Seminar Nasional Ilmu Teknik dan Aplikasi Industri (SINTA)

Homepage: sinta.eng.unila.ac.id

Penerapan teknik *oversampling* pada dataset indikator kesehatan diabetes menggunakan SMOTE untuk klasifikasi penderita diabetes

Tiya Muthia [a,1,*], Nurrahma Nurrahma [b], Nadia Julian Putri [aa]

[a] *Program Studi Teknik Elektro, Universitas Lampung, Jl. Prof. Dr. Ir. Sumantri Brojonegoro No.1, Kota Bandar Lampung, Lampung*
[b] *Program Studi Teknik Informatika, Universitas Lampung, Jl. Prof. Dr. Ir. Sumantri Brojonegoro No.1, Kota Bandar Lampung, Lampung*

I N F O R M A S I A R T I K E L

Riwayat artikel:
Diterima 19/11/2025
Direvisi 19/12/2025

Kata kunci:
diabetes,
klasifikasi,
machine learning,
SMOTE,
supervised learning,.

A B S T R A K

Perkembangan teknologi informasi telah mendorong penerapan *machine learning* dalam bidang kesehatan, khususnya untuk mendukung diagnosis dan klasifikasi penyakit secara otomatis. Salah satu tantangan utama dalam analisis data kesehatan adalah ketidakseimbangan kelas (*class imbalance*), yang dapat menurunkan performa model klasifikasi. Penelitian ini bertujuan untuk menerapkan teknik *oversampling* SMOTE (*Synthetic Minority Over-sampling Technique*) guna menyeimbangkan distribusi data pada dataset indikator kesehatan diabetes dari *Behavioral Risk Factor Surveillance System* (BRFSS) 2021, serta membandingkan kinerja beberapa algoritma *supervised learning* dalam mengklasifikasikan penderita diabetes. *Dataset* terdiri dari 21 fitur dengan variabel target biner yang menunjukkan status diagnosis diabetes. Tahapan penelitian meliputi studi literatur, pengumpulan dan pra-pemrosesan data (pembersihan, penerapan SMOTE, pembagian dataset, dan normalisasi fitur), pelatihan model, serta evaluasi kinerja menggunakan metrik akurasi. Enam algoritma yang diuji meliputi *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), *Decision Tree* (DT), *Random Forest* (RF), *Logistic Regression* (LR), dan *Naive Bayes* (NB). Hasil pengujian menunjukkan bahwa seluruh model memiliki akurasi di atas 67%, dengan *Random Forest* memberikan performa terbaik, yaitu akurasi pelatihan sebesar 91,96% dan akurasi pengujian sebesar 81,13%. Penerapan SMOTE terbukti efektif dalam meningkatkan proporsionalitas data dan kinerja model klasifikasi. Kombinasi antara penyeimbangan data menggunakan SMOTE dan penggunaan model *ensemble* seperti *Random Forest* menghasilkan sistem klasifikasi diabetes yang lebih akurat dan andal, sehingga berpotensi mendukung pengembangan sistem deteksi dini berbasis data kesehatan di masa mendatang.

1. Pendahuluan

Perkembangan teknologi informasi telah memberikan dampak yang signifikan terhadap transformasi di berbagai sektor, termasuk bidang kesehatan. Salah satu bentuk penerapan teknologi tersebut adalah pemanfaatan analisis data berbasis *machine learning*, yang memungkinkan komputer untuk belajar dari data berlabel dan melakukan prediksi secara otomatis. Dalam konteks analisis data kesehatan, pendekatan *supervised learning* menjadi salah satu metode yang banyak digunakan karena kemampuannya dalam mengenali pola dan mengklasifikasikan data baru dengan tingkat akurasi yang tinggi. Berbagai penelitian telah menunjukkan efektivitas *machine learning* dalam membantu proses diagnosis dan klasifikasi penyakit

* Penulis korespondensi.
E-mail: tiyamuthia@eng.unila.ac.id

secara otomatis [1], antara lain pada kasus gangguan mental [2], [3], penyakit jantung [4], [5], penyakit Covid-19 [6], [7], serta penyakit diabetes [8], [9]. Di antara berbagai penyakit tersebut, diabetes menjadi salah satu fokus utama penelitian karena prevalensinya yang terus meningkat dan dampaknya yang signifikan terhadap kesehatan masyarakat.

Diabetes merupakan salah satu penyakit tidak menular yang menjadi tantangan serius bagi sistem kesehatan global, termasuk di Indonesia. Penyakit ini dapat menimbulkan berbagai komplikasi kronis yang memengaruhi kualitas hidup dan meningkatkan beban pembiayaan kesehatan. Berdasarkan data dari *International Diabetes Federation* (IDF), diperkirakan terdapat sekitar 589 juta orang dewasa berusia 20–79 tahun yang hidup dengan diabetes pada tahun 2025, dan jumlah tersebut diproyeksikan meningkat hingga mencapai 853 juta orang pada tahun 2050 [1]. Di Indonesia, hasil *Riset Kesehatan Dasar* (Riskesdas) tahun 2018 menunjukkan bahwa prevalensi diabetes meningkat dari 8,5% pada tahun 2011 menjadi 10,9% pada tahun 2015 [10].

Fakta ini menegaskan pentingnya pengembangan sistem deteksi dini berbasis data untuk mendukung diagnosis diabetes secara lebih cepat, akurat, dan efisien. Dalam pengembangan sistem klasifikasi medis berbasis *machine learning*, berbagai algoritma telah digunakan, seperti *Support Vector Machine* (SVM), *K-Nearest Neighbor* (KNN), *Decision Tree* (DT), *Random Forest* (RF), *Logistic Regression* (LR), dan *Naive Bayes* (NB) [8]. Masing-masing algoritma memiliki karakteristik serta keunggulan dan keterbatasannya sendiri. Sebagai contoh, SVM memiliki kemampuan yang baik dalam menangani data berdimensi tinggi namun sensitif terhadap pemilihan parameter dan fungsi kernel. Sementara itu, KNN bergantung pada distribusi data di sekitar sampel, sehingga performanya dapat menurun jika dataset tidak seimbang. *Decision Tree* unggul dalam hal interpretabilitas dan kemampuannya menangani data heterogen, sedangkan *Naive Bayes* bekerja efektif pada data yang diasumsikan memiliki independensi antar fitur [11].

Kinerja model *machine learning* tidak hanya ditentukan oleh algoritma yang digunakan, tetapi juga sangat dipengaruhi oleh karakteristik dataset, seperti jumlah fitur, distribusi kelas, serta tingkat *noise* atau keberadaan data anomali [12]. Salah satu tantangan yang sering ditemui dalam data medis adalah ketidakseimbangan kelas (*class imbalance*), yaitu kondisi di mana jumlah data antara kelas penderita dan non-penderita tidak seimbang secara signifikan. Ketidakseimbangan ini dapat menyebabkan model cenderung bias terhadap kelas mayoritas, sehingga kemampuan deteksi terhadap kelas minoritas menjadi kurang optimal.

Untuk mengatasi permasalahan tersebut, berbagai metode *oversampling* telah dikembangkan dengan tujuan menyeimbangkan distribusi kelas dalam *dataset*. Salah satu teknik *intelligent oversampling* yang paling populer dan banyak digunakan adalah SMOTE (Synthetic Minority Over-sampling Technique), yang diperkenalkan oleh Chawla et al. [13]. Teknik ini bekerja dengan menghasilkan sampel sintetis baru di ruang fitur (*feature space*) melalui interpolasi antara satu sampel kelas minoritas dengan *k nearest neighbors*-nya [14]. Pendekatan ini tidak sekadar menggandakan data yang sudah ada, melainkan menciptakan variasi baru yang lebih representatif, sehingga membantu model *machine learning* mengenali pola kelas minoritas secara lebih baik. Dengan demikian, penerapan SMOTE diharapkan meningkatkan performa model pada *dataset* yang tidak seimbang dan memperbaiki akurasi klasifikasi terhadap kelas minoritas.

Berdasarkan latar belakang tersebut, penelitian ini berfokus pada penerapan teknik *oversampling* SMOTE pada *dataset* indikator kesehatan diabetes. Penelitian ini juga bertujuan untuk mengevaluasi dan membandingkan kinerja beberapa model *supervised learning* dalam mengklasifikasikan penderita diabetes berdasarkan indikator kesehatan. Diharapkan hasil penelitian ini dapat memberikan kontribusi terhadap peningkatan akurasi sistem deteksi dini diabetes berbasis *machine learning*, serta menjadi referensi dalam penerapan teknik *data balancing* pada analisis data kesehatan.

2. Metodologi

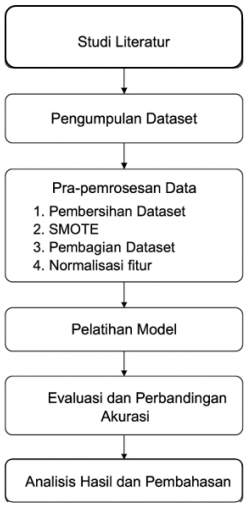
2.1. Alat dan Bahan

Perangkat dan spesifikasi yang digunakan untuk menunjang penelitian ini adalah Laptop Intel Core i3 10th Gen, RAM 8GB, OS Windows 11, Google Colab Pro, Python 3.11.13, dan Dataset dari Kaggle.

2.2. Prosedur Penelitian

Tahapan penelitian ini dilakukan secara sistematis melalui beberapa langkah utama yang disajikan dalam Gambar 1. Setiap tahapan dirancang untuk memastikan proses analisis berjalan terstruktur dan hasil yang diperoleh dapat dipertanggungjawabkan secara ilmiah. Adapun tahapan penelitian ini meliputi:

a) Studi Literatur.
   Tahapan awal penelitian dilakukan melalui studi literatur yang bertujuan untuk memperoleh pemahaman mendalam mengenai teori dasar yang berkaitan dengan penyakit diabetes, indikator kesehatan yang relevan dalam proses diagnosis, serta algoritma *supervised learning* yang umum digunakan untuk klasifikasi data medis.



Gambar 1. Diagram alir tahapan penelitian

b) Pengumpulan Dataset.
Dataset yang digunakan dalam penelitian ini berasal dari *Behavioral Risk Factor Surveillance System* (BRFSS), yaitu survei tahunan berbasis telepon yang diselenggarakan oleh *Centers for Disease Control and Prevention* (CDC), Amerika Serikat. Survei ini mengumpulkan data mengenai perilaku berisiko, kondisi kesehatan kronis, serta pemanfaatan layanan pencegahan pada individu dewasa berusia 18 tahun ke atas. Dalam penelitian ini digunakan data BRFSS tahun 2021 yang diperoleh dari platform Kaggle dalam format *.csv*. Dataset tersebut berisi berbagai fitur yang merepresentasikan pertanyaan survei dan variabel turunan hasil pengolahan respons peserta.

c) Pra-pemrosesan Data.
Tahapan pra-pemrosesan data bertujuan untuk mempersiapkan *dataset* agar layak digunakan dalam proses pelatihan model *machine learning*. Proses ini meliputi beberapa langkah, yaitu:
- Pembersihan data,
- Penyeimbangan kelas menggunakan SMOTE (*Synthetic Minority Over-sampling Technique*).
- Pembagian *Dataset Training* dan *Dataset Testing*
- Normalisasi fitur

d) Pelatihan Model.
Tahap ini bertujuan untuk melatih beberapa algoritma *supervised learning* menggunakan data latih yang telah melalui tahap pra-pemrosesan dan penyeimbangan kelas. Beberapa algoritma yang digunakan antara lain *Support Vector Machine* (SVM), *K-Nearest Neighbor* (KNN), *Decision Tree* (DT), *Random Forest* (RF), *Logistic Regression* (LR), dan *Naive Bayes* (NB). Setiap model dilatih untuk mengenali pola yang membedakan antara individu penderita dan non-penderita diabetes.

e) Evaluasi dan Perbandingan Akurasi.
Kinerja setiap model dievaluasi menggunakan metrik akurasi (*accuracy*). Hasil evaluasi digunakan untuk membandingkan efektivitas antar algoritma dan menentukan model yang paling sesuai dalam mengklasifikasikan data indikator kesehatan diabetes.

f) Analisis Hasil dan Pembahasan.
Berdasarkan hasil evaluasi, dilakukan analisis mendalam terhadap performa masing-masing model. Tahap ini mencakup interpretasi hasil, pembahasan faktor yang memengaruhi kinerja model, serta analisis peran SMOTE dalam peningkatan akurasi klasifikasi. Hasil analisis ini menjadi dasar dalam penarikan kesimpulan dan penyusunan rekomendasi terkait algoritma klasifikasi yang paling efektif untuk dataset diabetes yang digunakan.

3. Hasil dan Pembahasan

3.1. Dataset

Dataset yang digunakan diperoleh dari platform dataset publik Kaggle dengan nama *file "diabetes_binary_health_indicators_BRFSS2021.csv"*. *Dataset* ini berisi hasil survei yang telah diolah dalam bentuk numerik dan kategorikal, mencakup berbagai indikator kesehatan dan perilaku yang berpotensi memengaruhi risiko diabetes. Variabel target pada dataset ini adalah "Diabetes_binary", yang bernilai 1 untuk individu dengan diagnosis diabetes dan 0 untuk individu yang tidak terdiagnosis diabetes.

Secara keseluruhan, dataset ini memiliki 21 fitur (variabel independen) yang merepresentasikan berbagai aspek Kesehatan. Keberagaman fitur ini memungkinkan eksplorasi hubungan yang kompleks antara faktor-faktor gaya hidup dan kemungkinan seseorang menderita diabetes.

Selain itu, dataset BRFSS2021 ini memiliki ketidakseimbangan kelas (*class imbalance*), di mana jumlah sampel non-diabetes jauh lebih banyak dibandingkan dengan sampel diabetes. Kondisi ini berpotensi menyebabkan *bias* dalam proses pelatihan model klasifikasi, karena model cenderung lebih akurat dalam memprediksi kelas mayoritas.

3.2. Pra-pemrosesan Data

Pembersihan data (*data cleaning*) untuk menghapus atau mengganti nilai kosong (*missing values*), serta mengidentifikasi dan menangani *outlier* yang dapat memengaruhi distribusi data. Selanjutnya, penyeimbangan kelas menggunakan SMOTE (*Synthetic Minority Over-sampling Technique*) dilakukan. Salah satu permasalahan utama pada dataset medis adalah ketidakseimbangan antara jumlah sampel kelas positif

(penderita diabetes) dan kelas negatif (non-penderita). Untuk mengatasi hal tersebut, penelitian ini menerapkan teknik *oversampling* berbasis SMOTE. Metode ini menghasilkan sampel sintetis baru dari kelas minoritas. Dengan menambahkan data sintetis tersebut, distribusi kelas menjadi lebih seimbang sehingga model *machine learning* dapat belajar pola kedua kelas secara proporsional dan meningkatkan kemampuan klasifikasi terhadap penderita diabetes. Kemudian *dataset* dibagi menjadi dua bagian, yaitu data latih (*training set*) dan data uji (*testing set*) dengan proporsi tertentu untuk menjamin keadilan dalam proses pelatihan dan evaluasi model. Setelah itu, dilakukan normalisasi fitur untuk menyeragamkan skala antar variabel sehingga model tidak bias terhadap fitur dengan rentang nilai yang besar

3.3. Pelatihan Model

Pelatihan model dilakukan untuk membangun dan menguji kinerja berbagai algoritma *supervised learning* dalam mengklasifikasikan data indikator kesehatan penderita diabetes. Pada penelitian ini digunakan enam algoritma pembelajaran mesin, yaitu Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), dan Naive Bayes (NB). Pemilihan algoritma tersebut didasarkan pada karakteristiknya yang umum digunakan dalam penelitian klasifikasi data medis serta kemampuannya dalam menangani data dengan kombinasi fitur numerik dan kategorikal.

Seluruh model dibangun menggunakan bahasa pemrograman Python versi 3.11.13 dengan pustaka pendukung scikit-learn, yang menyediakan fungsi-fungsi siap pakai untuk pelatihan, pengujian, dan evaluasi model klasifikasi.

Proses pelatihan dilakukan dengan membagi dataset hasil pra-pemrosesan menjadi dua bagian, yaitu data latih (*training set*) dan data uji (*testing set*). Data latih digunakan untuk mengoptimalkan parameter model agar mampu mengenali pola hubungan antara fitur dan variabel target, sedangkan data uji digunakan untuk mengukur sejauh mana model dapat melakukan prediksi terhadap data baru yang belum pernah dilihat sebelumnya.

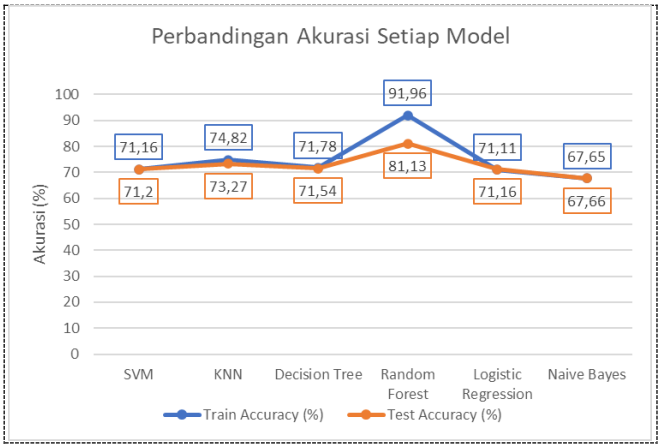
Setiap model dilatih pada data yang telah melalui proses penyeimbangan kelas menggunakan SMOTE, sehingga algoritma dapat belajar secara proporsional dari kedua kelas (penderita dan non-penderita diabetes). Setelah proses pelatihan selesai, masing-masing model diuji menggunakan data uji untuk memperoleh nilai akurasi. Nilai akurasi dihitung baik pada *training set* maupun *testing set* guna menilai tingkat generalisasi model serta mendeteksi kemungkinan terjadinya *overfitting*.

3.4. Hasil Evaluasi dan Perbandingan Akurasi Model

Hasil evaluasi kinerja masing-masing model pembelajaran mesin disajikan pada Tabel 1 dan Gambar 2. Pengujian dilakukan menggunakan *training set* dan *testing set* untuk mengukur kemampuan model dalam mempelajari pola dari data serta menggeneralisasikan pengetahuan tersebut terhadap data baru.

Tabel 1. Hasil akurasi *training* dan *testing* setiap model

Model	*Train Accuracy (%)*	*Test Accuracy (%)*
SVM	71,16	71,20
KNN	74,82	73,27
Decision Tree	71,78	71,54
Random Forest	**91,96**	**81,13**
Logistic Regression	71,11	71,16
Naive Bayes	67,65	67,66



Gambar 2. Hasil akurasi *training* dan *testing* setiap model

Berdasarkan hasil yang didapatkan, secara umum seluruh model yang diimplementasikan menunjukkan performa yang kompetitif dengan nilai akurasi pengujian berada di atas 67%. Model *Random Forest* (RF) memberikan hasil terbaik dengan nilai akurasi pengujian tertinggi sebesar 81,13%, menunjukkan kemampuannya dalam menangkap pola kompleks dari data yang digunakan. Hasil ini juga konsisten dengan akurasi pelatihan yang tinggi sebesar 91,96%, meskipun terdapat perbedaan yang mengindikasikan adanya potensi *overfitting* ringan karena model sedikit terlalu menyesuaikan diri dengan data latih.

Model *K-Nearest Neighbors* (KNN) menempati urutan kedua dengan akurasi pengujian sebesar 73,27%, diikuti oleh *Decision Tree* seebsar 71,54% dan *Support Vector Machine* sebesar 71,20% yang menunjukkan performa relatif stabil antara data latih dan uji. Model *Logistic Regression* memiliki akurasi pengujian sebesar 71,16%, yang cukup kompetitif dan menggambarkan kemampuannya dalam menangani data linier. Sementara itu, model *Naive Bayes* menunjukkan

performa terendah dengan akurasi pengujian 67,66%, yang kemungkinan disebabkan oleh asumsi independensi antar fitur yang tidak sepenuhnya terpenuhi dalam dataset ini.

Secara keseluruhan, hasil ini menunjukkan bahwa model berbasis *ensemble* seperti Random Forest memiliki keunggulan dalam menangani kompleksitas dan variasi data indikator kesehatan, sehingga lebih efektif dalam mengklasifikasikan penderita diabetes dibandingkan model linear atau probabilistik sederhana.

4. Kesimpulan

Penelitian ini telah berhasil menerapkan dan membandingkan kinerja enam algoritma *supervised learning* dalam klasifikasi penderita diabetes menggunakan *dataset Behavioral Risk Factor Surveillance System* (BRFSS) 2021. Proses pra-pemrosesan data melibatkan pembersihan data, penyeimbangan kelas menggunakan SMOTE (*Synthetic Minority Over-sampling Technique*), pembagian *dataset*, dan normalisasi fitur untuk meningkatkan kualitas data sebelum pelatihan model.

Berdasarkan hasil evaluasi, seluruh model menunjukkan performa yang cukup kompetitif dengan akurasi pengujian di atas 67%. Model Random Forest (RF) memperoleh hasil terbaik dengan akurasi pengujian sebesar 81,13%, diikuti oleh K-Nearest Neighbors (KNN) sebesar 73,27% dan Decision Tree (DT) sebesar 71,54%. Hasil ini menunjukkan bahwa pendekatan *ensemble learning* seperti Random Forest lebih mampu menangkap hubungan kompleks antar fitur dan memberikan performa prediksi yang lebih stabil dibandingkan model lainnya.

Penerapan teknik SMOTE terbukti efektif dalam mengatasi ketidakseimbangan kelas pada *dataset*, sehingga model dapat belajar secara lebih proporsional dari kedua kelas (penderita dan non-penderita diabetes). Hal ini berkontribusi terhadap peningkatan akurasi dan kemampuan generalisasi model.

Secara keseluruhan, penelitian ini menunjukkan bahwa kombinasi antara teknik penyeimbangan data dan pemilihan algoritma yang tepat memiliki peran penting dalam meningkatkan akurasi klasifikasi penyakit berbasis data kesehatan. Temuan ini diharapkan dapat menjadi dasar pengembangan sistem pendukung keputusan (*decision support system*) di bidang kesehatan, khususnya untuk deteksi dini penyakit diabetes.

Ucapan terima kasih

Penulis mengucapkan terima kasih kepada Universitas Lampung dan Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM) Universitas Lampung yang telah memberikan dukungan, baik dalam bentuk fasilitas, pendanaan, maupun arahan, sehingga penelitian ini dapat terlaksana dengan baik.

Daftar Pustaka

[1] IDF, "IDF Diabetes Atlas," *International Diabetes Federation*, 2025. https://diabetesatlas.org/

[2] N. Nurrahma, T. Muthia, and Y. E. Putra, "Perbandingan Akurasi Model Pembelajaran Mesin SVM , KNN , Decision Tree , dan Naive Bayes pada Klasifikasi Gangguan Kesehatan Mental," *Electr. – J. Rekayasa dan Teknol. Elektro*, vol. 18, no. 3, pp. 364–369, 2024.

[3] N. Moningka, Raynold, M. Hafidurrohman, W. A. T. R, and Kusrini, "Klasifikasi Mental Mahasiswa Menggunakan Metode Machine Learning," *J. Quancom*, vol. 1, no. 2, pp. 27–32, 2023, [Online]. Available: https://www.kaggle.com/datasets/shariful07/student-mental-

[4] T. Mythili, D. Mukherji, N. Padalia, and A. Naidu, "A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)," *Int. J. Comput. Appl. Technol.*, vol. 68, pp. 10–15, 2013, doi: 10.5120/11662-7250.

[5] M. Abdar, S. R. N. Kalhori, T. Sutikno, I. M. I. Subroto, and G. Arji, "Comparing Performance of Data Mining Algorithms in Prediction Heart Diseses," *Int. J. Electr. Comput. Eng.*, vol. 5, pp. 1569–1576, 2015, doi: 10.11591/ijece.v5i6.pp1569-1576.

[6] Nurrahma and R. Yusuf, "Comparing Different Supervised Machine Learning Accuracy on Analyzing COVID-19 Data using ANOVA Test," 2020. doi: 10.1109/ICIDM51048.2020.9339676.

[7] A. Narin, C. Kaya, and Z. Pamuk, "Automatic Detection of Coronavirus Disease (COVID-19) Using X-ray Images and Deep Convolutional Neural Networks," 2020.

[8] T. Vos *et al.*, "Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015," *Lancet*, vol. 388, no. 10053, pp. 1545–1602, 2016, doi: 10.1016/S0140-6736(16)31678-6.

[9] O. Odukoya *et al.*, "Development and Comparison of Three Data Models for Predicting Diabetes Mellitus Using Risk Factors in a Nigerian Population," vol. 28, no. 1, pp. 58–67, 2022.

[10] Kemenkes, "Laporan Riskesdas 2018," *Lembaga Penerbit Balitbangkes*. p. hal 156, 2018.

[11] P. Domingos and M. Pazzani, “On the Optimality of the Simple Bayesian Classifier underZero-One Loss. Machine Learning,” *Mach. Learn.*, vol. 29, pp. 103–130, 1997, [Online]. Available: https://link.springer.com/article/10.1023/A:1007413511361

[12] Pedro Domingos, “A Few Useful Things to Know About Machine Learning,” *Commun. ACM*, vol. 55, no. 10, pp. 79–88, 2012, [Online]. Available: https://dl.acm.org/citation.cfm?id=2347755

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, and Sandia, “SMOTE: Synthetic Minority Over-sampling Technique Nitesh,” *J. Artif. Intell. Res.*, vol. 16, no. 2, pp. 321–357, 2002, doi: 10.1002/eap.2043.

[14] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0197-0.