METODE ENSEMBLE WEIGHTED VOTING UNTUK DETEKSI RISIKO DIABETES

Ach.Diki Prasetyo[1], Fetty Tri Anggraeny [2], Retno Mumpuni[3]

[1,2,3] Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional "Veteran" Jawa Timur, Indonesia

[1]dikiprasetyo1234@gmail.com, [2]fettyanggraeny.if@upnjatim.ac.id, [3]retnomumpuni.if@upnjatim.ac.id

.

Abstrak

Diabetes Melitus (DM) adalah penyakit kronis yang ditandai dengan peningkatan kadar glukosa darah akibat gangguan produksi atau fungsi insulin. Secara global, prevalensi DM terus meningkat, dengan sekitar 537 juta penderita pada tahun 2021 dan proyeksi mencapai 783 juta pada tahun 2045 jika tidak ada penanganan yang lebih efektif. Deteksi dini penyakit ini sangat penating untuk mencegah komplikasi yang lebih serius. Namun, diagnosis manual konvensional sering kali memakan waktu dan biaya yang besar, sehingga menghambat upaya tersebut. Penelitian ini bertujuan mengembangkan model prediksi risiko diabetes yang efisien dan mudah diakses menggunakan metode ensemble *weighted voting*. Penelitian ini mengombinasikan tiga algoritma *machine learning*, yaitu *Logistic Regression, Support Vector Machine*, dan *Random Forest*. Data yang digunakan berasal dari survei publik "*Diabetes Health Indicators Datas*et" (BRFSS 2021) serta data primer lokal. Metodologi penelitian mencakup pengumpulan data, pra-pemrosesan, pelatihan model individual, dan pembentukan model *ensemble* dengan pembobotan berdasarkan akurasi model. Kinerja model dievaluasi menggunakan metrik akurasi, presisi, recall, dan *F1-score* melalui empat skenario pengujian. Hasil menunjukkan bahwa model *ensemble weighted voting* memberikan kinerja yang baik dengan akurasi tertinggi 90,00% pada skenario yang memadukan data latih terbatas dan data uji lokal. Penelitian ini menyimpulkan bahwa metode ensemble weighted voting merupakan metode yang cukup baik untuk pengembangan model prediksi risiko diabetes yang lebih akurat dan praktis

Kata kunci: Prediksi Risiko Diabetes, *Logistic Regression, Random Forest, SVM, Weighted Voting*.

1. Pendahuluan

Diabetes Melitus (DM) adalah penyakit kronis dengan ditandai peningkatan kadar glukosa darah yang diakibatkan oleh gangguan produksi insulin Secara global, prevalensi diabetes terus meningkat, dengan perkiraan 537 juta penderita pada tahun 2021 dan proyeksi peningkatan menjadi 783 juta pada tahun 2045 jika tidak ada intervensi efektif (Lestari et al., 2021). Deteksi dini menjadi krusial karena diabetes seringkali tidak teridentifikasi hingga timbul komplikasi serius pada organ vital. Pendekatan diagnosis manual konvensional seringkali memakan waktu dan biaya, serta memerlukan akses ke fasilitas medis. Oleh karena itu, pengembangan alat prediksi risiko diabetes yang efektif, akurat, dan mudah diakses sangat penting (Kusumastuti, 2024)

