

Analisis Perbandingan Akurasi Algoritma Naïve Bayes dan C4.5 untuk Klasifikasi Diabetes

Mursyid Ardiansyah¹, Andi Sunyoto^{*2}, Emha Taufiq Luthfi³.

^{1,2,3} Program Studi Teknik Informatika, Universitas Amikom Yogyakarta
email: mursyid.27@students.amikom.ac.id¹, andi@amikom.ac.id^{*2},
emhataufiqluthfi@amikom.ac.id³

(Received: 7 Mei 2021 / Accepted: 12 Juli 2021 / Published Online: 20 Desember 2021)

Abstrak

Diabetes merupakan penyakit metabolic dimana gula darah meningkat tinggi. Apabila gula darah tidak dikontrol dengan baik dapat menyebabkan berbagai macam penyakit yang kritis salah satunya yaitu diabetes. Tujuan penelitian ini dilakukan untuk mengetahui hasil perbandingan nilai performa algoritma Naïve Bayes dan C4.5 dengan 7 skenario berbeda pada klasifikasi penyakit diabetes yang akan diuji performa accuracy, precision, dan recall. Metode yang digunakan pada penelitian ini adalah deskriptif, dan sumber data skunder diperoleh dari data pasien diabetes yang tersedia di Kaggle dengan format .csv terbitan Ishan Dutta sebanyak 520 data dan 17 bidang. Tool yang digunakan untuk analisis data adalah Rapidminer untuk proses klasifikasi dan pengujian performa algoritma Naïve Bayes dan Algoritma C4.5. Hasil penelitian kami menunjukkan bahwa algoritma C4.5 (skenario 4) memiliki hasil yang baik dalam klasifikasi penyakit diabetes dibandingkan algoritma Naïve Bayes (skenario 2) dimana performa algoritma C4.5 memiliki accuracy 99.03%, precision 100%, dan recall 98.18%.

Kata kunci: Diabetes, Klasifikasi, Naïve Bayes, Decision Tree, C4.5

Abstract

Diabetes is a metabolic disease in which blood sugar rises high. If blood sugar is not controlled properly, it can cause a variety of critical diseases, one of which is diabetes. The purpose of this study was to find out the results of comparing the performance values of Naïve Bayes and C4.5 algorithms with 7 different scenarios in the classification of diabetes that will be tested for accuracy, precision, and recall performance. The method used in this study is descriptive, and the source of skunder data obtained from the data of diabetic patients available on Kaggle with the format .csv issued by Ishan Dutta as many as 520 data and 17 fields. The tool used for data analysis is Rapidminer for the process of classification and performance testing of Naïve Bayes algorithm and C4.5 Algorithm. Our results showed that the C4.5 algorithm (scenario 4) had good results in the classification of diabetes compared to Naïve Bayes' algorithm (scenario 2) where the performance of the C4.5 algorithm had an accuracy of 99.03%, precision 100%, and recall 98.18%.

Keywords: Diabetes, Classification, Naïve Bayes, Decision Tree, C4.5

PENDAHULUAN

Diabetes adalah penyakit kronis yang ditandai dengan glukosa yang tinggi atau di atas normal. Diabetes adalah penyakit metabolik dimana Glukosa meningkat tinggi. Glukosa sangat penting bagi kesehatan manusia karena merupakan sumber energi penting untuk sel dan jaringan. Jika tidak dikontrol dengan baik, dapat menyebabkan penyakit kritis seperti jantung, obesitas, penyakit ginjal, stroke, dan gangguan mata, dan diabetes (Maulida, 2020).

Diabetes adalah penyakit paling umum pada manusia. Setiap tahun terjadi peningkatan angka kematian yang besar akibat penyakit ini. Menurut World Health Organization (WHO),

penderita diabetes mendekati 350 juta. WHO memprediksikan pada tahun 2030 diabetes akan menjadi salah satu penyebab utama kematian (Hairani et al., 2018).

Dalam penelitian sebelumnya mengenai penyakit diabetes telah banyak dilakukan penelitian seperti algoritma K-Nearest Neighbour(KNN) untuk klasifikasi penyakit diabetes (Indrayanti et al., 2017), diagnosa diabetes menggunakan Bayesian Regularization Neural Network (Rahman et al., 2017), klasifikasi penyakit diabetes menggunakan algoritma Backpropagation (Brian, 2017), perbandingan algoritma C4.5, KNN dan Naïve Bayes untuk klasifikasi penanggung jawab BSI Entrepreneur Center (Hasan et al., 2018) dan penelitian yang membandingkan beberapa algoritma data mining menggunakan dataset pasien diabetes yang tersedia di Kaggle dengan format .csv terbitan Ishan Dutta sebanyak 520 data dan 17 kolom. Sehingga penelitian yang akan dibuat menggunakan dataset penelitian sebelumnya yang akan dibagi menjadi beberapa skenario berbeda dan algoritma Naïve Bayes dan C4.5 sebagai algoritma yang akan dibandingkan.

