Jurnal Teknik Informatika dan Teknologi Informasi
Volume. 5 Nomor. 3 Desember 2025
E-ISSN .: 2827-9387, P-ISSN .: 2827-9379, Hal 113-130
DOI: https://doi.org/10.55606/jutiti.v5i3.6213
Tersedia: https://journalshub.org/index.php/jutiti

Analisis Perbandingan Algoritma Random *Forest*, SVM, dan *Logistic Regression* untuk Menentukan Model Terbaik Prediksi Penyakit Diabetes

Alghifar Firgiawan[1*], Fauzan Nawwir Andriansyah[2], Raihan Naufal Ramadhan [3], Sumanto[4], Imam Budiawan[5], Roida Pakpahan[6]
[1-6]Prodi Informatika, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika, Indonesia
**Penulis korespondensi: alghifarfirgiawan258@gmail.com*

Abstract. *Diabetes is a chronic metabolic disorder characterized by elevated blood glucose levels caused by the body's inability to produce or effectively respond to insulin. The increasing prevalence of diabetes in Indonesia requires accurate data-driven early detection systems to assist the diagnostic process. This study aims to compare the performance of three machine learning algorithms—Support Vector Machine (SVM), Random Forest, and Logistic Regression—in predicting diabetes disease based on patient clinical data. The dataset used was obtained from the Kaggle repository titled 100,000 Diabetes Clinical Dataset. The research process was conducted using the Orange Data Mining software through several stages, including data preprocessing, One-Hot Encoding transformation, model training, and evaluation using the 10-Fold Cross Validation method. The results show that the Random Forest algorithm achieved the best performance with an accuracy of 97.1%, followed by Logistic Regression at 96.0% and SVM at 92.3%. These findings indicate that ensemble-based methods such as Random Forest outperform others in producing stable and accurate predictions for diabetes diagnosis*

Keywords: *Data Mining; Machine Learning; Random Forest; Support Vector Machine; Logistic Regression.*

Abstrak. Penyakit diabetes merupakan gangguan metabolik kronis yang ditandai oleh meningkatnya kadar glukosa darah akibat ketidakmampuan tubuh dalam memproduksi atau merespons insulin secara efektif. Peningkatan jumlah penderita diabetes di Indonesia menuntut adanya sistem deteksi dini berbasis data yang akurat untuk membantu proses diagnosis. Penelitian ini bertujuan untuk membandingkan performa tiga algoritma machine learning yaitu Support Vector Machine (SVM), Random Forest, dan Logistic Regression dalam memprediksi penyakit diabetes berdasarkan data klinis pasien. Dataset yang digunakan berasal dari repositori Kaggle berjudul 100,000 Diabetes Clinical Dataset. Proses penelitian dilakukan menggunakan perangkat lunak Orange Data Mining, melalui tahapan preprocessing, transformasi data dengan One-Hot Encoding, pelatihan model, serta evaluasi menggunakan metode 10-Fold Cross Validation. Hasil penelitian menunjukkan bahwa algoritma Random Forest memiliki performa terbaik dengan akurasi 97,1%, diikuti oleh Logistic Regression sebesar 96,0%, dan SVM sebesar 92,3%. Temuan ini menegaskan bahwa metode ensemble learning seperti Random Forest lebih unggul dalam menghasilkan prediksi yang stabil dan akurat terhadap diagnosis penyakit diabetes.

Kata kunci: Penambangan Data; Pembelajaran Mesin; Hutan Acak; Mendukung mesin vektor; Regresi Logistik.

1. LATAR BELAKANG

Diabetes merupakan penyakit metabolik kronis yang terjadi akibat gangguan pada proses pengaturan kadar gula darah di dalam tubuh. Kondisi ini muncul karena produksi insulin yang tidak mencukupi atau karena tubuh tidak mampu merespons insulin secara efektif. Akibatnya, kadar glukosa dalam darah meningkat melebihi batas normal dan dapat menimbulkan berbagai komplikasi serius pada organ tubuh. Beberapa dampak yang umum terjadi antara lain gangguan fungsi ginjal, kerusakan saraf, penurunan penglihatan, serta risiko amputasi pada anggota tubuh. Selain itu, penderita diabetes memiliki kemungkinan dua hingga tiga kali lebih tinggi mengalami penyakit jantung dan stroke dibandingkan individu yang sehat. Pada ibu hamil,

Naskah Masuk: 30 September 2025; Revisi: 18 Oktober 2025; Diterima: 30 Oktober 2025;
Tersedia: 19 November 2025

diabetes yang tidak terkendali juga dapat meningkatkan risiko kematian janin serta menyebabkan berbagai komplikasi kehamilan lainnya (Syamsudin dkk., t.t.)

Berdasarkan data *International Diabetes Federation* (IDF) tahun 2024, Indonesia termasuk dalam kawasan Pasifik Barat yang terdiri dari 38 negara. Di tingkat global, terdapat sekitar 589 juta penderita diabetes, sementara di kawasan Pasifik Barat jumlahnya mencapai 215 juta orang dan diproyeksikan meningkat menjadi 254 juta pada tahun 2050. Di Indonesia sendiri, dari total 185,2 juta penduduk dewasa, prevalensi diabetes mencapai 11,3 persen atau sekitar 20,4 juta kasus pada orang dewasa(International Diabetes Federation, 2024). Data ini menunjukkan bahwa Indonesia menghadapi beban penyakit diabetes yang cukup tinggi dan berpotensi meningkat di masa mendatang jika tidak dilakukan upaya penanganan yang optimal.

Selain faktor genetik, gaya hidup modern juga berperan besar dalam meningkatnya prevalensi diabetes di Indonesia. Pola makan tinggi karbohidrat sederhana, konsumsi makanan cepat saji, kurangnya aktivitas fisik, serta meningkatnya tingkat stres menjadi faktor utama yang mempercepat terjadinya resistensi insulin. Menurut penelitian (Dita Ayuningtiyas Tuti dkk., 2023), individu dengan kebiasaan sedentari dan pola makan tidak seimbang memiliki risiko dua kali lipat lebih tinggi mengalami diabetes tipe 2 dibandingkan mereka yang menjalani gaya hidup aktif. Sementara itu, hasil survei kesehatan nasional yang dikaji oleh (“Exploring the Non-Communicable Disease Burden in Indonesia – Findings from the 2023 Health Survey,” 2025) menunjukkan bahwa peningkatan kasus diabetes di Indonesia juga berkorelasi dengan pertambahan usia, obesitas, dan kebiasaan merokok. Fenomena ini menunjukkan bahwa pencegahan diabetes tidak hanya bergantung pada faktor medis, tetapi juga pada perubahan perilaku masyarakat secara menyeluruh. Oleh karena itu, pendekatan berbasis data melalui penerapan teknologi seperti *machine learning* sangat dibutuhkan untuk mendukung upaya deteksi dini dan pengendalian penyakit diabetes secara lebih efektif dan berkelanjutan.

2. KAJIAN TEORITIS

Machine learning merupakan cabang dari kecerdasan buatan yang berfokus pada kemampuan sistem komputer untuk belajar dari data dan meningkatkan kinerjanya tanpa perlu diprogram secara eksplisit. Menurut (Yusoff, 2024), machine learning bekerja dengan menganalisis pola, hubungan, serta tren dari sejumlah besar data melalui proses pembelajaran berulang (iterative learning process), sehingga model dapat menyesuaikan diri terhadap informasi baru dan menghasilkan prediksi yang lebih akurat. Pendekatan ini memungkinkan sistem untuk memperoleh pengetahuan secara otomatis dari pengalaman (experience-based

learning) tanpa intervensi manusia secara langsung. Oleh karena itu, machine learning menjadi teknologi penting dalam berbagai bidang, termasuk kesehatan, karena kemampuannya mengubah data kompleks menjadi informasi yang bermanfaat bagi pengambilan keputusan dan prediksi berbasis data.

Klasifikasi merupakan salah satu metode utama dalam *machine learning* yang berfungsi untuk mengelompokkan data ke dalam kelas tertentu berdasarkan pola yang ditemukan dari data pelatihan. Menurut (Lu dkk., 2022), proses klasifikasi dilakukan dengan membangun model prediktif yang mempelajari hubungan antara variabel input dan label output dari dataset berlabel (*supervised learning*). Model tersebut kemudian digunakan untuk menentukan kategori data baru yang belum pernah dianalisis sebelumnya. Proses ini biasanya melibatkan dua tahap, yaitu pelatihan model (*training phase*) dan pengujian model (*testing phase*) guna mengukur kemampuan prediksi dan generalisasi. Teknik klasifikasi banyak diterapkan di berbagai bidang, termasuk bidang medis, karena mampu mengidentifikasi pola tersembunyi dari data pasien dan membantu dalam pengambilan keputusan berbasis data, seperti mendeteksi kemungkinan penyakit atau menilai tingkat risikonya secara otomatis.

Support Vector Machine (SVM) merupakan algoritma *machine learning* yang digunakan untuk menyelesaikan permasalahan klasifikasi maupun regresi dengan cara mencari *hyperplane* terbaik yang memisahkan dua atau lebih kelas data. Menurut (Syahputra & Wibowo, 2023), prinsip utama SVM adalah memaksimalkan jarak atau *margin* antara kelas yang berbeda agar model memiliki kemampuan generalisasi yang baik terhadap data baru. SVM bekerja dengan memilih sejumlah kecil titik data penting yang disebut *support vectors* sebagai penentu posisi *hyperplane*. Untuk menangani data nonlinier, SVM menggunakan fungsi *kernel* seperti *radial basis function* (RBF) yang memetakan data ke ruang berdimensi lebih tinggi agar dapat dipisahkan dengan lebih optimal. Pendekatan ini membuat SVM sangat efektif dalam menangani data berdimensi tinggi seperti data medis atau citra digital.

Random Forest merupakan algoritma *ensemble learning* yang menggabungkan hasil dari banyak *decision tree* untuk meningkatkan stabilitas dan akurasi model prediksi. Menurut (Fadli Kurniawan & Ayu Megawaty, 2025) algoritma ini bekerja dengan membangun sejumlah pohon keputusan berdasarkan sampel data acak (*bootstrap sampling*) dan subset fitur yang dipilih secara acak pada setiap percabangan. Hasil dari seluruh pohon tersebut kemudian digabungkan menggunakan metode *majority voting* untuk klasifikasi atau *averaging* untuk regresi. Keunggulan utama Random Forest adalah kemampuannya dalam mengurangi

risiko *overfitting* serta tetap memberikan performa yang baik meskipun terdapat data yang bising atau hilang. Karena sifatnya yang tangguh dan fleksibel, algoritma ini banyak digunakan untuk analisis medis seperti prediksi penyakit diabetes atau kanker.

Logistic Regression adalah metode statistik yang digunakan dalam *machine learning* untuk memprediksi probabilitas suatu kejadian biner berdasarkan satu atau lebih variabel independen. Berdasarkan penelitian (Rahaman, 2024), algoritma ini bekerja dengan menggunakan fungsi logit atau *sigmoid function* untuk mengubah nilai input menjadi rentang antara 0 dan 1, yang kemudian diinterpretasikan sebagai peluang kejadian suatu kelas. Meskipun tergolong sederhana, *Logistic Regression* efektif dalam menghasilkan model yang mudah diinterpretasikan serta efisien dalam komputasi, terutama ketika hubungan antarvariabel bersifat linier. Metode ini sering dijadikan baseline model dalam klasifikasi medis, seperti memprediksi apakah seorang pasien berisiko atau tidak terhadap penyakit tertentu

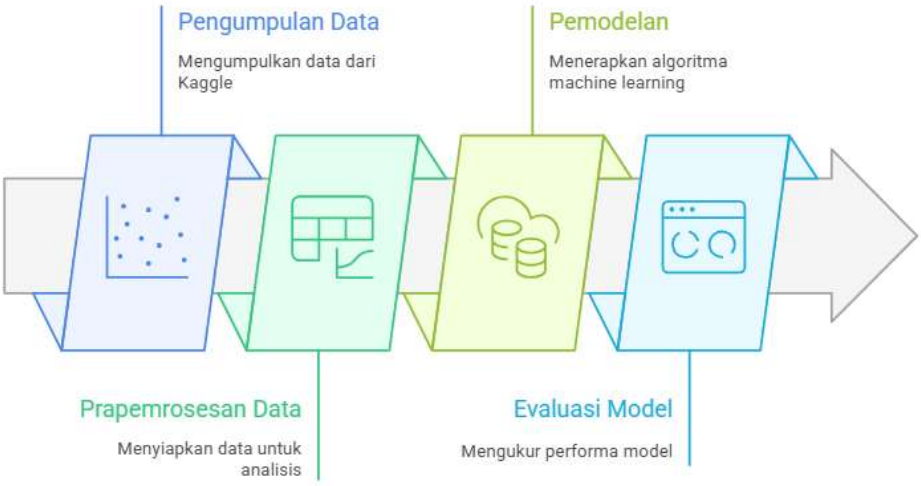
Dalam menghadapi permasalahan tersebut, penerapan metode *machine learning* banyak digunakan untuk membantu proses diagnosis dan prediksi penyakit diabetes. Berbagai algoritma telah diuji untuk menemukan model yang paling akurat. Penelitian oleh (Sanhaji dkk., t.t.).mengembangkan aplikasi *DIATECT* berbasis web menggunakan algoritma Support Vector Machine (SVM) dan menghasilkan akurasi 77%, dengan *precision* 77% serta *recall* dan *F1-score* mencapai 91%. Hal ini menunjukkan bahwa SVM cukup efektif untuk klasifikasi data pasien diabetes, meskipun kinerjanya masih dapat ditingkatkan dengan algoritma lain.

Penelitian oleh(Teknika & Ria Supriyatna, t.t.)menunjukkan bahwa algoritma Random Forest memiliki hasil yang lebih unggul, dengan akurasi 99,3%, *precision* 99,1%, *recall* 99,5%, dan nilai *AUC* 100%. Algoritma ini dinilai stabil, akurat, serta mampu mengurangi kesalahan klasifikasi. Sementara itu, penelitian (Khairunnisa, t.t.) menggunakan Regresi Logistik dengan metode *k-fold cross validation* dan memperoleh akurasi 93,7%. Model ini mudah dipahami dan efisien, namun performanya masih di bawah metode *ensemble* seperti Random Forest. Berdasarkan ketiga penelitian tersebut, Random Forest menjadi model paling optimal untuk prediksi diabetes karena tingkat akurasi dan kestabilannya yang tinggi. Oleh karena itu, peneliti tertarik untuk melakukan penelitian lanjutan dengan menggunakan metode Random Forest, serta membandingkannya dengan algoritma SVM dan Regresi Logistik, guna mengetahui performa terbaik dalam memprediksi penyakit diabetes secara lebih akurat dan efisien

3. METODE PENELITIAN

Penelitian ini dilakukan untuk membandingkan performa tiga algoritma *machine learning*, yaitu Support Vector Machine (SVM), Random Forest, dan Regresi Logistik, dalam memprediksi penyakit diabetes berdasarkan data klinis pasien. Metode penelitian ini menggunakan pendekatan kuantitatif dengan tahapan sistematis mulai dari pengumpulan data hingga evaluasi model.

Secara umum, tahapan penelitian ditunjukkan pada (Gambar 1) berikut:



Gambar 1. Diagram Alur Penelitian.

Tahapan Penelitian

Secara umum, tahapan penelitian ini terdiri dari empat langkah utama, yaitu pengumpulan data, prapemrosesan data, pemodelan, dan evaluasi model. Alur tahapan penelitian divisualisasikan pada (Gambar1).Tahapan tersebut dijelaskan sebagai berikut:

Pengumpulan Data: Mengambil dataset yang relevan dari sumber terpercaya, yaitu repositori daring *Kaggle*, dengan judul *"100,000 Diabetes Clinical Dataset."* Dataset berisi informasi klinis pasien seperti usia, jenis kelamin, BMI, tekanan darah, kadar glukosa, dan HbA1c (Priyam Choksi, 2023)

Prapemrosesan Data: Melakukan pembersihan data dari nilai kosong, transformasi variabel kategorikal ke numerik, dan normalisasi nilai agar berada dalam skala yang sama.

Pemodelan: Menerapkan tiga algoritma *machine learning* (SVM, Random Forest, dan Regresi Logistik) menggunakan perangkat lunak Orange Data Mining.

Evaluasi Model: Mengukur performa model menggunakan metode 10-Fold Cross Validation dengan metrik evaluasi *Accuracy, Precision, Recall,* dan *F1-Score.*

Pengumpulan Data

Tahap pengumpulan data merupakan langkah awal dalam proses penelitian ini. Data yang digunakan adalah data sekunder yang diperoleh dari repositori daring Kaggle dengan judul *"100,000 Diabetes Clinical Dataset."* Dataset ini dipilih karena memiliki ukuran yang besar dan memuat variabel-variabel yang relevan untuk analisis prediksi penyakit diabetes (Priyam Choksi, 2023).

Dataset terdiri dari 100.000 data pasien dengan atribut dapat dilihat pada Tabel 1:

Tabel 1. Deskripsi Atribut Dataset Diabetes.

Attribut	Tipe Data
Age	Numerik
Gender	Kategorikal
BMI (Body Mass Index)	Numerik
Blood Pressure	Numerik
Glucose Level	Numerik
HbA1C Level	Numerik
Smoking History	Kategorikal
Diabetes	Biner

Prapemoresesan Data

Prapemrosesan data merupakan tahapan penting untuk menyiapkan dataset agar dapat diolah secara optimal oleh model *machine learning*. Data mentah umumnya masih mengandung nilai kosong, duplikasi, format tidak seragam, atau rentang skala yang berbeda-beda. Oleh karena itu dilakukan langkah-langkah sistematis untuk meningkatkan kualitas data agar hasil pemodelan menjadi akurat dan representatif(Siswoyo & Iqbal Nurhafidz, t.t.).

Pemodelan

Tahap pemodelan merupakan proses utama dalam penelitian ini. Tiga algoritma *machine learning* diterapkan untuk membandingkan hasil klasifikasi, yaitu Support Vector Machine (SVM), Random Forest, dan Regresi Logistik (Logistic Regression).Pemodelan dilakukan di Orange.

Evaluasi Model

Setelah model selesai dibangun, tahap selanjutnya adalah mengevaluasi performa untuk menentukan model yang paling optimal. Evaluasi dilakukan dengan metode 10-Fold Cross Validation, yang membagi dataset menjadi sepuluh bagian. Sembilan bagian digunakan untuk pelatihan dan satu bagian untuk pengujian secara bergantian, sehingga seluruh data berperan

sebagai data uji (Citra Mawani dkk., 2023). Metrik yang digunakan untuk mengukur performa model adalah sebagai berikut:

Accuracy

Akurasi merupakan metrik dasar yang menunjukkan seberapa sering model memberikan prediksi yang benar. Semakin tinggi nilai akurasi, semakin besar proporsi prediksi yang sesuai dengan kondisi sebenarnya. Menurut (Yanti dkk., t.t.)rumus *Accuracy* sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \; X \; 100\%$$

Keterangan :
TP (True Positif) : Jumlah data positif yang berhasil diklasifikasikan benar sebagai positif.
TN (True Negatif) : Jumlah data negatif yang berhasil diklasifikasikan benar sebagai negatif .
FP (Falses Positif) : Jumlah data positif yang salah diklasifikasikan sebagai positif.
FN (False Negatif) : Jumlah data negatif yang salah diklasifikasikan sebagai negatif.

Precision

Precision digunakan untuk mengukur seberapa akurat model dalam memprediksi kelas positif. Metrik ini menjadi penting terutama ketika kesalahan prediksi positif (*false positive*) harus diminimalkan misalnya dalam diagnosis penyakit, di mana salah memprediksi pasien sehat sebagai sakit perlu dihindari. Menurut (Akhsani dkk., t.t.)rumus *Precision* sebagai berikut:

$$Precision = \frac{TP}{TP + FP} \; X \; 100\%$$

Keterangan :
TP (True Positif) : Jumlah data positif yang berhasil diklasifikasikan benar sebagai positif.
FP (Falses Positif) : Jumlah data positif yang salah diklasifikasikan sebagai positif.

Recall

Recall atau sensitivitas mengukur kemampuan model dalam mengenali semua data yang sebenarnya positif. Metrik ini penting dalam kasus di mana *false negative* (data positif yang tidak terdeteksi) berakibat fatal misalnya dalam deteksi penyakit. Menurut rumus (Hakim dkk., 2025) *Recall* sebagai berikut

$$Recall = \frac{TP}{TP + FN} \; X \; 100\%$$

Keterangan :
TP (True Positif) : Jumlah data positif yang berhasil diklasifikasikan benar sebagai positif.
FN (False Negatif) : Jumlah data negatif yang salah diklasifikasikan sebagai negatif.

F1-Score

F1-Score adalah rata-rata harmonis antara presisi dan recall, yang digunakan untuk menilai performa model secara seimbang. Metrik ini penting ketika data memiliki distribusi kelas yang tidak seimbang, karena F1-Score memberikan penilaian yang adil dengan

mempertimbangkan dua aspek ketepatan dan kelengkapan. Menurut rumus (Fadlianda dkk., t.t.)*F1-Score* sebagai berikut:

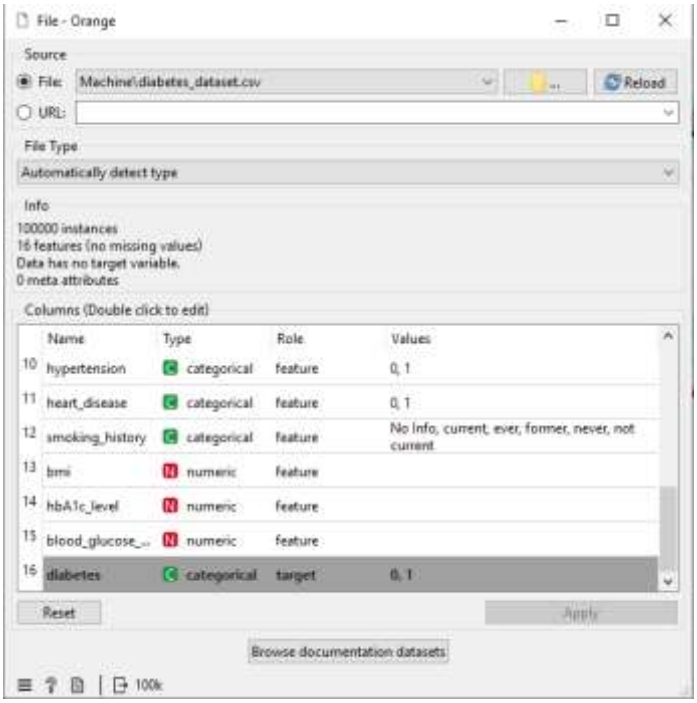
$$F1 - Score = 2X \frac{Precision\ X\ Recall}{Precision + Recall}\ X\ 100\%$$

4. HASIL DAN PEMBAHASAN

Import Dataset

Tahap awal adalah memasukkan dataset ke dalam *workspace* Orange menggunakan File Widget. Dataset yang digunakan adalah *100,000 Diabetes Clinical Dataset* dari Kaggle yang berisi 100.000 data pasien dengan berbagai atribut klinis dan demografis. Proses ini memastikan setiap variabel dikenali dengan benar oleh sistem, terutama untuk menentukan atribut mana yang akan dijadikan target klasifikasi, yaitu "Diabetes".

Selain itu, dilakukan verifikasi tipe data (numerik dan kategorikal), pengecekan jumlah data, serta pemeriksaan apakah terdapat *missing values*. Dengan langkah ini, dataset dapat dipastikan bersih dan siap diproses lebih lanjut.



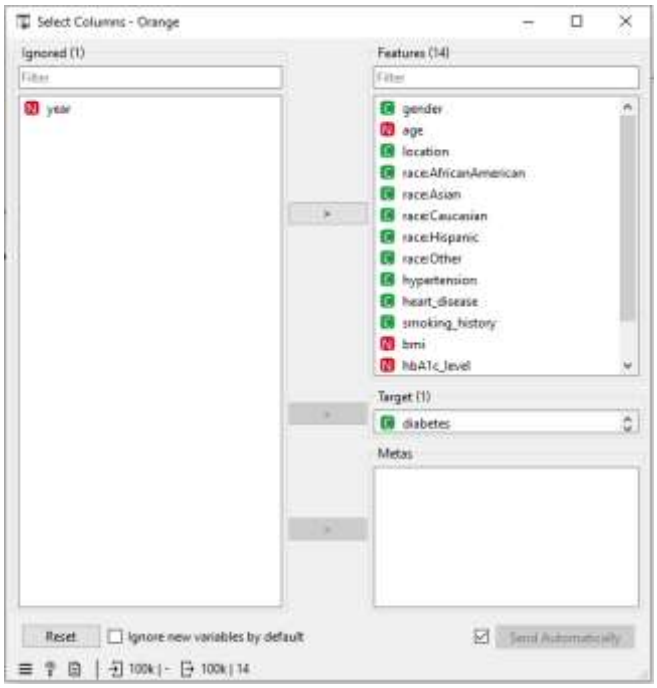
Gambar 3. Tampilan Dataset yang Diimpor pada Orange Data Mining.

Pembersihan Data dan Seleksi Atribut (Preprocessing)

Tahap selanjutnya adalah melakukan pembersihan data serta pemilihan atribut yang relevan. Dalam tahap ini digunakan Select Columns Widget untuk memilih fitur-fitur utama yang memiliki keterkaitan langsung terhadap risiko diabetes, seperti *age, BMI, glucose level, blood pressure,* dan *HbA1c level*.

Atribut yang tidak relevan seperti year dihapus karena tidak mengandung informasi klinis yang bermakna serta berpotensi mengganggu akurasi model. Selain itu, dilakukan pemeriksaan terhadap nilai kosong atau ekstrem yang dapat mempengaruhi proses pelatihan model.

Hasil dari proses ini adalah dataset yang telah terfilter dan hanya berisi atribut penting yang berkontribusi terhadap diagnosis diabetes.



Gambar 4. Proses Pemelihan Atribut Menggunakan Select Columns Widget.

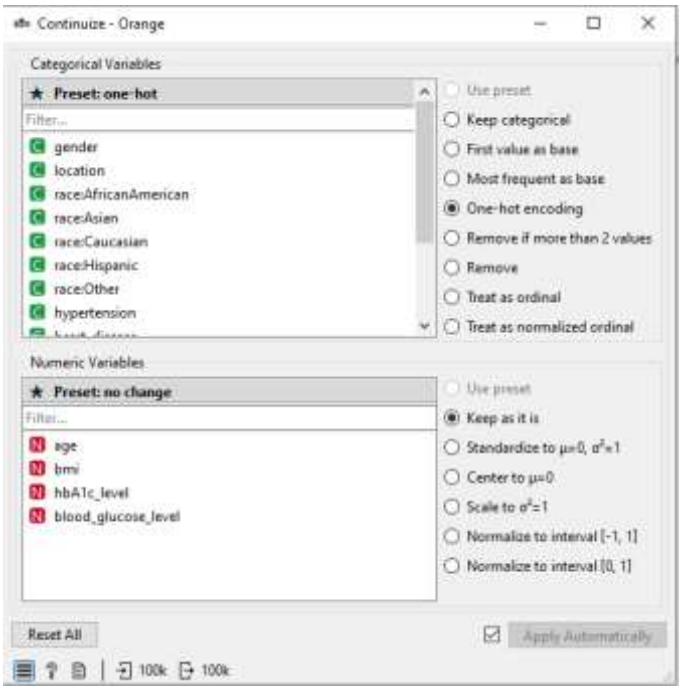
Transformasi Data Kategorikal

Tahapan selanjutnya adalah melakukan transformasi data kategorikal menjadi bentuk numerik agar dapat diproses oleh algoritma *machine learning*. Pada penelitian ini digunakan Continuize Widget untuk melakukan proses *encoding* terhadap atribut kategorikal seperti *gender* dan *smoking history*.

Metode yang diterapkan adalah One-Hot Encoding, yaitu teknik yang mengubah setiap nilai kategori menjadi kolom biner (0 dan 1).Dengan metode ini, setiap kategori akan direpresentasikan secara terpisah sehingga model dapat membedakan setiap nilai dengan lebih akurat. Misalnya, atribut *smoking history* yang berisi nilai "never", "current", dan "former" akan diubah menjadi tiga kolom biner yang masing-masing menunjukkan keberadaan atau ketiadaan kondisi tersebut.

Selain itu, untuk atribut numerik seperti *age*, *BMI*, *glucose level*, *blood pressure*, dan *HbA1c level*, pengaturan "Keep as it is" dipertahankan agar nilai numeriknya tidak berubah. Pendekatan ini penting untuk menjaga keaslian data kuantitatif, karena setiap angka

memiliki arti medis yang spesifik dan tidak boleh diubah skalanya. Secara keseluruhan, tahap transformasi ini memastikan bahwa seluruh variabel dalam dataset berada dalam format numerik yang konsisten dan kompatibel dengan algoritma klasifikasi yang digunakan.



Gambar 5. Transformasi Data kategorikal Menggunakan Continuize Widget dengan Metode One-hote encoding dan Pengaturan Keep as it is.

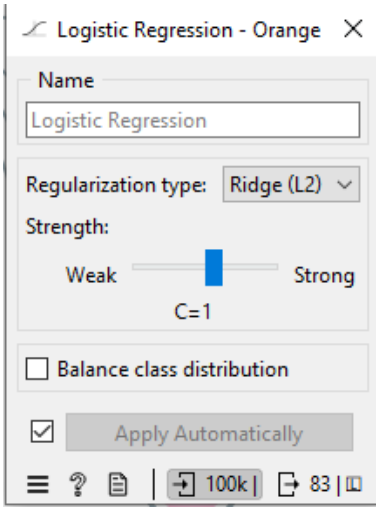
Pelatihan Dan Pengujian Model

Tahapan ini bertujuan untuk melakukan konfigurasi parameter (parameter tuning) terhadap masing-masing algoritma *machine learning* yang digunakan. Pengaturan dilakukan secara langsung melalui *widget* pada Orange Data Mining agar setiap model dapat beradaptasi dengan karakteristik data dan menghasilkan performa terbaik. Penyesuaian parameter ini penting karena secara langsung memengaruhi kemampuan model dalam mengenali pola, menyeimbangkan bias dan variansi, serta mengoptimalkan proses klasifikasi. Berikut ini merupakan pengaturan parameter untuk masing-masing model yang digunakan dalam penelitian:

Pengaturan Logisctic Regression

Algoritma *Logistic Regression* digunakan sebagai model linier untuk klasifikasi biner antara kelas *diabetes* dan *non-diabetes*. Pada penelitian ini, digunakan metode regularisasi Ridge (L2) dengan strength C = 1, yang berfungsi untuk mengurangi risiko *overfitting* dengan cara menyeimbangkan besar kecilnya koefisien regresi. Pendekatan ini menjadikan model lebih stabil, terutama saat berhadapan dengan atribut yang

saling berkorelasi seperti *BMI*, *Glucose Level*, dan *HbA1c Level*. Pengaturan ini dipilih karena mampu memberikan hasil yang konsisten tanpa mengubah struktur dasar model.



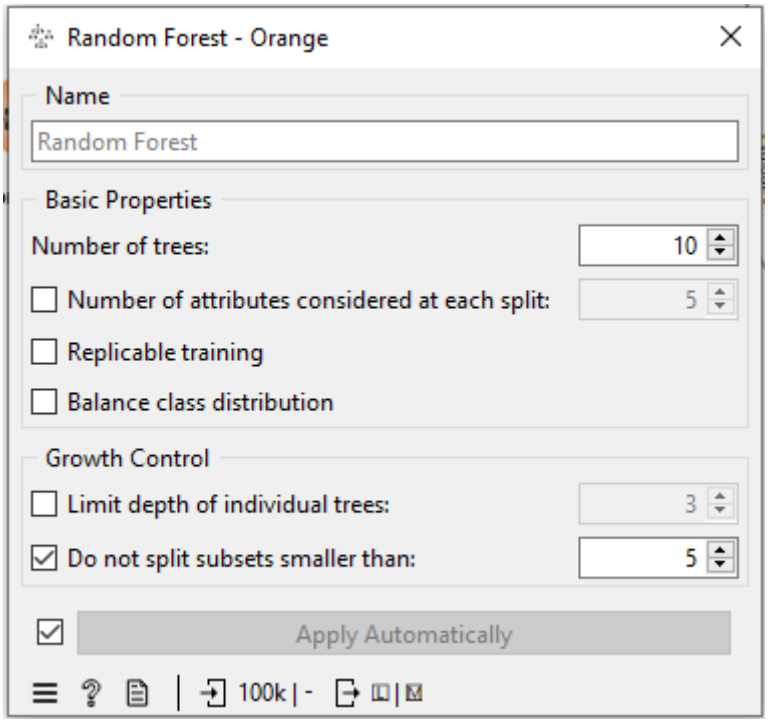
Gambar 6. Pengaturan Parameter *Logistic Regression* pada *Orange Data Mining.*

Pengaturan Random Forest

Untuk algoritma *Random Forest*, pengaturan yang digunakan adalah konfigurasi default bawaan *Orange Data Mining*, dengan beberapa parameter penting *Number of trees* 10,dan *Do not split subsets smaller than* 5.

Pengaturan ini membuat Random Forest lebih efisien dalam proses pelatihan tanpa mengorbankan stabilitas prediksi. Setiap pohon keputusan dibangun berdasarkan subset data dan fitur yang berbeda (*bootstrap sampling*), lalu hasil akhirnya digabungkan untuk menghasilkan keputusan kolektif yang lebih akurat.

Pendekatan *ensemble learning* ini sangat sesuai untuk dataset medis seperti diabetes, karena mampu mengurangi risiko kesalahan akibat variasi data dan menghasilkan model yang lebih kuat. Untuk contoh tampilan pengaturan model Random Forest dapat dilihat pada Gambar 7.



Gambar 7. Pengaturan Parameter *Random Forest* pada *Orange Data Mining.*

Pengaturan *Support Vector Machine (SVM)*

Pada algoritma *Support Vector Machine* (SVM), dilakukan sedikit penyesuaian terhadap beberapa parameter utama untuk meningkatkan performa model. Pengaturan ini dapat dilihat pada (Gambar 8).

Gambar 8. Pengaturan Parameter *Support Vector Machine (SVM)* pada Orange Data Mining.

Evaluasi dan Visualisasi Model

Setelah seluruh model selesai dikonfigurasi tahap berikutnya adalah melakukan evaluasi performa model untuk menilai seberapa baik algoritma mampu melakukan klasifikasi terhadap data pasien diabetes. Proses pengujian dilakukan menggunakan metode

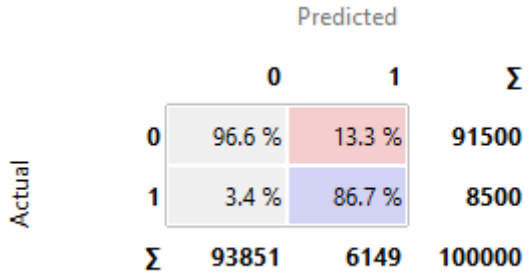
10-Fold Cross Validation pada *Widget Test & Score* di *Orange Data Mining*. Metode ini membagi dataset menjadi sepuluh bagian (*fold*), kemudian secara bergantian Sembilan bagian digunakan untuk training dan satu bagian digunakan untuk testing. Proses ini diulang hingga semua bagian data digunakan sebagai data uji setidaknya satu kali,sehingga hasil evaluasi menjadi lebih adil dan representatif.

Tabel 2. Hasil Evaluasi Ketiga Model Menggunakan Test & Score pada Orange Data Mining.

Model	AUC	CA	F1	Precision	Recall	MCC
Logistcic Regression	0.962	0.960	0.957	0.958	0.960	0.718
Random Forest	0.939	0.971	0.969	0.971	0.971	0.800
SVM	**0.893**	**0.923**	**0.926**	**0.923**	**0.923**	**0.527**

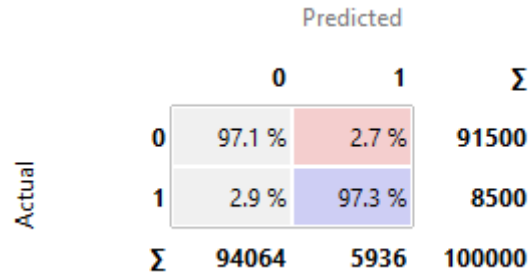
Berdasarkan hasil tersebut, algoritma Random Forest menunjukkan performa terbaik dengan akurasi tertinggi yaitu 97,1%, diikuti oleh Logistic Regression dengan 96,0%, dan SVM dengan 92,3%. Nilai *F1-Score* Random Forest yang mencapai 0,969 juga menunjukkan keseimbangan optimal antara ketepatan dan kelengkapan model dalam mengenali pasien dengan diabetes. Sedangkan Logistic Regression tetap menunjukkan performa yang stabil dan efisien, meskipun tidak seakurat Random Forest. Adapun SVM memberikan hasil yang cukup baik namun kurang maksimal pada data non-linier yang kompleks.

Pada algoritma Logistic Regression, model mampu mengklasifikasikan 96,6% data *non-diabetes* dengan benar, dan 86,7% data diabetes secara tepat. Hal ini menunjukkan bahwa Logistic Regression memiliki kemampuan yang sangat baik dalam mengenali pasien sehat, meskipun sensitivitasnya terhadap pasien dengan diabetes masih sedikit lebih rendah.Namun secara keseluruhan, model ini tetap memiliki performa yang konsisten, efisien, dan mudah diinterpretasikan, sehingga cocok digunakan untuk mendukung analisis medis yang membutuhkan hasil prediksi yang dapat dijelaskan secara logis.



Gambar 9. Confusion Matrix (Proportion of Predicted) untuk Algoritma Logistic Regression.

Sementara itu, pada algoritma Random Forest, hasil menunjukkan performa yang paling unggul di antara ketiga model. Model ini berhasil mengklasifikasikan 97,1% data *non-diabetes* dan 97,3% data *diabetes* dengan benar, menunjukkan keseimbangan klasifikasi yang sangat baik di kedua kelas. Kemampuan ini didukung oleh mekanisme *ensemble learning* yang menggabungkan hasil dari beberapa *decision tree* untuk mengurangi kesalahan prediksi. Random Forest juga lebih tahan terhadap *overfitting*, karena setiap pohon keputusan dibangun dari subset data dan fitur yang berbeda, sehingga hasil akhirnya lebih stabil dan akurat.

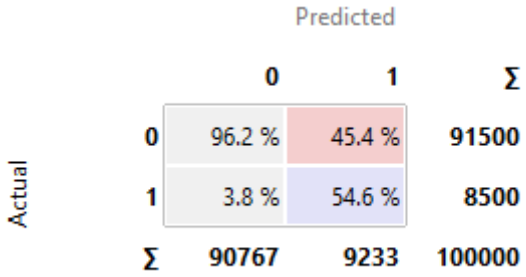


Gambar 10. Confusion Matrix (Proportion of Predicted) untuk Algoritma Random Forest.

Kinerja unggul Random Forest dalam penelitian ini membuktikan bahwa metode berbasis *ensemble* sangat efektif untuk kasus medis yang melibatkan banyak variabel dengan korelasi kompleks. Dengan tingkat akurasi yang tinggi serta keseimbangan yang baik antar kelas, model ini dapat dianggap sebagai pendekatan paling optimal dalam memprediksi potensi diabetes berdasarkan data pasien.

Berbeda dengan dua model sebelumnya, hasil Confusion Matrix pada algoritma Support Vector Machine (SVM) menunjukkan performa yang sedikit lebih rendah, terutama pada pengenalan kelas positif (*diabetes*). Model ini mampu mengklasifikasikan 96,2% data *non-diabetes* dengan benar, namun hanya 54,6% data *diabetes* yang berhasil diidentifikasi secara tepat. Perbedaan yang cukup

signifikan ini mengindikasikan bahwa SVM memiliki kecenderungan untuk lebih mudah mengenali pasien sehat dibandingkan pasien yang benar-benar mengidap diabetes.



Gambar 11. Confusion Matrix (Proportion of Predicted) untuk Algoritma Support Vector Machine (SVM).

Fenomena ini kemungkinan disebabkan oleh ketidakseimbangan data (imbalanced dataset) pada dataset diabetes, di mana jumlah data pasien *non-diabetes* jauh lebih banyak dibandingkan pasien *diabetes*. Selain itu, meskipun kernel RBF membantu menangani data non-linear, parameter yang digunakan mungkin belum sepenuhnya optimal untuk pola distribusi data tertentu.Akibatnya, model lebih cenderung menghasilkan prediksi konservatif yang meminimalkan kesalahan pada kelas mayoritas, namun mengorbankan sensitivitas pada kelas minoritas.

5. KESIMPULAN DAN SARAN

Kesimpulan

Berdasarkan hasil penelitian yang dilakukan menggunakan aplikasi Orange Data Mining, dapat disimpulkan bahwa penerapan algoritma *machine learning* mampu memberikan hasil yang efektif dan akurat dalam memprediksi penyakit diabetes berdasarkan data klinis pasien. Dataset yang digunakan terdiri dari beberapa atribut penting seperti *age*, *BMI*, *glucose level*, *blood pressure*, dan *HbA1c level*, yang telah melalui proses *preprocessing* dan transformasi data menggunakan *One-Hot Encoding* agar dapat diolah oleh model secara optimal. Tiga algoritma yang diterapkan, yaitu Logistic Regression, Random Forest, dan Support Vector Machine (SVM), diuji menggunakan metode 10-Fold Cross Validation untuk memastikan hasil yang adil dan representatif. Hasil evaluasi menunjukkan bahwa Random Forest memberikan performa terbaik dengan akurasi sebesar 97,1%, diikuti oleh Logistic Regression dengan 96,0%, dan SVM dengan 92,3%. Berdasarkan hasil *Confusion Matrix*, Random Forest menunjukkan keseimbangan yang baik dalam mendeteksi kelas *diabetes* maupun *non-diabetes*, sementara Logistic Regression tetap stabil dan mudah diinterpretasikan. Secara keseluruhan, algoritma Random Forest terbukti menjadi model paling

optimal untuk prediksi penyakit diabetes, karena mampu menghasilkan hasil klasifikasi yang akurat, seimbang, dan konsisten. Penelitian ini membuktikan bahwa penerapan *machine learning*, khususnya pendekatan *ensemble learning*, dapat menjadi dasar pengembangan sistem pendukung keputusan medis yang dapat membantu tenaga kesehatan dalam mendeteksi dini penyakit diabetes secara lebih efisien, akurat, dan berbasis data.

Saran

Penelitian selanjutnya disarankan untuk menggunakan dataset yang lebih besar, beragam, dan seimbang agar model prediksi yang dihasilkan lebih representatif terhadap kondisi nyata. Selain itu, perlu diterapkan teknik *data balancing* seperti *SMOTE* untuk mengatasi ketidakseimbangan antara data pasien diabetes dan non-diabetes. Penelitian mendatang juga dapat mempertimbangkan penggunaan metode *deep learning* seperti *Artificial Neural Network (ANN)*, *Convolutional Neural Network (CNN)*, atau *Gradient Boosting* guna meningkatkan akurasi dan kemampuan model dalam mengenali pola kompleks pada data medis. Integrasi model prediksi ke dalam sistem berbasis web atau aplikasi mobile juga menjadi langkah penting agar hasil penelitian dapat dimanfaatkan secara langsung oleh tenaga medis maupun masyarakat sebagai alat bantu deteksi dini diabetes yang cepat, efisien, dan mudah diakses.

UCAPAN TERIMA KASIH

Puji syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa atas terselesaikannya penelitian berjudul "Analisis Komparatif Algoritma Machine Learning untuk Prediksi Penyakit Diabetes." Penulis mengucapkan terima kasih kepada dosen pembimbing, seluruh dosen Program Studi Informatika Universitas Bina Sarana Informatika, rekan penelitian, serta keluarga atas bimbingan dan dukungan yang diberikan selama proses penelitian. Penelitian ini diharapkan dapat menjadi referensi dalam penerapan *machine learning* untuk deteksi dini penyakit diabetes. Untuk penelitian selanjutnya, disarankan menggunakan dataset yang lebih besar dan seimbang, serta mengembangkan metode *deep learning* atau sistem berbasis web agar hasil prediksi lebih akurat dan dapat diterapkan dalam lingkungan medis secara praktis.

DAFTAR REFERENSI

Akhsani, R., Prayoga, S., Basatha, R., Akbar, M. S., Aisyah Elfaiz, E., Putra, C. D., Surabaya, N., Kec, J. K., & Surabaya, G. (n.d.). Penerapan metode Naïve Bayes untuk klasifikasi performa siswa. Sistemasi: Jurnal Sistem Informasi. http://sistemasi.ftik.unisi.ac.id

Choksi, P. (2023). Comprehensive diabetes clinical dataset (100k rows). Kaggle.

Citra Mawani, A., Li Hin, L., & Anubhakti, D. (2023). Deteksi dini gejala awal penyakit diabetes menggunakan algoritma Random Forest. Idealis: Indonesia Journal Information System, 6(2). http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

Dita Ayuningtiyas Tuti, Fitriyani, N. L., & Maulana, J. (2023). Literature study: Risk factors for the incidence of diabetes mellitus in productive age in Indonesia. Journal of Multidisciplinary Science, 2(6), 288–296. https://doi.org/10.58330/prevenire.v2i6.413

Exploring the non-communicable disease burden in Indonesia – Findings from the 2023 health survey. (2025). Indonesia Journal of Public Health Nutrition, 5(2). https://doi.org/10.7454/ijphn.v5i2.1064

Fadli Kurniawan, M., & Ayu Megawaty, D. (2025). Comparison of logistic regression, random forest, support vector machine (SVM) and K-nearest neighbor (KNN) algorithms in diabetes prediction. Journal of Applied Informatics and Computing, 9(5). http://jurnal.polibatam.ac.id/index.php/JAIC

Fadlianda, D., Prananto, A., Eriska, C. A., Anjanira, S., Syadzwina, N., & Ula, M. (n.d.). Diagnosis penyakit jantung menggunakan algoritma Support Vector Machine (SVM). SENASTIKA Universitas Malikussaleh. https://www.kaggle.com/code/rafiromolo/prediksi-

Hakim, L., Sobri, A., Sunardi, L., & Nurdiansyah, D. (2025). Prediksi penyakit jantung berbasis machine learning dengan menggunakan metode K-NN. Jurnal Digital Teknologi Informasi, 7(2), 14. https://doi.org/10.32502/digital.v7i2.9429

International Diabetes Federation. (2024, Oktober). Indonesia – Western Pacific members. International Diabetes Federation.

Khairunnisa, A. (n.d.). Analisis perbandingan model regresi logistik dan probit dengan K-fold cross validation dalam mengidentifikasi faktor signifikan pada penyakit diabetes melitus. https://doi.org/10.26555/konvergensi.30879

Lu, W., Zhang, Y., Wen, W., Yan, H., & Li, C. (Eds.). (2022). Cyber security (Vol. 1506). Springer Nature Singapore. https://doi.org/10.1007/978-981-16-9229-1

Rahaman, M. J. (2024). A comprehensive review to understand the definitions, advantages, disadvantages and applications of machine learning algorithms. International Journal of Computer Applications, 186(31), 43–47. https://doi.org/10.5120/ijca2024923868

Sanhaji, G., Febrianti, A., & Teknik, F. (n.d.). Aplikasi DIATECT untuk prediksi penyakit diabetes menggunakan SVM berbasis web (Vol. 18, No. 1).

Siswoyo, B., & Iqbal Nurhafidz, M. (n.d.). Penerapan algoritma Random Forest untuk prediksi risiko diabetes berdasarkan data kesehatan pasien. JTID Integrasi Publikasi Digital, 1(1).

Syahputra, H., & Wibowo, A. (2023). Comparison of Support Vector Machine (SVM) and Random Forest algorithm for detection of negative content on websites. Jurnal Ilmiah Teknik Elektro Komputer dan Informatika, 9(1), 165–173. https://doi.org/10.26555/jiteki.v9i1.25861

Syamsudin, T., Handhayani, T., Muhammad, _____, & Syaifudin, I. (n.d.). Perbandingan klasifikasi penyakit diabetes menggunakan metode machine learning. Jurnal Ilmu Komputer dan Sistem Informasi. https://www.kaggle.com/datasets/nanditapore/healthcar

Teknika, J., & Supriyatna, A. R. (n.d.). Prediksi penyakit diabetes menggunakan algoritma Random Forest. Teknika, 17(1), 163–172.

Yanti, D. E., Framesti, L., & Desiani, A. (n.d.). Perbandingan algoritma C4.5 dan SVM dalam klasifikasi penyakit anemia. JIP (Jurnal Informatika Polinema). https://www.kaggle.com/datasets/biswaranjanrao/an

Yusoff, M. I. M. (2024). Machine learning: An overview. Open Journal of Modelling and Simulation, 12(3), 89–99. https://doi.org/10.4236/ojmsi.2024.123006