SMATIKA : STIKI Informatika Jurnal
Vol. 14, No. 2, Desember 2024, pp. 297~303
ISSN: 2087-0256, e-ISSN: 2580-6939

Komparasi Metode K-Nearest Neighbor dan Naïve Bayes untuk Mengklasifikasi Resiko Diabetes

Rizki Alifia Safitri [1*]
Rahmatina Hidayati[2]

[1,2]Sistem Informasi, Universitas Merdeka Malang, Terusan Dieng No. 62-64, Klojen, Pisang Candi, Sukun, Jawa Timur, 65146, Indonesia
[1]alifiasafitri54139@gmail.com, [2] rahmatina.hidayati@unmer.ac.id

***Penulis Korespondensi:**
Rizki Alifia Safitri
alifiasafitri54139@gmail.com

Abstrak

Diabetes melitus termasuk masalah kesehatan dengan pertumbuhan paling cepat di abad ke-21. Salah satu penyebabnya karena kurangnya kesadaran Masyarakat untuk chek-up kesehatan secara berkala, sedangkan pola hidup yang dijalaini terbilang tidak sehat. Pemeriksaan hemoglobin A1c (HbA1c) sangat dianjurkan untuk mendeteksi diabetes. Tapi layanan tersebut belum ada di Posbindu Desa Bulupitu. Oleh karena itu perlu pendekatan lain untuk mendeteksi dini resiko Masyarakat terkena diabetes yakni dengan data mining. Metode data mining yang digunakan dalam penelitian ini adalah metode klasifikasi Naïve Bayes dan KNN. Variabel untuk menentukan resiko diabetes antara lain: jenis kelamin, usia, keturunan diabetes, sering kencing, Body Mass Index (BMI), kadar gula, dan output resiko diabetes. Pembagian dataset testing dan training menggunakan cross-validation dan rasion (60:40, 70:30, 80:20, dan 90:10). Akurasi terbaik metode Naïve Bayes diperoleh diperoleh dengan pembagian dataset cross-validation k-fold=2 sebesar 96,1%. Sedangkan pada metode KNN hasil terbaik diperoleh dari rasio dataset 80:30. Manhattan distance menjadi perhitungan jarak terbaik dalam penelitian ini dibandingkan dengan Euclidean distance dan Chebyshev distance.
Kata Kunci: Diabetes; KNN; Naïve Bayes

Abstract

Diabetes mellitus is one of the fastest-growing health problems in the 21st century. One of the causes is the lack of public awareness for regular health check-ups, while the lifestyle being led is quite unhealthy. Hemoglobin A1c (HbA1c) examination is highly recommended to detect diabetes. However, this service is not yet available at Posbindu in Bulupitu Village. Therefore, another approach is needed to detect the risk of diabetes early, namely through data mining. The data mining methods used in this research are the Naïve Bayes and kNN classification methods. The variables to determine the risk of diabetes include gender, age, family history of diabetes, frequent urination, Body Mass Index (BMI), blood sugar levels, and diabetes risk output. The division of testing and training datasets uses cross-validation and ratio (60:40, 70:30, 80:20, and 90:10). The best accuracy of the Naïve Bayes method was obtained by dividing the dataset using k-fold cross-validation with k=2, achieving 96.1%. In the kNN method, the best results were obtained from the 80:20 dataset ratio. Manhattan distance was found to be the best distance calculation in this study compared to Euclidean distance and Chebyshev distance.
Keywords: Diabetes; KNN; Naïve Bayes

1. Pendahuluan

Diabetes melitus adalah penyakit gangguan metabolisme yang ditandai dengan kenaikan kadar gula dalam darah. Penyakit ini tergolong berisiko, karena jika terjadi dalam jangka waktu yang panjang dapat menyebabkan kerusakan ginjal, sistem saraf, disfungsi mata, dan pembuluh darah [1]. Menurut data *The International Diabetes Federation*, diabetes termasuk salah satu masalah kesehatan yang mengalami pertumbuhan tercepat di abat ke-21. Di Indonesia prevalensi diabetes pada usia antara 20 sampai 79 tahun sekitar 10,6% yang menandakan 1 dari 9 orang terkena diabetes [2].

