



Comparison of Classification Between Naive Bayes and K-Nearest Neighbor on Diabetes Risk in Pregnant Women

Perbandingan Klasifikasi Antara Naive Bayes dan K-Nearest Neighbor Terhadap Resiko Diabetes Pada Ibu Hamil

Bayu Delvika¹, Syahida Nurhidayarnis², Puji Dwi Rinada³,
Naufal Abror⁴, Assad Hidayat⁵

^{1,2,3,4}Program Studi Sistem Informasi, Fakultas Sains dan Teknologi,
UIN Sultan Syarif Kasim Riau, Indonesia
⁵Diskominfo Pekanbaru Indonesia

E-Mail: : ¹12050313245@students.uin-suska.sc.id ,
²12050313245@students.uin-suska.sc.id , ³12050321650@students.uin-suska.ac.id ,
⁴12050313965@students.uin-suska.ac.id , ⁵hidayatassad@gmail.com

Received Aug 30th 2021; Revised Sept 05th 2021; Accepted Sept 15th 2022
Corresponding Author: Naufal Abror

Abstract

Diabetes is also known as a chronic metabolic disease caused by the pancreas not producing enough insulin. During pregnancy, the body will increase hormones that also affect insulin work. In this study, the data used is the possible risk of pregnant women getting diabetes, the data is processed using data mining techniques, namely naive Bayes. After processing the data using the Naive Bayes data mining algorithm classification technique, the results obtained for data sharing using K-Fold Cross Validation $K = 10$ in the Naive Bayes algorithm obtained 75.78% results and for processing using knn with a value of $K = 25$ obtained yield 74.48%. From these results, Naive Bayes is better than K-Nearest Neighbor (KNN).

Keyword: Accuracy, Classification, Data Mining, Diabetes, K-Nearest Neighbor, Naive Bayes

Abstrak

Diabetes juga dikenal sebagai penyakit metabolik kronis yang dipicu karena pankreas tidak menghasilkan insulin yang cukup. Pada masa kehamilan, tubuh akan meningkatkan hormon-hormon yang juga mempengaruhi kerja insulin. Pada penelitian ini data yang digunakan adalah kemungkinan resiko ibu hamil terkena diabetes, data tersebut diolah memakai teknik data mining yaitu naive bayes. Setelah dilakukan pengolahan data memakai teknik klasifikasi data mining algoritma naive bayes, hasil yang didapat untuk pembagian data menggunakan K-Fold Cross Validation $K=10$ pada algoritma naive bayes didapat lah hasil 75,78% dan untuk pengolahan menggunakan knn dengan nilai $K=25$ didapat hasil 74,48%. Dari hasil tersebut naive bayes lebih baik dibandingkan K-Nearest Neighbor (KNN).

Kata Kunci: Akurasi, Data Mining, Diabetes, Klasifikasi, Naive Bayes, K-Nearest Neighbor.

1. PENDAHULUAN

Diabetes merupakan penyakit yang menyerang metabolisme tubuh dengan gejala hiperglikemia akibat efek sekresi insulin, aksi insulin, ataupun keduanya[1]. Diabetes atau disebut juga dengan diabetes mellitus (DM) adalah gangguan penyakit metabolisme jangka panjang yang disebabkan karena pankreas tidak menghasilkan cukup insulin atau insulin yang dihasilkan tidak dapat dipakai oleh tubuh secara baik. Diabetes mellitus merupakan penyakit komplikasi yang sering terjadi pada ibu hamil. Ada beberapa tipe diabetes mellitus diantaranya: tipe 1. (Bergantung pada pemakaian insulin). Tipe 2 (tidak bergantung pada insulin) dan diabetes mellitus gestasional[2].

