

**MODEL PREDIKSI DIABETES MENGGUNAKAN GAUSSIAN
PROCESS CLASSIFIER PADA DATA KLINIS**

SKRIPSI

DISUSUN OLEH

EDIL BERTO ZAMILI

2209020248



UMSU
Unggul | Cerdas | Terpercaya

**PROGRAM STUDI TEKNOLOGI INFORMASI
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA**

MEDAN

2026

**MODEL PREDIKSI DIABETES MENGGUNAKAN GAUSSIAN
PROCESS CLASSIFIER PADA DATA KLINIS**

SKRIPSI

**Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer
(S.Kom) dalam Program Studi Teknologi Informasi, pada Fakultas Ilmu Komputer
dan Teknologi Informasi, Universitas Muhammadiyah Sumatera Utara**

EDIL BERTO ZAMILI

2209020248

**PROGRAM STUDI TEKNOLOGI INFORMASI
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA**

MEDAN

2026

LEMBAR PENGESAHAN

Judul Skripsi : MODEL PREDIKSI DIABETES MENGGUNAKAN
GAUSSIAN PROCESS CLASSIFIER
Nama Mahasiswa : EDIL BERTO ZAMILI
NPM : 2209020248
Program Studi : TEKNOLOGI INFORMASI

Menyetujui

Komisi Pembimbing

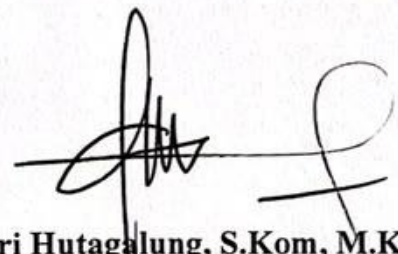


(Dr. Irvan, M.Si)

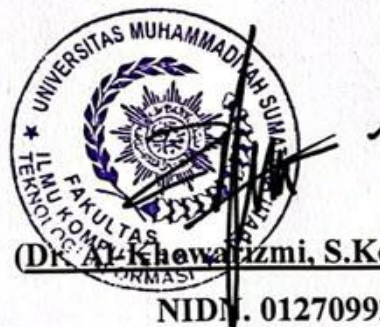
NIDN. 0117117301

Ketua Program Studi

Dekan



(Fatma Sari Hutagalung, S.Kom, M.Kom)
NIDN. 0117019301



(Dr. Al-Khawarizmi, S.Kom., M.Kom)
NIDN. 0127099201

PERNYATAAN ORISINALITAS

MODEL PREDIKSI DIABETES MENGGUNAKAN GAUSSIAN PROCESS CLASSIFIER PADA DATA KLINIS

SKRIPSI

Saya menyatakan bahwa karya tulis ini adalah hasil karya sendiri, kecuali beberapa kutipan dan ringkasan yang masing-masing disebutkan sumbernya.

Medan, 07 September 2026

Yang membuat pernyataan



EDIL BERTO ZAMILI

NPM. 2209020248

PERNYATAAN PERSETUJUAN PUBLIKASI

Sebagai sivitas akademika Universitas Muhammadiyah Sumatera Utara, saya bertanda tangan dibawah ini:

Nama : Edil Berto Zamili
NPM : 2209020248
Program Studi : Teknologi Informasi
Karya Ilmiah : Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Muhammadiyah Sumatera Utara Hak Bedas Royalti Non-Eksekutif (*Non-Exclusive Royalty free Right*) atas penelitian skripsi saya yang berjudul:

MODEL PREDIKSI DIABETES MENGGUNAKAN GAUSSIAN PROCESS CLASSIFIER PADA DATA KLINIS

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Non-Eksekutif ini, Universitas Muhammadiyah Sumatera Utara berhak menyimpan, mengalih media, memformat, mengelola dalam bentuk database, merawat dan mempublikasikan Skripsi saya ini tanpa meminta izin dari saya selama tetap mencantumkan nama saya sebagai penulis dan sebagai pemegang dan atau sebagai pemilik hak cipta.

Demikian pernyataan ini dibuat dengan sebenarnya.

Medan, 07 September 2026
Yang membuat pernyataan


Edil Berto Zamili
NPM.2209020248

RIWAYAT HIDUP

DATA PRIBADI

Nama Lengkap : EDIL BERTO ZAMILI
Tempat dan Tanggal Lahir : Telukdalam, 26 September 2005
Alamat Rumah : Jl. R.A. Kartini No.11
Telepon/Faks/HP : 082273411664
E-mail : edilzamili2005@gmail.com
Instansi Tempat Kerja : -
Alamat Kantor : -

DATA PENDIDIKAN

SD : SD Negeri No. 075069 Botohilisilambo TAMAT: 2016
SMP : SMP Negeri 1 Telukdalam TAMAT: 2019
SMA : SMK Negeri 1 Telukdalam TAMAT: 2022

KATA PENGANTAR



Penulis mengucapkan rasa syukur dan terima kasih kepada Allah SWT atas segala rahmat, karunia, serta kemudahan yang telah diberikan sehingga skripsi ini dapat diselesaikan. Shalawat dan salam semoga senantiasa tercurah kepada Nabi Muhammad SAW yang menjadi suri teladan bagi seluruh umat. Semoga segala jerih payah dan upaya yang telah penulis lakukan dalam menyelesaikan skripsi ini mendapat ridha dan berkah dari Allah SWT serta bermanfaat bagi diri penulis sendiri, bagi dunia pendidikan, dan bagi masyarakat luas. Selain itu, penulis juga mengucapkan terima kasih kepada:

1. Bapak Prof. Dr. Akrim, M.Pd., selaku Rektor Universitas Muhammadiyah Sumatera Utara (UMSU).
2. Bapak Assoc. Prof. Dr. Al-Khowarizmi, S.Kom., M.Kom., selaku Dekan Fakultas Ilmu Komputer dan Teknologi Informasi (FIKTI) UMSU.
3. Ibu Dr. Firahmi Rizky., M.Kom., selaku Wakil Dekan I Fakultas Ilmu Komputer dan Teknologi Informasi.
4. Bapak Mhd. Basri, S.Si., M.Kom., selaku Wakil Dekan III Fakultas Ilmu Komputer dan Teknologi Informasi.
5. Ibu Fatma Sari Hutagalung, S.Kom., M.Kom., selaku Ketua Program Studi Teknologi Informasi.
6. Bapak Okvi Nugroho, S.Kom., M.Kom., selaku Sekretaris Program Studi Teknologi Informasi.

7. Bapak Farid Akbar Siregar, S.Kom., M.Kom., selaku dosen pembimbing akademik.
8. Bapak Dr. Irvan, M.Si., selaku dosen pembimbing skripsi. Penulis mengucapkan terimakasih yang sebesar-besarnya atas waktu, ilmu, arahan, dukungan, serta bimbingan yang telah diberikan selama proses penyusunan skripsi ini. Terimakasih telah membimbing penulis dengan penuh kesabaran sehingga skripsi ini dapat diselesaikan dengan baik.
9. Seluruh Bapak dan Ibu Dosen Program Studi Teknologi Informasi, Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Muhammadiyah Sumatera Utara, atas ilmu dan didikan yang telah diberikan selama perkuliahan.
10. Penulis mengucapkan terima kasih yang tulus dan mendalam kepada teman-teman sekelas penulis di kelas F1 Pagi TI yang selalu peduli, mendukung dan senantiasa membantu penulis serta telah menjadi bagian penting dalam perjalanan akademik penulis.
11. Kepada kedua orang tua dan saudara/i yang senantiasa memberikan doa, kasih sayang, dukungan moral maupun materi, serta menjadi sumber semangat selama saya menempuh pendidikan hingga menyelesaikan Tugas Akhir ini. Terima kasih atas segala pengorbanan, perhatian, dan dukungan yang telah diberikan kepada saya selama proses pendidikan hingga sampai pada tahap ini.
12. Semua pihak yang terlibat langsung ataupun tidak langsung yang tidak dapat penulis ucapkan satu per satu yang telah membantu penyelesaian skripsi ini.

13. Dan terakhir untuk diri penulis sendiri karena selalu mau belajar dan berusaha memberikan yang terbaik disetiap langkah. Terimakasih kepada jiwa yang tetap kuat atas banyaknya proses yang telah dilalui, banyak lelah dan keluh kesah yang dipendam sendiri, banyak hal yang sudah dijalani, dihadapi dan diselesaikan sendiri. Sampai detik ini, kuat karena diri sendiri yang selalu mengusahakan semua hal agar terlihat baik-baik saja. Selamat berpetualang di level kehidupan selanjutnya. Tidak ada yang lebih indah dari menyaksikan proses pertumbuhan diri sendiri. Part of growing up and moving into new chapters is about catch and release. It means, knowing what things to keep and what things to release, I can't carry all things.

MODEL PREDIKSI DIABETES MENGGUNAKAN GAUSSIAN PROCESS CLASSIFIER PADA DATA KLINIS

ABSTRAK

Diabetes Mellitus merupakan penyakit metabolik kronis yang membutuhkan deteksi dini untuk mencegah komplikasi fatal. Penggunaan *Machine Learning* dalam prediksi klinis seringkali hanya memberikan keputusan deterministik tanpa estimasi ketidakpastian (*uncertainty estimation*), padahal informasi probabilitas sangat penting dalam pengambilan keputusan medis. Penelitian ini bertujuan untuk mengimplementasikan algoritma *Gaussian Process Classifier* (GPC) dalam membangun model prediksi risiko diabetes yang tidak hanya akurat tetapi juga menghasilkan probabilitas terkalibrasi. Data yang digunakan berasal dari *Diabetes prediction dataset* dengan memanfaatkan 1.500 sampel yang mencakup 8 fitur klinis. Tahapan pra-pemrosesan data meliputi imputasi, standardisasi (Z-score), *encoding*, dan penyeimbangan kelas data menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE). Berdasarkan hasil evaluasi pengujian dengan fungsi *Radial Basis Function* (RBF) Kernel serta penyesuaian ambang batas (*threshold tuning*) ke titik 0,3558, model berhasil mencapai tingkat Akurasi sebesar 90%, *Recall* 92,30%, Presisi 46,15%, dan *F1-Score* 61,54%. Model ini juga menunjukkan tingkat keandalan probabilitas yang luar biasa dengan skor ROC-AUC mencapai 0,9533 dan nilai *Brier Score* sebesar 0,0633. Hasil ini membuktikan bahwa algoritma GPC sangat layak dan andal untuk direkomendasikan sebagai instrumen skrining medis awal yang mampu meminimalkan rasio *False Negative*. Lebih lanjut, model prediktif ini telah diimplementasikan ke dalam purwarupa *Clinical Decision Support System* (CDSS) interaktif berbasis web.

Kata kunci: Diabetes, *Machine Learning*, *Gaussian Process Classifier*, *Clinical Decision Support System*, Prediksi Risiko.

A DIABETES PREDICTION MODEL USING A GAUSSIAN PROCESS CLASSIFIER ON CLINICAL DATA

ABSTRACT

Diabetes Mellitus is a chronic metabolic disease that requires early detection to prevent fatal complications. The application of Machine Learning in clinical prediction often provides deterministic decisions without uncertainty estimation, whereas probability information is crucial in medical decision-making. This study aims to implement the Gaussian Process Classifier (GPC) algorithm to build a diabetes risk prediction model that is not only accurate but also produces calibrated probabilities. The study utilized the Diabetes prediction dataset, extracting 1,500 samples comprising 8 clinical features. Data preprocessing stages included imputation, standardization (Z-score), encoding, and class balancing using the Synthetic Minority Over-sampling Technique (SMOTE). Based on the evaluation results utilizing the Radial Basis Function (RBF) Kernel and threshold tuning to 0.3558, the model successfully achieved an Accuracy of 90%, Recall of 92.30%, Precision of 46.15%, and F1-Score of 61.54%. The model also demonstrated outstanding probability reliability with an ROC-AUC score of 0.9533 and a Brier Score of 0.0633. These findings prove that the GPC algorithm is highly reliable and recommended as an initial medical screening instrument capable of minimizing False Negatives. Furthermore, this predictive model has been successfully implemented into an interactive web-based Clinical Decision Support System (CDSS) prototype.

Keywords: Diabetes, Machine Learning, Gaussian Process Classifier, Clinical Decision Support System, Risk Prediction.

