Gudang Jurnal Multidisiplin Ilmu
Volume 2 ; Nomor 10 ; Oktober 2024 ; Page 220-226
Doi : https://doi.org/10.59435/gjmi.v2i10.986
Website : https://gudangjurnal.com/index.php/gjmi
E-ISSN : 2988-5760

Prediksi Diabetes Menggunakan Algoritma Klasifikasi K-Nearest Neighbors (K-NN) pada Perempuan Indian Pima

Supri Arrohman[1*], Zaehol Fatah[2]

[1] Teknologi Informasi, Universitas Ibrahimy
[2]Sistem Informasi, Universitas Ibrahimy
[1*]supriarrohman23@gmail.com, [2]zaeholfatah@gmail.com

Abstrak

Diabetes merupakan penyakit kronis yang ditandai dengan tingginya kadar glukosa dalam darah, yang dapat menyebabkan komplikasi serius jika tidak dikelola dengan baik. Dalam penelitian ini, dilakukan prediksi diabetes pada perempuan suku Indian Pima menggunakan algoritma klasifikasi K-Nearest Neighbors (K-NN). Dataset yang digunakan diperoleh dari National Institute of Diabetes and Digestive and Kidney Diseases, yang terdiri dari 769 data pasien dengan 9 atribut. Penelitian ini menggunakan metode deskriptif kuantitatif dengan pendekatan rekayasa perangkat lunak dan Library Online Research. Data dianalisis menggunakan RapidMiner versi 10.3. Hasil dari model K-NN menunjukkan tingkat akurasi sebesar 70,13%, dengan presisi 59,09% untuk pasien positif dan 74,55% untuk pasien negatif. Nilai recall yang diperoleh adalah 48,15% untuk pasien positif dan 82,00% untuk pasien negatif. Meskipun akurasi model ini cukup baik, penelitian lebih lanjut diperlukan untuk meningkatkan kualitas prediksi. Algoritma K-NN terbukti efektif digunakan dalam klasifikasi diabetes, tetapi kualitas hasil sangat bergantung pada kualitas data yang dianalisis.

Kata Kunci: Diabetes, Klasifikasi, K-Nearest Neighbors, Data Mining, Perempuan Indian Pima

PENDAHULUAN

Diabetes adalah penyakit kronis yang ditandai dengan tingginya kadar gula (glukosa) dalam darah. Glukosa merupakan sumber energi utama bagi tubuh, yang berasal dari makanan yang kita konsumsi. Namun, untuk dapat digunakan oleh sel-sel tubuh, glukosa membutuhkan hormon insulin yang diproduksi oleh pankreas. Pada penderita diabetes, terjadi gangguan pada produksi atau fungsi insulin sehingga glukosa tidak dapat diserap dengan baik oleh sel, dan menyebabkan penumpukan dalam darah. Salah satu aspek yang paling mengkhawatirkan dari diabetes adalah komplikasi jangka panjang yang dapat timbul akibat pengendalian kadar gula darah yang tidak memadai (Viyan Qomarudin Noor et al., 2022).

Diabetes merupakan penyakit dengan resiko kematian yang tinggi. Dalamtahun 2019 WHO mencatat setidaknya 2 juta kematian akibat diabetes. Faktor utama pemicu diabetes adalah banyaknya kadar glukosa dalam darah sehingga tubuh tidak dapat mengontrol kadar glukosa dalam darah. Diabetes yang berbahaya bukan merupakan penyakit genetic yang dapat diturunkan orang tua ke anak keturunannya. Factor utama munculnya diabetes adalah pola makan yang tidak sehat (Ivandari, Much. Rifqi Maulana, Muhammad Faizal Kurniawan, & Al Karomi, 2023).

Data Mining adalah proses penggalian informasi dan pola yang bermanfaat dari suatu data yang sangat besar. Proses data mining terdiri dari pengumpulan data, ekstraksi data, analisa data, dan statistik data. Ia juga umum dikenal sebagai knowledge discovery, knowledge extraction, data/pattern analysis, information harvesting, dan lainnya (Iswadi Hamzah et al., 2024).