Diabetes mellitus gestasional (DMG) ditemukan pada saat kehamilan yang belum didiagnosa diabetes pada ibu yang menunjukkan kadar glukosa tinggi pada saat kehamilan[3]. Penyebab pemicu nya yaitu adanya pergantian hormon. Pada kondisi kehamilan, plasenta akan menghasilkan hormon baru yaitu seperti human placental lactogen (HPL), hormon estrogen, dan hormon yang menaikkan resistensi insulin. Selama masa

kehamilan, akan meningkatnya hormon-hormon yang juga mempengaruhi kerja insulin[4]. Biasanya keadaan ini akan terjadi pada saat umur kehamilan mencapai 24 minggu, dan separuh kadar gula darah penderita kembali normal pasca persalinan[5]. Dengan begitu, penanganan secara tepat sangat penting untuk dilakukan. proses diagnosis dapat menerapkan teknik data mining.

Data mining adalah pencarian beberapa data baru yang mencari pola dari beberapa data dengan aturan tertentu dalam kuantitas yang besar, yang dapat diharapkan untuk mengatasi kondisi tersebut[6]. Data mining mempunyai beberapa fungsi yaitu: Fungsi estimasi (estimation), Fungsi deskripsi (description), Fungsi klasifikasi (classification), Fungsi prediksi (prediction), Fungsi pengelompokan (cluster), Fungsi asosiasi (association)[7]. Untuk melakukan teknik pengolahan data mining salah satunya menggunakan algoritma klasifikasi.

Klasifikasi adalah cara pengelompokan benda berdasarkan model yang dimiliki oleh objek klasifikasi[8]. Klasifikasi membangun model pada data pelatihan yang ada dan kemudian menggunakan model tersebut untuk mengklasifikasikan data baru. Klasifikasi dapat didefinisikan sebagai suatu kegiatan yang melatih/mempelajari suatu fungsi objektif yang mengaitkan setiap himpunan atribut (karakteristik) dengan sejumlah label kelas yang tersedia[9]. Algoritma klasifikasi memiliki beberapa macam algoritma yaitu, C4.5, K-means, Priori, Support Vector Machines, dan Naive bayes. Algoritma yang dipergunakan pada penelitian ini adalah Naive Bayes dan KNN.

Metode NBC telah banyak digunakan dalam penelitian mengenai data mining, beberapa kelebihan NBC diantaranya adalah algoritma sederhana tapi memiliki akurasi yang tinggi[10]. Naive bayes ialah suatu algoritma data mining yang termasuk kedalam salah satu klasifikasi data mining yang sering digunakan di antara algoritma lainnya. Naive Bayes adalah algoritma yang menggunakan teori probabilitas, untuk memprediksi probabilitas masa mendatang berdasarkan kejadian masa lalu[11]. Selain NBC, terdapat algoritma yang dapat digunakan untuk mengklasifikasikan data yaitu KNN.

KNN adalah salah satu metode klasifikasi pada data mining yang terbaik dan banyak digunakan[12]. 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[13]. Setelah mengetahui kelebihan dari kedua algoritma diatas, maka kedua algoritma tersebut dapat digunakan untuk mengklasifikasi data resiko diabetes terhadap ibu hamil.

Jumlah kasus penderita Diabetes Mellitus (DM) di dunia semakin mencemaskan. Berdasarkan pernyataan dari World Health Organization (WHO) banyak yang mengidap diabetes mellitus di seluruh dunia pada tahun 2000 telah mencapai 171 juta orang, dan diperkirakan di tahun 2030 banyaknya akan naik sebesar 114% atau mencapai 366 juta orang. Peningkatan jumlah pengidap DM terjadi di negara-negara maju dan 80% nya di negara yang berkembang, terutama Negara pertumbuhan ekonominya cepat[14]. Maka dari pada itu kedua algoritma ini dapat mengklasifikasikan resiko diabetes pada ibu hamil. Berdasarkan penjelasan diatas maka perlu dilakukannya penelitian untuk mengklasifikasikan ibu hamil yang terkena diabetes dengan mengimplementasikan algoritma naive bayes.

2. METODOLOGI PENELITIAN

Metodologi yang dipakai pada penelitian ini dapat dilihat pada gambar 1.