DAFTAR ISI

LEMBAR PENGESAHAN	i
PERNYATAAN ORISINALITAS.....	ii
PERNYATAAN PERSETUJUAN PUBLIKASI.....	iii
RIWAYAT HIDUP	iv
KATA PENGANTAR.....	v
ABSTRAK	viii
ABSTRACT	ix
DAFTAR ISI.....	x
DAFTAR TABEL	xii
DAFTAR GAMBAR.....	xiii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Batasan Masalah.....	4
1.4 Tujuan Penelitian.....	4
1.5 Manfaat Penelitian.....	5
BAB II LANDASAN TEORI	6
2.1 Diabetes Mellitus.....	6
2.2 Parameter Klinis dalam Prediksi Diabetes	6
2.3 <i>Machine Learning</i> dan Klasifikasi	7
2.4 <i>Gaussian Process Classifier</i> (GPC)	7
2.5 Confusion Matrix	10
2.6 Tinjauan Pustaka	11
2.7 Analisis Gap	14
BAB III METODOLOGI PENELITIAN	15
3.1 Pendekatan Penelitian.....	15
3.2 Alur Penelitian.....	15
3.3 Alat dan Bahan Penelitian	17
3.3.1 Perangkat Keras (<i>Hardware</i>).....	17
3.3.2 Perangkat Keras (<i>Hardware</i>).....	17
3.4 Variabel Penelitian	18
3.5 Teknik Pengumpulan Data	18
3.6 Tahapan dan Proses Analisis Data	19
3.7 Alur Sistem Web	22
3.8 Rancangan Antarmuka	23
3.9 Proses Penafsiran dan Penyimpulan Hasil	25
BAB IV HASIL DAN PEMBAHASAN	27
4.1 Gambaran Umum	27
4.2 Import Dataset	27
4.3 Pembersihan Data (Data Cleaning)	28
4.4 Pra-pemrosesan Data	29
4.5 Modelling	31
4.6 Gaussian Process Classifier (GPC)	32
4.6.1 Pelatihan Model GPC dan Optimasi Kernel GPC	32
4.6.2 Validasi Silang (Cross-Validation) dan Penentuan Threshold	33
4.6.3 Evaluasi Pengujian Model GPC	34
4.6.4 Confusion Matrix GPC	36

4.6.5 Perhitungan Evaluasi Model GPC	36
4.6.6 Analisis Keandalan Probabilitas	38
4.7 Implementasi Antarmuka	39
4.7.1 Halaman Dashboard.....	39
4.7.2 Hasil Analisis dan Visualisasi.....	40
4.7.3 Tabel Riwayat Log Pasien	41
BAB V KESIMPULAN DAN SARAN	43
5.1 Kesimpulan.....	43
5.2 Saran	44
DAFTAR PUSTAKA	

DAFTAR TABEL

Tabel 2.1 Confusion Matrix	10
Tabel 2.2 Ringkasan Penelitian Terdahulu	12
Tabel 3.1 Perangkat Keras	17
Tabel 3.2 Perangkat Lunak	17
Tabel 4.1 Confusion Matrix GPC	36

DAFTAR GAMBAR

Gambar 3.1 Diagram Alur Penelitian.....	16
Gambar 3.2 Alur Sistem Web Prediksi Diabetes	22
Gambar 3.3 Tampilan Awal Web	23
Gambar 3.4 Tampilan dengan Hasil Positif.....	24
Gambar 3.5 Tampilan dengan Hasil Negatif.....	25
Gambar 4.1 Import Dataset.....	28
Gambar 4.2 Pembersihan Data.....	28
Gambar 4.3 Fitur Numerik.....	29
Gambar 4.4 Fitur Kategorikal	30
Gambar 4.5 Penyeimbangan Data.....	31
Gambar 4.6 Memisahkan Fitur	31
Gambar 4.7 Pelatihan Model GPC.....	32
Gambar 4.8 Melatih Model.....	33
Gambar 4.9 Hasil Threshold Tuning Data Latih.....	34
Gambar 4.10 Classification Report Threshold 0.5.....	34
Gambar 4.11 Classification Report Threshold 0.3558.....	35
Gambar 4.12 Kurva Kalibrasi (Reliability Diagram) GPC.....	38
Gambar 4.13 Halaman Dashboard (Input Pasien).....	40
Gambar 4.14 Hasil Klasifikasi dan Probabilitas Klinis	41

BAB I

PENDAHULUAN

1.1 Latar Belakang

Diabetes Mellitus merupakan salah satu penyakit metabolik kronis dengan prevalensi yang terus meningkat secara global dan menjadi ancaman serius bagi sektor kesehatan masyarakat. Penyakit ini ditandai dengan tingginya kadar glukosa dalam darah yang apabila tidak dideteksi dan ditangani sejak dini, dapat memicu berbagai komplikasi fatal seperti gagal ginjal, penyakit kardiovaskular, hingga kebutaan (Ali Küsmüş, 2025). Oleh karena itu, langkah deteksi dan skrining awal menjadi sangat krusial untuk mencegah perburukan kondisi pasien, menekan angka morbiditas, serta menurunkan beban biaya perawatan kesehatan.

Seiring dengan kemajuan teknologi informasi di bidang medis, pemanfaatan *Machine Learning* telah membuka dimensi baru dalam menganalisis rekam medis elektronik. Data klinis yang memuat berbagai indikator pasien seperti usia, Indeks Massa Tubuh (BMI), riwayat hipertensi, fungsi kardiovaskular, serta parameter kunci seperti tingkat HbA1c dan glukosa darah, menyimpan pola tersembunyi yang kompleks (Sahid et al., 2024). Pendekatan analitik berbasis komputasi memungkinkan tenaga medis untuk memproses fitur-fitur klinis tersebut guna memprediksi risiko diabetes secara lebih cepat, objektif, dan presisi dibandingkan dengan metode diagnosis konvensional secara manual.

Dalam ranah *Machine Learning*, tantangan utama pemodelan data medis tidak hanya terletak pada pencapaian akurasi tinggi, melainkan juga pada kemampuan model dalam merepresentasikan tingkat kepercayaan terhadap suatu

prediksi. Algoritma *Gaussian Process Classifier* (GPC) hadir sebagai metode probabilistik non-parametrik yang sangat tangguh dalam memodelkan hubungan non-linear antar variabel klinis (Yakut, 2023). Keunggulan utama GPC adalah kemampuannya dalam memberikan probabilitas prediksi kelas dan estimasi ketidakpastian (*uncertainty estimation*), sebuah fitur yang sangat dibutuhkan agar prediksi yang dihasilkan aman digunakan sebagai rujukan pengambilan keputusan klinis oleh dokter (Yothapakdee et al., 2022).

Namun, meskipun berbagai penelitian telah memanfaatkan algoritma Machine Learning untuk prediksi diabetes, sebagian besar pendekatan tersebut masih bersifat deterministik dan hanya menghasilkan keputusan kelas (positif/negatif) tanpa menyertakan tingkat keyakinan prediksi. Kondisi ini menjadi kelemahan signifikan dalam konteks medis yang sangat membutuhkan estimasi risiko yang lebih transparan dan terkalibrasi. Selain itu, metode probabilistik seperti Gaussian Process Classifier sebagian besar baru diuji pada dataset berskala kecil seperti PIMA Indian, sehingga efektivitasnya pada dataset klinis modern berskala besar belum banyak dieksplorasi. Celah penelitian ini menunjukkan perlunya penerapan GPC pada dataset yang lebih besar dan kaya fitur untuk menghasilkan model prediktif yang lebih informatif dan andal.

Penelitian ini akan mengimplementasikan algoritma GPC pada *Diabetes prediction dataset* yang disusun oleh Mohammed Mustafa, yang berisi ratusan ribu data klinis dan demografis pasien dengan rekam medis yang merepresentasikan kasus nyata (Mustafa, 2023). Kombinasi antara ketangguhan

algoritma GPC dengan kekayaan informasi dari dataset klinis tersebut diharapkan dapat menghasilkan model prediktif yang andal.

Dengan mempertimbangkan kompleksitas diabetes sebagai penyakit kronis dan meningkatnya kebutuhan akan sistem deteksi dini yang akurat serta aman digunakan dalam lingkungan klinis, penelitian ini menjadi penting untuk dilakukan. Implementasi Gaussian Process Classifier tidak hanya berpotensi meningkatkan akurasi prediksi, tetapi juga memberikan estimasi ketidakpastian yang sangat krusial untuk pengambilan keputusan berbasis risiko. Oleh karena itu, penelitian ini diharapkan memberikan kontribusi signifikan dalam mendukung pengembangan Clinical Decision Support System (CDSS) yang lebih transparan, presisi, dan berorientasi keselamatan pasien. Berdasarkan pemaparan tersebut, penelitian ini disusun dengan judul "**Model Prediksi Diabetes Menggunakan Gaussian Process Classifier pada Data Klinis**".

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah di atas, rumusan masalah dalam penelitian ini adalah:

1. Bagaimana mengimplementasikan algoritma Gaussian Process Classifier (GPC) untuk membangun model prediksi risiko diabetes berdasarkan parameter data klinis?
2. Bagaimana kinerja model Gaussian Process Classifier (meliputi akurasi, presisi, recall, dan F1-score) ketika dievaluasi menggunakan Diabetes prediction dataset?

1.3 Batasan Masalah

Untuk menjaga agar penelitian ini tetap fokus dan terarah, batasan masalah ditetapkan sebagai berikut:

1. Dataset: Penelitian ini menggunakan data sekunder yaitu Diabetes prediction dataset yang dipublikasikan oleh Mohammed Mustafa pada platform Kaggle.
2. Fitur Data: Variabel prediktor yang dianalisis mencakup 8 atribut: gender, age, hypertension, heart_disease, smoking_history, bmi, HbA1c_level, dan blood_glucose_level. Target klasifikasi adalah kelas biner (1 untuk positif diabetes, 0 untuk negatif).
3. Algoritma: Pemodelan Machine Learning sepenuhnya difokuskan pada penggunaan algoritma Gaussian Process Classifier (GPC).
4. Evaluasi: Kinerja model akan diukur berdasarkan Confusion Matrix dengan metrik Accuracy, Precision, Recall, F1-Score, serta Area Under Curve (AUC).

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dipaparkan, tujuan dari penelitian ini adalah:

1. Membangun model prediksi penyakit diabetes menggunakan metode probabilistik Gaussian Process Classifier berbasis fitur klinis.
2. Mengukur dan menganalisis performa klasifikasi algoritma Gaussian Process Classifier pada dataset diabetes untuk membuktikan kelayakannya sebagai instrumen skrining medis.

1.5 Manfaat Penelitian

Manfaat Teoritis: Penelitian ini diharapkan dapat memberikan kontribusi pada literatur ilmu komputer dan informatika medis, khususnya memperkaya kajian terkait efektivitas metode probabilistik Gaussian Process Classifier dalam menangani dataset klinis untuk penyakit kronis.

Manfaat Praktis: Bagi tenaga medis dan institusi kesehatan, model klasifikasi yang dihasilkan dapat dijadikan sebagai dasar pengembangan Clinical Decision Support System (CDSS). Sistem ini akan membantu mempercepat proses skrining awal dan memprioritaskan pasien yang memiliki risiko tinggi terkena diabetes berdasarkan indikator medis dasar mereka.

BAB II

LANDASAN TEORI

2.1 Diabetes Mellitus

Diabetes Mellitus (DM) adalah kelompok penyakit metabolik yang dicirikan oleh kondisi hiperglikemia (kadar gula darah tinggi) yang terjadi akibat kelainan sekresi insulin, kerja insulin, atau keduanya. Menurut World Health Organization (WHO), hiperglikemia kronis pada diabetes berkaitan erat dengan kerusakan jangka panjang, disfungsi, serta kegagalan berbagai organ tubuh, terutama mata, ginjal, saraf, jantung, dan pembuluh darah (WHO, 2024). Klasifikasi diagnosis sering kali merujuk pada standar American Diabetes Association (ADA, 2022).

2.2 Parameter Klinis dalam Prediksi Diabetes

Berdasarkan Diabetes prediction dataset (Mustafa, 2023), prediksi risiko diabetes diukur menggunakan beberapa parameter klinis kunci:

1. **HbA1c (*Hemoglobin A1c*)**: Parameter yang mengukur persentase hemoglobin yang berikatan dengan glukosa, merepresentasikan rata-rata kadar glukosa darah dalam 2-3 bulan terakhir. Standar diagnosis menetapkan angka $\geq 6.5\%$ sebagai indikator diabetes.
2. **Kadar Glukosa Darah (*Blood Glucose Level*)**: Konsentrasi gula di dalam darah. Glukosa puasa ≥ 126 mg/dL adalah indikator kuat DM.
3. **Indeks Massa Tubuh (BMI)**: Rasio berat dan tinggi badan yang merupakan faktor risiko utama resistensi insulin.
4. **Hipertensi dan Penyakit Jantung**: Penyakit komorbiditas yang sering memperburuk sindrom metabolik pasien.

