



# Klasifikasi Risiko Diabetes Menggunakan *Support Vector Machine* Berbasis *Feature Selection* dan *Hyperparameter Tuning*

Nova Ariansyah<sup>1\*</sup>, Dedy Hermanto<sup>2</sup>

<sup>1,2</sup>Universitas Multi Data Palembang, Palembang, Indonesia

## Informasi Artikel

*Sejarah Artikel:*  
 Submit: 18 Februari 2026  
 Revisi: 28 Februari 2026  
 Diterima: 04 Maret 2026  
 Diterbitkan: 24 Maret 2026

## Kata Kunci

Diabetes Mellitus, *Support Vector Machine*, *RelieFF*, *Mutual Information*, *Hyperparameter Tuning*

## Korespondensi

E-mail: [nariansyah23@mhs.mdp.ac.id](mailto:nariansyah23@mhs.mdp.ac.id)\*

## A B S T R A C T

*Diabetes Mellitus is a chronic metabolic disease with increasing global prevalence and significant clinical complications. Early risk detection is essential to reduce long-term cardiovascular and renal complications. This study aims to implement Support Vector Machine (SVM) for diabetes risk classification and to analyze the impact of Feature Selection and Hyperparameter Tuning on model performance. The dataset consists of 100,000 patient records obtained from Kaggle, with an imbalanced class distribution between diabetic and non-diabetic patients. Data Preprocessing was performed using Min-Max Normalization. Feature Selection was conducted using RelieFF and Mutual Information to reduce redundant attributes. Model evaluation employed 10-Fold Cross Validation to ensure reliable generalization performance. Hyperparameter Tuning was performed using Optuna by optimizing parameters C, gamma, and degree on Polynomial and Radial Basis Function (RBF) kernels. Experimental results show that Hyperparameter Tuning significantly improves the recall of the Diabetes class from 0.67 to 0.94 and increases overall accuracy from 94.76% to 98.90%. False negative cases decreased from 4,851 to 915 instances, indicating substantial improvement in sensitivity. These findings demonstrate that optimized SVM models enhance minority class detection in imbalanced medical datasets, with a reasonable computational trade-off.*

## Abstrak

Diabetes Mellitus merupakan penyakit metabolik kronis dengan prevalensi yang terus meningkat secara global. Deteksi dini risiko diabetes menjadi langkah strategis untuk mencegah komplikasi serius seperti penyakit kardiovaskular dan gagal ginjal. Penelitian ini bertujuan mengimplementasikan algoritma *Support Vector Machine* (SVM) dalam klasifikasi risiko diabetes serta menganalisis pengaruh *Feature Selection* dan *Hyperparameter Tuning* terhadap kinerja model. Dataset yang digunakan berjumlah 100.000 data pasien yang diperoleh dari Kaggle dengan distribusi kelas yang tidak seimbang antara pasien diabetes dan non-diabetes. *Preprocessing* dilakukan menggunakan *Min-Max Normalization*. Seleksi fitur diterapkan menggunakan metode *RelieFF* dan *Mutual Information*. Evaluasi model menggunakan *10-Fold Cross Validation* untuk meminimalkan bias estimasi performa. *Hyperparameter Tuning* dilakukan menggunakan *Optuna* dengan optimasi parameter *C*, *gamma*, dan *degree* pada kernel *Polynomial* dan *Radial Basis Function*. Hasil eksperimen menunjukkan bahwa *Hyperparameter Tuning* meningkatkan *recall* kelas Diabetes dari 0,67 menjadi 0,94 serta meningkatkan akurasi keseluruhan dari 94,76% menjadi 98,90%. Jumlah *false negative* menurun dari 4.851 menjadi 915 kasus. Temuan ini menunjukkan bahwa optimasi parameter dan seleksi fitur berperan penting dalam meningkatkan sensitivitas model terhadap kelas minoritas pada dataset medis yang tidak seimbang.



## 1. Pendahuluan

Diabetes Mellitus merupakan penyakit metabolik kronis yang ditandai dengan peningkatan kadar glukosa darah akibat gangguan sekresi insulin maupun resistensi insulin. Penyakit ini menjadi salah satu penyebab utama komplikasi kardiovaskular dan gangguan ginjal di berbagai negara [1]. Peningkatan prevalensi diabetes mendorong perlunya sistem deteksi dini berbasis teknologi untuk membantu proses identifikasi risiko secara lebih cepat dan efisien [2]. Pendekatan *machine learning* telah banyak digunakan dalam sistem pendukung keputusan klinis, termasuk untuk klasifikasi penyakit kronis [2], [14].

