Volume 19 Nomor 2 (2024) 155-164
E – Link P-ISSN 1858-2109 E-ISSN 2656-5676
Jurnal Teknik Elektro dan Informatika

PERBANDINGAN PERFORMA ALGORITMA KLASIFIKASI C4.5 DAN NAÏVE BAYES UNTUK PREDIKSI DIAGNOSA PENYAKIT DIABETES

Febri Basufi Bahtiarullah [1] , Ahmad Homaidi [2] , Firman Santoso [3]
[1,2,3] Jurusan Teknologi Informasi, Fakultas Sains dan Teknologi, Universitas Ibrahimy
Jl. KHR. Syamsul Arifin No.1-2, Sukorejo, Situbondo 68374, Jawa Timur, Indonesia
E-mail : [1]febri.basufi@gmail.com, [2]ahmadhomaidi@ibrahimy.ac.id,
[3]firman4bi@gmail.com

ABSTRAK

Di antara penyakit yang semakin meningkat di seluruh dunia, diabetes adalah masalah besar bagi kesehatan masyarakat. Untuk mencegah komplikasi serius yang dapat muncul akibat penyakit ini, deteksi dini dan intervensi yang tepat sangat penting. Penelitian ini bertujuan untuk menganalisa perbandingan performa algoritma C4.5 dan algoritma Naïve Bayes dalam mendeteksi penyakit diabetes . Untuk mengukur akurasi dan efektivitas masing-masing model, algoritma klasifikasi digunakan untuk pengumpulan data, pemrosesan, dan penerapan. Hasil analisis menunjukkan bahwa algoritma Naïve Bayes memiliki akurasi 87,50%, dan C4.5 memiliki akurasi 84,62%. Dalam hal prediksi diagnosis diabetes, Naive Bayes lebih baik daripada C4.5, menurut temuan ini . Meskipun Naïve Bayes tampak lebih baik dalam hal akurasi, kedua algoritma memiliki keunggulan masing-masing, yang harus dipertimbangkan dalam aplikasi klinis.. Studi ini memberikan wawasan berharga bagi tenaga kesehatan saat memilih diagnosis diabetes yang tepat. Oleh karena itu, diharapkan bahwa penelitian ini akan berkontribusi pada pengembangan sistem diagnosis yang lebih efektif untuk mendeteksi risiko diabetes.

Kata Kunci: C4.5, Diabetes, Klasifikasi, Naïve Bayes.

ABSTRACT

Among the increasing diseases around the world, diabetes is a major public health problem. To prevent serious complications that may arise as a result of this disease, early detection and proper intervention are essential. This study aims to analyse the comparison of the performance of the C4.5 algorithm and the Naïve Bayes algority in the detection of diabetes. To measure the accuracy and effectiveness of each model, classification algorithms are used for data collection, processing, and application. The results of the analysis showed that the Naïve Bayes algorithm has an accuracy of 87.50%, and the C4.5 has an exactitude of 84.62%. In terms of predicting the diagnosis of diabetes, Naive Bayes is better than C4.5, according to this finding. The study provides valuable insights to health professionals when choosing the right diabetes diagnosis. Therefore, it is expected that this research will contribute to the development of more effective diagnostic systems to detect the risk of diabetes.

Keywords: *C4.5, Diabetes, Classification, Naïve Bayes*.

I. PENDAHULUAN

1.1. Latar Belakang Penelitian

Setiap penyakit memiliki tanda-tanda dan gejala, kemudian gejala dari penyakit akan tersajikan dalam bentuk data-data. Lantas dari data tersebut sangat dibutuhkan oleh orang-orang yang berprofesi dibidang kesehatan untuk menentukan diagnosa penyakit yang diderita pasien yang sedang berobat. Lantaran penyakit ini dapat menyebabkan resiko kematian terhadap pasien, maka sangat diperlukan ketelitian dan kedetilan serta kecepatan dalam

melakukan diagnosa dini yang didasarkan pada data-data gejala penyakit yang tersajikan, agar terhindar dari kesalahan yang dapat menyebakan kerugian dan bahkan bisa menyebabkan kematian pada pasien yang berobat.

Salah satu penyakit yang menjadi penyebab kamatian tertinggi di Indonesia adalah penyakit diabetes. Hiperglikemia kronis yang disebabkan oleh kekurangan insulin absolut atau relatif, gangguan kerja insulin, atau peningkatan produksi glukosa dikenal sebagai diabetes melitus[1]. Menurut Internasional Diabetes Federation (IDF), Di antara negara-negara di dunia dengan jumlah penderita diabetes tertinggi, Dengan 19,5 juta penderita di tahun 2021, Indonesia berada di peringkat kelima, dan diperkirakan akan mencapai 28,6 juta pada tahun 2045 [2]. Namun, laporan survei yang dilakukan Kemenkes, yaitu survei (SKI) tahun 2023, yang dirilis oleh Kementerian Kesehatan, Menunjukkan bahwa prevalensi diabetes mellitus (DM) lebih tinggi di kalangan individu berusia di atas 15 tahun. Prevalensi diabetes di Indonesia pada tahun 2018 adalah 10,9%, menurut Riset Kesehatan Dasar (Riskesdas), tetapi pada tahun 2023 meningkat menjadi 11,7%[3].

Merujuk pada data tersebut, maka dapat ditarik kesimpulan bahwa penanganan penyakit diabetes ini sangat dibutuhkan untuk bisa memberikan dampak positif terhadap penanggulangan dan penurunan angka penderita diabetes di Indonesia. Lantas, agar penyakit diabetes dapat dilakukan diagnosa dengan tepat, maka perlu dilakukan klasifikasi data gejala secara cepat dan akurat, untuk melakukan hal tersebut tentu juga dibutuhkan data-data yang valid dengan metode serta model klasifikasi yang handal, agar kesalahan-kesalahan setiap proses klasifikasi data yang dilakukan dapat terminimalisir. Tentunya data-data tersebut harus dilakukan pemrosesan agar dapat diterjemahkan menjadi sebuah output diagnosa.

Dalam konteks ini, model data mining dengan metode klasifikasi berperan sangat penting untuk meprediksi diagnosa penyakit diabetes secara akurat dan efisien. Tentunya juga akan memberikan kemudahan kepada tenaga kesehatan selaku petugas yang berwenang melayani pasien yang berobat. Penggunaan teknologi data mining merupakan industri yang terus berkembang, dengan banyak aplikasi baru yang sedang dikembangkan. Data mining akan menjadi sangat penting untuk membuat keputusan yang lebih akurat tentang diabetes seiring dengan jumlah data yang dikumpulkan.

Metode komparasi performa algoritma C4.5 dan naïve bayes digunakan untuk pemodelan data mingin pada penelitian ini, serta dataset yang diperoleh dari situs Repositori Pembelajaran Mesin UC Irvine. Pemilihan kedua algoritma tersebut, bertujuan untuk membandingkan seberapa akurat serta efisien kedua algoritma tersebut dalam memprediksi penyakit diabetes. Untuk menguji kedua algoritma tersebut kedalam model data mining klasfikasi, kami menggunakan tool Rapidminer. Muara akhir pada penelitian ini adalah untuk menawarkan solusi bagi tenaga kesehatan untuk meningkatkan sistem kerja yang dapat diandalkan yang berbasis data mining. Ini juga diharapkan akan mempengaruhi jumlah hasil diagnosa diabetes yang tepat.

1.2. Tinjauan Penelitian

Melihat penelitian sebelumnya oleh Nurfazriah Attamami, Agung Triayudi, dan Rima Tamara Aldisa tentang analisis perbandingan prediksi penerima bantuan jaminan kesehatan dengan algoritma klasifikasi Naive Bayes dan C4.5. Tujuan dari penelitian ini adalah untuk membantu petugas dinas kesehatan menentukan seberapa layak orang yang memiliki masalah kesejahteraan sosial untuk menerima jaminan kesehatan sosial. Temuan dari penelitian ini mengungkapkan bahwa algoritma C4.5 berhasil mencapai tingkat akurasi sebesar 99,04%, menjadikannya yang tertinggi di antara algoritma yang diuji. Sebagai perbandingan, algoritma Naive Bayes hanya mencatatkan akurasi sebesar 92,97%.[4].

Dengan menggunakan model, karakteristik, dan hasil penelitian ini, peneliti disini memilih untuk membandingkan performa algoritma klasifikasi C4.5 dan Naïve Bayes untuk prediksi diagnosa penyakit diabetes. Penelitian sebelumnya, yang juga berkaitan dengan penelitian ini, menekankan bahwa komparasi

pemodelan analisa performa antara algoritma Naïve Bayes dan C4.5 diperlukan untuk memprediksi diagnosa diabetes. Sehingga diharapkan dapat menghasilkan model pengetahuan penggunaan algoritma klasifikasi data yang tepat dan efisien dalam memprediksi diagnosa penyakit diabetes yang akan berdampak pada kemudahan petugas kesehatan dalam melakukan diagnosa penyakit diabetes.