Di Pos Pembinaan Terpadu, Skrining Penyakit Tidak Menular (Posbindu) Desa Bulupitu pada kurun waktu Januari hingga Februari 2024, mencatat lebih dari 80 pasien mengidap diabetes. Sebagian dari pasien tersebut mengetahui dirinya terkena diabetes dalam kondisi yang lumayan

parah. Hal ini karena kurangnya kesadaran Masyarakat untuk *check-up* kesehatan secara berkala. Di Posbindu Desa Bulupitu Masyarakat hanya bisa mengecek gula darah. Sedangkan untuk medeteksi diabetes, perlu pemeriksaan hemoglobin A1c (*HbA1c*). Pemeriksaan ini memiliki kaitan terhadap kadar glukosa dalam darah pada penderita diabetes melitus [3]. Namun, layanan pengecekan *HbA1c* belum tersedia di Posbindu Desa Bulupitu, Kabupaten Malang.

Posbindu Desa Bulupitu memerlukan pendekatan lain untuk mendeteksi dini resiko Masyarakat terkena diabetes yakni dengan data mining. Metode data mining melibatkan penggunaan alat dan teknik untuk mengeksplorasi kumpulan data dan membantu penemuan pengetahuan [4]. Beberapa metode yang bisa digunakan untuk memprediksi kemungkinan diabetes antara lain: *random forest* [5], *support vector machine* [6], *K-Nearest Neighbor (KNN)* [7], dan *Naïve Bayes* [8]. Dalam penelitian ini penulis akan membandingkan metode KNN dan *Naïve Bayes* untuk memprediksi diabetes di Posbindu Desa Bulupitu.

Pada penelitian [7], peneliti menggunakan KNN untuk klasifikasi diabetes. Perhitungan jarang yang digunakan hanya *Euclidean distance*. Pada penelitian ini penulis akan membandingkan beberapa perhitungan jarak, antara lain *Euclidean distance*, *Mahanttan distance*, dan *Chebyshev distance*. Penelitian [9] membandingkan beberapa metode untuk mengklasifikasi indeks kedalaman kemiskinan provinsi Sulawesi Selatan. Hasil yang didapat metode KNN dan *Neural Network* menunjukkan performa paling baik.

Penelitian [10] mengklasifikasi diabetes melitus berdasarkan fakto-faktor penyebabnya. Faktor tersebut antara lain kehamilan, glukosa, BMI, dan usia. Pada penelitian [11] variabel yang digunakan untuk klasifikasi diabetes antara lain kehamilan, glukosa, tekanan darah diastolik, berat badan, umur, silsilah diabetes, ketebalan *triceps* pada lipatan kulit, dan serum insulin 2 jam. Dalam penelitian [12] terdapat 16 variabel untuk memprediksi diabetes. Penerapan metode Naïve Bayes dengan penambahan fitur selection menghasilkan 4 fitur tebaik yakni jenis kelamin, sering buang air kecil (*polyuria*), rasa haus yang berlebih (*polydipsia*), dan rambut rontok (*alopecia*). Perhitungan jarak pada kNN menjadi sangat penting untuk menentukan kedekatan antar data. Beberapa perhitungan jarak yang bisa digunakan antara lain *Euclidean, Chebyshev*, *Manhattan,* dan *Minkowski* [13].

Meskipun deteksi dini melalui sistem data mining ini dapat membantu mengidentifikasi risiko diabetes lebih awal, pemeriksaan *HbA1c* tetap perlu dilakukan untuk konfirmasi lebih lanjut. Deteksi dini hanya sebagai langkah awal untuk meningkatkan kesadaran masyarakat terhadap risiko diabetes dan mendorong mereka untuk melakukan pemeriksaan medis lebih mendalam, termasuk pengecekan *HbA1c*, yang memiliki akurasi lebih tinggi dalam diagnosis diabetes.

2. Metode Penelitian

Diabetes melitus merupakan penyakit gangguan metabolik yang diakibatkan pankreas memproduksi sedikit insulin. Diabetes juga dapat disebabkan oleh ketidakefektifan tubuh menggunakan insulin yang diproduksi [14]. Adapun model klasifikasi pada penelitian ini menggunakan *K-Nearest Neighbor* (*KNN*) dan *Naïve Bayes*.

K-Nearest Neighbor adalah metode klasifikasi terhadap objek berdasarkan data ketetanggaan (*neighbor*) yang memiliki jarak terdekat dengan objek tersebut. Perhitungan jarak yang akan digunakan dalam penetian ini antara lain [15]:

Euclidean distance

$$d_{euc}(x,y) = \sqrt{\sum_{j=1}^{d}(x_j - y_j)^2} \quad (1)$$

Manhattan distance

$$d_{man} = \sum_{j=1}^{d} |x_j - y_j| \quad (2)$$

Chebyshev distance

$$d_{che} = \max_{1 \le k \le d} |x_j - y_j| \quad (3)$$

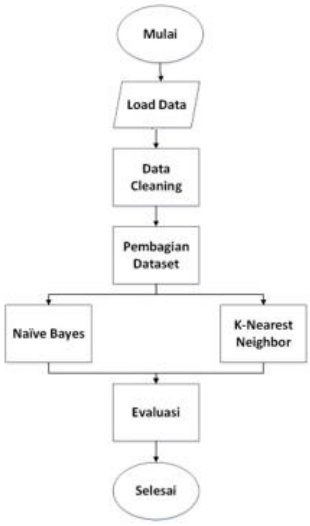
Naïve Bayes merupakan teknik untuk klasifikasi data dengan menggunakan prinsip teorema *Bayes*. Persamaan 4 menunjukkan rumus dari *Naïve Bayes* [16]:

$$P(H|X) = \frac{P(X|H) * P(H)}{P(X)} \quad (4)$$

Keterangan:

X	:	Data dengan kelas yang belum diketahui
H	:	Hipotesis data di kelas tertentu
P(H|X)	:	Probabilitas hipotesis H berdasarkan X *(probabilitas posteriori)*
P(H)	:	Probabilitas hipotesis H *(probabilitas prior)*
P(X|H)		Probabilitas X berdasarkan H *(likelihood probability)*

Berikut tahapan dalam penelitian ini:



Gambar 1. *Alur Penelitian*

Data yang digunakan dalam penelitian berupa data masyarakat yang berjumlah 100. Beberapa *dataset* ditampilkan pada Tabel 1. Data diperoleh dari (Posbindu) Desa Bulupitu. Varibael yang digunakan antara lain : jenis kelamin laki-laki/Perempuan, usia rentang 27 – 60 tahun, keturunan diabetes (ya/tidak), sering kencing (ya/tidak), *Body Mass Index* (*BMI*) (numerik), kadar gula (numerik) dan *output* resiko diabetes (ya/tidak).

Tabel 1*. Dataset*

No	Jenis Kelamin	Usia	Keturunan diabetes	Sering Kencing	*BMI*	Kadar Gula	Diabetes
1	P	45	Ya	Ya	27,6	241	Ya
2	L	35	Tidak	Tidak	20,5	138	Tidak
3	P	50	Ya	Tidak	30,6	201	Ya
:							
100	P	28	Tidak	Ya	28,7	218	Ya

Pada tahap *data cleaning* bertujuan untuk menghapus data yang rusak atau duplikat dan data yang memiliki variabel tidak lengkap. Pada pembagian *dataset* membagi dataset menjadi 2 yakni *training* dan *testing*. Proses pembagian data menggunakan beberapa model, antara lain (1) *Cross-validation* yaitu teknik validasi silang untuk membagi data menjadi k bagian set data (yang disebut *fold*) dengan sama ukuran [17]. (2) *Rasio training:testing* (60:40, 70:30, 80:20, dan 90:10). (3) Melakukan perbandingan klasifikasi *Naïve Bayes* dan *KNN*. Untuk metode *Naïve Bayes*, pemisahan data training dan testing menggunakan *cross-validation* dan rasio. Sedangkan pada metode *KNN* hanya menggunakan rasio. (4) Mengevaluasi hasil klasifikasi dengan *confusion matrix*.

Tabel 2. *Confusion Matrix*

		Predicted	
		Negatif	**Positif**
Actual	**Negatif**	TN	FP
	Positif	FN	TP

Sumber[18]

Confusion matrix terdiri dari 4 sel:

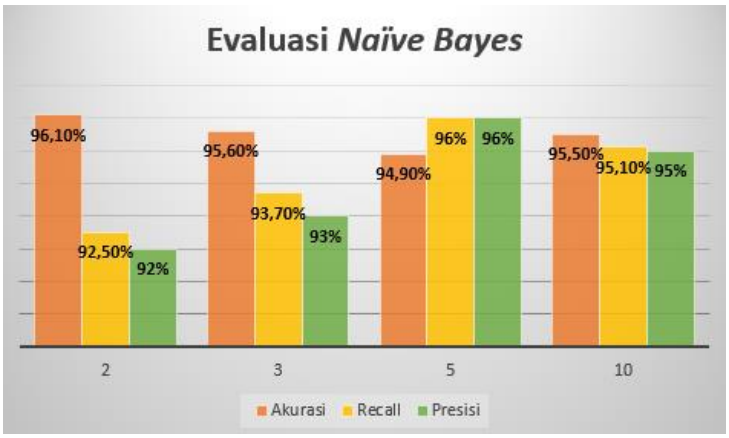
True Positive (TP): jumlah data *actual* positif dan dikenali sebagai positif
False Positive (FP): jumlah data *actual* negatif namun dikenali sebagai positif
True Negative (TN): jumlah data *actual* negatif dan dikenali sebagai negatif
False Negative (FN): jumlah data *actual* positif namun dikenali sebagai negatif

Perhitungan yang dilakukan berdasarkan *confusion matrix* antara lain: (1)Akurasi, yaitu menghitung berapa persen masyarakat yang benar diprediksi diabetes dan tidak dibetes dari keseluruhan data. (2) Presisi, yaitu menghitung berapa persen Masyarakat yang benar diabetes dari keseluruhan Masyarakat yang diprediksi diabetes. (3) *Recall,* perhitungan *recall* atau sensitivitas untuk mengetahui berapa persen masyarakat yang diprediksi diabetes dibandingkan keseluruhan masyarakat yang sebenarnya diabetes.

3. Hasil

Implementasi metode klasifikasi untuk memprediksi resiko diabetes Masyarakat Desa Bulupitu menggunakan aplikasi *Orange*. Pada *load* data *dataset* yang tersimpan dalam format *excel* dimasukkan ke dalam *orange.* Pada *data cleaning*, data yang digunakan tidak ditemukan duplikat data. Sedangkan untuk pengecekan variabel, semua data memiliki variabel yang lengkap. Sehingga, pada proses ini, tidak ada data yang dihapus karena sudah memenuhi syarat. Pada pembagian dataset berdasarkan *cross-validation* dan rasio *training:testing,* dan melakukan perbandingan metode.

Gambar 2 menampilkan hasil klasifikasi *Naïve Bayes* dengan *dataset cross-validation*. Pada nilai *k-fold* = 2 memiliki akurasi tertinggi 96,1% dibandingkan dengan *k-fold* = 3, *k-fold* = 5, dan *k-fold* = 10. Sedangkan *recall* dan presisi tertinggi berada pada *k-fold* = 5.



Gambar 2. *Evaluasi Naïve Bayes dengan Cross-Validation*

4. Pembahasan

Tabel 3 menampilkan hasil klasifikasi *Naïve Bayes* dengan pembagian dataset berdasarkan rasio. Akurasi tertinggi diperoleh dengan rasio data 70:30 dengan nilai 96%. Sedangkan presisi dan *recall* tertinggi berada di rasio 90:10.

Tabel 3. *Hasil Klasifikasi Naïve Bayes*

Rasio	Akurasi	Presisi	*Recall*
60:40	95,6%	93%	92,5%
70:30	96%	93,9%	93,3%
80:20	95,8%	93,8%	93%
90:10	94,8%	94,2%	94,1%

Dalam Tabel 4, nilai akurasi metode *KNN* yang diperoleh tidak jauh berbeda pada tiga perhitungan jarak. Dari percobaan 4 model rasio data *training:testing*, hasil terbaik berada pada model 80:20 di masing-masing jarak atau *distance*. Berdasarkan *Distance Eulidean* akurasi diperoleh 97,3%, sedangkan dengan *Distance Manhattan diperoleh* akurasi 97,4% dan pada *Distance Chebyshev* diperoleh akurasi 97,3%.

Tabel 4. *Hasil Klasifikasi Tertinggi Masing-Masing Distance*

Distance	Rasio	k	Akurasi
Euclidean	80:20	7	97,3%
Manhattan	80:20	7	97,4%
Chebyshev	80:20	5	97,3%

Evaluasi metode *Naive Bayes* dan *KNN* untuk mengklasifikasi resiko diabetes Masyarakat Desa Bulupitu menghasilkan *KNN* sebagai metode terbaik. Hal ini sesuai dengan penelitian yang dilakukan oleh[19]

5. Penutup

Berdasarkan hasil penelitian maka dapat disimpulkan bahwa performa terbaik metode *Naïve Bayes* diperoleh dengan pembagian *dataset cross-validation k-fold*=2. Sedangkan untuk *dataset* rasio, nilai tertinggi didapat dari model 70:30. Pada metode *KNN* hasil terbaik diperoleh dari rasio

dataset 80:30. Dan *Manhattan distance* menjadi perhitungan jarak terbaik dalam penelitian ini. Puskesmas dapat mempertimbangkan penggunaan model ini untuk mendeteksi dini penyakit diabetes. Jika hasil dari deteksi dini menyatakan warga tersebut berpotensi diabetes, maka puskemas bisa memberikan rekomendasi untuk pemeriksaan lebih lanjut.

Saran untuk penelitian selajutnya dapat menggunakan metode lain seperti *Random Forest* atau *Support Vector Machine*. Serta menambahkan fitur-fitur lain yang relevan dan signifikan dalam prediksi diabetes, seperti riwayat keluarga, gaya hidup, dan faktor genetik untuk meningkatkan akurasi prediksi.

Referensi

[1] D. Hardianto, "BIOTEKNOLOGI & BIOSAINS INDONESIA A Comprehensive Review of Diabetes Mellitus: Classification, Symptoms, Diagnosis, Prevention, and Treatment." [Online]. Available: http://ejurnal.bppt.go.id/index.php/JBBI

[2] M. Ratna Saraswati and I. Ngoerah, "Diabetes Melitus Adalah Masalah Kita," Kementerian Kesehatan Republik Indonesia. Accessed: May 15, 2024. [Online]. Available: https://yankes.kemkes.go.id/view_artikel/1131/diabetes-melitus-adalah-masalah-kita.

[3] S. Hartini, "Hubungan HBA1c Terhadap Kadar Glukosa Darah Pada Penderita Diabetes Mellitus Di RSUD. Abdul Wahab Syahranie Samarinda Tahun 2016," *Jurnal Husada Mahakam*, vol. IV, no. 3, pp. 171–180, 2016.

[4] L. Barreto Moreira and A. Amendoeira Namen, "A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia," *Comput Methods Programs Biomed*, pp. 139–149, 2018.

[5] W. Apriliah *et al.*, "Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest," *SISTEMASI:Jurnal Sistem Informasi*, vol. 10, no. 1, pp. 2540–9719, 2021, [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[6] A. Dwi Cahyani and A. Basuki, "Klasifikasi Diabetes Mellitus Menggunakan Support Vector Machine (Studi Kasus: Puskesmas Modopuro, Mojokerto)," *REKAYASA: Journal of Science and Technology*, vol. 12, no. 2, pp. 174–182, 2019.

[7] H. A. Dwi Fasnuari, H. Yuana, and M. T. Chulkamdi, "PENERAPAN ALGORITMA K-NEAREST NEIGHBOR UNTUK KLASIFIKASI PENYAKIT DIABETES MELITUS," *Antivirus : Jurnal Ilmiah Teknik Informatika*, vol. 16, no. 2, pp. 133–142, Oct. 2022, doi: 10.35457/antivirus.v16i2.2445.

[8] C. A. Rahayu, R. Hartono, and A. Sudiarjo, "Prediksi Penderita Diabetes Menggunakan Metode Naive Bayes," *JITET (Jurnal Informatika dan Teknik Elektro Terapan)*, vol. 11, no. 3, pp. 261–266, 2023.

[9] M. F. M. Khalik and F. Arifin, "Klasifikasi Indeks Kedalaman Kemiskinan Provinsi Sulawesi Selatan Berbasis Decision Tree, KNearest Neighbor, Naive Bayes, Neural Network, dan Random Forest," *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, vol. 9, no. 2, pp. 282–288, 2023.

[10] R. Putri Fadhillah *et al.*, "KLASIFIKASI PENYAKIT DIABETES MELLITUS BERDASARKAN FAKTOR-FAKTOR PENYEBAB DIABETES MENGGUNAKAN ALGORITMA C4.5," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 7, no. 4, pp. 1265–1270, 2022, [Online]. Available: www.kaggle.com

[11] P. Arsi and O. Somantri, "Deteksi Dini Penyakit Diabetes Menggunakan Algoritma Neural Network Berbasiskan Algoritma Genetika," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 3, no. 3, pp. 290–294, Oct. 2018, doi: 10.30591/jpit.v3i3.1008.

[12] F. Fitriyani, "Prediksi Diabetes Menggunakan Algoritma Naive Bayes dan Greedy Forward Selection," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 7, no. 2, pp. 61–69, Aug. 2021, doi: 10.25077/teknosi.v7i2.2021.61-69.

[13] W. I. N. P. Trisna, S. L. Sariwening, M. Fajar, and D. Wijayanto, "Perbandingan penghitungan jarak pada k-nearest neighbour dalam klasifikasi data tekstual," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 1, pp. 54–58, 2020.

[14] R. A. Siallagan and Fitriyani, "Prediksi Penyakit Diabetes Mellitus Menggunakan Algoritma C4.5," *JURNAL RESPONSIF*, vol. 3, no. 1, pp. 44–52, 2021.

[15] R. Hidayati, A. Zubair, A. H. Pratama, and L. Indana, "Analisis Silhouette Coefficient pada 6 Perhitungan Jarak K-Means Clustering," *Techno. Com*, vol. 20, no. 2, pp. 186–197, 2021.

[16] A. Zubair and M. Muksin, "Penerapan Metode Naive Bayes Untuk Klasifikasi Status Gizi (Studi Kasus Di Klinik Bromo Malang)," Malang, 2018.

[17] P. H. Azis, F. Fattah, and I. P. Putri, "Performa Klasifikasi K-NN dan Cross-validation pada Data Pasien Pengidap Penyakit Jantung," *ILKOM Jurnal Ilmiah* , vol. 12, no. 2, pp. 81–86, 2020.

[18] W. S. Hoar, A. Zubair, and L. Muflikhah, "Analisis sentimen kebijakan masuk sekolah pagi menggunakan algoritma Naïve Bayes," *Journal of Information System and Application Development (JISAD)*, vol. 2, no. 1, pp. 20–30, 2024.

[19] O. Nurdiawan, R. Herdiana, and S. Anwar, "Komparasi Algoritma Naïve Bayes dan Algoritma K-Nearst Neighbor terhadap Evaluasi Pembalajaran Daring," *SMATIKA JURNAL*, vol. 11, no. 02, pp. 126–135, Dec. 2021, doi: 10.32664/smatika.v11i02.621.