JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 3, September 2024, Pp. 1133-1144

PENGGABUNGAN K-NEAREST NEIGHBORS DAN LIGHTGBM UNTUK PREDIKSI DIABETES PADA DATASET PIMA INDIANS: MENGGUNAKAN PENDEKATAN EXPLORATORY DATA ANALYSIS

Arvi Pramudyantoro*[1)], Ema Utami[2)], Dhani Ariatmanto[3)]
1. Universitas Amikom Yogyakarta, Indonesia
2. Universitas Amikom Yogyakarta, Indonesia
3. Universitas Amikom Yogyakarta, Indonesia

Article Info
Kata Kunci: Diabetes Mellitus; EDA; KNN; LightGBM; Pima Indians Diabetes

Keywords: *Diabetes Mellitus; EDA; KNN; LightGBM; Pima Indians Diabetes*

Article history:
Received 7 June 2024
Revised 14 July 2024
Accepted 21 August 2024
Available online 1 September 2024

DOI :
https://doi.org/10.29100/jipi.v9i3.4966

* Corresponding author.
Corresponding Author
E-mail address:
arvipramudyantoro27@students.amikom.ac.id

ABSTRAK
Diabetes Melitus merupakan masalah kesehatan yang signifikan di seluruh dunia. Dengan menggabungkan algoritma K-Nearest Neighbors (KNN) dan Light Gradient Boosting Machine (LightGBM),penelitian ini menyajikan pendekatan baru untuk meningkatkan prediksi diabetes. Kumpulan data Indian Pima, yang terkenal dengan intrik dan signifikansinya dalam penelitian diabetes, menjadi subjek penelitian ini. Untuk menyelidiki pola dan hubungan dalam data, penelitian ini menggunakan analisis data eksploratif, atau EDA. Pra-pemrosesan data yang komprehensif, yang mencakup pengkodean, normalisasi, dan penanganan nilai yang hilang, adalah yang berikutnya. Karena KNN dan LightGBM cocok dengan fitur kumpulan data ini, maka keduanya dipilih. Performa model dioptimalkan melalui penggunaan teknik pengoptimalan seperti Pencarian Acak dan Pencarian Grid untuk mengubah hyperparameter. Metrik seperti skor F1, kurva ROC, analisis presisi-recall, dan akurasi-presisi digunakan untuk menilai model. Hasilnya menunjukkan peningkatan signifikan dalam keakuratan prediksi diabetes, yang menunjukkan bahwa penggunaan LightGBM bersama dengan KNN dan EDA secara hati-hati dapat meningkatkan akurasi prediksi. Khususnya bila dipertimbangkan dalam konteks data kesehatan yang rumit, temuan ini secara signifikan memajukan deteksi penyakit kronis. Menggunakan kumpulan data Pima Indians, algoritma KNN dan LightGBM bekerja sama untuk mencapai akurasi tertinggi sebesar 90,6%.

ABSTRACT
Diabetes mellitus is a significant worldwide health concern. By combining the K-Nearest Neighbors (KNN) dan Light Gradient Boosting Machine (LightGBM) algorithms, this study presents a novel approach to improving diabetes prediction. The Pima Indians dataset, which is renowned for its intrigue and significance in diabetes research, is the subject of this study. To investigate patterns and relationships in the data, this study employs exploratory data analysis, or EDA. Comprehensive data pre-processing, which includes coding, normalization, and handling of missing values, comes next. Because KNN and LightGBM match the features of this dataset, they were selected. The model performance is optimized through the use of optimization techniques like Random Search and Grid Search to modify hyperparameters. Metrics such as F1-score, ROC curves, precision-recall analysis, and accuracy-precision are used to assess models. The outcomes demonstrate a notable increase in the accuracy of diabetes predictions, indicating that the use of LightGBM in conjunction with KNN and cautiously EDA may enhance prediction accuracy. Particularly when considered within the context of intricate health data, these findings significantly advance the detection of chronic illnesses. Using the Pima Indians dataset, the KNN and LightGBM algorithms worked together to achieve the highest accuracy of 90.6%.

.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 3, September 2024, Pp. 1133-1144

I. PENDAHULUAN

Saat ini cara masyarakat hidup dalam masyarakat berubah drastis, baik pada orang dewasa maupun remaja. Kategori makanan yang paling banyak dikonsumsi adalah makanan cepat saji dan makanan instan. Diabetes melitus merupakan salah satu penyakit yang diakibatkan oleh banyak mengonsumsi makanan yang mengandung gula. Porsi makan dan jadwal makan harus dikontrol agar kadar gula darah tetap stabil. Porsi makan sebaiknya dikurangi untuk membantu mengontrol gula darah, sedangkan porsi makan lebih besar dapat memperburuk komplikasi diabetes melitus. [18]. Ciri utama diabetes, disebut juga diabetes melitus, adalah ketidakmampuan tubuh menyerap atau memetabolisme gula darah. Dalam keadaan normal, tubuh seseorang secara alami dapat memproduksi insulin, yang membantu menjaga kadar gula darah di atas ambang batas normal [2].

Resistensi insulin atau kurangnya respon sel terhadap insulin dapat menyebabkan berkembangnya diabetes. Kadar glukosa darah akan meningkat akibat hal ini. Lapar, rasa haus yang meningkat, dan sering buang air kecil adalah gejala umum diabetes. Federasi Diabetes Internasional (IDF) baru-baru ini merilis data yang menunjukkan bahwa, di antara 10 negara teratas, Indonesia memiliki persentase penderita diabetes tertinggi ketujuh. Pada tahun 2020, 10,8 juta penduduk Indonesia, atau 6,2% dari total penduduk negara ini, diperkirakan menderita diabetes. Pasien laki-laki tampaknya lebih banyak menderita diabetes dibandingkan pasien perempuan [1].

Berdasarkan data Riset Kesehatan Dasar Indonesia (Riskesdas) tahun 2018, diketahui bahwa masyarakat Indonesia persentase konsumsi makanan manis (87,9%) dan minuman manis (91,49%) sangat tinggi. Padahal ada pedoman untuk membatasi asupan gula harian. Peraturan Menteri Kesehatan Nomor 30 Tahun 2013 menyatakan bahwa 10% dari total energi (atau 200 kkal) harus dikonsumsi oleh setiap individu per hari dalam bentuk gula. Jumlah gula tersebut sama dengan empat sendok makan atau lima puluh gram per orang per hari. Mengonsumsi makanan dan minuman manis dan manis secara berlebihan setiap hari dapat menyebabkan sejumlah masalah kesehatan, termasuk kemungkinan lebih tinggi terkena diabetes melitus. Salah satu penyakit kronis yang paling banyak memakan korban jiwa di Indonesia adalah diabetes melitus. Institute for Health Metrics and Evaluation melaporkan bahwa, dengan sekitar 57,42 kematian per 100.000 penduduk, diabetes menempati peringkat ketiga penyebab kematian utama di Indonesia pada tahun 2019. Menurut data dari International Diabetes Federation (IDF), jumlah penduduk Indonesia yang menderita diabetes mencapai meningkat pesat selama sepuluh tahun sebelumnya. Diperkirakan pada tahun 2045 akan terdapat 28,57 juta jiwa, meningkat 47% dibandingkan tahun 2021 yang berjumlah 19,47 juta jiwa.

Langkah penting pertama dalam pengobatan diabetes adalah diagnosis dan klasifikasi kondisi tersebut. Pemeriksaan klinis, yang meliputi pengujian laboratorium seperti tes glukosa atau gula darah, digunakan untuk memeriksa pasien diabetes [4]. Namun, teknik berbasis pembelajaran mesin telah mendapat perhatian sebagai teknik alternatif untuk diagnosis dan kategorisasi diabetes karena kemajuan teknologi dan semakin tersedianya data medis. Menurut penelitian terbaru, metode pembelajaran mesin dapat secara signifikan meningkatkan prognosis penyakit kronis seperti diabetes [3]. Penggunaan pembelajaran mesin dalam diagnosis dan pengobatan diabetes melitus menghadirkan peluang baru untuk perbaikan. diabetes mellitus. Dalam karya ini, enam algoritma pembelajaran mesin yang berbeda digunakan untuk menganalisis kumpulan data Pima Indian [19]. Secara khusus, banyak penelitian telah dilakukan dengan menggunakan dataset Pima Indians Diabetes untuk membuat model prediktif [5]. Meskipun informasi klinis dalam kumpulan data ini sangat penting untuk menentukan risiko diabetes, kompleksitas dan sifat tidak seimbangnya membuat analisis menjadi sulit [6].

(EDA) adalah suatu metode atau prosedur untuk memeriksa data dengan tujuan mengidentifikasi ciri-ciri dan kecenderungan yang ada. EDA biasanya dilakukan sebelum penggunaan model yang lebih kompleks atau statistik. Mendapatkan pemahaman menyeluruh tentang data dan menemukan pola, anomali, ketergantungan, atau tren apa pun adalah tujuan utama analisis data eksplorasi (EDA). Teknik visualisasi data, seperti grafik, diagram, dan plot, sering digunakan dalam metode EDA untuk menyajikan data dengan cara yang mudah dipahami [20]. Langkah pertama dalam pemrosesan data, yang dikenal sebagai analisis data eksplorasi (EDA), melibatkan penyelidikan dan pemahaman menyeluruh terhadap kumpulan data yang perlukan. Kemudian EDA dapat untuk menemukan korelasi, wawasan, dan pola yang membantu terciptanya model klasifikasi yang lebih tepat dan efektif [7].

Salah satu algoritma yang populer dan mudah digunakan adalah K-Nearest Neighbors (KNN). Berdasarkan gagasan bahwa objek serupa biasanya merupakan tetangga terdekat satu sama lain dalam ruang fitur, KNN beroperasi [8]. Dalam klasifikasi diabetes, KNN dapat memprediksi kemungkinan seseorang terkena diabetes berdasarkan data dari pasien dengan gejala atau faktor risiko yang sebanding [9].

Regresi, klasifikasi, dan aplikasi *Machine Learning* lainnya dapat ditangani oleh Algoritma Light Gradient Boosting Machine (LightGBM) yang terdistribusi dan berkinerja tinggi berbasis *Decision Tree* [10].

Industri perawatan kesehatan mempunyai potensi besar untuk menggunakan kemajuan terkini dalam analisis data dan algoritma pembelajaran mesin, terutama dalam diagnosis dan klasifikasi diabetes. Algoritma K-NN adalah teknik klasifikasi yang populer dan sederhana. Sedangkan, algoritma LightGBM merupakan sebuah algoritma yang semakin banyak digunakan karena efisiensi dan kecepatannya untuk menangani kumpulan data yang cukup besar [11]. Meskipun K-NN dan LightGBM menunjukkan kinerja klasifikasi yang menjanjikan, belum ada penelitian komprehensif dan perbandingan kriteria mereka dalam kumpulan data Pima Indian untuk klasifikasi diabetes yang telah dilakukan [12].

Data klinis yang luas dari populasi dengan prevalensi diabetes tipe 2 yang tinggi, kumpulan data Pima Indians Diabetes bersifat khas dan signifikan dalam bidang penelitian diabetes. Delapan variabel klinis dimasukkan dalam dataset ini, yang terdiri dari 768 sampel: usia, ketebalan kulit, tekanan darah, indeks massa tubuh, (BMI), kadar insulin, riwayat diabetes keluarga dan jumlah kehamilan [4]. Kumpulan data ini istimewa karena tidak merata dan kompleksnya. Hal ini menyulitkan analisis, namun juga menawarkan peluang besar untuk menggali lebih dalam dan menciptakan model prediktif yang secara akurat memprediksi risiko diabetes.

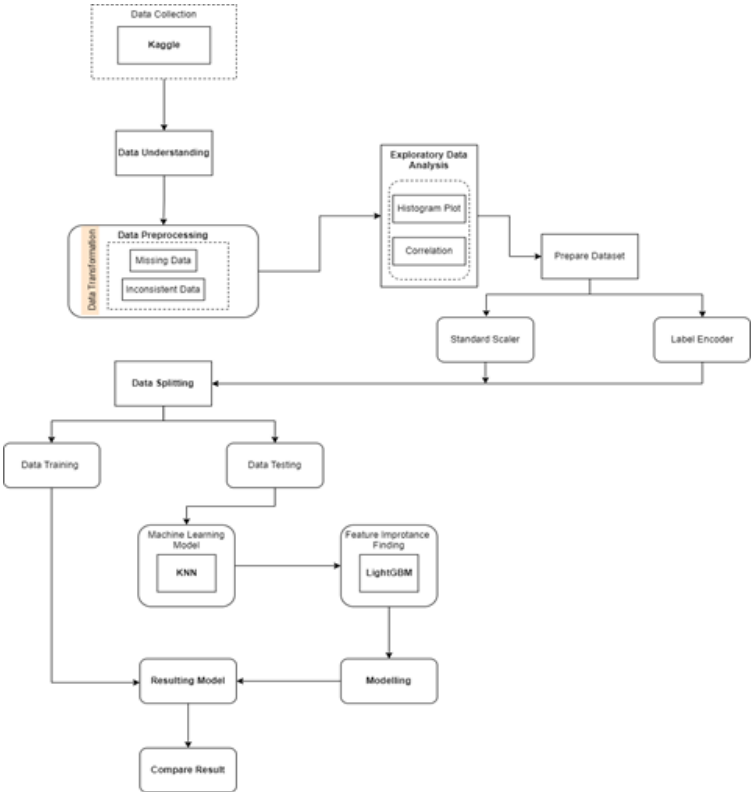
Penggunaan algoritma KNN dan LightGBM menggunakan pendekatan klasifikasi yang berbeda, sulit untuk menggabungkan keduanya. Sederhana dan berpusat pada kedekatan data adalah KNN; lebih kompleks dan menggunakan metode berbasis pohon keputusan adalah LightGBM. Kombinasi ini relevan dengan diabetes karena memperkuat akurasi prediksi dan relevansi klinis dengan memanfaatkan manfaat LightGBM untuk menangani kompleksitas data besar dan algoritma KNN untuk memahami pola lokal [13]. Penelitian ini juga menyoroti betapa pentingnya melakukan analisis data eksplorasi, atau EDA, untuk memahami data sepenuhnya sebelum menggunakan model ini.

II. METODE PENELITIAN

Pada penelitian ini menggunakan metode eksploratif. Oleh karena itu, pada tahap pada pemrosesan penelitian data dapat menghasilkan beberapa langkah meliputi :

A. Tahapan Penelitian

Pada tahapan ini berguna untuk menyelesaikan masalah pada penelitian yang ada secara terstruktur. Pada Gambar 1 dibawah menampilkan tahapan proses yang dilakukan oleh penelitian ini.



Gambar 1. Tahapan Penelitian

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 3, September 2024, Pp. 1133-1144

B. *Pengumpulan Data*

Diabetes Indian Pima yang digunakan dalam penelitian ini dapat diakses secara bebas dan digunakan secara luas dalam penelitian medis terkait diabetes.

Dataset yang akan digunakan ditunjukkan pada Tabel 1 di bawah ini. 768 sampel yang terdiri dari kumpulan data ini memiliki delapan karakteristik klinis: usia, ketebalan kulit, tekanan darah, indeks massa tubuh, (BMI), kadar insulin, riwayat diabetes keluarga dan jumlah kehamilan. Jika hasil akhirnya adalah 1, berarti ada diabetes; bila hasilnya 0, diabetes tidak ada.

TABEL I
DATASET

No.	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33,6	0,627	50	1
1	1	85	66	29	0	26,6	0,351	31	0
2	8	183	64	0	0	23,3	0,672	32	1
3	0	100	88	60	110	46,8	0,962	31	0
4	1	79	75	30	0	32	0,396	22	0
5	4	148	60	27	318	30,9	0,15	29	1
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…
767	0	137	40	35	168	43,1	2,288	33	1

C. *Pra-pemrosesan*

Pada tahapan pra-pemrosesan melibatkan sebuah langkah utama dalam mengimplementasikan data prepocessing:

- Pembersihan data : data yang salah atau tidak lengkap, akan dihapus
- Transformasi data : data akan diubah ke format atau skala yang sudah ditentukan
- Normalisasi dan standarisasi : mengubah data ke skala yang sama agar memudahkan dalam analisis

Teknik imputasi digunakan dalam penelitian ini, yang bertujuan untuk mengatasi *(missing value)* atau data yang kosong/hilang, yang melibatkan penggantian nilai dengan teknik imputasi dengan median atau mean. Hal ini menandakan jika nilai-nilai yang hilang dalam dataset diganti dengan nilai median (rata-rata). Proses imputasi menjadi hal penting agar bisa memastikan integritas dan kelengkapan dataset sebelum dilakukannya analisis lebih lanjut, dikhususkan pada konteks pembelajaran mesin.

D. *Metode Analisis Data*

Sebuah prosedur pada penelitian tentang data yang dikumpulkan ataupun diperoleh berfungsi dalam menyelesaikan suatu masalah penelitian. Dua metode, masing-masing dengan persetujuannya sendiri, digunakan dalam analisis penelitian ini.

Untuk memperoleh pemahaman data yang komprehensif, metode eksplorasi digunakan di awal alur penelitian. Menemukan pola, anomali, ketergantungan, atau tren dalam data adalah tujuan utama EDA. Untuk membantu memahami data, EDA sering menggunakan visualisasi data seperti grafik, diagram, dan plot. Melalui prosedur ini, investigasi dapat menghasilkan saran yang didasarkan pada pemahaman yang diperoleh dari informasi, sehingga mendukung kemajuan model klasifikasi yang lebih unggul dan tepat. Penggunaan model pembelajaran mesin yang dengan penelitian yang dilakukan yaitu dengan penggabungan algoritma KNN dengan LightGBM menjadi teknik analisis selanjutnya. Algoritma LightGBM berperan dalam mengoptimalkan akurasi hasil melalui teknik *Grid Search* dan *Random Search* untuk tuning hyperparameter. KNN diuji menggunakan varian nilai k dari jumlah tetangga terdekat.

E. *Evaluasi*

Untuk mengukur keakuratan prediksi, performa model akan dinilai menggunakan berbagai metrik. Langkah-langkah ini memberikan gambaran komprehensif tentang seberapa baik model yang dibuat mampu memprediksi diabetes dalam kumpulan data suku Indian Pima. Penerapan metrik ini diliput jelas dan informatif, seperti:

- Confusion Matrix

  Merupakan alat ukur yang sering digunakan dalam klasifikasi. Dengan menggunakan kumpulan data Pima Indians, metrik ini digunakan sebagai gambaran secara komprehensif atau mengukur seberapa baik model tersebut memprediksi diabetes 24 [21]. Penelitian ini bertujuan untuk mencari keseimbangan diantara akurasi

dan model agar mampu mengidentifikasi kasus dari diabetes dalam analisis data. Dibawah ini akan menunjukkan sebuah tabel *Confusion Matrix* yang telah dijelaskan pada Tabel 2.

TABEL II
CONFUSION MATRIX

		Prediction Class	
		+	-
Actual Class	+	True Positive (TP)	False Negative (FN)
	-	False Positive (FP)	True Negative (TN)

Yaitu:

1. *True Positive* (TP) = Hasil nilai positif yang sesuai dengan true positif yang diperoleh dari sistem prediksi.
2. *True Negative* (TN) = Sistem menghasilkan hasil negatif yang konsisten dengan hasil negatif sebenarnya.
3. *False Positive* (FP) = Ketika memprediksi sesuatu yang positif, tetapi hasil pada awalnya negatif.
4. *False Negative* (FN) = Hasil prediksi sistem negatif, namun sebelumnya positif.

Setelah perolehan hasil Confusion Matrix, matriks tersebut diukur menggunakan instrumen yang memberikan recall, akurasi, presisi, dan f1-score.

- Akurasi adalah kriteria mendasar untuk menentukan seberapa baik suatu model dapat menghasilkan prediksi umum. Ini memberikan hasil dari prediksi secara akurat (True Positives dan True Negatives) untuk setiap sampel.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

- Presisi merupakan sebuah metrik yang menilai seberapa akurat prediksi positif suatu model. Ini mewakili proporsi True Positives terhadap seluruh positif (True Positives dikurangi False Positives).

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

- Recall (sensitivitas) mengukur kemampuannya untuk mengenali setiap contoh positif yang sebenarnya. Ini mewakili proporsi True Positives terhadap semua kejadian positif (True Positives dikurangi False Negatives).

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

- F1-Score merupakan gabungan antara Recall dan Presisi. Hal ini dapat menunjukkan hasil keseimbangan pada tiap matrix untuk melihat semua hasil positif asli. F1-Score yaitu hasil pembagian dari Precision dan Recall.

$$F1\,Score = \frac{2(Recall*Precission)}{(Recall+Precission)} \quad (4)$$

Penelitian ini juga menggunakan Kurva Precision-Recall dan Kurva Karakteristik Operasi Penerima (*ROC Curve)* selain metrik evaluasi yang disebutkan di atas untuk memberikan penilaian yang lebih menyeluruh terhadap kinerja model.

- Kurva ROC: menggunakan ambang prediksi yang berbeda untuk membedakan antara kelas positif dan negatif. Untuk mengukur efektivitas diferensiasi kelas, digunakan area di bawah kurva ROC, atau AUC-ROC. Ketika AUC-ROC mendekati 1, performa spesifikasi antarkelas berada pada kondisi terbaiknya.
- Kurva Precision Recall: Ini menunjukkan bagaimana, pada ambang batas prediksi yang berbeda, kapasitas model untuk mengidentifikasi hal-hal positif yang sebenarnya diseimbangkan dengan kapasitasnya untuk mengidentifikasi semua hal-hal positif yang sebenarnya. Ketika kelas-kelas positif jarang terjadi atau tidak seimbang, kurva ini menjadi signifikan. Nilai AUC untuk Precision-Recall yang mendekati 1 menandakan kinerja luar biasa dalam mendeteksi hal positif asli.

Memahami kemampuan model untuk membedakan kelas menjadi lebih mudah dengan kurva ROC dan analisis Precision-Recall, terutama dalam hal prediksi diabetes.Penelitian ini dapat memberikan pemahaman menyeluruh tentang kualitas prediktif model kami dalam penelitian ini dengan menggunakan metrik dan analisis kurva ini.

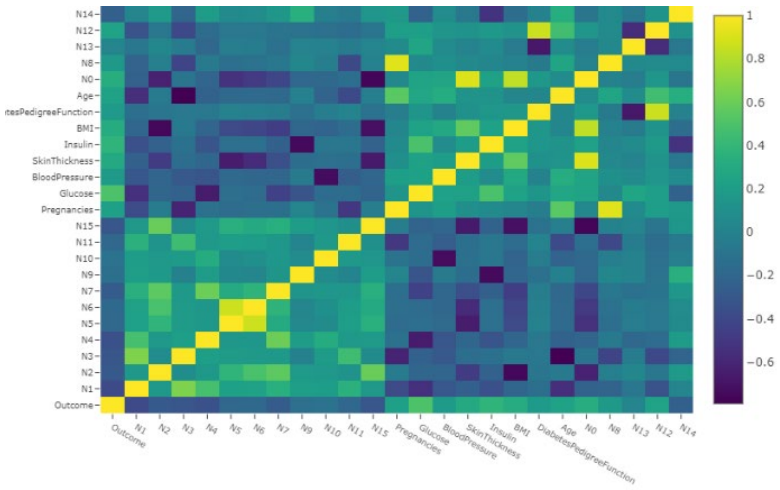
JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 3, September 2024, Pp. 1133-1144

III. HASIL DAN PEMBAHASAN

Tahap selanjutnya adalah pengolahan data menggunakan *machine learning* dengan algoritma KNN dan LightGBM setelah diolah, dianalisis, dan data yang hilang diganti dengan EDA. Namun semua data perlu dikonversi ke rentang 0 hingga 1 sebelum dapat diproses. Pada titik ini, normalisasi data juga dikenal sebagai fitur *Standard Scaler* digunakan.

A. Standard Scaler dan Label Encoder

Penggunaan pada *StandardScaler* akan membuat penyimpangan yang signifikan lebih kecil kemungkinannya. Min-Max Scaler dapat digunakan dalam metode ini. Gambar 2 di bawah menampilkan hasil yang dihasilkan *StandardScaler* bersama dengan *Heatmap*.



Gambar 2 Heatmap *StandarScaler*

Langkah berikutnya ialah mengolah data pengujian *(Data Testing)* dengan menggunakan K-Fold sebagai validasi model. K-Fold Validation, digunakan untuk mengacak data asli diubah menjadi nilai *k* hingga ditemukan ukuran yang sama. Untuk menguji model, data pelatihan k-1 dan satu sampel yang disimpan sebagai data validasi akan diambil dari sampel nilai *k* [21]. Selanjutnya dilakukan *k* iterasi proses validasi silang, dengan satu validasi data untuk setiap k subsampel. Data kemudian dapat dirata-ratakan untuk menyelesaikan proses. Manfaat dari metode ini daripada sub-sampling acak berulang yaitu seluruh observasi dapat dipakai untuk data latih dan validasi, dan dari observasi yang digunakan satu kali tiap validasi.

B. Random Search + LightGBM

Pada penelitian ini akan melewati tahap yaitu melakukan pengujian terhadap algoritma LightGBM dengan menggunakan *RandomSearch*, fungsi dari fitur ini adalah untuk membantu dalam mencari kombinasi pada setiap hyperparameter yang optimal secara acak agar memperoleh performa model yang lebih baik.

Untuk menentukan kombinasi parameter mana yang akan memberikan hasil terbaik untuk model LightGBM, metode ini digunakan untuk mencari kombinasi tersebut secara acak. Pencarian ini lebih efisien dibandingkan *Grid Search* karena menemukan kombinasi parameter yang optimal tanpa mengharuskan pengguna mencoba setiap kemungkinan kombinasi.

Hasil percobaan pertama diperoleh sebagai berikut, yang ditampilkan pada Tabel 3 *Confusion Matrix* di bawah ini:

TABEL III
CONFUSION MATRIX LIGHTGBM

		Prediction Class	
		+	-
Actual Class	+	226	42
	-	38	462

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 3, September 2024, Pp. 1133-1144

Langkah selanjutnya, apabila sebuah *Matrix* membuahkan hasil, adalah menggunakan *Randomize Search* dan LightGBM untuk menghitung hasil dari akurasi dan lain-lain. Pada Tabel 3 di bawah ini menunjukkan hasil dari LightGBM dan RandomSearch.

TABEL IV
HASIL PERHITUNGAN MATRIX LIGHTGBM

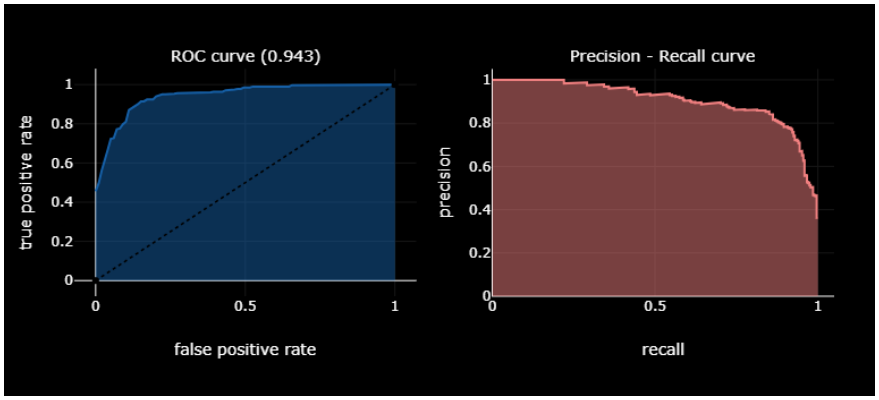
Accuracy	Precision	Recall	F1 Score
0,8496	0,8433	0,8561	0,8958

Pada Tabel 4 diatas menampilkan hasil perhitungan metrik evaluasi untuk algoritma LightGBM yang dioptimalkan dengan RandomSearch. Metrik-metrik yang disajikan dalam grafik batang ini mencakup F1-Score, Recall, Precision, dan Accuracy, dengan nilai-nilai berikut:

- F1-Score: 0,8496, yang mengindikasikan keseimbangan antara presisi dan recall dalam model. Nilai ini penting untuk konteks di mana kedua aspek penting dan tidak ada yang boleh dikorbankan untuk yang lain.
- Recall: 0,8433, yang menunjukkan seberapa baik model mengidentifikasi semua kasus positif yang sebenarnya. Ini kritis dalam kondisi medis di mana tidak mengidentifikasi kondisi positif (misalnya, penyakit) bisa memiliki konsekuensi serius.
- Precision: 0,8561, yang mengukur akurasi model dalam memprediksi kasus positif. Nilai tinggi di sini mengurangi jumlah false positives, yang dapat mengurangi risiko intervensi medis yang tidak perlu.
- Accuracy: 0,8958, ini adalah ukuran keseluruhan dari kinerja model dalam mengklasifikasikan kasus dengan benar dan mengindikasikan kinerja model yang solid pada dataset ini.

Dari nilai-nilai tersebut, dapat dilihat bahwa model menunjukkan tingkat akurasi yang tinggi dalam mengklasifikasikan data, yang menandakan efektivitasnya dalam konteks aplikasi, seperti dalam penelitian medis atau pengobatan diabetes, di mana model ini telah dioptimalkan untuk memberikan prediksi yang dapat diandalkan.

Pada percobaan untuk pengoptimalan algoritma LightGBM menggunakan RandomSearch, berbagai parameter penting seperti learning rate, jumlah pohon *(trees)*, kedalaman pohon *(tree depth)*, ukuran daun *(leaf size)*, *subsample data*, dan proporsi fitur yang digunakan *(colsample by tree)* diuji secara acak dalam batasan tertentu untuk menemukan kombinasi yang paling efektif. RandomSearch adalah metode yang efisien dalam mengeksplorasi ruang parameter yang luas, menawarkan keunggulan dalam menemukan konfigurasi yang mengoptimalkan metrik kinerja utama seperti F1-Score, Recall, Precision, dan Accuracy, yang penting untuk memvalidasi efektivitas model dalam klasifikasi data.



Gambar 3. Kurva ROC dan Kurva Precision - Recall

Kurva evaluasi kinerja dua model untuk klasifikasi ditunjukkan pada Gambar 3. Dengan nilai AUC *(Area Under the Curve)* sebesar 0,943, kurva *Receiver Operating Characteristic* (ROC) menunjukkan kinerja yang sangat baik. Nilai AUC berkisar antara 0 hingga 1, dimana 1 menunjukkan klasifikasi ideal dan 0,5 setara dengan kinerja pengacakan. Mengklasifikasikan positif sejati pada tingkat yang tinggi sambil menjaga tingkat positif palsu tetap rendah merupakan ciri model yang diwakili oleh kurva ini. Hal ini menunjukkan bahwa pembedaan kelas positif dan negatif dilakukan secara efektif oleh model.

Kurva Precision-Recall ditampilkan di sebelahnya, menunjukkan persentase hasil positif akurat yang terdeteksi model. Recall adalah kapasitas difungsikan sebagai pembeda semua kasus positif asli dan semua hasil positif. Pada Kurva ini menampilkan seberapa baik hasil dari model agar dapat mendeteksi kasus positif tanpa mengurangi

akurasi atau menghasilkan positif palsu dalam jumlah besar. Hal ini semakin menunjukkan bahwa meskipun terjadi peningkatan recall, model tersebut tetap mempertahankan tingkat presisi yang tinggi.

Secara umum, model ini memiliki performa yang sangat baik dalam mengklasifikasikan kumpulan data yang diuji. Kurva presisi-recall yang tetap tinggi pada rentang recall menunjukkan bahwa model tersebut dapat mempertahankan akurasi yang baik bahkan ketika mencoba mengklasifikasikan semua kasus positif yang sebenarnya. Nilai tinggi yang dihasilkan oleh nilai AUC akan menunjukkan bahwa model akhir baik. Untuk meminimalkan jumlah pasien yang tidak mengidap penyakit tersebut namun salah diidentifikasi sebagai mengidap penyakit tersebut (presisi tinggi), penting untuk mengidentifikasi sebanyak mungkin pasien yang benar-benar mengidap penyakit tersebut (recall tinggi). Khususnya dalam konteks medis seperti prediksi diabetes, ini adalah atribut yang diinginkan.

TABEL V. *Cross Validation* RandomSearch dan LughtGBM

Fold	Accuracy	Precision	Recall	F1 Score	ROC AUC
1	0,903	0,915	0,796	0,851	0,945
2	0,864	0,915	0,833	0,811	0,926
3	0,896	0,915	0,833	0,849	0,949
4	0,889	0,915	0,83	0,838	0,944
5	0,928	0,915	0,925	0,899	0,927
Mean	0,896	0,915	0,844	0,85	0,947
Std	0,021	0,915	0,043	0,029	0,015

Hasil dari validasi silang yang dilakukan 5 kali pada algoritma LightGBM pada Tabel 5. Dataset tersebut dipilah 5 bagian yang sama untuk validasi silang. Pada tiap bagian dilakukan pengujian sebanyak satu kali data uji, kemudian 4 bagian sisanya digunakan sebagai data latih*(Training)*. Hasilnya, evaluasi kinerja model diperkuat. Ini adalah sinopsis temuannya:

- Akurasi: Dengan deviasi standar (std) sebesar 2,1% dan akurasi rata-rata sebesar 89,6%, performa model cukup konsisten di berbagai subkumpulan data.
- Presisi: Model secara akurat memprediksi proporsi positif, dengan presisi rata-rata sebesar 85,8% dan deviasi standar sebesar 4,1%.
- Recall: Model mampu mengidentifikasi semua kasus positif yang sebenarnya, dengan rata-rata recall sebesar 84,4% dan standar deviasi sebesar 4,3%.
- F1 Score: Keserasian antara rata-rata presisi dan perolehan menghasilkan skor F1, yaitu rata-rata 85% dengan deviasi standar 2,9% dan mewakili satu metrik kinerja model.
- ROC AUC: Tingkat keberhasilan model dalam mengklasifikasikan kelas secara akurat rata-rata adalah 94,7%, dengan standar deviasi 1,5%.

Gambar tersebut menggambarkan seberapa baik dan konsisten performa model saat membuat prediksi pada kumpulan data, pada akurasi dan area yang stabil. Temuan ini menunjukkan performa keseluruhan model LightGBM yang luar biasa di semua metrik.

C. *LIGHTGBM + K-NEAREST NEIGHBORS*

Kemudian masuk ketahap penggabungan algoritma LightGBM dan KNN. Pada tahapan ini, menggunakan sebuah teknik *GridSearchCV* akan digunakan sebagai mesin pencari parameter KNN yang dapat dibuat dalam *Voting Classifier* dan mengidentifikasi kombinasi yang membuat hasil terbaik dalam hal akurasi. Sementara itu, Voting Classifier ini menghitung prediksi probabilitas dari kedua model dengan menggunakan teknik voting "lunak". Ini adalah ansambel LightGBM dan KNN. Mengingat bobotnya adalah [1,1], maka dapat disimpulkan bahwa kontribusi kedua model terhadap prediksi akhir adalah sama. Uji matriks yang sama dengan pengujian sebelumnya kemudian akan digunakan untuk mencari hasilnya. Tabel 5 di bawah menampilkan hasil dari penggabungan kedua algoritma yang dilakukan.

TABEL VI
CONFUSION MATRIX LIGHTGBM DAN KNN

		Prediction Class	
		+	-
Actual Class	+	230	38
	-	34	466

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 3, September 2024, Pp. 1133-1144

Pada Tabel 6 diatas untuk gabungan algoritma LightGBM dan KNN, berikut adalah penjelasan mengenai arti setiap elemen yang dihasilkan Confusion Matrix 6 [22]:

- True Positive (TP): Pada matriks konfusi diwakili oleh angka 230. Dalam hal ini, kelas positif (1) diprediksi positif oleh model yang akurat.
- False Negative (FN): 38 adalah indikatornya. Dalam contoh ini, model memprediksi kelas positif (1) sebagai kelas negatif (0) dalam kesalahan.
- True Negative (TN): 466 adalah indikator untuk ini. Dalam hal ini, model memprediksi kelas negatif (0) sebagai negatif dengan akurat.
- False Positive (FP): 34 merupakan indikator kondisi ini. Dalam contoh ini, model memprediksi kelas negatif (0) sebagai kelas positif (1) dalam kesalahan.
- 0 (Pred): Jumlah prediksi negatif (0) yang dibuat model ditampilkan pada kolom ini. Jumlah True Negatives (TN) dan False Positives (FP) pada kolom ini masing-masing adalah 466 dan 34.
- 1 (Pred): Banyaknya prediksi positif (1) yang dibuat model ditunjukkan pada kolom ini. Hitungan True Positive (TP) dan False Negative (FN) pada kolom ini masing-masing diwakili oleh angka 230 dan 38.

Dengan menggunakan nilai-nilai diatas, maka dapat dilakukan pehitungan sebagai berikut:

Accuracy = (TP+TN) / (TP+TN+FP+FN)
= (230+466) / (230+34+466+38)
= 696 / 768
= 1.464
= 0,90625 * 100%
= 90,625%

Dengan demikian berdasarkan hasil perhitungan akurasi di atas terlihat bahwa nilai akurasi yang diperoleh sebesar 90,6%. Nilai recall, presisi, dan f1-score harus dihitung selanjutnya:

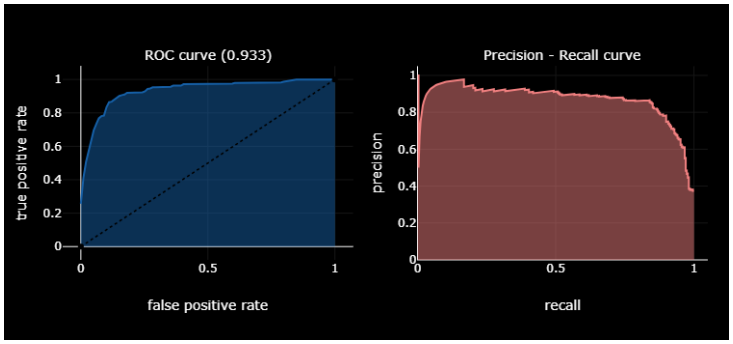
Recall = (TP) / (TP+FN)
= 230 / 268
= 0,852 * 100% = 85,2%

Precision = TP (TP+FP)
= 230 / 264
= 0,8712 * 100% = 87,12%

F1-Score = 2 (0,852 * 0,8712) / 0,852 + 0,8712
= 0,8647 * 100%
= 86,47%

Pada hasil perhitungan matrix diatas mendapatkan masing-masing sebesar, Presisi 87,12%, Recall 85,2%0 dan F1-Score dengan hasil akhir 86,47%

Pada perhitungan algoritma KNN menggunakan GridSearchCV yang digunakan untuk menentukan kombinasi parameter terbaik untuk KNN dalam *Voting Classifier*, yang merupakan bagian dari ensemble model yang juga termasuk LightGBM. Teknik ini melibatkan pencarian melalui berbagai kombinasi parameter yang telah ditetapkan untuk algoritma KNN untuk menemukan yang memberikan akurasi terbaik. *Voting Classifier* menggabungkan prediksi dari LightGBM dan KNN dengan menggunakan metode voting 'soft', yang berarti bahwa klasifikasi tidak hanya didasarkan pada hasil prediksi yang paling umum tetapi juga mempertimbangkan probabilitas prediksi dari kedua model tersebut. Tujuannya adalah untuk memanfaatkan kekuatan prediktif dari kedua model tersebut, dengan harapan bahwa kombinasinya akan memberikan kinerja yang lebih baik daripada setiap model secara terpisah.



Gambar 4. Kurva ROC LightGBM dan KNN

Gambar 5 di atas menampilkan kurva ROC dan Precision-recall yang dihasilkan dengan menggabungkan algoritma LightGBM dan KNN. Kurva ROC merupakan alat evaluasi penting dalam menilai kinerja model klasifikasi, menjelaskan hubungan pada ambang batas prediksi yang berbeda antara spesifisitas (rasio prediksi negatif yang benar terhadap total data negatif) dan sensitivitas (rasio prediksi positif yang benar terhadap total data positif). Model dapat dibedakan menjadi kelas positif dan negative dengan melakukan pengukuran terhadap nilai AUC, atau *Area Under the Curve;* nilai 1 yang menunjukkan bahwa kinerja yang didapatkan bahwa baik. Dalam penelitian ini, kurva ROC untuk model KNN dan LightGBM menunjukkan nilai AUC yang tinggi (0,933% pada penggabungan algoritma KNN dan LightGBM) mengindikasikan bahwa kemampuan pada kedua algoritma tersebut efektif dalam memprediksi diabetes pada pasien Pima Indian. Signifikansi nilai AUC yang tinggi ini sangat penting, menunjukkan potensi model dalam mendukung deteksi dini diabetes, yang berpotensi membantu dalam pengembangan aplikasi untuk pemantauan gejala, alat bantu diagnostik, dan sistem peringatan dini, memberikan kontribusi signifikan dalam pengobatan dan manajemen diabetes.

Kurva Precision-Recall merupakan grafik yang menilai kinerja model klasifikasi dalam konteks di mana ada ketidakseimbangan kelas, dengan fokus pada hubungan antara presisi, yang mengukur akurasi prediksi positif model, dan recall, yang mengukur kemampuan model menangkap kasus positif yang sebenarnya. Kurva yang mencapai pojok kanan atas, yang menandakan presisi dan recall sempurna, adalah ideal, namun pada kenyataannya, peningkatan recall biasanya menyebabkan penurunan presisi. Kurva ini sangat berharga dalam situasi di mana penting untuk mengidentifikasi semua kasus positif meskipun ada kemungkinan meningkatkan jumlah positif palsu, seperti dalam skenario deteksi penyakit serius di mana kesalahan negatif palsu dapat memiliki konsekuensi yang parah.

Performa model gabungan KNN dan LightGBM dalam penelitian ini dianalisis menggunakan kurva ROC yang menunjukkan nilai AUC yang tinggi sebesar 0,933. Kinerja ini mengindikasikan bahwa gabungan kedua algoritma efektif dalam memprediksi diabetes pada pasien Pima Indian. Temuan ini mempunyai konsekuensi klinis yang penting; Model ini memiliki nilai AUC yang tinggi, sehingga menunjukkan bahwa model ini dapat sangat membantu dalam mendorong deteksi dini diabetes. Ini penting karena deteksi dini merupakan faktor kunci dalam pengelolaan diabetes yang berhasil, membantu mencegah perkembangan komplikasi serius dan memungkinkan intervensi yang lebih tepat waktu. Hasil ini juga membuka peluang untuk pengembangan aplikasi seluler untuk pemantauan gejala, alat bantu diagnostik di klinik, dan sistem peringatan dini yang bisa digunakan dalam praktik klinis sehari-hari, sehingga dapat memberikan kontribusi penting dalam pengobatan dan manajemen diabetes.

Berdasarkan kurva Precision-Recall, performa model sebagai hasil penggabungan algoritme KNN dan LightGBM menunjukkan bahwa model tetap mempertahankan presisi tinggi meskipun recall meningkat, yang menunjukkan kemampuan model yang baik dalam mengidentifikasi kasus positif tanpa menghasilkan positif palsu dalam jumlah besar. hasil. Dengan kata lain, model ini mempertahankan tingkat positif palsu yang rendah dan tetap efektif dalam menangkap sebagian besar kasus positif sebenarnya. Hal ini menunjukkan perolehan yang sangat baik dan keseimbangan presisi, yang menunjukkan kinerja yang sangat baik dalam klasifikasi kumpulan data yang diuji. Dalam konteks prediksi diabetes, implikasi klinis dari hal ini sangatlah signifikan; Model yang dapat mengidentifikasi kasus diabetes secara tepat dan cepat dapat membantu profesional medis dalam mendiagnosis pasien, yang sangat penting untuk perencanaan pengobatan dan mencegah komplikasi yang lebih parah dari kondisi tersebut[22].

TABEL VII
PERBANDINGAN HASIL ALGORITMA

	LightGBM	KNN	LigthGBM+KNN
Accuracy	89,6%	83,12%	90,6%
Recall	84,4%	78%	85,2%
Precision	85,8%	79,3%	87,12%
F1-Score	85%	78,6%	86,47%

Pada Tabel 7 diatas merupakan metode gabungan Algoritma LightGBM dan KNN memberikan hasil yang lebih baik dibandingkan masing-masing metode yang digunakan secara terpisah. Gabungan kedua metode ini memberikan peningkatan pada semua metrik yang diukur: akurasi, recall, presisi, dan F1-score. Pemilihan untuk menggabungkan LightGBM dan KNN didasarkan pada kekuatan individu dari kedua algoritma. LightGBM adalah algoritma boosting yang efisien dan efektif untuk menangani data dalam skala besar, sedangkan KNN adalah algoritma pembelajaran berbasis *instance* yang sederhana dan efektif untuk klasifikasi berdasarkan kedekatan fitur.

Kombinasi dari kedua metode tersebut dapat memberikan pendekatan yang lebih robust dalam menangani variasi dalam data dan meningkatkan akurasi prediksi.

Penggabungan kedua model memungkinkan untuk memanfaatkan LightGBM untuk mengidentifikasi struktur kompleks dalam data dengan cepat dan KNN untuk menyempurnakan klasifikasi pada level lokal dengan mempertimbangkan kedekatan sampel-sampel data. Hal ini mungkin memberikan keseimbangan antara bias dan varians, mengurangi kemungkinan overfitting yang bisa terjadi jika hanya menggunakan salah satu metode, dan meningkatkan kemampuan generalisasi model. Kedua model tersebut memiliki kontribusi yang berbeda terhadap hasil akhir, LightGBM dapat mengurangi error bias dengan meningkatkan kompleksitas model, sedangkan KNN dapat mengurangi error varians dengan memastikan bahwa prediksi tidak terlalu spesifik terhadap data latih. Dengan menggabungkan keduanya, model akhir mungkin lebih baik dalam mengklasifikasikan data yang sebelumnya tidak terlihat, yang tercermin dalam peningkatan metrik performa seperti akurasi dan F1-score.

IV. KESIMPULAN

Penelitian ini berhasil meningkatkan prediksi penyakit diabetes dengan menggabungkan algoritma K-Nearest Neighbors (KNN) dan LightGBM menggunakan dataset Pima Indians dengan menghasilkan model yang efektif dan akurat. Penggunaan Exploratory Data Analysis (EDA) untuk memahami karakteristik data dan mengidentifikasi pola yang relevan mampu memprediksi diabetes dengan akurasi yang tinggi. Data diproses secara menyeluruh termasuk penanganan nilai yang hilang *(missing value)*, dan normalisasi pada data. Teknik optimisasi seperti *Random Search* dan *Grid Search* dengan menyesuaikan hyperparameter, agar meningkatkan kinerja model. Model yang dihasilkan mampu memprediksi diabetes dengan akurasi yang tinggi, mencapai 90,6% dengan nilai AUC sebesar 93,3%. Hasil ini menunjukkan potensi besar dalam mendukung deteksi dini diabetes dan dapat memberikan kontribusi penting dalam pengobatan dan manajemen penyakit ini. Kemudian pada penggunaan model ini untuk diagnosis dan pengelolaan diabetes yang lebih akurat dan cepat dapat dilakukan seperti, pengembangan aplikasi seluler untuk pemantauan gejala dan peringatan dini, peningkatan sistem peringatan dini di klinik, dan kemajuan dalam aplikasi diagnostik klinis. Model gabungan ini menawarkan kontribusi signifikan untuk deteksi dini, dan pencegahan komplikasi penyakit diabetes. Pada penelitian ini menunjukkan bahwa pemanfaatan model gabungan KNN dan LightGBM dapat menjadi sebuah alat dalam praktik klinis untuk meningkatkan deteksi dini dan pengelolaan diabetes. Penggunaan model prediktif yang akurat ini dapat membantu para profesional medis dalam mengidentifikasi pasien yang berisiko terkena diabetes lebih awal, memungkinkan intervensi yang tepat waktu dan potensial untuk mengurangi komplikasi serius. Selain itu, aplikasi berbasis teknologi ini dapat memberdayakan pasien untuk memonitor kondisi mereka secara proaktif dengan aplikasi seluler yang memberikan peringatan dini terhadap fluktuasi kadar gula darah, mendukung upaya manajemen kesehatan mandiri yang lebih efektif. Kemajuan ini tidak hanya meningkatkan kualitas hidup pasien tetapi juga mengoptimalkan alokasi sumber daya kesehatan, menandai langkah penting dalam evolusi pengobatan diabetes berbasis data.

DAFTAR PUSTAKA

[1] "Diabetes," World Health Organization. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes. [Accessed: Nov. 22, 2023].
[2] A. Perdana, A. Hermawan, and D. Avianto, "Analyze Important Features of PIMA Indian Database For Diabetes Prediction Using KNN," Jurnal SISFOKOM (Sistem Informasi dan Komputer), vol. 12, no. 1, pp. 70-75, 2022. [Online]. Available: https://doi.org/10.32736/sisfokom.v12i1.1598.
[3] M. Bergeron et al., "Episodic-Memory Performance in Machine Learning Modeling for Predicting Cognitive Health Status Classification," Journal of Alzheimer's Disease, vol. 70, no. 1, pp. 277-286, Jul. 2019. https://doi.org/10.3233/JAD-190165.
[4] G. Kaur et al., "Diagnostic accuracy of tests for type 2 diabetes and prediabetes: A systematic review and meta-analysis," Journal of PLoS ONE, vol. 15, no. 11, Art. no. e0242415, 2020. [Online]. Available: https://doi.org/10.1371/journal.pone.0242415.
[5] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," Journal Diabetes Metab Disord, vol. 19, no. 1, pp. 391-403. [Online]. Available: https://doi.org/10.1007/s40200-020-00520-5.
[6] A. Elsaddawy et al., "Predictive Analysis of Diabetes-Risk with Class Imbalance," Journal Comput Intell Neurosci, Oct. 2022. [Online]. Available: https://doi.org/10.1155/2022/3078025
[7] M. Hassan, S. Mollick, and F. Yasmin, "An unsupervised cluster-based feature grouping model for early diabetes detection," Healthcare Analytics, vol. 2, pp. 1-12, 2020. [Online]. Available: https://doi.org/10.1016/j.health.2022.100112.
[8] S. Uddin et al., "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," Journal Scientific reports, vol. 12, Art. no. 6256, 2022. [Online]. Available: https://doi.org/10.1038/s41598-022-10358-x.
[9] H. Wang, P. Xu, and J. Zhao, "Improved KNN Algorithm Based on Preprocessing of Center in Smart Cities," Journal of Hindawi Wiley, pp. 1-10, 2022. [Online]. Available: https://doi.org/10.1155/2021/552438.
[10] D. D. Rufo et al., "Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM)," Jurnal Diagnostics, vol. 11, no. 9, Art. no. 1714, Sep. 2019. [Online]. Available: https://doi.org/10.3390/diagnostics11091714.
[11] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," New England Journal of Medicine, vol. 80, no. 14, pp. 1347-1358, 2019. [Online]. Available: https://doi.org/10.1056/NEJMra1814259.
[12] S. Bhargava, M. K. Ali, and T. Rustagi, "Machine learning techniques for diabetes," in Machine Learning Techniques for Bioinformatics, pp. 83-1305, 2019. [Online]. Available: https://doi.org/10.1016/j.ejmech.2020.112457.
[13] I. Contreras and J. Vehi, "Artificial Intelligence for Diabetes Management and Decision Support: Literature Review," Journal of Medical Internet Research, vol. 20, no. 5, May 2018. [Online]. Available: https://doi.org/10.2196/10775.

[14] R. Saxena, D. D. Khumar, and M. Gupta, "Role of K-Nearest Neighbour in detection of Diabetes Mellitus," Turkish Journal of Computer and Mathematics Education, vol. 12, no. 10, pp. 373
[15] M. Nishom, "Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma KMeans Clustering Berbasis Chi-Square," Jurnal Pengembangan IT (JPIT), vol. 4, no. 1, pp. 20-24, 2019. [Online]. Available: https://10.30591/jpit.v4i1.1253
[16] J. Peng, W. Wu, B. Lockhart, and B. Song, "DataPrep.EDA: Task-Centric Exploratory Data Analysis for Statistical Modeling in Python," in Proc. 2021 Int. Conf. on Management of Data, 2021. [Online]. Available: https://doi.org/10.1145/3448016.3457330.
[17] V. Chang, J. Bailey, A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," Neural Computing and Applications, vol. 35, pp. 16147-16173, 2020. [Online]. Available: https://doi.org/10.1007/s00521-022-07049-z.
[18] Susilowati, A. A., & Waskita, K. N. (2019). Pengaruh Pola Makan Terhadap Potensi Resiko Penyakit Diabetes Melitus. Jurnal Mandala Pharmacon Indonesia, 5(01), 43–47. https://doi.org/10.35311/jmpi.v5i01.43
[19] Ridwan, A. M., & Setyawan, G. D. (2023). Perbandingan Berbagai Model Machine Learning Untuk Mendeteksi Diabetes. *Teknokom*, *6*(2), 127–132. https://doi.org/10.31943/teknokom.v6i2.152
[20] V. Khoirunnisa. (2023). IMPLEMENTASI KLASIFIKASI KEHAMILAN BERESIKO DENGAN METODE NAIVE BAYES PADA PUSKESMAS KELURAHAN MALAKA JAYA. *Jurnal Indonesia: Manajemen Informatika Dan Komunikasi*. 4(2),540–551.
[21] R. Rousiyati, A. N. Rais, N. Hasan, R. F. Amir, W. Warijono, "Komparasi Adaboost dan Bagging Dengan Naïve Bayes Pada Dataset Bank Direct Marketing," Bianglala Informatika, 2021. https://doi.org/10.31294/bi.v9i1.9890.g4731
[22] A. J. Taufiq, T. Pinandita, Susiyadi, & J. Juanita. (2023) Deteksi Suhu Tubuh dan Masker untuk Kendali Portal Otomatis Menggunakan Machine Learning. Techno. *Jurnal Fakultas Teknik, Universitas Muhammadiyah Purwokerto,* 109-116. https://10.30595/techno.v24i2.19267.