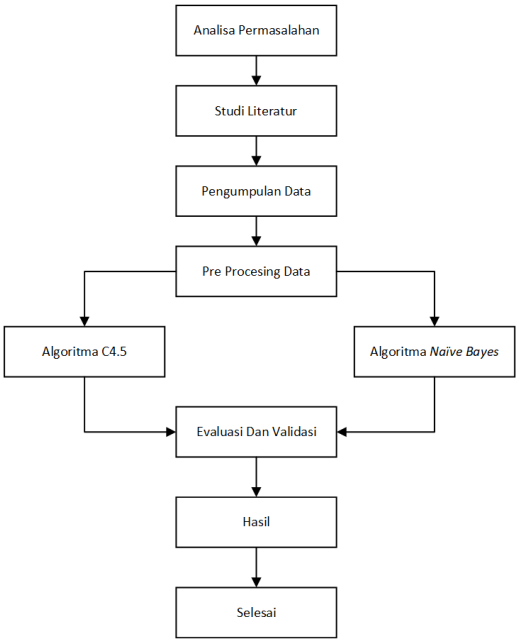
1.3. Landasan Teori

a. Diabetes
   Tubuh mengalami gangguan metabolisme yang disebut diabetes, yang menyebabkan peningkatan kadar gula dalam darah. Beberapa faktor menyebabkan peningkatan kadar gula ini, termasuk kekurangan insulin dan resistansi insulin[1].
b. Data Mining
   Proses penggalian data yang menggunakan kecerdasan buatan, matematika, dan statistika untuk menemukan pola, tren, atau anomali dalam data dikenal sebagai data mining [5].
c. Klasifikasi
   Data mining klasifikasi adalah proses mengkategorikan data ke dalam kategori yang berbeda berdasarkan karakteristiknya. Prediksi, pemahaman, pengambilan keputusan, deskripsi, dan eksplorasi data adalah tujuan dari klasifikasi data ini[6].
d. Algoritma C4.5
   Model klasifikasi data yang disebut Algoritma C4.5 bertujuan untuk membangun pohon keputusan secara rekursif. Dalam kontesk ini, dapat menemukan karakteristik yang paling cocok untuk membagi data pada setiap node dan kemudian mengulangi proses ini untuk setiap subnode[7].
e. Naïve Bayes
   Metode klasifikasi probabilitas yang sederhana, algoritma Naive Bayes berasumsi bahwa setiap fitur dalam data saling independen. Berdasarkan Teorema Bayes, algoritma ini menghitung probabilitas posterior untuk setiap kelas dan mengalokasikan instance data ke kelas yang memiliki probabilitas posterior tertinggi[8].

II. METODE PENELITIAN

2.1 Tahapan Penelitian

Alur penelitian secara keseluruhan digambarkan dalam langkah ini, yang dimulai dengan tahap studi literatur, pengumpulan data dan berakhir dengan penemuan hasil akhir. Gambar 1 menunjukkan hubungan antara langkah-langkah yang diambil dalam penelitian ini.



Gambar 1. Kerangka Penelitian

Alur dan penjelasan pada kerangka penelitian ini dijelaskan sebagai berikut.:

1. Analisis Masalah
   Analasis masalah pada penelitian ini adalah suatu proses untuk mengidentifikasi, memahami, dan mengevaluasi suatu masalah secara menyeluruh dan sistematis. Tujuan utama analisis masalah adalah untuk menemukan solusi yang tepat sasaran dan efektif.
2. Studi Literatur
   Peneliti melanjutkan studi pustaka setelah menyelesaikan analisis masalah. Pada tahap ini, peneliti mengumpulkan data dan informasi yang berkaitan dengan masalah yang dianalisa.
3. Pengumpulan Data
   Langkah selanjutnya adalah pengumpulan data; dalam hal ini, pengumpulan dataset adalah proses pengumpulan data yang terorganisir dari berbagai sumber untuk digunakan dalam analisis, pembelajaran mesin, atau pemodelan. Data publik

tentang penyakit diabetes dari UC Irvine Machine Learning Repository digunakan oleh peneliti sebagai dataset. Penelitian ini menggunakan dataset prediksi risiko diabetes awal. Dataset ini dapat diakses melalui URL berikut: https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset.

4. Pre-Procesing Data
   Tahap ini melibatkan pre-processing data untuk meningkatkan kualitas data dan mengurangi kesalahan variabel dan atribut dataset yang akan digunakan. Hal ini dapat berdampak pada hasil akhir pemrosesan dataset.
5. Pemodelan Algoritma Klasifikasi C4.5 dan Naïve Bayes
   Melanjutkan tahapan dari sebelumnya, peneliti menggunakan alat data mining Rapidminer untuk melakukan pengujian dataset untuk menentukan perbandingan nilai akurasi kedua algoritma untuk melakukan klasifikasi prediksi pola atribut dan label pada dataset.
6. Evaluasi dan Validasi
   Studi ini melakukan evaluasi dan validasi dengan menggunakan performance vector untuk menunjukkan nilai akurasi pengujian algoritma C4.5 dan Naïve Bayes dalam penentuan hasil deteksi penyakit diabetes.
7. Hasil
   Tahapan terakhir dari metode penelitian ini bertujuan untuk menyajikan hasil akhir dari seluruh rangkaian pengujian yang telah dilakukan. Semua data dan informasi yang dikumpulkan selama proses penelitian akan dianalisis dan disajikan pada tahap ini. Hasil dari pengujian ini diharapkan dapat memberikan wawasan yang jelas mengenai algoritma mana yang paling akurat dalam mendeteksi diagnosis penyakit diabetes. Selanjutnya peneliti dapat menentukan kelebihan dan kekurangan dari setiap metode yang diuji dengan menganalisis kinerja masing-masing algoritma

III. HASIL DAN PEMBAHASAN

3.1 Pre-Procesing Data

Tabel 1 menunjukkan variabel dataset. Untuk membuat data lebih mudah dibaca, nama atribut dan class harus disesuaikan dengan bahasa Indonesia. Selanjutnya, untuk memaksimalkan kinerja algoritma C4.5 dan Naive Bayes, penyesuaian tipe atribut dengan tipe data pada dataset juga diperlukan.

