Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 283−290
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5975

Optimasi Algoritma K-Nearest Neighbors Menggunakan Teknik Bayesian Optimization Untuk Klasifikasi Diabetes

Nur Kholis Sowabi*, Nur Aeni Widiastuti, Nadia Annisa Maori

Fakultas Sains dan Teknoligi, Teknik Informatika, Universitas Islam Nahdlatul Ulama Jepara, Jepara
Jl. Taman Siswa, Pekeng, Kauman, Tahunan, Kec. Tahunan, Kabupaten Jepara, Jawa Tengah, Indonesia
Email: [1,*]putratoang@gmail.com, [2]nuraeniwidiastuti@unisnu.ac.id, [3]nadia@unisnu.ac.id
Email Penulis Korespondensi: putratoang@gmail.com
Submitted: **26/09/2024;** Accepted: **09/10/2024;** Published: **19/10/2024**

Abstrak−Diabetes adalah salah satu penyakit kronis yang mempengaruhi jutaan orang di dunia. Diagnosis dini sangat penting untuk mencegah komplikasi jangka panjang, namun tantangan utama terletak pada kompleksitas data medis dan pemilihan parameter optimal dalam algoritma klasifikasi. Penelitian ini bertujuan untuk mengoptimalkan algoritma K-Nearest Neighbors (KNN) menggunakan teknik Bayesian Optimization guna meningkatkan akurasi dalam klasifikasi diabetes. Dataset yang digunakan adalah "Early-stage Diabetes Risk Prediction" dari UCI Machine Learning Repository yang diproses melalui normalisasi dan encoding fitur kategorikal. Bayesian Optimization diterapkan untuk menemukan parameter optimal, seperti jumlah tetangga (k) dan metrik jarak terbaik. Hasil penelitian menunjukkan bahwa KNN yang dioptimalkan mencapai akurasi 91,34%, precision 100%, dan F1-Score 93,23%, yang menunjukkan peningkatan signifikan dibandingkan dengan model KNN standar. Kesimpulannya, optimasi KNN dengan Bayesian Optimization terbukti efektif dalam meningkatkan kinerja klasifikasi diabetes, dan dapat berkontribusi signifikan pada deteksi dini serta manajemen penyakit ini.

Kata Kunci: K-Nearest Neighbor; Bayesian Optimization; Diabetes; Klasifikasi; Machine Learning

Abstract−Diabetes is one of the chronic diseases that affects millions of people worldwide. Early diagnosis is crucial to prevent long-term complications, but the main challenges lie in the complexity of medical data and selecting optimal parameters for classification algorithms. This research aims to optimize the K-Nearest Neighbors (KNN) algorithm using Bayesian Optimization to improve accuracy in diabetes classification. The dataset used is the "Early-stage Diabetes Risk Prediction" from the UCI Machine Learning Repository, preprocessed through normalization and categorical feature encoding. Bayesian Optimization was applied to find the optimal parameters, such as the number of neighbors (k) and the best distance metric. The results show that the optimized KNN achieved 91.34% accuracy, 100% precision, and a 93.23% F1-Score, demonstrating a significant improvement over the standard KNN model. In conclusion, KNN optimization with Bayesian Optimization proves effective in enhancing diabetes classification performance and can contribute significantly to early detection and disease management.

Keywords: K-Nearest Neighbor; Bayesian Optimization; Diabetes; Classification; Machine Learning

1. PENDAHULUAN

Diabetes adalah gangguan metabolisme kronis yang mempengaruhi jutaan orang di seluruh dunia. Menurut Organisasi Kesehatan Dunia (WHO), sekitar 422 juta orang di seluruh dunia menderita diabetes pada tahun 2021, dan angka ini terus meningkat setiap tahunnya. Penyakit ini ditandai oleh kadar gula darah yang tinggi akibat gangguan produksi atau penggunaan insulin dalam tubuh [1]. Diabetes tidak hanya menimbulkan dampak serius terhadap kesehatan fisik penderita, tetapi juga mempengaruhi kualitas hidup mereka secara keseluruhan [2]. Komplikasi jangka panjang dari diabetes termasuk penyakit kardiovaskular, kerusakan saraf, dan kerusakan ginjal, yang dapat mengurangi harapan hidup dan meningkatkan beban ekonomi bagi Masyarakat [3]. Dengan prevalensi yang terus meningkat, diabetes menjadi salah satu tantangan kesehatan global utama yang memerlukan pendekatan yang komprehensif untuk pencegahan, diagnosis, dan manajemen [4]. Penelitian dan pengembangan metode diagnostik serta terapeutik yang lebih efektif sangat penting untuk mengatasi epidemi ini dan meningkatkan kualitas hidup penderita diabetes [5].

Diagnosis awal dan akurat diabetes sangat penting untuk manajemen yang efektif dan pencegahan komplikasi yang lebih serius. Mengidentifikasi diabetes sejak dini memungkinkan intervensi yang tepat dan perencanaan pengobatan yang lebih baik, yang pada gilirannya dapat mengurangi risiko komplikasi jangka panjang seperti penyakit jantung, gangguan saraf, dan kerusakan ginjal [6].

Keterlambatan dalam diagnosis sering kali mengakibatkan kondisi yang lebih parah dan pengelolaan yang lebih sulit, yang dapat berdampak negatif pada kualitas hidup pasien dan meningkatkan beban sistem kesehatan. Oleh karena itu, pengembangan metode diagnostik yang cepat dan akurat sangat krusial untuk meningkatkan hasil kesehatan dan mengurangi dampak keseluruhan dari diabetes [7]. Teknologi dan teknik diagnostik terbaru, termasuk alat berbasis data dan algoritma pembelajaran mesin, berpotensi menawarkan solusi yang lebih efisien dalam mendeteksi diabetes sejak tahap awal, sehingga memungkinkan pencegahan dan pengelolaan yang lebih efektif dari penyakit ini [8].

Metode K-Nearest Neighbor (KNN) dipilih dalam penelitian ini karena beberapa alasan utama yang membuatnya sangat sesuai untuk tugas klasifikasi, terutama dalam prediksi risiko diabetes. Pertama, KNN merupakan algoritma yang sederhana namun efektif, di mana proses klasifikasi dilakukan berdasarkan jarak terdekat antara data uji dan data latih [9]. Kesederhanaan algoritma ini membuatnya mudah diimplementasikan

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 283−290
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5975

dan dipahami tanpa memerlukan asumsi yang kompleks tentang distribusi data. Kedua, KNN sangat cocok untuk dataset berukuran kecil hingga menengah, seperti dataset yang digunakan dalam penelitian ini.

Dengan fleksibilitas dalam penentuan parameter k, KNN dapat dioptimalkan untuk menghasilkan prediksi yang lebih akurat [10]. Ketiga, karena KNN adalah algoritma non-parametrik, ia tidak memerlukan asumsi distribusi tertentu, sehingga mampu menangani data yang bervariasi dan tidak terstruktur, seperti data medis yang sering kali heterogen. Selain itu, dengan menggunakan Bayesian Optimization dalam penelitian ini, KNN dapat dioptimalkan untuk menemukan kombinasi parameter terbaik, yang secara signifikan meningkatkan akurasi klasifikasi. Dengan keunggulan-keunggulan tersebut, KNN dipilih sebagai metode yang tepat untuk mengatasi permasalahan prediksi diabetes dalam penelitian ini [11].

Beberapa penelitian terdahulu menunjukkan efektivitas algoritma K-Nearest Neighbor (KNN) dalam klasifikasi dan prediksi diabetes. Dalam "Machine Learning Methods for Diabetes Prevalence Classification in Saudi Arabia" KNN mencapai akurasi tertinggi dalam memprediksi prevalensi diabetes dengan akurasi rata-rata 94,5% [12]. Penelitian "Type 2 Diabetes Prediction using K-Nearest Neighbor Algorithm" mengonfirmasi bahwa KNN efektif dalam mengklasifikasikan pasien berdasarkan risiko diabetes tipe 2 dari data medis elektronik [13].

Selain itu, "Diabetes Prediction using Machine Learning Algorithms" menunjukkan penggunaan KNN untuk menganalisis informasi pasien dalam memprediksi diabetes [14]. Penelitian lain, "Hyperglycemia Prediction Using Machine Learning," mengoptimalkan nilai K dengan metode elbow untuk meningkatkan akurasi KNN dalam memprediksi diabetes [15].

Artikel "Performance Evaluation of Various Classifiers for Diabetes Detection: A Comparative Approach" juga menunjukkan kinerja baik KNN dalam klasifikasi diabetes berdasarkan parameter akurasi, sensitivitas, dan spesifisitas [16]. Secara keseluruhan, hasil ini menegaskan peran penting KNN dalam deteksi dan prediksi diabetes serta perlunya pengoptimalan algoritma untuk meningkatkan akurasi prediksi.

Penelitian terdahulu menunjukkan berbagai pendekatan untuk mengoptimalkan algoritma K-Nearest Neighbor (KNN) dalam prediksi diabetes. Artikel "Optimized Computational Diabetes Prediction with Feature Selection Algorithms" membahas penggunaan Recursive Feature Elimination (RFE) dan algoritma Genetik (GA) untuk pemilihan fitur yang dikombinasikan dengan KNN dan Random Forest (RF). Hasil studi ini menunjukkan bahwa meskipun RF menunjukkan performa yang lebih baik daripada KNN, performa KNN meningkat secara signifikan ketika dikombinasikan dengan RFE atau GA [17].

Penelitian "Penerapan Particle Swarm Optimization Untuk Meningkatkan Kinerja Algoritma K-Nearest Neighbor Dalam Klasifikasi Penyakit Diabetes" menerapkan teknik Particle Swarm Optimization (PSO) untuk meningkatkan akurasi KNN, dengan peningkatan akurasi dari 75% menjadi 77,213% setelah pemilihan fitur menggunakan PSO [18]. Selain itu, penelitian "Machine Learning Methods for Diabetes Prevalence Classification in Saudi Arabia" mengeksplorasi berbagai algoritma pembelajaran mesin, termasuk KNN, untuk mengklasifikasikan prevalensi diabetes di Arab Saudi, dengan model KNN berbobot mencapai akurasi tinggi sebesar 94,5% [12].

Terakhir, Studi "An Integrated Classification and Association Rule Technique for Early-Stage Diabetes Risk Prediction" menemukan bahwa KNN memberikan akurasi tertinggi sebesar 97,36% dalam memprediksi diabetes tahap awal [19]. Keseluruhan penelitian ini menunjukkan bahwa berbagai teknik optimasi dan pemilihan fitur dapat secara signifikan meningkatkan kinerja KNN dalam prediksi diabetes.

Pengoptimalan algoritma K-Nearest Neighbor (KNN) dengan Bayesian Optimization dapat meningkatkan akurasi dan efisiensi klasifikasi diabetes. KNN, meskipun sederhana, efektif dalam deteksi diabetes, namun tantangannya terletak pada pemilihan parameter optimal. Bayesian Optimization menawarkan pendekatan sistematis untuk menemukan konfigurasi parameter terbaik, meningkatkan akurasi dan mengurangi waktu komputasi. Ini akan memperkuat diagnosis dini dan manajemen diabetes dalam praktik klinis.

Penggunaan Bayesian Optimization untuk mengoptimalkan algoritma K-Nearest Neighbor (KNN) adalah kunci dalam meningkatkan akurasi klasifikasi diabetes. KNN sering menghadapi tantangan dalam memilih parameter optimal, seperti jumlah tetangga (k) dan jarak metrik. Bayesian Optimization, dengan pendekatan berbasis model probabilistik, menawarkan solusi dengan mengeksplorasi ruang parameter secara sistematis. Ini memungkinkan penemuan kombinasi parameter optimal, meningkatkan akurasi dan efisiensi klasifikasi, serta mempercepat dan memperbaiki diagnosis diabetes.

Penelitian ini bertujuan untuk mengoptimalkan algoritma K-Nearest Neighbor (KNN) dengan Bayesian Optimization guna meningkatkan klasifikasi diabetes. KNN, yang sering digunakan dalam deteksi diabetes, menghadapi kendala dalam pemilihan parameter optimal. Dengan pendekatan Bayesian Optimization, penelitian ini berupaya menemukan konfigurasi parameter terbaik secara efisien, meningkatkan akurasi dan efisiensi KNN. Hasilnya diharapkan dapat mempercepat diagnosis dan meningkatkan hasil kesehatan bagi penderita diabetes.

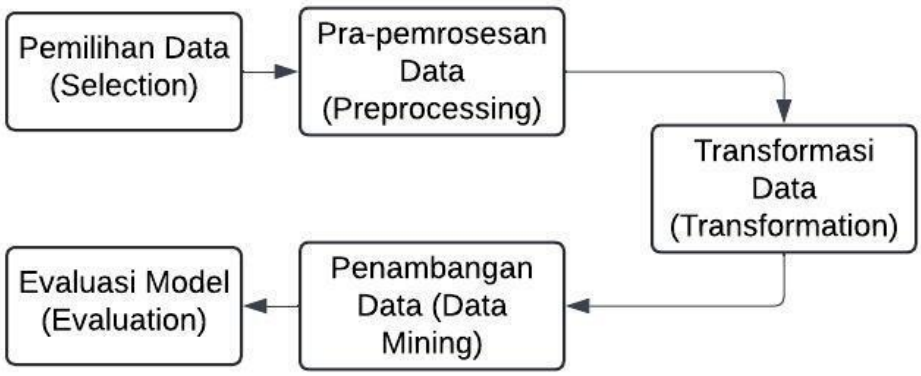
Mengoptimalkan algoritma K-Nearest Neighbor (KNN) dengan Bayesian Optimization dapat meningkatkan akurasi dan efisiensi klasifikasi diabetes dibandingkan dengan KNN standar. Bayesian Optimization, berbasis pendekatan probabilistik, menawarkan cara canggih untuk menemukan parameter optimal, mengatasi keterbatasan tuning manual, dan menghasilkan model yang lebih akurat. Peningkatan ini dapat memperbaiki diagnosis dan mempercepat analisis, memberikan manfaat signifikan dalam manajemen diabetes.

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 283−290
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5975

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini menggunakan pendekatan Knowledge Discovery in Databases (KDD) Berikut adalah gambaran tahapan dalam melakukan penelitian, dimulai dari pemilihan data hingga evaluasi hasil [20]. Tahapan penelitian ini ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian

a. Pemilihan Data (Selection)
Dataset yang digunakan adalah "Early-stage Diabetes Risk Prediction" dari UCI Machine Learning Repository. Dataset ini mencakup 520 sampel dengan 17 atribut, termasuk atribut demografis (seperti usia dan jenis kelamin) serta gejala klinis (seperti Polyuria, Polydipsia, dan penurunan berat badan mendadak). Kelas target yang digunakan adalah kategori risiko diabetes yang diklasifikasikan sebagai "Positive" atau "Negative". Pemilihan dataset ini sangat penting karena dataset yang kaya akan informasi medis klinis memiliki potensi untuk memberikan hasil prediksi yang akurat [21].

b. Pra-pemrosesan Data (Preprocessing)
Pada tahap pra-pemrosesan, data dibersihkan dan disiapkan untuk analisis lebih lanjut. Langkah pertama adalah menangani data kategorikal dengan mengonversi atribut seperti Gender dan class menjadi bentuk numerik, di mana Male dikodekan sebagai 1 dan Female sebagai 0; serta Positive dan Negative masing-masing dikodekan sebagai 1 dan 0. Selanjutnya, atribut numerik seperti Age dinormalisasi menggunakan teknik Min-Max Scaling untuk memastikan bahwa seluruh atribut berada dalam skala yang seragam. Normalisasi ini penting untuk meningkatkan kinerja algoritma, terutama pada algoritma seperti K-Nearest Neighbors (KNN) yang sensitif terhadap skala data. Selain itu, dataset dibagi menjadi 80% untuk data latih dan 20% untuk data uji menggunakan metode train_test_split. Pembagian ini bertujuan untuk menghindari masalah overfitting dan memastikan bahwa model dapat melakukan generalisasi dengan baik pada data baru yang belum pernah dilihat sebelumnya. Dengan langkah-langkah ini, dataset dipersiapkan secara optimal untuk membangun model machine learning yang akurat [22].

c. Transformasi Data (Transformation)
Pada tahap ini, analisis korelasi dilakukan untuk mengevaluasi hubungan antara variabel input dan target (class). Hasil analisis menunjukkan bahwa semua fitur memiliki relevansi klinis yang kuat dengan prediksi diabetes, sehingga tidak ada fitur yang dihilangkan. Penerapan metode analisis korelasi ini memastikan fitur-fitur yang digunakan dalam model KNN relevan dan dapat meningkatkan akurasi klasifikasi [23].

d. Penambangan Data (Data Mining)
Pada tahap ini, proses penambangan data dilakukan dengan membangun model klasifikasi menggunakan algoritma K-Nearest Neighbors (KNN). Untuk meningkatkan performa KNN, digunakan teknik Bayesian Optimization yang bertujuan untuk menemukan parameter optimal seperti jumlah tetangga (k) dan metrik jarak yang paling sesuai. Penerapan Bayesian Optimization ini merupakan inti dari penelitian dan bertujuan untuk memaksimalkan akurasi klasifikasi. Algoritma dioptimalkan menggunakan validasi silang K-fold dengan K=5 untuk menghindari overfitting dan memastikan model dapat melakukan generalisasi dengan baik [24].

e. Evaluasi Hasil (Evaluation)
Evaluasi dilakukan dengan menggunakan beberapa metrik kinerja, termasuk Akurasi, Precision, Recall, dan F1-Score. Selain itu, digunakan juga confusion matrix untuk memvisualisasikan performa klasifikasi model. ROC-AUC (Area Under Curve) digunakan untuk menilai kemampuan model dalam membedakan antara kelas positif dan negatif. Penerapan metrik evaluasi ini memberikan gambaran yang jelas tentang seberapa baik model dalam klasifikasi risiko diabetes [25].

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 283−290
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5975

3. HASIL DAN PEMBAHASAN

Bagian ini akan memaparkan hasil dari penelitian yang telah dilakukan beserta pembahasan mendalam tentang bagaimana optimasi model K-Nearest Neighbors (KNN) dengan Bayesian Optimization berhasil meningkatkan performa klasifikasi diabetes. Penjelasan ini mencakup pemilihan data, proses pra-pemrosesan, transformasi data, dan evaluasi kinerja model.

3.1 Pemilihan Data (Selection)

Data yang digunakan dalam penelitian ini berasal dari "Early-stage Diabetes Risk Prediction Dataset" yang tersedia di UCI Machine Learning Repository. Dataset ini dipilih karena mencakup data demografis yang penting dan gejala klinis yang relevan untuk prediksi dini diabetes. Dataset ini terdiri dari 520 sampel dengan 17 atribut, termasuk variabel demografis seperti usia dan jenis kelamin, serta gejala klinis seperti polyuria, polydipsia, dan penurunan berat badan mendadak. Variabel target dalam penelitian ini adalah klasifikasi risiko diabetes dengan kategori "Positive" atau "Negative". Pemilihan dataset ini sangat penting karena dataset yang kaya akan informasi medis klinis memiliki potensi untuk memberikan hasil prediksi yang akurat. Selain itu, dataset dari UCI Machine Learning Repository dikenal memiliki kualitas yang baik dan sering digunakan dalam penelitian klasifikasi penyakit, sehingga memungkinkan peneliti untuk membandingkan hasilnya dengan penelitian serupa yang sudah ada. Tabel berikut menunjukkan beberapa contoh data yang digunakan dalam penelitian ini:

Tabel 1. Dataset Diabetes

Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush
40	Male	No	Yes	No	Yes	No	No
58	Male	No	No	No	Yes	No	No
41	Male	Yes	No	No	Yes	Yes	No
45	Male	No	No	Yes	Yes	Yes	Yes
60	Male	Yes	Yes	Yes	Yes	Yes	No
55	Male	Yes	Yes	No	Yes	Yes	No
70	Male	No	Yes	Yes	Yes	Yes	No
44	Male	Yes	Yes	No	Yes	No	Yes
38	Male	Yes	Yes	No	No	Yes	Yes
35	Male	Yes	No	No	No	Yes	Yes
61	Male	Yes	Yes	Yes	Yes	Yes	Yes
60	Male	Yes	Yes	No	Yes	Yes	No
58	Male	Yes	Yes	No	Yes	Yes	No
54	Male	Yes	Yes	Yes	Yes	No	Yes
32	Female	No	No	No	Yes	No	No
42	Male	No	No	No	No	No	No

Dengan atribut-atribut ini, penelitian berfokus pada analisis hubungan antara gejala-gejala tersebut dan kemungkinan seseorang memiliki risiko diabetes.

3.2 Pra-pemrosesan Data (Preprocessing)

Pra-pemrosesan data adalah langkah krusial yang bertujuan untuk membersihkan data dan mempersiapkannya untuk analisis lebih lanjut. Langkah-langkah pra-pemrosesan dalam penelitian ini meliputi:

a. Penggantian Nilai Kategorikal dengan Numerik: Variabel kategorikal seperti Gender diubah menjadi nilai numerik, dengan Male dikodekan sebagai 1 dan Female sebagai 0. Begitu juga variabel target Risk, di mana Positive dikodekan sebagai 1 dan Negative sebagai 0. Langkah ini penting untuk memfasilitasi pemrosesan oleh algoritma machine learning yang lebih efektif dalam mengolah data numerik.
b. Normalisasi Data: Variabel-variabel numerik seperti usia dinormalisasi menggunakan teknik Min-Max Scaling. Normalisasi dilakukan untuk memastikan bahwa semua variabel berada dalam skala yang sama, sehingga setiap atribut memiliki bobot yang seimbang dalam model. Teknik ini penting karena atribut dengan skala yang berbeda dapat mendominasi model dan menurunkan kinerja klasifikasi.
c. Pembagian Data: Dataset dibagi menjadi dua bagian, yaitu 80% untuk data latih dan 20% untuk data uji. Pembagian ini dilakukan menggunakan metode train_test_split, yang bertujuan untuk menghindari masalah overfitting dan memastikan bahwa model dapat melakukan generalisasi dengan baik pada data baru yang belum pernah dilihat sebelumnya.

Pra-pemrosesan data ini memastikan bahwa dataset siap digunakan dalam pembangunan model machine learning yang andal dan akurat.

3.3 Transformasi Data (Transformation)

Dalam tahap ini, analisis korelasi dilakukan untuk mengevaluasi hubungan antara variabel-variabel input dengan target (class). Hasil analisis menunjukkan bahwa semua fitur memiliki relevansi klinis yang kuat dengan prediksi

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 283−290
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5975

diabetes, sehingga tidak ada fitur yang dihilangkan. Hal ini memastikan bahwa semua fitur digunakan untuk membangun model yang komprehensif dalam mendeteksi risiko diabetes.

3.4 Penambangan Data (Data Mining)

Pada tahap penambangan data, model klasifikasi dibangun menggunakan algoritma K-Nearest Neighbors (KNN). Algoritma KNN dipilih karena kesederhanaannya dalam mengklasifikasikan data berdasarkan kedekatan dengan data lain yang sudah diketahui klasifikasinya. Namun, untuk meningkatkan kinerja algoritma ini, dilakukan optimasi menggunakan teknik Bayesian Optimization.

Bayesian Optimization adalah teknik yang digunakan untuk menemukan konfigurasi parameter optimal secara efisien. Dalam konteks KNN, dua parameter utama yang dioptimalkan adalah jumlah tetangga (k) dan metrik jarak yang digunakan. Tujuannya adalah untuk memaksimalkan performa model KNN dalam mengklasifikasikan risiko diabetes.

Pelatihan model dilakukan menggunakan validasi silang K-fold dengan nilai K = 5. Validasi silang ini bertujuan untuk menghindari overfitting dan memberikan evaluasi yang lebih akurat terhadap performa model. Setelah parameter optimal ditemukan, model KNN dilatih menggunakan data latih yang telah dipra-proses dan dievaluasi menggunakan data uji.

3.5 Evaluasi Hasil (Evaluation)

Model K-Nearest Neighbor (KNN) dievaluasi dengan dan tanpa teknik pruning. Hasil evaluasi ditunjukkan pada tabel berikut:

Tabel 2. kinerja model

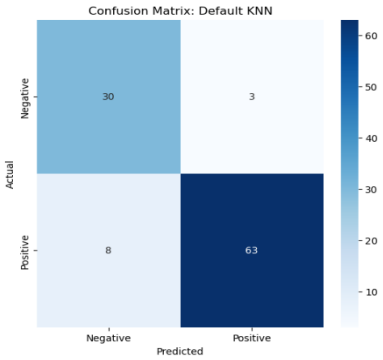
Model	Akurasi (%)	Presisi (%)	Recal (%)	F1-Score (%)
KNN Default	89.42	95.45	88.73	91.97
KNN Optimized	91.34	100	87.32	93.23

Dari hasil di atas, dapat dilihat bahwa model KNN yang dioptimalkan mencapai akurasi sebesar 91.34%, meningkat dari 89.42% pada model default. Selain itu, presisi model yang dioptimalkan mencapai 100%, yang berarti bahwa model ini sangat akurat dalam mengidentifikasi sampel positif diabetes tanpa membuat kesalahan dalam prediksi positif palsu (false positive).

Namun, recall pada model yang dioptimalkan sedikit menurun menjadi 87.32% dari 88.73% pada model default. Penurunan ini menunjukkan bahwa meskipun model lebih akurat dalam memprediksi kasus positif, ada beberapa sampel positif yang tidak terdeteksi (false negative). Namun, peningkatan pada F1-Score menjadi 93.23% menandakan bahwa model yang dioptimalkan lebih seimbang dalam memprediksi kasus positif dan negatif, serta memberikan performa keseluruhan yang lebih baik dibandingkan model default.

3.6 Confusion Matrix

Untuk memberikan gambaran yang lebih jelas tentang performa model, Gambar 2 menunjukkan confusion matrix dari model KNN default. Confusion matrix ini menggambarkan jumlah prediksi yang benar dan salah yang dilakukan oleh model dalam mengklasifikasikan sampel ke dalam kelas positif dan negatif.



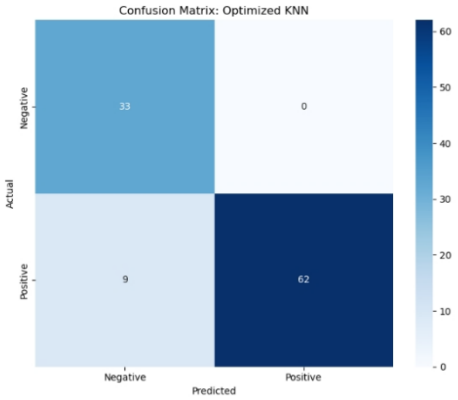
Gambar 2. confusion matrix KNN Default

Menunjukkan performa yang cukup baik dalam mengklasifikasikan data. Model ini berhasil memprediksi 30 sampel sebagai negatif yang benar-benar negatif dan 63 sampel sebagai positif yang benar-benar positif, menghasilkan akurasi sebesar 89.4%. Precision model mencapai 95.5%, yang berarti model ini jarang memberikan prediksi positif yang salah (false positive). Selain itu, recall sebesar 88.7% menunjukkan bahwa model ini cukup efektif dalam mendeteksi sampel positif, meskipun ada beberapa sampel positif yang tidak terdeteksi (false negative). Secara keseluruhan, dengan F1-Score sekitar 92%, model KNN ini mampu menyeimbangkan precision dan recall, sehingga cocok untuk digunakan dalam situasi di mana kedua metrik tersebut penting. Namun, perlu diingat bahwa terdapat beberapa kesalahan dalam prediksi, seperti 3 sampel yang diprediksi positif padahal negatif

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 283−290
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5975

dan 8 sampel yang diprediksi negatif padahal sebenarnya positif. Bagian ini juga wajib menyajikan mekanisme pengujian penelitian (sistem software) beserta pembahasan hasil pengujian. Pengujian dapat berupa: pengujian fungsionalitas dan non fungsionalitas sistem, dan atau jenis pengujian software lainnya. Hasil-hasil pengujian perlu disertai pembahasan, dengan mengulas mengenai sejauh mana permasalahan yang diidentifikasi pada awal tulisan telah diselesaikan pada akhir penelitian/kajian, dan atau sejauh mana kebutuhan-kebutuhan Fungsional dan Non Fungsional telah dipenuhi oleh sistem yang diusulkan. Pembahasan juga mesti mengaitkan sejauh mana relevansi antara hasil temuan penelitian/kajian saat ini terhadap hasil-hasil temuan penelitian terdahulu yang relevan, apakah saling menguatkan, saling mempertentangkan, atau merupakan temuan baru.

Selanjutnya, Gambar 3 menunjukkan confusion matrix untuk model KNN yang dioptimalkan. Penerapan Bayesian Optimization pada model ini diharapkan dapat meningkatkan performa klasifikasi secara signifikan.

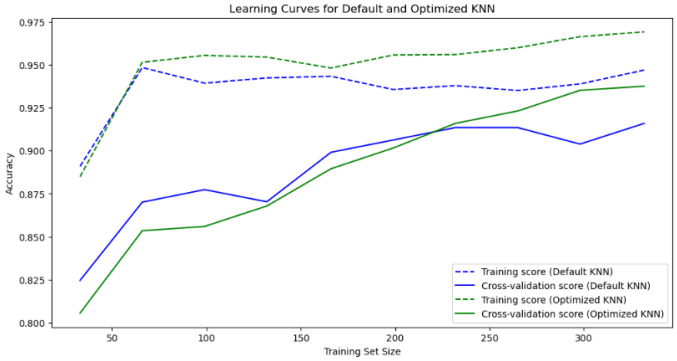


Gambar 3. confusion matrix KNN Optimized

Menunjukkan peningkatan kinerja yang signifikan. Dengan akurasi sebesar 91.3%, model ini berhasil memprediksi sampel dengan cukup akurat. Precision mencapai 100%, yang berarti model tidak membuat kesalahan dalam memprediksi sampel positif (tidak ada false positive). Namun, recall sebesar 87.3% menunjukkan bahwa meskipun model ini efektif dalam mendeteksi sampel positif, masih terdapat beberapa sampel positif yang tidak terdeteksi (false negative). Dengan F1-Score sekitar 93%, model ini menunjukkan keseimbangan yang baik antara precision dan recall. Dibandingkan dengan model KNN default, optimisasi ini secara signifikan meningkatkan precision dengan menghilangkan false positive, meskipun recall sedikit menurun. Secara keseluruhan, model KNN yang dioptimalkan ini menunjukkan performa yang lebih baik dalam menghindari kesalahan prediksi positif, namun tetap memiliki tantangan dalam mendeteksi semua sampel positif dengan benar.

3.7 Learning Curve

Gambar 4 menunjukkan grafik learning curve yang membandingkan model KNN default dan model KNN yang dioptimalkan. Learning curve memberikan gambaran tentang bagaimana akurasi model berubah seiring dengan bertambahnya ukuran data latih.



Gambar 4. Learning Curve KNN Defaulf dan KNN Optimized

Dari grafik learning curve, terlihat bahwa model K-Nearest Neighbors (KNN) yang dioptimalkan menunjukkan performa yang lebih baik dibandingkan dengan model default KNN. Model default KNN cenderung mengalami overfitting, di mana akurasi pada data training tinggi, tetapi akurasi pada validasi silang lebih rendah dan stabil setelah mencapai sekitar 100 sampel. Hal ini menunjukkan bahwa model default KNN tidak mampu melakukan generalisasi dengan baik pada data baru. Di sisi lain, model KNN yang dioptimalkan menunjukkan peningkatan performa, dengan skor validasi silang yang lebih tinggi dan stabil seiring bertambahnya ukuran data training. Meskipun akurasi pada data training juga tinggi, model yang dioptimalkan tetap mampu menjaga

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 283−290
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5975

kemampuan generalisasi yang lebih baik dibandingkan dengan model default. Secara keseluruhan, model yang dioptimalkan lebih andal untuk prediksi di dunia nyata karena dapat mempelajari pola pada data tanpa kehilangan kemampuan untuk mengeneralisasi ke data baru.

4. KESIMPULAN

Dari hasil penelitian dan analisis yang telah dilakukan, dapat disimpulkan bahwa optimasi model K-Nearest Neighbors (KNN) menghasilkan peningkatan performa yang signifikan dibandingkan dengan model default. Model KNN yang dioptimalkan berhasil mencapai akurasi sebesar 91.34%, meningkat dari 89.42% pada model default. Selain itu, presisi model yang dioptimalkan mencapai 100%, menunjukkan bahwa model ini sangat akurat dalam mengidentifikasi kasus positif diabetes, tanpa menghasilkan kesalahan prediksi positif. Meskipun recall sedikit menurun menjadi 87.32%, perbedaan ini tidak signifikan dan tetap menunjukkan kemampuan model dalam mendeteksi kasus positif dengan baik. Peningkatan F1-Score dari 91.97% pada model default menjadi 93.23% pada model yang dioptimalkan juga mengindikasikan bahwa keseimbangan antara presisi dan recall lebih baik dicapai setelah optimasi. Grafik learning curve juga menunjukkan bahwa model yang dioptimalkan memiliki kemampuan generalisasi yang lebih baik dibandingkan dengan model default, menjadikannya lebih andal untuk prediksi pada data baru. Berdasarkan hasil penelitian ini, disarankan agar peneliti mengeksplorasi algoritma lain seperti Random Forest, Support Vector Machine (SVM), atau Deep Neural Networks (DNN). Dengan mencoba algoritma yang berbeda, akan diperoleh gambaran yang lebih luas mengenai algoritma mana yang paling efektif dalam menangani dataset medis yang lebih kompleks dan beragam. Penggunaan algoritma yang berbeda juga dapat membantu mengidentifikasi kelebihan dan kelemahan masing-masing metode, sehingga hasil yang diperoleh dapat lebih optimal dalam aplikasi di dunia nyata.

REFERENCES

[1] J. Kumar, R. K. Tiwari, and V. Pandey, “Diabetes prediction using machine learning tools,” in 2021 4th International Conference on Recent Trends in Computer Science and Technology (ICRTCST), IEEE, Feb. 2022, pp. 263–267. doi: 10.1109/ICRTCST54752.2022.9781963.
[2] D. Mohajan and H. K. Mohajan, “Basic Concepts of Diabetics Mellitus for the Welfare of General Patients,” Stud. Soc. Sci. Humanit., vol. 2, no. 6, pp. 23–31, Jun. 2023, doi: 10.56397/SSSH.2023.06.03.
[3] J. Wu et al., “Associations Among Microvascular Dysfunction, Fatty Acid Metabolism, and Diabetes,” Cardiovasc. Innov. Appl., vol. 8, no. 1, 2023, doi: 10.15212/CVIA.2023.0076.
[4] A. Mishra, “Cardiovascular complications of diabetes mellitus,” InnovAiT Educ. Inspir. Gen. Pract., vol. 15, no. 6, pp. 354–361, Jun. 2022, doi: 10.1177/17557380221086012.
[5] C. G. Yedjou et al., “The Management of Diabetes Mellitus Using Medicinal Plants and Vitamins,” Int. J. Mol. Sci., vol. 24, no. 10, p. 9085, May 2023, doi: 10.3390/ijms24109085.
[6] Q. SAIHOOD and E. SONUÇ, “A practical framework for early detection of diabetes using ensemble machine learning models,” Turkish J. Electr. Eng. Comput. Sci., vol. 31, no. 4, pp. 722–738, Jul. 2023, doi: 10.55730/1300-0632.4013.
[7] H. Sharma, R. Kumar, and M. Gupta, “Optimised Machine Learning Algorithm for Detection And Diagnosis of Diabetes,” in 2023 2nd International Conference on Computational Modelling, Simulation and Optimization (ICCMSO), IEEE, Jun. 2023, pp. 21–27. doi: 10.1109/ICCMSO59960.2023.00018.
[8] A. R. Kulkarni et al., “Machine-learning algorithm to non-invasively detect diabetes and pre-diabetes from electrocardiogram,” BMJ Innov., vol. 9, no. 1, pp. 32–42, Jan. 2023, doi: 10.1136/bmjinnov-2021-000759.
[9] A. Asmarani et al., “Implementasi Algoritma K-Nearst Neighbor Untuk Memprediksi Penyakit Diabetes,” J. Inform. Dan Rekayasa Komputer(JAKAKOM), vol. 2, no. 2, pp. 231–239, 2022, doi: 10.33998/jakakom.2022.2.2.110.
[10] P. Sejati, Munawar, M. Pilliang, and H. Akbar, “Studi Komparasi Naive Bayes , K-Nearest Neighbor, dan Random Forest Untuk Prediksi Calon Mahasiswa Yang Diterima Atau Comparative Study Of Naive Bayes , K-Nearest Neighbor , And Random Forest For The Prediction Of Prospective Students,” J. Teknol. Inf. dan Ilmu Komput., vol. 9, no. 7, pp. 1341–1348, 2022, doi: 10.25126/jtiik.202296737.
[11] U. Hasanah, L. R. Mayangsari, A. Pratama, and I. Cholissodin, “Perbandingan Metode SVM, FUZZY-KNN, Dan BDT-SVM Untuk Klasifikasi Detak Jantung Hasil Elektrokardiografi,” J. Teknol. Inf. dan Ilmu Komput., vol. 3, no. 3, p. 201, 2016, doi: 10.25126/jtiik.201633196.
[12] E. S. Almutairi and M. F. Abbod, “Machine Learning Methods for Diabetes Prevalence Classification in Saudi Arabia,” Modelling, vol. 4, no. 1, pp. 37–55, Jan. 2023, doi: 10.3390/modelling4010004.
[13] S. Suriya and J. Joanish Muthu, “Type 2 Diabetes Prediction using K-Nearest Neighbor Algorithm,” J. Trends Comput. Sci. Smart Technol., vol. 5, no. 2, pp. 190–205, Jun. 2023, doi: 10.36548/jtcsst.2023.2.007.
[14] I. Journal, “Diabetes Prediction using Machine Learning Algorithms,” INTERANTIONAL J. Sci. Res. Eng. Manag., vol. 07, no. 02, Feb. 2023, doi: 10.55041/IJSREM17771.
[15] A. Modi, S. Kumar, and G. Geetha, “Hyperglycemia Prediction Using Machine Learning,” in 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), IEEE, Jul. 2023, pp. 1–8. doi: 10.1109/ICCCNT56998.2023.10306993.
[16] B. Sharma, “Performance Evaluation of Various Classifiers for Diabetes Detection: A Comparative Approach,” Int. J. Converg. Healthc., vol. 2, no. 1, 2022, doi: 10.55487/ijcih.v2i1.18.
[17] X. Li, M. Curiger, R. Dornberger, and T. Hanne, “Optimized Computational Diabetes Prediction with Feature Selection Algorithms,” in Proceedings of the 2023 7th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence, New York, NY, USA: ACM, Apr. 2023, pp. 36–43. doi: 10.1145/3596947.3596948.

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 283−290
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5975

[18] D. Susilowati, S. Sutrisno, and M. Yunus, "Penerapan Particle Swarm Optimization Untuk Meningkatkan Kinerja Algoritma K-Nearest Neighbor Dalam Klasifikasi Penyakit Diabetes," J-REMI J. Rekam Med. dan Inf. Kesehat., vol. 4, no. 3, pp. 176–184, Jun. 2023, doi: 10.25047/j-remi.v4i3.3980.

[19] D. S. Khafaga, A. H. Alharbi, I. Mohamed, and K. M. Hosny, "An Integrated Classification and Association Rule Technique for Early-Stage Diabetes Risk Prediction," Healthcare, vol. 10, no. 10, p. 2070, Oct. 2022, doi: 10.3390/healthcare10102070.

[20] Samsari, P. P. Irfana, and N. Zulkarnaim, "Prediction of Cocoa Productivity in Mamuju Regency with the K-Nearest Neighbor Algorithm," In Search, 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:228829578

[21] T. Palabaş, "EARLY-STAGE DIABETES RISK PREDICTION USING MACHINE LEARNING TECHNIQUES BASED ON ENSEMBLE APPROACH," Eskişehir Tek. Üniversitesi Bilim ve Teknol. Derg. - C Yaşam Bilim. Ve Biyoteknoloji, 2024, [Online]. Available: https://api.semanticscholar.org/CorpusID:271555897

[22] N. Sakib et al., "EEG-Driven Age Prediction: Advancements in Machine Learning Models," 2023 3rd Int. Conf. Electron. Electr. Eng. Intell. Syst., pp. 383–388, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:266194635

[23] G. F. Fahrudin, S. Suroso, and S. Soim, "Pengembangan Model Support Vector Machine untuk Meningkatkan Akurasi Klasifikasi Diagnosis Penyakit Jantung," J. Teknol. Sist. Inf. dan Apl., 2024, [Online]. Available: https://api.semanticscholar.org/CorpusID:272448734

[24] D. A. Anggoro, "Comparison of Accuracy Level of Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) Algorithms in Predicting Heart Disease," Int. J. Emerg. Trends Eng. Res., 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:225829420

[25] N. B, R. D. S, R. Annamalai, R. Bhuvaneswari, and S. S. Husain, "An Exploration of the Performance using Ensemble Methods Utilizing Random Forest Classifier for Diabetes Detection," 2023 Int. Conf. Network, Multimed. Inf. Technol., pp. 1–7, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:264293171