Jurnal Teknik Informatika (JUTIF)
Vol. 5, No. 4, August 2024, pp. 1165-1174

DOI: https://doi.org/10.52436/1.jutif.2024.5.4.2593
p-ISSN: 2723-3863
e-ISSN: 2723-3871

IMPLEMENTATION OF DIABETES PREDICTION MODEL USING RANDOM FOREST ALGORITHM, K-NEAREST NEIGHBOR, AND LOGISTIC REGRESSION

Rio Pratama[1], Amril Mutoi Siregar[2], Santi Arum Puspita Lestari[3], Sutan Faisal[4]

[1]Informatics Engineering, Faculty of computer science, Universitas Buana Perjuangan Karawang, Indonesia
[2,3,4]Departement Of Informatics, Faculty Of Computer Science, Buana Perjuangan Karawang University, Indonesia
Email: [1]if20.riopratama@mhs.ubpkarawang.ac.id, [2*]amrilmutoi@ubpkarawang.ac.id, [3]santi.arum@ubpkarawang.ac.id, [4]sutanfaisal@ubpkarawang.ac.id

(Article received: July 31, 2024; Revision: August 26, 2024; published: September 3, 2024)

Abstract

Diabetes is a serious metabolic disease that can cause various health complications. With more than 537 million people worldwide living with diabetes in 2021, early detection is crucial to preventing further complications. This research aims to predict the risk of diabetes using machine learning algorithms, namely Random Forest (RF), K-Nearest Neighbor (KNN), and Logistic Regression (LR), with the diabetes dataset from UCI. Previous research has explored a variety of algorithms and techniques, with results varying in accuracy. This research uses a dataset from Kaggle which consists of 768 data with 8 parameters, which are processed through pre-processing and data normalization techniques. The model was evaluated using metrics such as accuracy, confusion matrix, and ROC-AUC. The results showed that Logistic Regression had the best performance with 77% accuracy and AUC 0.83, compared to KNN (75% accuracy, AUC 0.81) and Random Forest (74% accuracy, AUC 0.81). These findings emphasize the importance of appropriate algorithm selection and good data pre-processing in diabetes risk prediction. This study concludes that Logistic Regression is the most effective method for predicting diabetes risk in the dataset used.

Keywords: *Diabetes, KNN, Random Forest, Logistic Regression.*

IMPLEMENTASI MODEL PREDIKSI DIABETES MENGGUNAKAN ALGORITMA RANDOM FOREST, K-NEAREST NEIGHBOR, DAN REGRESI LOGISTIK

Abstrak

Diabetes adalah penyakit metabolik serius yang dapat menyebabkan berbagai komplikasi kesehatan. Dengan lebih dari 537 juta orang di seluruh dunia yang hidup dengan diabetes pada tahun 2021, deteksi dini menjadi krusial untuk mencegah komplikasi lebih lanjut. Penelitian ini bertujuan memprediksi risiko diabetes menggunakan algoritma machine learning, yaitu *Random Forest* (RF*), K-Nearest Neighbor* (KNN), dan *Logistic Regression* (LR), dengan dataset diabetes dari UCI. Penelitian sebelumnya telah mengeksplorasi berbagai algoritma dan teknik, dengan hasil yang bervariasi dalam akurasi. Penelitian ini menggunakan dataset dari Kaggle yang terdiri dari 768 data dengan 8 parameter, yang diproses melalui teknik pra-pemrosesan dan normalisasi data. Model dievaluasi menggunakan metrik seperti akurasi, confusion matrix, dan ROC-AUC.Hasil menunjukkan bahwa *Logistic Regression* memiliki kinerja terbaik dengan akurasi 77% dan AUC 0,83, dibandingkan dengan KNN (75% akurasi, AUC 0,81) dan *Random Forest* (74% akurasi, AUC 0,81). Temuan ini menegaskan pentingnya pemilihan algoritma yang tepat dan pra-pemrosesan data yang baik dalam prediksi risiko diabetes. Penelitian ini menyimpulkan bahwa *Logistic Regression* adalah metode yang paling efektif untuk memprediksi risiko diabetes pada dataset yang digunakan.

Kata kunci: Diabetes, KNN, *Random Forest, Logistic Regression*.

1. PENDAHULUAN

Diabetes adalah penyakit metabolisme yang disebabkan oleh tingginya glukosa atau gula darah. Gula darah penting bagi kesehatan karena merupakan sumber energi penting bagi sel dan jaringan. Jika tidak ditangani dengan baik, diabetes dapat menyebabkan banyak komplikasi, termasuk penyakit jantung koroner, stroke, obesitas, serta masalah mata, ginjal, dan saraf [1].

Diabetes masih masuk dalam daftar penyebab kematian utama di dunia. Pada tahun 2021, Lebih

dari setengah miliar orang di seluruh dunia menderita diabetes, yaitu sekitar 537 juta orang. Angka ini diperkirakan akan meningkat menjadi 643 juta pada tahun 2030 dan 783 juta pada tahun 2045 [2].

Para dokter percaya bahwa kondisi ini berkaitan dengan gaya pola hidup dan gaya pola makan, yang berkontribusi terhadap penyakit tersebut. Penderita diabetes mempunyai risiko lebih tinggi terkena beberapa masalah tentang kesehatan sekunder, seperti penyakit jantung dan kerusakan saraf dll. Mengontrol perkembangan diabetes memerlukan deteksi dini untuk mencegah atau mengurangi komplikasi pada kesehatan tubuh. Namun dengan perkembangan teknologi terkini, prediksi penyakit diabetes menjadi mungkin dilakukan [3].

