INFORMATICS AND DIGITAL EXPERT (INDEX) - VOL. 6 NO. 1 (2024) 87-91
Terbit *online* pada laman web jurnal : https://e-journal.unper.ac.id/index.php/informatics
| ISSN (Print) 2775-2208 | ISSN (Online) 2715-0453 |

Penerapan Algoritma K Nearest Neighbor untuk Prediksi Akurasi Penyakit Diabetes

Ruhul Amin [1], Erika Tampubolon [2]

[1,2]*Universitas Nusa Mandiri, Jalan Jatiwaringin No.2, Cipinang Melayu, Jakarta Timur, 13620, Indonesia*
e-mail: ruhul.ran@nusamandiri.ac.id [1], erikatpbln10@gmail.com [2]

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi : 26 Maret 2024
Revisi Akhir : 1 Mei 2024
Diterbitkan *Online* : 30 Mei 2024

Kata Kunci :
Diabetes, K-Nearest Neighbor (KNN), Prediksi

Korespondensi :
Telepon / Hp : +62 (0265) 272727
E-mail : ruhul.ran@nusamandiri.ac.id

A B S T R A K

Diabetes mellitus merupakan salah satu penyakit kronis yang prevalensinya terus meningkat di seluruh dunia. Deteksi dini diabetes sangat penting untuk penanganan dan pencegahan komplikasi. Penelitian ini bertujuan untuk mengembangkan model prediksi diabetes menggunakan algoritma K-Nearest Neighbor (KNN) berdasarkan dataset Pima Indians Diabetes. Proses penelitian meliputi preprocessing data, implementasi algoritma KNN, dan evaluasi model. Preprocessing data mencakup penanganan nilai nol, imputasi nilai yang hilang, dan normalisasi fitur. Implementasi KNN dilakukan dengan mencari nilaiK optimal melalui cross-validation. Hasil penelitian menunjukkan bahwa model KNN dengan nilai K optimal 19 mencapai akurasi 75,32% dalam memprediksi diabetes. Analisisperforma model menunjukkan presisi 0,67 dan recall 0,57 untuk kasus positif diabetes. Meskipun model menunjukkan kinerja yang cukup baik, masih terdapat ruang untuk peningkatan, terutama dalam mengurangi false negatives. Penelitian ini menyoroti potensi penggunaan algoritma KNN dalam skrining diabetes dan memberikan dasar untuk pengembangan lebih lanjut dalam prediksi penyakit menggunakan teknik machine learning.

1. PENDAHULUAN

Diabetes melitus adalah gangguan kesehatan yang berkaitan dengan metabolisme tubuh. Karakteristik kunci dari kondisi ini adalah adanya kenaikan level glukosa yang abnormal dalam aliran darah. Fenomena ini dikenal secara umum dengan istilah hiperglikemia. Kondisi ini timbul akibat adanya masalah pada fungsi hormon insulin. Insulin berperan sebagai hormon yang mengatur keseimbangan tubuh dengan cara mengontrol penurunan kadar gula darah[1]. Diabetes merupakan salah satu jenis penyakit yang paling umum ditemui di seluruh dunia. Penyakit ini juga menjadi perhatian di banyak negara saat ini. Kasus diabetes terus meningkat dengan kecepatan yang mengkhawatirkan, dan prevalensinya telah bertambah selama beberapa dekade terakhir. Secara keseluruhan, diperkirakan bahwa jumlah orang dewasa yang didiagnosis menderita Diabetes mellitus mencapai 422 juta jiwa pada tahun 2014, yang merupakan peningkatan signifikan dari jumlah pada tahun 1980 (108 juta jiwa).

Peningkatan prevalensi kondisi ini berkaitan erat dengan makin maraknya faktor-faktor risiko tertentu. Di antaranya adalah kelebihan berat badan yang signifikan dan pola hidup yang cenderung minim aktivitas fisik[2]. Selama beberapa dasawarsa terakhir, telah tercatat adanya kenaikan yang substansial dalam jumlah kasus Diabetes mellitus yang terdiagnosis. Secara keseluruhan, diperkirakan bahwa jumlah orang dewasa yang didiagnosis menderita Diabetes mellitus mencapai 422 juta jiwa pada tahun 2014, yang merupakan peningkatan signifikan dari jumlah pada tahun 1980 (108 juta jiwa). Menurut data pada Pusat Data dan Informasi Kementerian Kesehatan RI, melalui Infodatin Diabetes Melitus 2020, mengungkapkan estimasi yang mengkhawatirkan dari International Diabetes Federation (IDF).

Menurut laporan tersebut, Pada 2019, sekitar 463 juta individu berusia 20-79 tahun di seluruh dunia hidup dengan diabetes. Angka ini menunjukkan prevalensi yang signifikan dalam kelompok usia produktif. Prediksi untuk masa mendatang menggambarkan lonjakan yang signifikan. Estimasi menunjukkan bahwa pada tahun 2030, populasi yang hidup dengan diabetes diproyeksikan akan meningkat secara dramatis, mencapai angka sekitar 578 juta individu. tahun 2045, Angka ini diprediksi akan melonjak hingga 700 juta individu. Peningkatan yang pesat ini mengindikasikan bahwa diabetes akan menjadi tantangan kesehatan global yang semakin besar dalam beberapa dekade mendatang. Hal ini menekankan pentingnya upaya pencegahan, deteksi dini, dan manajemen yang efektif untuk mengatasi epidemi diabetes yang terus berkembang. Tren peningkatan ini menggambarkan betapa pentingnya upaya pencegahan dan pengelolaan diabetes sebagai masalah kesehatan global yang semakin mendesak[3].

Diagnosis dini dan akurat sangat penting untuk pengobatan diabetes yang efektif dan pencegahan komplikasi lebih lanjut. Metode diagnostik tradisional seperti tes darah dan urin dapat memakan banyak waktu dan sumber daya. Mengingat situasi ini, terdapat kebutuhan mendesak untuk mengembangkan metode yang lebih tepat guna dan presisi dalam memperkirakan risiko serta mendiagnosis diabetes. Pendekatan- presisi

dalam memperkirakan risiko serta mendiagnosis diabetes. Pendekatan-pendekatan baru ini diharapkan dapat meningkatkan efektivitas dalam penanganan dan pencegahan penyakit tersebut.

Tujuan dari penelitian ini adalah sebagai berikut:

a. Mengevaluasi seberapa baik algoritma K-Nearest Neighbor dalam memprediksi akurasi diagnosis penyakit diabetes dibandingkan dengan metode prediksi lainnya.

b. Menilai seberapa stabil algoritma K-Nearest Neighbor dalam memprediksi akurasi diagnosis penyakit diabetes ketika diterapkan pada berbagai dataset.

2. METODE PENELITIAN

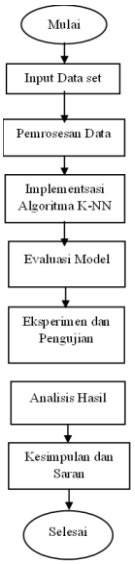
Algoritma K-Nearest Neighbor (KNN) adalah metode klasifikasi yang menentukan kategori suatu objek/topik baru beralaskan kedekatannya dengan beberapa objek lain yang sudah diketahui kategorinya (tetangga terdekat atau K). KNN merupakan bagian dari algoritma pembelajaran terawasi, di mana hasil klasifikasi dari instansi baru ditetapkan berlandaskan kategori yang terbanyak dari tetangga terdekatnya. Kategori yang paling sering muncul akan menjadi hasil klasifikasi untuk objek tersebut[5]. Metode Nearest Neighbor mencari data yang paling mirip dengan data baru dengan cara membandingkan ciri-ciri keduanya[6].

2.1 Metode Pengumpulan Data

Data merupakan bahan baku yang memerlukan proses pengolahan untuk diubah menjadi informasi yang bermakna. Informasi ini dapat berupa data kualitatif atau kuantitatif yang menggambarkan fakta-fakta tertentu, sehingga berguna dalam mendukung penelitian dan memberikan pemahaman yang lebih mendalam mengenai suatu fenomena[7].Peneliti menggunakan metode pengumpulan data untuk mendapatkan data penelitian. Angket, wawancara, pengamatan, dan dokumentasi adalah beberapa teknik yang umum digunakan.[8]. Sumber data yang digunakan peneliti adalah data public yaitu Pima Indians Diabetes Dataset. Kehamilan, glukosa, tekanan darah, ketebalan tubuh, insulin, BMI, riwayat diabetes keluarga, umur, dan kelas adalah semua variabel yang termasuk dalam data ini. Jumlah dataset yang digunakan yaitu 768 data.

2.2 Tahapan Penelitian

Penelitian ini secara sistematis menguraikan tahap demi tahap yang akan dilalui dalam upaya memprediksi akurasi penyakit diabetes dengan memanfaatkan algoritma KNN. Langkah-langkah dalam melakukan penelitian ini dapat dijelaskan sebagai berikut:



Gambar 1 Tahapan Penelitian

Berikut adalah pembahasan mengenai tahapan penelitian :

a. Input Data Set

Langkah pertama dalam penelitian ini adalah pengumpulan dataset yang relevan dan berkualitas untuk memprediksi penyakit diabetes. Dataset yang digunakan harus memiliki atribut yang sesuai dan representatif untuk mendukung proses analisis dan prediksi. Dataset yang digunakan dalam melakukan proses penelitian ini diperoleh dari sumber yang terpercaya, seperti Pima Indians Diabetes Database yang tersedia di UCI Machine Learning Repository. Dataset ini dipilih karena kelengkapan dan kualitas informasinya yang sesuai dengan kebutuhan penelitian. Dataset yang digunakan memiliki beberapa atribut penting yang berhubungan dengan faktor risiko diabetes, seperti jumlah kehamilan, glukosa, tekanan darah, ketebalan tubuh, insulin, BMI, resiko genetic terkena diabetes, umur dan hasil diagnosis (positif atau negatif untuk diabetes). Setiap atribut ini memainkan peran penting dalam memprediksi kemungkinan seseorang menderita diabetes.

b. Pemrosesan Data

Proses kedua dalam penelitian ini adalah pembersihan data, tahap ini mencakup identifikasi dan penanganan nilai yang hilang atau tidak valid dalam dataset. Nilai yang hilang dapat diisi dengan metode imputasi yang tepat atau dihapus jika jumlahnya minimal. Selanjutnya pada tahap ini, algoritma K-Nearest Neighbor (KNN) akan diimplementasikan untuk melakukan prediksi

penyakit diabetes berdasarkan dataset yang telah diproses sebelumnya.

c. Evaluasi Model

Setelah proses pelatihan, model diuji dengan data pengujian yang berbeda. Teknik validasi silang seperti cross-validation k-fold. memastikan hasil evaluasi yang lebih reliabel. Melalui teknik ini, kinerja model dapat dinilai secara komprehensif dan risiko overfitting dapat diminimalkan.Untuk mempertimbangkan hasil kinerja model, metrik seperti akurasi, presisi, recall, dan skor F1 digunakan. Hasil evaluasi ini akan memberikan gambaran yang jelas tentang kemampuan model KNN untuk memprediksi penyakit diabetes dan tingkat kehandalannya dalam praktiknya.

d. Eksperimen dan Pengujian

Pada saat ini, sejumlah eksperimen dan pengujian dilakukan untuk mengevaluasi akurasi algoritma K-Nearest Neighbor (K-NN) dalam mengidentifikasi penyakit diabetes.

e. Analisis Hasil

Informasi yang diperoleh dari penelitian ini, disajikan mulai dari dataset awal hingga hasil akhir prediksi penelitian. Data akhir disajikan dengan jumlah persentase dari hasil akhir pengujian.

f. Penarikan Kesimpulan

Pada fase ini, kesimpulan ditarik berdasarkan hasil prediksi data.yang telah dilakukan. Hasil penelitian presentase prediksi dan kesimpulan disusun berdasarkan dari data yang telah dikumpulkan, diolah, dan diuji terkait prediksi akurasi yang didapat.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Penelitian

Dalam penelitian ini menerapkan algoritma K-Nearest Neighbor (KNN) untuk mengkategorikan kasus diabetes melitus. Delapan faktor prediktor digunakan dalam klasifikasi ini: glukosa darah, tekanan darah diastolik, ketebalan lipatan kulit trisep, kadar insulin serum, indeks massa tubuh, riwayat diabetes dalam keluarga, dan usia. Dengan menggunakan variabel-variabel ini sebagai input, algoritma KNN berupaya untuk mengidentifikasi dan mengklasifikasikan kasus-kasus diabetes melitus secara akurat. Dalam penelitian ini, model data mining berbasis algoritma K-Nearest Neighbor (KNN) telah dikembangkan. Sebelum pemodelan, tiga metode (Simple Feature Scaling, min-max, dan z-score) digunakan untuk normalisasi data. Nilai K yang optimal untuk model dipilih berdasarkan hasil evaluasi kinerja model pada berbagai nilai K.

a. Pre Processing Data

Dataset Pima Indians Diabetes terdiri dari 768 sampel dengan 8 fitur dan 1 variabel target. Tabel 4.1 menunjukkan statistik deskriptif dari dataset sebelum preprocessing.

Train-Test Split

Dataset dibagi menjadi set pelatihan danpengujian dengan rasio 80:20:

Set Pelatihan: 614 sampel

Set Pengujian: 154 sampel

b. Hasil Pre processing Data

Setelah memeriksa dataset, ditemukan bahwa beberapa fitur memiliki nilai nol yang tidak valid secara medis. Tabel 4.2 menunjukkan jumlah nilai nol pada fitur-fitur tersebut sebelum preprocessing

Tabel 2. Jumlah nilai nol

Fitur	Jumlah nilai nol
Glucose	5
BloodPressure	35
SkinThickness	227
Insulin	374
BMI	11

Tabel tersebut menunjukkan jumlah nilai nol fitur dalam dataset. Fitur Insulin memiliki jumlah nilai nol terbanyak yaitu 374, diikuti oleh SkinThickness dengan 227 nilai nol. BloodPressure memiliki 35 nilai nol, BMI memiliki 11 nilai nol, dan Glucose hanya memiliki 5 nilai nol. Data ini mengindikasikan bahwa fitur Insulin dan SkinThickness memiliki banyak data yang hilang atau tidak terukur, sementara Glucose memiliki data yang paling lengkap di antara fitur-fitur tersebut

c. Imputasi Nilai yang Hilang

Setelah imputasi, dilakukan normalisasi fitur menggunakan StandardScaler.

Tabel 3. Statistik Deskriptif Setelah Normalisasi

Fitur	Mean	Std	Min	25%	50%	75%	Max
Pregnancies	0.0	1.0	-1.139	-0.844	-0.250	0.639	3.906
Glucose	0.0	1.0	-3.776	-0.684	-0.122	0.604	2.443
BloodPressure	0.0	1.0	-3.571	-0.367	-0.150	0.563	2.731
SkinThickness	0.0	1.0	-1.287	-1.287	-0.154	0.719	4.918
Insulin	0.0	1.0	-0.698	-0.693	-0.428	0.411	6.645
BMI	0.0	1.0	-4.057	-0.595	-0.001	0.584	4.452
DiabetesPedigreeFunction	0.0	1.0	-1.196	-0.690	-0.304	0.464	5.881
Age	0.0	1.0	-1.042	-0.785	-0.360	0.660	4.064

d. Hasil Analisis Korelasi

Setelah preprocessing data, dilakukan analisis korelasi untuk memahami hubungan antar variabel dalam dataset. Analisis ini penting untuk mengidentifikasifitur-fitur yang mungkin memiliki pengaruh signifikan terhadap outcome diabetes.

e. Interpretasi Hasil Korelasi

Korelasi dengan Outcome:

- Glucose menunjukkan korelasi positiftertinggi dengan Outcome (0.47), mengindikasikan bahwa tingkat glukosaadalah prediktor kuat untuk diabetes.
- BMI (0.29) dan Age (0.24) juga menunjukkan korelasi positif moderatdengan Outcome.
- Pregnancies (0.22) memiliki korelasi positif lemah dengan Outcome.

Korelasi antar Fitur:

- SkinThickness dan BMI menunjukkan korelasi positif kuat (0.66), yang mungkin mengindikasikan multikolinearitas.
- Pregnancies dan Age memiliki korelasi positif moderat (0.54), menunjukkan hubungan antara usia dan jumlah kehamilan.
- Insulin dan Glucose memiliki korelasi positif moderat (0.33), sesuai dengan pemahaman medis.

f. Feature Selection

Berdasarkan analisis korelasi danpentingnya medis, penulis memutuskan untuk mempertahankan semua fitur untuk model KNN.

g. Train test-split

Dataset dibagi menjadi set pelatihan dan pengujian dengan rasio 80:20:
Set Pelatihan: 614 sampel
Set Pengujian: 154 sampel

h. Kesimpulan pre processing :

- Nilai-nilai yang tidak valid telah ditangani.
- Data telah dinormalisasi untuk memastikan semua atribut memiliki skala yang sama.
- Analisis korelasi menunjukkan bahwa Glucose memiliki korelasi tertinggi dengan Outcome.
- Semua fitur dipertahankan untuk model KNN.

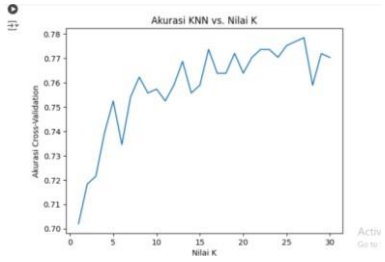
Dataset yang telah dipreprocessing ini siap untuk digunakan dalam implementasi algoritma KNN untuk prediksi penyakit diabetes.

3.2 Hasil Implementasi KNN

a. Pemilihan Nilai K Optimal

Setelah melakukan cross-validation dengan range nilai K dari 1 hingga 30, kami menemukan:

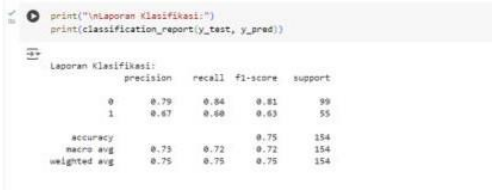
Nilai K optimal: 27



Gambar 2.Nilai K Optimal

3.3 Evaluasi Model

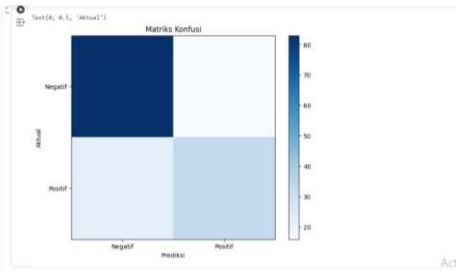
Menggunakan nilai K optimal (27), model KNN menghasilkan performa sebagai berikut:
Akurasi: 0.7532



```
print("\nLaporan Klasifikasi:")
print(classification_report(y_test, y_pred))

Laporan Klasifikasi:
              precision    recall  f1-score   support

           0       0.79      0.84      0.81        99
           1       0.67      0.60      0.63        55

    accuracy                           0.75       154
   macro avg       0.73      0.72      0.72       154
weighted avg       0.75      0.75      0.75       154
```

Gambar 3. Hasil Akurasi

Matriks Konfusi:
[83 15]
[24 32]



Gambar 4. Matriks Konfusi

3.4 Pembahasan

a. Performa Model

Model KNN dengan nilai K optimal 27 mencapai akurasi 75.32%. Ini menunjukkan bahwa model cukup baikdalam memprediksi diabetes, namun masih ada ruang untuk peningkatan.

b. Analisis Presisi dan Recall

Presisi untuk kelas positif(diabetes): 0.67
Recall untuk kelas positif: 0.57 Nilai presisi menunjukkan bahwa ketikamodel memprediksi seseorang memiliki diabetes, prediksi tersebut benar 67% dari waktu. Sementara itu, recall menunjukkan bahwa model berhasil mengidentifikasi57% dari semua kasus diabetes yangsebenarnya.

c. Interpretasi Matriks Konfusi

True Negatives (TN): 83

False Positives (FP): 15
False Negatives (FN): 24
True Positives (TP): 32

Model lebih baik dalam mengidentifikasi kasus negatif (non-diabetes) dibandingkandengan kasus positif (diabetes). Ini mungkin disebabkan olehketidakseimbangan kelas dalam dataset, di mana jumlah kasus non-diabetes lebih banyak daripada kasus diabetes.

d. Analisis Nilai K Optimal

Nilai K optimal yang ditemukan adala 27. Ini menunjukkan bahwa model memerlukan cukup banyak tetangga terdekat untuk membuat prediksi yang akurat. Hal ini mungkin mengindikasikan bahwa ada variabilitas yang cukup tinggi dalam dataset, dan menggunakan lebih banyak tetangga membantu dalam mengurangi pengaruh noise atau outlier.

4. KESIMPULAN

Algoritma K-Nearest Neighbor (KNN) digunakan untuk mengklasifikasikan diabetes melitus dengan melihat delapan indikator gejala. Ini termasuk jumlah kehamilan, glukosa, tekanan darah, ketebalan lipatan kulit, insulin, body mass index (BMI), fungsi silsilah diabetes, dan usia. Dataset yang digunakan terdiri dari 614 sampel untuk pelatihan dan 154 sampel untuk pengujian, yang telah melalui proses pembersihan dan normalisasi. Model KNN yang diimplementasikan menunjukkan performa yang cukup baik dalam memprediksi diabetes berdasarkan dataset Pima Indians. Dengan akurasi 75.32%, model ini dapat berfungsi sebagai alat skrining awal yang berguna. Namun, tingkat false negatives (24 kasus) menunjukkan bahwa model masih memiliki ruang untuk perbaikan, terutamadalam mengidentifikasi kasus diabetes yang sebenarnya.

5. DAFTAR PUSTAKA

[1] N. Maulidah, R. Supriyadi, D. Y. Utami, F. N. Hasan, A. Fauzi, and A. Christian, "Prediksi Penyakit Diabetes Melitus Menggunakan Metode Support Vector Machine dan Naive Bayes," *Indones. J. Softw. Eng.*, vol. 7, no. 1, pp. 63–68, 2021, doi: 10.31294/ijse.v7i1.10279.

[2] R. A. Siallagan and Fitriyani, "Prediksi Penyakit Diabetes Mellitus Menggunakan Algoritma C4.5," *J. Responsif Ris. Sains danInform.*, vol. 3, no. 1, pp. 44–52, 2021, doi: 10.51977/jti.v3i1.407.

[3] R. P. Kurniadi, R. R. Saedudin, and V. P. Widartha, "Perbandingan Akurasi Algoritma K-Nearest Neighbor Dan Logistic Regression Untuk Klasifikasi Penyakit Diabetes," *e- Proceeding Eng.*, vol. 8, no. 5, pp. 9757–9764, 2021.

[4] S. I. Fernanda, D. E. Ratnawati, and P. P. Adikara, "Identifikasi Penyakit Diabetes Mellitus Menggunakan Metode Modified K- Nearest Neighbor (MKNN)," *J. Pengemb. Teknol. Inf. dan Ilmu KomputerFernanda, S. I.,Ratnawati, D. E., Adikara, P. P. (2017). Identifikasi Penyakit Diabetes Mellit. Menggunakan Metod. Modif. K-Nearest Neighbor (MKNN). J. Pengemb. Teknol. Inf.* ,vol. 1, no. 6, pp. 507–513, 2017.

[5] M. Syukri Mustafa and I. Wayan Simpen, "Implementasi Algoritma K-Nearest Neighbor (KNN) Untuk Memprediksi Pasien Terkena Penyakit Diabetes Pada Puskesmas Manyampa. Kabupaten Bulukumba," *Februari*, vol. 2019,no. 1, pp. 1–10, 2019.

[6] E. W. Jumadi, "Penggunaan KNN (K-Nearest Neighbor) Untuk Klasifikasi Teks Berita yang Tak-Terkelompokkan pada Saat Pengklasteran Oleh STC (Suffix Tree Clustering)," *Istek*, vol.9, no. 1, pp. 50–81, 2015.

[7] M. Sholeh, D. Andayati, and R. Y.Rachmawati, "Data Mining Model Klasifikasi Menggunakan Algoritma K-Nearest Neighbor Dengan Normalisasi Untuk Prediksi Penyakit Diabetes," *TeIKa*, vol. 12, no. 02, pp. 77–87, 2022, doi: 10.36342/teika.v12i02.2911.

[8] M. Firdaus, "Intrumen Penelitian," *Metod. Penelit.*, pp. 15–20, 2010.

[9] M. F. Salim and S. Sugeng, "Analisis Rekam Medis Pasien Diabetes Mellitus Melalui Implementasi Teknik Data Mining di RSUP Dr. Sardjito Yogyakarta," *J. Kesehat. Vokasional*, vol. 2, no. 2, p. 167, 2018, doi: 10.22146/jkesvo.30331.

[10] M. Syukri Mustafa and I. Wayan Simpen, "Implementasi Algoritma K-Nearest Neighbor (KNN) Untuk Memprediksi Pasien Terkena Penyakit Diabetes Pada Puskesmas Manyampa Kabupaten Bulukumba," *Februari*, vol. 2019, no. 1, pp. 1–10, 2019.