2.3 Machine Learning dan Klasifikasi

Machine Learning adalah cabang dari kecerdasan buatan yang memungkinkan sistem komputer untuk belajar dari data dan meningkatkan kinerjanya dalam melakukan tugas spesifik tanpa diprogram secara eksplisit (Bishop, 2006). Dalam konteks rekam medis pasien yang memiliki label pasti (positif atau negatif diabetes), pendekatan yang digunakan adalah *Supervised Learning*.

Klasifikasi adalah sub-kategori dari Supervised Learning di mana tujuan utamanya adalah memetakan variabel input (fitur klinis) ke dalam variabel target diskrit (kelas label). Pada penelitian ini, masalah diformulasikan sebagai klasifikasi biner, di mana output $y \in \{0,1\}$.

2.4 Gaussian Process Classifier (GPC)

Gaussian Process (GP) pada dasarnya adalah pendekatan probabilistik non parametrik berbasis Teorema Bayes yang digunakan untuk regresi. Sebuah GP mendefinisikan distribusi atas fungsi-fungsi, dengan inferensi dilakukan langsung pada ruang fungsi (Rasmussen & Williams, 2006). GP secara matematis didefinisikan oleh fungsi mean $m(x)$ dan fungsi covariance atau kernel $k(x, x')$:

$$f(x) \sim GP(m(x), k(x, x')) \quad (1)$$

Keterangan:

$f(x) \sim$: Menyatakan bahwa fungsi target $f(x)$ berdistribusi secara *probabilistik* mengikuti **distribusi Gaussian (normal)** secara *joint* (gabungan)

GP : Singkatan dari **Gaussian Process**, yaitu sekumpulan variabel acak berhingga yang berdistribusi Gaussian (normal).

$m(x)$: **Fungsi Rata-rata (*Mean Function*)**, yang merepresentasikan nilai rata-rata yang diharapkan dari fungsi $f(x)$ pada masukan x . Seringkali dalam praktiknya, nilai *mean* ini diasumsikan bernilai nol ($m(x) = 0$) untuk mempermudah perhitungan.

$k(x, x')$: **Fungsi Kernel atau Kovarians (*Covariance/Kernel Function*)**, yang mengukur tingkat kemiripan atau keterkaitan antara dua titik data masukan x dan x' . Fungsi ini menentukan seberapa mulus (*smooth*) atau kompleks pola yang dapat dipelajari oleh model.

Untuk masalah klasifikasi biner, rentang keluaran fungsi laten $f(x)$ dari GP (yang bernilai kontinu dari $-\infty$ sampai ∞) tidak dapat langsung digunakan sebagai probabilitas. Oleh karena itu, Gaussian Process Classifier (GPC) menekan fungsi laten tersebut agar berada dalam rentang $[0,1]$ dengan memanfaatkan fungsi *link* seperti fungsi logistik atau *probit* (Murphy, 2012). Jika menggunakan fungsi logistik, probabilitas kelas target didefinisikan sebagai:

$$p(y = 1 | f) = \frac{1}{1 + \exp(-f)} \quad (2)$$

Keterangan :

$p(y = 1 | f)$: Probabilitas kondisional bahwa kelas target y bernilai 1, diberikan nilai fungsi laten f pada titik masukan tertentu.

$\exp(-f)$ (atau e^{-f}): Fungsi eksponensial natural dari negatif nilai fungsi laten f .

Secara keseluruhan, fungsi di sisi kanan adalah **fungsi sigmoid** ($\sigma(f)$) yang berfungsi untuk merentangkan atau memetakan nilai bilangan real apa pun dari fungsi laten $-\infty < f < +\infty$ agar menghasilkan nilai probabilitas di antara interval tertutup $[0,1]$.

Salah satu elemen krusial dalam GPC adalah pemilihan fungsi *covariance* (kernel). Kernel *Radial Basis Function* (RBF) atau *Squared Exponential* sering digunakan karena kemampuannya memodelkan fungsi yang mulus tak terhingga. Fungsi kernel RBF didefinisikan sebagai:

$$k(x_i, x_j) = \sigma^2 \exp\left(-\frac{\|x_i - x_j\|^2}{2l^2}\right) \quad (3)$$

Keterangan :

$k(x_i, x_j)$: Nilai kovarians atau tingkat kemiripan antara titik data ke- i (x_i) dan titik data ke- j (x_j). Nilainya berkisar antara 0 (tidak mirip sama sekali) hingga σ^2 (sangat identik).

σ^2 : Parameter **varians sinyal** (*signal variance*) atau amplitudo, yang menentukan skala vertikal dari variasi fungsi.

$\exp\left(-\frac{\|x_i - x_j\|^2}{2l^2}\right)$: Fungsi eksponensial yang menghasilkan nilai semakin mendekati 1 jika jarak antar-titik semakin dekat, dan mendekati 0 jika jaraknya sangat jauh.

$\|x_i - x_j\|^2$: Kuadrat dari **norma Euclidean (jarak)** antara vektor fitur x_i dan x_j , yang dihitung sebagai jumlah selisih kuadrat dari setiap fitur penyusunnya.

l : Parameter **panjang skala** (*length-scale*), yang menentukan seberapa jauh jangkauan pengaruh dari suatu titik data (mengontrol tingkat kemulusan/*smoothness* fungsi).

di mana σ^2 merepresentasikan varians dari sinyal, l adalah parameter *length-scale* yang mengontrol seberapa cepat fungsi dapat berubah, dan $\|x_i - x_j\|$ adalah jarak Euclidean antar fitur pasien (Rasmussen & Williams, 2006). Keunggulan utama GPC adalah kemampuannya dalam menyediakan estimasi ketidakpastian

(*uncertainty estimation*), di mana prediksi klasifikasi disertai dengan derajat keyakinan probabilitas, suatu hal yang esensial dalam domain medis.

2.5 Confusion Matrix

Confusion Matrix adalah tabel yang digunakan untuk mengevaluasi kinerja model klasifikasi dengan membandingkan antara hasil prediksi dan label aktual. Matriks ini menyajikan empat komponen utama yang menggambarkan jumlah prediksi benar dan salah untuk masing-masing kelas. Contoh *confusion matrix* ditunjukkan pada Tabel 2.1

Tabel 2.1 Confusion Matrix

	Prediksi Positive	Prediksi Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Keterangan:

TP (*True Positive*): data *positive* diprediksi *positive*

TN (*True Negative*): data *negative* diprediksi *negative*

FP (*False Positive*): data *Negative* diprediksi *Positive*

FN (*False Negative*): data *positive* diprediksi *negative*

Rumus *confusion matrix* untuk menghitung *accuracy*, *precision*, *recall* dan *F1-score* seperti berikut:

1. **Akurasi (*Accuracy*)**: Rasio prediksi benar atas keseluruhan data.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (4)$$

2. **Presisi (Precision):** Proporsi pasien yang diprediksi positif diabetes yang benar-benar menderita diabetes.

$$Precision = \frac{TP}{(TP + FP)} \quad (5)$$

3. **Recall (Sensitivitas):** Kemampuan model mendeteksi pasien yang menderita diabetes secara aktual. Dalam konteks medis, meminimalkan nilai *FN* sangat penting agar penderita tidak terabaikan.

$$Recall = \frac{TP}{(TP + FN)} \quad (6)$$

4. **F1-Score:** Rata-rata harmonik antara Presisi dan *Recall*, sangat relevan untuk dataset klinis yang seringkali tidak seimbang (*imbalanced*).

$$F1 - Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (7)$$

2.6 Tinjauan Pustaka

Pemanfaatan *Machine Learning* dalam upaya prediksi penyakit diabetes telah banyak dilakukan oleh para peneliti sebelumnya dengan mengeksplorasi berbagai algoritma dan dataset klinis. Kajian terhadap penelitian terdahulu sangat penting untuk mengetahui perkembangan terkini (*state of the art*) serta mengidentifikasi celah atau kekurangan yang dapat diperbaiki. Berikut adalah ringkasan beberapa penelitian terdahulu yang relevan dengan topik prediksi diabetes, disajikan pada Tabel 2.2.

Tabel 2.2 Ringkasan Penelitian Terdahulu

No	Judul dan Peneliti	Pembahasan	Metode	Kelebihan dan Kekurangan
1	<i>Application of the Random Forest Method for Classifying Diabetes</i> (Yuniarsih et al., 2025)	Menggunakan dataset klinis 100.000 data dari Kaggle (Mohammed Mustafa) untuk mengklasifikasi pasien diabetes.	<i>Random Forest</i>	Kelebihan: Mencapai akurasi 90,70% pada dataset klinis berskala besar. Kekurangan: Output berupa <i>hard classification</i> deterministik tanpa memberikan persentase probabilitas kepastian diagnostik.
2	<i>A Proposed Technique Using Machine Learning for the Prediction of Diabetes Disease through a Mobile App</i> (El-Sofany et al., 2024)	Fokus pada penanganan data medis yang tidak seimbang (<i>imbalanced</i>) dalam memprediksi status diabetes pasien.	<i>XGBoost + SMOTE</i>	Kelebihan: SMOTE sangat baik dalam menyeimbangkan data minoritas, mencapai F1-score tinggi. Kekurangan: <i>XGBoost</i> bertindak sebagai "kotak hitam" (<i>black box</i>) yang sulit diinterpretasi oleh dokter klinis.

3	<i>Hybrid Machine Learning Framework for Type 2 Diabetes Prediction (Zhang et al., 2025)</i>	Menerapkan beberapa metode klasifikasi tingkat lanjut pada dataset medis PIMA Indian.	<i>Gaussian Process Classification (GPC), Hard Voting Classifier</i>	Kelebihan: Membuktikan GPC sangat valid untuk diagnosis medis probabilistik. Kekurangan: Hanya dievaluasi pada dataset PIMA yang berukuran sangat kecil (768 baris) dan fiturnya kurang komprehensif.
4	<i>An Advanced Machine Learning Approach for Enhanced Diabetes Prediction (Nooraldaïm et al., 2024)</i>	Studi kombinasi algoritma klastering dan klasifikasi untuk meningkatkan reliabilitas model prediksi diabetes.	<i>KNN, Random Forest</i>	Kelebihan: Mempertahankan hubungan intrinsik antar <i>data points</i> medis dengan baik. Kekurangan: Kompleksitas komputasi yang tinggi dengan gabungan berbagai algoritma secara <i>pipeline</i>.

2.7 Analisis Gap

Berdasarkan tinjauan pustaka di atas, terlihat jelas sebuah celah (*gap*) dalam literatur prediksi diabetes terkini. Penelitian yang menggunakan dataset berskala besar dan komprehensif seperti *Diabetes prediction dataset* karya Mohammed Mustafa (2023) umumnya masih berfokus pada algoritma berbasis pohon keputusan (*tree-based*) yang bersifat deterministik, seperti *Random Forest* (sebagaimana diterapkan oleh penelitian (Yuniarsih et al., 2025) atau *XGBoost*. Algoritma ini handal dalam akurasi, namun hanya menghasilkan vonis kelas (1 atau 0) tanpa memberikan probabilitas matematis sejauh mana model tersebut "yakin" atas keputusannya.

Di sisi lain, penelitian dari jurnal *Informatica* (2025) telah membuktikan bahwa *Gaussian Process Classifier* (GPC) memiliki potensi besar untuk diagnosis medis karena kemampuannya menghasilkan probabilitas prediksi terkalibrasi (*calibrated uncertainty estimation*). Sayangnya, implementasi GPC tersebut baru diuji coba pada dataset tradisional berukuran kecil (seperti PIMA Indian) dan belum diuji pada dataset klinis modern berskala masif yang memuat fitur esensial seperti *HbA1c*.

Oleh karena itu, penelitian ini dirancang untuk **menjembatani celah tersebut**. Penelitian ini akan mengimplementasikan algoritma probabilistik non-parametrik *Gaussian Process Classifier* (GPC) pada dataset skala besar *Diabetes prediction dataset* (Mustafa, 2023). Hasilnya diharapkan tidak hanya membuahkan akurasi dan *recall* yang bersaing, tetapi juga menyajikan model diagnostik yang mampu mengukur *uncertainty* probabilitas, sehingga jauh lebih aman dan transparan jika kelak digunakan sebagai *Clinical Decision Support System* (CDSS) oleh praktisi medis.

