Journal of Artificial Intelligence and Software Engineering
Vol. 5, No. 1, Maret 2025, pp. 22-26
E-ISSN: 2777-001X, DOI: 10.30811/jaise.v5i1.6275

Comparison Of K-Nearest Neighbor And Naive Bayes In Diabetes Dataset

Anin Naba Anzila[1], Eko Budi Susanto[2*], Bambang Ismanto[3]
[1] Sistem Informasi, Fakultas Teknologi Informasi, Institut Widya Pratama, Kota Pekalongan, 51146, Indonesia
[2,3] Teknik Informatika, Fakultas Teknologi Informasi, Institut Widya Pratama, Kota Pekalongan, 51146, Indonesia

Informasi Artikel

Diterima : 13 Januari, 2025
Revisi : 26 Januari 2025
Publikasi : 20 Maret 2025

Kata Kunci:

K-Nearest Neighbor
Naïve Bayes
Confusion Matrix
K-Fold Cross Validation
Diabetes

ABSTRAK

Salah satu penyebab tingginya angka kematian akibat komplikasi pada penyakit diabetes adalah keterlambatan dalam melakukan diagnosis sebelum diagnosis tersebut ditegakkan. Dalam bidang medis, penerapan model machine learning telah membuka peluang signifikan dalam meningkatkan akurasi diagnosis dini diabetes. Penelitian ini bertujuan untuk membandingkan kinerja algoritma K-Nearest Neighbors (KNN) dan Naive Bayes Classifier (NBC) dengan menggunakan dataset sekunder berjumlah 128 record data dan memuat 10 variabel data yang relevan untuk prediksi diabetes. Hasil analisis menunjukkan bahwa algoritma KNN dengan parameter K=21, berdasarkan evaluasi confusion matrix mencapai akurasi sebesar 76,92%, recall 100%, precision 72% dan F1-Score 84%. Sementara itu algoitma naïve bayes memiliki akurasi 65,63%, recall 52%, precision 100% dan F1-Score 69%. Pada evaluasi dengan metode k-fold cross validation menggunakan K=10 menghasilkan rata-rata akurasi sebesar 73% untuk algoritma KNN dan 70% untuk algoritma naïve bayes. Dengan demikian, algoritma KNN lebih unggul dan direkomendasikan untuk klasifikasi penyakit diabetes.

ABSTRACT

Delay in diagnosis of diabetes is one of the causes of increasing mortality due to complications before diagnosis is made. In the medical field, the application of machine learning models has opened up significant opportunities in improving the accuracy of early diagnosis of diabetes. This study aims to compare the performance of the K-Nearest Neighbors (KNN) and Naive Bayes Classifier (NBC) algorithms using a secondary dataset of 128 data records and containing 10 data variables relevant to the prediction of diabetes. The results of the analysis show that the KNN algorithm with parameters K = 21 based on the evaluation of the confusion matrix obtained an accuracy of 76.92%, recall 100%, precision 72% and F1-Score 84%. Meanwhile, the naïve Bayes algorithm obtained an accuracy of 65.63%, recall 52%, precision 100% and F1-Score 69%. In the evaluation using the k-fold cross validation method with K = 10, the average accuracy for the KNN algorithm was 73% and for the Naïve Bayes algorithm was 70%. Thus, the KNN algorithm is superior and recommended for diabetes disease classification.

***Penulis Koresponden**

Email: ekobudi.s@stmik-wp.ac.id

Cara sitasi IEEE:

A. N. Anzila, E. B. Susanto, dan B. Ismanto, "Comparison Of K-Nearest Neighbor And Naive Bayes In Diabetes Dataset," *Journal of Artificial Intelligence and Software Engineering (J-AISE)*, vol. 5, no. 1, pp. 22-26, Maret 2025. doi:10.30811/ jaise.v5i1.6275

1. PENDAHULUAN

Salah satu penyebab tingginya angka kematian akibat komplikasi pada penyakit diabetes adalah keterlambatan dalam melakukan diagnosis sebelum diagnosis tersebut ditegakkan. Oleh karena itu, prediksi dini dengan memanfaatkan atribut-atribut pendukung menjadi sangat penting untuk mencegah dampak yang lebih parah [1].

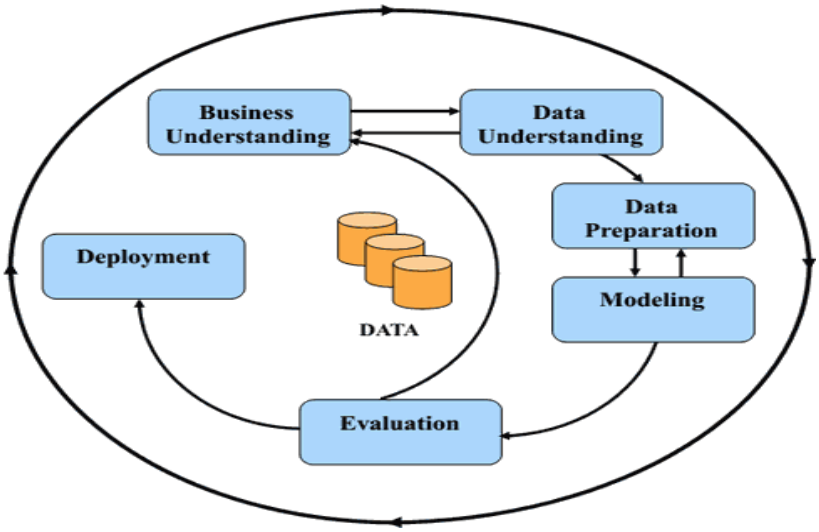
Naïve Bayes Classifier (NBC) merupakan metode klasifikasi yang menggunakan probabilitas untuk menghitung peluang berdasarkan frekuensi dan kombinasi nilai dalam dataset yang ada [2]. Sementara itu, K-Nearest Neighbor (KNN) adalah algoritma yang termasuk dalam kategori instance-based learning, sering disebut juga sebagai lazy learning. Algoritma ini melakukan analisis dengan mengidentifikasi K tetangga terdekat dalam data pelatihan yang memiliki karakteristik paling serupa dengan data baru atau data uji, kemudian menggunakan informasi tersebut untuk membuat prediksi [3]. Penelitian di bidang kesehatan telah memanfaatkan metode data mining untuk tujuan klasifikasi. Salah satunya adalah penelitian yang dilakukan oleh [2], yang menerapkan algoritma Naïve Bayes dan K-Nearest Neighbors (KNN) untuk memprediksi penderita gagal ginjal kronik. Hasilnya menunjukkan bahwa algoritma Naïve Bayes menghasilkan akurasi sebesar 94,25%, dengan rata-rata recall 94,23%, presisi 98,40%, dan Area Under Curve (AUC) sebesar 0,961.

Penelitian lain yang dilakukan oleh [4] menggunakan algoritma KNN dan Naïve Bayes untuk memprediksi klasifikasi penyakit jantung. Temuan penelitian ini mengungkapkan bahwa algoritma KNN pada nilai K=236 memperoleh akurasi 68%, presisi 66%, recall 74%, dan f-measure 70%, sementara pada nilai K=250 dengan metode jarak Euclidean, akurasi yang diperoleh adalah 65%, presisi 65%, recall 69%, dan f-measure 67%. Penelitian selanjutnya dilakukan oleh [5], yang menggunakan algoritma KNN dan Naïve Bayes untuk memprediksi klasifikasi penyakit diabetes gestasional. Temuan dari penelitian ini mengungkapkan bahwa performa algoritma KNN dengan feature selection pada iterasi K=5 mencapai akurasi sebesar 77%, sedangkan tanpa feature selection mencapai 80%. Sementara itu, Algoritma Naïve Bayes menunjukkan peningkatan akurasi menjadi 80% ketika diterapkan dengan feature selection, dibandingkan dengan akurasi 77% saat digunakan tanpa proses feature selection.

Penelitian ini bertujuan untuk menganalisis dan membandingkan kinerja algoritma K-Nearest Neighbors (KNN) dan Naive Bayes Classifier (NBC) dalam klasifikasi penyakit diabetes, guna menentukan algoritma yang lebih efektif dan akurat dalam mendeteksi penyakit diabetes. Dalam penelitian ini, peneliti menggunakan dataset berjudul Easiest Diabetes Classification yang diunduh dari situs kaggle.com untuk mendukung penelitian ini. [6]. Dataset ini terdiri atas 128 record data, yang mencakup 10 atribut dan 1 label, dengan tipe data yang bervariasi, yaitu integer, float dan object. Penilaian kinerja algoritma dilakukan melalui confusion matrix, yang digunakan untuk menghitung nilai akurasi, presisi, recall, dan F1-score.

2. METODE

Metode penelitian yang digunakan mengikuti standar pada Cross-Standard Industry for Data Mining (CRISP-DM) sebagai pendekatan untuk data mining. CRISP-DM adalah metode yang digunakan dalam pendekatan data mining. Metode CRISP-DM digunakan baik sebagai model proses dalam konteks teknis maupun sebagai metodologi proyek dalam konteks formal seperti penelitian, melibatkan setiap fase dari prosesnya [7].



Gambar 1. Tahapan Metode CRIPS-DM

Tahap business understanding bertujuan untuk memahami kontek bisnis, yaitu: penelitian ini bertujuan untuk mengevaluasi dan membandingkan efektivitas dua algoritma klasifikasi, yaitu K-Nearest Neighbors (KNN) dan Naive Bayes (NBC). Tujuan utama dari perbandingan ini adalah untuk menentukan algoritma yang lebih efisien dalam mengidentifikasi diabetes melitus, sehingga dapat meningkatkan ketepatan diagnosis dan pada akhirnya memperbaiki pengelolaan dan perawatan klinis pasien diabetes.

Tahap selanjutnya, data understanding untuk mengeksplorasi dan memahami karakteristik dataset. Dataset yang digunakan pada penelitian ini dari kaggle.com 10 variabel/atribut dan 1 label (tabel 1).

Tabel 1. Variabel Dataset Diabetes

No	Variabel	Jenis data	Tipe data
1	Age	Rasio	Int64
2	Gender	Binominal	Object
3	BMI	Rasio	Int64
4	Blood Pressure	Rasio	Object
5	FBS	Rasio	Int64
6	HbA1c	Rasio	Float64
7	Family History of Diabetes	Binominal	Object
8	Smoking	Binominal	Object
9	Diet	Ordinal	Object
10	Exercise	Ordinal	Object
11	Diagnosis	Binominal	Object

Data preparation melibatkan pengolahan variabel data diabetes untuk memastikan kualitas dan kesiapannya. Langkah-langkah yang dilakukan meliputi seleksi variabel penting, menangani data yang hilang dengan imputasi, normalisasi data untuk mengatasi perbedaan skala, serta pembagian dataset menjadi data latih dan data uji. Tahap pemodelan bertujuan untuk mengembangkan model yang dapat memprediksi hasil berdsarkan data yang ada. Penelitian ini menggunakan algoritma K-Nearest Neighbors (KNN) dan Naive Bayes Gausian. diterapkan untuk mengidentifikasi risiko diabetes pada individu berdasarkan dataset yang telah diproses. Pada tahap ini menggunakan library scikit-learn pada python.

Tahap pengujian dalam penelitian ini menggunakan pengujian confusion matrix dan K-fold cross validation [2]. Confusion matrix bertujuan untuk menilai kinerja model algoritma yang digunakan. Beberapa metrik evaluasi yang dihasilkan dari confusion matrix meliputi akurasi, presisi, recall, dan F1 Score. K-fold cross validation bertujuan untuk mengurangi resiko overfitting dan underfitting, serta memberikan estimasi performa model yang lebih akurat. Metode ini dilakukan dengan membagi dataset menjadi beberapa lipatan (folds) untuk keperluan evaluasi yang lebih representatif. Deployment merupakan tahap akhir dalam siklus pengembangan model ke dalam lingkungan operasional agar dapat digunakan secara praktis oleh pengguna akhir. Dalam proses integrasi ini, Flask digunakan sebagai framework utama untuk pengembangan aplikasi web.

3. HASIL DAN PEMBAHASAN

Tahap persiapan dataset diabetes menghasilkan data yang telah melalui proses pemeriksaan, di mana tidak ditemukan adanya nilai yang hilang, kosong, maupun menyimpang (outliers). Dengan demikian, dataset tersebut dinyatakan siap untuk digunakan dalam proses pelatihan dan pengujian model klasifikasi penyakit diabetes. Model yang diusulkan yaitu: algoritma K-Nearest Neighbors (KNN) dan Naive Bayes Gausian. Kedua algoritma tersebut dilatih dengan dataset diabetes dengan membagi menjadi 20 persen untuk data training, 80 persen untuk data testing.

Tabel 2 Hasil Pengujian K-Folds Cross Validation

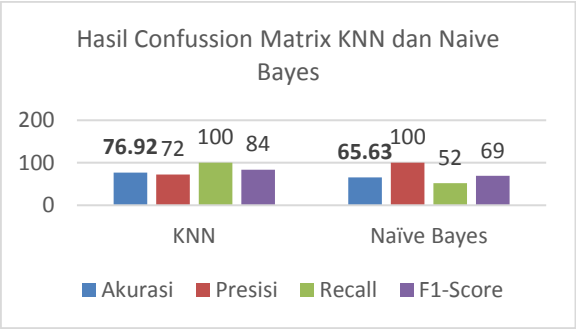
Iterasi	K-Nearest Neighbor (KNN)	Naïve Bayes
Iterasi 1	0.69230769	0.69230769
Iterasi 2	0.69230769	0.61538462
Iterasi 3	0.53846154	0.84615385
Iterasi 4	0.76923077	0.53846154
Iterasi 5	0.69230769	0.69230769
Iterasi 6	0.76923077	0.84615385
Iterasi 7	0.76923077	0.69230769
Iterasi 8	0.92307692	0.53846154
Iterasi 9	0.75	0.66666667
Iterasi 10	0.75	0.83333333
Mean Accuracy	0.73	0.70

Hasil rata-rata akurasi aglortima Naive Bayes Gausian yaitu: 70%. Pengukuran tingkat akurasi ini diambil dari nilai rata-rata pengujian K-Fold Validation, dengan nilai iterasi K=10. Sedangkan hasil rata akurasi aglortima K-Nearest Neighborn dengan menggunakan K-Fold Validation yaitu: 73%. Pengukuran tingkat akurasi ini diambil dari nilai rata-rata evaluasi K-Fold Validation. Hasil perbandingan evaluasi K-Fold Validation dapat dilihat di tabel 2.

Tabel 3 Akurasi Berdasarkan nilai K agloritma KNN

K	Akurasi (%)	K	Akurasi (%)
3	53,85	13	71,88
5	56,25	15	71,88
7	65,62	17	71,88
9	65,62	19	71,88
11	68,75	**21**	**76,92**

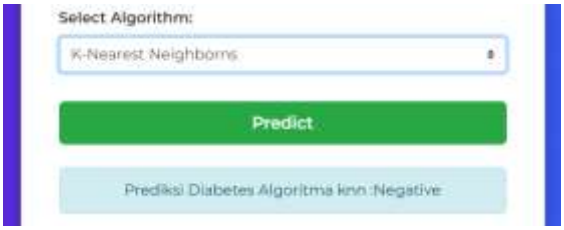
Pada algoritma KNN nilai paramater K yang terbaik yaitu pada nilai K=21 (tabel 3). Hasil confusion matrix untuk k=21 yaitu tingkat akurasi 76,92% presisi 72% recall 100% dan F1 Score 84%. Sementara itu algoitma naïve bayes memiliki akurasi 65,63%, recall 52%, precision 100% dan F1-Score 69% (gambar 2). Pada algoritma KNN nilai paramater K yang terbaik yaitu pada nilai K=21 (tabel 3). Hasil confusion matrix untuk k=21 yaitu tingkat akurasi 76,92% presisi 72% recall 100% dan F1 Score 84%. Sementara itu algoitma naïve bayes memiliki akurasi 65,63%, recall 52%, precision 100% dan F1-Score 69% (gambar 2).



Gambar 2. Hasil Confussion Matrix KNN dan Naive Bayes

Berdasarkan hasil dari evaluasi kinerja algoritma, maka akan dipilih algoritam terbaik yang akan dijadikan model untuk klafisikasi diabesetes. Algoritma yang dipilih yaitu algoritma K-NN. Pada tahap deployment, akan dikembangakan apliksi untuk memprediksi penyakit diabetes. Aplikasi dibangun dengan menggunakan framework flask, seperti yang terlihats di gambar 3 yang merupakan form input dari parameter penyakit diabetes. Pengguna akan menginputkan data seusai dengan parameter yang ada, kemudian memiliha tombol predict. Selanjutnya aplikasi akan menampilkan hasil prediksi, sebagaimana ditunjukkan pada gambar 4.

Gambar 3. Tampilan Interface



Gambar 4. Tampilan Interface Hasil Prediksi

4. KESIMPULAN

Berdasarkan hasil penelitian, Algoritma K-Nearest Neighbor (KNN) memiliki keunggulan kinerja dibandingkan algoritma Naïve Bayes (NB) dalam mendiagnosis diabetes.KNN mencapai akurasi tertinggi sebesar 76,92% pada nilai K=21, sementara NB hanya mencapai akurasi maksimum sebesar 65,63%. Dari segi evaluasi, KNN memiliki recall tertinggi sebesar 100%, menunjukkan kemampuan optimal dalam mendeteksi pasien positif, sedangkan NB unggul dalam precision sebesar 100%, mencerminkan akurasi tinggi dalam klasifikasi pasien negatif.

Variabel seperti kadar glukosa (FBS), BMI, dan riwayat keluarga terbukti signifikan dalam diagnosis diabetes. Penggunaan metode K-Fold Cross-Validation dengan 10 lipatan juga meningkatkan keandalan hasil analisis. Dengan demikian, algoritma KNN direkomendasikan sebagai pendekatan yang lebih efektif untuk diagnosis diabetes melitus dan dapat menjadi referensi untuk pengembangan algoritma klasifikasi lainnya di masa mendatang.

REFERENSI

[1] S. U. Putri, E. Irawan and R. Fitri, "Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4.5," *Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)*, p. 40, 2021.

[2] V. Wulandari, W. J. Sari, Z. Alfian, Legito and T. Arifianto, "Implementasi Algoritma Naive Bayes Classifier dan K-Nearest Neighbor untuk Klasifikasi Penyakit Ginjal Kronik," *Indonesia Journal of Machine Learning Computer Science*, p. 711, 2024.

[3] Sahar, "Analisis Perbandingan Metode K-Nearest Neighbor dan Naive Bayes Classifier pada Dataset Penyakit Jantung," *Indonesia Journal of Data amd Science*, p. 80, 2020.

[4] N. B. Sari and N. M. Purty, "Komparasi Algoritma KNN dan Naive Bayes Untuk Klasifikasi Diagnosis Penyakit Diabetes Melitus," *Jurnal Sains dan Manajemen*, pp. 45-57, 2022.

[5] A. K. E. Pily, Oktavianda, F. Aprilia, Rahmaddeni dan L. Efrizoni, "Komparasi Algoritma K-Nearest Neighborn dan Naive Bayes dalam Klasifikasi Penyakit Diabetes Gestasional," *Indonesia Journal of Computer Science*, p. 1196, 2024.

[6] S. K. Mandala, "https://www.kaggle.com/," 2022. [Online]. Available: https://www.kaggle.com/datasets/sujithmandala/easiest-diabetes-classification-dataset.

[7] A. Rianti, N. W. A. Majid and A. Fauzi, "CRIPS-DM: Metodologi Proyek Data Science," *SENARIB*, p. 108, 2023.