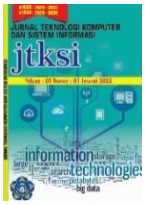


JTKSI (Jurnal Teknologi Komputer dan Sistem Informasi)
JTKSI, Volume 5, Nomor 3, September 2022
E ISSN: 2620-3030; P ISSN: 2620-3022, pp.178-182
Accredited SINTA 4 Nomor 200/M/KPT/2020
http://ojs.stmikpringsewu.ac.id/index.php/jtksi
Received: 25 Agustus 2022; Revised: 8 September 2022; Accepted: 28 September 2022

Diabetes Classification Analysis Using the Euclidean Distance Method Based on the K-Nearest Neighbors Algorithm

Galih Mahalisa[1], Nur Arminarahmah[2]
[1,2]Teknik Informatika, Universitas Islam Kalimantan Muhammad Arsyad Al Banjari Banjarmasin
[1,2]Jl. Adhyaksa, Jl. Kayu Tangi 1 Jalur 2 No.2, Sungai Miai, Kota Banjarmasin, Kalimantan Selatan, Indonesia
E-Mail: galih.mahalisa@gmail.com[1], nur.armina@gmail.com[2]

Abstrak

Diabetes adalah penyakit kronis yang ditandai dengan ciri-ciri berupa tingginya kadar gula (glukosa) darah. Diabetes dapat meningkatkan risiko sejumlah masalah mata, beberapa di antaranya dapat menyebabkan kehilangan penglihatan. Sekitar 9,1 juta penduduk Indonesia diperkirakan menderita penyakit diabetes. Berdasarkan kelompok usia, penderita diabetes paling banyak berada pada rentang usia 55–74 tahun. Meski demikian, penyakit ini juga dialami oleh orang muda di usia 20-an hingga 40-an. Salah satu cara untuk mendeteksi klasifikasi penyakit diabetes dalam machine learning yatiu menggunakan dataset sebagai data latih agar dapat dilakukan pengujian performa dengan metode klasifikasi yang tepat. Metode yang digunakan dalam penelitian ini yaitu algoritma K-Nearest Neighbor (KNN), dimana merupakan sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Hasil dari penelitian ini yaitu Pengenalan Pola dalam menentukan nilai K yang tepat sehingga menunjukan nilai akurasi yang baik. Pada tahapan ini diperoleh nilai K = 19 menunjukan nilai akurasi yang baik yaitu sekitar 76% dengan Tingkat kesalahan pada nilai K= 19 yaitu sekitar 24%.

Kata Kunci: Akurasi, Diabetes, Klasifikasi, K-Nearest Neighbor, KNN

Abstract

Diabetes is a chronic disease characterized by high blood sugar (glucose) levels. Diabetes can increase the risk of a number of eye problems, some of which can lead to vision loss. It is estimated that 9.1 million Indonesians suffer from diabetes. Based on age group, most people with diabetes are in the 55-74 year age range. However, this disease is also experienced by young people in their 20s to 40s. One way to detect the classification of diabetes in machine learning is to use a dataset as training data so that performance testing can be carried out with the right classification method. The method used in this study is the K-Nearest Neighbor (KNN) algorithm, which is a method for classifying objects based on learning data that is closest to the object. The results of this study are Pattern Recognition in determining the right K value so that it shows a good accuracy value. At this stage, the value of K = 19 shows a good accuracy value, which is about 76% with an error rate of K = 19, which is about 24%.

Keywords: Accuracy, Classification, Diabetes, K-Nearest Neighbor, KNN

I. PENDAHULUAN

Diabetes adalah penyakit kronis yang ditandai dengan ciri-ciri berupa tingginya kadar gula (glukosa) darah [1]. Diabetes dapat meningkatkan risiko sejumlah masalah mata, beberapa di antaranya dapat menyebabkan kehilangan penglihatan. Sekitar 9,1 juta penduduk Indonesia diperkirakan menderita penyakit diabetes. Berdasarkan kelompok usia, penderita diabetes paling banyak berada pada rentang usia 55–74 tahun [2]. Meski demikian, penyakit ini juga dialami

oleh orang muda di usia 20-an hingga 40-an. Pengklasifikasian digunakan sebagai alat bantu bahkan dijadikan suatu bahan pertimbangan dalam menghasilkan outcome yang akurat