BAB III

METODOLOGI PENELITIAN

3.1 Pendekatan Penelitian

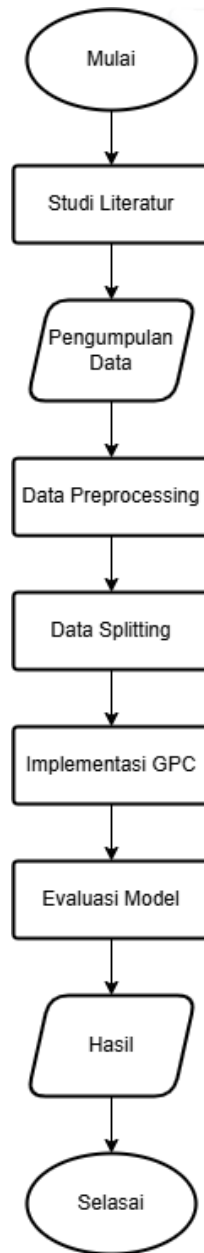
Penelitian ini menggunakan pendekatan **kuantitatif eksperimental** di bidang ilmu komputer dan informatika medis. Eksperimen difokuskan pada perancangan, pelatihan, dan pengujian model komputasi berbasis algoritma *Machine Learning* untuk memprediksi risiko penyakit diabetes. Kesimpulan ditarik berdasarkan hasil pengukuran metrik kinerja secara objektif dan matematis.

3.2 Alur Penelitian

Penelitian ini dilaksanakan melalui tahapan-tahapan yang sistematis untuk memastikan pemodelan klasifikasi berjalan dengan valid dan terukur.

1. **Pengumpulan Data:** Mengakuisisi dataset klinis sekunder dan mengambil subset data yang sesuai dengan batasan komputasi.
2. **Data Preprocessing:** Membersihkan, mentransformasi, menyeimbangkan, dan menormalkan data agar siap diproses oleh algoritma komputer.
3. **Pembagian Data:** Membagi dataset menjadi data latih untuk proses pembelajaran dan data uji untuk proses validasi.
4. **Implementasi GPC:** Membangun, melatih, dan mengoptimalkan model *Gaussian Process Classifier* menggunakan data latih untuk mengenali pola diabetes.
5. **Evaluasi Kinerja Model:** Mengukur performa model dalam memprediksi data uji menggunakan berbagai metrik evaluasi medis.

Secara visual, tahapan penelitian ini dapat digambarkan melalui diagram alur berikut:



Gambar 3.1 Diagram Alur Penelitian

Diagram Alur Penelitian pada Gambar 3.1 di atas memvisualisasikan langkah-langkah yang dilalui dalam penelitian ini, yang mencakup tahap studi literatur, pengumpulan data, pra-pemrosesan, pembagian data (data splitting), implementasi algoritma GPC, tahap evaluasi model, hingga penarikan hasil.

3.3 Alat dan Bahan Penelitian

Pada proses analisis ini dijelaskan alat apa saja yang digunakan dalam penelitian yang terbagi menjadi perangkat keras (*hardware*) dan perangkat lunak (*software*).

3.3.1 Perangkat Keras (*Hardware*)

Menggunakan Laptop ACER dengan spesifikasi sebagai berikut :

Tabel 3.1 Perangkat Keras

Komponen	Spesifikasi
Model Sistem	Nitro ANV15-51
Processor	13th Gen Intel(R) Core(TM) i5-13420H (2.10 GHz)
RAM/Storage	8 GB/477 GB SSD
Sistem Operasi	Windows 11

3.3.2 Perangkat Keras (*Hardware*)

Dalam penelitian ini digunakan beberapa perangkat lunak antara lain sebagai berikut:

Tabel 3.2 Perangkat Lunak

Jenis Perangkat Lunak	Nama Perangkat Lunak
Sistem Operasi	Windows 11
Platform Pemograman	<i>Google Colab</i> dan Visual Studio Code
Bahasa Pemograman	Python

3.4 Variabel Penelitian

Dalam konteks *Supervised Learning*, dataset klinis yang digunakan dipecah menjadi dua jenis variabel utama:

1. **Variabel Independen (Fitur Prediktor / X):** Atribut klinis dan demografis pasien yang mempengaruhi hasil prediksi. Variabel ini mencakup 8 fitur: *gender*, *age*, *hypertension*, *heart_disease*, *smoking_history*, *bmi*, *HbA1c_level*, dan *blood_glucose_level*.
2. **Variabel Dependen (Target Kelas / y):** Keluaran yang diprediksi oleh model. Variabel ini bersifat biner, yaitu kelas 0 (pasien diprediksi negatif/sehat) dan kelas 1 (pasien diprediksi positif menderita diabetes).

3.5 Teknik Pengumpulan Data

Metode pengumpulan data yang digunakan adalah **pengumpulan data sekunder**. Data diperoleh dengan cara mengunduh *Diabetes prediction dataset* secara publik melalui repositori Kaggle yang dipublikasikan oleh Mohammed Mustafa. Karena data telah dikumpulkan dan dianonimisasi oleh penyedia dataset, penelitian ini tidak memerlukan proses pengambilan data primer secara langsung ke rumah sakit.

Dataset asli *Diabetes prediction* memiliki total 100.000 baris data. Namun, algoritma *Gaussian Process Classifier* (GPC) melibatkan inversi matriks kovarians yang memiliki kompleksitas komputasi $O(N^3)$. Penggunaan seluruh populasi data akan menyebabkan beban komputasi yang tidak efisien dan melebihi kapasitas memori standar.

Oleh karena itu, penelitian ini menerapkan teknik pengurangan data menjadi sekitar 1.000 hingga 1.500 sampel. Untuk memastikan sampel yang diambil tetap merepresentasikan populasi asli dan tidak menghilangkan pola penyakit, teknik yang digunakan adalah ***Stratified Random Sampling***. Teknik ini menjamin bahwa rasio atau proporsi antara kelas minoritas (pasien positif diabetes) dan kelas mayoritas (pasien negatif) di dalam sampel 1.500 data tersebut tetap sama persis dengan proporsi pada dataset asli yang berjumlah 100.000 baris, sehingga model terhindar dari bias kelas.

3.6 Tahapan dan Proses Analisis Data

Proses komputasi dan analisis informasi dilakukan melalui alur sistematis (*pipeline*) sebagai berikut:

1. Pra-pemrosesan Data (*Data Preprocessing*)

Tahap ini krusial untuk memastikan algoritma *Gaussian Process Classifier* (GPC) dapat mencerna data dengan optimal:

- a. **Pembersihan Data (*Data Cleaning*):** Menghapus data ganda (*duplicates*) dan menangani *missing values* jika ditemukan.
- b. **Transformasi Data (*Encoding*):** Mengubah data kategorikal (seperti *gender* dan *smoking_history*) menjadi representasi numerik menggunakan teknik *Label Encoding* atau *One-Hot Encoding*.
- c. **Penanganan Data Tidak Seimbang (*Imbalanced Data Handling*):**
 Dalam data medis, jumlah pasien negatif diabetes umumnya jauh lebih besar dibandingkan pasien positif. Untuk mencegah model memiliki bias terhadap kelas mayoritas, diterapkan teknik *Synthetic Minority Over-sampling Technique* (SMOTE). SMOTE bekerja dengan

mensintesis data baru pada kelas minoritas (positif diabetes) berdasarkan jarak terdekat antar titik data (*k-nearest neighbors*), sehingga jumlah kelas 0 dan 1 menjadi seimbang sebelum dimasukkan ke dalam model.

- d. **Normalisasi Fitur (*Feature Scaling*):** Karena algoritma *Gaussian Process Classifier* (GPC) beroperasi berdasarkan perhitungan jarak menggunakan fungsi *kernel* untuk mengukur kemiripan antar titik data, maka seluruh variabel independen (fitur klinis) harus diskalakan agar memiliki rentang yang seragam. Jika tidak dilakukan penyesuaian skala, fitur dengan magnitudo nilai absolut yang besar, seperti kadar glukosa darah (*Blood Glucose Level*), akan mendominasi perhitungan jarak dan menyamarkan signifikansi fitur dengan magnitudo kecil, seperti persentase *HbA1c*. Oleh karena itu, pada tahap pra-pemrosesan data, diterapkan teknik standardisasi menggunakan perhitungan Z-score. Rumus matematis yang digunakan adalah sebagai berikut:

$$z = \frac{x - \mu}{\sigma} \quad (8)$$

Keterangan:

z : Representasi nilai standar (Z-score) pasca-transformasi.

x : Nilai observasi asli dari suatu fitur klinis.

μ : Nilai rata-rata (*mean*) dari keseluruhan observasi pada fitur terkait.

σ : Simpangan baku (*standard deviation*) dari distribusi fitur tersebut.

Penerapan Z-score secara fundamental mentransformasi data klinis sehingga memiliki distribusi dengan nilai rata-rata $\mu = 0$ dan standar

deviasi $\sigma = 1$. Standardisasi memastikan bahwa setiap variabel prediktor memberikan kontribusi dan bobot komputasi yang proporsional saat diproses oleh algoritma yang bersifat sensitif terhadap skala, seperti *Gaussian Processes*. Nilai Z-score secara inheren mengukur seberapa jauh jarak suatu titik data (dalam satuan standar deviasi) dari titik pusat rata-ratanya. Berdasarkan aturan empiris (*Rule of Thumb*), observasi medis yang memiliki rentang nilai $z < -3$ atau $z > 3$ diklasifikasikan sebagai pencilan ekstrem yang berpotensi mendistorsi konvergensi model, sehingga dapat dievaluasi lebih lanjut pada tahap pembersihan data. Transformasi ini memfasilitasi perbandingan analitis yang valid antar distribusi pengukuran indikator medis yang pada dasarnya memiliki satuan dan rentang nilai komparatif yang berbeda secara fundamental.

2. Pembagian Data (*Data Splitting*)

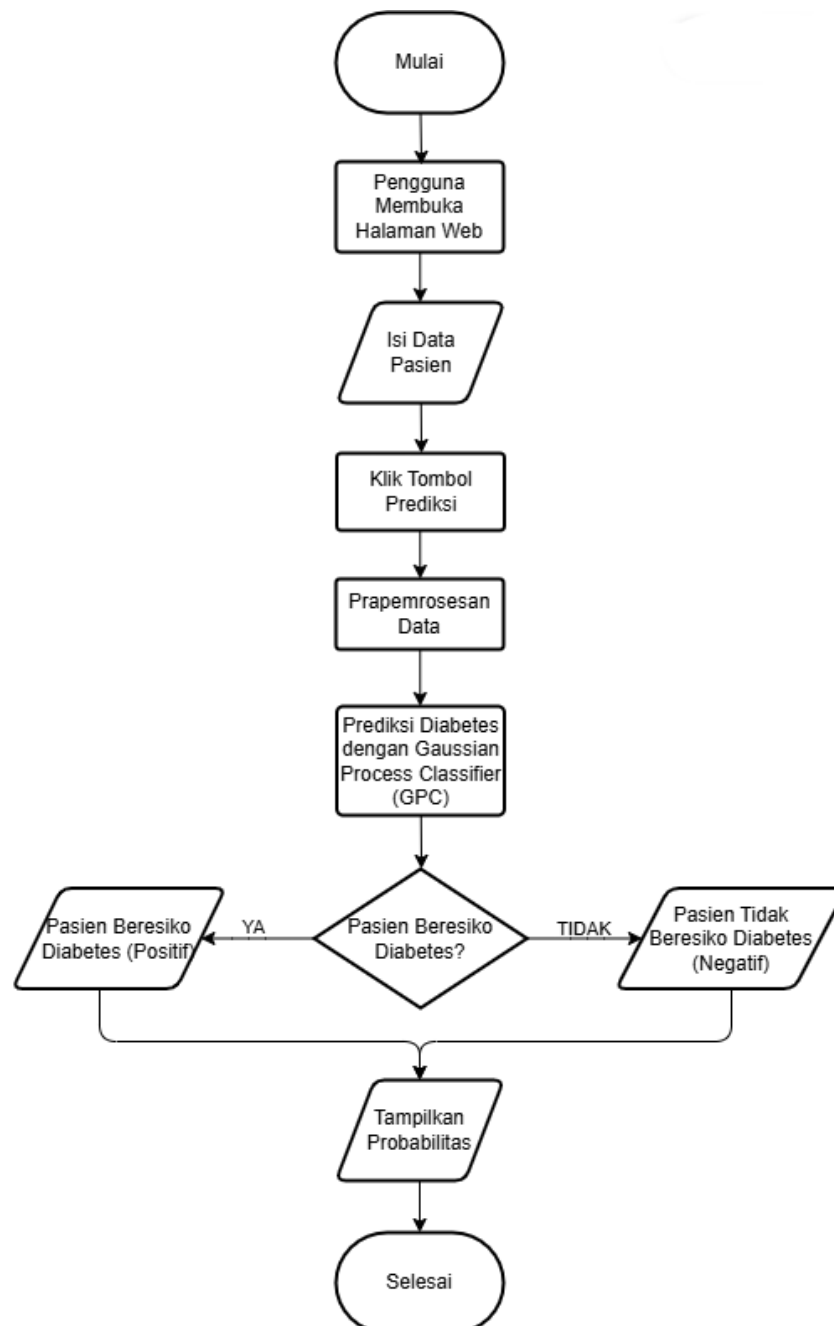
Dataset dibagi menjadi dua bagian terpisah untuk menghindari bias evaluasi (*overfitting*). Pendekatan yang umum digunakan adalah pembagian 80% sebagai **Data Latih** (*Training Set*) untuk mengajari model, dan 20% sebagai **Data Uji** (*Testing Set*) untuk mengevaluasi kinerja model pada data yang belum pernah dilihat sebelumnya.