Tabel 1. Variabel Dataset Penelitian

Kelas	Binominal	0	Positif	Negatif	Positif (320), Negatif (200)
Umur	Integer	0	Min: 16	Max: 90	Average: 48.029
Jenis Kelamin	Binominal	0	Laki-laki	Wanita	Laki-laki (328), Wanita (192)
Sering Buang Air Kecil	Binominal	0	Tidak	Ya	Tidak (262), Ya (258)
Sering Haus	Binominal	0	Ya	Tidak	Tidak (287), Ya (233)
Penurunan Berat Badan Secar...	Binominal	0	Tidak	Ya	Tidak (303), Ya (217)
Badan Lemas	Binominal	0	Ya	Tidak	Ya (305), Tidak (215)
Perasaan Lapar Yang Terus-me...	Binominal	0	Tidak	Ya	Tidak (283), Ya (237)
Infeksi Jamur Vagina	Binominal	0	Tidak	Ya	Tidak (404), Ya (116)
Penglihatan Kabur	Binominal	0	Tidak	Ya	Tidak (287), Ya (233)
Gatal	Binominal	0	Ya	Tidak	Tidak (267), Ya (253)
Gampang Marah	Binominal	0	Tidak	Ya	Tidak (394), Ya (126)
Gangguan Penyembuhan Luka	Binominal	0	Ya	Tidak	Tidak (281), Ya (239)
Sulit Menggerakkan Otot Secar...	Binominal	0	Tidak	Ya	Tidak (296), Ya (224)
Kekakuan Otot	Binominal	0	Ya	Tidak	Tidak (325), Ya (195)
Kerontokan Rambut Yang Berl...	Binominal	0	Ya	Tidak	Tidak (341), Ya (179)
Obesitas	Binominal	0	Ya	Tidak	Tidak (432), Ya (88)

Pada Tabel 1, terlihat bahwa terdapat 16 variabel dalam dataset serta 1 variabel yang berfungsi untuk menentukan klasifikasi.

3.2 Algoritma C4.5

Setelah data dikumpulkan, algoritma C4.5 dipakai dalam membuat pohon keputusan serta menghasilkan alur keputusan. Selanjutnya, perhitungan gain informasi dilakukan untuk memilih fitur sebagai akar. Pertama, rumus persamaan (2) digunakan untuk menghitung nilai entropy untuk masing-masing atribut setelah mengetahui entropy total. Selanjutnya, sesuai dengan landasan teori di atas, Nilai gain untuk setiap atribut dihitung dengan rumus persamaan. (1). Selanjutnya, nama atribut dan kelas harus disesuaikan dengan tipe data pada dataset.

Entropi adalah parameter yang digunakan untuk mengukur keragaman nilai atribut kriteria dan kemudian dibandingkan dengan atribut keputusan, atau atribut keputusan, dalam kumpulan data. Nilai entropi yang lebih rendah menunjukkan nilai keragaman kumpulan data yang lebih besar [9]. Rumus berikut digunakan untuk menghitung entropy:

$$Entropy\ (S) = \sum_{i=1}^{n} -pi * \log_2 pi$$

dengan :
S : Himpunan kasus
n : Jumlah partisi kasus
p_i : Proporsi dari S_1 terhadap S

Pada dataset yang diujikan terdapat 520 variabel data dengan dua kelas positif dan negatif diagnosa positif berjumlah 320 data sedangkan kelas diagnosa negatif berjumlah 200 data. Berdasarkan rumus diatas maka dapat dilakukan perhitungan manual nilai entropy sebagai berikut :

$$Entropy : \left(-\frac{320}{520} * \log_2\left(\frac{320}{520}\right)\right) + \left(-\frac{200}{520} * \log_2\left(\frac{200}{520}\right)\right) = 0{,}961236605$$

Maka nilai entropy total = 0,961236605

Nilai entropi total dikurangi dari nilai entropi masing-masing atribut kriteria, dikalikan dengan nilai proporsi nilai atribut, dan kemudian dibagi dengan jumlah sampel data adalah perbedaan nilai entropy total. Ini disebut gain (S, A). Nilai gain yang diperoleh digunakan untuk menentukan seberapa efektif setiap fitur kriteria dalam mengklasifikasikan data. Nilai gain digunakan untuk membentuk node dan cabang pohon keputusan dalam algoritma c4.5 [10]. Nilai gain informasi dapat dihitung dengan menggunakan rumus berikut :

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si)$$

dengan :
S : Himpunan kasus
A : Atribut
n : Jumlah partisi atribut A
$|S_i|$: Jumlah kasus pada partisi ke-i
$|S\ |$: Jumlah kasus dalam S

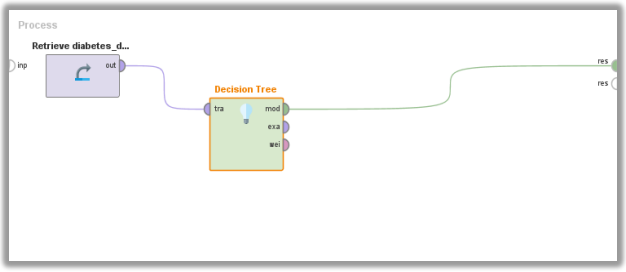
Dalam rangka mengimplementasikan rumus tersebut, untuk menghitung nilai gain informasi terhadap salah satu atribut data, pada konteks ini, atribut yang digunakan dalam perhitungan adalah jenis kelamin, dapat dilihat dirumus sebagai berikut:

$$Gain = 0{,}961236605 - \left(\frac{328}{520}(0{,}992235114)\right) + \left(\frac{192}{520}(0{,}465684865)\right) = 0{,}163420045$$

Nilai perhitungan gain pada atribut jenis kelamin sebesar = 0,163420045

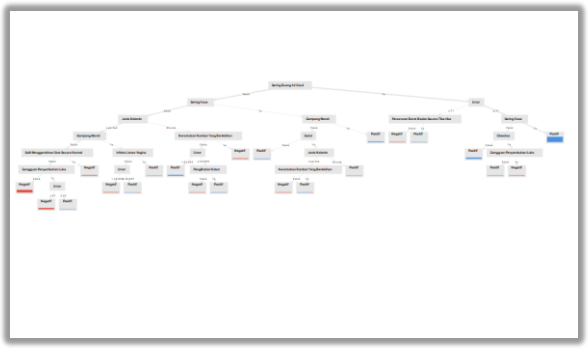
3.2.1 Model Algoritma C4.5 pada Rapidminer

Algoritma C4.5 diterapkan pada rapidminer setelah perhitungan nilai entropy dan nilai gain informasi selesai. Ini menunjukkan model proses datamining pada dataset, yang dapat dilihat sebagai berikut :



Gambar 2. Model Algoritma C4.5 pada Rapidminer

Gambar 2 menunjukkan bagaimana metode pemodelan Algoritma C4.5 akan digunakan untuk mengintegrasikan model data penyakit diabetes yang telah dikumpulkan ke dalam platform RapidMiner. Proses ini akan dilakukan menggunakan RapidMiner versi 9.10, yang merupakan salah satu versi terbaru dan memiliki berbagai fitur canggih untuk pemodelan data. Melalui proses ini, kita akan dapat menghasilkan sebuah desain pohon keputusan yang komprehensif. Desain ini kemudian akan divisualisasikan dan ditunjukkan pada Gambar 3, sehingga memberikan gambaran yang jelas tentang struktur keputusan yang dihasilkan dari data tersebut:



Gambar 3. Hasil klasifikasi *Decision Tree* pada aplikasi Rapidminer

Hasil klasifikasi *Decision Tree* pada gambar 3 menunjukan atribut paling penting yang mempengaruhi diagnosa penyakit diabetes ialah *Sering Buang Air Kecil*. Jika kondisi *Sering Buang Air Kecil* "Ya" maka kemudian atribut selanjutnya yang diperiksa ialah *Umur*, jika kondisi *Sering Buang Air Kecil* "Tidak" maka atribut yang akan diperiksa adalah *Haus Yang Berlebihan*, jika kondisi *Haus Yang Berlebihan* mengkondisikan "Ya" maka akan turun ke atribut *Gampang Marah* dan seterusnya sampai ditemukan hasil akhir keputusan yang bernilai "positif" dan keputusan yang bernilai "negatif".

3.3 Algoritma Naive Bayes

Metode klasifikasi Naive Bayes dapat digunakan untuk menemukan nilai peluang posterior tertinggi. Metode ini mengkomparasikan nilai posterior dengan nilai posterior lainnya. Kelas dengan nilai peluang posterior tertinggi diklasifikasikan sebagai kelas positif atau negatif[11].

Menghitung probabilitas hipotesis untuk setiap kelas P(H) adalah langkah pertama sebelum membangun model Naïve Bayes.. Pasien dengan diabetes positif dan negatif termasuk dalam hipotesis yang dianalisis. Data yang digunakan dalam metode Naïve Bayes terdiri dari total 520 sampel, di mana 320 di antaranya terdiagnosa positif diabetes dan 200 lainnya negatif. Teorema Bayes adalah dasar dari proses prediksi Naive Bayes. Ini gambarkan pada rumus sebagai berikut:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

dengan :

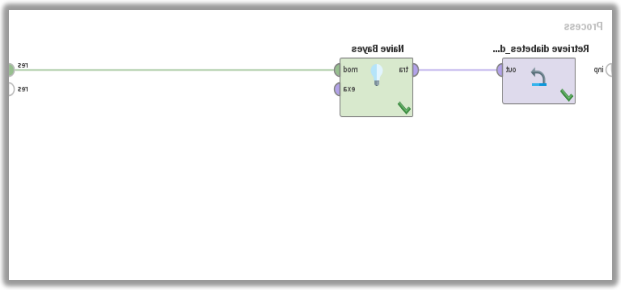
B : data dengan class yang belum diketahui
A : Hipotesis data B
P(A|B) : Probabilitas A berdasarkan B
P(B|A) : Probabilitas B berdasarkan A
P(A) : Probabilitas dari A
P(B) : Probabilitas dari B.

Penentuan probabilitas atribut kelas pada dataset terhadap hitung manual dapat dilihat sebagai berikut :

P(Positif, n) = 320/520 = 0,615384615
P(Negatif, n) = 200/520 = 0,384615385

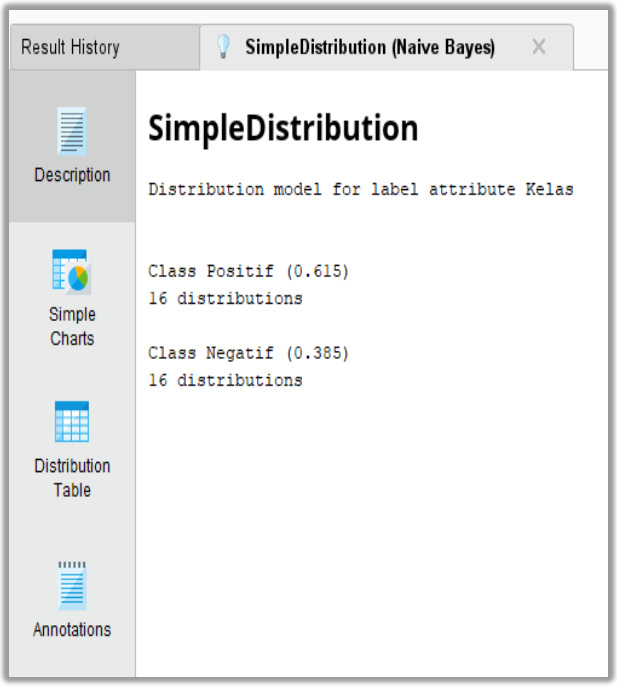
3.3.1 Model Algoritma Naive Bayes pada Rapidminer

Gambar 4 menunjukkan model pengujian Algoritma Naive Bayes untuk Rapidminer 9.10. :



Gambar 4. Model Algoritma Naïve Bayes pada Rapidminer

Data yang sudah melalui pre-procesing maka kemudian dilakukan pengujian algoritma Naïve Bayes di Rapidminer seperti pada gambar 4. Hasil perhitungan probabilitas kelas di Rapidminer dapat dilihat pada gamber berikut:

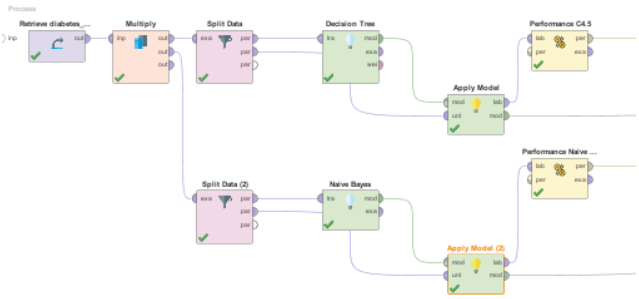


Gambar 5. Hasil probabilitas pada atribut kelas di Rapidminer

Hasil yang diperoleh dari perhitungan probabilitas di rapidminer bertujuan untuk mencari perbandingan perhitungan manual nilai probabilitas pada kelas dengan perhitungan yang dilakukan oleh rapidminer. Dari hasil perandingan tersebut tidak terdapat perbedaan perhitungan seperti pada gambar 5.

3.4 Evaluasi pengujian Algoritma C4.5 dan Naive Bayes di Rapidminer

Untuk menguji algoritma Naive Bayes dan model C4.5, operator split data digunakan untuk membagi data, seperti yang ditunjukkan pada gambar berikut :



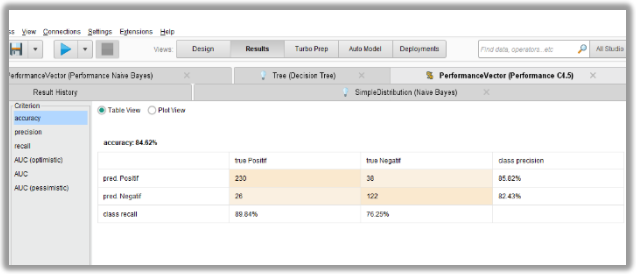
Gambar 6. Model perbandingan pengujian algoritma C4.5 dan Naïve Bayes pada Rapidminer

Dataset yang berbeda digunakan untuk menguji kinerja model. Dataset tersebut membagi 520 data menjadi dua bagian: 80% digunakan untuk pelatihan dan 20% untuk pengujian. Oleh karena itu, tujuh puluh persen dari total data, atau 416, digunakan untuk proses pelatihan, dan dua puluh persen, atau 104 data, digunakan untuk pengujian.

Analisa uji performa dilakukan dengan operator performance vector pada Rapidminer. Proses ini bertujuan untuk mengevaluasi dan meningkatkan performa dua model algoritma yang dilakukan pengujian di RapidMiner. Dengan memahami metrik yang ditawarkannya sehingga dapat membuat model yang lebih akurat.

3.4.1 Evaluasi Algoritma C4.5 di Rapidminer

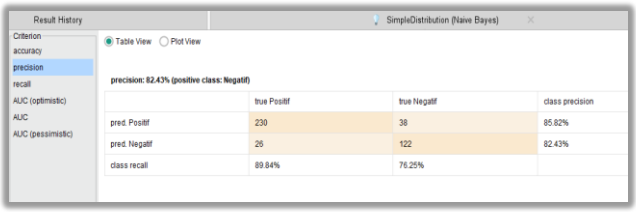
Dalam konteks ini dataset yang diterapkan kedalam model dilakukan uji performa algoritma C4.5 di rapidminer seperti pada gambar berikut :



Gambar 7. Hasil pengukuran *confusion matrix* kriteria *accuracy* pada algoritma C4.5

Tabel *confusion matrix accuracy* menunjukkan jumlah pengukuran prediksi akurasi yang dibuat oleh model algoritma C4.5 untuk setiap kelas target. Kelas aktual diwakili oleh baris matriks, dan prediksi model diwakili oleh kolom. Pada gambar 7 dihasilkan nilai *accuracy* algoritma C4.5 sebesar 84,62% dan 89,84% *true* Positif serta 76,25% *true* Negatif.

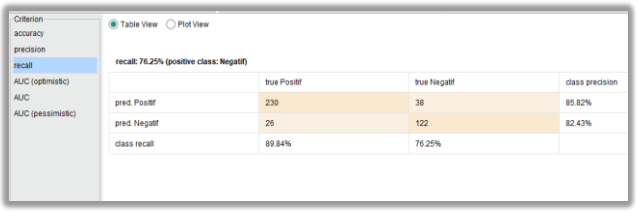
Gambar berikut menunjukkan pengukuran confusion matrix untuk kriteria presisi algoritma C4.5:



Gambar 8. Hasil pengukuran *confusion matrix* kriteria *precision* pada algoritma C4.5

Pada gambar 8 dihasilkan nilai *confusion matrix* kriteria *precision* algoritma C4.5 sebesar 82,43% dan 89,84% *true* Positif serta 76,25% *true* Negatif.

Gambar berikut menunjukkan hasil confusion matrix untuk kriteria recall:

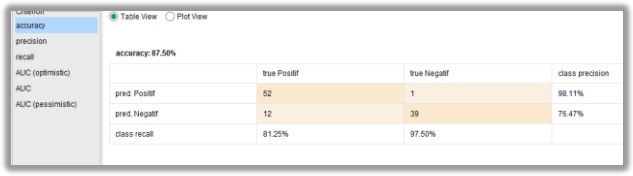


Gambar 9. Hasil pengukuran *confusion matrix* kriteria *recall* pada algoritma C4.5

Pada gambar 9 dihasilkan nilai *confusion matrix* kriteria nilai *recall* algoritma C4.5 sebesar 76,25% dan nilai 85,84% true Positif serta sebesar 76,25% *true* Negatif.

3.4.2 Evaluasi Algoritma Naïve Bayes di Rapidminer

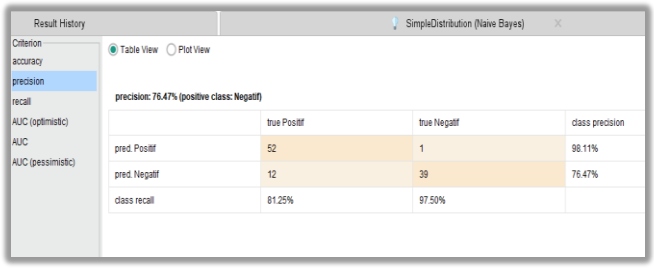
Sama halnya dengan pengujian diatas, evaluasi algoritma Naïve Bayes di Rapidminer juga menggunakan operator *Performance vector* yang bertujuan untuk mengukur confusion matrix dari 3 kriteria yaitu *accuracy, precision, recall.* Evaluasi menyeluruh menggunakan confusion matrix dan metrik akurasi, presisi, dan recall memberikan pemahaman mendalam tentang performa algoritma Naive Bayes di RapidMine. Pengukuran matrix confusion terhadap kriteria akurasi algoritma Naive Bayes ditunjukkan pada gambar berikut:



Gambar 10. Hasil pengukuran *confusion matrix* kriteria *accuracy* pada algoritma Naïve Bayes

Gambar 10 menunjukkan nilai confusion matrix kriteria *accuracy* kelas algoritma Naive Bayes sebesar 87,50% untuk nilai true positif, 81,25% untuk nilai true negatif, dan 97,50%.

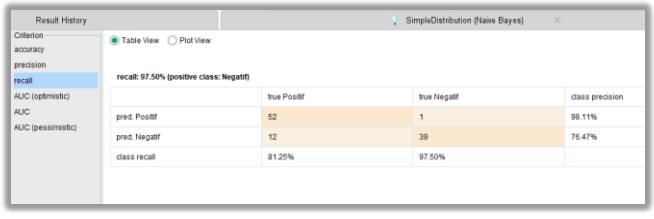
Gambar berikut menunjukkan pengukuran confusion matrix terhadap kriteria precision algoritma Naive Bayes:



Gambar 11. Hasil pengukuran *confusion matrix* kriteria *precision* pada algoritma Naïve Bayes

Gambar 11 menunjukkan nilai confusion matrix kriteria *precision* algoritma Naive Bayes sebesar 76,47%, 81,25% nilai true positif, dan 97,50% nilai true negatif.

Gambar berikut menunjukkan pengukuran confusion matrix terhadap kriteria *recall* pada algoritma Naive Bayes:



Gambar 12. Hasil pengukuran *confusion matrix* kriteria *recall* pada algoritma Naïve Bayes

Gambar 12 menunjukkan nilai confusion matrix kriteria *recall* algoritma Naive Bayes sebesar 97,50% nilai asli positif, 81,25% nilai true negatif, dan masing-masing 97,50% nilai true negatif..

3.5 Hasil Komparasi Dan Validasi

Tabel berikut menunjukkan hasil perbandingan proses klasifikasi dari model Algoritma C4.5 dan Algoritma Naïve Bayes yang diuji untuk memprediksi penyakit diabetes :

Tabel 2. Hasil Komparasi Klasifikasi

Algoritma	*Accuracy*	*Precision*		*Recall*	
		Positif	Negatif	Positif	Negatif
C4.5	84,62%	89,84%	76,25%	85,84%	76,25%
Naïve Bayes	87,50%	81,25%	97,50%	81,25%	97,50%

Tabel 2 menunjukkan proses perbandingan algoritma Naive Bayes dan C4.5 saat menguji performa klasifikasi di Rapidminer.

IV. KESIMPULAN DAN SARAN

4.1 KESIMPULAN

Penelitian ini berhasil membandingkan performa algoritma klasifikasi C4.5 dan Naïve Bayes dalam mendeteksi risiko penyakit diabetes, yang merupakan langkah penting dalam upaya deteksi dini dan pencegahan komplikasi diabetes. Keunggulan utama dari penelitian ini terletak pada penggunaan metode yang sistematis, dimulai dari studi literatur hingga pengujian model menggunakan dataset yang diambil dari UC Irvine Machine Learning Repository. Hasil penelitian menunjukkan bahwa algoritma Naive Bayes memiliki akurasi 87,50%, sedikit lebih tinggi dari 84,62% yang dimiliki C4.5.

Hasil ini menunjukkan bahwa Naive Bayes lebih baik dalam memprediksi diagnosis diabetes dan membantu mengembangkan sistem klasifikasi yang efektif. Selain itu, penelitian ini juga memberikan wawasan tentang pentingnya pemilihan algoritma yang tepat dalam konteks kesehatan, yang dapat membantu tenaga medis dalam pengambilan keputusan. Dengan hasil yang diperoleh, penelitian ini tidak hanya menambah pengetahuan dalam bidang teknologi informasi dan kesehatan, tetapi juga membuka peluang untuk penelitian lebih lanjut yang dapat mengeksplorasi algoritma lain dan faktor-faktor klinis yang mempengaruhi diagnosis. Secara keseluruhan, penelitian ini memberikan dasar yang kuat untuk pengembangan sistem diagnosis diabetes yang lebih akurat dan efektif.

4.2 SARAN

Untuk penelitian selanjutnya, disarankan agar peneliti mempertimbangkan beberapa aspek yang dapat meningkatkan kualitas dan relevansi hasil. Pertama, penggunaan dataset yang lebih besar dan beragam sangat penting untuk meningkatkan generalisasi model. Data yang mencakup berbagai demografi dan kondisi medis dapat memberikan wawasan yang lebih komprehensif tentang penyebab diabetes.

Kedua, melakukan penelitian tentang algoritma klasifikasi lain, seperti Random Forest, Support Vector Machine (SVM), atau algoritma berbasis pembelajaran mendalam, dapat menawarkan pemahaman baru tentang efektivitas prediksi diabetes. Masing-masing algoritma memiliki kelebihan yang dapat dimanfaatkan untuk meningkatkan akurasi.

Ketiga, analisis pre-processing data yang lebih mendalam, termasuk teknik pengisian nilai hilang dan normalisasi, dapat membantu meningkatkan kualitas data yang digunakan. Selain itu, integrasi faktor klinis dan sosial, seperti riwayat kesehatan dan gaya hidup, dalam model dapat memberikan hasil yang lebih akurat dan relevan.

Akhirnya, penelitian longitudinal yang mengevaluasi efektivitas model dalam jangka panjang akan sangat bermanfaat. Penelitian mendatang diharapkan dapat membantu mengembangkan sistem diagnosis diabetes yang lebih akurat dan efektif dengan mengikuti rekomendasi ini.

REFERENSI

[1] American Diabetes Association, “Diagnosis and classification of diabetes mellitus,” *Diabetes Care*, vol. 37, no. SUPPL.1, Jan. 2014, doi: 10.2337/dc14-S081.

[2] International Diabetes Federation (IDF), “Diabetes in Indonesia (2021).” Accessed: Jul. 19, 2024. [Online]. Available: https://idf.org/our-network/regions-and-members/western-pacific/members/indonesia/

[3] Kementerian Kesehatan RI, "Survei Kesehatan Indonesia Tahun 2023," 2023.

[4] U. Prediksi, P. Bantuan, J. Kesehatan, N. Attamami, A. Triayudi, and R. T. Aldisa, "Analisis Performa Algoritma Klasifikasi Naive Bayes dan C4.5 Teknologi Komunukasi dan Indormatika," vol. 7, no. 2, 2023, doi: 10.35870/jti.

[5] J. Han, M. Kamber, and J. Pei, "Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)," 2011.

[6] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. 2016.

[7] K. P. Murphy, "Machine Learning - A Probabilistic Perspective - Table-of-Contents," *The MIT Press*, 2012.

[8] C. Haruechaiyasak, "A Tutorial on Naive Bayes Classification," 2008. Accessed: Jul. 20, 2024. [Online]. Available: https://www.dit.uoi.gr/e-class/modules/document/file.php/184/A%20Tutorial%20on%20Naive%20Bayes%20Classification.pdf

[9] A. Asroni, B. Masajeng Respati, and S. Riyadi, "Penerapan Algoritma C4.5 untuk Klasifikasi Jenis Pekerjaan Alumni di Universitas Muhammadiyah Yogyakarta," *Semesta Teknika*, vol. 21, no. 2, 2018, doi: 10.18196/st.212222.

[10] D. Puspita, S. Aminah, and A. Arif, "Prediction System for Credit Eligibility Using C4.5 Algorithm," *JOURNAL OF INFORMATICS AND TELECOMMUNICATION ENGINEERING*, vol. 6, no. 1, pp. 148–156, Jul. 2022, doi: 10.31289/jite.v6i1.7311.

[11] R. A. Santoso, D. Syauqy, M. Hannats, and H. Ichsan, "Pengembangan Sistem Prediksi Hama Wereng Berdasarkan Data Cuaca Sensor Dan Cuaca Online Menggunakan Metode Naive Bayes," 2018. [Online]. Available: http://j-ptiik.ub.ac.id