HIJM
HORIZON: Indonesian Journal of Multidisciplinary

p-ISSN: 2987-7776
e-ISSN: 2987-7180
Vol. 4, No. 3, 2026

PERBANDINGAN METODE KLASIFIKASI SUPPORT VECTOR MACHINE DAN K-NEAREST NEIGHBOR DALAM PREDIKSI PENYAKIT DIABETES MELLITUS

MHD. Aldi Fathoni[1], Benni Purnama[2], Dodo Zaenal Abidin
[1, 2, 3]Universitas Dinamika Bangsa, Jl. Jend. Sudirman, Jambi, Indonesia
Email: fathonialdi849@gmail.com

Article History

Received: 19-04-2026

Revision: 07-05-2026

Accepted: 10-05-2026

Published: 13-05-2026

Abstract. This study aims to compare the performance of the Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) classification methods in predicting diabetes mellitus. The study employs a machine learning-based experimental approach, utilising a secondary dataset from Kaggle comprising 70,000 patient records and 34 variables, predominantly consisting of clinical data and individual health characteristics as the main variables. The data was split into training and test sets in an 80:20 ratio to ensure the model received sufficient training data whilst maintaining the objectivity of the testing. Model performance was evaluated using the accuracy, precision, recall, and F1-score metrics. The results of the study showed that SVM delivered superior performance with an accuracy of 78.21%, whilst KNN only achieved an accuracy of 47.23%. SVM also demonstrated better prediction stability across most classes. These findings indicate that SVM is more effective for predicting Diabetes Mellitus on datasets with a relatively large number of features and a complex class structure.

Keywords: Diabetes Mellitus, Machine Learning, Support Vector Machine, K-Nearest Neighbor, Classification

Abstrak. Penelitian ini bertujuan membandingkan kinerja metode klasifikasi *Support Vector Machine* (SVM) dan *K-Nearest Neighbor* (KNN) dalam memprediksi penyakit Diabetes Mellitus. Penelitian menggunakan pendekatan eksperimen berbasis *machine learning* dengan memanfaatkan dataset sekunder dari Kaggle yang berjumlah 70.000 data pasien dan 34 variabel, yang didominasi oleh data klinis dan karakteristik kesehatan individu sebagai variabel utama. Data dibagi menjadi data latih dan data uji dengan rasio 80:20 untuk memastikan model memperoleh data pelatihan yang cukup sekaligus menjaga objektivitas pengujian. Evaluasi kinerja model dilakukan menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score*. Hasil penelitian menunjukkan bahwa SVM memberikan performa yang lebih unggul dengan akurasi sebesar 78,21%, sedangkan KNN hanya mencapai akurasi 47,23%. SVM juga menunjukkan kestabilan prediksi yang lebih baik pada sebagian besar kelas. Temuan ini menunjukkan bahwa SVM lebih efektif digunakan untuk prediksi Diabetes Mellitus pada dataset dengan jumlah fitur yang relatif banyak dan struktur kelas yang kompleks.

Kata Kunci: Diabetes Mellitus, Machine Learning, Support Vector Machine, K-Nearest Neighbor, Klasifikasi

How to Cite: Fathoni, M. A., Purnama, B., & Abidin, D. Z. (2026). Perbandingan Metode Klasifikasi *Support Vector Machine* dan *K-Nearest Neighbor* dalam Prediksi Penyakit Diabetes Mellitus. *HORIZON: Indonesian Journal of Multidisciplinary*, 4 (3), 490-507. http://doi.org/10.54373/hijm.v4i3.5417

PENDAHULUAN

Diabetes Mellitus merupakan salah satu permasalahan kesehatan global yang memerlukan perhatian serius karena berpotensi menimbulkan komplikasi berat hingga kematian. Penyakit ini termasuk gangguan metabolik kronis yang ditandai dengan peningkatan kadar glukosa darah akibat gangguan sekresi insulin, kerja insulin, atau keduanya, dan menjadi salah satu

penyebab utama meningkatnya angka morbiditas dan mortalitas di dunia (World Health Organization, 2021). Peningkatan prevalensi Diabetes Mellitus mendorong perlunya metode prediksi yang akurat untuk mendukung deteksi dini dan mencegah komplikasi jangka panjang yang berdampak luas pada kualitas hidup pasien (Singh et al., 2020). Sejumlah penelitian menunjukkan bahwa pendekatan berbasis machine learning mampu memberikan tingkat akurasi prediksi yang lebih baik dibandingkan metode konvensional dalam analisis data medis pasien Diabetes Mellitus (Perveen et al., 2020).

Deteksi dini Diabetes Mellitus memiliki peran penting dalam menurunkan risiko komplikasi seperti penyakit kardiovaskular, gagal ginjal, dan neuropati (Kavakiotis et al., 2020). Namun, proses diagnosis konvensional yang bergantung pada pemeriksaan laboratorium sering kali membutuhkan waktu dan biaya yang relatif tinggi (Jain & Singh, 2021). Kondisi ini mendorong pemanfaatan algoritma klasifikasi dalam machine learning sebagai alternatif yang lebih efisien untuk memprediksi Diabetes Mellitus berdasarkan data klinis pasien. Model klasifikasi dinilai mampu membantu tenaga medis dalam pengambilan keputusan secara lebih cepat, objektif, dan berbasis data (Sneha & Gangil, 2020).

Support Vector Machine (SVM) merupakan algoritma supervised learning yang banyak digunakan dalam bidang medis karena kemampuannya menangani data berdimensi tinggi dan membentuk batas pemisah yang optimal antar kelas (Zheng et al., 2020). Sejumlah penelitian menunjukkan bahwa SVM mampu menghasilkan akurasi yang tinggi dan stabil dalam prediksi Diabetes Mellitus pada berbagai dataset medis, terutama dalam mengenali pola nonlinier yang kompleks (Sisodia & Sisodia, 2020; Swapna et al., 2021). Karakteristik tersebut menjadikan SVM sebagai salah satu algoritma yang potensial dalam pengembangan sistem prediksi penyakit berbasis data.

Selain SVM, algoritma *K-Nearest Neighbor* (KNN) juga banyak digunakan dalam klasifikasi data medis. KNN merupakan metode berbasis jarak yang mengklasifikasikan data baru berdasarkan kedekatannya dengan data latih terdekat (Patil et al., 2020). Algoritma ini dikenal sederhana dan cukup efektif untuk dataset dengan jumlah fitur yang terbatas. Beberapa studi menunjukkan bahwa KNN mampu memberikan performa prediksi yang kompetitif dalam klasifikasi Diabetes Mellitus, terutama ketika parameter k dan metrik jarak ditentukan secara optimal (Haq et al., 2020; Rahman et al., 2021).