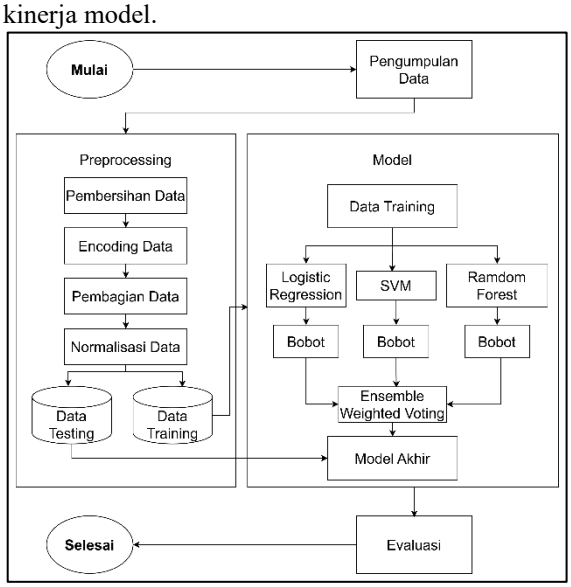
Machine learning (*ML*) merupakan sebuah solusi potensial untuk prediksi risiko diabetes. Berbagai algoritma *ML* memiliki keunggulan tersendiri dengan spesifik seperti *Logistic Regression efektif* untuk data *linear* dan interpretasi probabilitas (Yazar, n.d.). Lalu *Support Vector Machine (SVM)* mampu menangani data non-linear menggunakan *kernel trick* (Mase et al., 2018). Kemudian yang terakhir adalah *Random Forest* yang dapat meningkatkan akurasi serta mengurangi *overfitting* melalui *agregasi decision trees* (Marlina Haiza et al., 2023). Metode *ensembl*e, yang menggabungkan beberapa model dasar, telah terbukti dapat meningkatkan performa prediktif dan mengatasi masalah *overfitting* (Noreen et al., 2021)

Penelitian ini difokuskan untuk pada pengembangan dan evaluasi model prediksi risiko diabetes dengan menerapkan metode *ensemble weighted voting*, yang mengintegrasikan dari kekuatan prediksi dari *Logistic Regression, Support Vector Machine*, dan *Random Forest*. Kelebihan dari penelitian ini terletak pada pemanfaatan dataset berbasis survei yang mencakup faktor-faktor gaya hidup dan riwayat kesehatan, sehingga memungkinkan prediksi dilakukan tanpa perlu pemeriksaan medis langsung yang umumnya memerlukan waktu dan biaya besar. Diharapkan, dengan pemilihan algoritma dasar yang memiliki karakteristik unggul masing-masing dan pendekatan ensemble yang sesuai, model yang dibangun mampu berkontribusi dalam pengembangan alat prediksi risiko diabetes yang lebih akurat.

2. Metode

Metodologi penelitian ini dibuat untuk mencapai tujuan yang telah ditentukan. Keseluruhan tahapan

penelitian digambarkan pada Gambar 1. Proses dimulai dengan pengumpulan data, kemudian dilanjutkan dengan tahap pra-pemrosesan data. Selanjutnya, dilakukan pemodelan yang mencakup pelatihan model-model individual serta model *ensemble*, dan diakhiri dengan evaluasi terhadap kinerja model.



Gambar 1. Alur Metode Penelitian

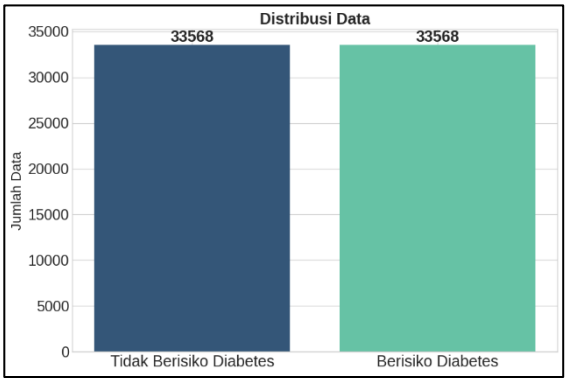
2.1 Pengumpulan Data

Dalam penelitian ini, menggunakan dua dataset utama. Pertama, dataset sekunder yang bernama "*Diabetes Health Indicators Dataset*" diperoleh dari survei *Behavioral Risk Factor Surveillance System (BRFSS)* tahun 2021 yang diakses melalui platform *Kaggle*. Dataset ini terdiri dari 67.136 data dengan dengan total 22 variabel, yang mencakup 21 fitur dan 1 variabel target (status risiko diabetes). Kemudian, untuk data primer dikumpulkan melalui survei lokal menggunakan *Google Form* yang mengikuti variabel serupa dengan dataset *Kaggle* dan telah disesuaikan kriteria pengisiannya sesuai dengan kondisi di indonesia, lalu menghasilkan 130 responden lokal. Rincian mengenai variabel fitur dan target dapat dilihat pada Tabel 1.

