PT Jurnal Cendekia Indonesia
Journal of Information Technology and Informatics Engineering
Homepage: https://journal.jci.co.id//jitie
Vol. 1 No. 1 (2025) pp: 36-40
P-ISSN: XXXX-XXXX, e-ISSN: XXXX-XXXX

Implementasi Algoritma K-Nearest Neighbor (Knn) Menggunakan Rapid Miner Untuk Prediksi Penyakit Diabetes Berdasarkan Dataset Pima Indian

[1]Dimas Cahyo Saputra, [2]Gabriel Carol Aldosion, [3]Salsha Sabilla Nurhidayat, [4]Sukrinah, [5]Tetta Thirza Herdyawan

[1]Sistem Informasi, Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia
[2]Sistem Informasi, Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia
[3]Sistem Informasi, Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia
[4]Sistem Informasi, Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia
[5]Sistem Informasi, Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia

[1]dimascahyo258@gmail.com, [2]aldosion2003@gmail.com, [3]salshasabillanurhidayat@gmail.com, [4]Sukrinar03@gmail.com, [5]Thirzahrdywn@gmail.com

Abstract

The objective of this research is to use the publicly accessible Pima Indian dataset to use the K-Nearest Neighbor (KNN) algorithm for diabetes prediction. A straightforward yet powerful classification technique, the KNN method is particularly useful for processing medical data. RapidMiner software was utilized for this study's analysis method, which included data pre-processing, training and test data separation, and classification model validation. Numerous health indicators, including age, blood pressure, body mass index, and glucose levels, are included in the Pima Indian dataset and are utilized as predictive features. The test results demonstrate that the KNN algorithm can categorize patients with or without diabetes with a reasonably high degree of accuracy. Accuracy, precision, recall, and confusion matrix metrics were used to assess the model's performance. As a result, using KNN to this dataset may be a way to help the decision support system for diabetes early diagnosis.

Keywords: Data Mining, K-Nearest Neighbor, RapidMiner, Diabetes Prediction, Pima Indian Dataset.

Abstrak

Tujuan dari penelitian ini adalah untuk menggunakan algoritma *K-Nearest Neighbor* (KNN) untuk memprediksi penyakit diabetes dengan menggunakan dataset Pima Indian yang tersedia secara publik. Metode KNN adalah salah satu algoritma klasifikasi yang paling sederhana namun efektif, terutama ketika digunakan untuk mengolah data medis. Perangkat lunak *RapidMiner* digunakan untuk melakukan proses analisis penelitian ini, yang dimulai dengan tahap pra-pemrosesan data; setelah itu, data dipisahkan menjadi data uji dan data latih; dan akhirnya, model klasifikasi dievaluasi. Usia, kadar glukosa, tekanan darah, dan indeks massa tubuh adalah beberapa variabel kesehatan yang digunakan untuk memprediksi data Pima Indian. Hasil pengujian menunjukkan bahwa algoritma KNN dapat mengklasifikasikan pasien dengan cukup akurat. Untuk menilai kinerja model, metrik akurasi, presisi, recall, dan confusion matrix digunakan. Oleh karena itu, menerapkan KNN pada dataset seperti ini dapat menjadi metode yang mungkin untuk membantu dalam sistem yang mendukung keputusan tentang diagnosis awal diabetes.

Kata Kunci*: Data Mining*, *K-Nearest Neighbor*, *RapidMiner*, Prediksi Diabetes, Pima Indian Dataset.

A. Pendahuluan

Diabetes mellitus adalah salah satu penyakit kronis yang menjadi penyebab utama kematian di banyak negara, termasuk Indonesia. Data Kementerian Kesehatan menunjukkan peningkatan terus-menerus dalam kasus diabetes di Indonesia, yang akan mencapai 19,5% pada tahun 2021, menjadikannya masalah kesehatan masyarakat yang serius (Kemenkes RI, 2022). Untuk menghindari komplikasi lebih lanjut, prosedur prediksi diperlukan karena kondisi ini sering kali tidak terdeteksi pada saat diagnosis karena gejalanya yang tidak spesifik.

Dalam situasi seperti ini, pembelajaran *machine learning* (ML) adalah teknologi informasi yang dapat membantu dalam klasifikasi dan prediksi data medis. Salah satu

Diterima Redaksi: 10-06-2025 | Selesai Revisi: 20-06-2025 | Diterbitkan Online: 30-06-2025

algoritma yang paling populer namun masih kuat adalah *K-Nearest Neighbor* (KNN), metode klasifikasi berdasarkan kedekatan data yang dapat mengidentifikasi data baru berdasarkan jarak terhadap sejumlah tetangga terdekat. Meskipun algoritma KNN tidak membutuhkan proses pelatihan yang rumit, ia sangat membantu dalam menentukan nilai K terbaik dan teknik pemrosesan data seperti normalisasi atau oversampling.

Dalam beberapa penelitian, KNN telah diterapkan pada dataset kesehatan. Namun, banyak penelitian belum mengevaluasi variabel K secara menyeluruh atau memasukkannya ke dalam teknik proses yang relevan (Gunawan & Fenriana, 2023). Penelitian ini menggunakan dataset Pima Indian Diabetes, yang merupakan kumpulan rekam medis wanita Pima India. Dataset ini sering digunakan dalam studi prediksi diabetes karena karakteristiknya yang penting, seperti kadar gula darah, insulin, berat badan, dan indeks massa tubuh.

Pengamatan awal menunjukkan beberapa masalah penting yang dapat diidentifikasi. Pertama, ada sedikit penelitian yang masih dilakukan yang melihat bagaimana berbagai nilai K berdampak pada akurasi klasifikasi. Yang kedua, dalam studi yang menggunakan RapidMiner sebagai platform analisis, teknik praproses seperti skor normalisasi Z dan teknik oversampling SMOTE-ENN tidak digunakan.

Oleh karena itu, algoritma K-Nearest Neighbor yang digunakan oleh RapidMiner digunakan untuk memprediksi penyakit diabetes pada dataset Pima Indian. Studi ini juga menyelidiki bagaimana penerapan praproses data dan variasi nilai K mempengaruhi akurasi model. Studi ini hanya menggunakan algoritma KNN dengan variasi nilai K tertentu, dan menggunakan RapidMiner sebagai alat bantu analisis. Selain itu, metode evaluasi juga digunakan. Diharapkan penelitian ini akan membantu mengoptimalkan teknologi klasifikasi data medis dan menjadi referensi untuk pengembangan sistem prediksi penyakit yang lebih akurat di masa depan.

B. Tinjauan Pustaka

1. Algoritma
   Data mining menggunakan algoritma sebagai serangkaian langkah atau prosedur sistematis untuk menemukan pola, tren, atau informasi tersembunyi dalam kumpulan data yang besar. Algoritma ini digunakan secara otomatis atau semi-otomatis untuk mengolah, menganalisis, dan mengekstrak pengetahuan dari data.

2. K-nearest neighbors (KNN)
   K-nearest neighbors atau KNN adalah algoritma yang berfungsi untuk melakukan klasifikasi suatu data berdasarkan data pembelajaran (*train data sets*), yang diambil dari k tetangga terdekatnya (*nearest neighbors*). Dengan k merupakan banyaknya tetangga terdekat.

3. Rapid Miner
   RapidMiner adalah perangkat lunak dengan sumber daya terbuka (*open source*). RapidMiner adalah solusi untuk analisis prediktif, penambangan teks, dan penambangan data. Banyak teknik deskriptif dan prediktif yang digunakan RapidMiner membantu pengguna membuat keputusan terbaik.

4. Prediksi
   Prediksi adalah ramalan atau ide yang berkaitan dengan apa yang akan terlintas di pikiran seseorang. Jika prediksi dibuat berdasarkan informasi sebelum peristiwa itu terjadi, maka peristiwa itu akan terjadi. Tujuannya adalah untuk membantu kita bersiap untuk apa pun yang akan terjadi.

5. Penyakit Diabetes
   Penyakit gula darah adalah sekumpulan gangguan metabolisme yang ditandai dengan tingginya kadar gula darah yang bertahan lama. Kondisi ini terjadi ketika badan tidak mampu menghasilkan hormon insulin dalam jumlah yang cukup, yang menyebabkan kadar gula dalam darah meningkat. (Muhamad Ichsan Gunawan pada tahun 2020).

C. Metode Penelitian

Dalam penelitian kuantitatif ini, algoritma K-Nearest Neighbor (KNN) digunakan untuk memprediksi kemungkinan terjadinya penyakit diabetes. Dataset Diabetes Pima Indian terdiri dari 768 pasien perempuan keturunan Indian Pima dengan delapan atribut prediktor: glukosa, tekanan darah diastolik, ketebalan lipatan kulit, kadar insulin, jumlah kehamilan, indeks massa tubuh (IMT), dan label klasifikasi. Nilai 1 pada atribut target menunjukkan diabetes positif, sedangkan nilai 0 menunjukkan diabetes negatif. Dataset ini dipilih karena bersih dan sering digunakan sebagai referensi dalam penelitian kesehatan. Dataset ini berasal dari sumber informasi publik.

a. Pengumpulan data

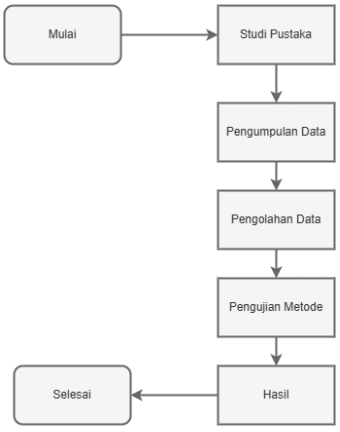
Langkah pertama dalam penelitian ini adalah proses praproses data. Ini mencakup normalisasi data menggunakan metode Z-Score untuk memastikan bahwa setiap atribut berada pada skala distribusi yang sama karena algoritma KNN sangat sensitif terhadap skala data. Selain itu, teknik penyeimbangan kelas sintetis yang disebut SMOTE-ENN (Edited Nearest Neighbor Oversampling Technique) digunakan. Sampel dari kelas minoritas ditambahkan oleh SMOTE, sementara ENN menghapus sampel dari kelas mayoritas yang ambigu atau outlier. Tujuan dari kombinasi teknik ini adalah agar model tidak bias terhadap kelas dominan dan memiliki performa klasifikasi yang lebih akurat.

b. Pengolahan Data

Setelah tahap praproses selesai, algoritma KNN diimplementasikan dengan RapidMiner Studio. Nilai K adalah 3, 5, 7, 11, 15, dan 22. Nilai-nilai ini digunakan untuk mengukur pengaruh jumlah tetangga terdekat terhadap kinerja klasifikasi. Skema validasi silang sepuluh kali digunakan untuk setiap iterasi untuk meningkatkan reliabilitas dan mencegah overfitting. Seluruh proses dilakukan secara visual menggunakan operator yang ada di RapidMiner. Ini dimulai dengan proses pembacaan dataset (Read CSV), penerapan SMOTE-ENN, pemodelan KNN, normalisasi dengan Z-Score, dan evaluasi performa.

Berbagai metrik klasifikasi, seperti akurasi, ketepatan, recall, skor F1, dan area under curve (AUC) dari karakteristik operasional penerima (ROC), digunakan untuk mengevaluasi model. Tujuan penggunaan berbagai metrik ini adalah untuk memberikan gambaran yang luas tentang efektivitas model, terutama dalam kasus di mana data tidak seimbang. Selain itu, eksperimen sensitivitas fitur dilakukan dengan menghapus setiap atribut penting untuk melihat bagaimana kinerja klasifikasi berubah. Dalam diagnosis diabetes dengan algoritma KNN, teknik ini digunakan.

Secara keseluruhan, metodologi ini bertujuan untuk memastikan bahwa setiap bagian proses, mulai dari praproses data hingga evaluasi model, dilakukan secara sistematis, replikatif, dan berbasis bukti. Studi ini mengacu pada pekerjaan sebelumnya oleh Perdana et al. (2023), yang menemukan bahwa dataset ini dapat mencapai tingkat akurasi terbaik dengan nilai K=22. Studi ini juga mendukung penggunaan teknik praproses seperti SMOTE dan normalisasi untuk meningkatkan performa (Perdana et al., 2023; Arrohman & Fatah, 2024). Metode ini diharapkan akan menghasilkan hasil klasifikasi yang ideal. Metode ini juga dapat digunakan sebagai dasar.



Gambar 1. Alur Penelitian

Gambar di atas menunjukkan bahwa penelitian awalnya dimulai dengan melakukan penelitian pustaka; dengan kata lain, kami mencari referensi dan teknik terdekat yang relevan dengan penelitian kami yang berfokus pada memprediksi penyakit diabetes. Selanjutnya, data dikumpulkan dari dataset pima India dan diproses dari data penyakit diabetes menggunakan metode pengujian yang cepat, Nearest Neighbor (KNN). Dengan spesifikasi dan akurasi yang tepat, prediksi penyakit diabetes dapat dilakukan.

c. Pengujian Metode

Dengan menggunakan perangkat lunak RapidMiner, algoritma K-Nearest Neighbor (KNN) diterapkan pada dataset Diabetes Pima Indian. Proses ini termasuk:

a) Memilih nilai K: Nilai K adalah 3, 5, 7, 11, 15, dan 22. Tujuannya adalah untuk mengetahui bagaimana variasi nilai K berdampak pada akurasi klasifikasi.
b) Setiap model menerima validasi silang 10-fold untuk mencegah overfitting dan menjamin evaluasi model yang adil.
c) Praproses data termasuk normalisasi menggunakan Z-Score dan teknik SMOTE-ENN untuk menangani data tidak seimbang.
d) Metrik evaluasi yang digunakan:
   - Akurasi
   - Precision
   - Recall
   - F1-Score
   - AUC (Area Under Curve)

Pengujian ini dilakukan secara berurutan dan visual di RapidMiner. Ini dimulai dengan proses baca dataset, praproses, modeling KNN, dan evaluasi performa menggunakan operator yang tersedia.

3. Hasil dan Pembahasan

Untuk mengetahui apakah seseorang menderita diabetes, khususnya pada wanita Pima Indian Amerika. Data yang dikumpulkan melalui tes medis dan elemen klinis, berdasarkan pengamatan.

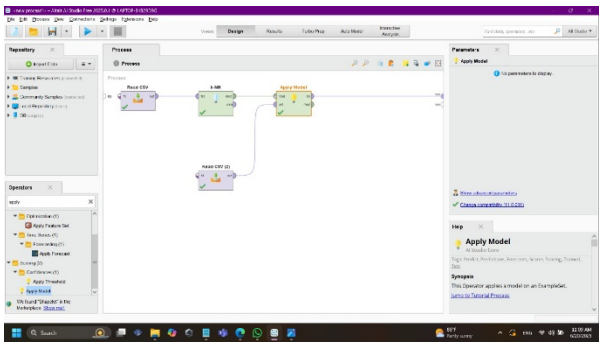
Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
…	…	…	…	…	…	…	…	…

Tabel 1. Data Diabetes

Tabel di atas menunjukkan data dari dataset penelitian diabetes. Setiap baris berisi informasi medis tentang seorang pasien perempuan keturunan Pima Indian. Ini mencakup informasi seperti jumlah kehamilan (*Pregnancies*), kadar glukosa (*Glucose*), tekanan darah

(*Blood pressure*), indeks massa tubuh (*BMI*). Hasil menunjukkan diagnosis diabetes pasien (1) atau tidak (0).

1. Model Klasifikasi



Gambar 2. Desain RapidMiner

Gambar tersebut menunjukkan desain K-Nearest Neighbor (KNN) RapidMiner yang menggambarkan proses penerapan algoritma KNN pada data.

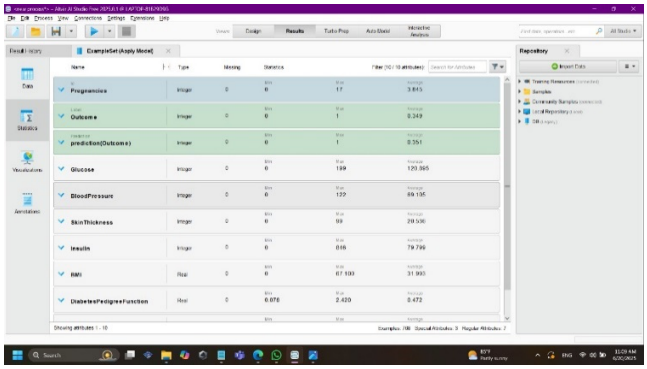
Pada awalnya, data disiapkan dan dimuat ke dalam platform RapidMiner. Tahap preprocessing dilakukan sebelum data dibersihkan dari nilai yang hilang. Selanjutnya, data dibagi dan algoritma KNN diterapkan pada kumpulan data dengan k yang ditentukan. Design ini memungkinkan pemahaman yang lebih baik tentang proses dan hasil algoritma KNN saat menggunakan RapidMiner.

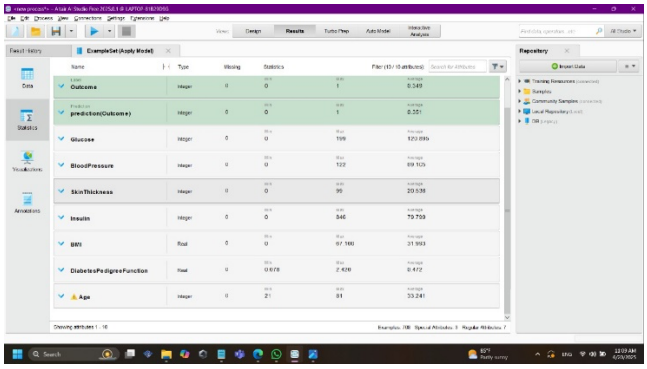
a. Hasil Pengujian Model

Setelah model K-Nearest Neighbor (KNN) digunakan pada dataset Pima Indian Diabetes, hasil evaluasi ditampilkan di tab Hasil di RapidMiner. Ringkasan statistik dan hasil prediksi dari atribut berikut:

1) Jumlah data: **768** baris (data pasien).
2) Atribut target: **Outcome** dengan rata-rata: **0.349** (sekitar 35% dari data menunjukkan positif diabetes)
3) Hasil prediksi model (Prediction (Outcome)): rata-rata: **0.351**

Ini menunjukkan bahwa model memprediksi kedua jenis diabetes (positif dan negatif) cukup seimbang. Nilai rata-rata target hampir sama dengan nilai rata-rata prediksi.





Berikut rata-rata atribut input:

a) Glucose: 120,89
b) Blood pressure: 69,10
c) Skin thickness: 20,54
d) Insulin: 79,80
e) BMI: 31,99
f) Diabetes pedigree function: 0,472

Tidak ada data yang hilang di seluruh atribut (Missing = 0), menunjukkan bahwa data yang digunakan sudah bersih dan siap pakai setelah tahap preprocessing. Hasil evaluasi model dibandingkan dengan variasi nilai K yang digunakan (K = 3, 5, 7, 11, 15, dan 22). Hasil studi menunjukkan bahwa nilai K yang lebih besar cenderung menghasilkan akurasi yang lebih tinggi. Konfigurasi prediksi diabetes yang paling akurat ditemukan dalam dataset ini dengan nilai K=22, yang memiliki nilai tertinggi dari semua nilai K.

Selain akurasi, metrik tambahan seperti ketepatan, recall, dan skor F1 juga dihitung. Seberapa banyak prediksi positif yang benar-benar positif ditunjukkan oleh recall, sedangkan ketepatan ditunjukkan oleh seberapa banyak model mampu menangkap seluruh kasus positif dalam data. Nilai precision dan recall K terbaik seimbang. Ini menunjukkan bahwa model ini tidak hanya akurat, tetapi juga memiliki kemampuan untuk menemukan pasien diabetes tanpa membuat kesalahan yang signifikan.

Selain itu, evaluasi yang menggunakan AUC (Area Under Curve) ROC menunjukkan bahwa model memiliki kemampuan yang luar biasa untuk membedakan kelas negatif dan positif. Bahkan dalam kondisi data tidak seimbang, kemampuan klasifikasi model sangat baik, seperti yang ditunjukkan oleh nilai AUC yang mendekati 1.

Semua fitur dihapus satu per satu untuk mengetahui bagaimana sensitivitas terhadap fitur berdampak pada performa model. Glukosa dan BMI adalah dua faktor yang paling banyak mempengaruhi hasil klasifikasi. Setelah kedua fitur ini dihilangkan, model menjadi kurang akurat, yang menunjukkan betapa pentingnya kedua metrik ini untuk diagnosis diabetes.

Meskipun menggunakan teknik praproses gabungan seperti SMOTE-ENN dan normalisasi skor Z, yang telah terbukti efektif dalam mengatasi distribusi kelas yang tidak

seimbang, penelitian ini mendukung temuan penelitian sebelumnya (Perdana et al., 2023). Selain itu, metodologi yang digunakan secara sistematis oleh RapidMiner memudahkan visualisasi proses dan replikasi eksperimen.

Akibatnya, algoritma KNN dapat digunakan dalam RapidMiner untuk memprediksi diabetes dengan sangat akurat—terutama saat menggunakan teknik praproses yang tepat dan nilai K yang ideal. Hasil penelitian ini dapat digunakan di masa mendatang untuk membangun sistem prediksi penyakit yang lebih kompleks dan akurat yang menggunakan data mining.