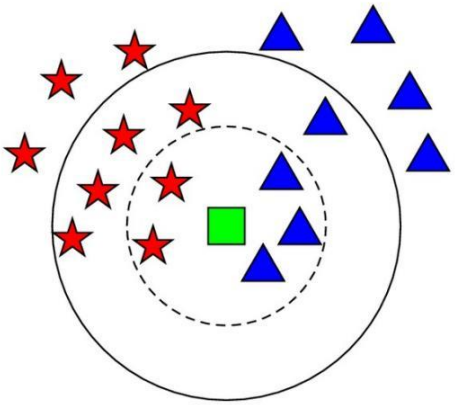
Referensi yang berkaitan dengan penelitian ini seperti berikut : 1)Peningkatan akurasi algoritma knn dengan seleksi fitur gain ratio untuk klasifikasi penyakit diabetes mellitus, menjelaskan bahwa penggunaan algoritma seleksi fitur gain ratio dapat meningkatkan akurasi dari klasifikasi penyakit diabetus mellitus dengan menggunakan algoritma knn hanya mempertahankan 4 atribut dari keseluruhan 8 atribut data [3]. 2) Analisis Perbandingan KNN dengan SVM untuk Klasifikasi Penyakit Diabetes Retinopati berdasarkan Citra Eksudat dan Mikroaneurisma, menjelaskan bahwa Tingkat akurasi yang diperoleh dengan menggunakan metode KNN lebih tinggi dibandingkan SVM, yaitu 65 % dan 62%. dan klasifikasi dengan algoritma KNN diperoleh hasil terbaik dengan parameter K=9 cityblock [4]. 3) Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes menjelaskan bahwa hasil akhir penelitian ini, telah dihitung akurasi tertinggi 39% pada K=3, presisi tertinggi 65% pada K=3 dan K=5 [5]. Aplikasi klasifikasi dapat digunakan, misalnya dalam bidang medis untuk mengklasifikasikan tingkat keparahan penyakit pasien, sehingga memudahkan dokter untuk memberikan solusi pengobatan yang tepat. Berbagai metode data mining dapat digunakan untuk menyelesaikan masalah klasifikasi. Akurasi klasifikasi objek sangat penting. Banyaknya atribut dapat mempengaruhi performa suatu algoritma [6]. Masalah klasifikasi pada dasarnya adalah sebagai berikut [7]: 1. Masalah Klasifikasi berangkat dari data training yang tersedia. 2. Data training akan diolah dengan menggunakan algoritma klasifikasi. 3. Masalah klasifikasi berakhir dengan dihasilkannya sebuah pengetahuan yang direpresentasikan dalam bentuk diagram, aturan atau pengetahuan.

Beberapa algoritma dapatdigunakan untuk melakukan tugas klasifikasi [8]. Salah satu algoritma klasifikasi data mining yang terbaik adalah K-Nearest Neighbour (KNN) [9]. Salah satu kelemahan dari algoritma KNN adalah dalam penentuan variabel K [3]. Nilai K yang terlalu besar akan membuat hasil klasifikasi semakin kabur. Sedangkan apabila nilai K yang digunakan adalah 1 akan mengakibatkan hasil klasifikasi terasa kaku. Penelitian ini menghitung nilai K paling optimal algoritma K-NN untuk Klasifikasi Penyakit Diabetes Mellitus. Dalam penelitian ini menggunakan metode Klasifikasi K-Nearest Neighbor dengan Euclidean distance dengan beberapa fitur yang ada agar prediksi yang didapatkan bisa lebih akurat dan spesifik.

II. LANDASAN TEORI

Data mining adalah teknik yang menggunakan sejumlah besar data untuk mendapatkan informasi yang sebelumnya tidak diketahui dan berharga yang dapat digunakan untuk membuat keputusan penting, salahsatu metode yang digunakan pada penelitian ini adalah K-nearest Neighbors [10].