Tabel 1 Variabel dalam Dataset

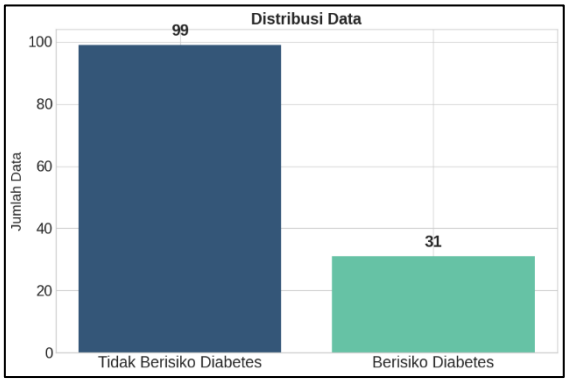
No	Nama Variabel	Tipe Data
1	*Diabetes_binary (Target)*	Biner
2	*HighBP*	Biner
3	*HighChol*	Biner
4	*CholCheck*	Biner
5	*BMI*	Numerik
6	*Smoker*	Biner
7	*Stroke*	Biner
8	*HeartDiseaseorAttak*	Biner
9	*PhysActivity*	Biner
10	*Fruits*	Biner
11	*Veggies*	Biner
12	*HvyAlcoholConsump*	Biner
13	*AnyHealthcare*	Biner
14	*NoDocbcCost*	Numerik
15	*GenHlth*	Biner
16	*MentHlth*	Numerik
17	*PhysHlth*	Numerik
18	*DiffWalk*	Biner
19	*Sex*	Biner
20	*Age*	Biner
21	*Education*	Biner
22	*Income*	Biner

Pada Gambar 2 menggambarkan distribusi jumlah data pada masing-masing kelas dalam dataset risiko diabetes yang digunakan dalam penelitian ini. Dataset terbagi menjadi dua kelas, yaitu kelas "Tidak Berisiko Diebetes" dan "Berisiko Diabetes", masing-masing berjumlah 33.568 data. Secara keseluruhan, dataset ini memiliki total data sekitar 67.136 entri.



Gambar 2 Distribusi dataset kaggle

Kemudian pada Gambar 3 memperlihatkan distribusi kelas pada data primer yang didapatkan langsung melalui pengumpulan data menggunakan Google Form dan diambil dari Indonesia. Sama seperti data sekunder, data primer ini juga terdiri dari 21 variabel fitur dan 1 variabel target. Terdapat dua kelas, yaitu kelas "Berisiko Diabetes" dengan 31 data dan kelas "Tidak Berisiko Diabetes" dengan 99 data, sehingga total data primer berjumlah 130 entri.



Gambar 3 Distribusi Dataset Form

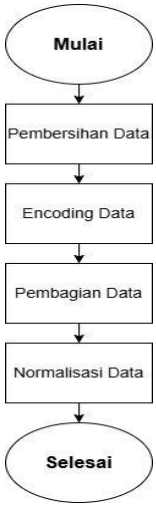
Tabel 2 merupakan pembagian dataset dalam empat skenario evaluasi model dengan variasi jumlah data latih dan uji. Skenario 1 dan 2 menggunakan dataset *Kaggle* dengan rasio data latih dan uji 80:20 dan 70:30 untuk mengukur pengaruh ukuran data latih terhadap kinerja model. Skenario 3 menguji kemampuan generalisasi model dengan melatih pada data *Kaggle* dan menguji pada data primer lokal, sedangkan Skenario 4 untuk mensimulasikan keterbatasan data latih untuk mengevaluasi kekuatan model saat diuji pada data primer lokal. Keempat

skenario penelitian ini bertujuan memberikan gambaran menyeluruh tentang performa model pada kondisi data yang bervariasi.