Algoritma Naïve Bayes merupakan istilah sederhana untuk klasifikasi probabilistik berdasarkan penerapan Teorema Bayes, dimana algoritma Naïve Bayes memprediksi peluang yang akan terjadi dimasa yang akan datang berdasarkan pengalaman yang telah terjadi di masa lalu (Parthiban et al., 2011; Sari et al., 2020). Sedangkan algoritma C4.5 merupakan pengembangan dari algoritma pohon keputusan ID3. Algoritma C4.5 digunakan untuk melakukan klasifikasi prediktif (Hssina et al., 2014). C4.5 sering digunakan untuk data mining yang bertujuan untuk memvalidasi keakuratan data (Banu & Gomathy, 2014).

Penelitian (Gerhana et al., 2019) menggunakan algoritma Naïve Bayes dan C4.5 untuk memprediksi masa belajar siswa. Naïve Bayes Classifier memiliki akurasi 88% sedangkan algoritma C4.5 memiliki akurasi 87%. Kelemahan penelitian ini adalah mengujinya hanya dengan menggunakan akurasi dan tidak membagi data sedangkan penelitian yang akan dilakukan memabagi data menjadi 7 skenario. Gata et al., (2019) membandingkan 3 algoritma yaitu Naïve Bayes, Random Forest, dan C4.5. Naïve Bayes memiliki akurasi 70,13%, Random Forest memiliki akurasi 64,51%, dan C4.5 memiliki akurasi 67,58%. Hasil penelitian ini menyimpulkan bahwa Naïve Bayes memiliki tingkat akurasi terbaik disusul C4.5 kedua dan random forest terakhir. Dalam penelitian ini memiliki kelemahan yaitu hanya menguji akurasi.

Dengan membandingkan 2 algoritma dalam penelitian prediksi kelulusan mahasiswa (Azizah et al., 2018), yaitu algoritma Naïve Bayes dan C4.5. Pada penelitian ini diperoleh kesimpulan bahwa Naïve Bayes memiliki akurasi 63,8% dan C4.5 memiliki 63,6%, pada bagian presisi Naïve Bayes memiliki hasil 64,3% dan C4.5 memiliki presisi 63,8%, sedangkan pada bagian presisi recall section Naïve Bayes memiliki 96,3% dan C4.5 memiliki 98,3%. Dapat disimpulkan bahwa Naïve Bayes lebih unggul di bagian Accuracy and Precision, sedangkan Algoritma C4.5 lebih unggul di bagian Recall. Penelitian ini tidak menggunakan skenario sedangkan penelitian yang akan dibuat akan menggunakan 7 skenario berbeda.

Penerapan Naïve Bayes untuk mendeteksi penyakit jantung dalam sebuah penelitian (Vembandasamy et al., 2015) mampu memberikan hasil klasifikasi hingga 86,4198% benar dan 13,5802% salah. Kelemahan penelitian ini adalah Naïve Bayes merupakan algoritma terbaik untuk mendeteksi penyakit jantung sedangkan penelitian ini hanya membahas algoritma Naïve Bayes saja tanpa ada perbandingan dengan algoritma lain. Penelitian Boukenze et al., (2017) menggunakan beberapa algoritma yang dibandingkan, yaitu Naïve Bayes, C4.5, Support Vector Machine, dan K-Nearest Neighbours. Kesimpulan dari penelitian mereka adalah C4.5 memiliki akurasi tertinggi yaitu 63% dan error terendah 0,37 sedangkan K-NN tercepat yaitu 0,01 detik. Perbandingan beberapa algoritma dapat menjadi keuntungan dalam penelitian ini sehingga dapat diketahui algoritma mana yang cocok untuk penelitian ini.

Naïve Bayes berfungsi untuk mendiagnosis penyakit Alzheimer (Bhagya Shree & Sheshadri, 2018). Hasil temuan mereka menunjukkan bahwa hasil pengujian tingkat akurasi dari Naïve Bayes adalah 96,69%, *Precision* 0,976, *Recall* 0,967, dan *F-Measure* 0,969.