Penelitian Rika Ajeng Finatih sebelumnya Prediksi Terkena Diabetes menggunakan Metode *K Nearest Neighbor* (KNN). Kemudian dievaluasi menggunakan matriks dan menggunakan metode KNN dengan K = 3 sehingga menghasilkan 55 *true positif* (pasien diabetes dengan diagnosis benar) dan 158 *true negative* (pasien diabetes dengan diagnosis benar) dan 5 *false positif* (pasien tanpa diabetes tetapi menderita diabetes) dan 7 penyakit serius (pasien yang menderita diabetes tetapi tidak menderita diabetes). Maka akurasi hasilya 66,5% [4].

Adapun penelitian oleh Naisah Marito Putry yaitu Komperasi Algoritma KNN Dan *Naive Bayes* untuk Klasifikasi Diagnosa Penyakit Diabetes Melitus. Pada penelitian ini menggunakan 5 *dataset* untuk mendapatkan nilai akurasi dari *Naive Bayes* dan KNN, Mendapatkan nilai absolut dari algoritma *Naive Bayes* adalah 80% dan algoritma KNN mempunyai nilai paling akurat yaitu 75%. Nilai recall tertinggi yang dihitung menggunakan algoritma KNN diketahui sebesar 0,92. Dan nilai paling akurat yang dihasilkan oleh algoritma *Naïve Bayes* adalah 0,86 [5].

Penelitian selanjutnya yaitu Gde Agung Brahmana Suryanegara dengan judul Peningkatkan Hasil Klasifikasi Algoritma *Random Forest* untuk deteksi Pasien Diabetes Menggunakan Metode Normalisasi. Menggunakan dua metode pengolahan data yaitu normalisasi *Min-max* dan normalisasi *Z-score* dan satu tanpa metode normalisasi data dengan *Random Forest* (RF). Menunjukkan bahwa model 1 (*min-max normalization*-RF) 95.45%, untuk model 2 (*Z-score normalization*-RF) dan 95%, dan model 3 (Tanpa normalisasi data-RF) 93%. Kesimpulanya bahwa model 1 (*regularisasi min-max*-RF) lebih baik dan dapat meningkatkan kinerja klasifikasi *Random Forest* sebesar 95,45% [6].

Adapun Penelitian oleh Dewi Nasien berencana fokus pada menganalisis kinerja tiga metode klasifikasi utama, yaitu *K-Nearest Neighbor* (KNN), *Naive Bayes*, dan *Logistic Regression*, dalam konteks pengklasifikasian data diabetes. dengan 2 pembagian data 70-30 dan 80-20, disajikan dalam bentuk confusion matrix. Dengan ekstraksi fitur dengan Analisis Komponen Utama yaitu (*Principal Component Analysis*) PCA dengan threshold 80%, menghasilkan 5 fitur utama terdapat beberapa temuan signifikan, KNN pada split data 70-30 nilai akurasi 70%, split 80-20, nilai akurasinya 71%. Metode *Naive Bayes* pada nilai akurasi 75% pada split 70-30, dan meningkat menjadi 79% pada split data 80-20. Dan *Logistic Regression* menunjukkan nilai akurasi 71% pada split data 70-30 dan 73% pada split data 80-20. Menunjukkan bahwa hasil *Naive Bayes* adalah akurasi terbaik, mencapai 79% pada pembagian data 80:20 [7].

Penilitian menurut Farkhina Dwi Utari yang berjudul Implementasi Algoritme *K-Nearest Neighboar* (KNN) untuk Prediksi Hasil Produksi. Yaitu Cara mengambil keputusan yang tepat dalam persiapan dan pengelolaan rencana produksi dan inspeksi Hal ini berdasarkan hasil statistik klasifikasi prediksi data keluaran PT. SKI menggunakan algoritma KNN (*K-Nearest Neighbor*) dengan 130 data latih dan 1 data uji, ditentukan K = 5 dan diperoleh nilai akurasi sebesar 100% [8].

Adapun penelitian yang di bawakan oleh Alma Hidayanti tentang Model Analisis Kasus Covid-19 Di Indonesia Menggunakan Algoritma Regresi Linier Dan Random Forest. Tujuannya untuk menganalisis data kasus COVID-19 yang tidak memiliki perkembangan yang jelas. Menganalisisnya akan menghasilkan informasi baru dengan mengetahui nilai akurasinya. Nilai paling akurat yang dihitung menggunakan metode *regresi linier* adalah 99,73% menggunakan analisis statistik manual menggunakan *Microsoft Excel*, 92,4% menggunakan analisis *Google Collaboration*, dan 85,7% menggunakan analisis *RapidMiner*. Sedangkan nilai paling akurat yang dihitung menggunakan metode *random forest* untuk data COVID-19 Indonesia adalah 98,4% dan 93,9% menggunakan analisis komputer *RapidMiner* [9].

Dalam penilitian Benedictus Mario Wendhi T berjudul Penerapan Algoritma *K-Nearest Neighbor* Dengan Pengolahan Citra Digital Untuk Mengidentifikasi Jenis Kayu dengan Perhitungan dilakukan untuk mengidentifikasi jenis kayu yang digunakan pada furnitur dan perabotan. Metode GLCM digunakan untuk mengekstraksi citra guna menghitung jenis pohon, dan algoritma K-Nearest Neighbor (KNN) digunakan untuk mengidentifikasi jenis pohon. Dengan data berupa 100 gambar tiga jenis pohon. Hasil akurasi terbaik Untuk nilai parameter K=5, hasil akurasi terbaik untuk mengidentifikasi jenis pohon ini adalah 91,6%, sedangkan hasil akurasi terendah adalah 61,1%. Dengan demikian, rata-rata akurasi adalah 75,54% [10].