Tabel 2 Pembagian Data

Data	Skenario 1	Skenario 2	Skenario 3	Skenario 4
Latih	53.119	46.479	46.479	3.357
Uji	13.280	19.920	130 (Primer)	130 (Primer)

2.2 Pra-pemrosesan Data



Gambar 4 Alur Pra-pemrosesan Dataset

Setelah data terkumpul, dilakukan tahap pra-pemrosesan untuk meningkatkan kualitas dan kesesuaian data untuk tahap pemodelan (Khan et al., 2021). Langkah-langkah pra-pemrosesan meliputi:

- Pembersihan Data: Mengidentifikasi dan menghapus data duplikat dari dataset Kaggle. Data kosong (*missing values*), jika ada, akan ditangani dengan metode imputasi yang sesuai (misalnya, modus untuk data kategorikal).
- Encoding Data: Mengubah variabel kategorikal (misalnya, jawaban "Ya"/"Tidak", tingkat pendidikan) menjadi representasi numerik yang dapat diproses oleh algoritma *machine learning*. Encoding ini hanya digunakan pada dataset lokal saja.
- Pembagian Data: Dataset dibagi menjadi data latih (*training set*) dan data uji (*testing set*) dengan proporsi tertentu seperti 80:20, 70:30, atau dapat dilihat pada Tabel 2 untuk semua pembagian datanya. Beberapa skenario pembagian data, termasuk penggunaan data Kaggle untuk pelatihan dan data lokal untuk pengujian, akan dieksplorasi. Data latih digunakan untuk melatih model, sedangkan data uji digunakan untuk evaluasi model akhir.
- Normalisasi Data: Variabel numerik yang memiliki rentang nilai yang berbeda sepert *BMI*, *age,* dan sebagainya. Pada data latih dan data uji dinormalisasi menggunakan teknik *Min-Max Scaling* untuk mengubah nilainya ke dalam rentang yang sama seperti 0 hingga 1. Normalisasi ini bertujuan untuk mencegah fitur dengan skala besar mendominasi proses pelatihan model.

2.3 Logistic Rgression

Logistic Regression merupakan salah satu algoritma klasifikasi yang digunakan untuk memperkirakan probabilitas dari variabel dependen yang bersifat kategorikal, biasanya dengan dua kelas (biner), berdasarkan satu atau lebih variabel independen. Algoritma ini memanfaatkan fungsi *logistik* atau *sigmoid* untuk mengubah hasil prediksi menjadi nilai probabilitas yang berada dalam rentang antara 0 hingga 1 (Rajendra & Latifi, 2021). Persamaan 1 untuk kombinasi linear :

$$P(X) = \frac{1}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}} \quad (1)$$

Keterangan :

P(X)	:	Probabilitas seseorang berisiko diabetes
$x_1, x_2, \ldots, x_n$	:	Variabel fitur
$\beta_0, \beta_1, \beta_2, .. \beta_n$	:	Koefisien regresi

2.4 Random Forest

*Random Fores*t adalah metode *ensemble learning* yang terdiri dari banyak pohon keputusan (*decision trees).* Setiap pohon dibangun menggunakan sampel bootstrap dari data pelatihan, dan pada setiap node, pemisahan data dilakukan berdasarkan subset fitur yang dipilih secara acak. Prediksi akhir dalam klasifikasi biasanya ditentukan melalui mekanisme majority voting dari seluruh pohon yang ada (G et al., 2022). Salah satu kriteria yang umum digunakan untuk menentukan pemisahan pada node adalah *Gini impurity,* yang dirumuskan pada persamaan 2 sebagai berikut

$$Gini(X) = 1 - \sum_{i-1}^{c} p_i^2 \quad (2)$$

Keterangan

p_i : Proposisi data pada ke-i dalam node tersebut.

c : merupakan jumlah kelas dalam klasifikasi.

Setelah seluruh pohon dalam *Random Forest* terbentuk, masing-masing pohon memberikan prediksi berdasarkan data uji y ang diberikan. Setiap pohon menghasilkan prediksi berupa kelas 0 atau 1 untuk input tersebut. Selanjutnya, hasil akhir dari model *Random Forest* diperoleh melalui proses majority voting dari seluruh prediksi pohon, yang secara matematis dapat dinyatakan pada persamaan 3.

$$\hat{y} = model(h_1(x), h_2(x), \ldots, h_k(x)\}) \quad (3)$$

Keterangan :

$\hat{y}$: Hasil prediksi akhir ensemble.

$(h_i(x)$: Hasil prediksi dari pohon ke-i.

k : Merupakan total jumlah pohon didalam random forest.

2.5 Support Vector Machine

Untuk data yang bisa dipisah secara *linear*, *Support Vector Machine (SVM)* akan mencari sebuah *hyperlane* yang memaksimalkan *margin* atau jarak terdekat antara titik data dari kedua kelas ke *hyperlane* tersebut. *Margin* ini merupakan indikator seberapa baik model memisahkan dua kelas secara jelas (Viloria et al., 2020). Pada fungsi *linear* yang digunakan untuk dataset dengan banyak fitur dinyatakan sebagai persamaan 4.

$$f(x) = w_1x_1 + w_2x_2 + \cdots + w_nx_n + b = 0 \quad (4)$$

Keterangan :

$x_1, x_2, \ldots, x_n$: Merupakan fitur input.

$w_1, w_2, \ldots, w_n$: Merupakan bobot koefisien dari setiap fitur.

b : Merupakan bias yang menggeser posisi *hyperlane*.

$f(x)$: Merupakan nilai dari keputusan yang dihasilkan oleh model.

Untuk mengatasi hubungan *non-linear* antar fitur, *kernel Radial Basis Function (*RBF) dapat digunakan. Kernel ini berperan dalam menghitung jarak antar dua titik data di ruang fitur yang lebih kompleks (Sidharth, 2022). Persamaan kernel RBF ditunjukkan sebagai berikut :

$$K(x_ix_j) = exp\,(-\gamma||x_i - x_j||^2) \quad (5)$$

x_i, x_j : Data fitur pada dataset

$||x_i - xj||$: Jarak *euclidean* antar dua titik data

γ : Parameter yang digunakan untuk mengontrol pengaruh dari masing-masing sample

Selain itu, kernel *Polynomial* juga digunakan untuk menangkap hubungan *non-linear* dengan mentransformasikan ruang fitur ke bentuk *polynomial* (Fremmuzar & Baita, 2023). Persamaan 6 kernel *Polynomial* dinyatakan sebagai berikut :

$$k(x_i, x_j) = (x_i \cdot x_j + c)^d \quad (6)$$

Keterangan :

x_i, x_j : Data fitur dalam dataset

$x_i \cdot x_j$: Hubungan *linear* antara dua sample dalam ruang fitur

c : Bias atau koefisien kontan untuk menyesuaikan skala transformasi

d : Derajat *polynomial* untuk menentukan nilai kompleksitas transformasi fitur

Kernel *sigmoid* berfungsi untuk mentransformasikan data fitur ke bentuk non-linear dengan memanfaatkan fungsi *sigmoid*, yang membatasi output dalam rentang antara -1 hingga 1 (Fremmuzar & Baita, 2023). Proses perhitungannya ditunjukkan pada Persamaan 7.

$$k(x, y) = tanh\,(\alpha x^T y + c) \quad (7)$$

Keterangan :

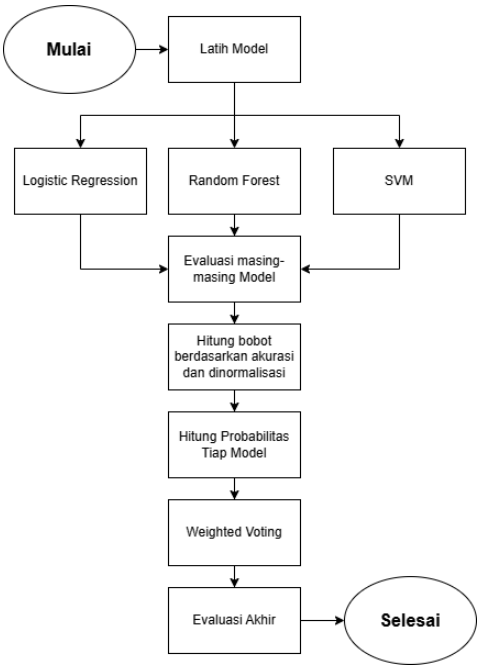
$x^T y$: Hubungan data berdasarkan fitur

α : Skala untuk mengontrol sensitivitas hubungan antar fitur

c : Bias yang digunakan untuk menentukan

$tangh$: Fungsi yang digunakan untuk membatasi output antara -1 dan 1

2.6 Ensemble Weighted Voting



Gambar 5 Alur Penbuatan Model *Ensemble Weighted Voting*

Ensemble learning merupakan teknik yang menggabungkan sejumlah model prediktif *(weak learners)* untuk memperoleh hasil klasifikasi yang lebih andal dibandingkan dengan model individu. Dalam metode *weighted voting*, kontribusi setiap model terhadap prediksi akhir ditentukan oleh bobot yang mencerminkan kinerja model tersebut (Triyana et al., 2024). Pada penelitian, model dasar yang digunakan *Logistic Regression* (LR), *Support Vector Machine* (SVM), dan *Random Forest* (RF) yang dapat dilihat pada Gambar 5. Penentuan bobot dilakukan dengan memperhitungkan akurasi masing-masing model menggunakan persamaan berikut:

$$W_i = \frac{accuracy}{\sum_{q=1}^{n} accuracy_q} \quad (8)$$

Keterangan :

W_i : Bobot yang diberikan untuk

model ke-i

$accuracy_q$: Nilai akurasi yang didapatkan oleh model ke-i

n : Jumlah model dalam *ensemble*

$$P(X) = arg\ max_i \sum_{j=1}^{N} w_j \cdot p_{ij} \quad (9)$$

Perhitungan akhir prediksi dilakukan dengan menjumlahkan estimasi probabilitas dari setiap model yang telah dikalikan dengan bobotnya, sebagaimana ditunjukkan pada persamaan 9 (Garrochamba Peñafiel, 2024).

P(X) : Hasil prediksi akhir yang didapatkan oleh *weighted voting.*

N : Jumlah model yang di *ensembl*e

w_j : Bobot dari klasifikasi ke-j

p_{ij} : Estimasi probabilitas dari klasifikasi ke-j untuk kelas ke-i

2.7 Confusion Matriks

Tabel 3 *Confusion Matriks*

	Positive (Predicted)	*Negative (Predicted)*
Positive/ Aktual	*True Positive*	*False Negative*
Negative/ Aktual	*False Positive*	*True Negative*

Evaluasi kinerja model klasifikasi dalam penelitian ini dilakukan dengan memanfaatkan *confusion matrix* serta sejumlah metrik evaluasi turunan, sebagaimana ditampilkan pada Tabel 3. *Confusion matrix* digunakan sebagai dasar untuk memperoleh nilai-nilai penting seperti *True Positive (TP), True Negative (TN), False Positive (FP),* dan *False Negative (FN*) (Garrochamba Peñafiel, 2024). Nilai-nilai tersebut kemudian digunakan untuk menghitung berbagai indikator performa model yang mencerminkan kemampuan klasifikasi secara keseluruhan menggunakan persamaan 10,11, 12 dan 13 yang dapat dilihat sebagai berikut:

- *Accuracy* : Mengukur total prediksi yang benar dari semua data uji.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

- *Precision*: Mengukur prediksi positif yang aslinya memang positif.

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

- *Recall*: Untuk mengukur kemampuan model dalam deteksi seluruh hasil prediksi positif yang sebenarnya.

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

- *F1-Score*: Memberikan keseimbangan antara presisi dan recall dengan cara menggabungkan kedua metrik tersebut dalam satu nilai.

$$F1 - score = 2 \times \frac{Recall \times Precision}{Precision + Recall} \quad (13)$$

3. Hasil dan Pembahasan

Evaluasi model dilakukan melalui empat skenario pengujian dengan variasi pada proporsi data latih dan data uji. Fokus utama evaluasi ini adalah untuk menilai performa model *Ensemble Weighted Voting* dalam memprediksi risiko diabetes. Hasil akurasi dari model *ensemble* pada masing-masing skenario dapat dilihat pada Tabel 4.

Pada Skenario 1, dengan menggunakan 80% data Kaggle untuk pelatihan dan 20% sisanya untuk pengujian, model *ensemble* memberikan akurasi sebesar 74,89%. Skenario 2, yang menggunakan 70% data *Kaggle* dengan menggunakan pelatihan dan 30% untuk pengujian, mendapatkan nilai akurasi sebesar 74,57%.

Kemudian pada skenario 3 yang menguji kemampuan model terhadap data primer sebagai data uji, sementara data latih tetap menggunakan dari 70% data *Kaggle*. Dalam skenario ini, model *ensemble* berhasil mencapai akurasi sebesar 89,23%. Performa tertinggi ditunjukkan pada Skenario 4, ketika hanya 5% data *Kaggle* digunakan sebagai data latih dan sisanya adalah data primer sebagai data uji, di mana akurasi model *ensemble* mencapai 90,00%.

Tabel 4 Evaluasi Setiap Skenario

Skenario	**Rasio**	***Accuracy***
Skenario 1	80:20	74.89%
Skenario 2	70:30	74.57%
Skenario 3	Kaggle (70%) : Data Primer	89.23%
Skenario 4	Kaggle (5%) : Data lokal	90.00%

Hasil ini menunjukkan bahwa model *ensemble* memiliki performa yang konsisten dan dapat meningkatkan akurasi pada beberapa skenario, terutama dalam skenario dengan keterbatasan data latih atau ketika diuji pada data uji dari sumber yang berbeda. Dengan menggunakan metode *weighted voting* bisa meningkatkan stabilitas dan akurasi prediksi, serta menjadikannya metode yang potensial untuk mendukung deteksi dini risiko diabetes secara lebih akurat.

4. Kesimpulan

Berdasarkan hasil evaluasi pada empat skenario pengujian, dapat disimpulkan bahwa metode *Ensemble Weighted Voting* memberikan kinerja prediksi yang baik dan konsisten dalam tugas prediksi risiko diabetes. Pada Skenario 1 dan 2 yang menggunakan dataset Kaggle secara penuh, model ensemble berhasil mencatat akurasi tertinggi masing-masing sebesar 74,89% dan 74,57%. Hal ini menunjukkan bahwa integrasi dari tiga algoritma *(Logistic Regression, Random Forest, dan SVM)*

mampu menghasilkan prediksi yang lebih stabil dibandingkan pendekatan tunggal.

Keunggulan metode ini semakin terlihat pada Skenario 4, di mana hanya 5% data *Kaggle* digunakan sebagai data latih dan sisanya berupa data primer lokal. Dalam kondisi terbatas tersebut, model ensemble tetap mampu mencapai akurasi tertinggi sebesar 90,00%. Temuan ini menunjukkan bahwa pendekatan ensemble tidak hanya efektif pada data homogen, tetapi juga tangguh terhadap keterbatasan data dan perbedaan distribusi antara data latih dan uji. Dengan demikian, *Ensemble Weighted Voting* dapat menjadi alternatif yang andal dalam pengembangan alat bantu prediksi risiko diabetes.

