Evaluasi Performa Random Forest, XGBoost, dan LightGBM dalam Diagnosis Dini Diabetes Mellitus

Hendra Kurniawan*[1], Asmaul Dwi Akbar[2], Nicholas Svensons[3], Yandi Jaya Antonio[4], Sri Karnila[5], Egi Safitri[6], Nurjoko[7]
[1,2,3,4,5,6,7]Institut Informatika dan Bisnis Darmajaya; Jl. Z.A Pagar Alam No. 93 Bandar Lampung 35142 Indonesia, Telp. (0721) 787214/Fax. (0721) 700261

e-mail: [1]**hendra.kurniawan@darmajaya.ac.id**, [2]akbaraza.2311080013@mail.darmajaya.ac.id, [3]2311080002.2311080002@mail.darmajaya.ac.id, [4]yandi.2311080002@mail.darmajaya.ac.id, *[5]srikarnila_dj@darmajaya.ac.id, [6]egisafitri@darmajaya.ac.id, [7]nurjoko@darmajaya.ac.id

Abstrak

Diabetes mellitus merupakan penyakit kronis yang ditandai dengan meningkatnya kadar gula darah, yang berisiko menimbulkan komplikasi serius seperti gangguan jantung, gagal ginjal, dan kebutaan. Deteksi dini memegang peran penting dalam menurunkan risiko komplikasi serta meningkatkan kualitas hidup penderita. Penelitian ini berfokus pada evaluasi kinerja tiga algoritma machine learning, yaitu Random Forest, XGBoost, dan LightGBM, dalam memprediksi potensi risiko diabetes. Dataset yang digunakan diperoleh dari National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), yang mencakup 768 sampel dengan 9 fitur utama. Metodologi penelitian terdiri dari beberapa tahap, meliputi pengumpulan data, pra-pemrosesan, penanganan ketidakseimbangan data dengan metode SMOTE, pemisahan data menjadi set pelatihan dan pengujian, penerapan algoritma, serta evaluasi model menggunakan metrik akurasi, presisi, recall, F1-score, dan Area Under the Curve (AUC). Hasil penelitian menunjukkan bahwa Random Forest memberikan performa terbaik dengan skor AUC sebesar 86%, disusul oleh XGBoost dengan 83% dan LightGBM sebesar 82%. Dengan akurasi yang tinggi, model ini berpotensi menjadi alat pendukung dalam proses diagnosis dini diabetes, sehingga dapat membantu mempercepat dan meningkatkan ketepatan pengambilan keputusan medis.

Kata kunci*—Diabetes Mellitus, Machine Learning, Random Forest, XGBoost, LightGBM*

Abstract

Diabetes mellitus is a long-term condition marked by elevated blood sugar levels, which can lead to serious complications such as heart disease, kidney failure, and vision impairment. Early detection plays a vital role in minimizing these risks and enhancing patients' quality of life. This research focuses on assessing the performance of three machine learning algorithms—Random Forest, XGBoost, and LightGBM—in predicting diabetes risk. The dataset utilized originates from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), comprising 768 samples with 9 key features. The research methodology involves multiple stages, including data collection, preprocessing, addressing data imbalance using SMOTE, data splitting for training and testing, algorithm implementation, and model evaluation through accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) metrics. Findings reveal that Random Forest delivers the highest performance with an AUC score of 86%, followed by XGBoost (83%) and LightGBM (82%). With its strong accuracy, this model holds potential as a valuable tool for early diabetes diagnosis, contributing to faster and more precise medical decision-making.

Keywords— *Diabetes Mellitus, Machine Learning, Random Forest, XGBoost, LightGBM*

1. PENDAHULUAN

Diabetes Mellitus, yang kerap dikenal sebagai kencing manis di kalangan masyarakat Indonesia, merupakan kondisi kronis yang dapat memicu berbagai kerusakan serius seperti stroke, penyakit jantung, kebutaan, gagal ginjal, hingga risiko kematian [1], [2]. Peningkatan kadar gula tersebut diakibatkan oleh sintetis insulin yang tidak mencukupi atau responsifitas sel tubuh yang tidak memadai terhadap insulin [3]. Diabetes adalah penyakit yang umum menyerang individu di seluruh dunia [4] dan penyakit yang memiliki keterkaitan erat dengan genetika dan epigenetika [5]. International Diabetes Federation melaporkan, Pada tahun 2021, ada 537 juta orang dewasa (20-79 tahun) di seluruh dunia yang mengalami diabetes. Diperkirakan jumlah ini akan tumbuh menjadi 643 juta pada tahun 2030 dan 783 juta pada tahun 2045 [6]. Lebih lanjut, berdasarkan data Kementerian Kesehatan Republik Indonesia tahun 2024, Jumlah penderita diabetes saat ini telah mencapai 19,5 juta jiwa, dan angka ini diproyeksikan akan meningkat hingga 28,5 juta jiwa pada tahun 2045 [7].

Upaya Pencegahan diabetes tidak selalu dapat dilakukan dengan cepat, tetapi pendeteksian secara dini diharapkan akan memberikan hasil yang optimal. Pencegahan harus dilakukan dimulai sejak dini terutama pada masa kanak-kanak, dimana anak cenderung mempunyai kebiasaan mengkonsumsi makanan manis yang dapat meningkatkan resiko terkena penyakit diabetes melitus [8]. Salah satu cara yang baik untuk mencegah adalah dengan hidup sehat, seperti makan-makanan yang seimbang [9], olahraga secara teratur, dan membatasi konsumsi makanan serta minuman yang mengandung gula tinggi [10]**,** [11]. Langkah-langkah ini diharapkan tidak hanya dapat mencegah diabetes melitus, tetapi juga memperkuat kesadaran akan pentingnya diagnosis dini sebagai upaya untuk mengendalikan diabetes. Oleh karena itu, diperlukan suatu peningkatan kemampuan untuk memprediksi secara dini resiko diabetes. Membuat prediksi dini terhadap risiko penyakit diabetes telah menjadi fokus penelitian dalam pembelajaran mesin, dimana hasil dari prediksi akan bermanfaat bagi masyarakat agar memiliki kesadaran akan risiko diabetes.

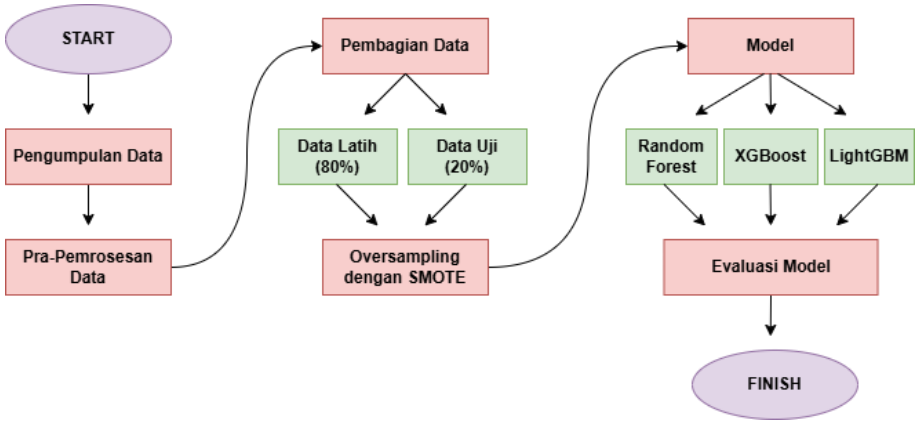
Penelitian ini memiliki urgensi dalam mendeteksi secara dini diabetes untuk mengurangi potensi kematian akibat diabetes dari prediksi nyata lembaga nasional dan internasional dengan memanfaatkan algoritma pembelajaran mesin. Dalam penelitian sebelumnya, berbagai algoritma telah diimplementasikan untuk mendukung diagnosis dini diabetes mellitus dan menghasilkan tingkat akurasi yang bervariasi. Algoritma C4.5 memiliki tingkat akurasi sebesar 76%, melampaui akurasi algoritma Support Vector Machine yang mencapai 70% [12]. Dilanjutkan algoritma C4.5 memiliki akurasi 75,32% [13]. Support Vector Machine (SVM) menunjukkan kinerja yang lebih baik daripada Naive Bayes, dengan akurasi sebesar 78,04% [14]. Adapun algoritma regresi logistik, memberikan hasil paling optimal, dengan tingkat akurasi mencapai 82% [15].

Studi ini bertujuan untuk membandingkan tiga algoritma machine learning, yakni Random Forest, LightGBM, dan XGBoost, dalam membantu prediksi awal penyakit diabetes mellitus. Penelitian ini akan menganalisis sejauh mana kinerja model dalam hal akurasi, presisi, recall, F1-Score, dan *Area Under the Curve* (AUC) untuk menemukan algoritma yang paling efisien dalam mendeteksi dini diabetes mellitus. Adapun tujuan dari penelitian ini untuk mengidentifikasi model dengan performa terbaik dalam memprediksi risiko diabetes, sehingga dapat membantu meningkatkan kualitas diagnosis dini dan mendukung pengambilan keputusan medis yang lebih tepat dan efektif. Dengan hasil evaluasi ini, diharapkan dapat memberikan kontribusi yang signifikan dalam menciptakan model-model prediktif yang lebih tepat dan efisien untuk membantu dalam pencegahan serta penanganan awal penyakit diabetes mellitus.

2. METODE PENELITIAN

Penelitian ini memanfaatkan *tools python* sebagai alat untuk menerapkan algoritma Random Forest dalam mendukung diagnosis dini dan prediksi penyakit diabetes. Metode yang digunakan mencakup langkah - langkah sistematis, seperti pengumpulan data, pra-pemrosesan

data, pembagian data untuk pelatihan dan pengujian, penanganan ketidakseimbangan data dengan teknik *oversampling* menggunakan SMOTE, pemodelan dengan algoritma Random Forest, serta evaluasi kinerja model. Proses penelitian ini dapat dilihat pada Gambar 1.



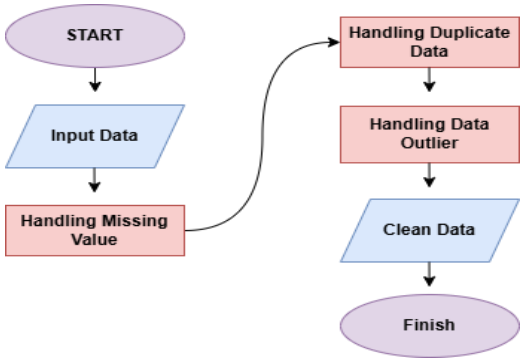
Gambar 1. Proses Penelitian

2.1 Pengumpulan Data

Pada tahap ini, data yang digunakan dalam penelitian ini bersumber dari *National Institute of Diabetes and Digestive and Kidney Diseases* (NIDDK). *Dataset* ini mencakup 768 entri dan 9 fitur, yang mencakup berbagai variabel penting yang berhubungan dengan faktor risiko diabetes. Variabel - variabel tersebut meliputi informasi seperti usia, jenis kelamin, kadar glukosa darah, tekanan darah, riwayat kehamilan, *body mass index* (BMI) serta faktor lainnya yang mempengaruhi perkembangan diabetes. Data ini menjadi landasan utama dalam pengembangan model prediksi yang tepat untuk mendeteksi dan memperkirakan risiko diabetes melitus.

2.2 Pra-Pemrosesan Data

Pra-pemrosesan data adalah langkah krusial dalam analisis data mining yang bertujuan untuk membersihkan data, menyesuaikan format, serta mempersiapkan data agar proses analisis dapat dilakukan dengan lebih mudah dan menghasilkan hasil yang lebih akurat [16]. Pra-pemrosesan data merupakan tahap krusial dalam menerapkan algoritma pembelajaran mesin dan analisis teks. Fokus utama dari pra-pemrosesan adalah untuk membersihkan dan menyiapkan data agar lebih mudah dianalisis, sehingga menghasilkan hasil yang lebih akurat [17]. Untuk lebih jelas perhatikan Gambar 2.



Gambar 2 Tahapan Pra-Pemrosesan Data

2.3 Pembagian Data

Pembagian data merupakan proses memisahkan dataset menjadi dua atau lebih subset yang berbeda, sering kali digunakan untuk pelatihan dan pengujian model *machine learning*. Tujuan utama dari pembagian data adalah untuk menilai seberapa baik model bekerja pada data yang belum terlihat saat di latih [18]. Dalam penelitian ini, dataset dipisah menjadi dua bagian, dimana 80% digunakan sebagai data latih yang mencakup 614 entri data, sementara 20% digunakan sebagai data uji dengan 154 entri data.

2.4 Oversampling dengan SMOTE

Synthetic Minority Oversampling Technique (SMOTE) merupakan metode yang termasuk dalam kategori teknik oversampling. SMOTE bekerja dengan cara memperbanyak data pada kelas minoritas melalui proses sintesis data baru [19] . SMOTE adalah metode *oversampling* yang sangat populer. Ini melibatkan penambahan data sintetis ke kelas minoritas. SMOTE adalah salah satu jenis metode oversampling yang dapat membantu mengatasi masalah ketidakseimbangan kelas. Dengan menggunakan SMOTE, performa pembelajaran mesin dapat meningkat dibandingkan dengan hasil yang diperoleh dari dataset asli [20]. Kelebihan utama dari metode ini adalah tidak menghasilkan kehilangan informasi karena tidak ada pengurangan data seperti yang terjadi pada metode *undersampling* [21].

2.5 Model

Dalam penelitian ini, digunakan tiga algoritma machine learning, yaitu Random Forest, XGBoost, dan LightGBM. Algoritma-algoritma ini dipilih karena mereka dikenal memiliki kinerja yang baik untuk mengolah data yang memiliki karakteristik yang rumit serta dapat menghasilkan prediksi yang tepat. Penjelasan tentang masing-masing algoritma akan dijabarkan di subbab berikut.

2.5.1 Random Forest

Random Forest (RF) merupakan metode pembelajaran mesin yang serbaguna dan kuat [22] yang menggunakan kumpulan pohon keputusan untuk membuat prediksi. RF memiliki beberapa manfaat, termasuk varians yang rendah, akurasi yang tinggi, ketahanan relatif terhadap *overfitting*, kemampuan untuk menangani hubungan nonlinier, dan kemampuan interpretasi dengan menggunakan kepentingan fitur [23]. Random Forest adalah sebuah kemajuan dari teknik pohon keputusan, yang melibatkan banyak pohon keputusan. Setiap pohon dibentuk dengan sampel yang berbeda-beda dan pemilihan atribut yang akan dipecah menjadi setiap pohon dilakukan secara acak dari subset atribut tersedia [24]. Random Forest menawarkan berbagai keuntungan, di antaranya meningkatkan akurasi meskipun ada data yang hilang, efektif dalam menangani outliers, dan efisien dalam hal penyimpanan data. Selain itu, metode ini juga memiliki mekanisme seleksi fitur yang memungkinkan dapat melaukan pemilihan fitur terbaik, sehingga dapat meningkatkan performa model klasifikasi secara keseluruhan [25].

2.5.2 XGBoost

XGBoost merupakan metode *boosting* yang dibentuk oleh sejumlah Decision Tree, di mana setiap pohon yang baru dibangun bergantung pada kinerja pohon yang telah ada. Pada tahap awal, Decision Tree pertama dalam XGBoost mungkin tidak sangat baik dalam melakukan klasifikasi, dan probabilitas awal ditentukan oleh para peneliti. Namun, seiring pembangunan pohon-pohon berikutnya, bobot akan diperbarui, sehingga akhirnya terbentuklah kumpulan pohon klasifikasi yang lebih kuat [26] .

2.5.3 LightGBM

LightGBM adalah algoritma yang masuk kedalam kelompok *gradient boosting*, yang memanfaatkan pendekatan pembelajaran yang berfokus pada pohon keputusan. Kelebihan dari LightGBM terletak pada penggunaan teknik *Exclusive Feature Bundling (EFB)* serta *Gradient-Based One Side Sampling (GOSS)*. Teknik-teknik ini dirancang untuk meningkatkan akurasi tanpa mengorbankan kompleksitas algoritmanya [27].

2.6 Evaluasi Model

Setelah menyelesaikan pemodelan, langkah selanjutnya adalah menilai kinerja model untuk memastikan keandalan dan ketepatan hasil prediksi. Penilaian ini dilakukan dengan menggunakan berbagai metrik, termasuk akurasi, presisi, recall, F1-Score, dan *Area Under the Curve* (AUC). Pemilihan metrik-metrik ini dimaksudkan untuk memberikan pemahaman yang menyeluruh tentang kemampuan model dalam mengklasifikasikan data secara akurat, khususnya dalam kasus di mana terdapat ketidakseimbangan pada data.

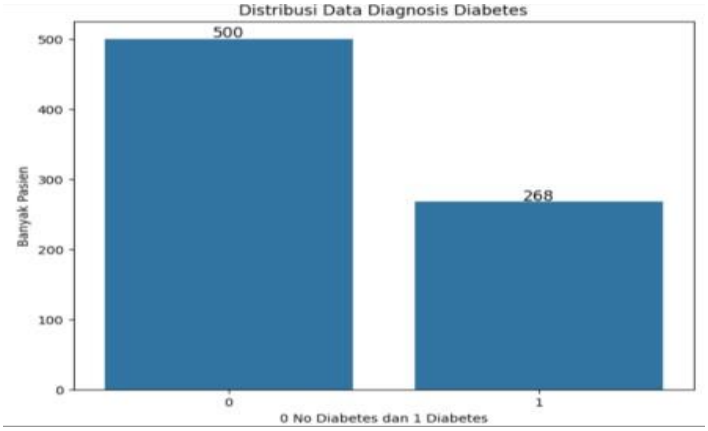
3. HASIL DAN PEMBAHASAN

3.1 Analisis Eksplorasi Data

EDA atau analisis eksplorasi data digunakan untuk menggali dan memahami struktur serta karakteristik data dalam memberikan pemahaman pemahaman yang lebih mendalam tentang distribusi target (diagnosis diabetes) serta hubungan antar variabel dalam dataset. Dengan menggunakan visualisasi distribusi target, hasil diagnosis diabetes terbagi antara positif dan negatif. Korelasi antar variabel dieksplorasi menggunakan *heatmap* untuk mengidentifikasi hubungan signifikan antara fitur-fitur seperti kadar glukosa darah, BMI, usia, dan faktor lainnya.

3.1.1 Distribusi Target

Distribusi target menggambarkan pembagian hasil diagnosis, dimana nilai 0 mengindikasikan pasien yang tidak menderita diabetes, sementara nilai 1 mengindikasikan pasien yang didiagnosis mengidap diabetes. Visualisasi distribusi target menggunakan barplot, untuk mengevaluasi proporsi pasien yang didiagnosis dengan diabetes dibandingkan dengan yang tidak, yang membantu dalam memperoleh pemahaman tentang ketidakseimbangan kelas dalam dataset. Visualisasi distribusi target dapat pada Gambar 3.



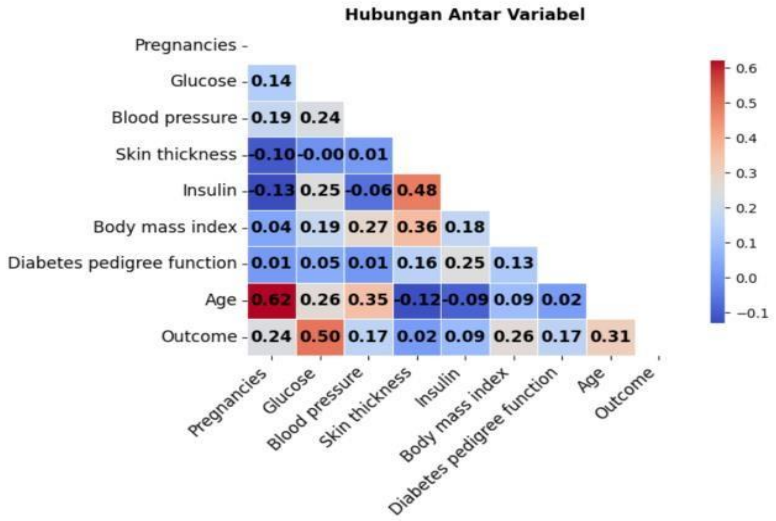
Gambar 3 Distribusi Diagnosis Diabetes

Berdasarkan hasil distribusi data pada Gambar 3 dapat diambil kesimpulan bahwasanya pada distribusi variabel target (*outcome*) terjadi ketidakseimbangan data yang dimana jumlah pasien yang tidak terdiagnosis diabetes lebih dominan yaitu sebesar 65.1% dibandingkan jumlah pasien yang terdiagnosis diabetes sebesar 34.9%.

3.1.2 Analisis Korelasi

Studi ini menggunakan analisis korelasi untuk melihat hubungan antara variabel seperti kadar glukosa, indeks massa tubuh, usia, tekanan darah, dan faktor kesehatan lainnya. Analisis ini membantu mengidentifikasi apakah hubungan antar variabel bersifat positif, negatif, atau tidak ada. Selain itu, analisis korelasi berguna untuk memahami faktor-faktor yang memengaruhi diagnosis diabetes dan membantu dalam pemilihan fitur model prediksi. Hasilnya juga

menunjukkan interaksi antar faktor yang dapat meningkatkan risiko diabetes, seperti ditunjukkan pada Gambar 4.



Gambar 4 Korelasi Terhadap Diabetes

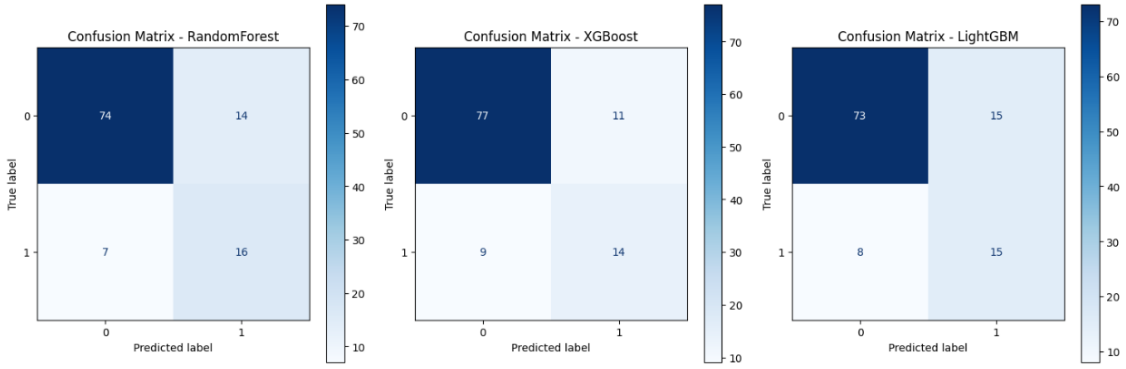
Gambar 4 menunjukkan matriks korelasi antar variabel dalam dataset. Warna merah menandakan korelasi positif, biru untuk negatif, dengan nilai berkisar dari -1 hingga 1. Glukosa memiliki korelasi tertinggi dengan Outcome (0,50), diikuti Usia (0,31) dan BMI (0,26). Variabel lain seperti Tekanan Darah dan Ketebalan Kulit memiliki korelasi rendah. Usia dan Kehamilan menunjukkan korelasi negatif kuat (-0,62). Hasil ini membantu mengidentifikasi fitur penting untuk model prediksi diabetes.

3.2 Hasil dan Evaluasi Model

Dalam diskusi ini, kami akan menunjukkan hasil dan penilaian dari model yang telah digunakan untuk memprediksi diabetes. Penilaian dilakukan dengan beberapa metrik, termasuk Matriks Kebingungan, kurva *Receiver Operating Characteristic (ROC)*, dan analisis perbandingan yang meliputi akurasi, presisi, recall, F1-Score, dan AUC. Kurva ROC berfungsi untuk menilai sejauh mana model mampu membedakan antara kelas positif dan negatif, sedangkan analisis komparasi dilakukan untuk membandingkan kinerja ketiga model yang digunakan, yaitu Random Forest, XGBoost, dan LightGBM. Dari hasil evaluasi ini, diharapkan kita bisa mendapatkan pemahaman mengenai model yang terbaik untuk memprediksi diabetes.

3.2.1 Confusion Matrix

Pada tahap evaluasi model, digunakan beberapa metrik untuk menilai kinerjanya. *Confusion matrix* menunjukkan jumlah prediksi benar dan salah untuk masing-masing kelas. Precision, recall, dan F1-score digunakan untuk mengukur kemampuan model dalam menangani data yang tidak seimbang. Akurasi dihitung untuk melihat seberapa baik model memprediksi secara keseluruhan. Terakhir, kurva ROC dan nilai AUC digunakan untuk menilai keseimbangan antara sensitivitas dan spesifisitas pada berbagai threshold, memberikan gambaran menyeluruh tentang performa model.

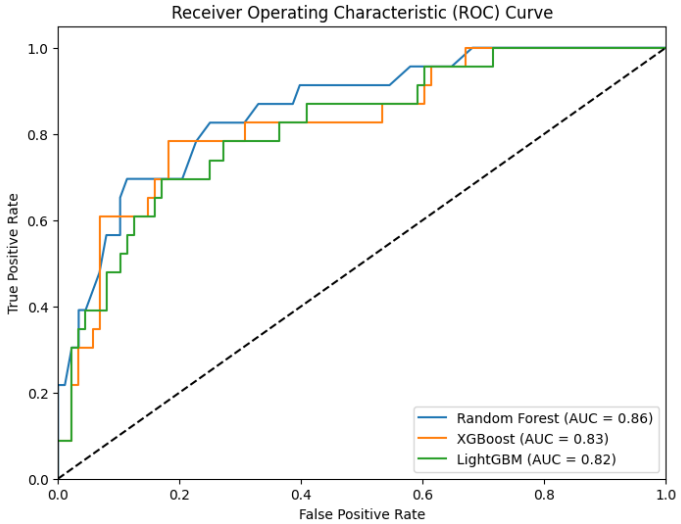


Gambar 5 Confusion Matrix

Gambar 5 menunjukkan *confusion matrix* untuk tiga algoritma *machine learning*: Random Forest, XGBoost, dan LightGBM. Matrix ini menggambarkan jumlah prediksi benar dan salah pada masing-masing kelas. Pada Random Forest, terdapat 74 True Negative (TN), 14 False Positive (FP), 7 False Negative (FN), dan 16 True Positive (TP). Pada XGBoost, terdapat 77 TN, 11 FP, 9 FN, dan 14 TP, sementara LightGBM memiliki 73 TN, 15 FP, 8 FN, dan 15 TP. XGBoost menunjukkan jumlah TN tertinggi, sedangkan Random Forest memiliki FN paling rendah, menandakan keunggulan masing-masing model pada aspek tertentu.

3.2.2 Kurva ROC

Untuk mengevaluasi kinerja model yang digunakan dalam prediksi diabetes, dilakukan analisis menggunakan kurva *Receiver Operating Characteristic (ROC)*. Kurva ROC adalah alat yang digunakan untuk menilai kemampuan model dalam membedakan antara kelas positif dan negatif pada berbagai ambang batas keputusan. Nilai *Area Under the Curve (AUC)* yang dihasilkan mencerminkan tingkat keandalan model dalam melakukan klasifikasi, dengan nilai AUC yang hampir mencapai 1 menunjukkan kinerja model yang sangat unggul. Gambar berikut menampilkan kurva ROC untuk tiga model yang dibandingkan, yaitu Random Forest, XGBoost, dan LightGBM.



Gambar 6 Kurva ROC

Berdasarkan kurva ROC yang terlihat pada gambar 6, model Random Forest menunjukkan hasil terbaik dengan nilai AUC mencapai 0,86. Setelah itu, model XGBoost mengikuti dengan AUC sebesar 0,83, dan LightGBM memiliki AUC 0,82. Nilai AUC yang lebih tinggi pada model Random Forest menandakan bahwa model ini lebih efektif dalam membedakan

antara orang yang memiliki diabetes dan yang tidak. Ketiga model ini memiliki kurva yang cukup baik, dengan nilai AUC di atas 0,8 yang menandakan kinerja model yang baik secara keseluruhan. Perbedaan kinerja yang relatif kecil antara model XGBoost dan LightGBM mengindikasikan bahwa kedua model ini memiliki kemampuan prediksi yang serupa, meskipun sedikit lebih rendah dibandingkan Random Forest.

3.2.3 Analisis Komparasi

Untuk membandingkan kinerja ketiga model yang digunakan dalam prediksi diabetes, dilakukan analisis komparasi berdasarkan metrik evaluasi seperti *accuracy*, AUC, *precision*, *recall*, dan *F1-score*. Tabel berikut menyajikan hasil evaluasi dari model Random Forest, XGBoost, dan LightGBM.

Tabel 1 Performa Model

Classifier	Accuracy	AUC	Precision	Recall	F1-Score
Random Forest	81.08%	86%	72.14%	77.20%	74.40%
XGBoost	81.98%	83%	73.20%	74.50%	73.40%
LightGBM	79.28%	82%	70.20%	74%	71,10%

Tabel 1 menunjukkan kinerja tiga model pembelajaran mesin yang diterapkan untuk meramalkan diabetes, yaitu LightGBM, XGBoost, dan Random Forest. Berdasarkan tabel tersebut, model Random Forest menunjukkan performa terbaik dengan nilai AUC tertinggi sebesar 86%, precision 72,14%, recall 77,20%, dan F1-score 74,40%. Model XGBoost memiliki akurasi tertinggi sebesar 81,98%, meskipun nilai AUC-nya sedikit lebih rendah dibandingkan Random Forest, yaitu 83%. Model LightGBM memiliki performa paling rendah dengan akurasi 79,28% dan F1-score 71,10%.

Berdasarkan hasil evaluasi, dapat disimpulkan bahwa model Random Forest menjadi pilihan terbaik untuk prediksi diabetes dalam penelitian ini, karena menunjukkan keseimbangan terbaik antara berbagai metrik evaluasi.

4. KESIMPULAN

Penelitian ini mengevaluasi kinerja tiga algoritma machine learning, yaitu Random Forest, XGBoost, dan LightGBM, dalam memprediksi risiko diabetes mellitus. Hasil evaluasi menggunakan metrik akurasi, presisi, recall, F1-score, dan AUC menunjukkan bahwa algoritma Random Forest memberikan performa terbaik dengan nilai AUC sebesar 86%, diikuti oleh XGBoost yang mencapai 83% dan LightGBM dengan 82%. Random Forest menunjukkan keseimbangan optimal dalam mendeteksi pasien dengan dan tanpa diabetes, sehingga lebih andal dalam diagnosis dini. Dengan tingkat akurasi yang tinggi, model ini dapat menjadi alat bantu yang efektif dalam pengambilan keputusan medis untuk mendukung pencegahan dan penanganan dini diabetes mellitus.

5. SARAN

Untuk meningkatkan akurasi dan cakupan penelitian, disarankan penggunaan dataset yang lebih besar dan beragam, mencakup berbagai populasi dan faktor risiko. Eksplorasi algoritma lain atau pendekatan hybrid juga dapat meningkatkan performa model. Teknik penanganan data tidak seimbang seperti kombinasi SMOTE dan undersampling bisa diterapkan untuk hasil prediksi yang lebih akurat, khususnya pada kelompok minoritas. Selain itu, implementasi model dalam bentuk aplikasi atau web direkomendasikan agar dapat dimanfaatkan langsung oleh tenaga medis dan masyarakat sebagai alat diagnosis dini yang praktis.

UCAPAN TERIMA KASIH

Penulis menyampaikan terima kasih kepada Tim Redaksi Jurnal atas dukungan dalam publikasi artikel ini, serta kepada NIDDK atas penyediaan dataset. Penghargaan juga disampaikan kepada Institut Informatika dan Bisnis Darmajaya atas fasilitas dan dukungan teknis, serta semua pihak yang turut berkontribusi dalam penelitian ini. Semoga hasilnya bermanfaat bagi perkembangan ilmu, khususnya di bidang kesehatan dan teknologi prediktif.

DAFTAR PUSTAKA

[1] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," *Comput Methods Programs Biomed*, vol. 220, Jun. 2022, doi: 10.1016/j.cmpb.2022.106773.

[2] World Health Organization, "Diabetes." Accessed: Apr. 17, 2025. [Online]. Available: https://www.who.int/health-topics/diabetes#tab=tab_1

[3] M. Kawarkhe and P. Kaur, "Prediction of Diabetes Using Diverse Ensemble Learning Classifiers," *Procedia Comput Sci*, vol. 235, pp. 403–413, 2024, doi: 10.1016/j.procs.2024.04.040.

[4] N. Ahmed *et al.*, "Machine learning based diabetes prediction and development of smart web application," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 229–241, 2021, doi: 10.1016/j.ijcce.2021.12.001.

[5] D. Fernández-Edreira, J. Liñares-Blanco, and C. Fernandez-Lozano, "Machine Learning analysis of the human infant gut microbiome identifies influential species in type 1 diabetes," *Expert Syst Appl*, vol. 185, p. 115648, 2021, doi: 10.1016/j.eswa.2021.115648.

[6] International Diabetes Federation, "IDF Diabetes Atlas," Jan. 23, 2025. [Online]. Available: https://diabetesatlas.org/

[7] Portal Informasi Indonesia, "Cegah Dini Ancaman Diabetes." Accessed: Apr. 17, 2025. [Online]. Available: https://indonesia.go.id/kategori/editorial/8401/cegah-dini-ancaman-diabetes?lang=1

[8] A. R. Hidayat, H. Hanipah, A. Nurjanah, and R. Farizki, "Upaya untuk Mencegah Penyakit Diabetes pada Usia Dini," *Jurnal Forum Kesehatan : Media Publikasi Kesehatan Ilmiah*, vol. 11, no. 2, pp. 63–69, 2021, doi: 10.52263/jfk.v11i2.229.

[9] N. C. for C. D. P. and H. P. Center for Disease Control and Prevention, "On Your Way to Preventing Type 2 Diabetes," Jan. 2025, [Online]. Available: https://www.cdc.gov/diabetes/pdfs/prevent/On-your-way-to-preventing-type-2-diabetes.pdf

[10] National Institute of Diabetes and Digestive and Kidney Diseases, "Preventing Type 2 Diabetes." Accessed: Apr. 17, 2025. [Online]. Available: https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-type-2-diabetes

[11] S. Adi Soelistijo *et al., Pedoman Pengelolaan dan Pencegahan Diabetes Melitus Tipe 2 Dewasa - 2019*, Desember 2019. Indonesia: PB PERKENI, 2019. [Online]. Available: https://pbperkeni.or.id/wp-content/uploads/2021/06/Pedoman-Pengelolaan-DM-Tipe-2-Dewasa-di-Indonesia-eBook-PDF.pdf

[12] R. P. Fadhillah, R. Rahma, A. Sepharni, R. Mufidah, B. N. Sari, and A. Pangestu, "Klasifikasi Penyakit Diabetes Mellitus Berdasarkan Faktor-Faktor Penyebab Diabetes menggunakan Algoritma C4.5," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 7, no. 4, pp. 1265–1270, 2022, doi: 10.29100/jipi.v7i4.3248.

[13] D. Sri Rahayu, J. Afifah, and S. Intan, "Klasifikasi Penyakit Diabetes Melitus Menggunakan Algoritma C4.5, Support Vector Machine (SVM) dan Regresi Linear : Classification Diabetes Mellitus Using C4.5 Algorithm, Support Vector Machine (SVM) and Linear Regression," 2023. [Online]. Available: https://journal.irpi.or.id/index.php/sentimas

[14] 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,"

Indonesian Journal on Software Engineering (IJSE), vol. 7, no. 1, pp. 63–68, 2021, [Online]. Available: http://ejournal.bsi.ac.id/ejurnal/index.php/ijse63

[15] A. Pratama, A. C. Nurcahyo, and L. Firgia, "Penerapan Machine Learning Dengan Algoritma Logistik Regresi Untuk Memprediksi Diabetes," 2023.

[16] A. Agung, A. Daniswara, I. Kadek, and D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," *Journal of Informatics and Computer Science*, vol. 05, 2023.

[17] M. Z. Haq, C. S. Octiva, A. Ayuliana, U. W. Nuryanto, and D. Suryadi, "Algoritma Naïve Bayes untuk Mengidentifikasi Hoaks di Media Sosial," *Jurnal Minfo Polgan*, vol. 13, no. 1, pp. 1079–1084, Jul. 2024, doi: 10.33395/jmp.v13i1.13937.

[18] F. Putra, H. F. Tahiyat, R. M. Ihsan, R. Rahmaddeni, and L. Efrizoni, "Penerapan Algoritma K-Nearest Neighbor Menggunakan Wrapper Sebagai Preprocessing Untuk Penentuan Keterangan Berat Badan Manusia," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 1, pp. 273–281, Jan. 2024, doi: 10.57152/malcom.v4i1.1085.

[19] T. Zulhaq Jasman, E. Hasmin, C. Susanto, and W. Musu, "Perbandingan Logistic Regression, Random Forest, dan Perceptron pada Klasifikasi Pasien Gagal Jantung," *Oktober*, vol. 14, no. 3, pp. 271–286, 2022, doi: 10.22303/csrid.14.3.2022.271-286.

[20] A. G. P. Sidhawara, "Male Fertility Classification using Machine Learning and Oversampling Techniques," *Jurnal Buana Informatika*, vol. 15, no. 01, pp. 1–10, 2024, doi: 10.24002/jbi.v15i1.8718.

[21] N. P. Y. T. Wijayanti, E. N. Kencana, and I. W. Sumarjaya, "SMOTE: Potensi dan Kekurangannya Pada Survei," *E-Jurnal Matematika*, vol. 10, no. 4, p. 235, Nov. 2021, doi: 10.24843/mtk.2021.v10.i04.p348.

[22] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[23] S. L. Cichosz, M. H. Jensen, O. Hejlesen, S. D. Henriksen, A. M. Drewes, and S. S. Olesen, "Prediction of pancreatic cancer risk in patients with new-onset diabetes using a machine learning approach based on routine biochemical parameters," *Comput Methods Programs Biomed*, vol. 244, p. 107965, 2024, doi: 10.1016/j.cmpb.2023.107965.

[24] A. Arisusanto, N. Suarna, and G. Dwilestari, "Analisa Klasifikasi Data Harga Handphone Menggunakan Algoritma Random Forest Dengan Optimize Parameter Grid," *Jurnal Teknologi Ilmu Komputer*, vol. 1, no. 2, pp. 43–47, 2023, doi: 10.56854/jtik.v1i2.51.

[25] S. Devella and F. Novia Rahmawati, "Implementasi Random Forest Untuk Klasifikasi Motif Songket Palembang Berdasarkan SIFT," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 7, no. 2, 2020, [Online]. Available: http://jurnal.mdp.ac.id

[26] M. Syukron, R. Santoso, and T. Widiharih, "Perbandingan Metode SMOTE Random Forest dan SMOTE XGBoost untuk Klasifikasi Tingkat Penyakit Hepatitis C pada Imbalance Class Data," 2020, [Online]. Available: https://ejournal3.undip.ac.id/index.php/gaussian/

[27] S. Diantika, "Penerapan Teknik Random Oversampling untuk Mengatasi Imbalance Class dalam Klasifikasi Website Phishing Menggunakan Algoritma LightGBM," 2023.