2.1 Deskripsi Metodologi

a. Data Collection

Data Collection merupakan suatu kegiatan yaitu mengumpulkan data yang digunakan. Pada paper ini, data didapat dari situs kaggle. Data yang digunakan dalam penelitian ini yaitu data diabetes terhadap Ibu hamil.

b. Data preprocessing

Data Preprocessing merupakan kegiatan untuk mempersiapkan data sebelum diproses menggunakan algoritma. Tahapannya adalah cleaning data, transformasi data dan normalisasi data.

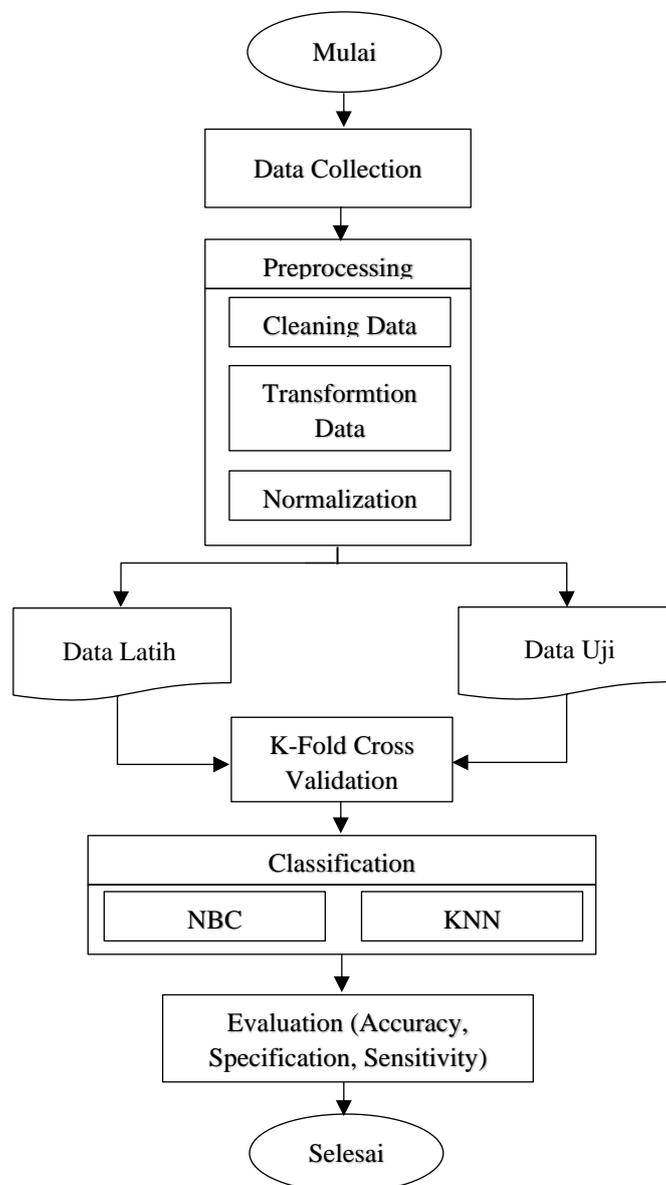
c. Pembagian Data

Pembagian Data merupakan teknik yang dipakai dalam membagi data menjadi data latih dan data uji. Teknik yang dipakai adalah *Hold Out* yaitu dengan membagi data latih sebesar 70% dan data uji sebesar 30%.

d. Pengolahan Menggunakan Naïve Bayes

Pengolahan menggunakan algoritma Naive Bayes merupakan kegiatan untuk mengolah data menggunakan algoritma naive bayes setelah data di processing.

- e. Evaluasi
Evaluasi merupakan kegiatan yang dilakukan untuk mencari nilai akurasi algoritma.
- f. K-Fold
Cross-validasi merupakan sebuah metode validasi model untuk menilai bagaimana hasil dari sebuah statistik analisis akan meyemarakakan kumpulan data independen[15].
- g. Rapid Miner
RapidMiner adalah sebuah perangkat lunak untuk mengelolah data. Dengan memakai prinsip algoritma data mining, RapidMiner mengekstrak pola-pola dari data set yang besar dengan menggabungkan metode statistika, kecerdasan buatan dan database[16].
- h. Pengolahan Menggunakan Naïve Bayes
Pengolahan menggunakan algoritma Naive Bayes merupakan kegiatan untuk mengolah data menggunakan algoritma naive bayes setelah data di processing.
- i. Evaluasi
Evaluasi merupakan kegiatan yang dilakukan untuk mencari nilai akurasi algoritma.