Daftar Pustaka:

Fremmuzar, P., & Baita, A. (2023). Uji Kernel SVM dalam Analisis Sentimen Terhadap Layanan Telkomsel di Media Sosial Twitter. *Komputika : Jurnal Sistem Komputer*, *12*(2), 57–66. https://doi.org/10.34010/komputika.v12i2.9460

G, A., Ganesh, B., Ganesh, A., Srinivas, C., Dhanraj, & Mensinkal, K. (2022). Logistic regression technique for prediction of cardiovascular disease. *Global Transitions Proceedings*, *3*(1), 127–130. https://doi.org/10.1016/j.gltp.2022.04.008

Garrochamba Peñafiel, B. D. (2024). Factores de Riesgo Asociados a Diabetes Mellitus Tipo 2. *Revista Científica de Salud y Desarrollo Humano*, *5*(2), 101–115. https://doi.org/10.61368/r.s.d.h.v5i2.123

Khan, F. A., Zeb, K., Al-Rakhami, M., Derhab, A., & Bukhari, S. A. C. (2021). Detection and Prediction of Diabetes Using Data Mining: A Comprehensive Review. *IEEE Access*, *9*, 43711–43735. https://doi.org/10.1109/ACCESS.2021.3059343

Kusumastuti, R. (2024). *PREDIKSI RISIKO DIABETES MENGGUNAKAN ALGORITMA DECISION TREE DENGAN APLIKASI RAPID MINER*. *November*, 14–24.

Lestari, Zulkarnain, Sijid, & Aisyah, S. (2021). Diabetes Melitus: Review Etiologi, Patofisiologi, Gejala, Penyebab, Cara Pemeriksaan, Cara Pengobatan dan Cara Pencegahan. *UIN Alauddin Makassar*, *1*(2), 237–241. http://journal.uin-alauddin.ac.id/index.php/psb

Marlina Haiza, Elmayati, Zulius Antoni, & Wijaya Harma Oktafia Lingga. (2023). Penerapan Algoritma Random Forest Dalam Klasifikasi Penjurusan Di SMA Negeri Tugumulyo. *Penerapan Kecerdasan Buatan*, *4*(2), 138–143.

Mase, J., Furqon, M. T., & Rahayudi, B. (2018). Penerapan Algoritme Support Vector Machine (SVM) Pada Pengklasifikasian Penyakit Kucing. *Jurnal Pegembangan Teknologi Informasi Dan Ilmu Komputer*, *2*(10), 3648–3654.

Noreen, N., Palaniappan, S., Qayyum, A., Ahmad, I., & Alassafi, M. O. (2021). Brain Tumor Classification Based on Fine-Tuned Models and the Ensemble Method. *Computers, Materials and Continua*, *67*(3), 3967–3982. https://doi.org/10.32604/cmc.2021.014158

Rajendra, P., & Latifi, S. (2021). Computer Methods and Programs in Biomedicine Update Prediction of diabetes using logistic regression and ensemble techniques. *Computer Methods and Programs in Biomedicine Update*, *1*, 100032. https://doi.org/10.1016/j.cmpbup.2021.100032

Sidharth. (2022). *The Kernel RBF in SVM*. Pycodemates. https://www.pycodemates.com/2022/10/the-rbf-kernel-in-svm-complete-guide.html

Triyana, D., Muharrom Al Haromainy, M., & Maulana, H. (2024). Implementasi Metode Ensemble Majority Vote Pada Algoritma Naive Bayes Dan Random Forest Untuk Analisis Sentimen Twitter Harga Tiket Pesawat Domestik. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(4), 7885–7894. https://doi.org/10.36040/jati.v8i4.10475

Viloria, A., Herazo-Beltran, Y., Cabrera, D., & Pineda, O. B. (2020). Diabetes Diagnostic Prediction Using Vector Support Machines. *Procedia Computer Science*, *170*, 376–381. https://doi.org/10.1016/j.procs.2020.03.065

Yazar, K. (n.d.). *Definition logistic regression*. Tech Target. Retrieved March 2, 2025, from https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression