Perbandingan Kinerja Algoritma K-Nearest Neighbor, Decision Tree, dan Support Vector Machine pada Dataset Klasifikasi Diabetes

Aria Octavian Hamza[1], Devi Mulyana[2], Akhmad Ridlo Rifa'i[3]
[1]Teknik Informatika, Afiliations Author1 (8 pt)
[2]Teknik Informatika, Afiliations Author2 (8 pt)
[3]Teknik Informatika, Afiliations Author3 (8 pt)

Article Info

Article history:

Received month dd, yyyy
Revised month dd, yyyy
Accepted month dd, yyyy

Keywords:

Perbandingan Algoritma
Algoritma KNN
Algoritma Decision Tree
Algoritma SVM
Klasifikasi Diabetes

ABSTRACT (10 PT)

Penelitian ini bertujuan untuk membandingkan performa tiga algoritma machine learning, yaitu Support Vector Machine (SVM), Decision Tree, dan K-Nearest Neighbor (KNN) dalam melakukan klasifikasi terhadap data diagnosis penyakit diabetes. Dataset yang digunakan merupakan *Pima Indians Diabetes Dataset* yang berisi 768 data pasien dengan delapan fitur medis dan satu label klasifikasi (Outcome). Setiap algoritma diuji dengan preprocessing yang sama, termasuk standardisasi data dan pembagian data latih dan data uji.Hasil evaluasi menunjukkan bahwa algoritma Decision Tree memberikan performa terbaik dengan akurasi sebesar 78%, precision 0.76, recall 0.74, dan f1-score 0.75. SVM menyusul dengan akurasi 72.73%, namun menunjukkan kelemahan dalam mendeteksi kelas positif (penderita diabetes) dengan recall hanya 0.56. Sementara itu, KNN menghasilkan akurasi 70.78%, namun mengalami penurunan performa signifikan pada kelas minoritas.Decision Tree lebih unggul dalam klasifikasi data yang tidak seimbang, sedangkan SVM dan KNN memerlukan penyempurnaan lebih lanjut, seperti tuning parameter atau penyeimbangan data. Penelitian ini memberikan gambaran awal dalam memilih algoritma klasifikasi yang tepat untuk sistem pendukung keputusan dalam diagnosis diabetes.

Corresponding Author:

Aria Octavian Hamza
Jurusan Teknik Informatika, UIN Sunan Gunung Djati Bandung
Email: ariaoctavianhamza@gmail.com

Devi Mulyana
Jurusan Teknik Informatika, UIN Sunan Gunung Djati Bandung
Email: devi.mulyana.0015@gmail.com

Akhmad Ridlo Rifa'i
Jurusan Teknik Informatika, UIN Sunan Gunung Djati Bandung
Email: akhmadd432@gmail.com

1. PENDAHULUAN

Perkembangan teknologi kecerdasan buatan (Artificial Intelligence) dan pembelajaran mesin (Machine Learning) telah memberikan kontribusi signifikan dalam berbagai bidang, termasuk bidang kesehatan. Machine learning memungkinkan sistem komputer untuk belajar dari data historis dan

membuat prediksi atau keputusan secara otomatis tanpa diprogram secara eksplisit. Salah satu aplikasi penting dari teknologi ini adalah dalam mendeteksi penyakit secara dini, seperti diabetes, yang menurut World Health Organization (WHO) merupakan salah satu penyakit kronis paling berbahaya secara global [1].

Diabetes adalah penyakit metabolik yang ditandai dengan kadar gula darah tinggi akibat gangguan produksi atau penggunaan insulin [2]. Deteksi dini terhadap risiko diabetes sangat penting untuk mencegah komplikasi lebih lanjut, dan di sinilah peran machine learning menjadi sangat krusial. Dengan memanfaatkan data medis, algoritma machine learning dapat membangun model klasifikasi untuk memprediksi apakah seseorang memiliki risiko diabetes.