Salah satu algoritma yang memiliki performa stabil dalam berbagai studi klasifikasi medis adalah *Support Vector Machine (SVM)* [3]. SVM bekerja dengan membangun *hyperplane* optimal untuk memisahkan dua kelas dengan *margin* maksimum. Pada data non-linear, penggunaan kernel *Polynomial* dan *Radial Basis Function (RBF)* terbukti mampu meningkatkan performa klasifikasi [4]. Dalam konteks klasifikasi diabetes, Berdasarkan berbagai penelitian terdahulu, algoritma *Support Vector Machine (SVM)* dan *K-Nearest Neighbor (KNN)* telah terbukti memiliki kinerja yang baik dalam mendeteksi dan mengklasifikasikan penyakit diabetes. Penelitian oleh [23] menunjukkan bahwa metode SVM mampu melakukan klasifikasi diabetes dengan tingkat akurasi yang cukup baik pada data Puskesmas Modopuro Mojokerto.

Sementara itu, penelitian membuktikan bahwa algoritma KNN dan SVM, setelah diterapkan teknik *SMOTE* untuk menangani ketidakseimbangan data, mampu mencapai akurasi di atas 80%, sehingga efektif dalam mendeteksi risiko diabetes. Hasil-hasil ini menunjukkan bahwa pendekatan *machine learning* berpotensi besar dalam mendukung deteksi dini penyakit diabetes, namun sebagian besar penelitian sebelumnya masih menggunakan dataset publik dan belum mempertimbangkan faktor gaya hidup secara komprehensif [20]. Selain itu, optimasi model melalui *Hyperparameter Tuning* menjadi faktor penting dalam meningkatkan akurasi dan sensitivitas sistem [6], [8]. Meskipun demikian, sebagian penelitian terdahulu masih menggunakan dataset berskala kecil atau belum mengintegrasikan *Feature Selection* dan *Hyperparameter Tuning* secara sistematis dalam satu kerangka evaluasi terkontrol [9], [15].

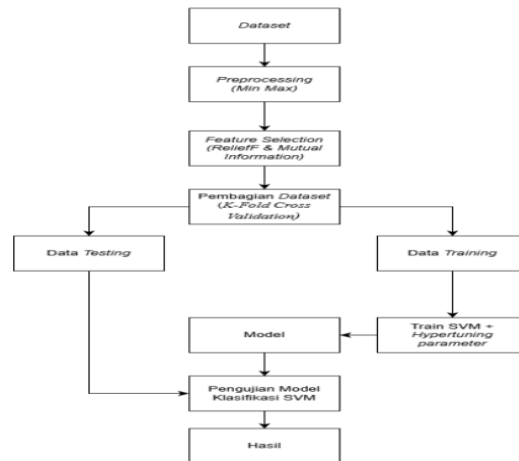
Selain itu, belum banyak studi yang memanfaatkan optimasi berbasis *Bayesian* seperti *Optuna* pada dataset diabetes berskala besar. Oleh karena itu, penelitian ini bertujuan mengintegrasikan seleksi fitur *ReliefF* dan *Mutual Information* dengan optimasi *hyperparameter* berbasis *Optuna* pada algoritma SVM untuk meningkatkan performa klasifikasi risiko diabetes. Oleh karena itu, penelitian ini akan mengembangkan sistem klasifikasi risiko diabetes berbasis web dengan menggunakan algoritma *Support Vector Machine (SVM)* yang dilatih berdasarkan data gaya hidup dan riwayat medis pengguna. Sistem ini akan memfasilitasi pengisian data gaya hidup pengguna, melakukan klasifikasi tingkat risiko diabetes melalui model SVM yang telah dibangun, serta menampilkan hasil klasifikasi berupa label risiko dan rekomendasi gaya hidup sehat sederhana. Pendekatan ini diharapkan dapat menjadi solusi dalam meningkatkan deteksi dini risiko diabetes di kalangan masyarakat muda serta mendukung upaya preventif melalui edukasi pola hidup sehat..

## 2. Metode Penelitian

### 2.1. Tahapan Penelitian

Tahapan penelitian dimulai dari studi literatur, pengumpulan dataset, *Preprocessing* data, penerapan *feature selection*, pelatihan model SVM, *Hyperparameter Tuning*, serta evaluasi model

menggunakan *cross validation*. Proses ini dirancang untuk memastikan hasil klasifikasi yang optimal dan dapat digeneralisasi dengan baik



Gambar 1. Skema Perancangan SVM

## 2.2. Dataset dan Preprocessing

Dataset terdiri dari 100.000 data pasien yang diperoleh dari *Kaggle* dan mencakup atribut usia, jenis kelamin, hipertensi, penyakit jantung, riwayat merokok, BMI, HbA1c level, dan kadar glukosa darah. Label target berupa status diabetes. *Preprocessing* dilakukan menggunakan *Min-Max Normalization* untuk menyetarakan skala fitur numerik agar stabil dalam proses pelatihan SVM [3].

