Original Article

e-ISSN: 2774-2016 - https://journal.itera.ac.id/index.php/indojam/
p-ISSN: 2774-2067

Prediksi Terkena Diabetes menggunakan Metode K-Nearest Neighbor (KNN) pada Dataset UCI Machine Learning Diabetes

Received 2nd September 2023
Accepted 7th November 2023
Published 11th November 2023

Open Access

DOI:
10.35472/indojam.v3i2.1577

Rika Ajeng Finatih[a], Muhammad Farhan Athaullah[a], Anisa Cahyani Surya[a], Pramudya Wibowo[a], Kiwit Novitasari[a], Muhammad Shauqi Athallah[a], Mika Alvionita S*[a], Febri Dwi Irawati[a]

[a] *Program Studi Sains Data, Jurusan Sains, Institut Teknologi Sumatera*

* *Corresponding E-mail: mika.alvionita@sd.itera.ac.id*

Abstract:

This research uses the K-Nearest Neighbor (KNN) algorithm to predict a person's risk of developing diabetes. The variables used in the prediction are pregnancy, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age. Analysis shows that glucose, BMI, and age have a high correlation with diabetes diagnosis, making them strong indicators for prediction. Using the KNN method with $K = 3$, model evaluation was carried out using the Confusion Matrix. The result shows an accuracy of 66,5%. These finding indicates that KNN with $K = 3$ is effective in predicting diabetes based on clinical variables. This information can provide benefits in the prevention and treatment of diabetes more effectively.

Keywords: ***Diabetes, K-Nearest Neighbor, UCI Machine Learning, Confusion Matrix.***

Abstrak:

Penelitian ini menggunakan algoritma *K-Nearest Neighbor* (KNN) untuk memprediksi resiko seseorang terkena diabetes. Variabel yang digunakan dalam prediksi adalah *pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function*, dan *age*. Analisis menunjukkan bahwa g*lucose, BMI,* dan a*ge* memiliki korelasi tinggi dengan diagnosis diabetes, menjadikannya indikator yang kuat untuk prediksi. Melalui metode KNN dengan $K = 3$, dilakukan evaluasi model menggunakan *Confusion Matrix*. Hasilnya menunjukkan akurasi sebesar 66,5%. Temuan ini menunjukkan bahwa KNN dengan $K = 3$ efektif dalam memprediksi diabetes berdasarkan variabel klinis. Informasi ini dapat memberikn manfaat dalam pencegahan dan pengobatan diabetes secara lebih efektif.

Kata Kunci*: Diabetes, K-Nearest Neighbor, UCI Machine Learning, Confusion Matrix.*

Pendahuluan

Kemajuan zaman membawa sebuah perubahan gaya hidup dalam masyarakat. Pada Tahun 2021, *International Database Federation* (IDF) memperkirakan bahwa sebanyak 537 juta masyarakat dunia hidup dengan diabetes. Beberapa pakar melihat bahwa urbanisasi, modernisasi, dan westernisasi menjadi faktor-faktor yang menyebabkan timbulnya diabetes. Namun, faktor resiko utama timbulnya penyakit

Original Article

diabetes ini dapat disebabkan oleh usia seseorang yang sudah berada di atas 40 tahun, riwayat keturunan, serta badan yang terlalu gemuk. Diabetes adalah suatu kondisi yang terjadi ketika kadar gula darah meningkat. Kadar gula darah penting untuk kesehatan karena merupakan sumber energi penting bagi sel dan jaringan tubuh. Namun, jika tidak dikelola dengan baik, diabetes dapat menyebabkan berbagai komplikasi yang serius seperti penyakit jantung, stroke, obesitas, dan gangguan pada organ tubuh seperti mata, ginjal,dan saraf. [1]