Menurut Siti Nurjanah dalam penelitiannya yang berjudul "Penerapan Algoritma *K-Nearest Neighbor* (KNN) untuk Klasifikasi Pencemaran

Udara di Kota Jakarta", data ISPU harian Jakarta diolah menggunakan metode klasifikasi data mining dengan algoritma K-*Nearest Neighbor* (KNN). Hasil klasifikasi tingkat pencemaran udara di Jakarta dengan algoritma KNN, menggunakan 304 data latih dan 1 data uji, menunjukkan bahwa nilai K = 7 memberikan akurasi sebesar 95,78% [11].

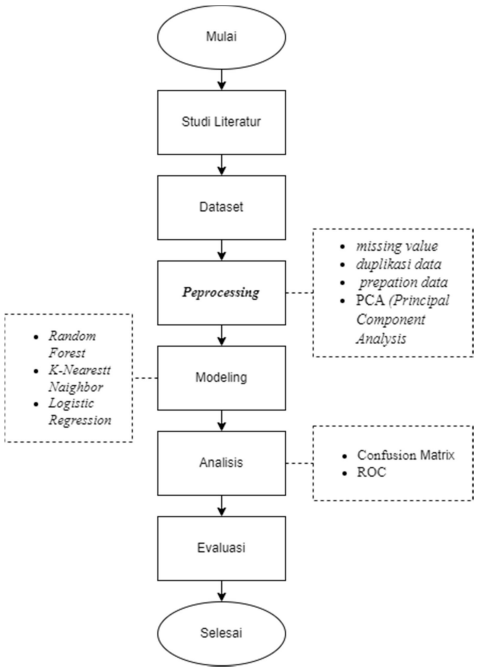
Penelitian yang dilakukan oleh Yana Cahyana Pada artikel berjudul "Penerapan Algoritma Random Forest untuk Klasifikasi Kabupaten/Kota di Provinsi Jawa Barat Berdasarkan Pendapatan dari Sektor Pertanian", algoritma data mining diterapkan. RapidMiner memanfaatkan kecerdasan buatan, metode statistik, dan basis data untuk mengidentifikasi pola dari kumpulan data besar. Hasil klasifikasi kabupaten dan kota berdasarkan pendapatan sektor pertanian mengindikasikan bahwa algoritma random forest berhasil mencapai akurasi sebesar 87,50%. [12].

Tujuan dari penelitian ini adalah untuk memprediksi risiko seseorang terkena dan tidak terkena diabetes menggunakan algoritma *Machine learning* berdasarkan metode *Random Forest, K-Nearest Neighbor (KNN),* dan *Logistic Regression* serta hasil akurasi dari beberapa model.

Penelitian ini melakukan prediksi hasil data diabetes dengan algoritma *machine* learning, dan *dataset* yang digunakan adalah *dataset* diabetes *machine learning* UCI, dan teknik yang digunakan yaitu membandingkan nilai akurasi dengan penerapan algoritma 3 yaitu *Random Forest* (RF), *K-Nearest Neighbor* (KNN) dan *Logistic Regression* (LR) kemudian memberikan nilai akurasi total.

2. METODE PENELITIAN

Studi implementasi dilakukan untuk menghasilkan algoritma yang baik dan hasil yang akurat untuk memprediksi risiko awal penyakit diabetes. [13].Tahapan studi pada penelitian ini dapat dilihat pada gambar 1. Flowchart tahapan studi.



Gambar 1 Flowchat Tahapan Studi

2.1 Studi Literatur

Sumber informasi penelitian ini adalah jurnal yang mempelajari tentang penyakit diabetes. Banyak peneliti menggunakan data diabetes untuk mempelajari seperti masalah, memprediksi dan membandingkan algoritma untuk menemukan hasil terbaik untuk data diabetes.

2.2 Dataset

Dataset penelitian ini diperoleh dari kaggle, sebuah situs web yang tersedia untuk umum [14]. (https://www.kaggle.com/code/mragpavank/pima-indians-diabetes-database). Jumlah sampel yang digunakan berjumlah 768 data, termasuk 8 parameter: glukosa, tekanan darah, ketebalan kulit, insulin, BMI, fungsi Pedigree, usia, dan hasil yang lengkap dengan nilai parameternya. *Random Forest* (RF), *K-Nearest Neighbor* (KNN), dan *Logistic Regression* (LF) adalah tiga teknik yang digunakan dalam penelitian ini.

2.3 Peprocessing

Pada tahapan *Preprocessing ini dataset* dilakukan pra-pemrosesan *(pre-processing)*. Langkah ini menggunakan metode *pre-processing* seperti : *missing value, duplikasi data, prepation* data dan PCA *(Principal Component Analysis)*. PCA (*Principal Component Analysis)* adalah salah satu teknik dalam analisis data yang termasuk dalam tahap *preprocessing* data. Proses ini menghasilkan data yang siap digunakan pada langkah pemodelan prediktif [15]. Parameter yang diahasilkan menjadi 7 yaitu : *Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction dan Age.*

2.4 Cleaning Data

Cleaning Data data adalah penghapusan data seperti data null, data yang salah dimasukkan, data tidak relevan, data duplikat, dan lain-lain, karena mengurangi data tersebut dalam kualitas atau keakuratan hasil data mining. Pembersihan data mempengaruhi kinerja sistem penambangan data. karena mengurangi jumlah dan kompleksitas data yang akan diproses [16].

2.5 Modeling

1) Random Forest

Algoritma Random Forest (RF) memiliki banyak keunggulan, termasuk tingkat kesalahan yang rendah, kinerja klasifikasi yang baik, kemampuan menangani data pelatihan yang besar, dan efisiensi dalam memperkirakan data yang hilang. Algoritma ini mencapai node terakhir dalam struktur pohon berdasarkan klasifikasi dan pohon yang disederhanakan dengan menggunakan metode analisis biner. *Random Forest* menghasilkan banyak pohon independen dari sampel pelatihan dan subset yang dipilih secara acak dengan melakukan bootstrapping pada variabel input di setiap node [17].

2) K-Nearestt Naighbor

Metode yang disebut KNN (K-Nearest Neighbor) mengklasifikasikan objek dengan memanfaatkan data dari tetangga terdekat. Jarak *Euclidean* digunakan untuk menentukan seberapa dekat atau jauhnya suatu tetangga. Metode ini terdiri dari dua fase: pelatihan (training) dan klasifikasi (testing). Pada tahap awal, algoritma menyimpan fitur dan vektor klasifikasi dari data pelatihan. Tahap kedua melibatkan perhitungan fitur serupa untuk data uji yang klasifikasinya belum diketahui. Jarak antara vektor baru dan semua vektor data pelatihan dihitung, kemudian dipilih *k* tetangga terdekat [15].

3) Logistic Regression

Logistic Regression adalah kerangka kerja yang berguna dan fleksibel yang dapat memenuhi kebutuhan banyak analis. Namun perlu diingat bahwa model ini mungkin tidak cocok untuk semua situasi. Dalam bidang teknik dan sains, banyak permasalahan yang berkaitan dengan variabel respon dan prediktor yang terkait dengan fungsi yang tidak diketahui. Dalam penelitian ini, pertama-tama kami menggunakan PCA untuk mereduksi dimensi dan kemudian melakukan regresi linier untuk memprediksi data eksperimen. Hitung dan cetak hasil akurat, matriks kebingungan, dan laporan klasifikasi. Rumus regresi logistik fungsi sigmoid. Rumus untuk Regresi Logistik Fungsi Sigmoid.

$$h\sigma(x) = \frac{1}{1+e-n} \quad (1)$$

$h\sigma(x)$ adalah fungsi logistik yang mengambil nilai antara 0 dan 1 [18]

2.6 Analisis

1) Confusion Matrix

False positive (FP), true negative (TN), true positive (TP), dan false negative (FN)—berdasarkan hasil klasifikasi—digunakan Confusion Matrix untuk mengevaluasi kinerja metode klasifikasi. Nilai negatif yang ditemukan dengan benar digambarkan oleh TN, sedangkan nilai negatif yang salah ditemukan digambarkan oleh FP sebagai positif. [19]. Model ini menghitung presisi, akurasi, recall, dan skor F1. Akurasi mengukur persentase prediksi positif yang benar dari total prediksi positif. Ini juga memberikan gambaran keseluruhan tentang seberapa baik kinerja model. [20].

Evaluasi Model

$$Accuracy = \frac{Number\ of\ correct\ Predictions\ Total\ Number\ of\ Data\ Accuracy}{Total\ Number\ Of\ Data\ Number\ of\ Correct\ Predictions} \quad (2)$$

Confusion Matrix

$$\begin{pmatrix} True\ Negative\ (TN) & False\ Positive\ (FB) \\ False\ Negative\ (FN) & Truse\ Positif\ (TP) \end{pmatrix}$$

2) Receiver Operator Characteristics

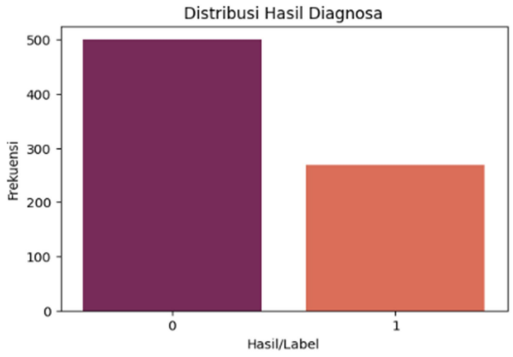
Menurut Vercellis, kurva ROC (*Receiver Operating Characteristics*) adalah metode yang menggambarkan, mengatur, dan melacak klasifikasi keterampilan berdasarkan operasi. ROC mewakili matriks ketidakpastian. ROC adalah grafik dua dimensi yang menampilkan tingkat positif palsu di sumbu horizontal dan tingkat positif benar di sumbu vertikal. [21].

2.7 Evaluasi

Hasil data yang telah di propeprocessing mencakup delapan parameter, antara lain *Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age* dan *Outcome*. setelah diolah ada 7 parameter yang akan diproses yaitu *Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction* dan *Age*. Parameter yang diekstraksi adalah *Outcome*.

3. HASIL DAN PEMBAHASAN

Pada tahap ini dilakukan untuk mengetahui hubungan setiap pasangan variabel, atribut kualitas yang tidak memiliki korelasi kuat satu sama lain dihilangkan. Untuk meningkatkan pemahaman tentang dataset dan fitur-fiturnya, dipelajari distribusi atribut dan keseimbangan data menggunakan berbagai pendekatan visualisasi data. Berikutnya melibatkan pengkodean label, yang mengubah data kategorikal menjadi numerik untuk pemahaman yang lebih baik dan kesederhanaan pemrosesan.



Gambar 2. Diagram Penderita Diabetes dan Tanpa Diabetes

Meningkatkan pemahaman tentang kumpulan data dan fitur-fiturnya EDA dilakukan untuk mengidentifikasi potensi nilai yang hilang dan duplikasi. Visualisasi kemudian dilakukan untuk mendapatkan wawasan berharga sebelum melanjutkan pemrosesan data lebih lanjut. Atribut seperti riwayat merokok yang tidak memberikan kontribusi terhadap jawaban yang akurat akan dibuang. Data visualisasi yang disajikan pada Gambar 1.

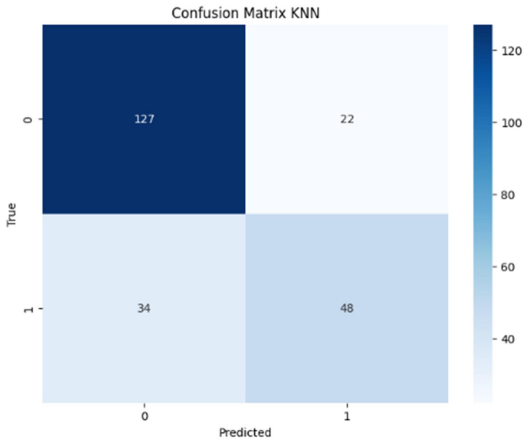
Tabel 1. Pengukuran Kinerja Beberapa Metode *Machine Learning*

Metode	Accuracy
KNN	75%
Random Forest	74%
Logistic Regression	77%

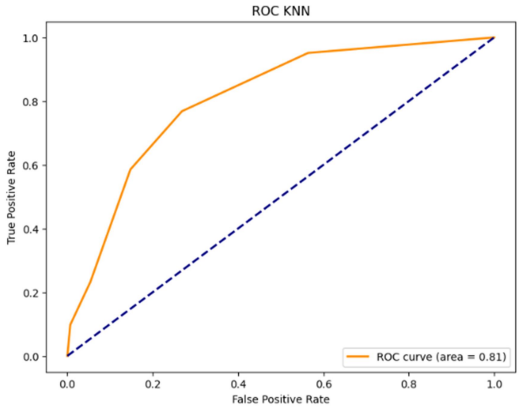
Analisis terhadap hasil eksperimen menunjukkan bahwa kinerja algoritma K-Nearest Neighbors (KNN) mengalami penurunan ketika data diolah terlebih dahulu menggunakan teknik Principal Component Analysis (PCA). Hal ini dapat dilihat secara visual pada grafik Confusion Matrix dan kurva ROC yang disajikan pada Gambar 3 dan 4. Penurunan performa ini mengindikasikan bahwa proses reduksi dimensi yang dilakukan oleh PCA telah menghilangkan informasi penting yang sangat dibutuhkan oleh algoritma KNN untuk melakukan klasifikasi dengan akurat. Sederhananya, ketika kita mengurangi dimensi data menggunakan PCA, kita seperti membuang beberapa potongan puzzle penting yang membuat gambar keseluruhan menjadi tidak lengkap. Akibatnya, algoritma KNN kesulitan dalam mengenali pola yang ada dalam data dan membuat prediksi yang tepat.

3.1 KNN *(K-Nearest Neighbor)*

Hasil eksperimen menunjukkan bahwa kinerja algoritma K-Nearest Neighbors (KNN) mengalami penurunan ketika data terlebih dahulu diolah menggunakan teknik Principal Component Analysis (PCA). Hal ini terlihat jelas pada grafik Confusion Matrix dan kurva ROC yang disajikan pada Gambar 3 dan 4. Penurunan performa ini mengindikasikan bahwa reduksi dimensi yang dilakukan oleh PCA telah menghilangkan informasi penting yang dibutuhkan oleh algoritma KNN untuk melakukan klasifikasi dengan akurat.



Gambar 3 Confusion Matrix Algortima KNN



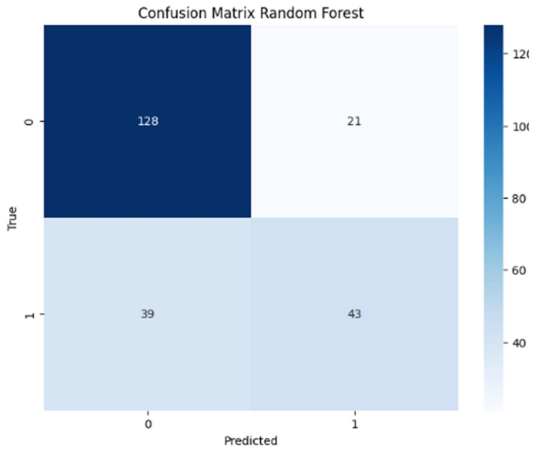
Gambar 4. ROC KNN

Gambar 4 menampilkan kurva ROC (Receiver Operating Characteristic) dari algoritma K-Nearest Neighbors (KNN). Kurva ROC ini menggambarkan kinerja model KNN dalam membedakan antara dua kelas. Nilai AUC (*Area Under the Curve*) pada kurva ROC ini adalah 0,81. Nilai AUC sebesar 0,81 mengindikasikan bahwa model KNN yang digunakan dalam eksperimen ini memiliki kemampuan yang cukup baik dalam melakukan klasifikasi. Semakin dekat nilai AUC ke 1, semakin baik kinerja model dalam membedakan kedua kelas. Dengan kata lain, model KNN dapat cukup diandalkan untuk membuat prediksi yang akurat berdasarkan data yang diberikan.

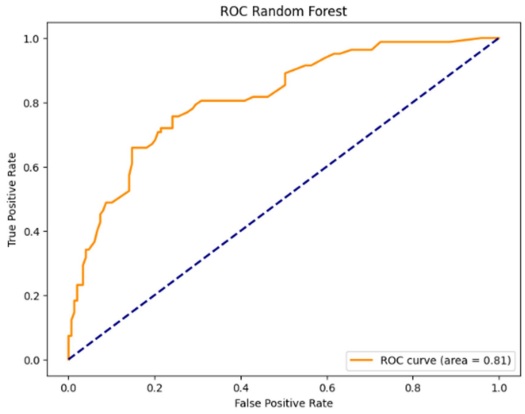
3.2 *Randorm Forest*

Algoritma *Random Forest* mengalami penurusan performa ketika diuji dengan menggunakan PCA seperti yang di tunjukan oleh gambar 5. dibawah ini

yang di tunjuikan oleh *Confusion Matrix* dan gambar 6. ROC *Random Forest* dibawah ini.



Gambar 5. *Confusion Matrix* Algoritma Random Forest

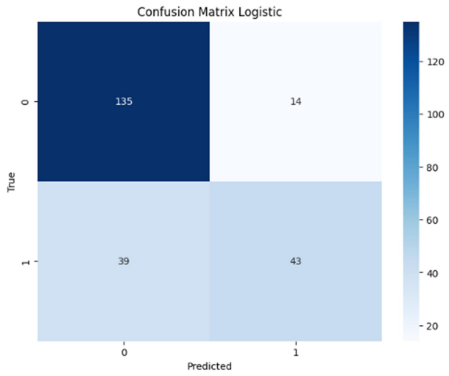


Gambar 6. ROC Random Forest

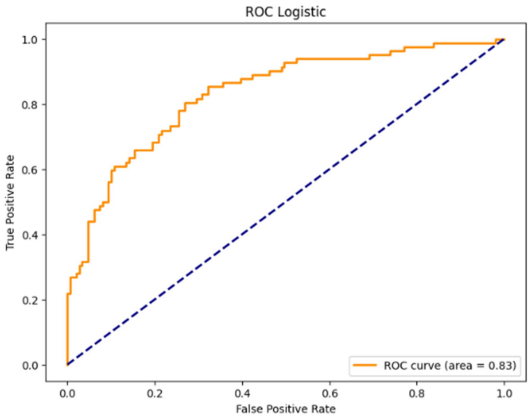
Gambar nomor 6 menunjukkan grafik ROC dari model klasifikasi yang menggunakan algoritma Random Forest. Nilai AUC dari grafik ini adalah 0,81. Ini berarti model Random Forest yang digunakan dalam eksperimen ini memiliki kemampuan yang cukup baik dalam membedakan antara dua kelas yang sedang diprediksi.

3.3 *Logistic Regression*

Algoritma *Logistic Regression* mengalami penurusan performa ketika diuji dengan menggunakan PCA seperti yang di tunjukan oleh gambar 6. dibawah ini yang di tunjuikan oleh *Confusion Matrix* dan gambar 7. ROC *Random Forest* dibawah ini.



Gambar 6. *Confusion Matrix* Algoritma Logistic Regression



Gambar 7. ROC Logistic Regression

Gambar nomor 7 menunjukkan grafik ROC dari model klasifikasi yang menggunakan algoritma Logistic Regression. Nilai AUC dari grafik ini adalah 0,83. Ini berarti model Logistic Regression yang digunakan dalam eksperimen ini memiliki kemampuan yang cukup baik dalam membedakan antara dua kelas yang sedang diprediksi.

Gambar 4, 6, dan 7 menampilkan hasil perbandingan kinerja dari tiga model machine learning yang berbeda, yaitu KNN, Random Forest, dan Logistic Regression. Ketiga model ini digunakan untuk melakukan klasifikasi data. Hasil evaluasi menunjukkan bahwa model KNN dan Random Forest memiliki performa yang serupa, ditunjukkan oleh nilai AUC yang sama yaitu 0,81. Meskipun kedua model ini memiliki akurasi yang cukup baik, nilai AUC yang diperoleh mengindikasikan bahwa model-model ini mungkin kurang optimal dalam membedakan antara kelas positif dan negatif secara keseluruhan. Sebaliknya, model Logistic Regression menunjukkan hasil yang berbeda, yang mungkin mengindikasikan bahwa model ini lebih cocok untuk data yang digunakan dalam eksperimen ini.

4. DISKUSI

Dalam penelitian ini, kami telah mengevaluasi beberapa algoritma machine learning untuk memprediksi risiko diabetes, menggunakan dataset yang tersedia secara publik di Kaggle. Tiga algoritma utama yang digunakan adalah Random Forest, K-Nearest Neighbor (KNN), dan Logistic Regression. Hasil evaluasi menunjukkan bahwa Logistic Regression memberikan akurasi tertinggi sebesar 77%, diikuti oleh KNN dengan 75%, dan Random Forest dengan 74%.

Hasil ini konsisten dengan penelitian sebelumnya, yang menunjukkan bahwa Logistic Regression cenderung memberikan hasil yang lebih akurat dalam konteks klasifikasi data diabetes. Sebagai contoh, penelitian oleh Dewi Nasien yang membandingkan KNN, Naive Bayes, dan Logistic Regression juga menunjukkan bahwa Logistic Regression memberikan hasil akurasi yang lebih tinggi pada pembagian data tertentu.

Namun, ketika algoritma-algoritma tersebut diuji dengan menggunakan *Principal Component Analysis* (PCA) untuk reduksi dimensi, kinerja KNN dan Random Forest mengalami penurunan. Grafik Confusion Matrix dan kurva ROC mengindikasikan bahwa reduksi dimensi dengan PCA dapat menghilangkan informasi penting yang diperlukan untuk klasifikasi akurat oleh KNN dan Random Forest.

Penelitian oleh Rika Ajeng Finatih menunjukkan bahwa algoritma KNN dengan K=3K = 3K=3 memberikan akurasi 66,5%, yang merupakan hasil yang relatif moderat[4]. Di sisi lain, Naisah Marito Putry membandingkan KNN dan Naive Bayes, menemukan bahwa Naive Bayes memberikan akurasi yang lebih tinggi yaitu 80% dibandingkan dengan KNN yang memiliki akurasi 75% [5].

Gde Agung Brahmana Suryanegara mengeksplorasi penggunaan normalisasi data dengan algoritma Random Forest. Hasilnya menunjukkan bahwa penggunaan normalisasi min-max memberikan kinerja terbaik dengan akurasi 95,45%, yang menegaskan pentingnya teknik pra-pemrosesan data dalam meningkatkan kinerja model [6].

Penelitian oleh Dewi Nasien menyoroti bahwa Naive Bayes juga menunjukkan performa terbaik dengan akurasi mencapai 79% pada pembagian data 80:20. Logistic Regression juga menunjukkan kinerja yang cukup baik dengan akurasi hingga 73% pada pembagian data yang sama [7].

Farkhina Dwi Utari menunjukkan bahwa KNN dapat mencapai akurasi 100% dalam prediksi hasil produksi [8], sementara Alma Hidayanti menyoroti bahwa Random Forest dapat mencapai akurasi 98,4% dalam analisis data COVID-19 di Indonesia [9].

Penilitian oleh Benedictus Mario Wendhi Tranose mendapatkan Hasil akurasi tertinggi dari identifikasi ini adalah 91.6% [10], Sementara Siti Nurjanah meyoroti bahwa KNN dapat mencapai nilai akurasi sebanyak 95.78% [11].

Penelitian ini bertujuan untuk memprediksi risiko diabetes menggunakan dataset dari Kaggle dengan tiga algoritma utama: *Random Forest*, KNN, dan *Logistic Regression*. Hasil menunjukkan bahwa *Logistic Regression* memberikan akurasi terbaik sebesar 77%, dibandingkan dengan KNN (75%) dan *Random Forest* (74%). Analisis lebih lanjut dengan kurva ROC menunjukkan bahwa *Logistic Regression* juga memberikan hasil AUC yang lebih baik (0,83) dibandingkan dengan KNN dan *Random Forest* (0,81).

Secara keseluruhan, penelitian ini mengindikasikan bahwa *Logistic Regression* adalah metode yang paling efektif untuk memprediksi risiko diabetes dalam dataset yang digunakan. Ini menyoroti pentingnya memilih algoritma yang tepat dan melakukan pra-pemrosesan data yang baik untuk mendapatkan hasil yang akurat dalam prediksi penyakit diabetes.

5. KESIMPULAN

Hasil dari penelitian ini menunjukkan bahwa, *Logistic Regression* adalah metode paling efektif untuk memprediksi risiko diabetes menggunakan dataset dari Kaggle. Logistic Regression memberikan akurasi tertinggi sebesar 77% dan nilai AUC terbaik sebesar 0,83 dibandingkan dengan *K-Nearest Neighbor* (KNN) dan *Random Forest*, yang masing-masing memiliki akurasi 75% dan 74% serta nilai AUC 0,81. Penelitian ini menekankan pentingnya pemilihan algoritma yang tepat dan pra-pemrosesan data yang baik untuk mencapai hasil akurat dalam prediksi penyakit diabetes. Untuk penelitian lebih lanjut ada beberapa yang berdasarkan pada temuan penelitian sebelumnya. arah penelitian yang menjanjikan untuk pengembangan model prediksi diabetes yang lebih akurat dan komprehensif. Pertama, perlu dilakukan eksplorasi terhadap algoritma machine learning yang lebih canggih seperti deep learning dan ensemble methods. Kedua, pengolahan data yang lebih mendalam, seperti feature engineering dan penanganan data tidak seimbang, dapat meningkatkan kinerja model. Ketiga, evaluasi model yang lebih komprehensif dengan menggunakan berbagai metrik dan teknik visualisasi akan memberikan pemahaman yang lebih baik tentang kekuatan dan kelemahan model. Selain itu, interpretasi model yang lebih mendalam, penelitian longitudinal, pengembangan aplikasi, dan kolaborasi multi-pusat juga merupakan arah penelitian yang menarik. Secara spesifik, penelitian dapat difokuskan pada prediksi komplikasi diabetes, personalisasi prediksi, pengaruh faktor gaya hidup, perbandingan algoritma untuk data berdimensi tinggi, dan pengembangan aplikasi mobile untuk deteksi dini diabetes. Dengan demikian, penelitian

lebih lanjut di bidang upaya pencegahan dan pengendalian diabetes dapat memperoleh manfaat besar dari hal ini.

DAFTAR PUSTAKA

[1] A. M. Argina, "Indonesian Journal of Data and Science Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," vol. 1, no. 2, pp. 29–33, 2020.

[2] Dr. Made Ratna Saraswati, "Diabetes Melitus Adalah Masalah Kita," yankes.kemkes.go.id, 2022.

[3] I. And and D. Expert, "Prediksi Penyakit Diabetes Menggunakan Algoritma Support Vector Machine (SVM) INFORMASI ARTIKEL ABSTRAK," 2022. [Online]. Available: https://e-journal.unper.ac.id/index.php/informatics

[4] M. A. S, "Prediksi Terkena Diabetes menggunakan Metode K-Nearest Neighbor (KNN) pada Dataset UCI Machine Learning Diabetes," *Indonesian Journal of Applied Mathematics*, vol. 3, no. 2, p. 15, Nov. 2023, doi: 10.35472/indojam.v3i2.1577.

[5] N. Marito Putry and B. Nurina Sari, "Komparasi Algoritma Knn Dan Naïve Bayes Untuk Klasifikasi Diagnosis Penyakit Diabetes Melitus," *Jurnal Sains dan Manajemen*, vol. 10, no. 1, 2022.

[6] Gde Agung Brahmana Suryanegara, Adiwijaya, and Mahendra Dwifebri Purbolaksono, "Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 1, pp. 114–122, Feb. 2021, doi: 10.29207/resti.v5i1.2880.

[7] D. Nasien *et al.*, "Perbandingan Implementasi Machine Learning Menggunakan Metode KNN, Naive Bayes, Dan Logistik Regression Untuk Mengklasifikasi Penyakit Diabetes," 2024.

[8] F. D. Utari, A. M. Siregar, and D. Wahiddin, "Implementasi Algoritme K-Nearest Neighboar (KNN) untuk Prediksi Hasil Produksi," vol. 1, no. 1, 2020.

[9] A. Hidayanti, A. M. Siregar, S. A. P. Lestari, and Y. C. Cahyana, "Model Analisis Kasus Covid-19 Di Indonesia Menggunakan Algoritma Regresi Linier Dan Random Forest," *PETIR*, vol. 15, no. 1, pp. 91–101, Dec. 2021, doi: 10.33322/petir.v15i1.1487.

[10] Benedictus Mario Wendhi Tranose, "Penerapan Algoritma K-Nearest Neighbor Dengan Pengolahan Citra Digital Untuk Mengidentifikasi Jenis Kayu ," *Scientific Student Journal for Information, Technology and Science*, vol. 4, no. 2715–2766, Jul. 2023.

[11] Siti Nurjanah, "Penerapan Algoritma K-Nearest Neighbor (KNN) untuk Klasifikasi Pencemaran Udara Di Kota Jakarta," *Scientific Student Journal for Information, Technology and Science*, vol. 1, no. 2715–2766, 2020.

[12] Yana Cahyana and Amril Mutoi Siregar, "Penerapan Algoritma Random Forest Untuk Klasifikasi KAB kota Provinsi Jawa Barat Berdasarkan Pertanian," *Konferensi Nasional Penelitian dan Pengabdian Universitas Buana Perjuangan Karawang*, vol. 1, no. 2798–2580, Jul. 2021.

[13] G. Abdurrahman, "Jurnal Sistem dan Teknologi Informasi Klasifikasi Penyakit Diabetes Melitus Menggunakan Adaboost Classifier," vol. 7, no. 1, 2022, [Online]. Available: http://jurnal.unmuhjember.ac.id/index.php/JUSTINDO

[14] PAVAN KUMAR D, "Pima Indians Diabetes Database," https://www.kaggle.com/code/mragpavank/pima-indians-diabetes-database.

[15] A. P. Silalahi, G. Simanullang, and M. I. Hutapea, "METHOMIKA: Jurnal Manajemen Informatika & Komputerisasi Akuntansi Supervised Learning Metode K-nearest Neighbor Untuk Prediksi Diabetes Pada Wanita," vol. 7, no. 1, 2023, doi: 10.46880/jmika.Vol7No1.pp144-149.

[16] H. Rifa, R. Hamonangan, and D. Ade Kurnia, "KOPERTIP: Jurnal Ilmiah Manajemen Informatika dan Komputer Implementasi Algoritma Decision Tree Dalam Klasifikasi Kompetensi Siswa", [Online]. Available: http://jurnal.kopertipindonesia.or.id/

[17] F. Yulian Pamuji, V. Puspaning Ramadhan, and R. Artikel, "Jurnal Teknologi dan Manajemen Informatika Komparasi Algoritma Random Forest Dan Decision Tree Untuk Memprediksi Keberhasilan Immunotheraphy Info Artikel ABSTRAK," vol. 7, pp. 46–50, 2021, [Online]. Available: http://http://jurnal.unmer.ac.id/index.php/jtmi

[18] I. Maulana, A. Mutoi Siregar, and A. Fauzi, "Optimization of Machine Learning Model Accuracy for Brain Tumor Classification with Principal Component Analysis," *Jurnal Teknik Informatika (JUTIF)*, vol. x, No. y, pp. x-y, 2023, doi: 10.52436/jutif.

[19] H. Yun, "Prediction model of algal blooms using logistic regression and confusion matrix," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 3,

pp. 2407–2413, Jun. 2021, doi: 10.11591/ijece.v11i3.pp2407-2413.

[20] 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.

[21] D. Y. Utami, E. Nurlelah, and F. N. Hasan, "Comparison of Neural Network Algorithms, Naive Bayes and Logistic Regression to predict diabetes," *JOURNAL OF INFORMATICS AND TELECOMMUNICATION ENGINEERING*, vol. 5, no. 1, pp. 53–64, Jul. 2021, doi: 10.31289/jite.v5i1.5201.