$$x' = (x - x_{min}) / (x_{max} - x_{min}) \quad (1)$$

## 2.3. Support Vector Machine

*Support Vector Machine (SVM)* merupakan salah satu algoritma *supervised learning* yang digunakan untuk klasifikasi dan regresi. SVM bekerja dengan mencari sebuah hyperplane optimal yang dapat memisahkan data dari dua kelas dengan *margin* maksimum. Jika diberikan dataset pelatihan  $(x_i, y_i)$ , di mana  $x_i \in \mathbb{R}^n$  merupakan vektor fitur dan  $y_i \in \{-1, +1\}$  adalah label kelas, maka fungsi keputusan yang digunakan SVM didefinisikan pada Fungsi keputusan SVM dirumuskan sebagai:

$$f(x) = w^T x + b \quad (2)$$

Untuk kasus *soft margin*, fungsi objektif dirumuskan sebagai:

$$\min (1/2) \|w\|^2 + C \sum \xi_i \quad (3)$$

Parameter  $C$  mengontrol keseimbangan antara *margin* maksimum dan kesalahan klasifikasi. Kernel yang digunakan dalam penelitian ini adalah *Polynomial* dan *Radial Basis Function*.

## 2.4 Feature Selection dan Hyperparameter Tuning

*Feature Selection* dilakukan menggunakan metode *ReliefF* dan *Mutual Information* untuk mengidentifikasi atribut yang paling relevan terhadap variabel target. *ReliefF* menggunakan parameter jumlah tetangga terdekat ( $K$ ) sebesar 10 untuk menjaga keseimbangan antara sensitivitas terhadap variasi lokal dan stabilitas hasil seleksi fitur.

*Hyperparameter Tuning* dilakukan menggunakan *Optuna*, yaitu *framework* optimasi berbasis *Bayesian* yang efisien dalam mengeksplorasi ruang parameter secara adaptif. Parameter yang dioptimalkan meliputi:

1.  $C$  (*regularization parameter*): 0.01 – 100 (*log-uniform distribution*)
2.  $\gamma$  (*kernel coefficient* pada RBF): 1e-4 – 10 (*log-uniform distribution*)
3.  $degree$  (*kernel Polynomial*): 2 – 5 (*integer uniform distribution*)

Pemilihan rentang pencarian (*search space*) tersebut didasarkan pada praktik umum optimasi SVM dalam literatur klasifikasi medis serta untuk memastikan eksplorasi parameter yang cukup luas tanpa menyebabkan eksplorasi berlebihan yang tidak efisien.

Proses optimasi dilakukan dengan memaksimalkan *F1-Score* kelas Diabetes sebagai *objective function*, karena metrik ini lebih representatif dalam kondisi dataset yang tidak seimbang dibandingkan akurasi semata.

Untuk menghindari potensi data *leakage*, proses *Feature Selection* dan *Hyperparameter Tuning* dilakukan di dalam setiap fold pada skema *10-Fold Cross Validation*. Dengan demikian, pada setiap iterasi:

1. Dataset dibagi menjadi data pelatihan dan data pengujian sesuai *fold*.
2. *Feature Selection* hanya dilakukan pada data pelatihan.
3. *Hyperparameter Tuning* dilakukan menggunakan data pelatihan pada *fold* tersebut.
4. Model terbaik kemudian diuji pada data pengujian *fold* terkait.

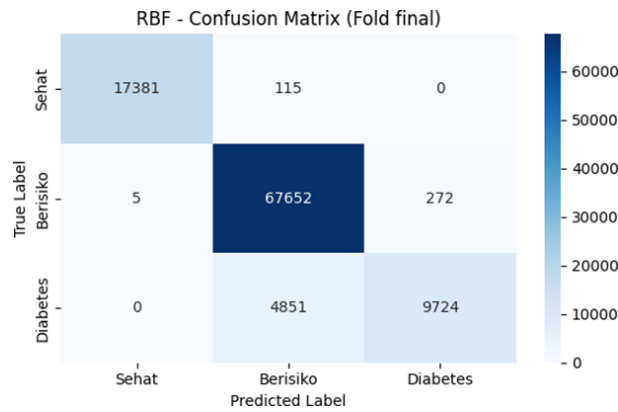
Pendekatan ini memastikan bahwa data pengujian tidak mempengaruhi proses seleksi fitur maupun optimasi parameter, sehingga estimasi performa model mencerminkan kemampuan generalisasi yang lebih realistis dan tidak mengalami overestimasi.