Penelitian ini menggunakan beberapa pengujian seperti akurasi, presisi, *recall*, dan F-measure dimana hal tersebut dapat menjadi nilai tambah. Kelemahan dari penelitian ini adalah beberapa gambar tidak dijelaskan sehingga pembaca kurang memahami gambar tersebut. Selain itu Amra & Maghari, (2017) membahas tentang prediksi kinerja siswa dengan menggunakan dua algoritma yaitu K-Nearest Neighbor dan Naïve Bayes. Penelitian ini menyimpulkan bahwa algoritma Naïve Bayes memiliki akurasi yang lebih tinggi yaitu 93,17%, *recall* 93,6%, dan presisi 94,65% dibandingkan K-NN.

Penelitian oleh Harisinghaney et al., (2014) membandingkan beberapa algoritma yaitu K-NN, Naïve Bayes, dan Reverse DBSCAN. Algoritma Naïve Bayes adalah yang terbaik dengan Precision 0,7446, Sensitivity 0,4962, Specificity 0,5163, dan *Accuracy* 0,4772. Penelitian Singh et al., (2017) membahas dampak dari tipe data yang berbeda terhadap kinerja klasifikasi menggunakan tiga algoritma yang dibandingkan, yaitu Random Forest, Naïve Bayes, K-Nearest Neighbours. Dari segi akurasi, K-NN memiliki nilai tertinggi yaitu 93,3 disusul *Random Forest* dan Naïve Bayes sebagai yang terendah. Namun pada bagian *Computation Time* (detik), Naïve Bayes memiliki waktu terpendek yaitu 0,124, sedangkan K-NN kedua dan Random Forset ketiga.

Perbandingan algoritma C4.5 dan K-NN dalam penelitian (Atma & Setyanto, 2018) untuk mengidentifikasi mahasiswa yang berpotensi *drop out*. Pada penelitian ini algoritma C4.5 memiliki akurasi sebesar 96,28% sedangkan K-NN memiliki akurasi sebesar 95,07% sehingga algoritma C4.5 merupakan yang terbaik dalam penelitian. Pada perbandingan performansi algoritma C4.5 dan CART (Rahmayuni, 2014), algoritma C4.5 menghasilkan akurasi sebesar 85,61% yang merupakan yang terbaik dibandingkan algoritma CART yang memiliki akurasi 84,95%. Menurut penelitian ini, hal ini terjadi karena algoritma C4.5 cocok untuk data non-numerik. Selanjutnya, penelitian yang dilakukan oleh Agustina & Wijanarto, (2016) membahas perbandingan algoritma ID3 dan C4.5 untuk klasifikasi. Pada pengujian diambil data sebesar 10% dari 90% dari total data 1473. Algoritma ID3 memiliki akurasi sebesar 99,33% sedangkan algoritma C4.5 memiliki akurasi sebesar 99,40% yang menjadikan algoritma C4.5 sebagai yang terbaik. algoritma dalam penelitian ini.

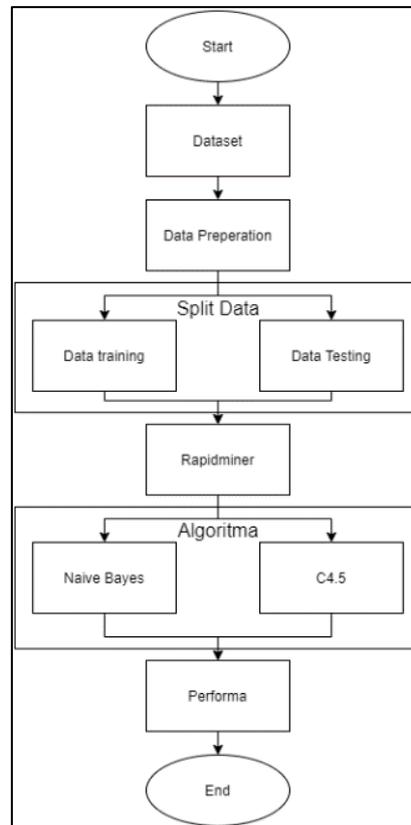
Berdasarkan beberapa penelitian di atas, algoritma Naïve Bayes dan C4.5 memiliki hasil yang baik sehingga tujuan dari penelitian ini adalah membandingkan dua algoritma dalam klasifikasi diabetes yaitu algoritma Naïve Bayes dan C4.5 dengan 7 skenario berbeda untuk mengetahui algoritma yang mana memiliki performa terbaik menggunakan pengujian *Accuracy*, *Precision*, dan *Recall*.

METODE

Metode yang digunakan dalam penelitian ini meliputi pengumpulan dataset, pembagian dataset menjadi data latih dan data pengujian, model klasifikasi yang menerapkan algoritma Naïve Bayes dan Algoritma C4.5, serta pengujian dengan melihat hasil klasifikasi menggunakan confusion matrix. Alur penelitian berikut dapat dilihat pada Gambar 1. Pada Gambar 1 dapat dilihat urutan alur penelitian mulai dari dataset, penyiapan data, membagi data menjadi dua yaitu Data Training dan Data Pengujian, Penerapan Algoritma Naïve Bayes dan C4.5 pada Rapidminer, dan Pengujian Akurasi. Precision, dan Recall dengan melihat hasil klasifikasi menggunakan Confusion Matrix.

Dataset diambil dari data pasien diabetes yang tersedia di Kaggle dengan format .csv terbitan Ishan Dutta sebanyak 520 data dan 17 bidang / kolom. Dimana dataset akan dibagi menjadi dua bagian yaitu data latih, dan data uji dan dibuat menjadi 7 skenario berbeda. Data Preparation adalah mengubah data mentah menjadi data yang dapat dianalisis, dari 17 field yang akan digunakan adalah 16 field, age, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle, stiffness, alopecia, obesity, dan class. Data training digunakan untuk proses pelatihan

algoritma Naïve Bayes dan C4.5 dengan menggunakan 20% atau 103 data dari total data yang ada di dataset. Data testing digunakan untuk menguji algoritma Naïve Bayes dan C4.5 menggunakan 80% atau 417 data dari total data pada dataset. Selanjutnya data training dan data testing dibagi menjadi 7 skenario dimana 7 skenario.



Gambar 1. Alur penelitian

Rapidminer adalah *software* yang digunakan untuk melakukan pemrosesan data mining (Uska et al., 2020). Perangkat lunak ini merupakan solusi untuk menganalisis data mining, text mining, dan analisis prediktif (Fitriani, 2020). Lakukan validasi terpisah untuk membagi dataset menjadi data training dan data testing. Penerapan algoritma Naïve Bayes dan C4.5 yang sudah tersedia pada software atau tools Rapidminer kemudian melakukan penghitungan pada dataset yang tersedia. Selain perhitungan, Rapidminer juga digunakan untuk melakukan tes untuk algoritma Naïve Bayes dan C4.5.

Naïve Bayes adalah salah satu algoritma yang paling efisien dan efektif untuk pembelajaran mesin dan penambahan data yang menggunakan perhitungan probabilitas dan statistik yang diusulkan oleh ilmuwan Inggris Thomas Bayes (Zhang, 2005). Pada tahap ini algoritma diimplementasikan pada dataset penelitian ke dalam sistem yang mengimplementasikan Naïve Bayes.

Algoritma C4.5 merupakan pengembangan dari algoritma ID3 dimana salah satu algoritma yang digunakan untuk mengklasifikasikan berupa pohon keputusan (Rahim et al., 2018). Tahap ini mengimplementasikan algoritma pada dataset yang ada ke dalam sistem yang mengimplementasikan C4.5. Pada tahap performa algoritma dibandingkan jumlah akurasi dan jumlah error pada algoritma Naïve Bayes dan C4.5 menggunakan confusion matrix. Untuk pengujian algoritma menggunakan Precision, Recall, dan Accuracy sehingga dapat diketahui algoritma mana yang terbaik dari kedua algoritma tersebut (Harianto et al., 2020; Pramadhana, 2021). Dalam *Machine Learning*, terutama masalah klasifikasi statistik, confusion matrix juga dikenal sebagai matriks kesalahan, tata letak tabel yang memungkinkan visualisasi kinerja

suatu algoritma. Terdapat empat representasi dari proses klasifikasi confusion matrix, yaitu *true positif*, *true negatif*, *false positif*, dan *false negatif* (Powers, 2020).

HASIL DAN PEMBAHASAN

Hasil

Pada bagian hasil dan pembahasan akan dilakukan pengujian algoritma Naïve Bayes dan C4.5 dengan 7 skenario berbeda selanjutnya akan dilakukan pengujian performa algoritma Naïve Bayes dan C4.5 berdasarkan hasil accuracy, precision, dan recall sesuai hasil klasifikasi pada confusion matrix. Untuk melakukan pengujian algoritma Naïve Bayes dan C4.5 sebelumnya dataset dibagi menjadi 7 skenario dimana setiap skenario memiliki atribut-atribut yang berbeda yang akan hitung menggunakan tools rapidminer. Skenario yang telah dibuat dapat dilihat pada tabel 1.

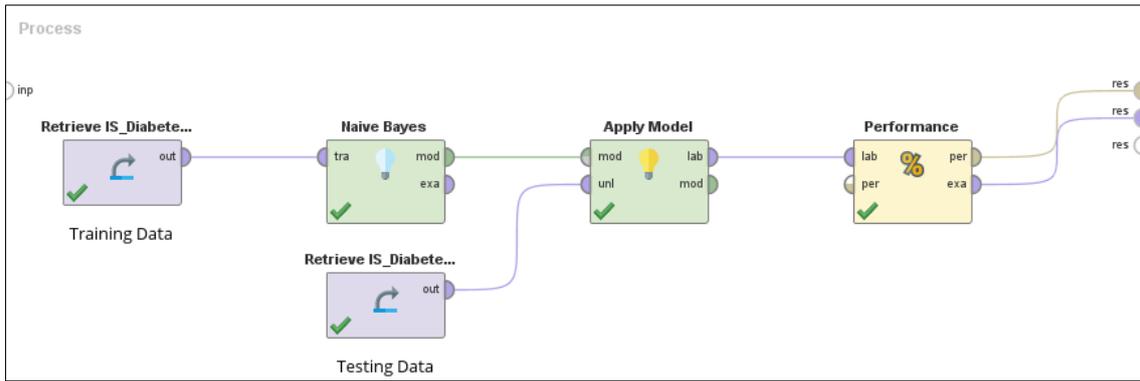
Tabel 1. Skenario

	Atribut
Skenario 1	age, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity, dan class
Skenario 2	polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity, dan class
Skenario 3	sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity, dan class
Skenario 4	polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity, dan class
Skenario 5	visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity, dan class
Skenario 6	irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity, dan class
Skenario 7	partial paresis, muscle stiffness, alopecia, obesity, dan class

Tabel 1 menampilkan skenario yang telah dibuat dimana pada Skenario 1 terdapat 16 atribut yang lengkap, Skenario 2 dikurangi 1 atribut, Skenario 3 dikurangi 3 atribut, Skenario 4 dikurangi 5 atribut, Skenario 5 dikurangi 7 atribut, Skenario 6 dikurangi 9 atribut, dan Skenario 7 dikurangi 11 atribut. Setiap Skenario diuji performanya baik akurasi, presisi, dan recall kemudian membandingkan performa algoritma Naïve Bayes dan C4.5.

Perhitungan algoritma Naïve Bayes pada software Rapidminer dapat dilihat pada gambar 2. Gambar 2 merupakan perhitungan menggunakan model algoritma Naïve Bayes dengan menggunakan tools Rapidminer. Proses pertama adalah penginputan data training, kemudian masuk ke model Naïve Bayes, setelah itu masuk ke Apply Model untuk menghubungkan data testing dengan model Naïve Bayes, dan yang terakhir adalah performa untuk menguji algoritma Naïve Bayes.

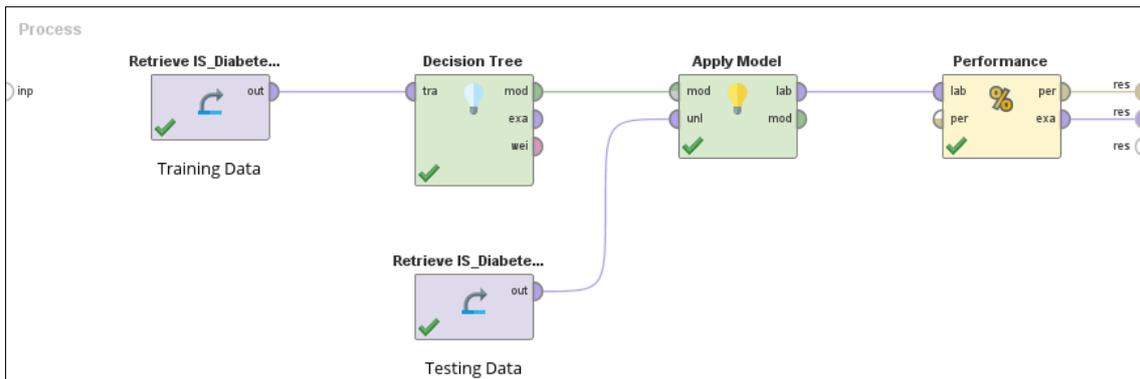
Performa algoritma Naïve Bayes dengan 7 skenario berbeda menggunakan tools Rapidminer diperoleh hasil seperti tabel 2. Pada Tabel 2 dapat dilihat performa setiap skenario menggunakan algoritma Naïve Bayes dimana skenario 2 memiliki hasil accuracy dan precision terbaik sedangkan skenario 7 memiliki nilai recall terbaik.



Gambar 2. Perhitungan algoritma Naïve Bayes

Tabel 2. Performa 7 skenario algoritma Naïve Bayes

	Accuracy	Precision	Recall
Skenario 1	87,38%	90,38%	85,45%
Skenario 2	88,35%	92,16%	85,45%
Skenario 3	79,61%	78,33%	85,45%
Skenario 4	77,67%	76,67%	83,64%
Skenario 5	77,67%	78,57%	80%
Skenario 6	78,64%	78,95%	81,82%
Skenario 7	65,05%	62,34%	87,27%



Gambar 3. Perhitungan algoritma C4.5

Tabel 3. Performa 7 skenario algoritma C4.5

	Accuracy	Precision	Recall
Skenario 1	95,15%	96,30%	94,55%
Skenario 2	97,09%	100%	94,55%
Skenario 3	98,06%	100%	96,36%
Skenario 4	99,03%	100%	98,18%
Skenario 5	91,26%	92,59%	90,91%
Skenario 6	89,32%	95,83%	83,64%
Skenario 7	81,55%	90,91%	72,73%

Gambar 3 menunjukkan perhitungan menggunakan algoritma C4.5 menggunakan tools Rapidminer. Hal pertama yang dilakukan adalah memasukkan data training kemudian masuk ke *Decision Tree* yang mewakili C4.5 setelah itu pada bagian *Apply Model* menerapkan model C4.5 dengan data testing untuk melakukan pengujian, dan terakhir performa untuk mengetahui hasil pengujian algoritma C4.5.

Performa algoritma C4.5 dengan 7 skenario berbeda menggunakan tools Rapidminer diperoleh hasil seperti tabel 3. Hasil performa setiap skenario menggunakan algoritma C4.5 dimana skenario 4 memiliki hasil *accuracy* dan *recall* terbaik sedangkan skenario 2, skenario 3 dan skenario 4 memiliki hasil *precision* terbaik.

Pembahasan

Berdasarkan hasil pengujian algoritma Naïve Bayes dan C4.5 dengan 7 skenario menggunakan tools rapidminer dapat dilihat perbandingan Accuracy, Precision, dan Recall pada tabel 4. hasil perhitungan dari algoritma Naïve Bayes dan C4.5 dengan 7 skenario menggunakan tools rapidminer dimana untuk algoritma Naive Bayes skenario 2 adalah skenario terbaik dengan hasil *accuracy* 88,35%, *precision* 92,16%, dan *recall* 85,45%. Untuk algoritma C4.5 skenario 4 adalah yang terbaik dengan hasil *accuracy* 99,03%, *precision* 100%, dan *recall* 98,18%. Masing-masing algoritma tersebut baik algoritma Naïve Bayes maupun Algoritma C4.5 memiliki hasil yang baik dibandingkan dengan dataset awal atau skenario 1.

Tabel 4. Perbandingan Performa Naïve Bayes dan C4.5

		Accuracy	Precision	Recall
Skenario 1	Naïve Bayes	87,38%	90,38%	85,45%
	C4.5	95,15%	96,30%	94,55%
Skenario 2	Naïve Bayes	88,35%	92,16%	85,45%
	C4.5	97,09%	100%	94,55%
Skenario 3	Naïve Bayes	79,61%	78,33%	85,45%
	C4.5	98,06%	100%	96,36%
Skenario 4	Naïve Bayes	77,67%	76,67%	83,64%
	C4.5	99,03%	100%	98,18%
Skenario 5	Naïve Bayes	77,67%	78,57%	80%
	C4.5	91,26%	92,59%	90,91%
Skenario 6	Naïve Bayes	78,64%	78,95%	81,82%
	C4.5	89,32%	95,83%	83,64%
Skenario 7	Naïve Bayes	65,05%	62,34%	87,27%
	C4.5	81,55%	90,91%	72,73%

Untuk perbandingan performa algoritma Naïve Bayes dan C4.5 dari hasil perhitungan yang telah dilakukan dapat diambil kesimpulan bahwa algoritma C4.5 memiliki performa terbaik karena algoritma C4.5 memiliki tingkat kesalahan yang kecil ketika proses perhitungan hal tersebut menunjukkan bahwa dalam melakukan klasifikasi penyakit diabetes algoritma C4.5 adalah yang terbaik.

Hasil perbandingan tersebut berbeda dengan penelitian yang dilakukan (Hasan et al., 2018) yang membuktikan algoritma Naïve Bayes adalah yang terbaik dibandingkan C4.5. Hal tersebut dapat dipengaruhi oleh studi kasus yang berbeda, atribut yang digunakan, dan jumlah data yang dikelolah. Pada penelitian ini menggunakan dataset diabetes sedangkan penelitian sebelumnya menggunakan dataset penanggung jawab jawab BSI, penelitian ini membagi data menjadi 7 skenario untuk algoritma Naïve Bayes dan C4.5 sedangkan penelitian sebelumnya tidak membagi data.

SIMPULAN

Pada penelitian ini dibandingkan dua algoritma data mining, yaitu algoritma Naïve Bayes dan C4.5 dengan 7 skenario menggunakan tools rapidminer untuk melakukan klasifikasi

diabetes. Dataset diambil dari situs kaggle.com. Data tersebut terbagi menjadi dua data yaitu data latih dan data uji dengan 7 skenario berbeda. Kemudian dilakukan pengujian menggunakan algoritma Naïve Bayes dan C4.5 menggunakan tools Rapidminer dimana tool tersebut menyediakan model klasifikasi Naïve Bayes dan C4.5 serta pengujian performa untuk algoritma tersebut. Dari hasil penelitian perbandingan performa algoritma Naïve Bayes dan C4.5 untuk klasifikasi penyakit diabetes yang telah dilakukan dapat diambil kesimpulan bahwa algoritma C4.5 memiliki nilai performa yang baik dibandingkan algoritma Naïve Bayes dimana algoritma C4.5 unggul di setiap skenario dengan skenario 4 sebagai skenario yang tinggi dengan hasil *accuracy* 99,03%, *precision* 100%, dan *recall* 98,18% dimana 3,88% peningkatan *accuracy* dari *accuracy* skenario 1.

REFERENSI

- Agustina, D. melina, & Wijanarto. (2016). Analisis Perbandingan Algoritma ID3 Dan C4 . 5 Untuk Klasifikasi Penerima Hibah Pemasangan Air Minum pada PDAM Kabupaten Kendal. *Journal of Applied Intelligent System*, 1(3), 234–244.
- Amra, I. A. A., & Maghari, A. Y. A. (2017). Students performance prediction using KNN and Naïve Bayesian. *ICIT 2017 - 8th International Conference on Information Technology, Proceedings*, 909–913. <https://doi.org/10.1109/ICITECH.2017.8079967>
- Atma, Y. D., & Setyanto, A. (2018). Perbandingan algoritma c4.5 dan k-nn dalam identifikasi mahasiswa berpotensi drop out. *Metik Jurnal*, 2(2), 31–37.
- Azizah, E. N., Pujiyanto, U., Nugraha, E., & Darusalam. (2018). Comparative performance between C4.5 and Naive Bayes classifiers in predicting student academic performance in a Virtual Learning Environment. *2018 4th International Conference on Education and Technology, ICET 2018, 1*, 18–22. <https://doi.org/10.1109/ICEAT.2018.8693928>
- Banu, M. N., & Gomathy, B. (2014). Disease forecasting system using data mining methods. *International Conference on Intelligent Computing Applications, ICICA 2014*, 130–133. <https://doi.org/10.1109/ICICA.2014.36>
- Bhagya Shree, S. R., & Sheshadri, H. S. (2018). Diagnosis of Alzheimer’s disease using Naive Bayesian Classifier. *Neural Computing and Applications*, 29(1), 123–132. <https://doi.org/10.1007/s00521-016-2416-3>
- Boukenze, B., Haqiq, A., & Mousannif, H. (2017). Predicting chronic kidney failure disease using data mining techniques. *Lecture Notes in Electrical Engineering*, 397, 701–712. https://doi.org/10.1007/978-981-10-1627-1_55
- Brian, T. (2017). Analisis Learning Rates Pada Algoritma Backpropagation Untuk Klasifikasi Penyakit Diabetes. *Edutic - Scientific Journal of Informatics Education*, 3(1), 21–27. <https://doi.org/10.21107/edutic.v3i1.2557>
- Fitriani, E. (2020). Perbandingan Algoritma C4.5 Dan Naïve Bayes Untuk Menentukan Kelayakan Penerima Bantuan Program Keluarga Harapan. *Sistemasi*, 9(1), 103. <https://doi.org/10.32520/stmsi.v9i1.596>
- Gata, W., Basri, H., Hidayat, R., Patras, Y. E., Baharuddin, B., Fatmasari, R., Tohari, S., & Wardhani, N. K. (2019). *Algorithm Implementations Naïve Bayes, Random Forest. C4.5 on Online Gaming for Learning Achievement Predictions*. 258(Icream 2018). <https://doi.org/10.2991/icream-18.2019.1>
- Gerhana, Y. A., Fallah, I., Zulfikar, W. B., Maylawati, D. S., & Ramdhani, M. A. (2019). Comparison of naive Bayes classifier and C4.5 algorithms in predicting student study period. *Journal of Physics: Conference Series*, 1280(2). <https://doi.org/10.1088/1742-6596/1280/2/022022>
- Hairani, H., Nugraha, G. S., Abdillah, M. N., & Innuddin, M. (2018). Komparasi akurasi metode correlated naive Bayes classifier dan naive Bayes classifier untuk diagnosis penyakit diabetes. *InfoTekJar: Jurnal Nasional Informatika Dan Teknologi Jaringan*,

3(1), 6–11.

- Hariato et al. (2020). Optimasi Algoritma Naïve Bayes Classifier untuk Mendeteksi Anomaly dengan Univariate Fitur Selection. *Edumatic: Jurnal Pendidikan Informatika*, 4(2), 40–49. <https://doi.org/10.29408/edumatic.v4i2.2433>
- Harisinghaney, A., Dixit, A., Gupta, S., & Arora, A. (2014). Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm. *ICROIT 2014 - Proceedings of the 2014 International Conference on Reliability, Optimization and Information Technology*, 153–155. <https://doi.org/10.1109/ICROIT.2014.6798302>
- Hasan, F. N., Hikmah, N., & Utami, D. Y. (2018). Perbandingan Algoritma C4.5, KNN, dan Naive Bayes untuk Penentuan Model Klasifikasi Penanggung jawab BSI Entrepreneur Center. *Jurnal Pilar Nusa Mandiri*, 14(2), 169. <https://doi.org/10.33480/pilar.v14i2.908>
- Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4.5. *International Journal of Advanced Computer Science and Applications*, 4(2), 13–19. <https://doi.org/10.14569/specialissue.2014.040203>
- Indrayanti, Sugianti, D., & Al Karomi, M. A. (2017). Optimasi Parameter K pada Algoritma K-Nearest Neighbour untuk Klasifikasi Penyakit Diabetes Mellitus. *Prosiding SNATIF*, 823–829.
- Maulida, A. (2020). Penerapan Metode Klasifikasi K-Nearest Neighbor pada Dataset Penderita Penyakit Diabetes. *Indonesian Journal of Data and Science*, 1(2), 29–33.
- Parthiban, G., S.K.Srivatsa, A., & Rajesh, A. (2011). Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method. *International Journal of Computer Applications*, 24(3), 7–11. <https://doi.org/10.5120/2933-3887>
- Powers, D. M. W. (2020). *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*. January 2008.
- Pramadhana, D. (2021). Klasifikasi Penyakit Diabetes Menggunakan Metode CFS Dan ROS dengan Algoritma J48 Berbasis Adaboost. *Edumatic: Jurnal Pendidikan Informatika*, 5(1), 89–98.
- Rahim, R., Zufria, I., Kurniasih, N., Simargolang, M. Y., Hasibuan, A., Sutiksno, D. U., Nanuru, R. F., Anamofa, J. N., Ahmar, A. S., & Achmad Daengs, G. S. (2018). C4.5 classification data mining for inventory control. *International Journal of Engineering and Technology(UAE)*, 7, 68–72. <https://doi.org/10.14419/ijet.v7i2.3.12618>
- Rahman, M. F., Alamsah, D., Darmawidjadjaja, M. I., & Nurma, I. (2017). Klasifikasi Untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (RBNN). *Jurnal Informatika*, 11(1), 36. <https://doi.org/10.26555/jifo.v11i1.a5452>
- Rahmayuni, I. (2014). Perbandingan performansi algoritma c4.5 dan cart dalam klasifikasi data nilai mahasiswa prodi teknik komputer politeknik negeri padang. *Teknoif*, 2(1), 40–46.
- Sari, V. R., Firdausi, F., & Azhar, Y. (2020). Perbandingan Prediksi Kualitas Kopi Arabika dengan Menggunakan Algoritma SGD, Random Forest dan Naive Bayes. *Edumatic: Jurnal Pendidikan Informatika*, 4(2), 1–9. <https://doi.org/10.29408/edumatic.v4i2.2202>
- Singh, A., N., M., & Lakshmganthan, R. (2017). Impact of Different Data Types on Classifier Performance of Random Forest, Naïve Bayes, and K-Nearest Neighbors Algorithms. *International Journal of Advanced Computer Science and Applications*, 8(12), 1–11. <https://doi.org/10.14569/ijacsa.2017.081201>
- Uska, M., Wirasmita, R., Usuluddin, U., & Arianti, B. (2020). Evaluation of Rapidminer-Application in Data Mining Learning using PeRSIVA Model. *Edumatic: Jurnal Pendidikan Informatika*, 4(2), 164–171. <https://doi.org/10.29408/edumatic.v4i2.2688>
- Vembandasamy, K., Sasipriya, R., & Deepa, E. (2015). Heart Diseases Detection Using Naive Bayes Algorithm. *International Journal of Innovative Science, Engineering & Technology*, 2(9), 441–444.
- Zhang, H. (2005). Exploring conditions for the optimality of naïve bayes. *International Journal*

of Pattern Recognition and Artificial Intelligence, 19(2), 183–198.
<https://doi.org/10.1142/S0218001405003983>