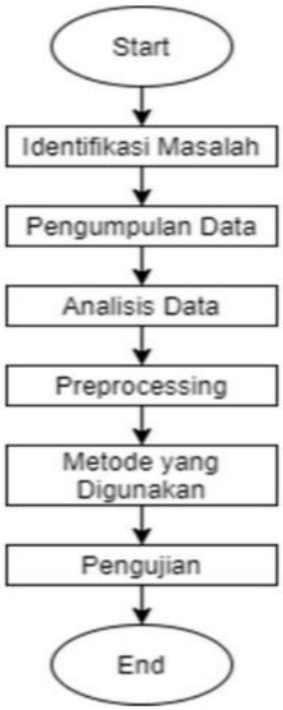
K-Nearest Neighbors atau KNN adalah algoritma yang berfungsi untuk melakukan klasifikasi suatu data berdasarkan data pembelajaran (train data sets), yang diambil dari k tetangga terdekatnya (nearest neighbors). Dengan k merupakan banyaknya tetangga terdekat [11]. Jarak Euclidean mengukur kedekatan antara dua objek. Ini digambarkan sebagai garis lurus atau pengukuran garis lurus. Metode pengukuran ini cocok diterapkan pada 14 data dengan nilai atribut numerik, terutama data dengan atribut kontinu.



Gambar 1. Struktur Umum KNN [12]

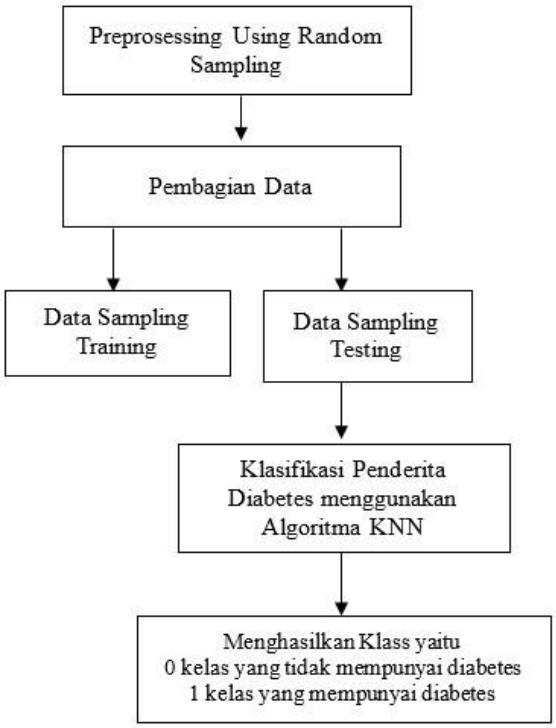
Data baru yang diklasifikasi selanjutnya diproyeksikan pada ruang dimensi banyak yang telah memuat titik-titik c data pembelajaran. Proses klasifikasi dilakukan dengan mencari titik c terdekat dari c-baru (nearest neighbor). Teknik pencarian tetangga terdekat yang umum dilakukan dengan menggunakan formula jarak euclidean, yaitu formula untuk mencari jarak antara 2 titik dalam ruang dua dimensi [13].

Untuk mengklasifikasikan penderita diabetes, pendekatan penelitian menggunakan metode kuantitatif, dan metode penelitian dilakukan dengan menggunakan data yang memiliki karakteristik dataset relevan untuk mengklasifikasi penderita diabetes seperti Pregnancies Glucose,Blood Pressure,Skin thickness, Insulin, BMI, Diabetes Pedigress Function, Age dan Outcome kemudian diterapkan algoritma KNN pada kumpulan data untuk menggambarkan penelitian yang dilakukan. Caranya seperti gambar 2 di bawah ini.



Gambar 2. Tahap Metode Penelitian [14]

Pengolahan data dimulai dengan mengidentifikasi kumpulan data yang berisi tipe data integer dan data floating-point. Setelah menganalisis data, langkah selanjutnya adalah preprocessing dan splitting dataset. Artinya, sebaran data latih dan uji seperti terlihat pada gambar di bawah ini:



Gambar 3. Pengolahan Data [9]

III. HASIL DAN PEMBAHASAN

A. Pembahasan Data Set

Pengumpulan dataset dengan cara pengunduhan pada API Command kaggle (kaggle datasets download -d mathchi/diabetes-data-set). Karakteristik dataset terdiri dari 9 variabel. 8 fitur yang dijadikan variabel X atau disebut variabel dependen dan 1 variabel yang dijadikan variabel Y atau variabel independen, dan data sebanyak 768, selanjutnya melakukan prepocesssing, kemudian melakukan split dataset yaitu pembagian data training dan data testing.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

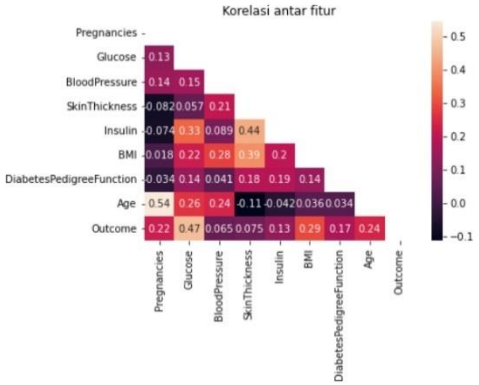
Gambar 4. Dataset

B. Tahap Penerapan KNN

Outcome yang ingin dicapai adalah menggunakan kelas 0 dan 1, nilai 0 sebagai kelas yang tidak mempunyai diabetes pada sample, dan nilai 1 sebagai kelas yang mempunyai diabetes sample. Terdapat beberapa fitur yang digunakan yaitu pregnancies, glucose, blood presure, bmi, Skin Thickness, Insulin, dan Diabetes Pedigree Function. Algoitma KNN terdiri dari beberapa tahapan, yaitu:

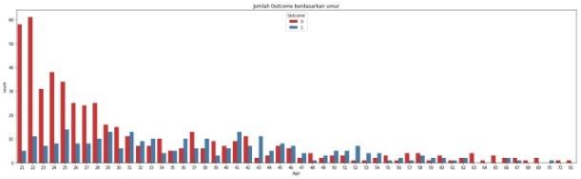
1. Pengukuran Jarak, dalam penelitian ini menggunakan Euclidean Distance
2. Memilih Nilai K, akan dilakukan dengan beberapa percobaan agar mendapatkan nilai K yang terbaik
3. Mengurutkan jarak dari yang paling kecil.

C. Tahap Pengujian



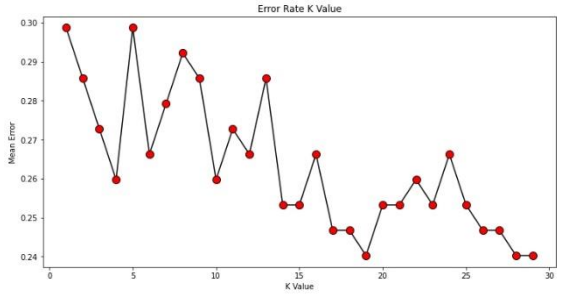
Gambar 5. Korelasi Antar Fitur

Dari data visualisasi gambar 5 dapat dilihat bahwa semakin besar angkanya maka semakin besar nilai dari fitur tersebut dan semakin mempengaruhi prediksi analisa penyakit diabetes.



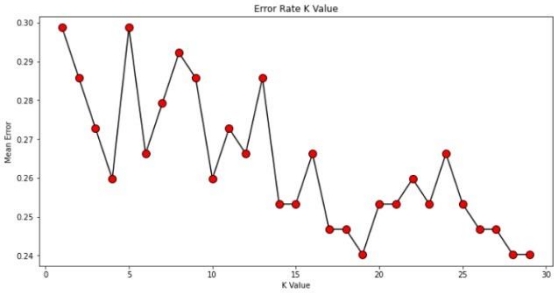
Gambar 6. Jumlah Outcome berdasarkan usia

Setelah dilakukan percobaan analisis data dapat dilihat pada gambar 6 bahwa penderita diabetes paling banyak pada rentang usia 20 sampai 40



Gambar 7. Tingkat kesalahan.

Untuk mengetahui K yang terbaik, maka dilakukan percobaan dengan range k=1 hingga k=29. pada gambar 7 menunjukan tingkat kesalahan pada nilai K= 19 yaitu sekitar 24%.



Gambar 8. Akurasi Nilai K

Pada gambar 8 terlihat bahwa nilai K=19 menunjukan nilai akurasi yang baik yaitu sekitar 76%.

IV. KESIMPULAN

Hasil klasifikasi menggunakan data sampling secara random untuk penderita diabetes dan dengan karakteristik dataset terdiri dari 9 variabel. 8 fitur yang dijadikan variabel X atau disebut variabel dependen dan 1 variabel yang dijadikan variabel Y atau variabel independent serta data sebanyak 768 hasil yang didapat yaitu bahwa Penderita diabetes berdasarkan jumlah outcome berdasarkan usia yaitu paling banyak pada rentang usia 20 sampai 40 tahun.Pengenalan Pola dalam menentukan nilai K yang tepat sehingga menunjukan nilai akurasi yang baik dan pada tahapan ini diperoleh nilai K = 19. Tingkat kesalahan pada nilai K= 19 yaitu sekitar 24%. Nilai K=19 menunjukan nilai akurasi yang baik yaitu sekitar 76%.

DAFTAR PUSTAKA

[1] Y. Nurdiansyah, "Informal : informatics journal.," vol. 2, no. 2, pp. 114–122, Jul. 2017.

[2] D. R. Ente, S. A. Thamrin, S. Arifin, H. Kuswanto, and A. Andreza, "KLASIFIKASI FAKTOR-FAKTOR PENYEBAB PENYAKIT DIABETES MELITUS DI RUMAH SAKIT UNHAS MENGGUNAKAN ALGORITMA C4.5," *Indones. J. Stat. Its Appl.*, vol. 4, no. 1, pp. 80–88, Feb. 2020, doi: 10.29244/ijsa.v4i1.330.

[3] D. Sugianti, M. Adib Al Karomi, and S. Widya Pratama Pekalongan, *OPTIMASI PARAMETER K PADA ALGORITMA K-NEAREST NEIGHBOUR UNTUK KLASIFIKASI PENYAKIT DIABETES MELLITUS*.

[4] S. Aulia, S. Hadiyoso, and D. Nur Ramadan, "Analisis Perbandingan KNN dengan SVM untuk Klasifikasi Penyakit Diabetes Retinopati berdasarkan Citra Eksudat dan Mikroaneurisma."

[5] A. M. Argina, "Indonesian Journal of Data and Science Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," *jurnal.yoctobrain.org*, vol. 1, no. 2, pp. 29–33, 2020.

[6] A. Homaidi, "Aplikasi Pengusulan dan Pemantauan Pelaksanaan Penelitian dan Pengabdian Masyarakat Universitas Ibrahimy," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 20, no. 2, pp. 225–236, May 2021, doi: 10.30812/matrik.v20i2.942.

[7] H. J. Suryanto, A. R. C., and Y. Lukito, "Indoor Positioning System dengan Algoritma K-Means dan KNN," *J. Tek. Inform. dan Sist. Inf.*, vol. 2, no. 3, Dec. 2016, doi: 10.28932/JUTISI.V2I3.641.

[8] Kusrini and L. Taufiq Emha, "Algoritma Data Mining Yogyakarta," *Algoritm. Data Min.*, no. February, pp. 149–176, 2009.

[9] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2888, pp. 986–996, 2003, doi: 10.1007/978-3-540-39964-3_62.

[10] D. A. Fauziah, A. Maududie, and I. Nuritha, "Klasifikasi Berita Politik Menggunakan Algoritma K-nearst Neighbor," *Berk. SAINSTEK*, vol. 6, no. 2, p. 106, Dec. 2018, doi: 10.19184/bst.v6i2.9256.

[11] R. K. Dinata, H. Akbar, and N. Hasdyna, "Algoritma K-Nearest Neighbor dengan Euclidean Distance dan Manhattan Distance untuk Klasifikasi Transportasi Bus," *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 104–111, Aug. 2020, doi: 10.33096/ilkom.v12i2.539.104-111.

[12] S. Mutro□, A. Izzah, A. Kurniawardhani, and M. Masrur, "OPTIMASI TEKNIK KLASIFIKASI MODIFIED K NEAREST NEIGHBOR MENGGUNAKAN ALGORITMA GENETIKA Optimization Techniques Modi□ ed k Nearest

Neighbor Classi cation Using Genetic Algorithm," *ejournal.umm.ac.id*, pp. 130–134, 2014.

[13] D. Sugianti, M. Adib Al Karomi, and S. Widya Pratama Pekalongan, *OPTIMASI PARAMETER K PADA ALGORITMA K-NEAREST NEIGHBOUR UNTUK KLASIFIKASI PENYAKIT DIABETES MELLITUS*, vol. 0, no. 0. 2017.

[14] V. Z. Kamila and E. Subastian, "ANALISIS DAN PERANCANGAN SISTEM EVALUASI PELATIHAN TENAGA KEPENDIDIKAN," *Sebatik*, vol. 24, no. 2, Dec. 2020, doi: 10.46984/sebatik.v24i2.1125.