### 3. Hasil dan Pembahasan

Proses normalisasi data dan *feature selection*, dilakukan secara terpisah pada setiap *fold* dalam skema *10-Fold Cross Validation*. Pada setiap iterasi, data dibagi menjadi data pelatihan dan data pengujian. Proses *Min-Max Normalization* dihitung berdasarkan parameter minimum dan maksimum dari data pelatihan pada *fold* tersebut, kemudian diterapkan pada data pengujian tanpa menggunakan informasi *global* dari keseluruhan dataset dan proses seleksi fitur menggunakan *ReliefF* dan *Mutual Information* hanya dilakukan pada data pelatihan di masing-masing *fold*. Fitur terpilih dari data pelatihan kemudian digunakan untuk melatih model dan mengevaluasi performa pada data pengujian *fold* terkait. Pendekatan ini memastikan bahwa tidak terjadi kebocoran informasi (*data leakage*) antara data pelatihan dan data pengujian, sehingga estimasi performa model mencerminkan kemampuan generalisasi yang lebih realistis.

Berdasarkan *confusion matrix RBF* pada *fold final*, model menunjukkan performa klasifikasi yang sangat baik pada kelas mayoritas, khususnya kelas Berisiko, dengan jumlah prediksi benar sebanyak 67.652 data. Pada kelas Sehat, model berhasil mengklasifikasikan dengan benar sebanyak 17.381 data, dengan kesalahan klasifikasi sebanyak 115 data yang diprediksi sebagai Berisiko dan tidak ada yang salah diklasifikasikan sebagai Diabetes. Hal ini menunjukkan bahwa model memiliki kemampuan separasi yang baik antara kelas Sehat dan Diabetes, namun demikian, pada kelas Diabetes, masih terdapat kesalahan klasifikasi yang cukup signifikan. Dari total data Diabetes, sebanyak 9.724 berhasil diprediksi dengan benar sebagai Diabetes, tetapi sebanyak 4.851 data salah diklasifikasikan sebagai Berisiko. Meskipun tidak terdapat kasus Diabetes yang salah diklasifikasikan sebagai Sehat, jumlah kesalahan ke kelas Berisiko menunjukkan bahwa model masih mengalami kesulitan dalam membedakan batas keputusan antara kelas Berisiko dan Diabetes.

Secara keseluruhan, *confusion matrix* ini menunjukkan bahwa model memiliki tingkat akurasi yang tinggi pada kelas mayoritas, tetapi sensitivitas terhadap kelas Diabetes masih belum optimal pada konfigurasi ini. Kesalahan terbesar berasal dari pergeseran prediksi antara kelas Berisiko dan Diabetes, yang mengindikasikan adanya kemiripan pola fitur antara kedua kelas tersebut. Oleh karena itu, diperlukan optimasi lebih lanjut, seperti *Hyperparameter Tuning* atau penyesuaian *threshold* keputusan, untuk meningkatkan kemampuan deteksi pada kelas Diabetes dan menurunkan jumlah *false negative*.

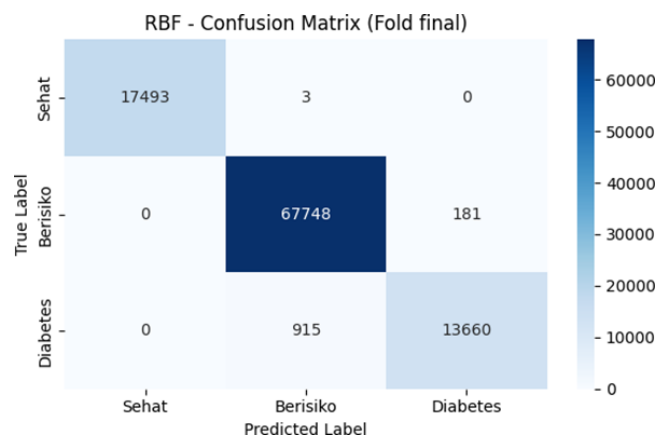


**Gambar 2.** Confusion Matrix SVM tanpa Tuning

Setelah dilakukan *Hyperparameter Tuning*, terlihat peningkatan performa klasifikasi yang signifikan pada seluruh kelas. Pada kelas Sehat, model berhasil mengklasifikasikan dengan benar sebanyak 17.493 data, dengan hanya 3 data yang salah diprediksi sebagai Berisiko dan tidak terdapat kesalahan prediksi ke kelas Diabetes. Hal ini menunjukkan kemampuan model yang sangat baik dalam membedakan individu sehat dari kategori risiko lainnya. Pada kelas Berisiko, model menunjukkan performa yang sangat tinggi dengan 67.748 data berhasil diklasifikasikan secara tepat, tanpa kesalahan ke kelas Sehat dan hanya 181 data yang salah diprediksi sebagai Diabetes. Angka ini menunjukkan peningkatan presisi dan stabilitas model pada kelas mayoritas.