Gambar 1. Metodologi Penelitian

2.2 Naïve Bayes

Klasifikasi naïve bayes merupakan suatu algoritma di teknik data mining yang menggunakan teori Bayes untuk mengklasifikasi[17][18]. Naïve Bayes mempunyai tingkat kecepatan dan akurasi yang tinggi saat dipergunakan ke suatu wadah data dengan data yang cukup besar. Rumus Naïve Bayes:

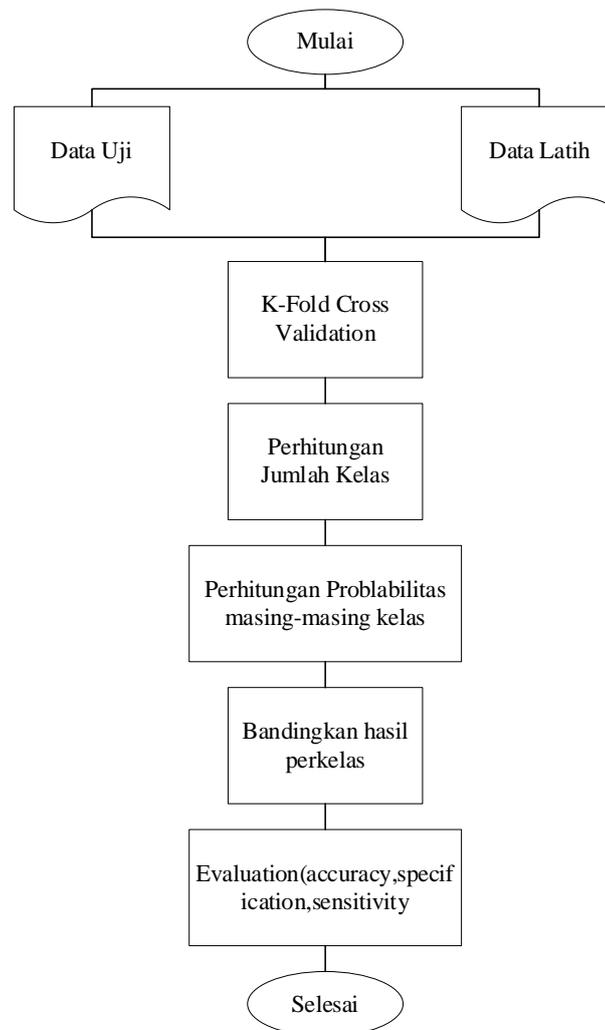
$$P(C_i|X) = \frac{P(X|C_i) \cdot P(C_i)}{P(X)} \quad (1)$$

Merupakan suatu cara untuk mengolah data dengan menetapkan kelas apa yang sesuai untuk contoh data yang dianalisis[19]. Naïve Bayes ialah salah satu metode klasifikasi yang sederhana tetapi memiliki kemampuan dan akurasi tinggi[20].

Dengan:

- X : class data yang tidak diketahui
- C_i : kemungkinan data X berada disuatu kelas tertentu
- P(C_i|X) : Peluang kemungkinan C_i sesuai keadaan X
- P(C_i) : Peluang kemungkinan C_i
- P(X|C_i) : Peluang X sesuai keadaan pada kemungkinan C_i
- P(X) : Peluang dari X

Gambar 2 merupakan alur processing menggunakan naïve bayes:

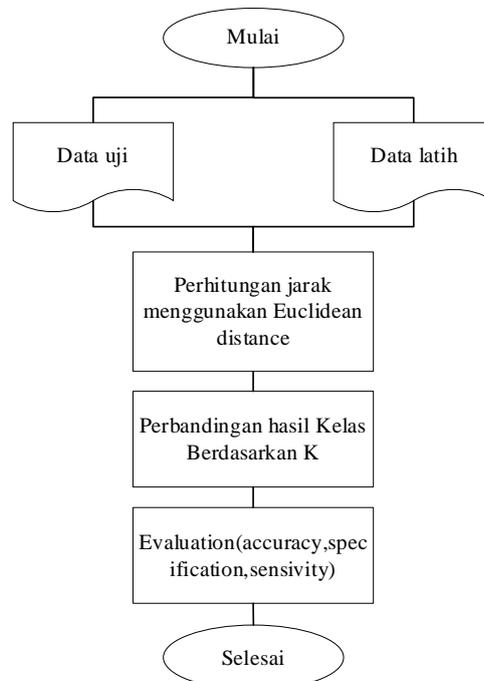


Gambar 2. Tahapan Naive Bayes

2.3 K-Nearest Nighbor

KNN adalah metode pengelompokan data untuk menentukan kategori berdasarkan sebagian besar kategori pada K-Nearest Neighbor [22], KNN dilakukan dengan mencari kelompok K objek pada data latih yang paling dekat dengan objek pada data baru atau data uji [21].

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (2)$$



Gambar 3. Flowchat K-NN

3. HASIL DAN PEMBAHASAN

Berikut ini adalah hasil dan pembahasan menggunakan Naïve Bayes dan K-Neares Neigbor.

3.1 Data

Data ini diambil dari website Kaggle.com. Data ini berjumlah 768.

Tabel 1. Dataset

No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	Diabetes Pedegree Function	Age	Outcome	Class
1	6	148	72	35	0	33,6	0,627	50	1
2	1	85	66	29	0	26,6	0,351	31	0
3	8	183	64	0	0	23,3	0,672	32	1
4	1	89	66	23	94	28,1	0,167	21	0
...
768	1	93	70	31	0	30,4	0,315	23	0

Pada data diatas telah dilakukan normalisai dengan atribut Glucose, Diabetes Pedegree Function, Blood Pressure, Skin Thickness, Pregnancies, Age, BMI, Insulin, Outcome. Penjelasannya adalah sebagai berikut:

- Pregnancies adalah Berapa kali hamil
- Glucose adalah konsentrasi plasma gula dalam 2 jam pada tes glukosa oral.
- Blood Pressure adalah Tekanan darah dengan satuan mm Hg.
- Skin Thickness Ketebalan lipatan kulit trisep dengan satuan mm.
- Insulin adalah Pemberian serum insulin setiap 2 jam dengan satuan mu U/ml.
- BMI (Body Mass Index) adalah pengukuran presentase lemak di dalam tubuh.
- Diabetes Pedegree Function adalah indikasi diabetes turunan.

- h) Age usia (Tahun).
- i) Outcome Variabel kelas (0 atau 1) dimana 1 merupakan kelas terindikasi terkena diabetes dan 0 merupakan kelas terindikasi tidak terkena diabetes.

3.2 Hasil Pada Naïve Bayes

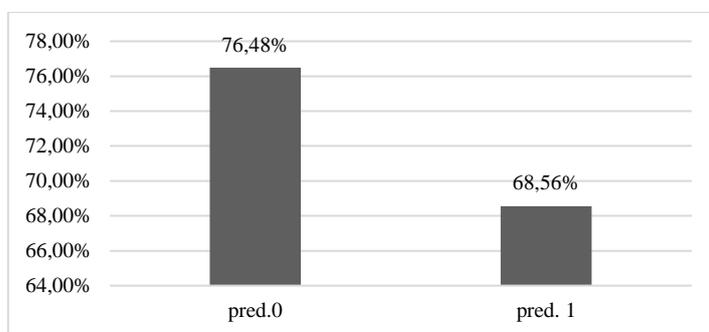
Berikut ini adalah hasil dan pembahasan pada proses pengolahan data menggunakan Naïve Bayes dan KNN, dimana data dibagi menggunakan Teknik K-Fold 10 K, dan dataset yang digunakan berjumlah 768 data yang sudah dilakukan proses cleaning, transformasi, dan normalisasi.

Tabel. 2 Hasil Percobaan Naive Bayes K-Fold K=10

Accuracy: 75.78% +/-5.44& (micro average: 75.78%)

	True 1	True 0	Class precision
Pred. 1	160	78	67.23%
Pred. 0	108	422	79.62%
Class recall	59.70	84.40%	

Tabel 2 merupakan hasil yang didapatkan dari perhitungan algoritma naïve bayes, menggunakan 10 K K-Fold maka didapatkan akurasi sebesar 75,78%. Class precision prediksi kelas 1 adalah 67,23% dan prediksi kelas 0 adalah 79,62%.



Gambar 3 Grafik Hasil Percobaan Naive Bayes K-Fold K=10

Gambar 4 merupakan diagram hasil percobaan pada setiap skenario proses koleksi data menggunakan naïve bayes dengan pembagian data menggunakan K-Fold sebanyak 10 K dengan hasil prediksi pada kelas 0 adalah 79,62% dan kelas 1 adalah 67,23%.

3.3 Hasil Pada K-Nearest Neighbor

Berikut hasil pengolahan algoritma K-Nearest Neighbor dengan K=25 menggunakan K-Fold dengan jumlah K=10, dimana penentuan kelas berdasarkan hitungan jarak terdekat data baru ke masing-masing data point di data latih.

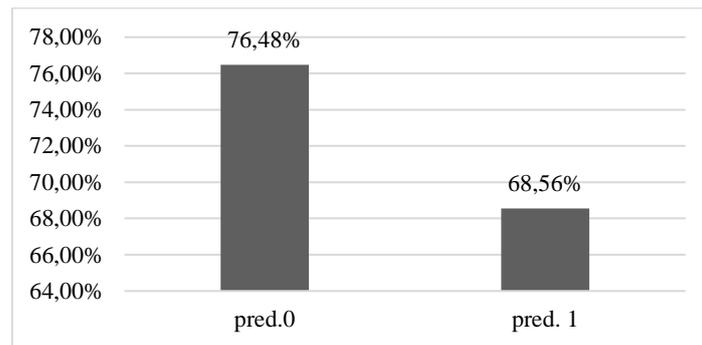
Tabel 3. Hasil Percobaan K-Nearest Neighbor K=25 menggunakan K-Fold K=10

Accuracy: 74.48% +/- 3.87% (micro average:74.48%)

	True 1	True 0	Class Precision
Pred. 1	133	61	68.56%
Pred. 0	135	439	76.48%
Class recall1	49.63%	87.80%	

Tabel 3 merupakan hasil yang didapatkan dari perhitungan algoritma KNN, setelah beberapa percobaan didapatlah akurasi tertinggi pada K=25 yaitu sebesar 74.48%, menggunakan K-Fold K=10 Class precision prediksi kelas 1 adalah 68,56% dan prediksi kelas 0 adalah 76,48%.

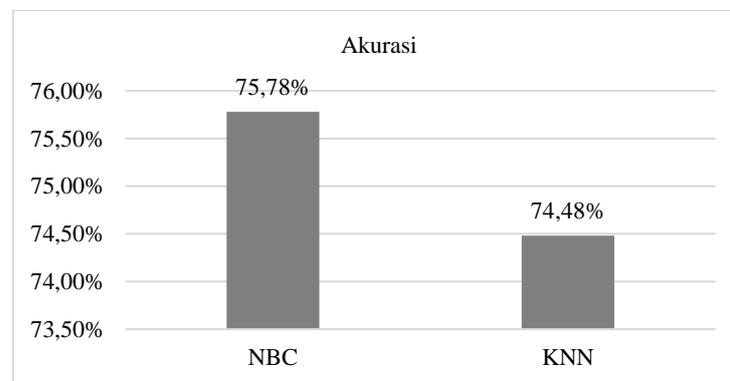
Gambar 5 merupakan diagram hasil percobaan pada setiap skenario proses koleksi data menggunakan KNN dengan pembagian data menggunakan K-Fold sebanyak 10 K dengan hasil prediksi pada kelas 0 merupakan label kelas yang tidak terkena diabetes dengan akurasi adalah 76,48% dan prediksi 1 adalah merupakan label kelas yang terkena diabetes dengan akurasi adalah 68,56%.



Gambar 4. Grafik Hasil Percobaan K-Nearest Neighbor K=25 menggunakan K-Fold K=10

3.4 Hasil Perbandingan NBC dan KNN

Setelah data diolah menggunakan kedua algoritma, maka langkah selanjutnya adalah membandingkan akurasi antara naïve bayes dan KNN untuk menemukan hasil yang terbaik. Berikut hasil perbandingannya, hasil tersebut dapat dilihat pada gambar 6.



Gambar 5. Hasil Perbandingan Akurasi NBC dan KNN

Dari gambar diatas dapat dilihat Naïve Bayes memiliki akurasi lebih tinggi dibandingkan KNN yaitu Naïve Bayes 75,78% sedangkan knn 74,48%. Dari hasil tersebut naïve bayes lebih baik dibandingkan knn terhadap data yang digunakan.

4. KESIMPULAN

Setelah menerapkan metode NBC dan KNN untuk mengklasifikasikan tingkat akurasi resiko ibu hamil yang terkena diabetes, diketahui bahwa kinerja dari algoritma NBC mempunyai hasil akurasi yang lebih tinggi. terbukti Setelah dilakukan pengolahan data menggunakan K-Fold Cross Validation K=10 dalam penerapan algoritma Naïve Bayes didapat lah hasil 75,78% dan untuk pengolahan menggunakan KNN dengan nilai K=25 didapat hasil 74,48%. Hal ini dipengaruhi jumlah data yang digunakan dan preprocessing yang dilakukan.

REFERENSI

- [1] F. Eryuda and T. U. Soleha, "Ekstrak Daun Kluwih (*Artocarpus camansi*) Dalam Menurunkan Kadar Glukosa Darah Pada Penderita Diabetes Melitus Kluwih Leaf Extract (*Artocarpus camansi*) In Lowering Blood Glucose Levels In Patients With Diabetes Melitus," *Majority*, vol. 5, no. 4, pp. 71–75, 2016.
- [2] M. D. M. Tito Putri, P. Wahjudi, and I. Prasetyowati, "Gambaran Kondisi Ibu Hamil dengan Diabetes Mellitus di RSD dr. Soebandi Jember Tahun 2013-2017," *Pustaka Kesehat.*, vol. 6, no. 1, p. 46, 2018.
- [3] H. Indeks, M. Tubuh, D. Resiko, L. Back, P. Pada, and P. Di, "Jurnal Keperawatan Muhammadiyah," vol. 7, no. 1, pp. 3–6, 2022.
- [4] M. Munawaroh and Hafizzurachman, "Konfirmasi Lima Faktor yang Berpengaruh terhadap Pencegahan Diabetes Mellitus pada Ibu Hamil 1Madinah," *J. Ilmu Kesehat.*, vol. 19, no. 2, pp. 34–38, 2018.
- [5] N. Djamaluddin and V. M. O. Mursalin, "Gambaran Diabetes Mellitus Gestasional Pada Ibu Hamil di RSUD Prof. Dr. H. Aloi Saboe Kota Gorontalo," *Jambura Nurs. J.*, vol. 2, no. 1, pp. 124–130, 2020, doi: 10.37311/jnj.v2i1.6858.

-
- [6] S. Haryati, A. Sudarsono, and E. Suryana, "Implementasi Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: Universitas Dehasen Bengkulu)," *J. Media Infotama*, vol. 11, no. 2, pp. 130–138, 2015.
- [7] S. Susanto, *PENGANTAR DATA MINING Menggenali Pengetahuan dan Bongkahan data*. 2010.
- [8] F. A. D. Aji Prasetya Wibawa, Muhammad Guntur Aji Purnama, Muhammad Fathony Akbar, "Metode-metode Klasifikasi," *Pros. Semin. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 1, p. 134, 2018.
- [9] D. P. Utomo and M. Mesran, "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung," *J. Media Inform. Budidarma*, vol. 4, no. 2, p. 437, 2020, doi: 10.30865/mib.v4i2.2080.
- [10] D. Ariadi and K. Fithriasari, "Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer," *J. SAINS DAN SENI ITS Vol. 4, No.2*, vol. 4, no. 2, pp. 248–253, 2015.
- [11] Y. Yuliana, P. Paradise, and K. Kusriani, "Sistem Pakar Diagnosa Penyakit Ispa Menggunakan Metode Naive Bayes Classifier Berbasis Web," *CSRID (Computer Sci. Res. Its Dev. Journal)*, vol. 10, no. 3, p. 127, 2021, doi: 10.22303/csr.10.3.2018.127-138.
- [12] Indriyanti, D. Sugianti, and M. A. Al Karomi, "Peningkatan Akurasi Algoritma KNN dengan Seleksi Fitur Gain Ratio untuk Klasifikasi Penyakit Diabetes Mellitus," *IC-Tech*, vol. 7, no. 2, pp. 1–6, 2017, [Online]. Available: <https://ejournal.stmik-wp.ac.id/index.php/ictech/article/view/3>
- [13] R. N. Whidhiasih, N. A. Wahanani, and Supriyanto, "Klasifikasi Buah Belimbing Berdasarkan Citra RED-GREEN-BLUE," *J. Penelit. Ilmu Komputer, Syst. Embed. Log.*, vol. 1, no. 1, pp. 29–35, 2013.
- [14] I. Luthfa, "Family Support in Patients of Type 2 Diabetes Mellitus Bangetayu Health Center in Semarang, Rasch Model Analysis," *Nurscope J. Penelit. dan Pemikir. Ilm. Keperawatan*, vol. 2, no. 1, p. 12, 2016, doi: 10.30659/nurscope.2.1.12-23.
- [15] F. Tempola, M. Muhammad, and A. Khairan, "Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, p. 577, 2018, doi: 10.25126/jtiik.201855983.
- [16] B. Rahmat C.T.I. et al., "Implementasi k-means clustering pada rapidminer untuk analisis daerah rawan kecelakaan," *Semin. Nas. Ris. Kuantitatif Terap.* 2017, no. April, pp. 58–60, 2017.
- [17] M. S. Mustafa, M. R. Ramadhan, and A. P. Thenata, "Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," *Creat. Inf. Technol. J.*, vol. 4, no. 2, p. 151, 2018, doi: 10.24076/citec.2017v4i2.106.
- [18] M. Syukri Mustafa, M. Rizky Ramadhan, and A. P. Thenata, "Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," *Citec J.*, vol. 4, no. 2, pp. 151–162, 2017.
- [19] H. Pramudia and A. Nugroho, "Sistem Informasi Kerusakan Laptop Menggunakan Metode Naive Bayes," *Teknol. Elektro, Univ. Mercu Buana*, vol. 8, no. 3, pp. 206–214, 2017.
- [20] H. F. Putro, R. T. Vulandari, and W. L. Y. Saptomo, "Penerapan Metode Naive Bayes Untuk Klasifikasi Pelanggan," *J. Teknol. Inf. dan Komun.*, vol. 8, no. 2, 2020, doi: 10.30646/tikomsin.v8i2.500.
- [21] M. Reza Noviansyah, T. Rismawan, and D. Marisa Midyanti, "Penerapan Data Mining Menggunakan Metode K-Nearest Neighbor Untuk Klasifikasi Indeks Cuaca Kebakaran Berdasarkan Data Aws (Automatic Weather Station) (Studi Kasus: Kabupaten Kubu Raya)," *J. Coding, Sist. Komput. Untan*, vol. 06, no. 2, pp. 48–56, 2018, [Online]. Available: <http://jurnal.untan.ac.id/index.php/jcskommipa/article/view/26672>
- [22] M. Mustakim, R. Hastarimasuci, P. Papilo dan A. Nazir, "Variable Selection to Determine Majors of Student using K-Nearest Neighbor and Naive Bayes Classifier Algorithm", *Journal of Physics: Conference Series*, vol. 1363, no. 1, 2019.
-