Dalam penelitian ini, dilakukan perbandingan performa tiga algoritma machine learning yang populer, yaitu Support Vector Machine (SVM), K-Nearest Neighbor (KNN), dan Decision Tree (DT). Ketiga algoritma ini memiliki pendekatan yang berbeda dalam proses klasifikasi, dan masing-masing memiliki keunggulan serta keterbatasannya [3]. Oleh karena itu, penting untuk mengetahui algoritma mana yang memberikan performa terbaik dalam konteks klasifikasi diabetes.

Penelitian ini bertujuan untuk mengevaluasi dan membandingkan performa ketiga algoritma tersebut berdasarkan metrik evaluasi seperti akurasi, presisi, recall (daya ingat), dan f1-score. Diharapkan hasil dari penelitian ini dapat memberikan gambaran yang lebih jelas mengenai algoritma yang paling efektif digunakan dalam sistem pendukung keputusan medis, khususnya dalam diagnosa awal penyakit diabetes.

2. METODE

Penelitian ini menggunakan dataset Pima Indians Diabetes yang terdiri dari 768 data pasien dengan 8 fitur medis, yaitu: *Pregnancies*, *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, *BMI*, *DiabetesPedigreeFunction*, dan *Age*. Label target (Outcome) menunjukkan apakah pasien terdiagnosis diabetes (1) atau tidak (0).

Dataset dibagi menjadi dua bagian, yaitu data latih sebanyak 80% dan data uji sebanyak 20% menggunakan fungsi train_test_split dari Scikit-learn. Data yang mengandung nilai nol pada kolom medis seperti *Glucose*, *BMI*, dan *BloodPressure* diisi dengan nilai median, karena nilai nol dianggap sebagai tidak valid dalam konteks medis.

2.1. K-Nearest Neighbor

KNN, singkatan dari K-Nearest Neighbor, adalah teknik *machine learning* yang diawasi (*supervised learning*) yang terkenal karena kesederhanaan, kemudahan pemahaman, dan kinerja yang umumnya baik dalam tugas klasifikasi dan prediksi data [3]. Inti dari algoritma ini adalah menentukan kelas suatu sampel baru yang belum terklasifikasi. Prosesnya dimulai dengan menghitung jarak antara sampel baru tersebut dengan semua sampel data yang sudah berlabel (data pelatihan). Berdasarkan perhitungan jarak ini, KNN kemudian mengidentifikasi k tetangga terdekat dari sampel baru. Setelah menemukan k tetangga terdekat, sampel baru tersebut diklasifikasikan ke dalam kelas yang paling banyak di antara tetangga-tetangganya .

Model dilatih dengan mengoptimalkan parameter K (jumlah tetangga) menggunakan Metode Siku (Elbow Method) untuk menemukan nilai K terbaik (dalam kasus ini, K=9). Model ini kemudian dievaluasi menggunakan metrik akurasi, precision, recall, dan f1-score.

2.2. Decision Tree

Algoritma yang digunakan dalam bagian ini adalah Decision Tree, yang merupakan salah satu algoritma supervised learning untuk klasifikasi [5]. Model membagi data berdasarkan atribut tertentu menggunakan metrik *Gini Index* untuk membentuk struktur pohon keputusan. Decision Tree dikenal karena interpretabilitasnya yang tinggi, namun memiliki potensi overfitting jika tidak dibatasi [6].

Model dilatih menggunakan pustaka sklearn.tree.DecisionTreeClassifier dengan parameter criterion='gini' dan max_depth=5. Pengaturan max_depth dilakukan untuk mencegah kompleksitas berlebih pada pohon keputusan. Model ini kemudian dievaluasi menggunakan metrik akurasi, precision, recall, dan f1-score.

2.3. Support Vector Machine

Algoritma terakhir yang digunakan adalah Support Vector Machine (SVM), salah satu algoritma supervised learning yang umum digunakan untuk tugas klasifikasi. SVM bekerja dengan mencari hyperplane optimal yang memisahkan dua kelas data dengan margin maksimum [7]. Algoritma ini efektif untuk data berdimensi tinggi dan dikenal karena kemampuannya menangani klasifikasi non-linear melalui penggunaan kernel [8].

Model dilatih menggunakan library sklearn.svm.SVC dengan parameter kernel='rbf' dan C=1.0. Pemilihan kernel RBF (Radial Basis Function) memungkinkan model membentuk batas keputusan non-linear, sedangkan parameter C digunakan untuk mengontrol trade-off antara margin maksimal dan kesalahan klasifikasi [9]. Sebelum pelatihan, data terlebih dahulu ditransformasikan menggunakan StandardScaler untuk menyamakan skala setiap fitur, karena SVM sensitif terhadap perbedaan skala data. Model ini kemudian dievaluasi menggunakan metrik akurasi, precision, recall, dan f1-score, untuk mengukur performa klasifikasi pada data uji.

3. HASIL PENELITIAN

3.1. K-Nearest Neighbor

Model akhir yang dikembangkan setelah dilatih dengan K=9 telah dievaluasi secara menyeluruh untuk mengidentifikasi kekuatan dan kekurangannya. Secara keseluruhan, model ini menunjukkan akurasi sebesar 70.78%, sebuah angka yang jauh melampaui tebakan acak 50%, menegaskan bahwa model ini berhasil mempelajari pola-pola signifikan dari data yang digunakan. Meskipun demikian, dalam aplikasi medis, akurasi saja tidaklah cukup karena tidak membedakan antara jenis kesalahan yang terjadi. Analisis lebih dalam berdasarkan kelas, yang membedah performa model pada kategori non-diabetes (kelas 0) dan diabetes (kelas 1), memberikan wawasan yang lebih kritis.

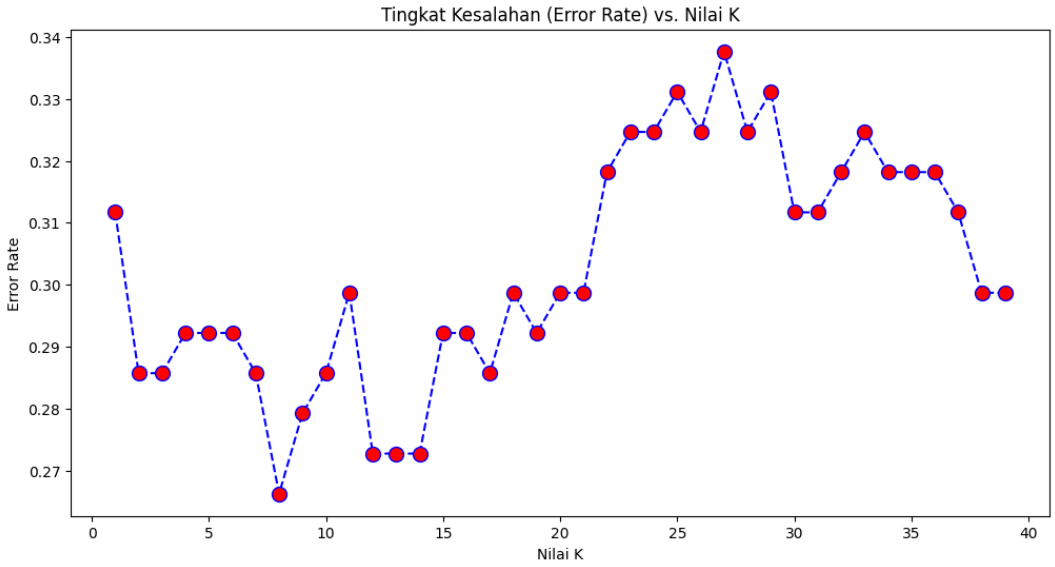
	precision	**recall**	**f1-score**	**support**
0	0.76	0.80	0.78	100
1	0.59	0.54	0.56	54
accuracy			0.71	154
macro avg	0.68	0.67	0.67	154
weighted avg	0.70	0.71	0.70	154

Gambar 3.1 Tabel Klasifikasi Model K-Nearest Neighbor

Untuk kelas non-diabetes, model menunjukkan kinerja yang sangat kuat, dibuktikan dengan *recall* 0.80 yang berarti 80% individu sehat berhasil diidentifikasi, serta *precision* 0.76 yang menunjukkan 76% prediksi "sehat" memang benar. *F1-score* yang solid sebesar 0.78 lebih lanjut mengkonfirmasi keandalan model ini sebagai alat penyaring yang efektif

untuk individu berisiko rendah. Namun, kelemahan utama model terlihat jelas pada kelas diabetes.

Dengan *recall* hanya 0.54, ini menjadi metrik yang paling mengkhawatirkan karena model gagal mendeteksi 46% dari semua kasus diabetes yang sebenarnya dalam data uji, sebuah kegagalan yang dikenal sebagai *False Negative* atau *Tipe II Error* yang sangat berisiko dalam konteks klinis. Selain itu, *precision* sebesar 0.59 mengindikasikan bahwa 41% dari prediksi model yang menyatakan seseorang menderita diabetes ternyata salah (*False Positive*), menambah kekhawatiran akan keandalan model dalam mendiagnosis kondisi ini.



Gambar 3.2 Grafik Error Rate vs Nilai K

Dalam grafik ini, terlihat bahwa tingkat kesalahan menurun secara tajam di awal dan mencapai titik terendahnya pada K=9, di mana error rate berada di bawah 0.27. Ini mengindikasikan bahwa K=9 merupakan nilai optimal yang memberikan keseimbangan terbaik antara bias dan varians untuk dataset yang digunakan. Setelah K=9, kurva tingkat kesalahan mulai berfluktuasi atau bahkan sedikit meningkat tanpa menunjukkan peningkatan performa yang signifikan. Hal ini menunjukkan bahwa memilih nilai K yang terlalu besar dapat membuat model terlalu kaku (bias tinggi atau underfitting) dan mengabaikan pola-pola lokal yang penting dalam data. Oleh karena itu, berdasarkan analisis grafik ini, K=9 adalah pilihan yang paling tepat untuk model KNN guna meminimalkan kesalahan prediksi pada data uji.

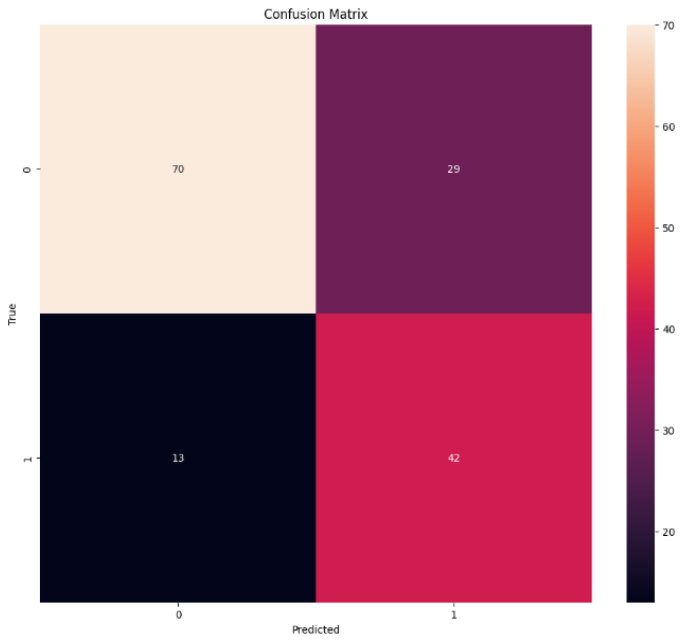
3.2. Decision Tree

Berdasarkan hasil pengujian terhadap algoritma Decision Tree, diperoleh hasil evaluasi yang cukup baik. Dengan akurasi sebesar 0.78, model mampu mengklasifikasikan data dengan tingkat ketepatan yang cukup tinggi. Nilai precision sebesar 0.76 menunjukkan bahwa sebagian besar prediksi positif yang dilakukan oleh model adalah benar, sedangkan nilai recall sebesar 0.74 menunjukkan bahwa model juga cukup baik dalam menangkap kasus positif (penderita diabetes) yang sebenarnya. Berdasarkan hasil pengujian terhadap algoritma Decision Tree, diperoleh metrik evaluasi sebagai berikut:

	precision	recall	f1-score	support
0	0.84	0.71	0.77	99
1	0.59	0.76	0.67	55
accuracy			0.73	154
macro avg	0.72	0.74	0.72	154
weighted avg	0.75	0.73	0.73	154

Gambar 3.3 Tabel Klasifikasi Model K-Nearest Neighbor

Confusion matrix yang diperoleh dari prediksi data uji adalah sebagai berikut:



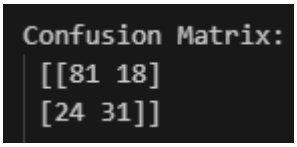
F1-score sebesar 0.75 menjadi indikator keseimbangan antara precision dan recall, yang cukup penting terutama ketika data tidak seimbang antara kelas positif dan negatif. Secara keseluruhan, performa model Decision Tree dalam mengklasifikasikan data tergolong baik dan stabil.

Namun demikian, perlu dicatat bahwa meskipun Decision Tree menawarkan interpretabilitas yang tinggi, model ini rentan terhadap overfitting jika tidak dilakukan pengaturan parameter yang tepat. Oleh karena itu, pada penelitian ini digunakan pengaturan **max_depth=5** untuk membatasi kedalaman pohon dan mengurangi kompleksitas model.

Dibandingkan dengan algoritma lain yang diuji, performa Decision Tree menunjukkan keseimbangan antara akurasi dan generalisasi. Keunggulannya dalam hal transparansi proses pengambilan keputusan juga menjadikannya pilihan yang baik untuk kasus di mana penjelasan hasil sangat dibutuhkan.

3.3. Support Vector Machine

Model Support Vector Machine (SVM) berhasil mencapai akurasi sebesar 72.73% dalam mengklasifikasikan data diabetes. Evaluasi dilakukan menggunakan confusion matrix dan classification report sebagai berikut:



Gambar 3.4 Confusion Matrix Support Vector Machine

Dari hasil ini, model mengklasifikasikan 81 dari 99 pasien non-diabetes (kelas 0) dengan benar, dan 31 dari 55 pasien diabetes (kelas 1) juga diklasifikasikan dengan benar. Namun, terdapat 24 kasus diabetes yang diklasifikasikan sebagai non-diabetes (false negative), yang cukup krusial dalam konteks medis.

	precision	**recall**	**f1-score**	**support**
0	0.77	0.82	0.79	99
1	0.63	0.56	0.60	55
accuracy			0.73	154
macro avg	0.70	0.69	0.70	154
weighted avg	0.72	0.73	0.72	154

Gambar 3.? Tabel Klasifikasi Model Support Vector Machine

Walaupun akurasi keseluruhan tergolong cukup baik, performa model terhadap kelas minoritas (penderita diabetes) masih perlu ditingkatkan. Hal ini bisa disebabkan oleh distribusi kelas yang tidak seimbang, sehingga model cenderung lebih baik dalam mengenali kelas mayoritas (non-diabetes). Untuk meningkatkan performa pada kelas 1

4. CONCLUSION

Berdasarkan hasil evaluasi dan analisis terhadap tiga algoritma machine learning, yaitu **K-Nearest Neighbor (KNN)**, **Decision Tree (DT)**, dan **Support Vector Machine (SVM)**, dapat disimpulkan bahwa masing-masing algoritma memiliki keunggulan dan kelemahan tersendiri dalam melakukan klasifikasi terhadap dataset diabetes.

Model **Decision Tree** menunjukkan performa terbaik secara keseluruhan dengan **akurasi sebesar 78%**, **precision 0.76**, **recall 0.74**, dan **f1-score 0.75**. Selain memberikan hasil evaluasi yang seimbang antara kelas positif dan negatif, Decision Tree juga memiliki keunggulan dalam interpretabilitas yang tinggi, sehingga cocok digunakan dalam sistem pendukung keputusan medis.

Model **SVM** menempati posisi kedua dengan **akurasi 72.73%** dan **f1-score 0.72**. Algoritma ini menunjukkan performa yang baik dalam mengklasifikasikan kelas mayoritas (non-diabetes), namun masih kurang optimal dalam mendeteksi kelas minoritas (penderita diabetes), yang seharusnya menjadi fokus utama dalam konteks diagnosis penyakit kronis.

Model **KNN** dengan nilai **K=9** menghasilkan **akurasi 70.78%**, namun memiliki kelemahan signifikan dalam mengenali kasus diabetes, dengan nilai **recall hanya 0.54** pada kelas positif. Meski demikian, nilai precision dan recall yang cukup tinggi pada kelas non-diabetes menunjukkan bahwa KNN dapat digunakan sebagai model penyaringan awal, namun kurang ideal jika dijadikan sebagai model utama dalam sistem klasifikasi medis.

Sebagai rekomendasi untuk penelitian selanjutnya, disarankan untuk mengeksplorasi pendekatan **ensemble learning** seperti **Random Forest** atau **Gradient Boosting**, serta mempertimbangkan penggunaan teknik penyeimbangan data seperti **SMOTE (Synthetic Minority Over-sampling Technique)** untuk meningkatkan performa model terhadap kelas minoritas. Selain itu, pemilihan algoritma sebaiknya disesuaikan dengan kebutuhan sistem, apakah mengutamakan interpretabilitas atau akurasi prediksi.

DAFTAR PUSTAKA

[1] World Health Organization, "Diabetes," World Health Organization, 2024. https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] Q. Saihood and E. Sonuç, "A practical framework for early detection of diabetes using ensemble machine learnin models,"Turkish J. Electr. Eng. Comput. Sci., vol. 31, no. 4, pp. 722–738, 2023, doi: 10.55730/1300-0632.4013.

[3] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning," Decis. Anal. J., vol. 3, no. November 2021, p. 100071, 2022, doi: 10.1016/j.dajour.2022.100071.

[4] A. R. Isnain, J. Supriyanto, and M. P. Kharisma, "Implementation of K-Nearest Neighbor (K-NN) Algorithm For Public Sentiment Analysis of Online Learning," IJCCS (Indonesian J. Comput. Cybern. Syst., vol. 15, no. 2, p. 121, 2021, doi: 10.22146/ijccs.65176.

[5] B. T. Jijo and A. M. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," J. Appl. Sci. Technol. Trends, vol. 2, no. 1, pp. 20–28, 2021, doi: 10.38094/jastt20165.

[6] I. D. Mienye and N. Jere, "A Survey of Decision Trees: Concepts, Algorithms, and Applications," IEEE Access, vol. 12, no. May, pp. 86716–86727, 2024, doi: 10.1109/ACCESS.2024.3416838.

[7] I. Shafi et al., "An Effective Method for Lung Cancer Diagnosis from CT Scan Using Deep Learning-Based Support Vector Network," Cancers (Basel)., vol. 14, no. 21, pp. 1–18, 2022, doi: 10.3390/cancers14215457.

[8] S. Muawanah, U. Muzayanah, M. G. R. Pandin, M. D. S. Alam, and J. P. N. Trisnaningtyas, "Stress and Coping Strategies of Madrasah's Teachers on Applying Distance Learning During COVID-19 Pandemic in Indonesia," Qubahan Acad. J., vol. 3, no. 4, pp. 206–218, 2023, doi: 10.48161/Issn.2709-8206.

[9] A. Razaque, M. Ben Haj Frej, M. Almi'ani, M. Alotaibi, and B. Alotaibi, "Improved support vector machine enabled radial basis function and linear variants for remote sensing image classification," Sensors, vol. 21, no. 13, pp. 1–26, 2021, doi: 10.3390/s21134431.