3. Pemodelan dengan Gaussian Process Classifier (GPC)

Tahap inti di mana Data Latih diumpankan ke algoritma GPC. Model akan mempelajari pola hubungan non-linear antara fitur klinis dan status diabetes menggunakan fungsi *covariance* berbasis *Radial Basis Function* (RBF) Kernel.

3.7 Alur Sistem Web

Selain alur metode penelitian, dirancang juga sebuah alur kerja untuk sistem web prediksi diabetes guna menggambarkan urutan interaksi pengguna dengan aplikasi dari awal hingga hasil prediksi muncul.



Gambar 3.2 Alur Sistem Web Prediksi Diabetes

Gambar 3.2 menjelaskan alur (flowchart) dari sistem web secara rinci. Proses tersebut dimulai dari tindakan pengguna membuka halaman web, mengisi formulir data indikator pasien, proses komputasi prediksi menggunakan model GPC, hingga sistem akhirnya menampilkan output berupa status berisiko (positif/negatif) beserta persentase probabilitasnya.

3.8 Rancangan Antarmuka

Untuk mendemonstrasikan hasil penelitian secara praktis, model GPC yang telah dilatih dapat diintegrasikan ke dalam prototipe antarmuka sistem berbasis web sederhana (misalnya menggunakan *framework* Streamlit). Rancangan antarmuka akan mencakup **formulir input** bagi tenaga medis untuk memasukkan 8 parameter klinis pasien (Usia, BMI, HbA1c, dll), serta **panel output** yang berfungsi menampilkan hasil klasifikasi akhir (Positif/Negatif) berdampingan dengan persentase tingkat probabilitas.

Prototipe Sistem Prediksi Diabetes - GPC

Petunjuk Penggunaan:
Masukkan 8 parameter klinis pada formulir di sebelah kiri, lalu klik tombol "Prediksi" untuk melihat hasil klasifikasi.

Input Data Pasien

1. Jenis Kelamin:
2. Usia (tahun):
3. Hipertensi:
4. Penyakit Jantung:
5. Riwayat Merokok:
6. BMI (kg/m²):
7. HbA1c (%):
8. Blood Glucose Level (mg/dL):

Prediksi

Hasil Prediksi

Hasil Klasifikasi:

Belum Ada Hasil Prediksi
Masukkan data pasien pada formulir di sebelah kiri lalu klik tombol "Prediksi" untuk melihat hasil.

Gambar 3.3 Tampilan Awal Web

Gambar 3.3 di atas menampilkan Tampilan Awal Web di mana pengguna dihadapkan pada sebuah formulir kosong di sisi kiri yang berfungsi untuk menginputkan 8 (delapan) parameter klinis pasien sebelum tombol "Prediksi" dijalankan.

Setelah data pasien berhasil dimasukkan dan diproses oleh model, sistem akan langsung menampilkan hasil klasifikasinya. Berikut adalah contoh rancangan tampilan jika hasil prediksi menunjukkan status pasien berisiko.

The screenshot displays a web application titled "Prototipe Sistem Prediksi Diabetes - GPC". It features a "Petunjuk Penggunaan" (Usage Guide) section at the top, followed by an "Input Data Pasien" (Patient Data Input) form on the left and a "Hasil Prediksi" (Prediction Result) section on the right.

Input Data Pasien:

- 1. Jenis Kelamin: Laki-laki
- 2. Usia (tahun): 55
- 3. Hipertensi: Ya
- 4. Penyakit Jantung: Tidak
- 5. Riwayat Merokok: Ya
- 6. BMI (kg/m²): 28.5
- 7. HbA1c (%): 6.9
- 8. Blood Glucose Level (mg/dL): 185

A "Prediksi" button is located below the input fields.

Hasil Prediksi:

Hasil Klasifikasi:

Positif ✓

Probabilitas Positif: 87%

Gambar 3.4 Tampilan dengan Hasil Positif

Gambar 3.4 di atas menunjukkan antarmuka (interface) sistem ketika hasil prediksi ditarik sebagai "Positif" diabetes. Tampilan ini juga dilengkapi dengan persentase probabilitas (confidence) model terhadap kepastian diagnosis.

Sebaliknya, sistem pendukung keputusan juga akan memberikan informasi visual yang informatif apabila pasien diprediksi tidak memiliki risiko yang signifikan terhadap diabetes.

Prototipe Sistem Prediksi Diabetes - GPC

Petunjuk Penggunaan:
Masukkan 8 parameter klinis pada formulir di sebelah kiri, lalu klik tombol "Prediksi" untuk melihat hasil klasifikasi.

Input Data Pasien

1. Jenis Kelamin	Laki-laki
2. Usia (tahun)	55
3. Hipertensi	Ya
4. Penyakit Jantung	Tidak
5. Riwayat Merokok	Ya
6. BMI (kg/m ²)	28.5
7. HbA1c (%)	6.9
8. Blood Glucose Level (mg/dL)	185

Hasil Prediksi

Hasil Klasifikasi:

Negatif ❌

Probabilitas Negatif: 89%

Gambar 3.5 Tampilan dengan Hasil Negatif

Gambar 3.5 di atas memperlihatkan tampilan dashboard dengan hasil "Negatif", yang mengindikasikan bahwa sistem memprediksi pasien tersebut memiliki risiko diabetes yang rendah berdasarkan indikator-indikator yang dimasukkan, ditambahkan dengan angka probabilitas prediksi.

3.9 Proses Penafsiran dan Penyimpulan Hasil

Penafsiran hasil eksperimen tidak hanya berfokus pada seberapa banyak model menebak dengan benar, tetapi juga implikasi medisnya. Analisis dilakukan melalui langkah berikut:

1. **Analisis *Confusion Matrix*:** Menilai penyebaran prediksi model. Fokus utama akan diberikan pada metrik **Recall (Sensitivitas)** untuk meminimalkan nilai *False Negative* (pasien yang sebenarnya sakit diabetes, namun diprediksi sehat oleh model), karena kesalahan ini sangat fatal dalam dunia medis.

2. **Analisis Probabilitas (Ketidakpastian):** Menginterpretasikan keluaran probabilitas dari algoritma GPC. Prediksi dengan probabilitas antara 40%–60% akan ditafsirkan sebagai *area abu-abu (uncertainty)* yang mengindikasikan bahwa pasien memerlukan pemeriksaan laboratorium lebih lanjut.
3. **Penyimpulan:** Menarik kesimpulan apakah GPC layak dan memenuhi standar metrik performa untuk direkomendasikan sebagai inti dari *Clinical Decision Support System (CDSS)*.

BAB IV

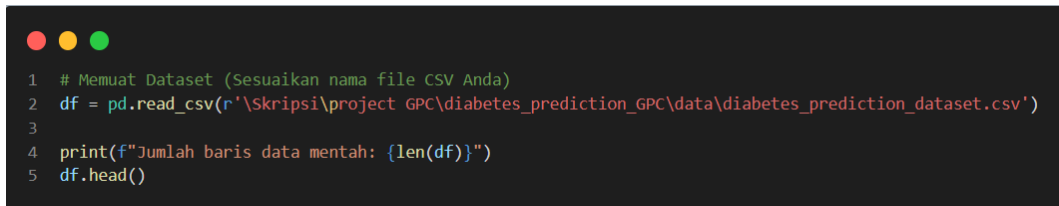
HASIL DAN PEMBAHASAN

4.1 Gambaran Umum

Penelitian ini bertujuan untuk mengklasifikasikan risiko penyakit diabetes berdasarkan parameter klinis pasien menjadi dua kategori, yaitu positif (berisiko) dan negatif (rendah risiko), dengan menggunakan metode *Gaussian Process Classifier* (GPC). Data yang digunakan dalam penelitian ini dikumpulkan dari *Diabetes prediction dataset* sebanyak 1.500 sampel (melalui *Stratified Random Sampling*) yang mencakup 8 fitur: jenis kelamin, usia, riwayat hipertensi, penyakit jantung, riwayat merokok, BMI, kadar HbA1c, dan kadar glukosa darah. Seluruh data melalui tahap pra-pemrosesan menggunakan *Pipeline* yang meliputi imputasi nilai kosong (*SimpleImputer*), *encoding* kategorikal (*OneHotEncoder*), standardisasi skala (*StandardScaler*), serta penyeimbangan kelas minoritas menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE). Data dibagi menjadi data latih sebesar 80% (1.200 data) dan data uji sebesar 20% (300 data). Model dilatih, dioptimasi menggunakan *Cross Validation* untuk mencari *threshold* terbaik (0.3558), lalu dievaluasi kinerjanya menggunakan metrik akurasi, presisi, *recall*, *F1-score*, *ROC-AUC*, dan *Brier Score*. Hasil akhir diimplementasikan ke dalam *dashboard* berbasis web.

4.2 Import Dataset

Pada langkah ini, dataset yang berisi indikator kesehatan pasien dimuat ke dalam *dataframe* Pandas. Dataset ini memuat fitur-fitur klinis yang diperlukan untuk proses penapisan awal.



```
1 # Memuat Dataset (Sesuaikan nama file CSV Anda)
2 df = pd.read_csv(r'\Skripsi\project GPC\diabetes_prediction_GPC\data\diabetes_prediction_dataset.csv')
3
4 print(f"Jumlah baris data mentah: {len(df)}")
5 df.head()
```

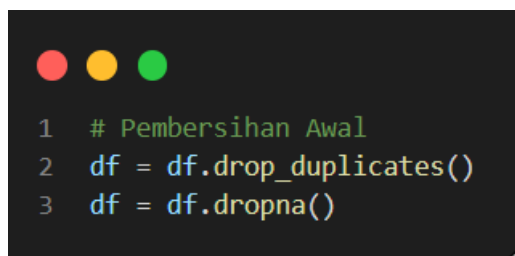
Gambar 4.1 Import Dataset

Gambar 4.1 di atas menampilkan potongan kode (snippet) untuk mengimpor file data berformat CSV ke dalam dataframe, serta fungsi untuk mencetak jumlah total baris data mentah yang diolah dalam penelitian.

Setelah dimuat, lima baris pertama dari *dataframe* ditampilkan (`df.head()`) untuk memverifikasi bahwa data telah berhasil diimpor dan untuk melihat sekilas struktur serta isi datanya. Ini adalah langkah awal yang penting untuk memastikan semua data yang diperlukan tersedia untuk analisis lebih lanjut.

4.3 Pembersihan Data (Data Cleaning)

Setelah dataset berhasil dimuat ke dalam program, tahapan selanjutnya adalah melakukan proses pembersihan data (data cleaning) agar analisis dan model machine learning tidak terdistorsi oleh data yang tidak valid.



```
1 # Pembersihan Awal
2 df = df.drop_duplicates()
3 df = df.dropna()
```

Gambar 4.2 Pembersihan Data

Gambar 4.2 memaparkan barisan kode Python yang digunakan untuk menghapus baris-baris data yang ganda atau duplikat (menggunakan

`drop_duplicates()` dan membersihkan baris data yang bernilai kosong atau *missing values* (menggunakan `dropna()`). Berikut penjelasan:

1. **Penghapusan Duplikat** : Semua baris duplikat dari dataset dihapus (`df = df.drop_duplicates()`) untuk menghindari bias dalam analisis dan pelatihan model.
2. **Penanganan Nilai yang Hilang** : Semua baris yang memiliki nilai hilang (NaN) dihapus (`df = df.dropna()`) untuk memastikan tidak ada data yang tidak lengkap yang memengaruhi model.