Dari sekian banyak penyakit, salah satu penyakit degeneratife yang dapat diprediksi dengan menggunakan metode data mining adalah penyakit diabetes. Penyakit diabetes atau yang sering dikenal dengan sebutan kencing manis ini merupakan penyakit di mana kadar glukosa (gula sederhana) di dalam darah menjadi tinggi karena tubuh tidak dapat memproduksi atau mengeluarkan insulin secara cukup. Penyakit diabetes dapat disebabkan oleh penderita yang memiliki riwayat penyakit diabetes turunan yang kita sebut Tipe Adan karena factor pemicu lainnya yang tidak berkaitan dengan factor turunan yang kita sebut Tipe B (Faizal Aris et al., 2019).

Banyak sekali algoritma yang dapat digunakan dalam proses data mining, algoritma tersebut juga turut membantu untuk menganalisis serta mengekstraksi data menjadi pengetahuan baru yang berguna. diantara banyaknya algoritma yang sering digunakan dan populer digunakan oleh peneliti di bidang ilmu data mining adalah algoritma K-Nearest Neighbors. KNN adalah salah satu metode klasifikasi pada data mining yang terbaik dan banyak digunakan. Algoritma KNN merupakan teknik klasifikasi yang sering digunakan, yang dikenalkan oleh Fix dan Hodges pada tahun 1951, dan telah diakui sebagai algoritma sederhana yang terbaik. KNN adalah salah satu metode yang dipakai dalam pengelompokan data yang menggunakan algoritma supervised (Rinanda, Delvika, Nurhidayarnis, Abror, & Hidayat, 2022). algoritma ini terbilang sederhana serta bisa dibilang efektif untuk data yang ukurannya besar, dan digunakan untuk mengklasifikasi data dengan menentukan kedekatan objek yang baru atau tetangga terdekat. tetangga terdekat

biasa disimbolkan dengan (K) yang berpengaruh terhadap proses pengambilan keputusan dari metode KNN (Yogianto, Homaidi, & Fatah, 2024). Metode ini merupakan bagian dari algoritma supervised learning, yang berarti dataset diharuskan memiliki target, pada penelitian kali ini penentuan nilai K diukur berdasarkan perhitungan dengan rumus Euclidean Distance [4]. Penelitian ini bertujuan untuk mengimplementasikan metode K-NN dalam Klasifikasi penyakit diabetes yang berpotensi terjadinya komplikasi dan mengetahui tingkat akurasi dari metode tersebut.

METODE

Jenis Penelitian

Jenis penelitian ini adalah deskriptif kuantitatif dengan pendekatan rekayasa software serta *Library Online Research*. Pengumpulan data dilakukan dengan teknik kepustakaan. Penelitian deskriptif kuantitatif dipilih agar dapat menggambarkan, mendeskripsikan serta menjelaskan sesuatu secara objektif serta dapat menarik kesimpulan dari data berupa angka-angka yang telah disajikan, dan dengan cara ini peneliti menghimpun, mengelola data serta menganalisis data-data yang terkait dengan klasifikasi komplikasi penyakit diabetes dengan metode K-nearest Neighbors (Zulfikar, Podungge, Saleh, & ..., 2022).

Metode Pengumpulan Data

Pada bagian ini merupakan suatu tata cara yang dilakukan untuk dapat mengumpulkan data terkait dengan penelitian yang sedang dilakukan, metode pengumpulan data ini merupakan suatu cara independen untuk melakukan analisis data atau bisa dijadikan alat utama dalam menganalisis data (Ocal, Gokcek, Colak, & Korkanc, 2021). Untuk mendapatkan data-data yang berkaitan dengan penyakit diabetes pada perempuan khususnya suku Indian Pima, penulis mengumpulkan data yang terkait dengan penelitian ini menggunakan cara Study Literature dengan jenis data sekunder yang didapatkan dari Website Kaggle. Kumpulan dataset ini berasal dari *the National Institute of Diabetes and Digestive and Kidney Diseases*.

Tautan yang dapat diakses adalah sebagai berikut https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. Data terdiri dari 769 dataset dengan format csv. Data ini dihimpum dengan 9 attribut atau fitur berkenan dengan prediksi secara diagnostik pasien yang memiliki penyakit diabetes atau tidak. Berikut adalah dataset beserta penjelasan dari attribut-attributnya disajikan pada gambar berikut:

Tabel 1. Dataset

No	Pregnancies	Glucose	BloodPress	ST	Ins	BMI	DP	Age	OutCome
1	6	148	72	35	0	336	627	50	1
2	1	85	66	29	0	266	351	31	0
3	8	183	64	0	0	233	672	32	1
4	1	89	66	23	94	281	167	21	0
5	0	137	40	35	168	431	2288	33	1
6	5	116	74	0	0	256	201	30	0
7	3	78	50	32	88	31	248	26	1
8	10	115	0	0	0	353	134	29	0
9	2	197	70	45	543	305	158	53	1
10	8	125	96	0	0	0	232	54	1
11	4	110	92	0	0	376	191	30	0
12	10	168	74	0	0	38	537	34	1
13	10	139	80	0	0	271	1441	57	0
14	1	189	60	23	846	301	398	59	1
15	5	166	72	19	175	258	587	51	1
756	8	154	78	32	0	324	443	45	1
757	1	128	88	39	110	365	1057	37	1
758	7	137	90	41	0	32	391	39	0
759	0	123	72	0	0	363	258	52	1
760	1	106	76	0	0	375	197	26	0
761	6	190	92	0	0	355	278	66	1
762	2	88	58	26	16	284	766	22	0
763	9	170	74	31	0	44	403	43	1
764	9	89	62	0	0	225	142	33	0
765	10	101	76	48	180	329	171	63	0

766	2	122	70	27	0	368	34	27	0
767	5	121	72	23	112	262	245	30	0
768	1	126	60	0	0	301	349	47	1
769	1	93	70	31	0	304	315	23	0

Keterangan:

a. Pregnancies : Jumlah kehamilan yang dialami.
b. Glucose : Kadar glukosa dalam darah.
c. BloodPress : Tekanan darah dalam arteri.
d. ST : Pengukuran ketebalan lipatan kulit, sering digunakan untuk estimasi lemak tubuh.
e. Ins : Jumlah hormon insulin dalam darah yang mengatur gula darah.
f. BMI : Rasio berat badan terhadap tinggi badan, digunakan untuk mengukur obesitas.
g. DP : Indikator risiko diabetes berdasarkan riwayat kesehatan keluarga.
h. Age : Umur individu.
i. OutCome : Apakah individu menderita diabetes atau tidak.

Data Mining

Data mining adalah proses ekstraksi atau penggalian data serta informasi dari database berukuran besar yang belum diketahui sebelumnya, namun dapat dipahami dan berguna. Proses ini digunakan untuk mendukung pengambilan keputusan bisnis yang penting. Data mining mengacu pada serangkaian teknik yang bertujuan menemukan pola tersembunyi dalam data yang telah dikumpulkan. Teknologi ini memungkinkan pengguna menemukan pengetahuan yang sebelumnya tidak terdeteksi dalam database. Secara otomatis, data mining mencari informasi berharga dari tempat penyimpanan data yang sangat besar (Rayuwati, Husna Gemasih, & Irma Nizar, 2022).

Klasifikasi

Klasifikasi adalah proses menilai objek data untuk memasukkannya ke dalam salah satu kelas yang tersedia. Klasifikasi membangun model berdasarkan data latih, yang kemudian digunakan untuk mengklasifikasikan data baru. Klasifikasi dapat didefinisikan sebagai proses pembelajaran atau pelatihan terhadap fungsi target yang memetakan setiap kumpulan atribut (fitur) ke salah satu label kelas yang ada. Sistem klasifikasi diharapkan dapat mengklasifikasikan semua data dengan akurat, namun, kinerjanya tidak selalu 100% tepat. Oleh karena itu, performa sistem klasifikasi perlu diukur. Biasanya, pengukuran kinerja klasifikasi dilakukan menggunakan matriks konfusi (Utomo & Mesran, 2020).

K-Nearest Neighbors (K-NN)

Algoritma klasifikasi K-NN memprediksi kategori sampel uji berdasarkan nilai k sampel latih yang merupakan tetangga terdekat dari sampel uji, kemudian mengklasifikasikannya ke dalam kategori dengan probabilitas tertinggi (Andrian, Naufal, Hermanto, Junaidi, & Lumbanraja, 2019). Kedekatan atau jarak antara titik-titik tetangga dihitung menggunakan jarak Euclidean, yang dinyatakan sebagai berikut:

$$D(a,b) = \sqrt{\sum_{k=1}^{d} (a_k - b_k)^2} \quad (1)$$

Keterangan:

D = Jarak antara titik
a = Titik yang diketahui
b = Titik yang tidak diketahui
d = Dimensi titik yang diukur
k = Nilai data tetangga yang diukur

Suku Indian Pima

Benua Amerika yang sangat luas awalnya dihuni oleh suku Indian, yang merupakan penduduk asli di wilayah tersebut. Suku Indian ini berasal dari masyarakat urban (pendatang) yang berasal dari Asia, khususnya dari rumpun Mongoloid. Migrasi rumpun Mongoloid dari Asia ke Amerika terjadi sekitar 20.000 hingga 25.000 tahun yang lalu, melalui rute barat laut Siberia, melintasi Selat Bering menuju Alaska, dan kemudian bergerak ke selatan. Perpindahan ini berlangsung dalam beberapa gelombang kelompok, yang kemudian menyebar ke berbagai wilayah di Amerika.

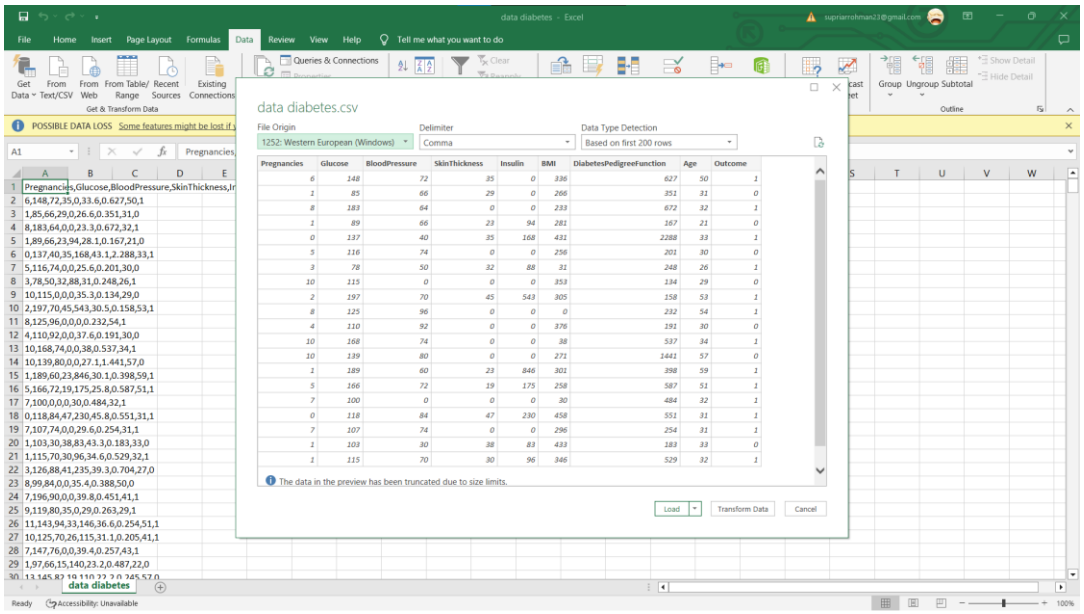
Penyebaran ini menyebabkan mereka menyesuaikan perilaku dengan lingkungan alam masing-masing, sehingga menghasilkan keragaman budaya yang berbeda-beda (Pamungkas, 2019).

HASIL DAN PEMBAHASAN

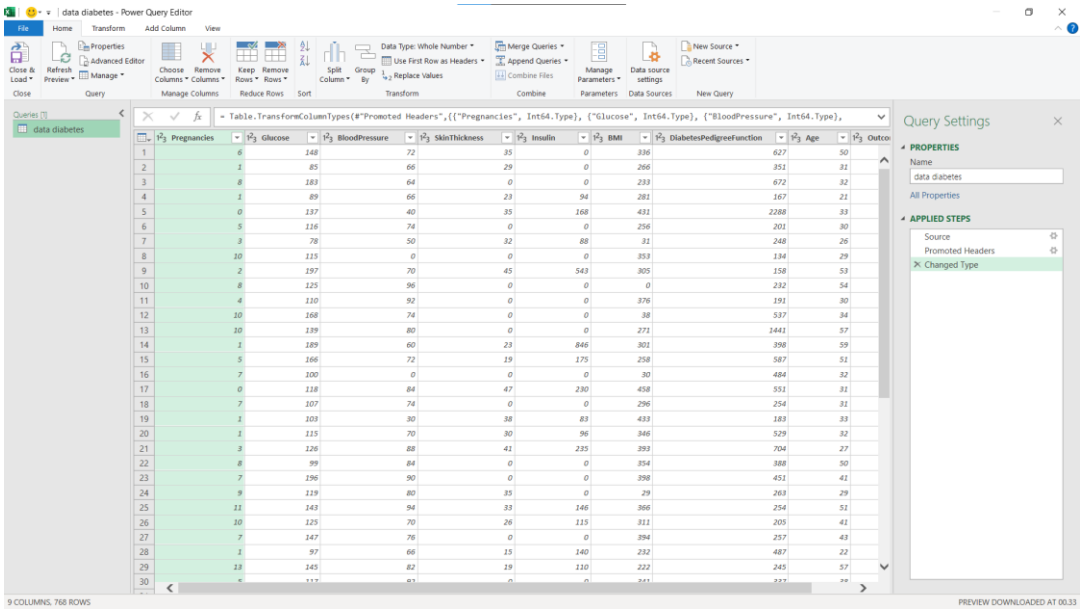
Penelitian kali ini dilakukan menggunakan salah satu model klasifikasi yaitu K-Nearest Neighbors (K-NN). Meachine Learning yang digunakan untuk mengimplementasikan model tersebut adalah Rapidminer versi 10.3. Dataset penyakit diabetes yang didapatkan pada penelitian kali ini bersumber dari website kaggle.com dengan judul "Diabetes Dataset". Dataset ini diperoleh dari *The National Institute of Diabetes and Digestive and Kidney Diseases.* Data ini dihimpun dengan 9 attribut dengan klasifikasi pasien yang memiliki penyakit diabetes atau tidak yang dapat digunakan sebagai parameter. Proses dengan menggunakan visualisasi Rapidminer dilakukan dengan langkah-langkah sebagai berikut:

Transformasi Data

Transformasi adalah serangkaian instruksi untuk mengubah input menjadi output yang diharapkan (input-proses-output). Proses ini melibatkan perubahan data ke dalam bentuk yang sesuai, disesuaikan dengan algoritma klasifikasi yang akan diterapkan (Widaningsih, 2022). Untuk itu dilakukan normalisasi data dengan mengubah data excel menjadi data csv. Transformasi data dilakukan dengan mengimport data excel menjadi data csv menggunakan tools yang ada pada microsoft excel. Berikut adalah normalisasi data yang dilakukan:

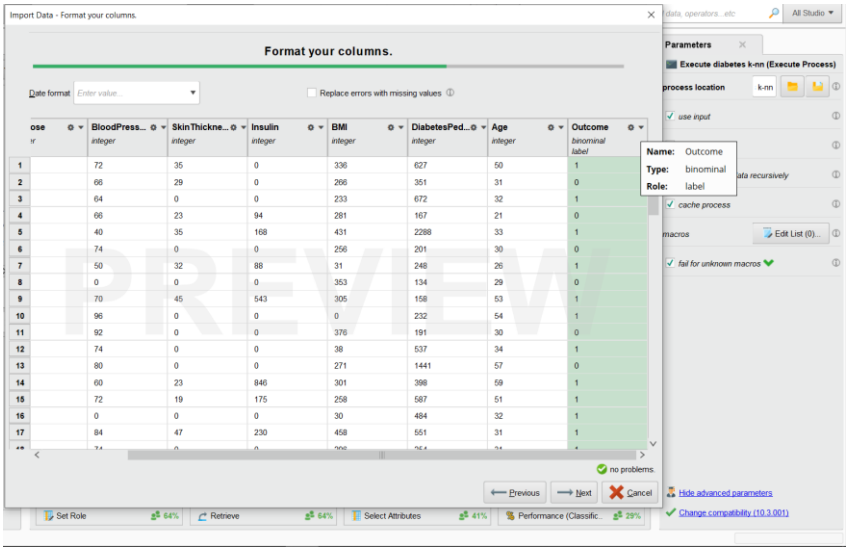


Gambar 1. Transformasi Data



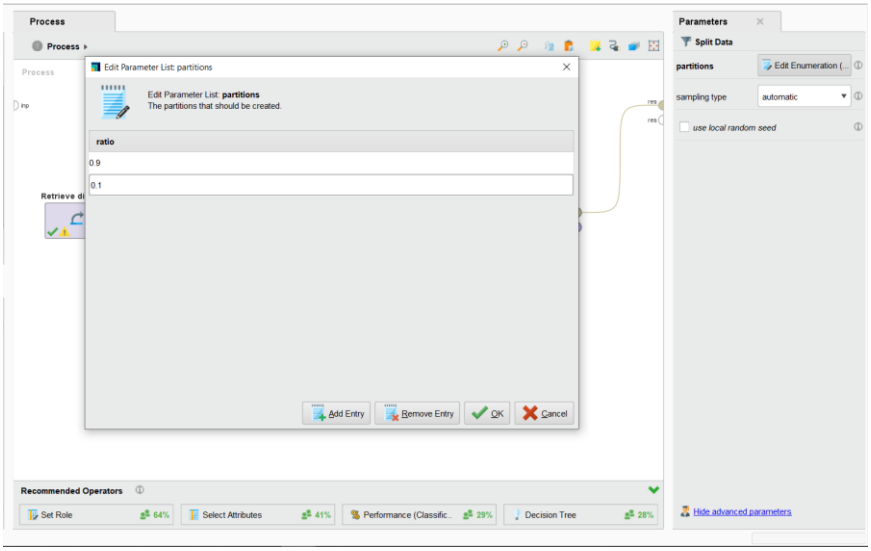
Gambar 2. Hasil Transformasi Data

Model Data Mining



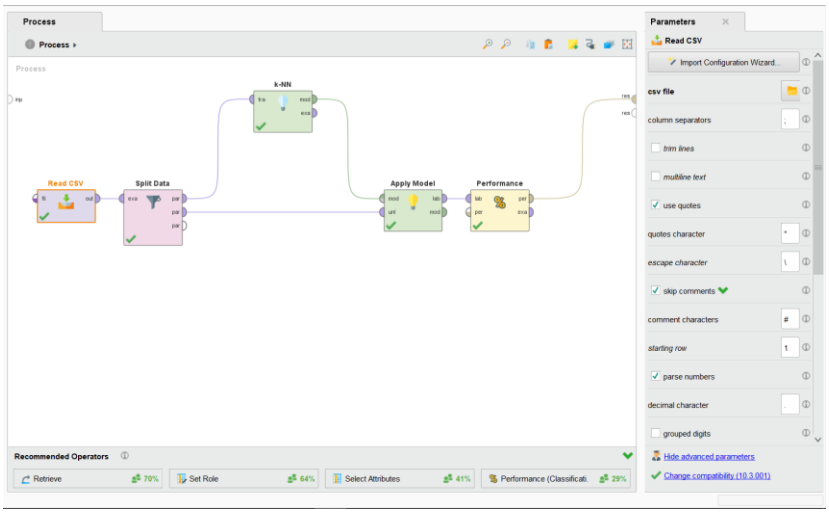
Gambar 3. Operator Pemanggil Data

Operator Read csv digunakan untuk membaca dataset yang disimpan dengan format csv. Pada bagian ini, ditentukan juga attribut yang akan dijadikan sebagai label pada dataset yang akan digunakan.



Gambar 4. Split Data

Split data pada K-Nearest Neighbors (K-NN) adalah proses memisahkan dataset menjadi dua bagian: data latih (training data) dan data uji (testing data). Ini adalah langkah penting dalam membangun model meachine learning untuk menghindari overfitting dan mengukur kinerja model pada data yang belum pernah dilihat sebelumnya. Operator Split Data digunakan untuk memisahkan data training (0.9) dan data testing (0.1). Split data membantu memastikan bahwa model K-NN bekerja dengan baik dan dapat menggeneralisasi dengan baik pada data baru, bukan hanya pada data latih.

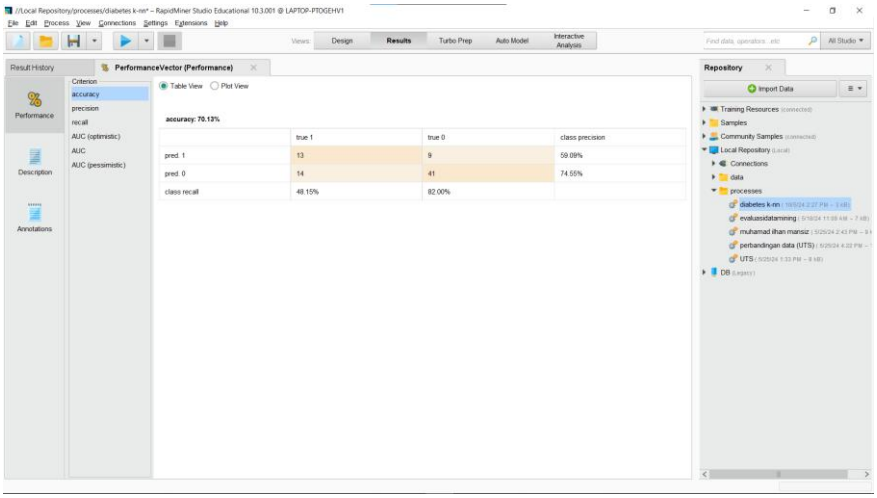


Gambar 5. Model Metode K-Nearest Neighbors

Implementasi data mining menggunakan algoritma K-Nearest Neighbors (K-NN) dengan aplikasi Rapidminer versi 10.3 seperti pada Gambar 5. Tahap berikutnya adalah menyambungkan semua operator yang telah disiapkan. Operator Apply Model dalam RapidMiner adalah salah satu operator penting yang digunakan untuk menerapkan model yang sudah dilatih ke dataset baru. Ini memungkinkan prediksi atau klasifikasi dilakukan pada data yang tidak dikenal berdasarkan model yang sebelumnya dibuat menggunakan operator training. Setelah melatih model K-NN dengan data latih menggunakan operator K-NN, Anda dapat menggunakan operator Apply Model untuk menerapkan model tersebut pada data uji dan mendapatkan prediksi diabetes pada dataset perempuan Indian Pima.

Operator Performance digunakan untuk mengukur kinerja model yang telah dilatih dan diterapkan pada data ini. Ini adalah alat yang sangat penting untuk menilai seberapa baik model tersebut bekerja dalam memprediksi atau mengklasifikasi data. Operator ini memberikan berbagai metrik evaluasi kerja. Dalam penelitian kali ini, operator performance dapat digunakan untuk mengevaluasi akurasi, presisi dan recall dari model K-NN setelah diterapkan ke data uji.

Hasil Akurasi



Gambar 6. Hasil Akurasi

Berdasarkan gambar di atas, penerapan metode K-NN dengan nilai k = 5 menghasilkan nilai akurasi sebesar 70,13%. Dari hasil prediksi, terdapat 13 pasien yang diprediksi positif terkena diabetes dan benar-benar positif, serta 9 pasien yang diprediksi positif tetapi sebenarnya negatif. Disisi lain, terdapat 14 pasien yang diprediksi negatif tetapi sebernya positif, dan 41 pasien yang diprediksi negatif dan memang negatif.

Dengan demikian, total prediksi pasien yang memiliki diabetes adalah 22 pasien. Selain itu model ini menghassilkan presisi tertinggi sebesar 59,09% untuk pasien positif terkena diabetes dan 74,55% untuk pasien negatif terkena diabetes. Recall untuk pasien positif adalah 48,15%, sedangkan untuk pasien negatif adalah 82,00%. Hasil ini memberikan gambaran tentang seberapa efektif model K-NN dalam mengidentifikasi pasien dengan diabetes di antara perempuan Indian Pima.

KESIMPULAN

Implementasi metode *K-Nearest Neighbors* (KNN) terhadap data klasifikasi penyakit diabetes pada perempuan Indian Pima dilakukan dengan menggunakan dataset yang berjumlah 769 data dengan 9 atribut. Aplikasi yang digunakan adalah RapidMiner dengan berbagai operator seperti *Read CSV*, *Split Data*, *K-Nearest Neighbors*, *Apply Model*, dan *Performance*. Hasil dari model ini menunjukkan tingkat akurasi yang cukup baik, namun masih dapat ditingkatkan.

Metode K-Nearest Neighbors ini cocok digunakan untuk klasifikasi penyakit diabetes, namun kualitas akurasi sangat dipengaruhi oleh kualitas data yang dianalisis. Oleh karena itu, diperlukan evaluasi dan pengembangan lebih lanjut untuk meningkatkan kualitas model secara keseluruhan. Pada penelitian ini, akurasi yang didapatkan sebesar 70,13%, dengan presisi sebesar 59,09% untuk pasien positif dan 74,55% untuk pasien negatif. *Recall* yang diperoleh untuk pasien positif adalah 48,15% dan untuk pasien negatif sebesar 82,00%.

UCAPAN TERIMA KASIH

Puji dan syukur saya panjatkan kepada Tuhan Yang Maha Esa, karena atas berkat dan rahmat-Nya, saya dapat menyelesaikan jurnal ini. Terima kasih yang sebesar-besarnya kepada dosen pembimbing yang telah mengarahkan saya dengan penuh kesabaran. Dan juga terima kasih untuk segenap dukungan, khususnya kepada teman-teman yang selalu mengingatkan saya akan terselesaikannya penelitian ini. Tanpa adanya semua dukungan ini, sulit bagi kami untuk menyelesaikan penelitian ini.

DAFTAR PUSTAKA

Andrian, R., Naufal, M. A., Hermanto, B., Junaidi, A., & Lumbanraja, F. R. (2019). K-Nearest Neighbor (k-NN) Classification for Recognition of the Batik Lampung Motifs. *Journal of Physics: Conference Series*, *1338*(1). https://doi.org/10.1088/1742-6596/1338/1/012061

Hasil, P., Merdeka, S., Kampus, B., Di, M., Bhayangkara, U., Raya, J., … Clustering, D. K. (2022). Pelita teknologi, *17*(2), 1–11.

Ivandari, Much. Rifqi Maulana, Muhammad Faizal Kurniawan, & Al Karomi, M. A. (2023). Komparasi Algoritma Data Mining untuk Klasifikasi Penyakit Diabetes. *Bulletin of Computer Science Research*, *3*(5), 343–350. https://doi.org/10.47065/bulletincsr.v3i5.280

Nasional, J., Komputer, T., Informasi, M. T., Pembangunan, U., Budi, P., Medan, K., … Komputer, T. (2024). Analisa Classification Decision Tree C45 dan Naïve Bayes Pada Indikasi Penyakit Diabetes Menggunakan Rapid Miner Data Mining adalah proses penggalian informasi dan pola yang bermanfaat dari suatu data yang sangat besar . Proses data mining terdiri dari pe, *4*, 25–33.

Ocal, S., Gokcek, M., Colak, A. B., & Korkanc, M. (2021). A COMPREHENSIVE AND COMPARATIVE EXPERIMENTAL ANALYSIS ON THERMAL CONDUCTIVITY OF TiO2-CaCO3/WATER HYBRID NANOFLUID: PROPOSING NEW CORRELATION AND ARTIFICIAL NEURAL NETWORK OPTIMIZATION. *Heat Transfer Research*, *52*(17), 55–79. https://doi.org/10.1615/HeatTransRes.2021039444

Pamungkas, P. (2018). Caritas pro Serviam, 2018. *ASMI Santa Maria Yogyakarta*, (November), 64–77.

Rayuwati, Husna Gemasih, & Irma Nizar. (2022). IMPLEMENTASI AlGORITMA NAIVE BAYES UNTUK MEMPREDIKSI TINGKAT PENYEBARAN COVID. *Jural Riset Rumpun Ilmu Teknik*, *1*(1), 38–46. https://doi.org/10.55606/jurritek.v1i1.127

Rinanda, P. D., Delvika, B., Nurhidayarnis, S., Abror, N., & Hidayat, A. (2022). Perbandingan Klasifikasi Antara Naive Bayes dan K-Nearest Neighbor Terhadap Resiko Diabetes pada Ibu Hamil. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *2*(2), 68–75. https://doi.org/10.57152/malcom.v2i2.432

Sistem Komputer dan Sistem Informasi, J., Studi Teknologi Komputasi dan Informatika Stmik Bina Bangsa Kendari, P., Aris, F., Program Studi Sistem Komputer, D., Studi Sistem Komputer, P., & Bina Bangsa Kendari, S. (2019). Penerapan Data Mining untuk Identifikasi Penyakit Diabetes Melitus dengan Menggunakan Metode Klasifikasi. *Router Research*, *1*(1), 1–6.

Utomo, D. P., & Mesran, M. (2020). Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung. *Jurnal Media Informatika Budidarma*, *4*(2), 437. https://doi.org/10.30865/mib.v4i2.2080

Widaningsih, S. (2022). Penerapan Data Mining untuk Memprediksi Siswa Berprestasi dengan Menggunakan Algoritma K Nearest Neighbor. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *9*(3), 2598–2611. https://doi.org/10.35957/jatisi.v9i3.859

Yogianto, A., Homaidi, A., & Fatah, Z. (2024). Implementasi Metode K-Nearest Neighbors (KNN) untuk Klasifikasi Penyakit Jantung. *G-Tech: Jurnal Teknologi Terapan*, *8*(3), 1720–1728. https://doi.org/10.33379/gtech.v8i3.4495

Zulfikar, Z., Podungge, E. S., Saleh, M. I., & ... (2022). Penerapan Data Mining Untuk Memprediksi Tingkat Kelulusan Siswa Menggunakan Algoritma Neural Network. *Jurnal Elektronik Sistem …*, *5*(1), 7–13. Retrieved from http://jesik.web.id/index.php/jesik/article/view/91