Perbedaan karakteristik data medis menyebabkan tidak ada satu algoritma klasifikasi yang selalu unggul pada semua kondisi (Ahmed et al., 2022). Oleh karena itu, studi perbandingan kinerja algoritma menjadi penting untuk menentukan metode yang paling sesuai dengan karakteristik dataset yang digunakan (Ali et al., 2020). Penelitian terdahulu menunjukkan hasil

yang beragam. Pratama et al. (2024) menemukan bahwa Random Forest lebih unggul dibandingkan KNN dan Logistic Regression pada dataset Pima Indians dengan akurasi 77%. Kurniawan dan Megawaty (2025) melaporkan bahwa Random Forest memberikan performa tertinggi pada dataset berskala besar dengan penerapan teknik SMOTE. Sementara itu, Hidayat et al. (2024) menunjukkan bahwa Logistic Regression memiliki stabilitas dan akurasi terbaik dibandingkan beberapa algoritma lainnya. Variasi hasil tersebut menunjukkan bahwa performa algoritma sangat bergantung pada karakteristik data, ukuran dataset, dan teknik prapemrosesan yang digunakan.

Berdasarkan kondisi tersebut, perbandingan kinerja SVM dan KNN dalam prediksi Diabetes Mellitus masih relevan untuk diteliti, khususnya pada dataset dengan jumlah data dan fitur yang relatif besar. Penelitian ini diharapkan dapat memberikan kontribusi dalam pemilihan algoritma klasifikasi yang tepat serta mendukung pengembangan sistem pendukung keputusan medis berbasis machine learning. Dengan demikian, tujuan penelitian ini adalah menganalisis dan membandingkan kinerja metode klasifikasi *Support Vector Machine* dan *K-Nearest Neighbor* dalam memprediksi penyakit Diabetes Mellitus.

METODE

Penelitian ini menggunakan pendekatan eksperimen berbasis machine learning untuk membandingkan kinerja algoritma *Support Vector Machine* (SVM) dan *K-Nearest Neighbor* (KNN) dalam memprediksi penyakit Diabetes Mellitus. Data yang digunakan merupakan dataset sekunder yang diperoleh dari platform Kaggle, terdiri atas 70.000 data pasien dengan 34 fitur yang merepresentasikan atribut kesehatan dan demografis. Kelas target dalam dataset ini bersifat biner, yaitu status Diabetes Mellitus (positif dan negatif). Distribusi kelas menunjukkan ketidakseimbangan moderat, sehingga diperlukan evaluasi model yang tidak hanya bergantung pada akurasi, tetapi juga pada metrik lain seperti *precision*, *recall*, dan *F1-score*.

Tahap prapemrosesan data dilakukan secara singkat dan terfokus pada kesiapan data untuk pemodelan. Proses ini meliputi pemeriksaan data hilang, penyesuaian tipe data, pemisahan variabel fitur dan label, serta normalisasi fitur numerik untuk mendukung kinerja algoritma berbasis jarak dan margin. Dataset kemudian dibagi menjadi data latih dan data uji dengan rasio 80:20, yang dipilih untuk memberikan proporsi data latih yang cukup besar agar model dapat mempelajari pola secara optimal, sekaligus menyediakan data uji yang representatif untuk evaluasi kinerja.

Tahap pemodelan dilakukan dengan mengimplementasikan algoritma SVM dan KNN menggunakan pustaka *scikit-learn* pada bahasa pemrograman *Python*. Model dilatih menggunakan data latih dan diuji menggunakan data uji untuk menilai kemampuan generalisasi terhadap data yang belum pernah dilihat sebelumnya. Parameter dasar digunakan pada masing-masing algoritma untuk memastikan perbandingan kinerja yang adil dan terkontrol.

Evaluasi kinerja model dilakukan menggunakan metrik akurasi, presisi, *recall*, dan *F1-score*, serta didukung dengan confusion matrix untuk menganalisis kesalahan klasifikasi secara lebih rinci, termasuk false positive dan false negative. Hasil evaluasi dari kedua algoritma kemudian dibandingkan untuk menentukan metode yang memiliki performa paling optimal dalam prediksi Diabetes Mellitus berdasarkan karakteristik dataset yang digunakan.

Seluruh proses analisis dilakukan menggunakan lingkungan *Google Colaboratory*. Hasil penelitian disajikan dalam bentuk tabel dan uraian analitis untuk menjawab tujuan penelitian, yaitu menentukan algoritma klasifikasi yang lebih efektif dalam memprediksi penyakit Diabetes Mellitus serta memberikan dasar empiris bagi pengembangan sistem pendukung keputusan medis berbasis machine learning.

HASIL DAN DISKUSI

Deskripsi Umum Dataset

Variabel dalam dataset terbagi menjadi dua jenis, yaitu numerik dan kategorikal. Variabel numerik meliputi usia (*age*), indeks massa tubuh (*body mass index*), tekanan darah, kadar glukosa darah, serta beberapa indikator klinis lainnya. Variabel kategorikal mencakup informasi riwayat keluarga diabetes, status merokok, pola makan, serta hasil pemeriksaan medis tertentu seperti *glucose tolerance test* dan *genetic testing*. Variabel target dalam penelitian ini adalah jenis diabetes melitus yang terdiri atas 13 kelas, antara lain *tipe 1 diabetes*, *tipe 2 diabetes*, *prediabetes*, *diabetes gestasional*, *MODY*, *LADA*, dan jenis diabetes lainnya.

Distribusi data pada setiap kelas target relatif seimbang, sehingga dataset ini tidak mengalami permasalahan ketidakseimbangan kelas (*imbalanced data*). Hasil pemeriksaan awal menunjukkan bahwa dataset tidak mengandung nilai kosong (*missing values*) maupun data duplikat, sehingga kualitas data tergolong baik dan siap digunakan untuk analisis lebih lanjut. Secara keseluruhan, dataset memiliki karakteristik yang kompleks karena melibatkan jumlah fitur yang cukup banyak, tipe data yang beragam, serta kelas target yang multikelas. Kondisi ini menjadikan tahapan prapemrosesan dan pemilihan algoritma klasifikasi sebagai aspek penting untuk memperoleh model prediksi yang optimal dan andal.

Visualisasi Struktur Dataset

Struktur dataset pada penelitian ini terdiri dari 34 kolom yang mencakup berbagai informasi terkait kondisi kesehatan, gaya hidup, serta faktor medis yang berhubungan dengan penyakit Diabetes Mellitus. Setiap kolom memiliki tipe data yang berbeda, yaitu *numerical data* (data numerik) dan *categorical data* (data kategorikal). Kolom target dalam dataset ini adalah variabel *Target*, yang berisi informasi mengenai jenis penyakit diabetes yang diderita oleh setiap individu. Sementara itu, 33 kolom lainnya merupakan fitur (*features*) yang digunakan sebagai variabel prediktor dalam proses klasifikasi.