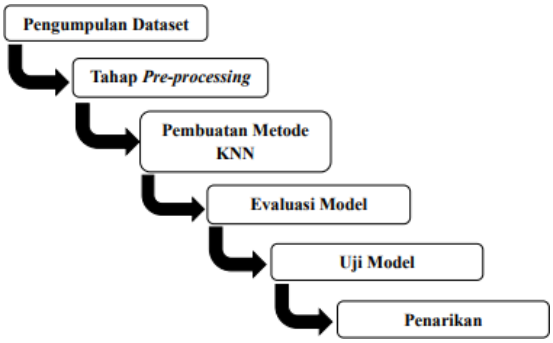
Diabetes adalah penyakit metabolik yang mempengaruhi kesehatan banyak orang di seluruh dunia [2]. Penyakit ini mempengaruhi kualitas hidup dan dapat menyebabkan komplikasi serius jika tidak dikelola dengan baik. Dalam rangka meningkatkan deteksi dini dan pengobatan diabetes, teknologi *machine learning* dapat digunakan untuk melakukan prediksi resiko seseorangterkena diabetes berdasarkan data klinis dan genetik.

Salah satu metode *machine learning* yang dapat digunakan untuk memprediksi penyakit diabetes adalah *K-Nearest Neighbor* (KNN). Metode ini berfungsi untuk mengklasifikasikan data berdasarkan kategori yang telah ditentukan dengan mempertimbangkan jarak antar data yang akan diklasifikasikan dengan data latihan yang sudah diketahui kategorinya. Pada penelitian ini, metode KNN akan diterapkan pada dataset UCI *Machine Learning* Diabetes untuk melakukan analisis prediksi kemungkinan terkena diabetes. Dataset ini berisikan informasi klinis dari pasien diabetes, seperti usia, BMI, tekanan darah, dan kadar glukosa darah. Dengan menggunakan metode KNN, akan dilakukan klasifikasi pasien diabetes berdasarkan variabel-variabel klinis yang tersedia.

Hasil dari analisis ini diharapkan dapat memberikan informasi yang bermanfaat bagi dokter dan pasien untuk melakukan tindakan pencegahan dan pengobatan diabetes secara lebih efektif. Selain itu, penerapan metode KNN pada dataset diabetes ini juga dapat menjadi salah satu contoh penerapan teknologi *machine learning* dalam bidang kesehatan.

Metode

Penelitian ini menggunakan data sekunder dari *repository* UCI *Machine Learning* dan menggunakan *software* RStudio [3]. UCI *Machine Learning* adalah sebuah *repository* daringyang berisikan berbagai jenis dataset dan algoritma *machine learning* yang dapat digunakan untuk keperluan penelitian dan pembelajaran mesin. *Repository* ini disediakan oleh Universitas California, Irvine (UCI). UCI *Machine Learning* menyediakan berbagai dataset yang bervariasi dari berbagai bidang, seperti ilmu sosial, ilmu biologi, ilmu kesehatan, ilmu komputer, dan sebagainya. Selain itu, repository ini juga menyediakan berbagai algoritma *machinelearning* dan *software* untuk membantu para peneliti dan pengembang dalam membuat model *machine learning*. **Gambar 1** menunjukkan tahapan penelitian yang akan dilakasanakan.



Gambar 1. Tahapan Penelitian

Tahapan pada **Gambar 1** dijelaskan sebagai berikut.

1. Pengumpulan Data
   Penelitian ini menggunakan data sekunder dari repository database UCI *Machine Learning* yang terdiri dari 2000 data klinis. Semua pasien dalam dataset adalah perempuan yang tinggal di Phoenix, Arione, Amerika Serikat dan berumur minimal 21 tahun. Dataset ini terdiri dari dua kelas yang diwakili oleh variabel biner dengan nilai "0" atau "1". Jika hasil tes menunjukan bilangan "1", maka pasien tersebut terkena diabetes. Sedangkan jika pasien tersebut menunjukan angka "0" menunjukan pasien tersebut negative terkena diabetes. Dataset ini terdiri dari 2000 pasien dengan sembilan variabel numerik. Dalam dataset ini terdapat 684 (34,9%) kasus positif

diabetes yang ditandai dengan kelas "1" dan 1316 (65,1%) kasus negative diabetes yang ditandai dengan kelas "0". Dataset ini tidak mengandung *missing values* namun terdapat lima pasien dengan nilai glukosa "0", sebelas pasien dengan nilai BMI "0", 28 pasien dengan nilai blood pressure "0", 192 pasien dengan nilai skin fold thickness "0",dan 140 pasien dengan nilai serum insulin level "0" [3].

2. Tahap *Pre-processing*
   Tahap *Pre-processing* data meliputi identifikasi dan pemilihan atribut (*attribute identification and selection*), penanganan nilai atribut yang hilang atau kurang lengkap *(handling missing value),* dan proses diskritisasi nilai. Kemudian pada tahap ini juga akan dilakukan korelasi pada dataset untuk mencari tahu sejauh mana hubungan antara variabel. Nantinya akan diambil 3 variabel yang memiliki korelasi terhadap diabetes. Setelah mendapatkan korelasi dari tiap variabel, kemudian dilakukan *split data*.

3. Analisis menggunakan *Metode K-Nearest Neighbor* (KNN)
   K-Nearest Neighbor (KNN) adalah algoritma yang digunakan untuk mengklasifikasikan objek baru dengan mempertimbangkan atribut dan sampel pelatihan yang sudah ada. Klasifikasi dilakukan tanpa menggunakan model dan hanya berdasarkan pada memori. Dalam algoritma ini, titik *query* akan dicari sejumlah K titik pelatihan terdekat, dengan klasifikasi dilakukan berdasarkan mayoritas dari K titik tersebut. KNN menggunakan metode klasifikasi berdasarkan ketetanggaan dengan menghitung jarak terpendek dari *query instance* ke sampel pelatihan untuk menentukan KNN. Algoritma KNN sangat sederhana dan efektif dalam memprediksi klasifikasi objek baru. Langkah-langkah dalam menggunakan metode KNN yaitu:
   a. Menentukan nilai jumlah tetangga terdekat atau K
   b. Menghitung jarak *Euclidean,* sebagai berikut
      $$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$
      Keterangan:
      n : Dimensi data
      i : Variabel data
      x, y :Dua titik dalam ruang Eculidean
      d : Jarak
   c. Mengelompokkan objek berdasarkan jarak *Euclidean* terkecil.
   d. Mengumpulkan kelompok Y dalam KNN
   e. Dari kelompok *Nearest Neighbor* yang paling banyak dapat diprediksi nilai K yang telah dihitung
   f. Membagi data menjadi data latih dan data uji dengan menggunakan nilai K yang berbeda, kemudian mencari *confusion matrix.*
   g. Dari hasil *confusion matrix* dapat dihitung nilai *precision, recall, dan F-measure* dengan menjabarkan *confusion matrix.*

4. Evaluasi Model

   Pada evaluasi model ini dilihat nilai akurasi model dengan melihat derajat kedekatan dari pengukuran kuantitas untuk nilai sebenarnya.

5. Uji Model
   Dalam tahap pengujian model ini dilakukan validasi dan pengukuran keakuratan hasil yang dicapai oleh model menggunakan c*onfusion matrix.*

6. Kesimpulan
   Tahap selanjutnya yaitu menyimpulkan hasil yang diperoleh dari penelitian. Pada penarikan kesimpulan ini dapat diketahui variabel mana saja yang banyak mempengaruhi terjangkitnya penyakit diabetes. Kemudian pada penarikan kesimpulan, akan dilihat akurasi dalam mendiagnosis penyakit diabetes berdasarkan nilai *precision, recall, F-measure.*

Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia

Original Article

Hasil dan Diskusi

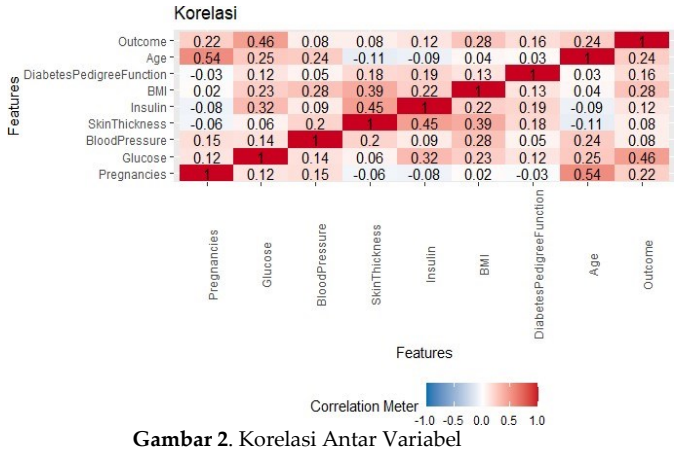
Penelitian ini menggunakan *outcome* sebagai variabel prediktor karena didalamnya terdapat informasi mengenai apakah pasien tersebut didiagnosis diabetes atau tidak. Variabel lainnya adalah *pregnancy* (banyaknya kehamilan), *glucose* (kadar glukosa darah), *blood pressure* (tekanan darah), *skin thickness* (ketebalan kulit), insulin, BMI (Indeks Massa Tubuh), *diabetes pedigree function* (faktor keturunan diabetes), dan *age* (usia). Selanjutnya dilakukan pegecekan terhadap data dengan mengidentifikasi *missing value* dan menghilangkan nilai tersebut menggunakan *na.omit()* dalam RStudio (ditunjukkan pada **Tabel 1).**

Skim_variable <chr>	N_missing	Complete_rate	Mean <dlb>	sd <dlb>	p0 <dlb>	p25 <dlb>
1. Pregnancies	0	1	3.70350	3.3060630	0.000	1.000
2. Glucose	0	1	121.1825	32.0686356	0.000	99.000
3. Blood Preassure	0	1	69.14550	19.1883148	0.000	63.000
4. Skin Thickness	0	1	20.93500	16.1032429	0.000	0.000
5. Insulin	0	1	80.25400	111.1805335	0.000	0.000
6. BMI	0	1	32.19300	8.1499007	0.000	27.375
7. Diabetes Pedigree Function	0	1	0.47093	0.3235526	0.078	0.244
8. Age	0	1	33.09050	11.78	21.000	24.000
9. Outcome	0	1	0.34200	0.4744982	0.000	0.000

Tabel 1. Dataset setelah menghilangkan *missing value*

Setelah penanganan *missing value* maka ditentukan korelasi pada penelitian ini. **Gambar 2** menunjukkan korelasi antar variabel. Berdasarkan **Gambar 2** variabel yang memiliki korelasi dalam mengdiagnosis seseorang terkena diabetes yaitu Glukosa (kadar gula darah) dengan korelasi sebesar 0.46, BMI (IndeksMasa Tubuh) dengan korelasi sebesar 0.28, dan Age (Usia) dengan korelasi sebesar 0.24.

Selanjutnya dilakukan *split data* UCI *Machine Learning* dengan membagi variabel prediktor yaitu *outcome* (80% data latih dan 20% data uji). Metode Klasifikasi K-Nearest Neighbor (KNN) digunakan untuk memprediksi kemungkinan seseorang terkena diabetes



Gambar 2. Korelasi Antar Variabel

pada data uji. KNN adalah algoritma yang bekerja dengan mencari K tetangga terdekat dari data baru berdasarkan jarak yang diukur, seperti jarak Euclidean. Uji coba ini, dilakukan 5 kali percobaan dengan menggunakan nilai K-tetangga yang berbeda beda, diantaranya: $K-1, K-3, K-5, K-7, K-9$ dan $K-11$. Hal ini dilakukan untuk melihat keakuratan dari setiap k yang dicoba**.Table 1** menunjukkan hasil dari akurasi dari setiap K yangberbeda-beda.

Tabel 2. Hasil akurasi dari setiap K

Nilai k yang digunakan	Hasil Akurasi
$K-3$	66,5 %
$K-5$	53,25 %
$K-7$	43,5 %
$K-9$	35,5 %
$K-11$	25,5 %

Berdasarkan **Tabel 2** menunjukkan bahwa nilai K sangat berpengaruh terhadap nilai akurasi. Hal ini dikarenakan semakin besar nilai K, maka akan semakin banyak tetangga yang digunakan untuk proses Klasifikasi dan kemungkinan untuk terjadinya *noise* juga semakin besar. Penelitian ini dilakukan sebanyak lima kali dengan nilai K yang berbeda- beda. Hasil akurasi yang diperoleh pada **Tabel 2** didapatkan bahwa akurasi terbaik yaitu 66,5% dengan pemilihan $K=3$. Selanjutnya, dievaluasi model KKN dengan $K=3$, menggunakan *confusion matrix*. Hasil *confusion matrix* yang diperoleh dapat dilihat pada **Gambar 3.**

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 158   7
         1   5  55
```

Gambar 3. *Confussion Matrix*

Hasil *confusion matrix* pada **Gambar 3** menunjukkan bahwa terdapat 55 *True Positive* (pasien yang sebenarnya memiliki diabetes dan diprediksi benar), 158 *True Negative* (pasien yang sebenarnya tidak memiliki diabetes dan diprediksi benar), lima *False Positive* (pasien yang sebenarnya tidak memiliki diabetes tetapi diprediksi memiliki diabetes), dan tujuh *False Negative* (pasien yang sebenarnya memiliki diabetes tetapi diprediksi tidak memiliki diabetes). Model KNN yang diperoleh menghasilkan akurasi sebesar 66,5% sehingga Model KNN yang digunakan dalam penelitian ini menggunakan K sebesar 3.

Kesimpulan

Dalam penelitian ini, dilakukan analisis menggunakan algoritma K-Nearest Neighbor (KNN) untuk memprediksi kemungkinan seseorang terkena diabetes. Variabel yang digunakan dalam prediksi tersebut adalah *pregnancies, glucose, blood pressure, skin thickness,* insulin, BMI, *diabetes pedigree function,* dan *age.* Dari hasil penelitian, ditemukan bahwa *Glucose,* BMI, dan *ag*e merupakan variabel yang memiliki korelasi yang tinggi dalam mendiagnosis diabetes.

Selanjutnya, dengan menggunakan metode KNN dengan $K = 3$, dilakukan evaluasi model menggunakan *confusion matrix*. Hasilnya menunjukkan bahwa terdapat 55*True Positive* (pasien yang sebenarnya memiliki diabetes dan diprediksi benar), 158 *True Negative* (pasien yang sebenarnya tidak memiliki diabetes dan diprediksi benar), lima *False Positive* (pasien yang sebenarnya tidak memiliki diabetes tetapi diprediksi memiliki diabetes), dan tujuh *False Negative* (pasien yang sebenarnya memiliki diabetes tetapi diprediksi tidak memiliki diabetes). Dari informasi tersebut, diperoleh nilai akurasi sebesar 66,5%.

Berdasarkan temuan tersebut, dapat disimpulkan bahwametode KNN dengan $K = 3$ relatif baik dalam memprediksi kemungkinan terkena diabetes berdasarkan variabel klinis yang digunakan. Informasi ini diharapkan dapat memberikanmanfaat bagi dokter dan pasien dalam melakukanpencegahan dan pengobatan diabetes secara lebih efektif.

Referensi

[1] A. M. Argina, "Penerapan Metode Klasifikasi K-NearestNeighbor pada Dataset Penderita Penyakit Diabetes," Indonesian Journal of Data and Science, Vol. 1(2), pp. 29-33, Juli 2020.

[2] N. A. M. M. Jan Ahmadi Z, " Hydroalcoholic extract of Allium eriophyllum leaves attenuates cardiac impairment in rats with simultaneous type 2 diabetes and renal hypertension," Res Pharm Sci., Vol. 10(2), pp. 123-133, Mar-Apr 2015.

[3] M. P. W. U. S. L. M. Michael Kahn, UCI Machine Learning Repository, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes. [Accessed 16 5 2023].

Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia