TEKNOKOM : Jurnal Teknologi dan Rekayasa Sistem Komputer
Vol. 6, No. 2, September 2023, pp. 127 - 132
ISSN: 2686-3219, DOI: 10.31943/teknokom.v6i2.152

PERBANDINGAN BERBAGAI MODEL MACHINE LEARNING UNTUK MENDETEKSI DIABETES

Ahmad Maulid Ridwan[1], Gilang Dwi Setiawan[1]
[1]Department of Computer Science, Universitas Nusa Mandiri, Indonesia

A R T I C L E I N F O

History of the article:
Received July 13, 2023
Revised July 28, 2023
Accepted August 1, 2023
Published August 4, 2023

Keywords:
Diabetes
Machine Learning
Prediction

ABSTRACT

Diabetes mellitus, often known as diabetes, is a significant metabolic illness that has a negative impact on living organisms. It causes high blood sugar levels by either creating inadequate insulin or using it inefficiently. Diabetes that is not effectively treated raises the risk of heart attacks, retinopathy, vision loss, skin disorders, and other ailments. Early detection is critical for guiding essential actions. In this setting, machine learning (ML) has emerged as a potent tool. We used Python data manipulation tools to develop ML techniques for discovering patterns and risk factors in the Pima Indian diabetes dataset in our study. We correctly identified patients as diabetes or non-diabetic using K-Nearest Neighbors (KNN), AdaBoost, Logistic Regression (LR), Light Gradient Boosting, Random Forest (RF), dan Support Vector Machine (SVM). Notably, we used the Synthetic Minority Over-sampling Technique (SMOTE) to solve class imbalance, which enhanced model performance. By efficiently utilizing ML and SMOTE in diabetes categorization, our work greatly adds to the scientific area. We suggest studying cutting-edge technology and undertaking external validation and clinical studies to assure trustworthy and generalized models for diabetic patient care in the future. With diabetes's increasing prevalence, such improvements have enormous promise for improving early identification and management, eventually leading to better health outcomes.

Kata Kunci :

ABSTRAK

Diabetes
Machine Learning
Prediksi

Diabetes melitus, sering dikenal sebagai kencing manis, adalah penyakit metabolik yang signifikan yang berdampak negatif pada organisme hidup. Ini menyebabkan kadar gula darah tinggi dengan menciptakan insulin yang tidak memadai atau menggunakannya secara tidak efisien. Diabetes yang tidak diobati secara efektif meningkatkan risiko serangan jantung, retinopati, kehilangan penglihatan, kelainan kulit, dan penyakit lainnya. Deteksi dini sangat penting untuk memandu tindakan penting. Dalam pengaturan ini, pembelajaran mesin (ML) telah muncul sebagai alat yang ampuh. Kami menggunakan alat manipulasi data Python untuk mengembangkan teknik ML untuk menemukan pola dan faktor risiko dalam kumpulan data diabetes Pima Indian dalam penelitian kami. Kami dengan benar mengidentifikasi pasien sebagai diabetes atau non-diabetes menggunakan K-Nearest Neighbors (KNN), AdaBoost, Logistic Regression (LR), Light Gradient Boosting, Random Forest (RF), dan Support Vector Machine (SVM). Khususnya, kami menggunakan Teknik Over-sampling Minoritas Sintetis (SMOTE) untuk mengatasi ketidakseimbangan kelas, yang meningkatkan kinerja model. Dengan memanfaatkan ML dan SMOTE secara efisien dalam kategorisasi diabetes, pekerjaan kami sangat menambah bidang ilmiah. Kami menyarankan untuk mempelajari teknologi mutakhir dan melakukan validasi eksternal dan studi klinis untuk memastikan model perawatan pasien diabetes yang dapat dipercaya dan digeneralisasikan di masa depan. Dengan prevalensi diabetes yang meningkat, peningkatan tersebut memiliki harapan besar untuk meningkatkan identifikasi dan manajemen dini, yang pada akhirnya mengarah pada hasil kesehatan yang lebih baik.