Tabel 1. *Feature Dataset*

No	Nama Kolom	Tipe Data	Penjelasan
1	*Genetic Markers*	Kategorikal	Menunjukkan apakah terdapat penanda genetik yang berhubungan dengan diabetes.
2	*Autoantibodies*	Kategorikal	Menunjukkan adanya antibodi yang menyerang sel tubuh sendiri, sering terkait dengan diabetes tipe 1.
3	*Family History*	Kategorikal	Menunjukkan apakah terdapat riwayat diabetes dalam keluarga.
4	*Environmental Factors*	Kategorikal	Faktor lingkungan seperti pola hidup atau kondisi sekitar yang dapat memicu diabetes.
5	*Insulin Levels*	Numerik	Jumlah insulin dalam tubuh, hormon yang mengatur kadar gula darah.
6	*Age*	Numerik	Usia individu dalam tahun.
7	*BMI*	Numerik	Indeks massa tubuh yang menunjukkan apakah berat badan normal, kurang, atau berlebih.
8	*Physical Activity*	Kategorikal	Tingkat aktivitas fisik individu (rendah, sedang, tinggi).
9	*Dietary Habits*	Kategorikal	Pola makan individu, apakah sehat atau tidak sehat.
10	*Blood Pressure*	Numerik	Tekanan darah individu.
11	*Cholesterol Levels*	Numerik	Kadar kolesterol dalam darah.
12	*Waist Circumference*	Numerik	Lingkar pinggang yang berkaitan dengan risiko obesitas.
13	*Blood Glucose Levels*	Numerik	Kadar gula dalam darah, indikator utama diabetes.
14	*Ethnicity*	Kategorikal	Kelompok etnis yang dapat mempengaruhi risiko diabetes.
15	*Socioeconomic Factors*	Kategorikal	Kondisi sosial dan ekonomi individu (rendah, sedang, tinggi).
16	*Smoking Status*	Kategorikal	Status merokok individu (perokok atau tidak).
17	*Alcohol Consumption*	Kategorikal	Tingkat konsumsi alkohol individu.
18	*Glucose Tolerance Test*	Kategorikal	Hasil tes toleransi glukosa untuk melihat kemampuan tubuh mengolah gula.
19	*History of PCOS*	Kategorikal	Riwayat sindrom ovarium polikistik yang berhubungan dengan risiko diabetes.
20	*Previous*	Kategorikal	Riwayat diabetes saat kehamilan sebelumnya.

	Gestational Diabetes		
21	*Pregnancy History*	Kategorikal	Riwayat kehamilan, apakah normal atau terdapat komplikasi.
22	*Weight Gain During Pregnancy*	Numerik	Penambahan berat badan selama kehamilan.
23	*Pancreatic Health*	Numerik	Kondisi kesehatan pankreas, organ penghasil insulin.
24	*Pulmonary Function*	Numerik	Fungsi paru-paru individu.
25	*Cystic Fibrosis Diagnosis*	Kategorikal	Apakah individu didiagnosis dengan *cystic fibrosis*.
26	*Steroid Use History*	Kategorikal	Riwayat penggunaan obat steroid.
27	*Genetic Testing*	Kategorikal	Hasil tes genetik terkait risiko diabetes.
28	*Neurological Assessments*	Numerik	Hasil pemeriksaan saraf.
29	*Liver Function Tests*	Kategorikal	Hasil tes fungsi hati.
30	*Digestive Enzyme Levels*	Numerik	Kadar enzim pencernaan dalam tubuh.
31	*Urine Test*	Kategorikal	Hasil tes urin untuk mendeteksi zat tertentu seperti glukosa atau protein.
32	*Birth Weight*	Numerik	Berat badan saat lahir.
33	*Early Onset Symptoms*	Kategorikal	Apakah gejala diabetes muncul sejak usia dini.

Target Kelas pada Dataset

Seluruh kolom dalam dataset memiliki jumlah data yang lengkap (70.000 *non-null*), sehingga tidak terdapat nilai kosong (*missing values*). Selain itu, tipe data telah sesuai dengan karakteristik masing-masing variabel, dimana data numerik digunakan untuk nilai kuantitatif dan data kategorikal digunakan untuk label atau kategori. Selanjutnya, berdasarkan variabel *Target*, dataset ini terdiri dari 13 kelas jenis Diabetes Mellitus. Berdasarkan distribusi tersebut, dapat diketahui bahwa jumlah data pada setiap kelas relatif seimbang. Hal ini menunjukkan bahwa dataset tidak mengalami masalah ketidakseimbangan kelas (*class imbalance*), yang sering menjadi kendala dalam proses klasifikasi

Tabel 2. Penjelasan kelas target diabetes mellitus

No	Target Kelas	Jumlah Data	Penjelasan
1.	*MODY (Maturity Onset Diabetes of the Young)*	5.553	Jenis diabetes yang disebabkan oleh faktor genetik dan biasanya muncul pada usia muda. Tidak selalu membutuhkan insulin.
2.	*Secondary Diabetes*	5.479	Diabetes yang muncul akibat penyakit lain atau

			penggunaan obat tertentu, bukan karena faktor utama seperti gaya hidup atau genetik.
3.	*Cystic Fibrosis-Related Diabetes (CFRD)*	5.464	Diabetes yang terjadi pada penderita penyakit cystic fibrosis, yaitu gangguan genetik yang mempengaruhi paru-paru dan sistem pencernaan.
4.	*Type 1 Diabetes*	5.446	Diabetes yang terjadi karena sistem imun menyerang sel penghasil insulin. Biasanya terjadi sejak usia muda dan membutuhkan insulin seumur hidup.
5.	*Neonatal Diabetes Mellitus (NDM)*	5.408	Diabetes langka yang muncul pada bayi sejak lahir atau dalam 6 bulan pertama kehidupan.
6.	*Wolcott-Rallison Syndrome*	5.400	Penyakit genetik langka yang menyebabkan diabetes sejak usia sangat dini dan disertai gangguan organ lain.
7.	*Type 2 Diabetes*	5.397	Jenis diabetes paling umum yang terjadi karena tubuh tidak mampu menggunakan insulin dengan baik (insulin resistance). Biasanya terkait gaya hidup.
8.	*Prediabetic*	5.376	Kondisi sebelum diabetes, dimana kadar gula darah lebih tinggi dari normal tetapi belum masuk kategori diabetes.
9.	*Gestational Diabetes*	5.344	Diabetes yang terjadi selama masa kehamilan dan biasanya akan hilang setelah melahirkan.
10.	*Type 3c Diabetes (Pancreatogenic Diabetes)*	5.320	Diabetes yang disebabkan oleh kerusakan pankreas, misalnya karena penyakit atau cedera.
11.	*Wolfram Syndrome*	5.315	Penyakit genetik langka yang menyebabkan diabetes disertai gangguan penglihatan dan saraf.
12.	*Steroid-Induced Diabetes*	5.275	Diabetes yang muncul akibat penggunaan obat steroid dalam jangka waktu tertentu.
13.	*LADA (Latent Autoimmune Diabetes in Adults)*	5.223	Jenis diabetes yang mirip Type 1, tetapi muncul lebih lambat pada orang dewasa dan sering awalnya terlihat seperti Type 2.

Pembersihan Data (*Preprocessing*)

Data Cleaning

Tahap *data cleaning* merupakan proses awal dalam *preprocessing* yang bertujuan untuk memastikan kualitas data sebelum digunakan dalam proses pemodelan. Data yang bersih dan konsisten sangat penting agar model *machine learning* dapat bekerja secara optimal dan menghasilkan prediksi yang akurat. Pada penelitian ini, proses *data cleaning* dilakukan dengan beberapa langkah utama, yaitu pengecekan data duplikat dan pemeriksaan konsistensi data kategorikal.

- Melakukan pengecekan terhadap data duplikat. Data duplikat merupakan baris data yang memiliki nilai yang sama pada seluruh kolom. Keberadaan data duplikat dapat menyebabkan bias pada model karena informasi yang sama dihitung lebih dari satu kali. Berdasarkan hasil pemeriksaan, tidak ditemukan data duplikat dalam dataset, sehingga tidak dilakukan penghapusan data. Jumlah data tetap sebanyak 70.000 baris setelah proses ini.
- Melakukan pemeriksaan terhadap konsistensi data, khususnya pada kolom kategorikal. Hal ini dilakukan untuk memastikan bahwa tidak terdapat perbedaan penulisan yang dapat menyebabkan kategori dianggap berbeda, seperti perbedaan huruf besar dan kecil atau variasi penulisan yang tidak konsisten.
- Pengecekan terhadap nilai unik pada setiap kolom kategorikal untuk memastikan bahwa seluruh kategori telah sesuai dengan definisi yang diharapkan. Hasilnya menunjukkan bahwa seluruh nilai dalam dataset telah konsisten dan tidak terdapat anomali data.

Tabel 3. Hasil data cleaning

No	Nama Kolom	Nilai Unik
1	*Target*	Steroid-Induced Diabetes, Neonatal Diabetes Mellitus (NDM), Prediabetic, Type 1 Diabetes, Wolfram Syndrome, LADA, Type 2 Diabetes, Wolcott-Rallison Syndrome, Secondary Diabetes, Type 3C Diabetes, Gestational Diabetes, CFRD, MODY
2	*Genetic Markers*	Positive, Negative
3	*Autoantibodies*	Positive, Negative
4	*Family History*	Yes, No
5	*Environmental Factors*	Present, Absent
6	*Physical Activity*	High, Moderate, Low
7	*Dietary Habits*	Healthy, Unhealthy
8	*Ethnicity*	Low Risk, High Risk
9	*Socioeconomic Factors*	Low, Medium, High
10	*Smoking Status*	Smoker, Non-Smoker
11	*Alcohol Consumption*	Low, Moderate, High
12	*Glucose Tolerance Test*	Normal, Abnormal
13	*History of PCOS*	Yes, No
14	*Previous Gestational Diabetes*	Yes, No
15	*Pregnancy History*	Normal, Complications
16	*Cystic Fibrosis Diagnosis*	Yes, No
17	*Steroid Use*	Yes, No

	History	
18	*Genetic Testing*	Positive, Negative
19	*Liver Function Tests*	Normal, Abnormal
20	*Urine Test*	Ketones Present, Glucose Present, Protein Present, Normal
21	*Early Onset Symptoms*	Yes, No

Berdasarkan hasil *data cleaning*, dapat disimpulkan bahwa dataset tidak memiliki data duplikat serta memiliki kategori yang konsisten pada setiap variabel. Hal ini menunjukkan bahwa dataset memiliki kualitas yang baik dan siap untuk digunakan pada tahap *preprocessing* selanjutnya.

Encoding

Tahap *data encoding* merupakan proses mengubah data kategorikal (teks) menjadi data numerik agar dapat diproses oleh algoritma *machine learning*. Hal ini diperlukan karena sebagian besar algoritma, seperti *Support Vector Machine (SVM)*, dan *K-Nearest Neighbor (KNN)*, hanya dapat bekerja dengan data dalam bentuk numerik. Pada penelitian ini, sebagian besar fitur dalam dataset bertipe kategorikal, seperti *Genetic Markers*, *Family History*, dan *Smoking Status*. Oleh karena itu, dilakukan proses *encoding* menggunakan metode *One Hot Encoding*.

Metode *One Hot Encoding* bekerja dengan cara mengubah setiap kategori menjadi kolom baru yang berisi nilai biner (*True/False* atau 1/0). Sebagai contoh, kolom *Genetic Markers* yang memiliki nilai *Positive* dan *Negative* diubah menjadi satu kolom baru yaitu *Genetic Markers_Positive*. Nilai *True* menunjukkan kategori tersebut aktif, sedangkan *False* menunjukkan sebaliknya. Penggunaan *One Hot Encoding* dipilih karena mampu menghindari asumsi hubungan ordinal (tingkatan) antar kategori. Hal ini penting karena sebagian besar variabel dalam dataset tidak memiliki urutan tertentu, sehingga tidak tepat jika menggunakan *Label Encoding*. Perlu diperhatikan bahwa pada penelitian ini, kolom *Target* tidak dilakukan proses *encoding*. Hal ini karena algoritma klasifikasi yang digunakan dapat langsung menangani label dalam bentuk teks (*multi-class classification*), serta untuk menjaga interpretasi hasil agar tetap mudah dipahami. Setelah proses *encoding* dilakukan, jumlah kolom dalam dataset mengalami perubahan. Hal ini terjadi karena setiap kategori diubah menjadi kolom baru.

Tabel 4. hasil data *encoding*

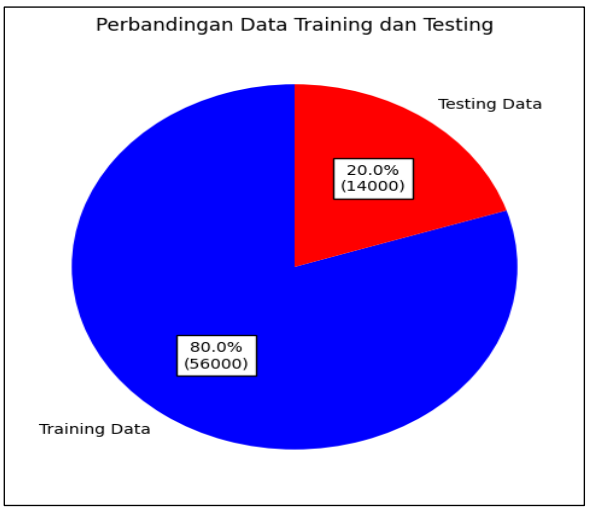
No	Nama Kolom	Hasil Encoding
1	*Insulin Levels*	Numerik
2	*Age*	Numerik
3	*BMI*	Numerik
4	*Blood Pressure*	Numerik
5	*Cholesterol Levels*	Numerik
6	*Waist Circumference*	Numerik
7	*Blood Glucose Levels*	Numerik
8	*Weight Gain During Pregnancy*	Numerik
9	*Pancreatic Health*	Numerik
10	*Pulmonary Function*	Numerik
11	*Neurological Assessments*	Numerik
12	*Digestive Enzyme Levels*	Numerik
13	*Birth Weight*	Numerik
14	*Genetic Markers_Positive*	True / False
15	*Autoantibodies_Positive*	True / False
16	*Family History_Yes*	True / False
17	*Environmental Factors_Present*	True / False
18	*Physical Activity_Low*	True / False
19	*Physical Activity_Moderate*	True / False
20	*Dietary Habits_Unhealthy*	True / False
21	*Ethnicity_Low Risk*	True / False
22	*Socioeconomic Factors_Low*	True / False
23	*Socioeconomic Factors_Medium*	True / False
24	*Smoking Status_Smoker*	True / False
25	*Alcohol Consumption_Low*	True / False
26	*Alcohol Consumption_Moderate*	True / False
27	*Glucose Tolerance Test_Normal*	True / False
28	*History of PCOS_Yes*	True / False
29	*Previous Gestational Diabetes_Yes*	True / False
30	*Pregnancy History_Normal*	True / False
31	*Cystic Fibrosis Diagnosis_Yes*	True / False
32	*Steroid Use History_Yes*	True / False
33	*Genetic Testing_Positive*	True / False
34	*Liver Function Tests_Normal*	True / False
35	*Urine Test_Ketones Present*	True / False
36	*Urine Test_Normal*	True / False
37	*Urine Test_Protein Present*	True / False
38	*Early Onset Symptoms_Yes*	True / False

Split Data

Tahap *split data* merupakan proses pembagian dataset menjadi dua bagian, yaitu data latih (*training data*) dan data uji (*testing data*). Data latih digunakan untuk melatih model *machine learning*, sedangkan data uji digunakan untuk mengevaluasi performa model terhadap data yang belum pernah dilihat sebelumnya.

Pada penelitian ini, pembagian data dilakukan menggunakan metode *train-test split* dengan perbandingan 80:20. Artinya, sebanyak 80% data digunakan sebagai data latih dan 20% sisanya digunakan sebagai data uji. Pemilihan proporsi ini bertujuan untuk memberikan cukup data bagi model untuk belajar, sekaligus menyediakan data uji yang representatif untuk evaluasi. Selain itu, digunakan parameter *stratify* pada proses pembagian data. Tujuan dari penggunaan *stratify* adalah untuk menjaga distribusi kelas pada data latih dan data uji tetap seimbang. Hal ini sangat penting karena dataset memiliki 13 kelas (*multi-class classification*), sehingga setiap kelas harus tetap terwakili secara proporsional pada kedua subset data.

Berdasarkan hasil proses *split data*, diperoleh sebanyak 56.000 data sebagai data latih dan 14.000 data sebagai data uji. Jumlah fitur yang digunakan dalam pemodelan adalah sebanyak 38 fitur, yang merupakan hasil dari proses *encoding* sebelumnya. Dengan demikian, data telah siap digunakan pada tahap selanjutnya, yaitu proses *feature scaling* sebelum dilakukan pelatihan model *machine learning*.



Gambar 1. Hasil Split Data

Scaling

Tahap *feature scaling* merupakan proses menyamakan skala nilai pada setiap fitur dalam dataset agar berada pada rentang yang sebanding. Proses ini diperlukan karena dalam dataset terdapat perbedaan nilai yang cukup jauh antar fitur. Sebagai contoh, fitur seperti *Age* memiliki nilai dalam puluhan, sedangkan *Blood Glucose Levels* dapat mencapai ratusan, dan beberapa fitur hasil *encoding* hanya bernilai 0 dan 1. Perbedaan skala ini dapat menyebabkan model *machine learning* menjadi tidak seimbang dalam membaca data, karena fitur dengan nilai yang lebih besar akan dianggap lebih penting dibandingkan fitur dengan nilai kecil.

Untuk mengatasi hal tersebut, dilakukan proses *feature scaling* menggunakan metode *StandardScaler*. Metode ini bekerja dengan cara mengubah nilai data sehingga memiliki rata-rata (*mean*) sebesar 0 dan standar deviasi (*standard deviation*) sebesar 1. Dengan demikian, seluruh fitur akan berada pada skala yang sama dan tidak saling mendominasi satu sama lain. Secara sederhana, proses ini dapat diibaratkan seperti menyamakan satuan ukuran agar semua data dapat dibandingkan secara adil. Setelah dilakukan proses *scaling*, nilai data berubah menjadi berada di sekitar angka nol, dengan rentang nilai yang relatif kecil, baik dalam bentuk positif maupun negatif. Hal ini menunjukkan bahwa data telah berhasil dinormalisasi dan siap digunakan dalam proses pelatihan model. Tahap ini sangat penting terutama untuk algoritma seperti *Support Vector Machine (SVM)* dan *K-Nearest Neighbor (KNN)* yang sangat bergantung pada perhitungan jarak antar data. Tanpa proses *feature scaling*, model dapat menghasilkan prediksi yang kurang akurat karena dipengaruhi oleh perbedaan skala antar fitur.

```
Contoh data setelah scaling (training):
[[-0.33641123  1.56948792  1.53237306  1.23621965 -0.33279933  1.31341642
   1.0681126   1.19669316  2.42629341  0.72905966  1.76226907  0.69866128
   0.61259229  0.9948703   1.00235993 -1.00318364 -0.99575899 -0.70663332
   1.41203821  0.99946443  1.00021431  1.41755363 -0.70987389 -1.00261055
   1.41022653 -0.70731514  1.00666507 -0.9978238  -1.0006788   0.99167749
  -0.99618585 -0.99689767  0.99157124 -1.00057159 -0.57482091  1.73610172
  -0.58372881 -1.00117927]
 [-0.15113087  1.0464051   1.36608351  1.18608146  1.62163218  0.57945189
   0.46504309  1.71646024  0.7234186   0.05989885  1.76226907  0.33774448
  -0.1291945  -1.00515615  1.00235993  0.99682646 -0.99575899  1.41516111
  -0.70819613 -1.00053586  1.00021431 -0.70544068  1.40870092  0.99739625
   1.41022653 -0.70731514  1.00666507 -0.9978238   0.99932166  0.99167749
  -0.99618585 -0.99689767  0.99157124  0.99942873 -0.57482091  1.73610172
  -0.58372881 -1.00117927]
 [-0.05849069 -0.99837319 -0.13052245 -0.66903142  0.04910108 -1.03527007
  -0.65791393 -0.46656149  1.52477146  1.31457537 -1.18106559 -0.6934464
  -0.65082906  0.9948703  -0.99764563  0.99682646  1.00425907 -0.70663332
   1.41203821 -1.00053586 -0.99978574 -0.70544068  1.40870092 -1.00261055
   1.41022653 -0.70731514  1.00666507 -0.9978238  -1.0006788  -1.00839236
  -0.99618585 -0.99689767  0.99157124  0.99942873 -0.57482091 -0.57600312
   1.71312429 -1.00117927]
 [ 0.68263073  0.42821632  1.36608351 -0.41834049 -0.24294041 -0.30130555
  -0.20041292 -0.05074783 -0.32835701  0.47812436 -1.18106559 -0.79656549
   1.0388743   0.9948703  -0.99764563 -1.00318364  1.00425907 -0.70663332
  -0.70819613 -1.00053586 -0.99978574  1.41755363 -0.70987389 -1.00261055
  -0.70910593  1.41379698 -0.99337906 -0.9978238   0.99932166  0.99167749
   1.00382876 -0.99689767 -1.00850041  0.99942873  1.73967227 -0.57600312
  -0.58372881  0.99882212]
 [-0.89225229 -1.23613811 -0.296812   -0.51861686 -1.32124745 -1.18206298
   0.6937936  -1.19423541 -1.88097816 -1.36206786 -1.18106559 -0.89968458
  -1.07570883  0.9948703  -0.99764563  0.99682646  1.00425907  1.41516111
  -0.70819613 -1.00053586 -0.99978574  1.41755363 -0.70987389  0.99739625
  -0.70910593  1.41379698  1.00666507 -0.9978238  -1.0006788   0.99167749
   1.00382876 -0.99689767  0.99157124 -1.00057159 -0.57482091 -0.57600312
  -0.58372881  0.99882212]]
```

Gambar 2. Hasil *Scalling*

Evaluasi Model

Support Vector Machine (SVM)

Pada tahap ini dilakukan pemodelan menggunakan algoritma Support Vector Machine (SVM), yaitu metode klasifikasi yang bekerja dengan menentukan *hyperplane* optimal sebagai batas pemisah antar kelas. SVM bertujuan memaksimalkan margin antar kelas sehingga proses klasifikasi dapat dilakukan secara lebih akurat. Algoritma ini sesuai digunakan pada dataset dengan jumlah fitur yang relatif banyak dan data yang telah melalui proses *feature scaling*.

Hasil pengujian menunjukkan bahwa model SVM menghasilkan akurasi sebesar 0,7821 atau 78,21%, yang menandakan kemampuan klasifikasi yang cukup baik pada dataset multikelas. Berdasarkan *classification report*, nilai precision, recall, dan F1-score pada sebagian besar kelas berada pada tingkat yang relatif seimbang. Kelas Neonatal Diabetes Mellitus (NDM) menunjukkan performa paling tinggi dengan nilai precision 0,98, recall 0,99, dan F1-score 0,99, yang mengindikasikan bahwa model hampir selalu mengklasifikasikan kelas ini dengan benar. Kinerja yang tinggi juga ditunjukkan oleh beberapa kelas lain, seperti Type 1 Diabetes dengan nilai precision 0,84, recall 0,82, dan F1-score 0,83, Wolfram Syndrome dengan F1-score 0,81, serta Wolcott–Rallison Syndrome dengan F1-score 0,82. Nilai-nilai tersebut menunjukkan bahwa model SVM mampu mengenali pola data pada kelas-kelas tersebut secara konsisten. Sebaliknya, beberapa kelas seperti Steroid-Induced Diabetes dan Secondary Diabetes memiliki performa yang lebih rendah, dengan nilai F1-score masing-masing sebesar 0,60 dan 0,63, yang mengindikasikan adanya kesulitan model dalam membedakan karakteristik kelas-kelas tersebut dari kelas lain yang memiliki pola serupa.

Secara keseluruhan, nilai macro average F1-score sebesar 0,78 dan weighted average F1-score sebesar 0,78 menunjukkan bahwa model SVM memiliki performa yang stabil dan relatif merata dalam mengklasifikasikan seluruh kelas. Temuan ini menegaskan bahwa algoritma Support Vector Machine mampu memberikan hasil klasifikasi yang cukup baik dan konsisten dalam prediksi penyakit Diabetes Mellitus pada dataset yang digunakan dalam penelitian ini.

Akurasi SVM: 0.7821428571428571

Classification Report SVM:

	precision	recall	f1-score	support
Cystic Fibrosis-Related Diabetes (Cfrd)	0.86	0.77	0.81	1093
Gestational Diabetes	0.71	0.73	0.72	1069
Lada	0.84	0.81	0.82	1044
Mody	0.78	0.79	0.79	1111
Neonatal Diabetes Mellitus (Ndm)	0.99	1.00	1.00	1082
Prediabetic	0.80	0.84	0.82	1075
Secondary Diabetes	0.66	0.60	0.63	1096
Steroid-Induced Diabetes	0.64	0.58	0.61	1055
Type 1 Diabetes	0.82	0.88	0.85	1089
Type 2 Diabetes	0.74	0.72	0.73	1079
Type 3C Diabetes (Pancreatogenic Diabetes)	0.63	0.76	0.69	1064
Wolcott-Rallison Syndrome	0.85	0.84	0.85	1080
Wolfram Syndrome	0.85	0.85	0.85	1063
accuracy			0.78	14000
macro avg	0.78	0.78	0.78	14000
weighted avg	0.78	0.78	0.78	14000

Gambar 3. *Support Vector Machine (SVM)*

K-Nearest Neighbor (KNN)

Pada tahap ini dilakukan proses pemodelan menggunakan algoritma *K-Nearest Neighbor (KNN)*. Algoritma ini bekerja dengan cara mengklasifikasikan suatu data berdasarkan kedekatan jarak dengan data lain di sekitarnya. Dengan kata lain, suatu data akan dimasukkan

ke dalam kelas yang paling banyak muncul dari sejumlah tetangga terdekatnya. Metode ini sangat bergantung pada perhitungan jarak antar data, sehingga sangat dipengaruhi oleh proses *feature scaling* yang telah dilakukan sebelumnya.

Berdasarkan hasil pengujian, model *KNN* menghasilkan nilai akurasi sebesar 0.4723 atau sekitar 47,23%. Nilai ini tergolong cukup rendah jika dibandingkan dengan model lainnya. Hal ini menunjukkan bahwa kemampuan model dalam mengklasifikasikan data masih kurang optimal pada dataset yang digunakan. Dilihat dari hasil *classification report*, sebagian besar kelas memiliki nilai *precision*, *recall*, dan *f1-score* yang relatif rendah, yaitu berada di kisaran 0.30 hingga 0.50. Hal ini menunjukkan bahwa model mengalami kesulitan dalam membedakan antar kelas yang memiliki karakteristik yang mirip. Namun demikian, terdapat beberapa kelas yang memiliki performa cukup baik, seperti *Neonatal Diabetes Mellitus (NDM)* dengan nilai *f1-score* sebesar 0.88, serta *Wolfram Syndrome* dan *Wolcott-Rallison Syndrome* yang memiliki nilai di atas 0.60. Rendahnya performa model *KNN* pada penelitian ini kemungkinan disebabkan oleh jumlah kelas yang cukup banyak, yaitu sebanyak 13 kelas, serta kompleksitas data yang tinggi. Algoritma *KNN* cenderung kurang efektif ketika digunakan pada dataset dengan banyak kelas dan fitur, karena perhitungan jarak menjadi lebih sulit dalam membedakan masing-masing kelas secara jelas.

Secara keseluruhan, nilai *macro average* dan *weighted average* yang berada di sekitar 0.47 menunjukkan bahwa performa model *KNN* masih kurang baik dalam mengklasifikasikan seluruh kelas. Dengan demikian, dapat disimpulkan bahwa algoritma *K-Nearest Neighbor (KNN)* kurang optimal digunakan pada dataset ini dibandingkan dengan metode lainnya.

Akurasi KNN: 0.4722857142857143

Classification Report KNN:

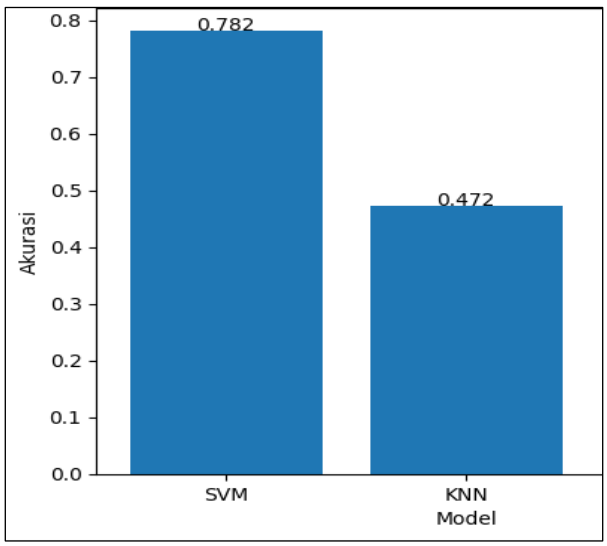
	precision	recall	f1-score	support
Cystic Fibrosis-Related Diabetes (Cfrd)	0.41	0.46	0.43	1093
Gestational Diabetes	0.29	0.35	0.32	1069
Lada	0.39	0.35	0.37	1044
Mody	0.36	0.42	0.39	1111
Neonatal Diabetes Mellitus (Ndm)	0.84	0.94	0.88	1082
Prediabetic	0.39	0.45	0.42	1075
Secondary Diabetes	0.35	0.36	0.36	1096
Steroid-Induced Diabetes	0.38	0.33	0.35	1055
Type 1 Diabetes	0.58	0.49	0.53	1089
Type 2 Diabetes	0.53	0.33	0.41	1079
Type 3C Diabetes (Pancreatogenic Diabetes)	0.38	0.42	0.40	1064
Wolcott-Rallison Syndrome	0.61	0.61	0.61	1080
Wolfram Syndrome	0.72	0.63	0.67	1063
accuracy			0.47	14000
macro avg	0.48	0.47	0.47	14000
weighted avg	0.48	0.47	0.47	14000

Gambar 4. *K-Nearest Neighbor* (KNN)

Perbandingan Model

Pada tahap ini dilakukan perbandingan performa keuda algoritma yang telah digunakan, yaitu *Support Vector Machine (SVM)*, dan *K-Nearest Neighbor (KNN)*. Perbandingan dilakukan berdasarkan nilai akurasi serta metrik evaluasi lainnya seperti *precision*, *recall*, dan *f1-score*, baik secara keseluruhan maupun per kelas. Berdasarkan hasil pengujian yang disajikan pada Gambar 5 (Perbandingan Model Akurasi), model Support Vector Machine (SVM) memperoleh akurasi tertinggi sebesar 78,21%, sedangkan K-Nearest Neighbor (KNN) menunjukkan akurasi terendah yaitu 47,23%. Hasil ini menegaskan bahwa secara umum SVM merupakan model dengan kemampuan klasifikasi terbaik pada dataset yang digunakan dalam penelitian ini.

Perbedaan performa ini berkaitan erat dengan karakteristik masing-masing algoritma. SVM mampu menemukan hyperplane pemisah yang optimal antar kelas, sehingga lebih efektif pada data berdimensi tinggi dan memiliki banyak fitur. Sementara itu, KNN sangat bergantung pada perhitungan jarak antar data, yang menjadi kurang optimal ketika jumlah fitur dan kelas meningkat.



Gambar 5. Perbandingan Model Akurasi

Jika ditinjau lebih rinci melalui tabel perbandingan kinerja per kelas, SVM menunjukkan performa yang relatif stabil pada hampir seluruh kelas. Nilai precision, recall, dan f1-score pada model ini cenderung tinggi dan konsisten, yang menandakan kemampuan SVM dalam mengklasifikasikan data lintas kelas secara seimbang. Sebaliknya, KNN memperlihatkan nilai yang lebih rendah dan fluktuatif pada sebagian besar kelas, sehingga menunjukkan keterbatasan dalam mengenali pola data secara akurat.

Analisis lebih lanjut terhadap nilai precision, recall, dan f1-score per kelas yang ditampilkan pada Gambar 6 (Perbandingan Model Precision, Recall, dan F1-Score) menunjukkan bahwa SVM memiliki kinerja unggul pada hampir seluruh kelas, seperti Cystic Fibrosis-Related Diabetes, LADA, dan Wolfram Syndrome. Nilai recall yang tinggi juga menunjukkan bahwa SVM mampu mengenali sebagian besar data aktual dengan baik, termasuk pada kelas Type 1 Diabetes dan Prediabetic. Sebaliknya, KNN menunjukkan nilai precision dan recall yang relatif rendah dan tidak stabil.

	SVM_Precision	SVM_Recall	SVM_F1	KNN_Precision	KNN_Recall	KNN_F1
Cystic Fibrosis-Related Diabetes (Cfrd)	0.855556	0.774931	0.813250	0.410049	0.462946	0.434895
Gestational Diabetes	0.714679	0.728718	0.721630	0.287250	0.349860	0.315479
Lada	0.839482	0.806513	0.822667	0.386581	0.347701	0.366112
Mody	0.784121	0.791179	0.787634	0.364770	0.421242	0.390977
Neonatal Diabetes Mellitus (Ndm)	0.994480	0.999076	0.996773	0.838576	0.936229	0.884716
Prediabetic	0.798408	0.840000	0.818676	0.392799	0.446512	0.417936
Secondary Diabetes	0.664300	0.597628	0.629203	0.354373	0.358577	0.356463
Steroid-Induced Diabetes	0.644515	0.579147	0.610085	0.381421	0.330806	0.354315
Type 1 Diabetes	0.820447	0.876951	0.847759	0.578603	0.486685	0.528678
Type 2 Diabetes	0.737689	0.721965	0.729742	0.525849	0.329935	0.405467
Type 3C Diabetes (Pancreatogenic Diabetes)	0.627527	0.758459	0.686809	0.382727	0.424812	0.402673
Wolcott-Rallison Syndrome	0.848231	0.843519	0.845868	0.610956	0.609259	0.610107
Wolfram Syndrome	0.846805	0.847601	0.847203	0.715048	0.630292	0.670000

Gambar 6. Perbandingan Model Precision, Recal, F1

Secara keseluruhan, berdasarkan hasil akurasi pada Gambar 5 serta analisis detail pada Tabel X dan Gambar 6, dapat disimpulkan bahwa Support Vector Machine (SVM) merupakan metode yang paling direkomendasikan dalam penelitian ini karena memiliki performa paling tinggi dan stabil di berbagai kelas. Sementara itu, K-Nearest Neighbor (KNN) kurang sesuai untuk dataset dengan karakteristik banyak kelas dan fitur, sebagaimana ditunjukkan oleh rendahnya nilai akurasi, precision, recall, dan f1-score.

KESIMPULAN

Berdasarkan hasil penelitian, penerapan metode machine learning menggunakan *Support Vector Machine* (SVM) dan *K-Nearest Neighbor* (KNN) terbukti dapat digunakan untuk memprediksi penyakit Diabetes Mellitus melalui tahapan preprocessing data yang mencakup pembersihan, pengodean, dan penskalaan data medis. Evaluasi kinerja menggunakan metrik accuracy, *precision*, *recall*, dan *F1-score* menunjukkan bahwa masing-masing algoritma memiliki karakteristik performa yang berbeda dalam menangani klasifikasi multikelas. Hasil analisis juga menunjukkan bahwa kinerja model sangat dipengaruhi oleh karakteristik data,

jumlah fitur, serta teknik *preprocessing* yang diterapkan, sehingga pemilihan metode klasifikasi perlu disesuaikan dengan kebutuhan dan kondisi data agar menghasilkan prediksi yang lebih akurat dan relevan.

REKOMENDASI

Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa saran yang diharapkan dapat menjadi bahan pertimbangan untuk pengembangan penelitian selanjutnya:

- Penelitian selanjutnya disarankan untuk menambahkan metode klasifikasi lain seperti *Random Forest*, *Logistic Regression*, *XGBoost*, atau *Neural Network* agar diperoleh perbandingan yang lebih luas dan komprehensif. Dengan menambahkan variasi algoritma, diharapkan dapat ditemukan model dengan performa yang lebih optimal dalam prediksi penyakit *Diabetes Mellitus*.
- Disarankan untuk melakukan proses *hyperparameter tuning* pada setiap algoritma yang digunakan, seperti penentuan kernel pada *Support Vector Machine (SVM)*, dan pemilihan nilai *k* pada *K-Nearest Neighbor (KNN)*. Proses ini bertujuan untuk meningkatkan performa model sehingga hasil prediksi menjadi lebih akurat dan optimal.
- Penelitian selanjutnya juga disarankan untuk menggunakan dataset yang lebih beragam serta berasal dari data nyata seperti rumah sakit atau institusi kesehatan, serta mengembangkan hasil penelitian ke dalam bentuk sistem atau aplikasi. Hal ini bertujuan agar penelitian tidak hanya bersifat akademis, tetapi juga dapat diimplementasikan secara nyata sebagai sistem pendukung keputusan dalam bidang kesehatan

REFERENSI

Ahmed, S., & others. (2022). Probabilistic prediction of diabetes using logistic regression. *Journal of Medical Systems*. https://doi.org/10.1007/s10916-022-01801-7

Ali, A., & others. (2020). Comparative study of machine learning algorithms for diabetes prediction. *IEEE Access*, *8*, 219990–220004. https://doi.org/10.1109/ACCESS.2020.3044276

Dwivedi, S. (2020). Comparative analysis of classification algorithms for diabetes prediction. *International Journal of Data Mining and Knowledge Management Process*. https://doi.org/10.5121/ijdkp.2020.10301

Fahrezi, I., Taufiq, M., Akhwani, A., & Nafia'ah, N. (2020). Meta-Analisis Pengaruh Model Pembelajaran Project Based Learning Terhadap Hasil Belajar Siswa Pada Mata Pelajaran IPA Sekolah Dasar. *Jurnal Ilmiah Pendidikan Profesi Guru*, *3*(3), 408. https://doi.org/10.23887/jippg.v3i3.28081

Haq, A. U., & others. (2020). Performance analysis of KNN for diabetes prediction. *Computers in Biology and Medicine*. https://doi.org/10.1016/j.compbiomed.2020.103638

Jain, R., & Singh, P. (2021). Predictive analysis of diabetes using machine learning. *International Journal of Information Technology*. https://doi.org/10.1007/s41870-021-00645-9

Kavakiotis, I., & others. (2020). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*, *15*, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005

Kavitha, R. J., & Kannan, V. R. (2022). Diabetes prediction using machine learning classifiers. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-021-03442-6

Kumar, P., & Singh, R. K. (2021). Performance comparison of SVM, KNN, and logistic regression for diabetes prediction. *International Journal of Advanced Computer Science and Applications*. https://doi.org/10.14569/IJACSA.2021.0120312

Patil, S. R., & others. (2020). K-nearest neighbor based diabetes prediction. *International Journal of Computer Applications*. https://doi.org/10.5120/ijca2020920545

Perveen, S., Shahbaz, M., Guergachi, A., & Keshavjee, K. (2020). Performance analysis of data mining classification techniques to predict diabetes. *Procedia Computer Science*, *82*, 115–121. https://doi.org/10.1016/j.procs.2016.04.016

Rahman, M., & others. (2021). Diabetes prediction using K-nearest neighbor algorithm. *SN Computer Science*. https://doi.org/10.1007/s42979-021-00538-8

Shouman, M., & others. (2020). Using machine learning techniques for diabetes prediction. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2019.112949

Singh, A. K., & others. (2020). Machine learning-based diabetes prediction: A review. *Journal of Healthcare Engineering*. https://doi.org/10.1155/2020/8892046

Sisodia, M., & Sisodia, D. S. (2020). Prediction of diabetes using classification algorithms. *Procedia Computer Science*, *132*, 1578–1585. https://doi.org/10.1016/j.procs.2018.05.122

Sneha, S., & Gangil, N. (2020). Analysis of diabetes mellitus for early prediction using optimal feature selection. *Journal of Big Data*, *6*(13). https://doi.org/10.1186/s40537-019-0266-6

Swapna, H., Soman, S., & Vinayakumar, R. (2021). Automated detection of diabetes using support vector machine. *Biomedical Signal Processing and Control*. https://doi.org/10.1016/j.bspc.2021.102432

World Health Organization. (2021). *Global Report on Diabetes*. WHO.

Zheng, Y., & others. (2020). Diabetes prediction using support vector machine. *Applied Soft Computing*, *93*. https://doi.org/10.1016/j.asoc.2020.106383