BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 3, April 2025 | Hal 219-227
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i3.502

Pendekatan Hybrid K-Means SMOTE dan Logistic Regression Untuk Deteksi Dini Diabetes Mellitus Pada Imbalanced Data

Abdus Salam[1,*], Lukman Azhari[2], Ri Sabti Septarini[2], Nofitri Heriyani[2]

[1] Program Studi Informatika, Fakultas Teknik, Universitas 17 Agustus 1945 Jakarta, Jakarta Utara, Indonesia
[2] Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Tangerang, Tangerang, Indonesia
Email: [1,*]abdus.salam@uta45jakarta.ac.id, [2]lukman.azhari@ft-umt.ac.id, [3]risabtis@ft-umt.ac.id, [4]nofitri.heriyani@ft-umt.ac.id
Email Penulis Korespondensi: abdus.salam@uta45jakarta.ac.id

Abstrak−Peningkatan prevalensi Diabetes Mellitus secara global menuntut upaya deteksi dini yang lebih akurat, terutama melalui pendekatan berbasis machine learning. Namun, salah satu tantangan utama dalam klasifikasi medis adalah ketidakseimbangan data, di mana jumlah sampel penderita jauh lebih sedikit dibandingkan non-diabetes. Penelitian ini bertujuan untuk mengembangkan model *hybrid* dengan mengintegrasikan *Logistic Regression* dan *K-Means* SMOTE guna meningkatkan sensitivitas deteksi dini Diabetes Mellitus terhadap kelas minoritas. *Logistic Regression* dipilih karena efisien secara komputasi dan mudah diinterpretasikan, sedangkan *K-Means* SMOTE berperan dalam menyeimbangkan distribusi kelas dengan membangkitkan sampel sintetis secara terstruktur berdasarkan klaster data minoritas. Dataset yang digunakan terdiri dari 2.000 data dengan 9 fitur kesehatan, yang diperoleh dari *platform* Kaggle. Hasil evaluasi menunjukkan bahwa model dengan *K-Means* SMOTE memberikan performa terbaik, dengan akurasi sebesar 82,00%, F1-score sebesar 72,73% untuk kelas Diabetes, serta nilai ROC-AUC tertinggi sebesar 87,48%. Dibandingkan dengan metode tanpa oversampling dan SMOTE standar, pendekatan ini mampu meningkatkan generalisasi model dan sensitivitas terhadap kasus positif. Temuan ini memiliki implikasi praktis dalam pengembangan sistem deteksi dini berbasis *machine learning* yang lebih adil dan efektif, khususnya untuk diterapkan di fasilitas kesehatan dengan sumber daya terbatas.

Kata Kunci: Diabetes Mellitus; K-Means SMOTE; Logistic Regression; Klasifikasi Medis; Ketidakseimbangan Data

Abstract−The increasing global prevalence of Diabetes Mellitus necessitates more accurate early detection efforts, particularly through machine learning-based approaches. However, one of the main challenges in medical classification lies in data imbalance, where the number of diabetic cases is significantly lower than that of non-diabetic ones. This study aims to develop a hybrid model by integrating Logistic Regression and K-Means SMOTE to enhance the sensitivity of early detection for Diabetes Mellitus, especially toward the minority class. Logistic Regression is chosen for its computational efficiency and interpretability, while K-Means SMOTE plays a role in balancing class distribution by generating synthetic samples in a structured manner based on clusters of minority class data. The dataset used consists of 2,000 records with 9 health-related features, obtained from the Kaggle platform. Evaluation results indicate that the model utilizing K-Means SMOTE achieves the best performance, with an accuracy of 82.00%, an F1-score of 72.73% for the Diabetes class, and the highest ROC-AUC score of 87.48%. Compared to models without oversampling and with standard SMOTE, this approach improves model generalization and sensitivity to positive cases. These findings have practical implications for the development of fairer and more effective machine learning-based early detection systems, particularly for implementation in healthcare facilities with limited resources.

Keywords: Diabetes Mellitus; K-Means SMOTE; Logistic Regression; Medical Classification; Imbalanced Data

1. PENDAHULUAN

Diabetes Mellitus merupakan salah satu penyakit kronis dengan prevalensi yang terus meningkat secara global. Menurut data terbaru dari *World Health Organizatoin* (WHO), jumlah penderita diabetes meningkat dari 200 juta pada tahun 1990 menjadi sekitar 830 juta pada tahun 2022 [1]. Prevalensi global juga mengalami peningkatan signifikan, dari 7% menjadi 14% dalam periode yang sama, dengan lonjakan tertinggi terjadi di negara-negara berpenghasilan rendah dan menengah [2]. Di Indonesia sendiri, prevalensi diabetes telah menjadi perhatian serius karena menimbulkan komplikasi kesehatan jangka panjang serta membebani sistem layanan kesehatan nasional [3].

Deteksi dini terhadap penderita Diabetes Mellitus sangat krusial untuk mencegah komplikasi serius yang dapat mengancam nyawa. Seiring dengan pesatnya perkembangan teknologi informasi, pendekatan *machine learning* telah banyak dimanfaatkan dalam pengembangan sistem pendukung keputusan untuk menunjang diagnosis awal penyakit ini. Berbagai pendekatan modern seperti *deep learning* memang menawarkan tingkat akurasi yang tinggi, namun metode ini sering kali memerlukan jumlah data yang sangat besar, waktu pelatihan yang lama, serta sumber daya komputasi yang tinggi [4]. Oleh karena itu, algoritma *machine learning* klasik masih banyak digunakan karena lebih sederhana, efisien secara komputasi, mudah diinterpretasikan, dan tetap mampu memberikan performa yang kompetitif, khususnya pada dataset berskala kecil hingga menengah [5].

Berbagai algoritma *machine learning* klasik telah diterapkan dalam penelitian sebelumnya untuk membangun model klasifikasi diabetes. Salah satu penelitian menggunakan *Decision Tree* yang menghasilkan akurasi sebesar 65,06% dengan presisi 32,53% dan *recall* 50,00% [6]. Model *Support Vector Machine* (SVM) juga telah digunakan, menghasilkan akurasi sebesar 70,78% [7]. Studi lain membandingkan SVM dengan *Naive Bayes*, dan menunjukkan bahwa SVM memberikan akurasi lebih tinggi sebesar 78,04% dibandingkan *Naive Bayes* yang mencapai 76,98% [8]. Penelitian berikutnya membandingkan algoritma *Random Forest* dan *Gradient Boosting Classifier*, yang masing-masing mencatatkan akurasi sebesar 79% dan 81% pada rasio data latih dan uji 80:20 [9]. Selain itu, *K-Nearest Neighbor* (KNN) juga telah diuji pada beberapa skenario data uji, menghasilkan akurasi sebesar 82%, 84%, dan 82% untuk ukuran data uji 150, 200, dan 300 secara berurutan [10]. Meskipun algoritma-algoritma tersebut menunjukkan hasil yang menjanjikan, sebagian besar penelitian belum secara eksplisit mengatasi masalah ketidakseimbangan data (*imbalanced data*) yang

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 3, April 2025 | Hal 219-227
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i3.502

umum dijumpai pada dataset medis, termasuk dalam kasus diabetes. Kondisi ini terjadi ketika jumlah sampel penderita diabetes jauh lebih sedikit dibandingkan non-diabetes, yang menyebabkan model cenderung bias terhadap kelas mayoritas dan mengabaikan kelas minoritas. Akibatnya, performa model dalam mendeteksi pasien dengan risiko diabetes menjadi rendah, terutama dari sisi *recall* atau sensitivitas.