Correspondece:
Ahmad Maulid Ridwan,
Department of Computer Science,
Universitas Nusa Mandiri, Indonesia,
Email : 14210252@nusamandiri.ac.id

PENDAHULUAN

Diabetes melitus, sering dikenal sebagai diabetes, adalah kondisi metabolisme serius yang berdampak buruk pada manusia[1]. Topik biologi komputasi kini menjadi bagian penting dalam industri data besar berkat kemajuan dalam bioteknologi dan produksi data yang besar. Sumber data dalam industri kesehatan meliputi spektrometri massa, MRI, dan teknologi lainnya, namun seringkali data ini tidak dianalisis atau dimanfaatkan sepenuhnya.

Penemuan pengetahuan dari data bionik memiliki kepentingan yang besar di dunia modern [2]. Tujuan utamanya adalah menganalisis data bionik yang melimpah dan menciptakan model untuk meningkatkan respons medis yang tepat waktu. Kemanjuran dan keakuratan suatu pendekatan akan sangat tergantung pada seberapa baik sistem tersebut dapat menemukan pola dalam data dan membantu membangun model prediktif [3].

Proliferasi sumber data mengarah pada penguatan penelitian berbasis data di bidang biologi. Penggunaan yang paling signifikan dari data ini adalah dalam mendeteksi dini penyakit yang mengancam jiwa manusia [4], termasuk Diabetes Mellitus (DM). Diabetes, atau yang lebih dikenal sebagai Diabetes Melitus, adalah penyakit kronis yang mengenai jutaan orang di seluruh dunia. Salah satu tanda utama diabetes adalah peningkatan kadar glukosa darah.

Gejala hiperglikemia tinggi termasuk buang air kecil yang berlebihan dan peningkatan nafsu makan. Diabetes dapat menyebabkan sejumlah masalah kesehatan, termasuk risiko kematian dini, jika tidak terdeteksi pada tahap awal. Ini akhirnya dapat menyebabkan masalah lain seperti penyakit kulit dan masalah mata.

Di dalam tubuh manusia, pankreas mengeluarkan hormon yang dikenal sebagai insulin untuk membantu mengangkut glukosa dari darah ke dalam sel untuk digunakan sebagai sumber energi [5]. Diabetes disebabkan oleh ketidakmampuan tubuh untuk memproduksi cukup insulin atau oleh ketidakmampuan jaringan tubuh untuk menggunakan insulin dengan efektif.

Selain itu, terdapat jenis diabetes lain yang dikenal sebagai diabetes gestasional yang dapat berkembang pada tubuh seorang wanita jika kadar gula darahnya meningkat selama kehamilan dan diabetes belum terdiagnosis [6].

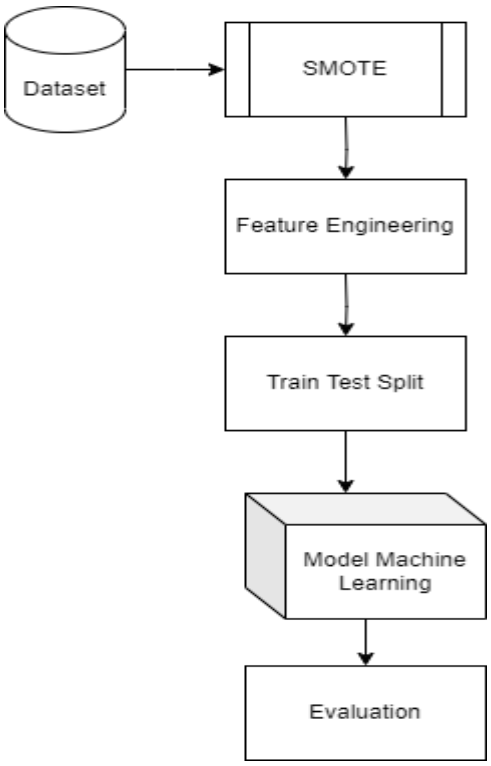
Tingginya prevalensi diabetes melitus dan tantangan dalam pengelolaannya menjadi fokus penting dalam studi ini. Di tengah kemajuan bioteknologi dan produksi data besar, industri kesehatan telah menghasilkan banyak data bionik yang berpotensi bernilai besar. Namun, masih ada kesenjangan dalam memanfaatkan data ini secara optimal untuk mengatasi masalah diabetes.

Metode konvensional dalam mendeteksi dan mengelola diabetes juga menghadapi keterbatasan, sehingga diperlukan pendekatan yang lebih inovatif. Oleh karena itu, tujuan utama studi ini adalah menerapkan teknik machine learning untuk menggali pola dan faktor risiko dalam dataset Pima Indian terkait diabetes. Dengan menggunakan analisis kesenjangan ini, penelitian ini berharap dapat meningkatkan deteksi dini, pengelolaan yang lebih baik, dan penggunaan optimal data bionik dalam upaya pencegahan dan pengobatan diabetes. Dengan kontribusi dari kecerdasan buatan dan analisis data, diharapkan penelitian ini dapat memberikan dampak positif bagi jutaan orang yang menderita diabetes, membawa perubahan yang signifikan dalam cara kita menghadapi penyakit ini di dunia moden.

Pada penelitian [1] tentang diabetes melitus dan penerapan Machine Learning menawarkan harapan baru dalam mendeteksi dan mengelola kondisi kesehatan ini secara lebih efektif. Diabetes melitus. Dalam penelitian ini, para peneliti berfokus pada analisis dataset Pima Indian menggunakan enam algoritma Machine Learning yang berbeda. Hasil dari penelitian ini Random forest mengungguli yang lain dengan akurasi 84%, presisi 83, recall 76, skor f1 86, dan skor ROC-AUC 83. Teknik klasifikasi dan machine learning digunakan dalam analisis ini. Kecerdasan buatan dikenal sebagai alat bantu machine learning dalam pengembangan sistem yang dapat membuat prediksi berdasarkan informasi yang diperoleh sebelumnya. Setiap bidang semakin banyak menggunakan teknik machine learning seiring berjalannya waktu. Di setiap industri, machine learning akan berkontribusi pada lebih banyak otomatisasi dan mengurangi kebutuhan akan tenaga kerja manusia. Berbagai tes laboratorium, termasuk tes toleransi glukosa oral (OGTT), tes urin, tes RPG, dan lainnya digunakan untuk mendiagnosis diabetes. Penelitian ini terutama berfokus pada penggunaan beberapa metode machine learning untuk membangun model prediksi diabetes.

METODE PENELITIAN

Penelitian ini diawali dengan pengumpulan dataset diabetes Pima Indian[7] sebagai langkah awal. Dilakukan analisis eksplorasi data untuk memahami sumber dan karakteristik dataset yang terkumpul. Setelah itu, tahap pra-proses dilakukan untuk membersihkan data dengan menghapus duplikat, data yang hilang, atau data yang tidak biasa. Selanjutnya, kami memilih model untuk melatih dan mengimplementasikan model prediksi kami. Berbagai parameter performa, seperti recall, precision, skor f1, dan akurasi, digunakan untuk membandingkan dan mengevaluasi model yang telah dibuat. Seluruh proses implementasi direpresentasikan dalam Gambar 1.



Gambar 1 Metodology

Metodologi yang kami terapkan dalam penelitian ini dimulai dengan menggunakan dataset diabetes Pima Indian sebagai dasar untuk membangun model prediksi dengan menggunakan machine learning. Untuk mengatasi ketidakseimbangan kelas dalam dataset, kami menggunakan teknik SMOTE untuk menciptakan sampel sintetis pada kelas minoritas[8].

Selanjutnya, kami melakukan tahap feature engineering untuk memilih fitur-fitur yang relevan, menghapus fitur yang tidak berguna, dan mengubah data kategorikal menjadi bentuk yang dapat diterima oleh model. Data kemudian dibagi

menjadi train set dan test set, di mana train set digunakan untuk melatih berbagai algoritma machine learning seperti K-Nearest Neighbors (KNN), AdaBoost, Logistic Regression (LR), Light Gradient Boosting, Random Forest (RF), dan Support Vector Machine (SVM). Pada tahap akhir, Penelitian ini mengevaluasi performa model menggunakan data uji dengan berbagai metrik, seperti akurasi, presisi, recall, F1-score, dan AUC-ROC. Dengan demikian, metodologi yang diterapkan memungkinkan pengembangan model prediksi yang dapat diandalkan untuk identifikasi dini dan manajemen diabetes, yang berpotensi memberikan dampak positif pada kesehatan masyarakat.

HASIL DAN PEMBAHASAN

Diagnosis diabetes dini dapat memperpanjang hidup seseorang. Menggunakan metode supervised machine learning, beberapa model telah dibuat. Python digunakan untuk membuat model ini. Dataset dibagi menjadi bagian pelatihan dan pengujian. Sisa 20% data digunakan untuk menguji model setelah dilatih dengan 80% data. Enam algoritma machine learning yang berbeda K-Nearest Neighbors, Light Gradient Boosting, Random Forest, Logistic Regression, AdaBoost dan Support Vector Machine digunakan untuk mengkategorikan pasien sebagai diabetes atau non-diabetes.

Lima parameter performa yaitu akurasi, presisi, recall, skor f1, dan Auc (Area di Bawah Kurva) digunakan untuk membandingkan semua model. Akurasi adalah proporsi waktu model secara akurat memprediksi apakah seorang pasien menderita diabetes atau tidak.

Persentase pasien diabetes yang dikenali secara akurat oleh model dikenal sebagai ingatan. Persentase pasien non-diabetes yang diidentifikasi dengan benar oleh model dikenal sebagai spesifisitas.

Untuk deskripsi data dapat dilihat pada Tabel 1.

Table 1. Tabel Deskripsi Data

No	Atribut	Tipedata	Missing Value
1	Pregnancy	int	0
2	Glucose Concentration	int	0
3	Blood Pressure(mmHg)	int	0
4	Skin Thickness(mm)	int	0
5	Insulin(U/ml)	Int	0
6	BMI (kg/m)	Int	0
7	Diabetes Pedigree Function	int	0
8	Age	Int	0
9	Outcome Class	Int	0

Pada deskripsi data di Tabel 1 menunjukan bahwa data tidak ada missing value dan tipe data menunjukan semua data mempunyai tipe data integer.

F1-score adalah rata-rata harmonik dari recall dan akurasi model. Seberapa baik model mampu membedakan antara kelas ditunjukkan oleh karakteristik operasi penerima (ROC) dan kurva area di bawah kurva (AUC). Tabel 2 menampilkan hasil yang diperoleh setelah menggunakan beberapa metode machine learning pada dataset diabetes PIMA.

Table 2. Perbandingan model klasifikasi

Model	Akurasi	Precision	Recall	F1 Scorre	AUC - ROC
K-Nearest Neighbors	82%	84%	82%	82%	82%
Light Gradient Boosting	79%	79%	78%	78%	78%
Random Forest	79%	78%	78%	78%	78%
Logistic Regression	78%	75%	83%	79%	73%
Gradient Boosting	78%	75%	75%	75%	81%
AdaBoost	76%	76%	76%	76%	76%
Support Vector Machine	75%	75%	75%	75%	75%

Penelitian ini merupakan langkah penting dalam menganalisis performa beberapa model machine learning yang berbeda untuk memprediksi target tertentu. Hasil penelitian menyoroti keunggulan model K-Nearest Neighbors (KNN) dengan akurasi mencapai 82%. KNN menunjukkan hasil yang sangat baik dalam kinerja prediksi dengan nilai Precision, Recall, dan F1-Score yang seimbang[15], menandakan kemampuan model ini dalam meminimalkan false positive dan false negative. Dengan demikian, KNN sangat cocok digunakan dalam kasus-kasus di mana penting untuk memiliki tingkat akurasi yang tinggi dan menjaga keseimbangan dalam memprediksi kedua kelas.

Selanjutnya, model Light Gradient Boosting (LGB) dan Random Forest juga menunjukkan performa yang baik dengan akurasi dan F1-Score mencapai 79%. LGB dan Random Forest, meskipun memiliki akurasi yang setara, mungkin memiliki karakteristik yang berbeda dalam situasi yang lebih kompleks. Karenanya,

pemilihan model antara LGB dan Random Forest harus mempertimbangkan faktor-faktor lain seperti waktu komputasi dan kemudahan interpretasi.

Di sisi lain, meskipun model Logistic Regression menunjukkan recall yang tinggi sebesar 83%, namun precision-nya hanya sebesar 75%. Hal ini menandakan model ini lebih cenderung untuk memberikan false positive. Dalam konteks tertentu, di mana penting untuk menghindari kesalahan prediksi false positive, mungkin perlu dipertimbangkan trade-off antara precision dan recall ini. Kemudian, model AdaBoost dan Support Vector Machine (SVM) juga menunjukkan hasil yang cukup baik dengan akurasi dan F1-Score mencapai 76% dan 75% secara berurutan.

KESIMPULAN

Dari hasil penelitian ini, dapat disimpulkan bahwa model K-Nearest Neighbors (KNN) menunjukkan performa terbaik dengan akurasi 82%, serta nilai Precision, Recall, dan F1-Score yang seimbang. Model ini merupakan pilihan yang kuat untuk kasus-kasus di mana tingkat akurasi yang tinggi dan keseimbangan dalam prediksi kelas positif dan negatif sangat krusial. Selain itu, model Light Gradient Boosting (LGB) dan Random Forest juga memberikan hasil yang baik dengan akurasi dan F1-Score sekitar 79%, sehingga kedua model ini dapat dijadikan alternatif yang layak tergantung pada kompleksitas tugas dan kebutuhan interpretasi.

Model KNN menjadi pilihan unggul untuk tugas prediksi dengan performa yang konsisten, sementara LGB dan Random Forest juga merupakan pilihan yang kuat dalam situasi tertentu. Namun, perlu diingat bahwa penelitian ini memiliki beberapa keterbatasan. Salah satunya adalah ukuran dataset yang digunakan, karena ukuran dataset yang lebih besar mungkin dapat memberikan hasil yang lebih stabil dan menggambarkan lebih baik kinerja model. Selain itu, evaluasi performa model mungkin dapat ditingkatkan dengan menggunakan teknik validasi silang (cross-validation) untuk menghindari overfitting dan memperoleh estimasi performa yang lebih konsisten.

Untuk pengembangan penelitian lebih lanjut, disarankan untuk mempertimbangkan beberapa hal. Pertama, eksplorasi lebih lanjut pada parameter-model yang digunakan untuk masing-masing model, seperti jumlah tetangga dalam KNN, jumlah pohon dalam Random Forest, atau hyperparameter lainnya untuk LGB dan SVM, untuk meningkatkan performa model. Kedua, penelitian dapat di perluas untuk mencakup lebih banyak model dan algoritma machine learning yang lebih kompleks untuk mengevaluasi kemampuan prediktif mereka secara lebih komprehensif. Terakhir, perlu diadakan penelitian lebih lanjut dengan menggunakan dataset yang lebih besar dan lebih bervariasi untuk menguji generalisasi model pada berbagai jenis data. Dengan demikian, penelitian lebih lanjut ini dapat memberikan wawasan yang lebih mendalam dan aplikatif dalam penerapan machine learning di berbagai bidang.

REFERENSI

[1] R. Katarya and S. Jain, “Comparison of different machine learning models for diabetes detection,” in Proceedings of 2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering, ICADEE 2020, Institute of Electrical and Electronics Engineers Inc., Dec. 2020. doi: 10.1109/ICADEE51157.2020.9368899.