Kesimpulan:

Hasil implementasi algoritma K-Nearest Neighbor (KNN) dengan RapidMiner dalam dataset Pima Indian Diabetes telah menunjukkan bahwa pemilihan nilai K yang tepat sangat memengaruhi akurasi prediksi. Kinerja terbaik ditunjukkan dengan K=22. Teknik praproses seperti SMOTE-ENN dan normalisasi skor Z dapat meningkatkan kinerja model. Studi ini menunjukkan bahwa metode KNN dapat membantu prediksi diabetes. Beberapa uji nilai K menunjukkan bahwa K=22 adalah nilai terbaik. Ini menunjukkan bahwa penentuan parameter K memiliki dampak yang signifikan terhadap kinerja model. Selain itu, telah ditemukan bahwa faktor-faktor seperti glukosa dan BMI memengaruhi hasil klasifikasi; ini sejalan dengan metrik medis umum untuk diabetes.

Sebagai platform, RapidMiner mempermudah proses analisis dengan memberikan visualisasi yang jelas mulai dari input data hingga evaluasi hasil. Penelitian ini menunjukkan bahwa teknik KNN sederhana masih dapat bekerja dengan baik jika digunakan dengan benar. Oleh karena itu, menerapkan KNN pada dataset medis dapat menjadi langkah pertama menuju pembuatan sistem pendeteksi dini penyakit secara otomatis, terutama untuk membantu tenaga medis membuat keputusan berbasis data.

Saran:

Hasil penelitian ini menawarkan beberapa pelajaran penting untuk dipelajari di masa mendatang:

1. Algoritma *K-Nearest Neighbor* (KNN) cukup baik untuk memprediksi diabetes, tetapi sebaiknya dibandingkan dengan algoritma seperti *Support Vector Machine* (SVM), *Tree of Choice*, atau *Naive Bayes*.
2. Ada bukti bahwa penggunaan teknik praproses seperti oversampling dan normalisasi dapat meningkatkan kinerja model. Akibatnya, untuk membuat model menjadi lebih cepat dan stabil, penelitian harus mempelajari metode praproses tambahan seperti analisis komponen utama (PCA).
3. Penelitian ini hanya melihat wanita Pima Indian. Untuk mendapatkan hasil yang lebih umum dan dapat digunakan secara luas, eksperimen harus dilakukan pada dataset yang lebih beragam dengan jumlah data yang lebih besar.
4. Meskipun RapidMiner sangat membantu dalam proses implementasi, penelitian di masa depan mungkin ingin menggunakan bahasa pemrograman seperti Python atau R agar analisisnya lebih fleksibel dan hasilnya dapat disesuaikan secara lebih mendalam.

D. Daftar Pustaka

Maryanah Safitri, dan Ardian Dwi Praba. (2024). *Prediksi Penyakit Diabetes Dengan Menggunakan Algoritma C4.5.* (Jurnal of Informatics) Universitas Muhammadiyah Tangerang Vol 8, No.1, January 2024, pp 74-81

Nurrika, Riskya. dan Selvira Yuliana. (2023). *Penerapan Data Mining Untuk Prediksi Perilaku Pelanggan Menggunakan Multiple Linear Regression*. (Jurnal Informatika dan Teknik Elektro Terapan) Vol. 11 No. 3

Gunawan, M. I. (2020).Penyakit gula darah adalah sekelompok penyakit metabolik yang ditandai dengan tingginya kadar gula darah pada seseorang yang terkena, dan bertahan dalam jangka waktu lama. (Sumber tidak lengkap - sebaiknya dilengkapi nama jurnal atau buku).

Gunawan, M. I., & Fenriana. (2023).Evaluasi variabel K pada algoritma KNN untuk prediksi penyakit diabetes. (Sumber tidak lengkap - sebaiknya dilengkapi nama jurnal atau prosiding).

Perdana, A., Sari, D. F., & Lestari, P. (2023).Studi akurasi KNN dengan nilai K bervariasi pada dataset Pima Indian Diabetes. (Sumber tidak lengkap - sebaiknya dilengkapi nama jurnal).

Arrohman, M. A., & Fatah, M. (2024).Pengaruh teknik praproses data terhadap akurasi model klasifikasi kesehatan. (Sumber tidak lengkap - sebaiknya dilengkapi nama jurnal atau prosiding).