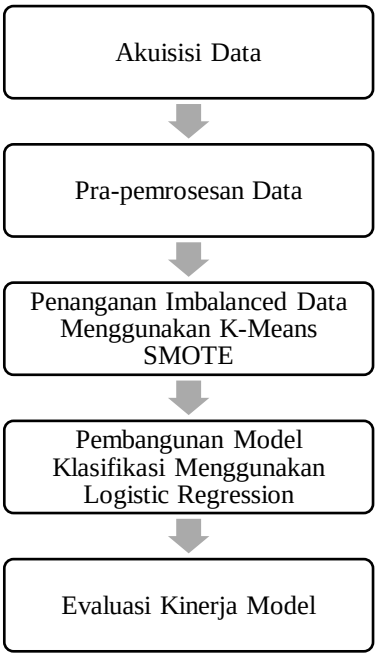
Penelitian ini mengusulkan penggunaan algoritma klasifikasi *machine learning* klasik yang sederhana namun kuat, yaitu *Logistic Regression*, sebagai solusi terhadap ketidakseimbangan data dalam deteksi dini Diabetes Mellitus. Algoritma ini memiliki keunggulan dalam interpretabilitas, efisiensi komputasi, serta kemampuannya menghasilkan output dalam bentuk probabilitas yang berguna untuk penilaian risiko. *Logistic Regression* juga mudah diperluas untuk klasifikasi multi-kelas dan tidak bergantung pada asumsi distribusi normal [11]. Namun demikian, *Logistic Regression* tetap rentan terhadap ketidakseimbangan data. Ketika distribusi kelas tidak seimbang, model cenderung bias terhadap kelas mayoritas dan mengabaikan kelas minoritas [12]. Untuk mengatasi permasalahan tersebut, penelitian ini mengintegrasikan teknik *oversampling* berbasis *K-Means* SMOTE. Berbeda dari SMOTE biasa yang hanya mengandalkan tetangga terdekat, *K-Means* SMOTE terlebih dahulu mengelompokkan data minoritas menggunakan algoritma *K-Means Clustering*, lalu melakukan *oversampling* secara selektif di dalam masing-masing *cluster* [13]. Hal ini memungkinkan penciptaan sampel sintetis yang lebih representatif, terstruktur, dan kontekstual terhadap distribusi fitur, terutama di area batas keputusan yang kompleks [14].

Berdasarkan hal tersebut, penelitian ini bertujuan untuk mengembangkan model deteksi dini Diabetes Mellitus menggunakan pendekatan *hybrid K-Means* SMOTE dan *Logistic Regression*. Kontribusi utama dari penelitian ini terletak pada integrasi yang sistematis antara teknik penanganan ketidakseimbangan data menggunakan *K-Means* SMOTE dengan model klasifikasi *Logistic Regression* untuk menghasilkan sistem deteksi dini Diabetes Mellitus yang lebih akurat dan sensitif. Penelitian ini memberikan solusi terhadap tantangan klasik dalam klasifikasi penyakit, khususnya dalam meningkatkan kemampuan model dalam mengenali kasus positif dengan jumlah yang terbatas.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Pembangunan model deteksi dini Diabetes Mellitus berbasis *machine learning* dilakukan melalui serangkaian tahapan sistematis yang saling terintegrasi. Tahapan penelitian ini dirancang secara metodis agar pendekatan yang diusulkan dapat diterapkan secara terstruktur dan dapat direproduksi dalam konteks serupa [15]. Penelitian ini dilakukan melalui beberapa tahapan utama yang ditunjukkan pada Gambar 1.



Gambar 1. Alur Penelitian

Mengacu pada Gambar 1, berikut adalah penjelasan rinci dari masing-masing tahapan penelitian yang telah diimplementasikan.

a. Akuisisi Data

Tahap awal dari penelitian ini dimulai dengan pengumpulan dataset yang relevan untuk membangun model klasifikasi Diabetes Mellitus. Dataset yang digunakan diperoleh dari *platform* Kaggle dengan nama "Diabetes" dan file sumber bernama *diabetes.csv* (https://www.kaggle.com/datasets/johndasilva/diabetes) [16]. Dataset ini berisi rekam medis pasien yang mencakup sembilan atribut, yaitu: jumlah kehamilan (*Pregnancies*), kadar glukosa (*Glucose*), tekanan

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 3, April 2025 | Hal 219-227
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i3.502

darah (*BloodPressure*), ketebalan kulit (*SkinThickness*), kadar insulin (*Insulin*), indeks massa tubuh (*BMI*), riwayat genetik diabetes (*DiabetesPedigreeFunction*), usia (*Age*), serta label keluaran (*Outcome*), di mana nilai 1 menunjukkan pasien menderita diabetes dan nilai 0 menunjukkan sebaliknya. Dataset terdiri dari 2000 baris data dan dipilih karena memiliki fitur yang representatif serta kualitas data yang memadai untuk mendukung proses analisis dan prediksi dalam konteks penelitian ini.

b. Pra-pemrosesan Data
Pra-pemrosesan dilakukan untuk menjamin kualitas dan konsistensi data sebelum digunakan dalam pelatihan model [17]. Beberapa fitur seperti *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, dan *BMI* memiliki nilai nol yang tidak valid secara medis, sehingga nilai-nilai tersebut dianggap sebagai data hilang dan diimputasi menggunakan metode median untuk menjaga kestabilan terhadap outlier. Selanjutnya, seluruh fitur numerik dinormalisasi menggunakan *StandardScaler* agar memiliki distribusi dengan rata-rata nol dan standar deviasi satu, sehingga meminimalkan dominasi skala fitur dan meningkatkan efektivitas algoritma klasifikasi berbasis linier seperti *Logistic Regression*. Selanjutnya, dataset dibagi menjadi dua bagian: 80% untuk data latih dan 20% untuk data uji, menggunakan metode *stratified split* untuk menjaga proporsi kelas. Rasio ini dipilih karena umum digunakan dalam praktik pembelajaran mesin dan memberikan keseimbangan antara ukuran pelatihan dan evaluasi model secara adil [18].

c. Penanganan Imbalanced Data Menggunakan *K-Means* SMOTE
Dataset yang digunakan memiliki ketidakseimbangan kelas, di mana jumlah sampel penderita diabetes (*Outcome* = 1) jauh lebih sedikit dibandingkan non-diabetes. Kondisi ini dapat menyebabkan model terlalu bias terhadap kelas mayoritas, sehingga banyak kasus positif yang tidak terdeteksi (*false negative* tinggi). Untuk mengatasi hal ini, digunakan teknik *K-Means* SMOTE, yaitu varian dari metode *Synthetic Minority Oversampling Technique* (SMOTE). Berbeda dengan SMOTE standar, *K-Means* SMOTE terlebih dahulu mengelompokkan data minoritas menggunakan algoritma *K-Means Clustering*, lalu melakukan *oversampling* secara selektif di dalam masing-masing *cluster* [13]. Pendekatan ini memungkinkan penciptaan sampel sintetis yang lebih representatif terhadap distribusi data minoritas [19]. Teknik ini juga efektif dalam menghindari *oversampling* di daerah *noise* dan membantu mempertajam batas keputusan antar kelas.

d. Pembangunan Model Klasifikasi Menggunakan *Logistic Regression*
Model klasifikasi yang diterapkan dalam penelitian ini adalah *Logistic Regression*, salah satu algoritma *machine learning* klasik yang digunakan untuk memprediksi probabilitas kejadian dari dua kelas, yaitu penderita dan non-diabetes. Model ini membangun hubungan antara fitur prediktor dan kelas target melalui fungsi logistik (*sigmoid*), yang mengubah hasil linier menjadi nilai probabilitas antara 0 dan 1 [20]. *Logistic Regression* dipilih karena efisien secara komputasi, memberikan *output* berbasis probabilitas yang sesuai untuk analisis risiko medis, serta mudah diinterpretasikan, di mana setiap koefisien dapat ditafsirkan sebagai pengaruh relatif dari fitur terhadap probabilitas kelas minoritas [21]. Proses pembangunan model dilakukan dengan melatih *Logistic Regression* pada data latih yang telah dipra-pemrosesan, baik dalam kondisi asli maupun setelah dilakukan *oversampling* dengan *K-Means* SMOTE.

e. Evaluasi Kinerja Model
Evaluasi dilakukan untuk mengukur kemampuan model dalam membedakan kelas penderita dan non-diabetes. Metrik yang digunakan meliputi *confusion matrix*, *precision*, *recall* (sensitivitas), *F1-score*, serta *ROC curve* dan *AUC score*. *Confusion matrix* memberikan rincian prediksi benar dan salah dari masing-masing kelas, sedangkan *precision* dan *recall* mengukur ketepatan serta kelengkapan deteksi kasus positif. *F1-score* merepresentasikan keseimbangan antara keduanya [22]. *ROC curve* menunjukkan performa model pada berbagai ambang klasifikasi, dan *AUC score* mencerminkan kemampuan diskriminatif antar kelas [23]. Selain itu, evaluasi dilakukan dengan perbandingan antara tiga model klasifikasi, yaitu: *Logistic Regression* tanpa *oversampling*, dengan SMOTE standar, dan dengan *K-Means* SMOTE. Perbandingan ini bertujuan untuk mengetahui pengaruh teknik *oversampling* terhadap peningkatan performa model, khususnya dalam mendeteksi kasus positif yang jumlahnya lebih sedikit.

2.2 Metode *K-Means* SMOTE

K-Means SMOTE merupakan pengembangan dari teknik *oversampling* klasik SMOTE (*Synthetic Minority Oversampling Technique*) yang bertujuan untuk mengatasi ketidakseimbangan kelas pada data pelatihan [12]. Berbeda dari SMOTE standar yang menghasilkan sampel sintetis dengan melakukan interpolasi acak antara titik minoritas dan tetangga terdekatnya, *K-Means* SMOTE terlebih dahulu mengelompokkan data menggunakan algoritma *K-Means Clustering*, lalu melakukan *oversampling* secara selektif di dalam masing-masing *cluster* [13]. Pendekatan ini memungkinkan proses sintesis sampel dilakukan pada area yang representatif dan padat dari distribusi data minoritas, sehingga menghindari penciptaan data sintetis di area *noise* atau *outlier* [24].

Secara umum, proses *K-Means* SMOTE terdiri dari tiga tahap utama: (1) menerapkan *K-Means* pada seluruh data pelatihan untuk membentuk *k cluster*; (2) mengidentifikasi *cluster* yang memiliki proporsi tinggi dari kelas minoritas; dan (3) menerapkan proses *oversampling* dalam *cluster* tersebut menggunakan metode interpolasi [13]. Proses interpolasi ini mengikuti prinsip yang sama dengan SMOTE, yaitu dengan membentuk sampel sintetis x_{new} berdasarkan persamaan (1).

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 3, April 2025 | Hal 219-227
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i3.502

$$x_{new} = x_i + \delta \times (x_{zi} - x_i) \quad (1)$$

di mana x_i adalah sampel minoritas asli, x_{zi} adalah salah satu tetangganya dalam *cluster*, dan $\delta \in [0,1]$ adalah bilangan acak. Persamaan ini menghasilkan sampel baru yang berada pada garis antara dua titik minoritas, sehingga tetap menjaga karakteristik lokal dari kelas minoritas.

2.3 Metode *Logistic Regression*

Logistic Regression merupakan salah satu algoritma *machine learning* klasik yang umum digunakan dalam masalah klasifikasi biner [11]. Berbeda dengan regresi linier yang menghasilkan nilai kontinu, *Logistic Regression* memodelkan hubungan antara variabel independen dan probabilitas kejadian suatu kelas target melalui fungsi logistik (*sigmoid*) [21]. Fungsi ini memetakan hasil dari kombinasi linier fitur menjadi nilai probabilitas antara 0 dan 1. Model ini didasarkan pada persamaan (2).

$$P(y = 1|x) = \frac{1}{1+e^{(\beta_0+\beta_1 x_1+\beta_1 x_1+\cdots+\beta_n x_n)}} \quad (2)$$

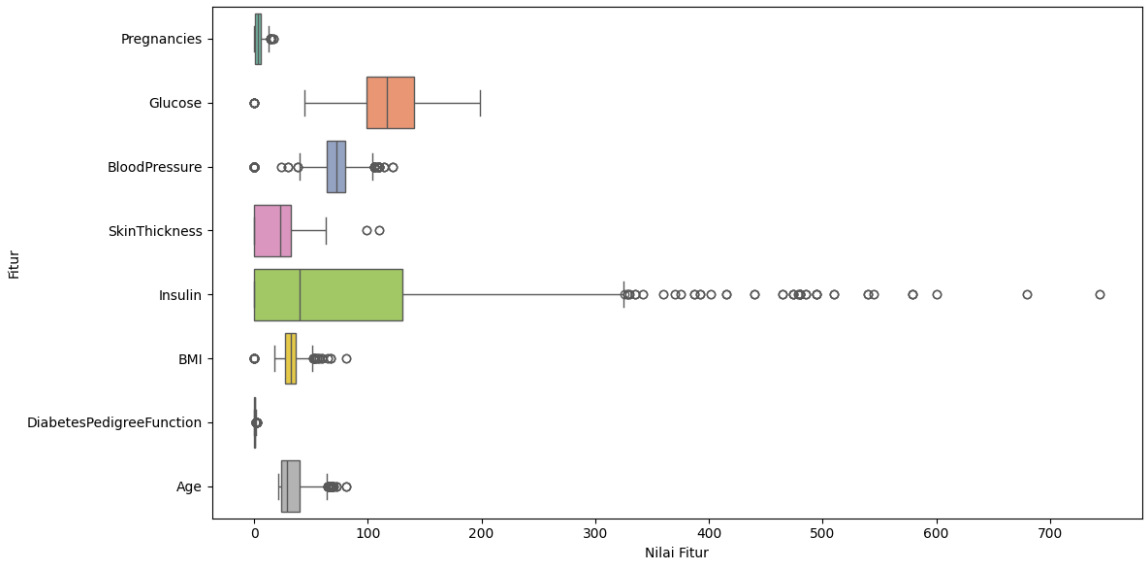
di mana $P(y = 1|x)$ adalah probabilitas suatu data termasuk ke dalam kelas 1 (positif), β_0 adalah bias/intersep, dan β_i adalah koefisien regresi dari masing-masing fitur x_i. Model ini dilatih dengan meminimalkan fungsi *log-loss* atau *binary cross-entropy* untuk mendapatkan parameter terbaik.

Keunggulan utama dari *Logistic Regression* terletak pada sifatnya yang interpretable, di mana nilai koefisien dapat ditafsirkan sebagai besarnya pengaruh masing-masing fitur terhadap kemungkinan klasifikasi positif [25]. Selain itu, algoritma ini juga efisien secara komputasi, tidak memerlukan asumsi distribusi normal pada fitur, dan dapat dengan mudah diperluas untuk klasifikasi multikelas menggunakan pendekatan *one-vs-rest* atau *Multinomial Logistic Regression*.

3. HASIL DAN PEMBAHASAN

Pengembangan model prediksi penyakit Diabetes Mellitus berbasis pendekatan *hybrid K-Means* SMOTE dan *Logistic Regression* diawali dengan penyiapan dataset yang digunakan dalam proses pelatihan dan pengujian. Dataset yang digunakan dalam penelitian ini berasal dari *platform* Kaggle dengan nama "Diabetes" (https://www.kaggle.com/datasets/johndasilva/diabetes) [16]. Dataset ini terdiri dari 2.000 data dan mencakup 9 vitur, yaitu: jumlah kehamilan (*Pregnancies*), kadar glukosa (*Glucose*), tekanan darah (*BloodPressure*), ketebalan lipatan kulit (*SkinThickness*), kadar insulin (*Insulin*), indeks massa tubuh (*BMI*), skor riwayat genetik diabetes (*DiabetesPedigreeFunction*), usia (*Age*), serta label diagnosis (*Outcome*), dengan nilai 1 menunjukkan penderita diabetes dan 0 menunjukkan non-diabetes.

Sebelum dilakukan pra-pemrosesan, eksplorasi awal terhadap karakteristik fitur numerik dilakukan untuk mengidentifikasi *outlier* dan sebaran nilai. Visualisasi *boxplot* terhadap seluruh fitur disajikan pada Gambar 2.



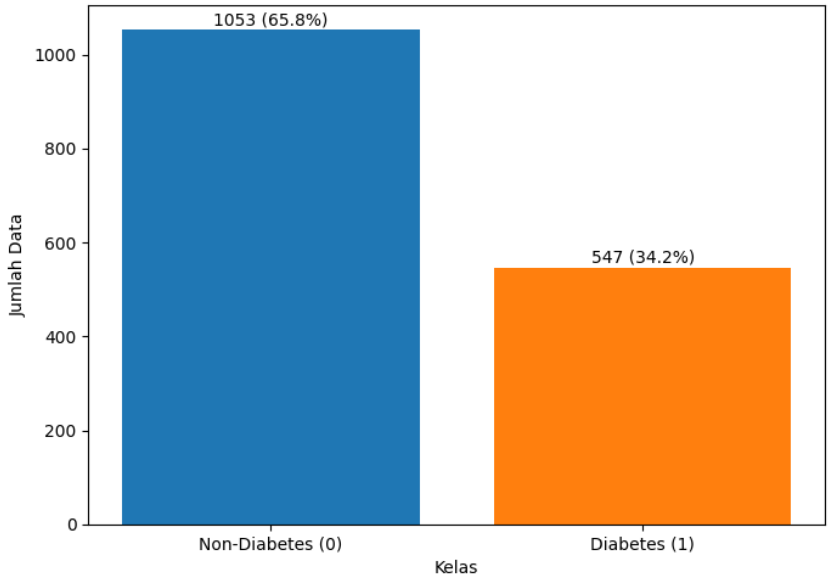
Gambar 2. *Boxplot* Seluruh Fitur Pada Dataset

Gambar 2 memperlihatkan visualisasi *boxplot* dari seluruh fitur numerik dalam dataset. Terlihat adanya sebaran nilai yang tidak merata, dengan keberadaan nilai ekstrem dan nol yang tidak valid secara medis, khususnya pada fitur *Insulin*, *Glucose*, dan *SkinThickness*. *Outlier* juga tampak jelas pada beberapa fitur lain seperti *BMI* dan *DiabetesPedigreeFunction*, yang mengindikasikan distribusi data yang tidak normal dan cenderung miring. Sebagai bagian dari tahapan pra-pemrosesan, kondisi ini diatasi dengan menggunakan teknik imputasi median untuk menggantikan nilai-nilai nol yang tidak valid, karena metode median lebih *robust* terhadap *outlier* dibandingkan rata-rata.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 3, April 2025 | Hal 219-227
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i3.502

Selain itu, perbedaan skala yang signifikan antar fitur menjadi alasan diterapkannya normalisasi menggunakan *StandardScaler*, agar setiap fitur memiliki kontribusi yang seimbang dalam pelatihan model.

Langkah selanjutnya adalah menganalisis distribusi data target untuk memastikan apakah data berada dalam kondisi seimbang atau tidak. Analisis ini penting untuk memahami proporsi antara kelas positif dan negatif dalam variabel *Outcome*, yang akan memengaruhi strategi pemodelan yang digunakan. Visualisasi distribusi data target pada dataset yang digunakan ditampilkan pada Gambar 3.



Gambar 3. Distribusi Data Target Sebelum Dilakukan *Oversampling*

Gambar 3 menampilkan distribusi kelas target sebelum dilakukan Teknik *oversampling*, di mana kelas 0 (non-diabetes) mendominasi sebesar 65,81% dari total 1.600 data, sementara kelas 1 (diabetes) hanya sebesar 34,19%. Ketidakseimbangan ini ditangani dengan menerapkan teknik *K-Means* SMOTE pada data pelatihan. Setelah melalui proses imputasi dan normalisasi, algoritma *K-Means* digunakan untuk mengelompokkan data minoritas ke dalam beberapa *cluster*. Selanjutnya, sampel sintetis dibangkitkan melalui interpolasi linier antar titik dalam setiap *cluster*. Dengan pendekatan ini, *K-Means* SMOTE mampu menghasilkan data sintetis yang lebih representatif terhadap distribusi lokal kelas minoritas dan menghindari pembentukan sampel pada area yang berpotensi *noise*. Untuk memberikan gambaran yang lebih jelas mengenai cara kerja *K-Means* SMOTE, berikut disajikan ilustrasi perhitungan sederhana menggunakan dua fitur dari dataset Diabetes, yaitu *Glucose* dan *BMI*. Dalam contoh ini, diasumsikan bahwa kita hanya fokus pada sampel dari kelas minoritas (penderita diabetes), seperti yang ditunjukkan pada Tabel 1.

Tabel 1. Sampel Data Kelas Minoritas

Sample	***Glucose***	***BMI***
A	140	35.0
B	148	36.5
C	135	34.2
D	160	38.0

Tabel 1 menunjukkan empat sampel dari kelas minoritas (penderita diabetes) dengan dua fitur utama yang digunakan dalam ilustrasi, yaitu *Glucose* dan *BMI*. Langkah pertama dalam *K-Means* SMOTE adalah melakukan *clustering* terhadap data minoritas menggunakan algoritma *K-Means*. Misalnya, jika jumlah cluster ditentukan sebanyak 2 (k = 2), maka hasil klasterisasi dapat membagi data menjadi dua kelompok sebagai berikut:

Cluster 1: Sample A, B
Cluster 2: Sample C, D

Setelah *cluster* terbentuk, proses selanjutnya adalah menghasilkan data sintetis melalui interpolasi dalam masing-masing *cluster*. Di *cluster* 1, interpolasi dilakukan antara sample A dan B:

A = [140, 35.0]
B = [148, 36.5]
δ = 0.4

Berdasarkan persamaan (1) maka nilai $Synthetic_1$ adalah:

$$Synthetic_1 = A + \delta \times (B - A) = [140{,}35.0] + 0.4 \times ([8{,}1.5]) = [143.2{,}35.6]$$

Sedangkan di *cluster* 2, interpolasi dilakukan antara sampel C dan D:

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 3, April 2025 | Hal 219-227
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i3.502

C = [135, 34.2]
D = [160, 38.0]
δ = 0.6

Berdasarkan persamaan (1) maka nilai $Synthetic_2$ adalah:

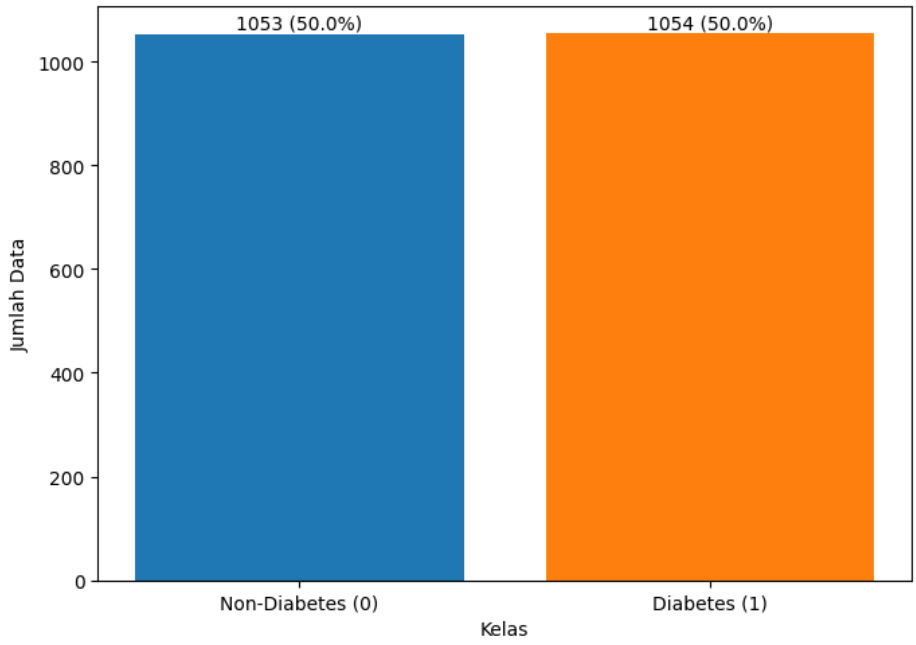
$$Synthetic_2 = C + \delta \times (D - C) = [135,34.2] + 0.6 \times ([25,3.8]) = [150.0,36.48]$$

Dengan demikian, dua sampel sintetis berhasil dibentuk dan ditambahkan ke dalam dataset sebagai representasi baru dari kelas minoritas yang ditampilkan Tabel 2.

Tabel 2. Hasil Interpolasi Sampel Sintetis

Synthetic Sample	***Glucose***	***BMI***
S1	143.2	35.6
S2	150.0	36.48

Tabel 2 menyajikan hasil perhitungan interpolasi linier dalam masing-masing *cluster* yang terbentuk. Contoh ini menggambarkan bagaimana *K-Means* SMOTE bekerja dengan lebih selektif dan terstruktur dibandingkan SMOTE standar, karena proses *oversampling* dilakukan dalam wilayah *cluster* yang lebih homogen, sehingga meminimalkan risiko pembentukan data sintetis di area *outlier* atau *noise*. Hasil distribusi kelas target dengan teknik *oversampling* menggunakan *K-Means* SMOTE pada dataset yang digunakan ditampilkan pada Gambar 4.

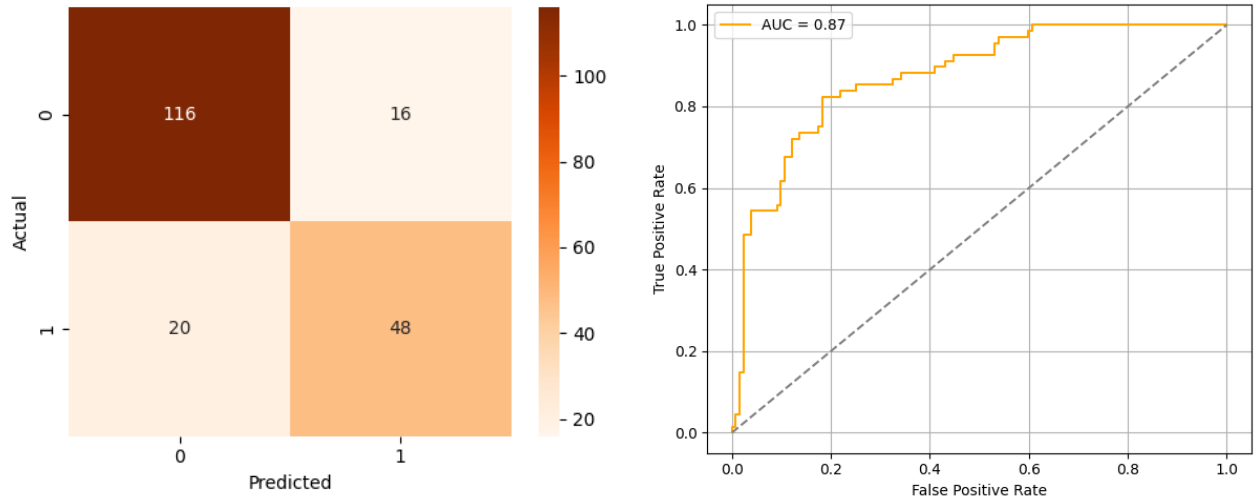


Gambar 4. Distribusi Data Target Setelah Dilakukan *Oversampling* Menggunakan *K-Means* SMOTE

Gambar 4 menunjukkan distribusi data target setelah dilakukan proses *oversampling* menggunakan teknik *K-Means* SMOTE. Terlihat bahwa jumlah data antara kelas non-diabetes (0) dan diabetes (1) telah berhasil diseimbangkan, masing-masing dengan proporsi sekitar 50%.

Proses selanjutnya adalah membangun model klasifikasi menggunakan algoritma *Logistic Regression*, yang diterapkan untuk proses pelatihan dan pengujian. Dataset dibagi menjadi data pelatihan dan data pengujian dengan rasio 80:20 menggunakan teknik *stratified sampling* untuk memastikan bahwa proporsi kelas target tetap seimbang pada kedua subset data. *Logistic Regression* bekerja dengan memetakan kombinasi linier dari fitur-fitur input ke dalam fungsi sigmoid, yang menghasilkan nilai probabilitas antara 0 dan 1. Kemampuan ini menjadikan *Logistic Regression* sangat sesuai untuk kasus klasifikasi biner, seperti dalam studi deteksi dini penyakit Diabetes Mellitus. Proses pelatihan dilakukan menggunakan fungsi *fit*, di mana model mempelajari hubungan antara fitur dan kelas target. Setelah pelatihan selesai, model menghasilkan prediksi kelas dan probabilitasnya melalui fungsi *predict* dan *predict_proba*. Kinerja model kemudian dievaluasi menggunakan sejumlah metrik, antara lain *confusion matrix*, *classification report*, dan *ROC-AUC score*, yang digunakan untuk menilai kemampuan model dalam membedakan antara kelas positif dan negatif. Hasil evaluasi berupa *confusion matrix* dan kurva ROC untuk model *Logistic Regression* menggunakan *K-Means* SMOTE yang ditampilkan pada Gambar 5.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 3, April 2025 | Hal 219-227
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i3.502



(a) (b)

Gambar 5. (a) *Confusion Matrix* dan (b) *ROC curve* Dari Model *Logistic Regression* Dengan *K-Means* SMOTE

Gambar 5 menyajikan hasil *Confusion Matrix* dan *ROC curve* dari model *Logistic Regression* tanpa *oversampling* serta model dengan *oversampling* menggunakan *K-Means* SMOTE. Pada Gambar 5(a) menunjukkan bahwa jenis kesalahan prediksi yang paling dominan pada model tanpa *oversampling* adalah *false negative*, yaitu ketika pasien yang sebenarnya menderita diabetes tidak terdeteksi oleh model. Hal ini mengindikasikan bahwa model kurang sensitif terhadap kelas minoritas. Setelah diterapkan teknik SMOTE dan *K-Means* SMOTE, jumlah *false negative* berkurang secara signifikan, yang ditunjukkan oleh peningkatan nilai recall pada kelas diabetes. Namun, pada model dengan SMOTE standar, terdapat sedikit peningkatan *false positive* yang menyebabkan penurunan *precision*. Sementara itu, model dengan K-Means SMOTE menunjukkan keseimbangan yang lebih baik, dengan penurunan *false negative* yang tetap terjaga tanpa lonjakan signifikan pada *false positive*. Hal ini penting, karena dalam konteks medis, *false negative* memiliki dampak yang lebih serius dibanding *false positive*, karena dapat menyebabkan keterlambatan diagnosis dan pengobatan bagi pasien yang seharusnya segera ditangani.

Tahap selanjutnya adalah melakukan evaluasi dan perbandingan kinerja dari ketiga model yang digunakan, yaitu *Logistic Regression* tanpa *oversampling*, dengan penerapan SMOTE standar, dan dengan *K-Means* SMOTE. Untuk menilai performa masing-masing model, digunakan sejumlah metrik evaluasi, antara lain *precision*, *recall*, *F1-score*, *accuracy*, dan ROC-AUC *score*. Hasil perbandingan lengkap antar model berdasarkan metrik tersebut ditampilkan pada Tabel 2.

Tabel 2. Ringkasan Performa Evaluasi Pada Masing-Masing Model

Model	Kelas	*Precision*	*Recall*	*F1-Score*	*Accuracy*	ROC-AUC *Score*
Logistic Regression	Non-Diabetes	81,38%	89,73%	85,35%	79,75%	85,49%
	Diabetes	75,45%	60,58%	67,21%		
Logistic Regression + SMOTE	Non-Diabetes	88,52%	81,82%	85,04%	81,00%	86,66%
	Diabetes	69,23%	79,41%	73,97%		
Logistic Regression + *K-Means* SMOTE	Non-Diabetes	85,29%	87,88%	86,57%	82,00%	87,48%
	Diabetes	75,00%	70,59%	72,73%		

Berdasarkan hasil evaluasi pada Tabel 2, terlihat bahwa model *Logistic Regression* tanpa *oversampling* memiliki performa yang cukup baik pada kelas mayoritas yaitu non-diabetes, dengan nilai *F1-score* sebesar 85,35%. Namun, model ini menunjukkan kelemahan dalam mendeteksi kelas minoritas, yaitu Diabetes, yang ditandai dengan rendahnya *recall* sebesar 60,58% dan *F1-score* sebesar 67,21%. Hal ini mencerminkan adanya bias model terhadap kelas mayoritas, yang merupakan permasalahan umum pada data yang tidak seimbang.

Setelah diterapkan teknik *oversampling* menggunakan SMOTE standar, terjadi peningkatan signifikan pada performa kelas Diabetes. *Recall* meningkat dari 60,58% menjadi 79,41%, sedangkan *F1-score* naik dari 67,21% menjadi 73,97%. Peningkatan ini menunjukkan bahwa SMOTE berhasil meningkatkan sensitivitas model terhadap kelas minoritas, meskipun disertai sedikit penurunan nilai *precision*. Akurasi keseluruhan juga meningkat menjadi 81,00%, dan skor ROC-AUC naik menjadi 86,66%.

Model *Logistic Regression* dengan *K-Means* SMOTE menunjukkan kinerja yang paling seimbang. Model ini mempertahankan *precision* dan *recall* yang kompetitif pada kedua kelas, dengan *F1-score* sebesar 86,57% untuk kelas

non-diabetes dan 72,73% untuk kelas Diabetes. Akurasi model tercatat sebesar 82,00%, dan skor ROC-AUC mencapai nilai tertinggi yaitu 87,48%. Hasil ini menunjukkan bahwa *K-Means* SMOTE mampu menghasilkan data sintetis yang lebih representatif dengan mempertimbangkan struktur distribusi lokal melalui proses klasterisasi, sehingga meningkatkan kinerja model secara keseluruhan dalam menghadapi ketidakseimbangan data.

Secara keseluruhan, pendekatan *Logistic Regression* dengan *K-Means* SMOTE terbukti paling efektif dalam mengatasi ketidakseimbangan data dan meningkatkan generalisasi model dalam klasifikasi awal Diabetes Mellitus. Kinerja unggul ini diperoleh karena *K-Means* SMOTE tidak hanya menyeimbangkan jumlah data antar kelas, tetapi juga mempertimbangkan distribusi spasial data dengan mengelompokkan sampel minoritas sebelum menghasilkan data sintetis. Strategi ini menghasilkan sampel yang lebih representatif dan mampu mengurangi risiko *overfitting* pada area yang berpotensi *noise*. Namun demikian, efektivitas *K-Means* SMOTE sangat dipengaruhi oleh pemilihan jumlah *cluster* (nilai *k*), yang bersifat sensitif terhadap struktur data. Nilai *k* yang terlalu kecil dapat menyebabkan *oversampling* yang kurang bervariasi dan berisiko *underfitting*, sementara nilai *k* yang terlalu besar justru dapat memicu *overfitting* akibat terbentuknya *cluster* kecil yang mengandung *noise*. Hal ini dapat berdampak negatif pada kualitas generalisasi model, terutama jika data sintetis yang dihasilkan terlalu spesifik terhadap *cluster* yang tidak representatif. Oleh karena itu, diperlukan pemilihan parameter *k* yang optimal, misalnya melalui metode *elbow* atau validasi silang, serta eksplorasi lebih lanjut terhadap integrasi teknik ini dengan pendekatan lainnya seperti *ensemble* atau *boosting* untuk meningkatkan akurasi dan stabilitas pada data medis yang tidak seimbang.

4. KESIMPULAN

Penelitian ini berhasil membangun model klasifikasi untuk deteksi dini Diabetes Mellitus dengan mengintegrasikan *Logistic Regression* dan *K-Means* SMOTE sebagai solusi terhadap permasalahan data yang tidak seimbang. Dataset yang digunakan memiliki distribusi kelas yang timpang, di mana jumlah data penderita diabetes jauh lebih sedikit dibandingkan non-diabetes. Penerapan *K-Means* SMOTE memungkinkan pembangkitan data sintetis yang tidak hanya menyeimbangkan jumlah antar kelas, tetapi juga mempertimbangkan struktur distribusi lokal kelas minoritas melalui proses klasterisasi. Hasil evaluasi menunjukkan bahwa pendekatan hybrid ini mampu meningkatkan performa model, khususnya dalam mendeteksi kelas minoritas (penderita diabetes), tanpa mengorbankan akurasi pada kelas mayoritas. Hasil evaluasi menunjukkan bahwa pendekatan hybrid ini mampu meningkatkan performa model, khususnya dalam mendeteksi kelas minoritas (penderita diabetes), tanpa mengorbankan akurasi pada kelas mayoritas. Dibandingkan dengan model *Logistic Regression* tanpa *oversampling* dan model dengan SMOTE standar, integrasi *K-Means* SMOTE mencatatkan *accuracy* tertinggi sebesar 82%, *F1-score* yang lebih seimbang antar kelas, serta nilai *ROC-AUC* tertinggi sebesar 87,48%. Hasil ini menegaskan efektivitas *K-Means* SMOTE dalam menghasilkan data sintetis yang representatif dan meningkatkan generalisasi model. Temuan ini memberikan dasar yang kuat bagi implementasi model klasifikasi yang lebih adil dan responsif terhadap kasus minoritas dalam sistem diagnosa awal berbasis *machine learning*. Meskipun demikian, metode ini memiliki tantangan dalam kompleksitas perhitungan dan pemilihan jumlah *cluster* yang optimal. Oleh karena itu, penelitian selanjutnya disarankan untuk mengeksplorasi strategi pemilihan parameter *k* yang lebih adaptif, serta mengintegrasikan pendekatan ini dengan teknik *ensemble* atau *boosting* guna meningkatkan akurasi dan stabilitas model secara keseluruhan.

REFERENCES

[1] N. Singh, A. Kumari, and L. Kishore, "New-insight Management Implications of Diabetic Autonomic Neuropathy: Future Perspectives," *Int. J. Res. Pharm. Allied Sci.*, vol. 3, no. 6, pp. 63–71, 2024, doi: 10.71431/IJRPAS.2025.4106.

[2] Reuters, "More than 800 million adults have diabetes globally, many untreated, study suggests," reuters.com. Accessed: Apr. 15, 2025. [Online]. Available: https://www.reuters.com/business/healthcare-pharmaceuticals/more-than-800-million-adults-have-diabetes-globally-many-untreated-study-2024-11-13

[3] A. Aminuddin, Yenny Sima, Nurril Cholifatul Izza, Nur Syamsi Norma Lalla, and Darmi Arda, "Edukasi Kesehatan Tentang Penyakit Diabetes Melitus bagi Masyarakat," *Abdimas Polsaka*, pp. 7–12, 2023, doi: 10.35816/abdimaspolsaka.v2i1.25.

[4] R. Rianto and P. I. Santosa, *Data Preparation untuk Machine Learning & Deep Learning*. Yogyakarta: Penerbit Andi, 2024.

[5] V. R. Konasani and S. Kadre, *Machine Learning and Deep Learning Using Python and TensorFlow*. New York: McGraw Hill LLC, 2021.

[6] L. Safitri and Z. Fatah, "Implementasi Prediksi Penyakit Diabetes Menggunakan Metode Decision Tree," *JUSIFOR J. Sist. Inf. dan Inform.*, vol. 2, no. 2, pp. 125–132, 2023, doi: 10.70609/jusifor.v3i2.5788 .

[7] A. W. Mucholladin, F. A. Bachtiar, and M. T. Furqon, "Klasifikasi Penyakit Diabetes menggunakan Metode Support Vector Machine," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 2, pp. 622–633, 2021.

[8] N. Maulidah, R. Supriyadi, D. Y. Utami, F. N. Hasan, A. Fauzi, and A. Christian, "Prediksi Penyakit Diabetes Melitus Menggunakan Metode Support Vector Machine dan Naive Bayes," *Indones. J. Softw. Eng.*, vol. 7, no. 1, pp. 63–68, 2021, doi: 10.31294/ijse.v7i1.10279.

[9] S. P. Nainggolan and A. Sinaga, "Comparative Analysis of Accuracy of Random Forest and Gradient Boosting Classifier Algorithm for Diabetes Classification," *Sebatik*, vol. 27, no. 1, pp. 97–102, 2023, doi: 10.46984/sebatik.v27i1.2157.

[10] A. P. Silalahi and H. G. Simanullang, "Supervised Learning Metode K-Nearest Neighbor Untuk Prediksi Diabetes Pada Wanita," *METHOMIKA J. Manaj. Inform. dan Komputerisasi Akunt.*, vol. 7, no. 1, pp. 144–149, 2023, doi: 10.46880/jmika.vol7no1.pp144-149.

[11] S. Sutarman, R. Siringoringo, D. Arisandi, E. Kurniawan, and E. B. Nababan, "Model Klasifikasi Dengan Logistic Regression

Dan Recursive Feature Elimination Pada Data Tidak Seimbang," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 4, pp. 735–742, 2024, doi: 10.25126/jtiik.1148198.

[12] C. Haryawan and Y. M. K. Ardhana, "Analisa Perbandingan Teknik Oversampling SMOTE Pada Imbalanced Data," *J. Inform. dan Rekayasa Elektron.*, vol. 6, no. 1, pp. 73–78, 2023, doi: 10.36595/jire.v6i1.834.

[13] N. Indrani *et al.*, "Classification of Natural Disaster Reports from Social Media using K-Means SMOTE and Multinomial Naïve Bayes," *J. Comput. Sci. Informatics Eng.*, vol. 7, no. 1, pp. 60–67, 2023, doi: 10.29303/jcosine.v7i1.503.

[14] C. V. Angkoso, M. A. N. Thrisna, B. D. Satoto, and A. Kusumaningsih, "Optimasi Klasifikasi Sentimen Menggunakan Random Forest dengan Preprocessing K-Means Clustering dan SMOTE," *JEPIN (Jurnal Edukasi dan Penelit. Inform.*, vol. 10, no. 3, pp. 389–400, 2024.

[15] R. I. Borman, F. Rossi, Y. Jusman, A. A. A. Rahni, S. D. Putra, and A. Herdiansah, "Identification of Herbal Leaf Types Based on Their Image Using First Order Feature Extraction and Multiclass SVM Algorithm," in *International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS)*, IEEE, 2021, pp. 12–17.

[16] J. Dasilva, "Diabetes Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/johndasilva/diabetes

[17] R. I. Borman, D. E. Kurniawan, Styawati, I. Ahmad, and D. Alita, "Classification of Maturity Levels of Palm Fresh Fruit Bunches Using the Linear Discriminant Analysis Algorithm," *AIP Conf. Proc.*, vol. 2665, no. 1, pp. 30023.1-30023.8, 2023, doi: 10.1063/5.0126513.

[18] A. Bisri and M. Man, "Machine Learning Algorithms Based on Sampling Techniques for Raisin Grains Classification," *Int. J. Informatics Vis.*, vol. 7, no. 1, pp. 7–14, 2023, doi: 10.30630/joiv.7.1.970.

[19] X. Zhu *et al.*, "An automatic identification method of imbalanced lithology based on Deep Forest and K-means SMOTE," *Geoenergy Sci. Eng.*, vol. 224, no. February, p. 211595, 2023, doi: 10.1016/j.geoen.2023.211595.

[20] W. F. Hidayat, T. Asra, and A. Setiadi, "Klasifikasi Penyakit Daun Kentang Menggunakan Model Logistic Regression," *Indones. J. Softw. Eng.*, vol. 8, no. 2, pp. 173–179, 2022.

[21] S. Suhliyyah, H. H. Handayani, and K. A. Baihaqi, "Implementasi Algoritma Logistic Regression Untuk Klasifikasi Penyakit Stroke," *Syntax J. Inform.*, vol. 12, no. 01, pp. 15–23, 2023.

[22] Z. Abidin, R. I. Borman, F. B. Ananda, P. Prasetyawan, F. Rossi, and Y. Jusman, "Classification of Indonesian Traditional Snacks Based on Image Using Convolutional Neural Network (CNN) Algorithm," in *International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS)*, IEEE, 2022, pp. 18–23.

[23] Y. Liu, Y. Li, and D. Xie, "Implications of imbalanced datasets for empirical ROC-AUC estimation in binary classification tasks," *J. Stat. Comput. Simul.*, vol. 94, no. 1, pp. 183–203, Jan. 2024, doi: 10.1080/00949655.2023.2238235.

[24] H. Hairani, "Peningkatan Kinerja Metode SVM Menggunakan Metode KNN Imputasi dan K-Means-SMOTE untuk Klasifikasi Kelulusan Mahasiswa Universitas Bumigora," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 4, pp. 713–718, 2021, doi: 10.25126/jtiik.2021843428.

[25] S. rahmah Jabir, H. Azis, D. Widyawatia, and A. U. Tenripada, "Prediksi Potensi Donatur Menggunakan Model Logistic Regression," *Indones. J. Data Sci.*, vol. 4, no. 1, pp. 31–37, 2023, doi: 10.56705/ijodas.v4i1.64.