4.4 Pra-pemrosesan Data

Pra-pemrosesan data numerik dan kategorikal merupakan serangkaian langkah penting untuk mengubah data klinis mentah menjadi format yang lebih terstruktur dan siap untuk dianalisis oleh model *machine learning*. Untuk mencegah kebocoran data (*data leakage*), proses transformasi fitur dilakukan menggunakan *ColumnTransformer* yang diintegrasikan ke dalam *ImbPipeline*.

1. **Fitur Numerik**: Untuk fitur-fitur medis yang berbentuk numerik (angka riil), diperlukan penanganan dan transformasi khusus agar rentang ukurannya menjadi seragam sebelum dieksekusi ke dalam algoritma Gaussian Process.



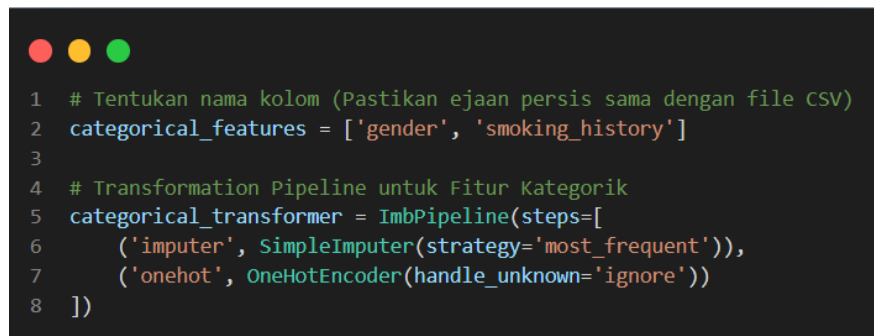
```

1 # Tentukan nama kolom (Pastikan ejaan persis sama dengan file CSV)
2 numeric_features = ['age', 'bmi', 'HbA1c_level', 'blood_glucose_level']
3
4 # Transformation Pipeline untuk Fitur Numerik
5 numeric_transformer = ImbPipeline(steps=[
6     ('imputer', SimpleImputer(strategy='median')),
7     ('scaler', StandardScaler())
8 ])
  
```

Gambar 4.3 Fitur Numerik

Gambar 4.3 menunjukkan proses pembuatan pipeline transformasi khusus untuk data numerik (seperti umur, bmi, HbA1c, glukosa). Proses ini menggunakan teknik pengisian nilai tengah (SimpleImputer berstrategi median) dan penyesuaian skala menggunakan standarisasi Z-score (StandardScaler).

2. **Fitur Kategorikal:** Sama halnya dengan variabel numerik, fitur-fitur yang bersifat kategorikal (contoh: jenis kelamin dan riwayat merokok) juga harus diubah ke dalam bentuk angka yang bisa diproses oleh komputer.



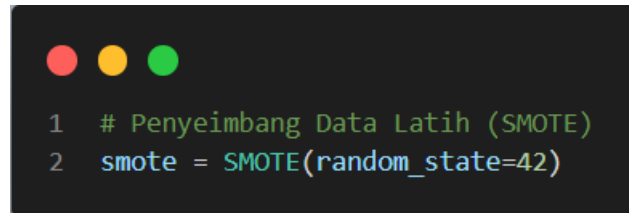
```

1 # Tentukan nama kolom (Pastikan ejaan persis sama dengan file CSV)
2 categorical_features = ['gender', 'smoking_history']
3
4 # Transformation Pipeline untuk Fitur Kategorik
5 categorical_transformer = ImbPipeline(steps=[
6     ('imputer', SimpleImputer(strategy='most_frequent')),
7     ('onehot', OneHotEncoder(handle_unknown='ignore'))
8 ])
  
```

Gambar 4.4 Fitur Kategorikal

Gambar 4.4 di atas menampilkan implementasi pipeline untuk atribut kategorikal. Skema ini mengkombinasikan pengisian kekosongan data dengan nilai yang paling sering muncul (*most_frequent*) serta pengubahan teks menjadi matriks biner menggunakan fungsi *OneHotEncoder*.

3. **Penyeimbangan Data:** Teknik SMOTE diterapkan di dalam *pipeline* khusus pada data latih untuk mengatasi ketidakseimbangan jumlah pasien diabetes dan non-diabetes.



```

1 # Penyeimbang Data Latih (SMOTE)
2 smote = SMOTE(random_state=42)

```

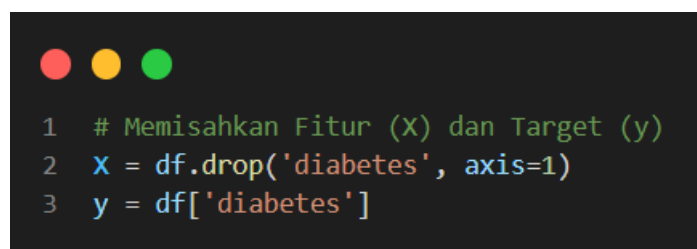
Gambar 4.5 Penyeimbangan Data

Gambar 4.5 memperlihatkan inisialisasi teknik penyeimbangan data latih bernama SMOTE (Synthetic Minority Over-sampling Technique). Kode ini berfungsi menyintesis data virtual pada kelas minoritas sehingga jumlah pasien positif menjadi setara dengan jumlah pasien negatif.

4.5 Modelling

Tahap *Modelling* (Pemodelan) dalam penelitian ini adalah proses inti di mana kita membangun dan melatih sebuah model *machine learning* untuk memprediksi kategori target dari data klinis yang telah diproses. Langkah-langkah utama dalam tahap ini meliputi:

1. **Memisahkan Fitur (X) dan Target (y):** Memisahkan 8 data masukan (fitur klinis) dari variabel target (status diabetes). Ini adalah langkah pertama untuk mempersiapkan data bagi algoritma pembelajaran mesin.



```

1 # Memisahkan Fitur (X) dan Target (y)
2 X = df.drop('diabetes', axis=1)
3 y = df['diabetes']

```

Gambar 4.6 Memisahkan Fitur

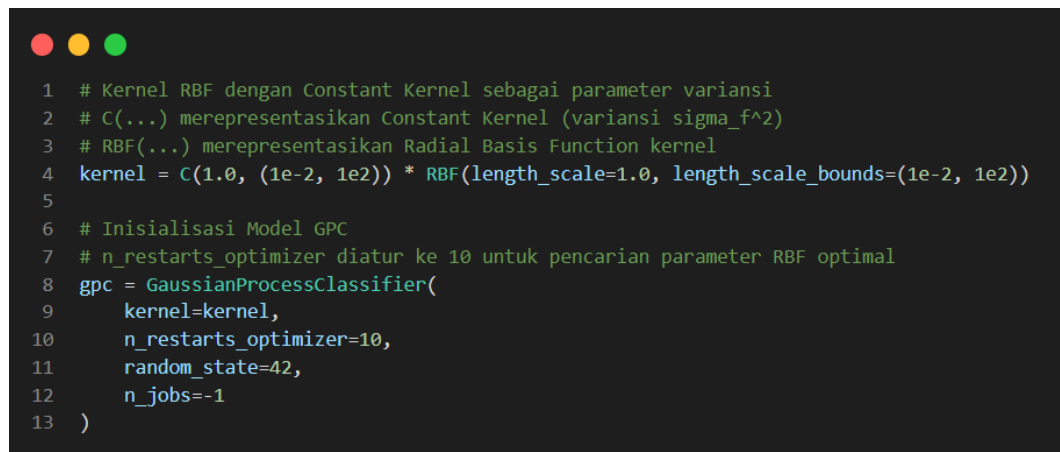
Gambar 4.6 merupakan cuplikan kode yang memisahkan tabel dataset awal menjadi dua variabel inti, yaitu X (berisi 8 parameter prediktor) dan y (berisi label atau kelas target status diabetes).

2. **Pembagian Data:** Membagi dataset menjadi set pelatihan (*training set*) dan set pengujian (*testing set*) dengan rasio 80:20. Set pelatihan digunakan untuk melatih model, sedangkan set pengujian digunakan untuk mengevaluasi kinerja model pada data yang belum pernah dilihat sebelumnya.

4.6 Gaussian Process Classifier (GPC)

4.6.1 Pelatihan Model GPC dan Optimasi Kernel GPC

Setelah melalui tahap pembagian data menjadi data latih dan data uji dengan rasio 80:20, langkah selanjutnya adalah melakukan pelatihan model menggunakan metode *Gaussian Process Classifier*. Model GPC diinisialisasi menggunakan kombinasi fungsi *Constant Kernel* dan *Radial Basis Function* (RBF) dengan batas pencarian parameter (1e-2, 1e2).



```

1 # Kernel RBF dengan Constant Kernel sebagai parameter variansi
2 # C(...) merepresentasikan Constant Kernel (variansi sigma_f^2)
3 # RBF(...) merepresentasikan Radial Basis Function kernel
4 kernel = C(1.0, (1e-2, 1e2)) * RBF(length_scale=1.0, length_scale_bounds=(1e-2, 1e2))
5
6 # Inisialisasi Model GPC
7 # n_restarts_optimizer diatur ke 10 untuk pencarian parameter RBF optimal
8 gpc = GaussianProcessClassifier(
9     kernel=kernel,
10    n_restarts_optimizer=10,
11    random_state=42,
12    n_jobs=-1
13 )

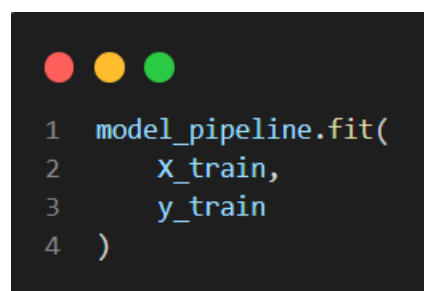
```

Gambar 4.7 Pelatihan Model GPC

Berdasarkan hasil pelatihan, algoritma optimasi L-BFGS-B menemukan titik konvergensi dengan nilai kernel optimal: $10 \times 2 \times \text{RBF}(\text{length_scale}=1.39)$. Parameter *length-scale* mencapai titik stabil pada nilai 1.39, sedangkan parameter variansi menyentuh batas atas maksimal (100 atau 10×2). Menyentuhnya

algoritma pada batas pencarian ini memunculkan `ConvergenceWarning` pada sistem log komputasi, yang membuktikan bahwa variansi data medis sangat kompleks. Pembatasan nilai variansi di angka 100 ini terbukti berhasil bertindak sebagai bentuk regularisasi untuk mencegah *overfitting*.

Setelah instrumen model GPC dan tahapan-tahapan pra-pemrosesan data terkonfigurasi dengan matang, perintah pelatihan akan diberikan agar komputer mulai memetakan pola klasifikasi penyakit. Proses pelatihan dilakukan dengan memanggil fungsi `fit()` pada objek model yang telah diinisialisasi, dengan memasukkan data latih (`X_train`) dan label target (`y_train`) sebagai argumen.

A screenshot of a code editor with a dark background and three colored window control buttons (red, yellow, green) at the top left. The code is written in a light blue font and consists of four lines:

```
1 model_pipeline.fit(  
2     x_train,  
3     y_train  
4 )
```

Gambar 4.8 Melatih Model

Gambar 4.8 memperlihatkan cuplikan baris kode `model_pipeline.fit(X_train, y_train)`. Eksekusi dari perintah ini menandakan dimulainya tahapan pembelajaran aktif dari model GPC terhadap data set latih.

4.6.2 Validasi Silang (Cross-Validation) dan Penentuan Threshold

Sebelum pengujian akhir, dilakukan *5-Fold Cross Validation* pada data latih untuk menguji stabilitas model sekaligus mencari nilai ambang batas (*threshold*) yang optimal. Berdasarkan analisis kurva *Precision-Recall* pada probabilitas

prediksi *Out-of-Fold* (OOF), *threshold* digeser dari standar 0.50 menjadi **0.3558** guna memenuhi target medis yaitu $Recall \geq 80\%$.

```
Menjalankan 5-Fold Cross Validation...
Rata-rata Akurasi CV : 93.75%
Rata-rata Recall CV : 70.65%
Rata-rata F1-Score CV: 66.88%

=== HASIL THRESHOLD TUNING ===
Target Recall      : 0.80
Threshold Terpilih : 0.3558
Precision          : 0.5059
Recall             : 0.8113
F1-Score           : 0.6232
```

Gambar 4.9 Hasil Threshold Tuning Data Latih

Gambar 4.9 memunculkan output dari sistem hasil validasi silang (5-Fold Cross Validation). Log teks ini menegaskan bahwa untuk memenuhi target pencapaian sensitivitas ($Recall \geq 80\%$), nilai ambang (*threshold*) yang terbaik dan terpilih untuk prediksi akhir adalah 0.3558.

