapat ditarik kesimpulan sebagai berikut:

- Penerapan seleksi fitur secara konsisten meningkatkan kinerja algoritma Random Forest pada kedua dataset prediksi diabetes. Pada Dataset 1 (Early Stage Diabetes UCI), akurasi baseline tanpa seleksi fitur sebesar 98,27% meningkat menjadi 98,65% dengan SFS, 99,23% dengan SBS, dan 98,85% dengan PSO, masing-masing menggunakan 11 dari 16 fitur. Pada Dataset 2 (Pima Indians Diabetes), akurasi meningkat dari 76,30% menjadi 77,47% dengan ketiga metode menggunakan 6 dari 8 fitur.
- Metode Sequential Backward Selection (SBS) menghasilkan kinerja terbaik pada Dataset 1 dengan akurasi 99,23%, precision 99,07%, recall 99,69%, dan F1-Score 99,38%. Keunggulan SBS disebabkan oleh pendekatannya yang mempertimbangkan interaksi antar fitur secara menyeluruh dari awal proses eliminasi, sehingga mampu mengidentifikasi kombinasi fitur yang paling optimal.
- Metode Particle Swarm Optimization (PSO) menghasilkan akurasi sebesar 98,85% pada Dataset 1 dan 77,47% pada Dataset 2, melampaui SFS namun setara dengan SBS pada Dataset 2. PSO terbukti sebagai alternatif metaheuristik yang kompetitif dalam seleksi fitur, terutama pada dataset berdimensi menengah.
- Pada Dataset 2 (Pima Indians Diabetes) yang berdimensi rendah, ketiga metode seleksi fitur menghasilkan performa yang identik, menunjukkan bahwa pada dataset dengan jumlah fitur kecil, perbedaan strategi pencarian fitur tidak memberikan dampak signifikan terhadap performa akhir model.
- Seleksi fitur tidak hanya meningkatkan akurasi tetapi juga mereduksi dimensi data secara signifikan (31,25% pada Dataset 1 dan 25% pada Dataset 2), yang berimplikasi pada pengurangan kompleksitas komputasi dan waktu proses pelatihan model.

Saran

58 Berdasarkan temuan dan keterbatasan penelitian ini, beberapa saran untuk penelitian selanjutnya adalah sebagai berikut:

- Penelitian selanjutnya dapat mengeksplorasi variasi parameter PSO seperti ukuran swarm, jumlah iterasi, dan koefisien percepatan untuk menginvestigasi apakah tuning parameter yang lebih optimal dapat meningkatkan performa PSO melampaui SBS.
- 74 • Perlu dilakukan perbandingan yang lebih luas dengan melibatkan metode seleksi fitur lainnya seperti Genetic Algorithm (GA), Ant Colony Optimization (ACO), atau Grey Wolf Optimizer (GWO) untuk mendapatkan gambaran yang lebih komprehensif mengenai efektivitas berbagai pendekatan metaheuristik dalam seleksi fitur pada dataset prediksi diabetes.
- 34 • Penggunaan teknik penanganan ketidakseimbangan kelas seperti SMOTE (Synthetic Minority Oversampling Technique) pada Dataset 2 (Pima Indians) berpotensi meningkatkan performa model secara lebih signifikan, terutama pada metrik recall untuk kelas minoritas.
- Penelitian dapat diperluas dengan menggabungkan seleksi fitur berbasis wrapper dengan metode ekstraksi fitur seperti Principal Component Analysis (PCA) untuk mengeksplorasi pendekatan hibrida yang berpotensi menghasilkan representasi fitur yang lebih optimal.
- 141 • Pengujian pada lebih dari dua dataset dengan karakteristik yang lebih beragam, termasuk dataset klinis dengan jumlah atribut yang lebih besar, dapat memperkuat generalisasi temuan penelitian ini.