Perbaikan paling signifikan terlihat pada kelas Diabetes. Dari total data Diabetes, sebanyak 13.660 berhasil diprediksi dengan benar sebagai Diabetes, sedangkan jumlah kesalahan klasifikasi ke kelas Berisiko menurun drastis menjadi 915 kasus, dan tidak terdapat kesalahan prediksi ke kelas Sehat. Penurunan jumlah *false negative* dari ribuan kasus pada model sebelum *tuning* menjadi hanya 915 kasus menunjukkan peningkatan sensitivitas model yang sangat substansial dalam mendeteksi pasien Diabetes.

Secara keseluruhan, *confusion matrix* ini menunjukkan bahwa proses *Hyperparameter Tuning* berhasil meningkatkan kemampuan diskriminatif model, terutama dalam memisahkan kelas Berisiko dan Diabetes yang sebelumnya memiliki tumpang tindih cukup tinggi. Model hasil *tuning* tidak hanya meningkatkan akurasi keseluruhan, tetapi juga memperbaiki *recall* pada kelas minoritas, sehingga lebih layak untuk diimplementasikan dalam sistem deteksi dini berbasis *machine learning* pada konteks kesehatan.



**Gambar 3.** Confusion Matrix SVM dengan Tuning

**Tabel 1.** Perbandingan Matrix SVM dengan Tuning

Metrik	Tanpa Tuning (Default)	Dengan Tuning (Optimized)	Implikasi
Waktu Komputasi	44 menit 33 detik	2 jam 13 menit 1,47 detik	Model <i>tuning</i> memakan waktu hampir 3x lipat, namun menghasilkan performa yang jauh lebih andal.
Akurasi (Mean)	Total 0,9476 (~95%)	0,9890 (~99%)	Peningkatan akurasi keseluruhan sebesar $\pm 4\%$ .
Recall Diabetes	Kelas 0,67 (Rendah/Berbahaya)	0,94 (Sangat Baik)	<i>Tuning</i> meningkatkan kemampuan deteksi pasien Diabetes hingga 94%.
Total <i>false negative</i> Diabetes	<i>False</i> (FN) 4.851 kasus	915 kasus	Terjadi penurunan drastis pada kasus Diabetes yang salah diklasifikasikan sebagai berisiko.

Berdasarkan Waktu Komputasi Model tanpa *tuning* memerlukan waktu pelatihan selama 44 menit 33 detik. Sebaliknya, model dengan *tuning* membutuhkan waktu 2 jam 13 menit 1,47 detik. Hal ini menunjukkan bahwa proses *Hyperparameter Tuning* meningkatkan waktu komputasi hampir tiga kali lipat dibandingkan model default. Peningkatan waktu tersebut merupakan konsekuensi dari proses pencarian kombinasi parameter optimal yang dilakukan secara iteratif oleh Optuna. Meskipun waktu komputasi meningkat secara signifikan, proses *tuning* menghasilkan model dengan performa yang jauh lebih aman dan sensitif dalam mendeteksi kasus diabetes. Dengan demikian, tambahan waktu komputasi dapat dikategorikan sebagai *trade-off* yang rasional dalam konteks pengembangan sistem medis.

Berdasarkan akurasi total Model tanpa *tuning* menghasilkan akurasi sebesar 0,9476 ( $\pm 95\%$ ). Setelah dilakukan *tuning*, akurasi meningkat menjadi 0,9890 ( $\pm 99\%$ ). Peningkatan ini menunjukkan adanya kenaikan akurasi keseluruhan sebesar kurang lebih 4%.

Secara statistik, peningkatan ini mencerminkan kemampuan model yang lebih baik dalam melakukan klasifikasi secara umum. Dalam konteks sistem pendukung keputusan, peningkatan akurasi ini menunjukkan bahwa model hasil *tuning* memiliki tingkat kesalahan prediksi yang lebih rendah dibandingkan model *default*. Perbedaan paling signifikan terlihat pada metrik *recall* kelas Diabetes. Model tanpa *tuning* hanya mencapai nilai *recall* sebesar 0,67, yang tergolong rendah dan berisiko tinggi. Nilai ini menunjukkan bahwa sekitar 33% pasien diabetes tidak terdeteksi oleh sistem.

Setelah dilakukan *Hyperparameter Tuning*, nilai *recall* meningkat menjadi 0,94, yang termasuk kategori sangat baik. Artinya, model mampu mendeteksi 94% pasien diabetes secara benar. Peningkatan ini bersifat krusial karena dalam domain kesehatan, kesalahan dalam mendeteksi pasien positif (*false negative*) dapat berdampak serius terhadap keselamatan pasien. Dan Model tanpa *tuning* menghasilkan 4.851 kasus *false negative*, yaitu pasien diabetes yang salah diklasifikasikan sebagai non-diabetes atau berisiko rendah. Setelah dilakukan *tuning*, jumlah *false negative* menurun drastis menjadi 915 kasus.

Penurunan ini menunjukkan adanya peningkatan sensitivitas model yang sangat signifikan. Dalam konteks medis, pengurangan *false negative* jauh lebih penting dibandingkan sekadar peningkatan akurasi keseluruhan, karena *false negative* berpotensi menyebabkan keterlambatan diagnosis dan penanganan.

1. Secara keseluruhan, *Hyperparameter Tuning* memberikan peningkatan performa yang substansial, terutama pada metrik *recall* dan penurunan *false negative*. Meskipun waktu komputasi meningkat hampir tiga kali lipat, peningkatan sensitivitas model terhadap kelas minoritas menjadikan model hasil *tuning* lebih layak diterapkan dalam sistem deteksi dini berbasis *machine learning*.

## 4. Kesimpulan

Kesimpulan dari penelitian ini menunjukkan bahwa implementasi algoritma *Support Vector Machine (SVM)* berhasil diterapkan untuk klasifikasi risiko diabetes menggunakan dataset berskala besar. Model yang dikembangkan mampu mengolah data secara efektif dan menghasilkan performa klasifikasi yang tinggi. Hasil eksperimen membuktikan bahwa penerapan *Feature Selection* dan *Hyperparameter Tuning* memberikan peningkatan yang signifikan terhadap kinerja model. Peningkatan paling krusial terlihat pada nilai *recall* kelas Diabetes yang meningkat dari 0,67 menjadi 0,94, yang menunjukkan bahwa kemampuan model dalam mendeteksi pasien diabetes mengalami perbaikan yang sangat substansial. Hal ini penting dalam konteks medis, karena peningkatan *recall* berkontribusi langsung pada penurunan jumlah *false negative*, sehingga risiko kesalahan deteksi pasien dapat diminimalkan.

Selain itu, peningkatan akurasi keseluruhan hingga mencapai 98,90% mengindikasikan bahwa model memiliki kemampuan generalisasi yang baik terhadap data uji. Hasil ini menegaskan bahwa optimasi parameter dan pemilihan fitur yang tepat memainkan peran strategis dalam meningkatkan sensitivitas dan stabilitas model, terutama pada dataset medis yang memiliki karakteristik tidak seimbang. Dengan pendekatan optimasi yang sistematis, model mampu memisahkan kelas Berisiko dan Diabetes dengan lebih presisi dibandingkan konfigurasi *default*.

Meskipun demikian, penelitian ini masih memiliki keterbatasan. Dataset yang digunakan bersumber dari *Kaggle* sehingga representativitasnya terhadap populasi klinis nyata belum sepenuhnya terverifikasi [2]. Selain itu, penelitian ini belum melakukan validasi eksternal menggunakan dataset independen dari institusi kesehatan yang berbeda, sehingga potensi bias terhadap distribusi data tertentu masih mungkin terjadi. Oleh karena itu, penelitian lanjutan disarankan untuk mengintegrasikan pendekatan *cost-sensitive learning* guna menangani ketidakseimbangan kelas secara lebih adaptif serta melakukan validasi lintas populasi untuk meningkatkan robustness dan generalisasi model secara menyeluruh [6], [7]. Dengan pengembangan lebih lanjut, sistem ini berpotensi menjadi alat bantu deteksi dini berbasis *machine learning* yang dapat mendukung pengambilan keputusan preventif di bidang kesehatan..

