Vol. 2 No. (1) (Juni 2024) 36-42

Journal of Data Insights

http://journalnew.unimus.ac.id/index.php/jodi

Klasifikasi Dataset Diabetes menggunakan Algoritma K-Nearest Neighbors

Fitri Diana Musa [1], **Dannu Purwanto**[2], **Saeful Amri**[3], **Alwan Fadlurohman**[4], **Ariska Fitriyanan**[5]
[12345]FSTP Universitas Muhammadiyah Semarang, Indonesia

DOI: https://doi.org/10.26714/jodi

Info Artikel

Sejarah Artikel:
Disubmit 6 Mei 2024
Direvisi 16 Mei 2023
Disetujui 03 Juni 2024

Keywords:
Accuracy; Data Mining; KNN ; Precision; Recall.

Abstrak

Data mining merupakan suatu metode yang baik untuk menangani data skala besar. Performasi menjadi penting dalam metode data mining. Salah satu metode yang memiliki performasi terbaik adalah K-Nearest Neighbor (KNN). Artikel ini membahas terkait performasi K-NN. Data yang digunakan pada penelitian ini adalah Diabetes. Data dibagi menjadi 80% data trainingdan 20% data testing. Dengan menggunakan 11 tetangga terdekat, model menghasilkan akurasi sebesar 0.765625. Angka ini mencerminkan kinerja yang baik. Metrik kritis termasuk akurasi sebesar 0.77, presisi sebesar 0.80, dan recall sebesar 0.85. Hasil ini menunjukkan bahwa model KNN memiliki potensi untuk mengklasifikasikan pasien diabetes dengan akurasi yang baik.

Abstract

Data mining is a good method for handling large-scale data. Performance is important in data mining methods. One of the methods that has the best performance is K-Nearest Neighbor (KNN). This article discusses the performance of K-NN. The data used in this study is Diabetes. The data is divided into 80% training data and 20% testing data. By using 11 nearest neighbors, the model produces an accuracy of 0.765625. This figure reflects good performance. Critical metrics include accuracy of 0.77, precision of 0.80, and recall of 0.85. These results show that the KNN model has the potential to classify diabetic patients with good accuracy.

✉ Alamat Korespondensi:
E-mail: fitridianamusaa@gmail.com

e-ISSN: 2988 - 2109

PENDAHULUAN

Diabetes merupakan kelompok gangguan metabolik yang ditandai dengan tingginya kadar gula darah dan hiperglikemia kronis. Petersmann (2019) mengatakan penyebab seseorang dapat terkena diabetes adalah gangguan pada sekresi insulin atau efek insulin atau keduanya. Diabetes mellitus tipe 1 disebabkan oleh kerusakan sel beta yang menghasilkan defisiensi insulin absolut. Diabetes mellitus tipe 2 berkaitan dengan resistensi insulin dan defisiensi sekresi insulin yang relatif. Selain itu, terdapat juga jenis diabetes mellitus lainnya yang disebabkan oleh faktor-faktor seperti penyakit pankreas eksokrin, endokrinopati, atau penggunaan obat-obatan tertentu.

Berdasarkan data dari *World Health Organization* (WHO), pada tahun 2014, orang dewasa berusia 18 tahun keatas menderita diabetes. Dan pada tahun 2019, jumlah penderita diabetes mellitus di seluruh dunia mencapai 463 juta jiwa, dengan jumlah kematian sebanyak 4,2 juta jiwa. Menurut *World Health Organization* (WHO), gejala diabetes dapat terjadi tiba-tiba. Sedangkan menurut *American Diabetes Association* (2011) gejala seseorang terkena diabetes adalah *poliuria* (sering buang air kecil), *polidipsia* (sering haus), penurunan berat badan, *polifagia* (sering lapar), penglihatan kabur, infeksi yang sering, dan luka yang sulit sembuh.

Data mining merupakan istilah yang digunakan untuk mendeskripsikan penemuan pengetahuan dari daftar data. Data mining adalah proses mengekstraksi dan mengidentifikasi informasi berguna dan pengetahuan terkait dari berbagai daftar data besar menggunakan teknik statistik, matematika, kecerdasan buatan, dan pembelajaran mesin (Agustina et al., 2020). Pembelajaran mesin mengacu pada metode yang memberi komputer kemampuan untuk belajar dan melakukan pekerjaan secara otomatis. Proses pembelajaran mesin dilakukan melalui algoritma tertentu sehingga instruksi pekerjaan yang diberikan untuk komputer dapat diselesaikan secara otomatis (Primajaya & Sari, 2018).

Klasifikasi merupakan metode yang terdiri dari data superviced dan unsuperviced. Data superviced digunakan untuk kelompok data yang sudah diketahui kelasnya. Tujuan klasifikasi adalah untuk mendekripsikan suatu data atau objek kedalam kelas tertentu berdasarkan kemiripan karakteristik datanya. Klasifikasi dapat menggunakan clustering untuk mengelompokkan data yang didasari pada kemiripan antar data, sehingga data dengan kemiripan paling dekat berada dalam satu cluster sedangkan data yang berbeda berada dalam kelompok lainnya (Widyadhana et al., 2021). Salah satu algoritma klasifikasi yang sering digunakan dalam penelitian terkait klasifikasi yaitu menggunakan algoritma K-Neirs Neighbors. K-Nearest Neighbor (KNN) diperkenalkan oleh Fix dan Hodges pada tahun 1951. Algoritma KNN merupakan suatu algoritma non parametrik untuk klasifikasi atas dasar kebutuhan untuk melakukan analisis diskriminan ketika nilai estimasi parametrik yang fungsi peluangnya tidak diketahui atau sulit untuk ditentukan. Selain itu, algoritma ini menjadi terkenal karena sederhana dan tingkat konvergen yang relatif tinggi (Mumtaz et al., 2023). Metode KNN adalah satu yang paling banyak digunakan di dunia untuk persoalan pengklasifikasian (Fauzi, 2017).

Penelitian sebelumnya mengenai klasifikasi menggunakan algoritma KNN pernah dilakukan oleh Sharma (2016), dalam penelitian tersebut digunakan metode KNN untuk mengklasifikasikan Kanker Serviks Klinis. Hasil penelitian tersebut menunjukkan bahwa klasifikasi dengan metode KNN mampu menghasilkan akurasi klasifikasi yang cukup tinggi yaitu sekitar 84.3%. Selain itu, penelitian yang dilakukan oleh Hasan (2021), dimana penelitian dilakukan guna membandingkan akurasi algoritma klasifikasi K-Nearest Neighbor dan Random Forest dalam Seleksi Fitur Information Gain untuk Klasifikasi Lama Studi Mahasiswa. Hasil pengujian dalam penelitian diperoleh nilai akurasi dengan menggunakan KNN sebesar 86,67% dan untuk nilai akurasi dengan menggunakan RF diperoleh hasil 100%. Berdasarkan kedua penelitian diatas terlihat bahwa metode KNN memiliki performa yang baik. Oleh karena itu pada penelitian ini akan dilakukan klasifikasi menggunakan algoritma KNN.

Selanjutnya, hasil dari proses klasifikasi dengan logoritma KNN akan dibandingkan berdasarkan metrik-metrik penting seperti accuracy, precision, dan recall, yang akan membantu dalam

mengevaluasi performa dan kehandalan algoritma KNN. Dengan memahami perbandingan ini, penelitian ini bertujuan untuk memberikan wawasan yang lebih baik tentang algoritma klasifikasi KNN guna memberikan prediksi yang akurat terkait kerentanan penyakit diabetes.

METODE

Sumber Data

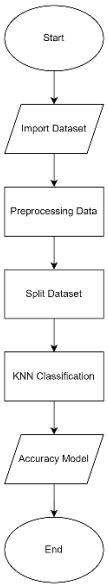
Penelitian yang dilakukan adalah klasifikasi menggunakan dataset medis dan demografis pasien untuk memprediksi diabetes dengan algoritma KNN. Hasilnya mendukung pengembangan prediksi risiko diabetes berdasarkan riwayat medis dan faktor demografis. Metode pengumpulan data pada penelitian ini menggunakan metode sekunder. Metode sekunder adalah menggunakan data yang telah dikumpulkan oleh orang lain atau lembaga sebelumnya.

Variabel Penelitian

Dataset prediksi Diabetes adalah koleksi data medis dan demografis pasien yang mencakup berbagai variabel penting. Pertama, usia pasien direkam sebagai faktor utama, mengingat risiko diabetes yang cenderung meningkat seiring bertambahnya usia. Jenis kelamin, variabel berikutnya, memungkinkan untuk mempertimbangkan perbedaan risiko antara laki-laki dan perempuan dalam prediksi diabetes. Indeks Massa Tubuh (BMI) adalah indikator obesitas yang berhubungan erat dengan diabetes, karena tingkat BMI yang tinggi sering menjadi faktor risiko. Data juga mencakup informasi tentang hipertensi, penyakit jantung, dan riwayat merokok sebagai faktor risiko potensial. Selain itu, kadar HbA1c dan glukosa darah mencerminkan kondisi kesehatan glikemik pasien, yang memiliki dampak signifikan pada prediksi diabetes. Variabel-variabel ini digunakan untuk memahami hubungan kompleks antara faktor medis dan demografis serta risiko diabetes, membantu dalam mengembangkan tingkat akurasi yang relevan.

Tahapan Penelitian

Tahapan penelitian yang dilakukan dalam penelitian ini disajikan dalam flowchart sebagai berikut:



Gambar 1. Flowchart Penelitian

HASIL DAN PEMBAHASAN

Import Dataset

Data yang digunakan merupakan data sekunder. Data sekunder adalah data yang diperoleh secara tidak secara langsung dari objeknya, tetapi melalui sumber lain, baik lisan maupun tulisan. Data didatapkan dari kaggle dan disimpan dalam format Comma Separated Value (CSV). Library pandas pada bahasa pemograman python digunakan untuk mengimport data dari csv. Data yang dikumpulkan sebanyak 100000 data dengan 9 variabel. Berikut adalah sampel data yang diperoleh.

Tabel 1. Sampel Data Diabetes

Pregnancies	*Glucose*	*Blood Pressure*	*Skin Thickness*	*Insulin*	*BMI*	*Diabetes PedigreeFunction*	*Age*	*Outcome*
6	148	72	35	0	33,6	0,627	50	1
1	85	66	29	0	26,6	0,351	31	0
8	183	64	0	0	23,3	0,672	32	1
1	89	66	23	94	28,1	0,167	21	0
0	137	40	35	168	43,1	2,288	33	1

Preprocessing Data

Preprocessing data adalah langkah penting dalam proses pengolahan data sebelum menggunakan model K-Nearest Neighbors (KNN). Tahap preprocessing data bertujuan untuk membersihkan, mengubah, dan mempersiapkan data mentah sehingga data tersebut lebih sesuai untuk digunakan dalam model KNN. Berikut adalah tahapan dalam preprocessing data menggunakan model KNN pada penelitian ini:

1. Pengumpulan data adalah mengumpulkan data mentah yang akan digunakan untuk model KNN.
2. Pembersihan data melibatkan penghapusan atau penanganan data yang hilang, duplikat, atau outlier. Outlier dapat mempengaruhi hasil KNN, sehingga perlu dikenali dan diatasi.
3. Penanganan data hilang yaitu dengan mengisi nilai yang hilang dengan nilai rata-rata atau median dari data yang bersangkutan.

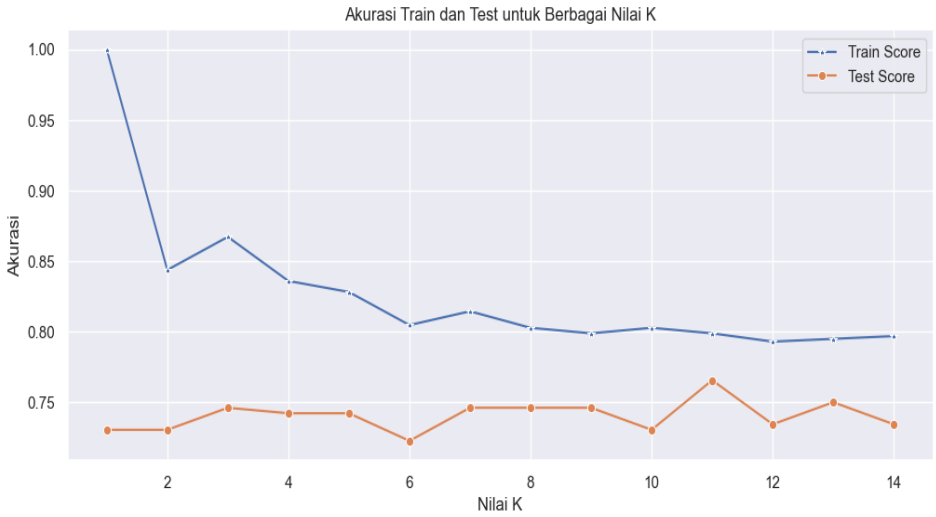
Split Data

Pembagian data (*Split Data*) dalam model K-Nearest Neighbors (KNN) adalah proses memisahkan dataset menjadi dua subset, data pelatihan dan data pengujian. Data pelatihan digunakan untuk melatih model KNN, sementara data pengujian digunakan untuk menguji kinerja model. Biasanya, data dibagi dengan perbandingan, seperti 80% data pelatihan dan 20% data pengujian. Data split digunakan untuk menghindari overfitting dan memastikan bahwa model dapat menggeneralisasi dengan baik. Dengan pembagian data, maka dapat mengukur sejauh mana model KNN dapat memprediksi data dengan akurat.

KNN Classification

Tahap ini diawali dengan pemilihan k tetangga terdekat, ini akan memengaruhi model dalam melakukan klasifikasi. Dalam penelitian ini, hasil akurasi pada data pelatihan mencapai 100% dengan penggunaan satu tetangga terdekat (K=1), menunjukkan bahwa model K-Nearest Neighbors (KNN) cocok dengan data pelatihan. Pada data pengujian, meskipun terjadi penurunan akurasi menjadi 76.5625% ketika K=11, akurasi tersebut masih dianggap cukup baik. Hasil ini mengindikasikan bahwa

model KNN memiliki kemampuan yang baik untuk menggeneralisasi data. Berikut adalah gambaran lengkap terkait nilai k dan akurasinya dalam data training dan data testing.



Gambar 2. Akurasi Nilai K dalam Data Training dan Data Testing

Gambar 2 menunjukkan bahwa nilai k tetangga terdekat yang akan dipakai untuk model KNN adalah 11. Setelah mendapatkan nilai k tetangga terdekat untuk model KNN, selanjutnya akan dihitung akurasi atau kebaikan model KNN dalam memprediksi penyekit diabetes. Setelah membagi data menjadi subset pelatihan dan pengujian, model K-Nearest Neighbors (KNN) akan dievaluasi untuk mengukur kinerjanya. Hasil evaluasi model adalah sebagai berikut.

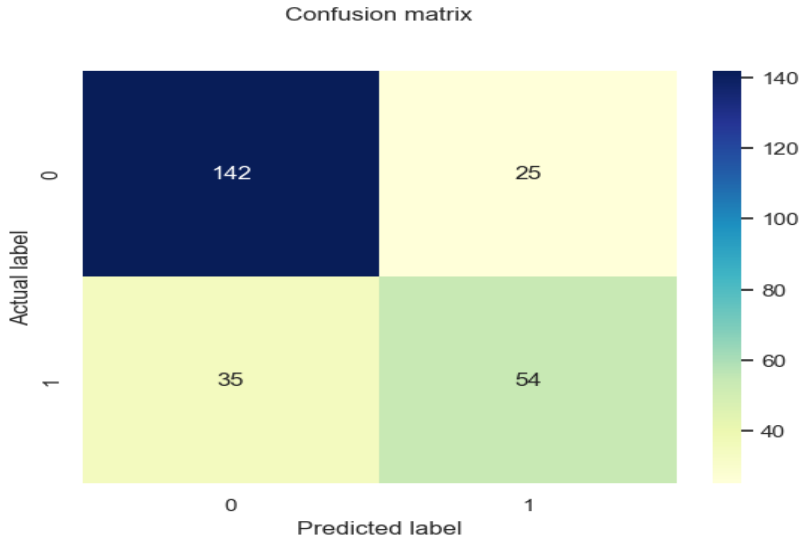
Tabel 2. Performa KNN

Performa	Nilai
F1-Score	0.83
Akurasi	0.77

Tabel 1 menunjukkan hasil akurasi model KNN yang didapat adalah sebesar 77%. Artinya model KNN cukup baik dalam melakukan prediksi terhadap data diabetes. Nilai F1-Score sebesar 83% menggambarkan bahwa sebaran hasil klasifikasi seimbang pada setiap kelas.

Accuracy Model

Dalam penelitian ini, hasil dari proses klasifikasi akan dibandingkan menggunakan pengukuran matriks performa yang meliputi tiga parameter utama, yaitu *precision*, *recall*, dan *accuracy*. Pengukuran performa model klasifikasi ini akan digunakan untuk menganalisis dan membandingkan sejauh mana kemampuan setiap algoritma dalam memberikan hasil yang akurat. Parameter-parameter ini memiliki peran penting dalam mengevaluasi performa model dan memberikan wawasan yang komprehensif tentang seberapa baik model-model tersebut dapat mengklasifikasikan data. *Confusion matrix* digunakan untuk menguji kemampuan model dalam memprediksi objek dengan benar atau salah. Berikut adalah tabel confusion matrix untuk penelitian ini:



Gambar 3. Confusion Matrix

Dari tabel confusion matrix diatas, selanjutnya akan dilihat hasil performa klasifikasi dengan menggunakan recall, accuracy dan precision.

Tabel 3. Hasil Performa KNN

Algoritma	*Accuracy (%)*	*Precision (%)*	*Recall (%)*
KNN	0.77	0.80	0.85

Tabel 2 menunjukkan hasil evaluasi model menggunakan algoritma K-Nearest Neighbors (KNN) menunjukkan akurasi sebesar 77%, presisi sebesar 80%, dan recall sebesar 85%. Akurasi mencerminkan sejauh mana model benar dalam klasifikasi keseluruhan, sedangkan presisi menggambarkan tingkat ketepatan dalam mengidentifikasi positif. Recall menunjukkan kemampuan model untuk menemukan semua kasus positif yang sebenarnya. Hasil ini mengindikasikan bahwa model KNN memiliki kemampuan yang baik dalam mengklasifikasikan data dengan tingkat akurasi yang memadai, presisi yang baik, dan kemampuan yang kuat dalam menemukan kasus positif. Model ini dapat berpotensi berguna dalam aplikasi klasifikasi dengan persyaratan tinggi terhadap identifikasi positif.

KESIMPULAN

Dalam konteks penggunaan model K-Nearest Neighbors (KNN) pada dataset Diabetes, ditemukan bahwa dengan menggunakan 11 tetangga terdekat, model menghasilkan akurasi sebesar 0.765625. Meskipun angka ini mencerminkan kinerja yang baik, evaluasi lebih lanjut mengungkapkan hasil yang lebih mendalam. Metrik kritis termasuk akurasi sebesar 0.77, presisi sebesar 0.80, dan recall sebesar 0.85. Akurasi mengukur sejauh mana model cocok dengan data, sedangkan presisi menggambarkan ketepatan dalam mengidentifikasi kelas positif, dan recall mengukur kemampuan model untuk menemukan kasus positif yang sebenarnya. Hasil ini menunjukkan bahwa model KNN memiliki potensi untuk mengklasifikasikan pasien diabetes dengan akurasi yang baik dan kemampuan yang baik dalam mengidentifikasi pasien yang berpotensi mengidap penyakit. Hasil ini dapat digunakan dalam aplikasi prediksi diabetes terhadap pasien.

DAFTAR PUSTAKA

Agustina, D., Putri, E., Fauzi, F., Alawiyah, S. N., & Wasono, R. 2020. METODE SUPPORT VECTOR MACHINE (SVM) UNTUK KLASIFIKASI DATA EKSPRESI GEN MICROARRAY. EDUSAINTEK 4, 1–10.

American Diabetes Association. (2011). Standards of medical care in diabetes—2011. Diabetes Care, 34(Supplement 1), S11-S61.

Fauzi, F. (2017). K-Nearset Neighbor (K-NN) dan Support Vector Machine (SVM) untuk Klasifikasi Indeks Pembangunan Manusia Provinsi Jawa Tengah Info Artikel. Jurnal MIPA, 40(2). http://journal.unnes.ac.id/nju/index.php/JM

Hasan, I. K., Resmawan, & Ibrahim, J. (2022). Perbandingan K-Nearest Neighbor dan Random Forest dengan Seleksi Fitur Information Gain untuk Klasifikasi Lama Studi Mahasiswa. Indonesian Journal of Applied Statistics, 5(1), 58-66. https://doi.org/10.13057/ijas.v5i1.58056

Mumtaz, G., Akram, S., Iqbal, W., Ashraf, M. U., Almarhabi, K. A., Alghamdi,A. M., & Bahaddad, A. A. (2023). Classification and Prediction of Significant Cyber Incidents (SCI) using Data Mining and Machine Learning (DM-ML). IEEE Access, 1. https://doi.org/10.1109/ACCESS.2023.3249663

Petersmann, A., Müller-Wieland, D., Landgraf, R., Nauck, M., Freckmann, G., & Heinemann, L. (2019). Screening for Diabetes Mellitus: Current Recommendations and Future Perspectives. Exp Clin Endocrinol Diabetes, 127(Suppl 1), S1-S7.

Primajaya, A., & Sari, B. N. 2018. Random Forest Algorithm for Prediction of Precipitation. Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM), 1(1), 27–31.

Sharma, M., Singh, S. K., Agrawal, P., & Madaan, V. (2016). Classification of Clinical Dataset of Cervical Cancer using KNN. Indian Journal of Science and Technology, 9(28). DOI: 10.17485/ijst/2016/v9i28/98380

Widyadhana, D., Hastuti, R. B., Kharisudin, I., & Fauzi, F. (2021). Perbandingan Analisis Klaster K-Means dan Average Linkage untuk Pengklasteran Kemiskinan di Provinsi Jawa Tengah. PRISMA, Prosiding Seminar Nasional Matematika, 4, 584–594. https://journal.unnes.ac.id/sju/index.php/prisma

World Health Organization. (2019). WHO Health Statistics Overview 2019 (pp. 1–16).