4.6.3 Evaluasi Pengujian Model GPC

Model yang telah dilatih kemudian diuji menggunakan 300 data uji (X_{test}). Evaluasi difokuskan pada perbandingan performa antara *threshold* standar (0.5) dan *threshold* hasil optimasi (0.3558).

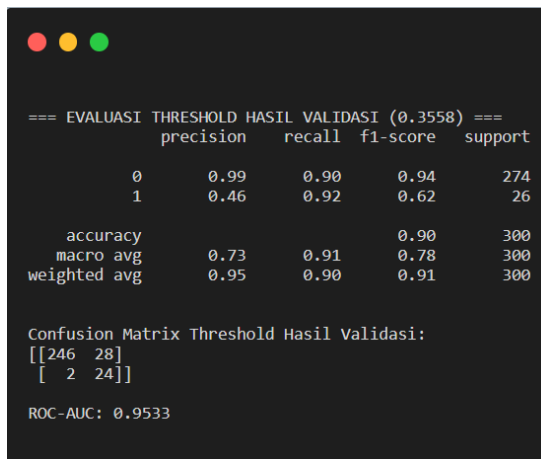
```
=== EVALUASI THRESHOLD STANDAR 0.5 ===
precision  recall  f1-score  support
0          0.98    0.94    0.96     274
1          0.54    0.77    0.63      26

accuracy          0.92    300
macro avg         0.76    0.85    0.80    300
weighted avg      0.94    0.92    0.93    300

Confusion Matrix Threshold 0.5:
[[257  17]
 [  6  20]]
```

Gambar 4.10 Classification Report Threshold 0.5

Gambar 4.10 menyajikan laporan evaluasi (classification report) dengan ambang 0.5. Dari tabel pada gambar tersebut, tingkat penangkapan penderita diabetes (Recall pada kelas 1) dilaporkan masih berada di nilai 77%. Menindaklanjuti evaluasi sebelumnya, performa akan kembali diukur dengan mengganti parameter threshold menjadi nilai optimal 0.3558.



```

=== EVALUASI THRESHOLD HASIL VALIDASI (0.3558) ===
              precision    recall  f1-score   support

      0               0.99       0.90       0.94         274
      1               0.46       0.92       0.62          26

 accuracy               0.90         300
 macro avg              0.73         0.91         0.78         300
 weighted avg           0.95         0.90         0.91         300

Confusion Matrix Threshold Hasil Validasi:
[[246  28]
 [  2  24]]

ROC-AUC: 0.9533

```

Gambar 4.11 Classification Report Threshold 0.3558

Gambar 4.11 merupakan classification report pemungkas setelah implementasi penyesuaian threshold ke angka 0.3558. Penerapan batas ini menyebabkan performa Recall melesat menjadi 92% dan berhasil meminimalisasi deteksi palsu negatif.

1. **Threshold Standar (0.50):** Menghasilkan Akurasi 92%, Presisi 54%, dan *Recall* 77%.
2. **Threshold Optimasi (0.3558):** Menghasilkan Akurasi 90%, Presisi 46%, dan *Recall* meningkat tajam menjadi 92%. Penyesuaian ini sangat krusial dalam domain klinis untuk menekan persentase penderita diabetes yang tidak terdeteksi. Kemampuan diskriminasi model secara keseluruhan sangat luar biasa, dibuktikan dengan nilai **ROC-AUC mencapai 0.9533**.

Hasil ini menunjukkan bahwa model GPC memiliki kinerja yang sangat baik dalam mengklasifikasikan pasien, terutama dalam mendeteksi penderita aktual (*Recall* 92%) untuk menekan angka *False Negative* pada kasus medis.

4.6.4 Confusion Matrix GPC

Confusion Matrix digunakan untuk memvisualisasikan jumlah prediksi benar dan salah untuk setiap kelas. Berikut adalah hasil dari penerapan *threshold* 0.3558:

Tabel 4.1 Confusion Matrix GPC

	Prediksi Positive	Prediksi Negative
Aktual Positive	24	2
Aktual Negative	28	246

Keterangan:

True Positives (TP): 24 pasien positif berhasil dideteksi dengan benar.

True Negatives (TN): 246 pasien negatif diprediksi benar sebagai negatif.

False Positives (FP): 28 pasien negatif salah diprediksi sebagai positif (toleransi skrining)

False Negatives (FN): 2 pasien positif salah diprediksi sebagai negatif (kesalahan fatal berhasil ditekan menjadi minimal).

4.6.5 Perhitungan Evaluasi Model GPC

Setelah *confusion matrix* dari GPC berhasil diperoleh, tahap selanjutnya adalah melakukan perhitungan evaluasi model secara manual menggunakan rumus-rumus yang telah ditentukan. Perhitungan manual ini bertujuan untuk memverifikasi hasil evaluasi yang diperoleh dari kode program serta memberikan pemahaman yang lebih mendalam.

$$1. \text{ Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

$$\begin{aligned}
&= \frac{24 + 246}{24 + 246 + 28 + 2} \times 100\% \\
&= \frac{270}{300} \times 100\% \\
&= 0,9 \times 100\% \\
&= 90\%
\end{aligned}$$

$$2. \text{ Presisi kelas positif} = \frac{TP}{TP + FP} \times 100\%$$

$$\begin{aligned}
&= \frac{24}{24 + 28} \times 100\% \\
&= 0,4615 \times 100\% \\
&= 46,15\%
\end{aligned}$$

$$3. \text{ Recall kelas positif} = \frac{TP}{TP + FN} \times 100\%$$

$$\begin{aligned}
&= \frac{24}{24 + 2} \times 100\% \\
&= 0,9231 \times 100\% \\
&= 92,31\%
\end{aligned}$$

$$4. \text{ F1 – Score kelas positif} = \frac{2 \times \text{Recall} \times \text{Presisi}}{\text{Recall} + \text{Presisi}} \times 100\%$$

$$= \frac{2 \times 0,9231 \times 0,4615}{0,9231 + 0,4615} \times 100\%$$

$$= \frac{2 \times 0,4260}{1,3845} \times 100\%$$

$$= \frac{0,8520}{1,3845} \times 100\%$$

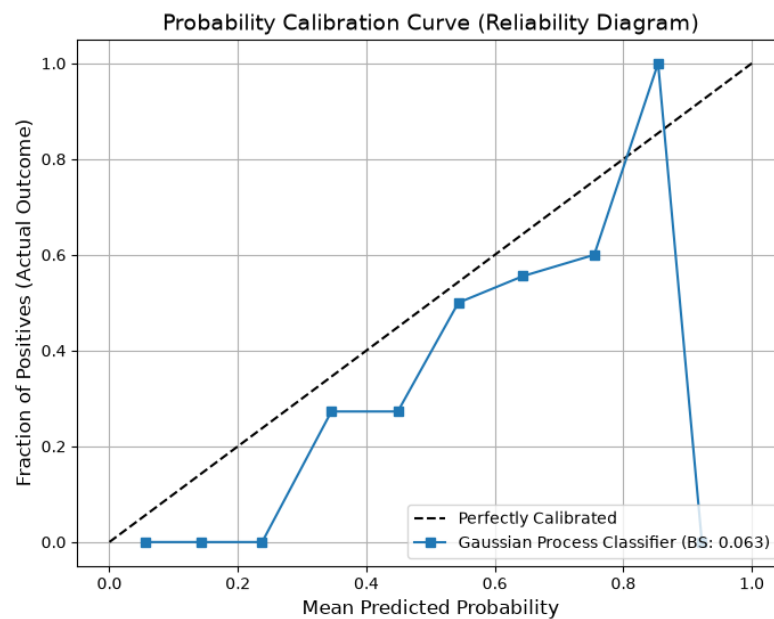
$$= 0,6154 \times 100\%$$

$$= 61,54\% \text{ atau } 62\%$$

Berdasarkan hasil evaluasi yang telah dilakukan, model ini menunjukkan performa yang sangat luar biasa dalam mengklasifikasikan kelas minoritas (positif diabetes) dengan *Recall* 92%, sehingga sangat layak digunakan sebagai sarana penapisan medis awal.

4.6.6 Analisis Keandalan Probabilitas

Selain prediksi kelas, model GPC dievaluasi kapabilitasnya dalam memberikan probabilitas terkalibrasi (*uncertainty estimation*) untuk membuktikan bahwa nilai persen probabilitas keluaran model GPC tidak menipu dan selaras dengan kejadian sesungguhnya (terkalibrasi), model diuji keandalannya menggunakan visualisasi kurva reliabilitas.



Gambar 4.12 Kurva Kalibrasi (Reliability Diagram) GPC

Hasil pengujian mencatatkan nilai **Brier Score sebesar 0.0633**. Nilai yang mendekati nol ini, didukung oleh garis kurva kalibrasi yang bergerak sejajar dengan garis referensi (*perfectly calibrated*), membuktikan bahwa persentase probabilitas

yang dikeluarkan oleh model dapat dipercaya secara medis dan tidak bersifat sembarang (*overconfident*).

4.7 Implementasi Antarmuka

Implementasi antarmuka terdiri dari halaman *dashboard* interaktif yang menampilkan *form* indikator klinis, visualisasi probabilitas risiko, dan riwayat pasien. Untuk mendemonstrasikan kapabilitas model secara praktis, model GPC dan metrik *threshold* (0.3558) yang tersimpan dalam format *artifact* diintegrasikan ke dalam prototipe *Clinical Decision Support System* (CDSS) berbasis web menggunakan *framework* Streamlit.

4.7.1 Halaman Dashboard

Halaman *Dashboard* merupakan halaman utama yang pertama kali muncul saat pengguna membuka *website*. Pada halaman ini, antarmuka yang sederhana dan informatif dimulai dengan judul utama "**Diabetes Risk Dashboard & Predictive Simulator**" dan deskripsi singkat yang menjelaskan tujuan dari *website* ini. Di bawahnya, terdapat *form* input untuk memasukkan 8 preferensi parameter klinis pasien.

Diabetes Risk Dashboard & Predictive Simulator
Platform analisis klinis interaktif berbasis ML dengan grafik visual parameter medis & tabel riwayat pasien.

Form Input Pasien

Jenis Kelamin Female	Riwayat Merokok never
Usia (Tahun) 80.00	BMI (Indeks Massa) 25.19
Hipertensi Tidak	Kadar HbA1c (%) 6.6
Penyakit Jantung Ya	Glukosa Darah (mg/dL) 140

Jalankan Analisis Prediksi

Live Hasil Analisis & Grafik

Isi Form Indikator Medis Pasien di sebelah kiri lalu klik **Jalankan Analisis Prediksi** untuk menampilkan live dashboard, grafik radar, & bar chart.

Tabel Riwayat Pasien & Analytics Log

Belum ada data riwayat pasien. Gunakan Form di atas atau klik 'Muat 3 Sample Data Pasien' di sidebar untuk mencoba.

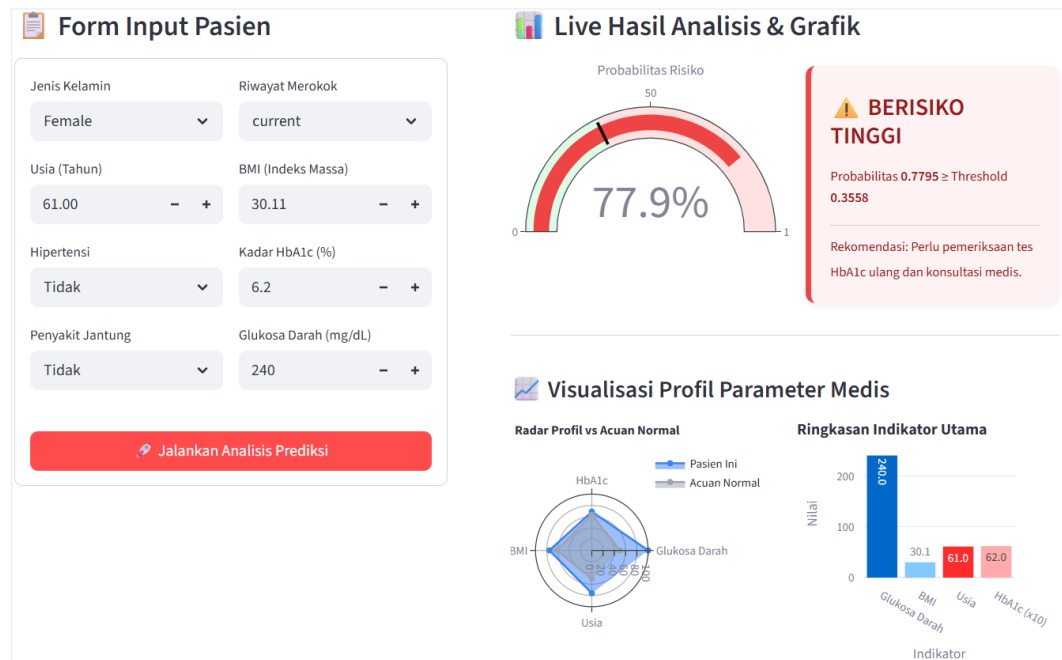
Mahasiswa UMSU Fakultas Ilmu Komputer dan Teknologi Informasi Stambuk 2022
Developed with Streamlit, Plotly & Scikit-Learn

Gambar 4.13 Halaman Dashboard (Input Pasien)

Gambar 4.13 mengilustrasikan halaman utama Dashboard web "Diabetes Risk Dashboard & Predictive Simulator" buatan penulis yang dibentuk dengan platform Streamlit. Tenaga medis dapat mengisi langsung kolom fitur untuk memulai simulasi diagnosis.

4.7.2 Hasil Analisis dan Visualisasi

Setelah memilih preferensi parameter medis pada halaman *dashboard* dan menekan tombol "Jalankan Analisis Prediksi", sistem akan memproses data yang telah dimasukkan dan menampilkan hasil rekomendasi diagnosis pada panel sebelah kanan. Setiap prediksi berisi informasi lengkap disertai dengan tingkat probabilitas (*confidence*) model dalam bentuk visualisasi *Gauge Chart* beserta perbandingan *Radar Chart* terhadap acuan normal.



Gambar 4.14 Hasil Klasifikasi dan Probabilitas Klinis

Gambar 4.14 merupakan ilustrasi panel keluaran hasil. Panel ini memiliki elemen grafis Gauge Chart untuk memberikan visual persentase indikasi risiko dan grafik Radar (Radar Chart) untuk melakukan perbandingan secara holistik terhadap nilai normal setiap pengujian laboratoriumnya.

4.7.3 Tabel Riwayat Log Pasien

Terakhir, aplikasi simulator prediktif ini dilengkapi dengan fungsionalitas perekaman histori sesi di bagian bawah *dashboard*, sistem menyediakan manajemen sesi interaktif (*Session State*) yang merekam riwayat seluruh penapisan yang telah dilakukan, sehingga pengguna dapat melacak hasil penapisan sebelumnya.

Tabel Riwayat Pasien & Analytics Log

Total Pasien Diperiksa

Positif (Berisiko)

Negatif (Rendah)

Rata-rata Probabilitas

4

2

2

0.4121

Data Tabel Log

Grafik Distribusi & Sebaran Sesi Ini

	gender	Usia	hypertension	heart_disease	smoking_history	BMI	HbA1c	Glukosa Darah	Probabilitas Risk	Hasil Diagnosa
0	Female	80	0	1	never	25.19	6.6	140	0.8068	Positif (Berisiko)
1	Female	54	0	0	No Info	27.32	6.6	80	0.0390	Negatif (Rendah)
2	Male	28	0	0	never	27.32	5.7	158	0.0232	Negatif (Rendah)
3	Female	61	0	0	current	30.11	6.2	240	0.7795	Positif (Berisiko)

📄 Unduh Seluruh Data Riwayat (.CSV)

Gambar 4.15 Tabel Riwayat Pasien dan *Analytics Log*

Gambar 4.15 di atas memperlihatkan blok Tabel Riwayat Log Pasien pada bagian dasar dashboard. Tabel tersebut merangkum rincian input form, hasil prediksi, rekap probabilitas keseluruhan pasien, serta memfasilitasi pengguna untuk mengekspor (unduh) log tersebut ke dalam berkas CSV guna keperluan dokumentasi klinis lebih lanjut.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian, perancangan model, hingga implementasi sistem mengenai prediksi penyakit diabetes menggunakan metode *Gaussian Process Classifier* (GPC) pada data klinis, dapat ditarik beberapa kesimpulan sebagai berikut:

1. Model *machine learning* berbasis *Gaussian Process Classifier* (GPC) berhasil diimplementasikan untuk memprediksi risiko diabetes berdasarkan 8 karakteristik data klinis. Keberhasilan ini dicapai melalui proses perakitan *pipeline* yang terstruktur, meliputi teknik imputasi data, *encoding* kategorikal, penyesuaian skala fitur menggunakan *Z-score* (*StandardScaler*), serta penyeimbangan kelas data (pasien positif dan negatif) menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE). Penggunaan *Radial Basis Function* (RBF) Kernel dengan pembatasan rentang pencarian $[1e-2, 1e2]$ terbukti bertindak sebagai regularisasi yang sangat efektif untuk mencegah *overfitting* pada data medis yang kompleks.
2. Berdasarkan hasil pengujian dan evaluasi model, metode GPC terbukti sangat andal digunakan sebagai instrumen skrining medis awal. Melalui optimasi ambang batas (*threshold* tuning) pada nilai 0.3558, model berhasil memprioritaskan keselamatan pasien dengan mencapai tingkat *Recall* (Sensitivitas) yang sangat tinggi yaitu **92.30%**, sehingga mampu menekan angka *False Negative* secara maksimal. Evaluasi keseluruhan juga mencatatkan Akurasi sebesar **90%**, Presisi 46.15%, dan *F1-Score* 61.54%.

3. Algoritma GPC tidak hanya menghasilkan vonis kelas yang kaku, tetapi terbukti mampu memberikan estimasi ketidakpastian (*uncertainty estimation*) yang sangat baik. Hal ini dibuktikan dengan nilai *ROC-AUC* sebesar **0.9533** dan *Brier Score Loss* yang sangat rendah yaitu **0.0633**. Probabilitas prediksi yang terkalibrasi ini kemudian berhasil diimplementasikan ke dalam antarmuka *Clinical Decision Support System* (CDSS) berbasis web interaktif menggunakan Streamlit, lengkap dengan visualisasi probabilitas dan pencatatan riwayat diagnostik (*analytics log*).

5.2 Saran

Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa saran yang dapat dipertimbangkan untuk pengembangan penelitian selanjutnya, yaitu:

1. Penelitian selanjutnya dapat melakukan studi komparatif dengan menguji model probabilistik lainnya (seperti *Bayesian Neural Networks* atau *Naive Bayes*) serta algoritma *Ensemble Learning* (seperti XGBoost atau Random Forest) untuk membandingkan tingkat presisi, waktu komputasi, dan keandalan probabilitas kalibrasi terhadap GPC.
2. Penambahan dimensi dataset klinis dari sumber lokal (seperti rekam medis rumah sakit di Indonesia) dengan fitur tambahan seperti riwayat keturunan (genetik), profil lipid (kolesterol), atau tekanan darah sistolik/diastolik secara rinci, guna meningkatkan kemampuan generalisasi model terhadap demografi pasien lokal.
3. Penerapan teknik optimasi parameter tingkat lanjut seperti *Bayesian Optimization* atau *GridSearchCV* dengan rentang pencarian batas atas yang

lebih lebar namun disertai dengan fungsi penalti (*penalty function*), untuk menangani ConvergenceWarning tanpa memicu masalah overfitting.

4. Pengembangan arsitektur sistem web (*dashboard*) lebih lanjut dengan mengintegrasikan sistem basis data relasional (seperti MySQL atau PostgreSQL) agar riwayat penapisan pasien dapat tersimpan secara permanen, serta menambahkan sistem autentikasi (Login/Register) agar privasi data medis pasien tetap terjaga dan hanya dapat diakses oleh staf medis yang berwenang.

DAFTAR PUSTAKA

- ADA. (2022). 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2022. *Diabetes Care*, 45, S17–S38. <https://doi.org/10.2337/dc22-S002>
- Ali Küsmüş. (2025). Diabetes Prediction Based on Health Indicators Using Machine Learning: Feature Selection and Algorithm Comparison. *International Journal of Advanced Engineering and Management Research*, 10(02), 281–292. <https://doi.org/10.51505/ijaemr.2025.1115>
- Butt, U. M., Letchmunan, S., Ali, M., Hassan, F. H., Baqir, A., & Sherazi, H. H. R. (2021). Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications. *Journal of Healthcare Engineering*, 2021. <https://doi.org/10.1155/2021/9930985>
- El Massari, H., Sabouri, Z., Mhammedi, S., & Gherabi, N. (2022). Diabetes Prediction Using Machine Learning Algorithms and Ontology. *Journal of ICT Standardization*, 10(2), 319–338. <https://doi.org/10.13052/jicts2245-800X.10212>
- El-Sofany, H., El-Seoud, S. A., Karam, O. H., Abd El-Latif, Y. M., & Taj-Eddin, I. A. T. F. (2024). A Proposed Technique Using Machine Learning for the Prediction of Diabetes Disease through a Mobile App. *International Journal of Intelligent Systems*, 2024. <https://doi.org/10.1155/2024/6688934>

- Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, 8, 76516–76531. <https://doi.org/10.1109/ACCESS.2020.2989857>
- Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., & Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific Reports*, 10(1). <https://doi.org/10.1038/s41598-020-68771-z>
- Mahmud, S., Islam, B. U., Haque Anik, N., & Ghosh, T. (2024). Diabetes Prediction: A Comparative Analysis of Machine Learning Algorithms with SMOTE. *2024 IEEE Conference on Computing Applications and Systems, COMPAS 2024*. <https://doi.org/10.1109/COMPAS60761.2024.10796405>
- Mustafa, M. (2023). *Diabetes prediction dataset*. <https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset>
- Nooraldaim, A. S., Ahmed Abdalla, A. E., & Seed, A. M. (2024). An Advanced Machine Learning Approach for Enhanced Diabetes Prediction. *International Journal of Current Science Research and Review*, 07(12). <https://doi.org/10.47191/ijcsrr/V7-i12-15>
- Nora Marlim, Y., Suryati, L., & Agustina, N. (2022). Deteksi Dini Penyakit Diabetes Menggunakan Machine Learning dengan Algoritma Logistic Regression. In *Jurnal Nasional Teknik Elektro dan Teknologi Informasi* | (Vol. 11, Issue 2).
- Rasmussen, C. Edward., & Williams, C. K. I. . (2006). *Gaussian processes for machine learning*. MIT Press.

- Sahid, M. A., Ul Hoque Babar, M., & Uddin, M. P. (2024). Predictive modeling of multi-class diabetes mellitus using machine learning and filtering iraqi diabetes data dynamics. *PLoS ONE*, 19(5 May). <https://doi.org/10.1371/journal.pone.0300785>
- Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018). Prediction of Diabetes Using Machine Learning Algorithms in Healthcare. *2018 24th International Conference on Automation and Computing (ICAC)*, 1–6. <https://doi.org/10.23919/ICoNAC.2018.8748992>
- Shashikant, R., Chaskar, U., Phadke, L., & Patil, C. (2021). Gaussian process-based kernel as a diagnostic model for prediction of type 2 diabetes mellitus risk using non-linear heart rate variability features. *Biomedical Engineering Letters*, 11(3), 273–286. <https://doi.org/10.1007/s13534-021-00196-7>
- World Health Organization. (2024). *Diabetes*. <https://www.who.int/news-room/fact-sheets/detail/diabetes>
- Yakut, Ö. (2023). Diabetes Prediction Using Colab Notebook Based Machine Learning Methods. *International Journal of Computational and Experimental Science and Engineering*, 9(1), 36–41. <https://doi.org/10.22399/ijcesen.1185474>
- Yothapakdee, K., Charoenkhum, S., & Boonnuk, T. (2022). Improving the efficiency of machine learning models for predicting blood glucose levels and diabetes risk. *Indonesian Journal of Electrical Engineering and Computer Science*, 27(1), 555–562. <https://doi.org/10.11591/ijeecs.v27.i1.pp555-562>

Yuniarsih, E. P., Estri, M. N., Istikaanah, N., & Sihwaningrum, I. (2025). *Application of the Random Forest Method for Classifying Diabetes* (pp. 376–387). https://doi.org/10.2991/978-2-38476-525-6_38

Zhang, N., Liu, Y., & Zhang, Z. (2025). Hybrid Machine Learning Framework for Type 2 Diabetes Prediction Using Metaheuristic Optimization Algorithms. *Informatica*, 49(12). <https://doi.org/10.31449/inf.v49i12.9298>