## Daftar Pustaka

- [1] I. D. A. E. C. Astutisari, A. Y. D. AAA Yulianti Darmini, and I. A. P. W. Ida Ayu Putri Wulandari, "Hubungan Pola Makan Dan Aktivitas Fisik Dengan Kadar Gula Darah Pada Pasien Diabetes Melitus Tipe 2 Di Puskesmas Manggis I," *Jurnal Riset Kesehatan Nasional*, vol. 6, no. 2, pp. 79–87, Oct. 2022, doi: 10.37294/jrkn.v6i2.350.
- [2] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve Bayes algorithm," *Knowl Based Syst*, vol. 192, p. 105361, Mar. 2020, doi: 10.1016/j.knosys.2019.105361.
- [3] D. A. Pisner and D. M. Schnyer, "*Support Vector Machine*," in *Machine Learning*, Elsevier, 2020, pp. 101–121. doi: 10.1016/B978-0-12-815739-8.00006-7.
- [4] J. N.P. and R. Aruna, "Big data analytics in health care by data mining and classification techniques," *ICT Express*, vol. 8, no. 2, pp. 250–257, Jun. 2022, doi: 10.1016/j.ict.2021.07.001.
- [5] H. Apriyani and K. Kurniati, "Perbandingan Metode Naïve Bayes Dan *Support Vector Machine* Dalam Klasifikasi Penyakit Diabetes Melitus," *Journal of Information Technology Ampera*, vol. 1, no. 3, pp. 133–143, Dec. 2020, doi: 10.51519/journalita.volume1.issue3.year2020.page133-143.
- [6] A. S. Munir, A. B. Saputra, A. Aziz, and M. A. Barata, "Perbandingan Akurasi Algoritma Naive Bayes dan Algoritma Decision Tree dalam Pengklasifikasian Penyakit Kanker Payudara," *Jurnal Ilmiah Informatika Global*, vol. 15, no. 1, pp. 23–29, Apr. 2024, doi: 10.36982/jiig.v15i1.3578.
- [7] H. Fathurahman, A. Ariwikri, G. A. Pratama, M. A. F. S. FIKRI, and M. F. ALRIZKI, "PERBANDINGAN AKURASI METODE NAIVE BAYES CLASSIFIER DAN RANDOM FOREST MENGGUNAKAN REDUKSI DIMENSI LINEAR DICRIMINANT ANALYSIS (LDA) UNTUK DIAGNOSIS PENYAKIT DIABETES," *Jurnal Rekayasa Elektro Sriwijaya*, vol. 4, no. 1, pp. 24–31, Jan. 2023, doi: 10.36706/jres.v4i1.58.
- [8] E. Ramon, A. Nazir, N. Novriyanto, Y. Yusra, and L. Oktavia, "KLASIFIKASI STATUS GIZI BAYI POSYANDU KECAMATAN BANGUN PURBA MENGGUNAKAN ALGORITMA *SUPPORT VECTOR MACHINE (SVM)*," *Jurnal Sistem Informasi dan Informatika (Simika)*, vol. 5, no. 2, pp. 143–150, Aug. 2022, doi: 10.47080/simika.v5i2.2185.

- [9] P. H. Prastyo, A. S. Sumi, A. W. Dian, and A. E. Permanasari, "Tweets Responding to the Indonesian Government's Handling of COVID-19: Sentiment Analysis Using SVM with Normalized Poly Kernel," *Journal of Information Systems Engineering and Business Intelligence*, vol. 6, no. 2, p. 112, Oct. 2020, doi: 10.20473/jisebi.6.2.112-122.
- [10] Y. Harni, I. Afrianty, S. Sanjaya, R. Abdillah, F. Yanto, and F. Syafria, "Performance Analysis of LVQ 1 Using *Feature Selection* Gain Ratio for Sex Classification in Forensic Anthropology," *Building of Informatics, Technology and Science (BITS)*, vol. 5, no. 1, Jun. 2023, doi: 10.47065/bits.v5i1.3625.
- [11] Aswin Ardiansyah, Enos C.O.Telaumbanua, Aron S. Gultom, and Angelita A. S. M. Limbong, "Klasifikasi Penyakit Diabetes Menggunakan Metode SVM Dan KNN," *Jurnal Penelitian Rumpun Ilmu Teknik*, vol. 3, no. 1, pp. 77–83, Dec. 2023, doi: 10.55606/juprit.v3i1.3151.
- [12] I. Afrianty, D. Nasien, and H. Haron, "Performance Analysis of *Support Vector Machine* in Sex Classification of The Sacrum Bone in Forensic Anthropology," *JURNAL TEKNIK INFORMATIKA*, vol. 15, no. 1, pp. 63–72, Jun. 2022, doi: 10.15408/jti.v15i1.25254.
- [13] R. Islam, A. Sultana, Md. N. Tuhin, Md. S. H. Saikat, and M. R. Islam, "Clinical Decision Support System for Diabetic Patients by Predicting Type 2 Diabetes Using *Machine learning* Algorithms," *J Healthc Eng*, vol. 2023, no. 1, Jan. 2023, doi: 10.1155/2023/6992441.
- [14] N. Aggarwal et al., "Mean based relief: An improved *Feature Selection* method based on ReliefF," *Applied Intelligence*, vol. 53, no. 19, pp. 23004–23028, Oct. 2023, doi: 10.1007/s10489-023-04662-w.
- [15] L. Silalahi, "Hubungan Pengetahuan dan Tindakan Pencegahan Diabetes Mellitus Tipe 2," *Jurnal PROMKES*, vol. 7, no. 2, p. 223, Dec. 2019, doi: 10.20473/jpk.v7.i2.2019.223-232.
- [16] Nafisfaturr, "Big Data - Diabetes classification," *Kaggle*. Accessed: Oct. 05, 2025. [Online]. Available: <https://www.Kaggle.com/code/nafisfaturr/big-data-diabetes-classification/input>
- [17] P. R. Putri and R. Alit, "Klasifikasi Penyakit Diabetes Melitus Menggunakan Metode *Support Vector Machine* (SVM)," *Journal of Informatics and Computer Science*, vol. 06, 2024.
- [18] F. Aninda Nurdila, U. A. Dahlan, and I. ; Yogyakarta, "Penggunaan Metode Classification And Regression Tree (CART) Dalam Mengklasifikasikan Faktor Yang Mempengaruhi Penyakit Diabetes", doi: 10.26555/konvergensi.30878.
- [19] N. al Rasyid, I. Afrianty, E. Budianita, and S. Kurnia Gusti, "Diabetes Classification using Gain Ratio *Feature Selection* in *Support Vector Machine* Method," *Bulletin of Informatics and Data Science*, vol. 4, no. 1, 2025, doi: 10.61944/bids.v4i1.114.
- [20] Asri Mulyani, Sarah Khoerunisa, and Dede Kurniadi, "Perbandingan Kinerja Algoritma KNN dan SVM Menggunakan SMOTE untuk Klasifikasi Penyakit Diabetes," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 14, no. 1, pp. 25–34, Feb. 2025, doi: 10.22146/jnteti.v14i1.15198.
- [21] A. Wantoro, A. Fitriya Yuliana, D. Yana, A. Andini, I. Awaliyani, and W. Caesarendra, "Optimizing Type 2 Diabetes Classification with *Feature Selection* and Class Balancing in Machine Learning," *Jurnal Teknik Informatika (JUTIF)*, vol. 6, no. 4, pp. 2723–3863, 2025, doi: 10.52436/1.jutif.2025.6.4.5166.
- [22] F. Refindha, A. Harianto, Z. Alawi, and I. A. Sa'ida, "PENGARUH KOMPOSISI SPLIT DATA PADA AKURASI KLASIFIKASI PENDERITA DIABETES MENGGUNAKAN ALGORITMA MACHINE LEARNING," *Jurnal Sistem Informasi dan Informatika (Simika)*, vol. 8, no. 1, 2025.
- [23] A. D. Cahyani and A. Basuki, "Klasifikasi Diabetes Mellitus Menggunakan *Support Vector Machine* (Studi Kasus: Puskesmas Modopuro, Mojokerto)," *Rekayasa*, vol. 12, no. 2, pp. 174–182, Oct. 2019, doi: 10.21107/rekayasa.v12i2.19763.
- [24] C. Z. V. Junus, T. Tarno, and P. Kartikasari, "KLASIFIKASI MENGGUNAKAN METODE *SUPPORT VECTOR MACHINE* DAN RANDOM FOREST UNTUK DETEKSI AWAL RISIKO DIABETES MELITUS," *Jurnal Gaussian*, vol. 11, no. 3, pp. 386–396, Jan. 2023, doi: 10.14710/j.gauss.11.3.386-396.
- [25] M. Muhammad, J. Samodro, M. Kunta Biddinika, A. Fadlil, A. Dahlan, and Y. J. Ringroad Selatan, "Klasifikasi Penyakit Diabetes dengan Algoritma Decision Tree dan Naïve Bayes," vol. 6, no. 2.
- [26] E. Ramadanti, D. A. Dinathi, C. Sri, K. Aditya, and R. Chandranegara, "Diabetes Disease Detection Classification Using Light Gradient Boosting (LightGBM) With *Hyperparameter Tuning* ," *Jurnal dan Penelitian Teknik Informatika*, vol. 8, no. 2, 2024, doi: 10.33395/v8i2.13530.
- [27] S. Rahmawati, A. Wibowo, and A. F. N. Masruriyah, "Improving Diabetes Prediction Accuracy in Indonesia: A Comparative Analysis of SVM, Logistic Regression, and Naive Bayes with SMOTE and ADASYN," *Jurnal RESTI*, vol. 8, no. 5, pp. 607–614, Oct. 2024, doi: 10.29207/resti.v8i5.5980.
- [28] N. Nurdiana and A. Algifari, "STUDI KOMPARASI ALGORITMA ID3 DAN ALGORITMA NAIVE BAYES UNTUK KLASIFIKASI PENYAKIT DIABETES MELITUS".