[2] A. Mujumdar and V. Vaidehi, “Diabetes Prediction using Machine Learning Algorithms,” in Procedia Computer Science, Elsevier B.V., 2019, pp. 292–299. doi: 10.1016/j.procs.2020.01.047.

[3] A. Z. Woldaregay et al., “Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes,” Artificial Intelligence in Medicine, vol. 98. Elsevier B.V., pp. 109–134, Jul. 01, 2019. doi: 10.1016/j.artmed.2019.07.007.

[4] H. M. Deberneh and I. Kim, “Prediction of type 2 diabetes based on machine learning algorithm,” Int J Environ Res Public Health, vol. 18, no. 6, Mar. 2021, doi: 10.3390/ijerph18063317.

[5] D. R. Nair et al., “Trend in the clinical profile of type 2 diabetes in India - Study from a diabetes care centre in South India,” Diabetes and Metabolic Syndrome: Clinical Research and Reviews, vol. 14, no. 6, pp. 1851–1857, Nov. 2020, doi: 10.1016/j.dsx.2020.09.018.

[6] M. M. Islam, M. J. Rahman, D. Chandra Roy, and M. Maniruzzaman, “Automated detection and classification of diabetes disease based on Bangladesh demographic and health

survey data, 2011 using machine learning approach," Diabetes and Metabolic Syndrome: Clinical Research and Reviews, vol. 14, no. 3, pp. 217–219, May 2020, doi: 10.1016/j.dsx.2020.03.004.

[7] S. C. Gupta and N. Goel, "Predictive Modeling and Analytics for Diabetes using Hyperparameter tuned Machine Learning Techniques," Procedia Comput Sci, vol. 218, pp. 1257–1269, 2023, doi: 10.1016/j.procs.2023.01.104.

[8] A. Arafa, N. El-Fishawy, M. Badawy, and M. Radad, "RN-SMOTE: Reduced Noise SMOTE based on DBSCAN for enhancing imbalanced data classification," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 8, pp. 5059–5074, Sep. 2022, doi: 10.1016/j.jksuci.2022.06.005.

[9] J. Huang, Y. Wei, J. Yi, and M. Liu, "An improved knn based on class contribution and feature weighting," in Proceedings - 10th International Conference on Measuring Technology and Mechatronics Automation, ICMTMA 2018, Institute of Electrical and Electronics Engineers Inc., Apr. 2018, pp. 313–316. doi: 10.1109/ICMTMA.2018.00083.

[10] L. Hao and G. Huang, "An improved AdaBoost algorithm for identification of lung cancer based on electronic nose," Heliyon, vol. 9, no. 3, Mar. 2023, doi: 10.1016/j.heliyon.2023.e13633.

[11] L. Yang and A. Shami, "On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice," Jul. 2020, doi: 10.1016/j.neucom.2020.07.061.

[12] T. O. Omotehinwa, D. O. Oyewola, and E. G. Dada, "A Light Gradient-Boosting Machine algorithm with Tree-Structured Parzen Estimator for breast cancer diagnosis," Healthcare Analytics, p. 100218, Dec. 2023, doi: 10.1016/j.health.2023.100218.

[13] S. Liu, H. Li, Y. Zhang, B. Zou, and J. Zhao, "Random forest-based track initiation method," The Journal of Engineering, vol. 2019, no. 19, pp. 6175–6179, Oct. 2019, doi: 10.1049/joe.2019.0180.

[14] D. Valero-Carreras, J. Alcaraz, and M. Landete, "Comparing two SVM models through different metrics based on the confusion matrix," Comput Oper Res, vol. 152, Apr. 2023, doi: 10.1016/j.cor.2022.106131.

[15] A. Theissler, M. Thomas, M. Burch, and F. Gerschner, "ConfusionVis: Comparative evaluation and selection of multi-class classifiers based on confusion matrices," Knowl Based Syst, vol. 247, Jul. 2022, doi: 10.1016/j.knosys.2022.108651.