Versi Online tersedia di : https://jurnal.buddhidharma.ac.id/index.php/algor/index

JURNAL ALGOR

|2715-0577 (Online)| 2715-0569 (Print)

Artikel

PERBANDINGAN ALGORITMA *DATA MINING* DALAM MENGKLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN MODEL *C4.5* DAN *NAÏVE BAYES*

Hansen[1], *Susanto Hariyanto*[2]

[1,2]*Fakultas Sains dan Teknologi, Universitas Buddhi Dharma, Banten, Indonesia*

SUBMISSION TRACK

Recieved: Jan 22, 2023
Final Revision: March 12,2023
Available Online: March 24, 2023

KEYWORD

C4.5, Database, Data Mining, Diabetes, Naive Bayes.

KORESPONDENSI

Phone: 082210439602

E-mail: hhansen984@gmail.com

A B S T R A C T

Diabetes adalah penyakit metabolic kronis yang ditandai dengan peningkatan kadar gula atau glukosa darah, yang dari waktu ke waktu menyebabkan kerusakan serius pada jantung, pembuluh darah, mata, ginjal dan saraf. Menurut *International Diabetes Federation* (IDF) Indonesia menjadi negara urutan kelima dengan jumlah penderita diabetes terbanyak didunia dan diabetes menjadi penyebab kematian atas 6,7 juta kematian yang terjadi pada tahun 2021 setiap 5 detik. Data mining adalah teknik dalam dunia komputer yang sering digunakan dalam memprediksi apa yang akan terjadi pada masa depan, ini menjadi salah satu metode yang banyak digunakan dalam memprediksi apakah suatu individu terdiagnosa positif atau negatif diabetes. Salah satu metode yang paling popular adalah C4.5 dan *Naïve Bayes*. Data sendiri didapatkan dari *website* kaggle dari Pima Indian heritage dengan 9 atribut dan 769 *records* yang nantikan akan di *cleaning* menjadi 220 data. Hasil pemrosesan data mining membuktikan algoritma *Naïve Bayes* menghasilkan akurasi yang lebih besar dibandingkan C4.5 dengan nilai 85.00% dibandingkan C4.5 dengan nilai akurasi 78.86%. *Naïve Bayes* juga menghasilkan nilai AUC 0.936 dari 1 yang membuktikan bahwa klasifikasi ini termasuk kedalam *excellent classification.* Dari tulisan diatas dapat disimpulkan bahwa algoritma *Naïve Bayes* merupakan metode yang lebih baik dibandingkan dengan algoritma C4.5 dalam memprediksi apakah seseorang terdiagnosa positif/negatif diabetes.

PENDAHULUAN

Seiring berkembangnnya zaman, masyarakat lebih menyukai pola hidup yang serba cepat dan instan. Salah satu penyakit yang dapat ditimbulkan adalah diabetes. Diabetes sendiri adalah penyakit metabolic kronis yang ditandai dengan peningkatan kadar gula glukosa dalam darah, yang dari waktu ke waktu menyebabkan kerusakan serius pada jantung, pembuluh darah, mata, ginjal, dan saraf [1]. Pada tahun 2021 sendiri terdapat 531 juta orang dewasa dengan rentan usia 20 – 73 tahun hidup dengan mengidap penyakit diabetes, dan diprediksi pada tahun 2030 pengidap diabetes akan meningkat hingga 643 juta, diabetes bertanggung jawab atas 6.7 juta kematian yang terjadi pada tahun 2021[2].

Banyaknya penderita diabetes yang tidak mengetahui bahwa mereka menderita diabetes sehingga terjadinya keterlambatan diagnosa yang menyebabkan nyawa tersebut tidak dapat tertolong. Dengan seiring perkembangannya zaman khususnya dalam bidang teknologi, dapat membantu pasien dalam memprediksi penyakit diabetes khususnya menggunakan teknik *data mining*. *Data Mining* adalah salah satu ilmu dalam komputer yang digunakan untuk mengumpulkan berbagai informasi yang penting dari berbagai macam sumber, yang nantinya kumpulan informasi tersebut akan diolah dan akan menghasilkan informasi yang lebih akurat [3]. Salah satu teknik yang digunakan dalam melakukan *data mining* adalah dengan menggunakan teknik klasifikasi.

Klasifikasi adalah proses menemukan model atau pola baru untuk menentukan data yang saling terikat kedalam suatu kelas yang sama dengan data yang saling bercocokan dan memasukannya kedalam pada kategori tertentu [4]. Dalam klasifikasi sendiri terdapat berbagai macam metode didalamnya khususnya adalah *C4.5* dan *Naïve Bayes*. *C4.5* adalah algoritma yang digunakan untuk menghasilkan sebuah pohon keputusan berdasarkan pemilihan atribut yang memiliki prioritas tertinggi atau *gain* tertinggi berdasarkan nilai *entrophy*. Sedangkan *Naïve Bayes* adalah metode yang bekerja dengan memprediksi masa depan berdasarkan data sebelumnya. Namun belum diketahui metode apa yang terbaik dalam mengklasifikasi gejala penyakit diabetes diantara kedua algoritma tersebut sehingga dilakukan perbandingan algoritma dalam memprediksi penyakit diabetes diantara *C4.5* dan *Naïve Bayes*.

I. METODE

1.1 *Data Mining*

Data Mining adalah proses pencarian mengenai pola yang berwawasan luas, menarik, dan baru, serta merupakan model deskriptif yang mudah dipahami dan prediktif dari data yang berskala besar [5].

Data Mining juga dapat dikatakan sebagai sebuah proses untuk mengajukan berbagai pertanyaan dan mengekstrak informasi, pola dan tren yang berguna yang belum diketahui dari sekumpulan data yang berjumlah besar dan disimpan dalam *database*. *Data mining* biasanya diimplementasikan untuk mencapai tujuan tertentu. [6]

1.2 Klasifikasi

Klasifikasi adalah suatu proses menemukan kumpulan atau pola atau fungsi yang mendeskripsikan serta memisahkan kelas data yang satu dengan yang lainnya untuk menyatakan objek tersebut masuk pada kategori tertentu [4].

Klasifikasi adalah proses menemukan model atau fungsi yang menggambarkan dan membedakan kelas data atau konsep , dengan tujuan dapat menggunakan model untuk memprediksi kelas objek yang label kelasnya tidak diketahui.

1.3 *C4.5*

C4.5 atau *Decision Tree* adalah sebuah metode klasifikasi yang menggunakan struktur pohon, yang dimana setiap *noed* mempresentasikan atribut dan akarnya mempresentasikan hasilnya, sedangkan daunnya mempresentasikan kelasnya [7].

Cara kerja algoritma *C4.5* adalah dengan menentukan akar dari pohon. Akar ini diambil dari atribut yang dipilih dengan cara menghitung nilai *gain* dari masing-masing atribut. Nilai *gain* tertinggi akan menjadi akar pertama, namun sebelum mendapatkan nilai

gain terlebih dahulu menghitung nilai *entrophy* dengan rumus sebagai berikut:

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

Setelah mencari *entrophy* tahap selanjutnya adalah mencari *gain* dengan rumus sebagai berikut:

$$Gain\ (\text{S, A}) = \text{Entrophy (S)} - \sum_{i=1}^{n} \frac{Si}{S} * Entrophy \quad 2$$

Setelah mendapatkan nilai *gain*, ulangi kembali proses perhitungan *entrophy* hingga ditemukan *gain* tertinggi untuk setiap atributnya. Proses perhitungan akan selesai saat memenuhi suatu kondisi yaitu, semua tupel dalam node N mendapatkan kelas yang sama, tidak ada atribut yang dapat dipartisi kembali dan tidak ada tupel didalam cabang yang kosong.

1.4 *Naïve Bayes*

Naïve Bayes merupakan algoritma *data mining* yang membagikan objek sehingga masing-masing objek ditugaskan ke salah satu dari sejumlah kategori yang saling melengkapi dan ekslusif yang dikenal sebagai kelas. Ekslusif disini adalah objek harus mewakili suatu objek dan tidak pernah mewakili kelas yang lain.

Naïve Bayes adalah algoritma yang digunakan untuk memecahkan masalah dengan cara mengklasifikasikan suatu data. *Naïve Bayes* juga merupakan salah satu algoritma dengan teknik klasifier dengan menggunakan metode probabilitas dan statistika yang dikemukakan oleh ilmuan Inggris bernama Thomas Bayes yaitu dengan cara memprediksi masa depan berdasarkan pengalaman sebelumnya (Teorema bayes)[8].

Berikut merupakan rumus *Naïve Bayes*:

$$\text{P (C|X)} = \frac{\text{P(X|C)}P\ (C)}{P(X)} \quad 3$$

Keterangan:

X = data dengan kelas yang belum diketahui.

C = hipotesis data yang merupakan kelas yang spesifik.

P (C|X) = probabilitas hipotesis berdasarkan kondisi.

P (C) = probabilitas hipotesis.

P (X|C) = probabilitas berdasarkan kondisi pada hipotesis.

P (X) = probabilitas C

1.5 Pengujian

Pada tahap ini pengujian dilakukan untuk mengetahui metode apa diantara *C4.5* dan *Naïve Bayes* yang memiliki tingkat keakuratan terbaik. Pengujian dilakukan dengan melihat hasil *confusion matrix* dan AUC yang dihasilkan melalui *software rapidminer*. *Confusion Matrix* suatu metode yang digunakan untuk melakukan perhitungan akurasi pada konsep *data mining*[9]. Hasil *confusion matrix* dibagi menjadi empat istilah yaitu, *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), *False Negative* (FN). Akurasi dihitung dengan rumus

$$\text{akurasi} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \text{ x } 100\% \quad 4$$

Selain menggunakan *confusion matrix*, pada thap pengujian juga dilakukan pengujian AUC. AUC atau yang disingkat dari *Area Under Curve* merupakan sebuah area dibawah kurva yang menggambarkan evaluasi mengenai kemampuan metode klasifikasi dalam membedakan antar kelas. Semakin tinggi AUC maka kinerja model tersebut semakin baik dalam membedakan diantara kelas positif dan negatif. Nilai AUC merupakan 0 – 1 [10].

Confusion Matrix dan AUC didapatkan melalui *software rapidminer*. *Rapidminer* adalah alat penambangan data yang berguna untuk membuat model prediktif dengan cepat. Alat ini juga menerapkan *all in one* yang dimana *software* dapat menampilkan segala macam algoritma *machine learning* untuk mendukung segala macam proyek *machine learning*[11]. *Rapidminer* juga merupakan *software open source* yang memiliki kurang

lebih 500 *operator data mining*. *Rapidminer* dibuat dengan bahasa pemrograman *java* sehingga dapat berjalan disegala macam *platform*[12].

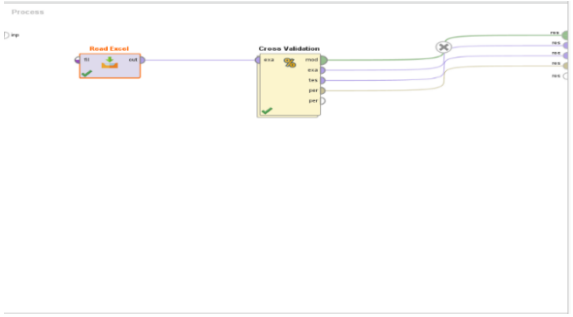
1.6 Pembuatan Aplikasi

Dalam pengaplikasiaan dan memudakan *user* dalam memprediksi diabetes, dibuatkan aplikasi berbasis java yang dibuat dengan *software netbeans.* Java adalah bahasa pemrograman yang bersifat OOP (*Object Oriented Programming*) yang dapat dijalankan diberbagai *platform* seperti Windiws, Linux, dll[13]. Java juga bahasa pemrograman yang mudah diingat karena Java bersifat *High Level Languange* yang dimana akan mudah bagi manusia untuk memahaminya karena banyak perintah atau fungsi yang menggunakan bahasa inggris dalam penulisannya [14].

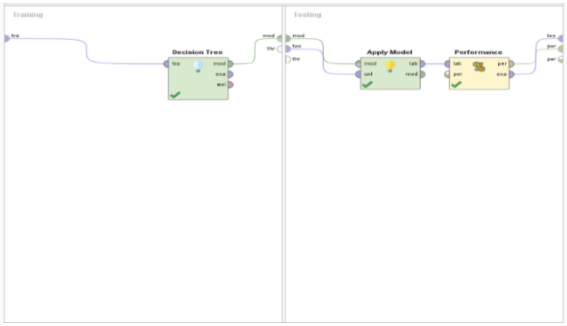
Netbeans IDE adalah sebuah IDE atau yang disingkat dari *Integrated Development Environment* yang berguna untuk membuat aplikasi dengan Java, PHP, C, C++, dan HTML. *Netbeans* IDE memiliki cara kerja yang mirip dengan *microsoft visual studio* dan *dreamweaver* sebagai aplikasi yang memiliki cakupan ruang lingkup kerja yang luas dalam bidang membangun aplikasi dan bersifat *open source*[15].

II. HASIL

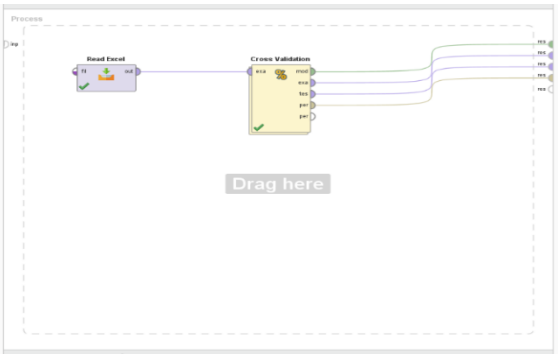
Setelah dilakukan *data cleaning* dihasilkan *dataset* dengan total 220 data yang terdiri dari 110 data positif dan 110 data negatif dengan 8 atribut yang terdiri dari kehamilan, gula darah, tekanan darah, ketebalan kulit, insulin, BMI, riwayat diabetes, umur dan 1 *label* yaitu Hasil. Setelah itu masukan *dataset* ke aplikasi *rapidminer.*



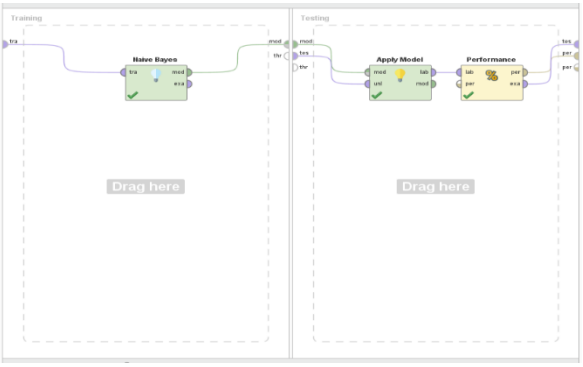
Gambar 1: Perancangan *Operator* di *Rapidminer*



Gambar 2: Perancangan *Operator* di *Rapidminer* 2 (*C4.5*)



Gambar 3: Perancangan *Operator* di *Rapidminer* (*Naïve Bayes*)



Gambar 4: Perancangan *Operator* di *Rapidminer* 2 (*Naïve Bayes*)

Setelah menerapkan kedua algoritma didalam aplikasi *rapidminer* didapatkan hasil *confusion matrix* untuk *C4.5* sebagai berikut:

Tabel 1. Hasil Confussion Matrix C4.5

Akurasi: 85.00%	*True* Positif Diabetes	*True* Negatif Diabetes	*Class Precission*
Pred. Positif Diabetes	87	24	78.38%
Pred. Negatif Diabetes	23	86	78.90%
Class Recall	79.09%	78.18%	

Hasil akurasi *confusion Matrix* pada *C4.5* menghasilkan akurasi 78.64% yang dihitung dengan cara sebagai berikut:

$$\left(\frac{87+86}{87+23+24+86}\right)\text{x}100\% = 78.64\% \quad 5$$

Sedangkan untuk algoritma *Naïve Bayes* menghasilkan hasil *confusion matrix* untuk *Naive Bayes* sebagai berikut:

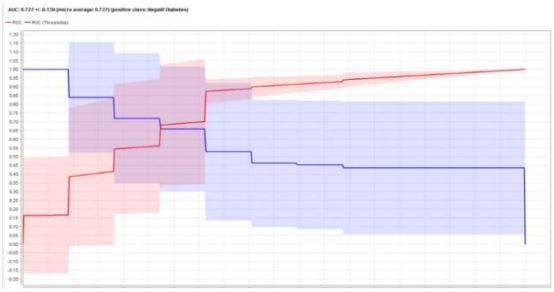
Tabel 2. Hasil Confusion Matrix Naive Bayes

Akurasi: 85.00%	*True* Positif Diabetes	*True* Negatif Diabetes	*Class Precission*
Pred. Positif Diabetes	92	15	85.95%
Pred. Negatif Diabetes	18	95	84.07%
Class Recall	83.64%	86.36%	

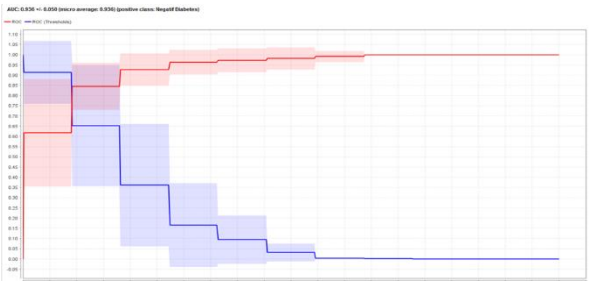
Algoritma *Naïve Bayes* menghasilkan akurasi *confusion matrix* 85.00% dengan cara sebagai berikut:

$$\left(\frac{92+95}{92+18+15+95}\right)\text{x}100\% = 85.00\% \quad 6$$

Setelah menerapkan pengujian *confusion matrix* untuk mencari akurasi terbaik, peneliti juga menggunakan pengujian AUC, dan didapatkan hasil *C4.5* dengan nilai 0.727 sedangkan *Naïve Bayes* menghasilkan nilai 0.936 dan hasil grafik ditunjukan dibawah ini:



Gambar 5: Hasil Grafik AUC C4.5



Gambar 6: Hasil Grafik AUC Naïve Bayes

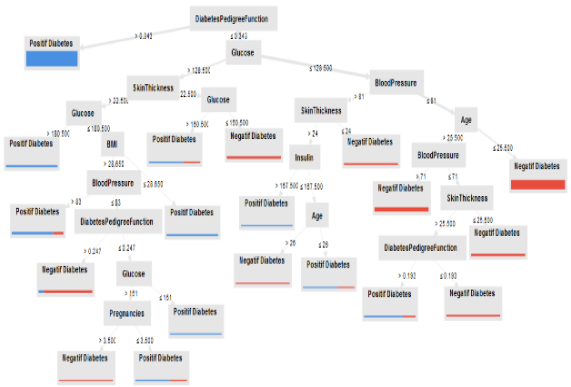
Setelah dilakukan pengujian terhadap kedua algoritma yang digunakan didapatkan tabel perbandingan sebagai berikut:

Tabel 3. Tabel Pengujian

	Naïve Bayes	C4.5
Confusion Matrix	85.00%	78.64%
AUC	0.936	0.727

Berdasarkan tabel pengujian diatas algoritma *Naïve Bayes* terbukti lebih baik dalam mengklasifikasi penyakit diabetes dengan jumlah 220 data dengan hasil tingkat akurasi 85.00% nilai AUC 0.936.

Dalam Penelitian ini juga didapatkan hasil model untuk algoritma *C4.5* berupa pohon keputusan, berikut merupakan gambarannya.



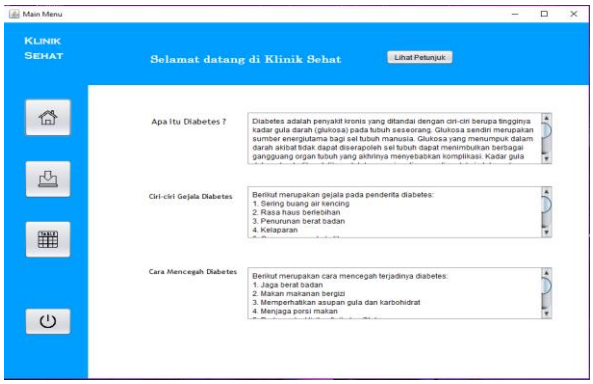
Gambar 7: Hasil Pohon Keputusan (*C4.5*)

Sedangkan untuk *Naive Bayes*, didapatkan hasil model pembagian untuk setiap datanya yang hasilnya akan digunakan pada perhitungan selanjutnya.

```
SimpleDistribution

Distribution model for label attribute Outcome


Class Positif Diabetes (0.500)
8 distributions

Class Negatif Diabetes (0.500)
8 distributions
```

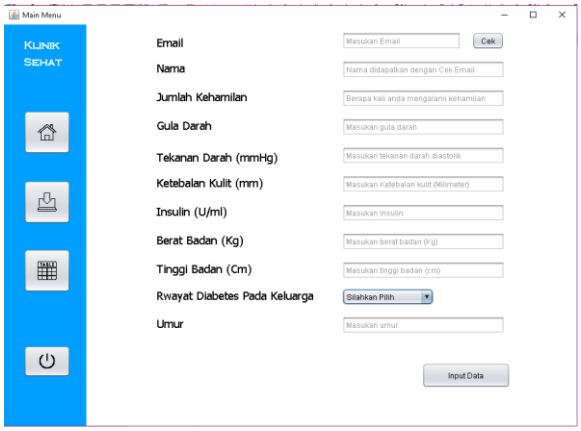
Gambar 8: Hasil Model *Naive Bayes*

Dalam penelitian ini peneliti mengembangkan aplikasi berbasis *desktop* yang mengimplementasikan ilmu *Data Mining* untuk memudahkan pengguna atau *user* dalam penggunaannya, dikarenakan *Naive Bayes* memiliki presentasi lebih tinggi maka aplikasi yang dibuat berbasis *Naive Bayes* Berikut merupakan tampilan aplikasi menu utama dan *input data*.



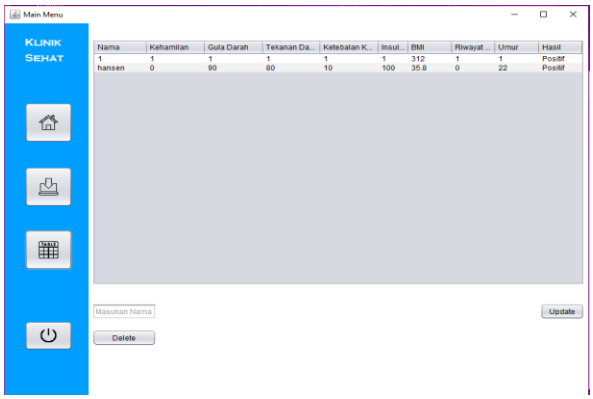
Gambar 9: Halaman Utama

Pada Halaman ini setelah *user login*, *user* akan masuk kedalam halaman utama. Halaman utama memiliki 3 fungsi utama yaitu *home* yang berisikan informasi mengenai seputar diabetes. Fungsi lainnya adalah menu *input* yang berguna untuk melakukan *input data* yang dimana *data user* akan dihitung dan akan mengeluarkan hasil prediksi diabetes, dan fungsi terakhir adalah fungsi untuk melihat tabel hasil *input user* untuk melihat riwayat atau *history* input *user*.



Gambar 10: Halaman Input

Halaman *Input* berisikan beberapa *fieldtext*, *button*, dan *label* yang dimana akan memberikan informasi mengenai data yang harus dimasukan oleh *user*. Disaat *user* menekan tombol *input* maka data dalam *textfield* akan dimasukan kedalam *database* dan akan dilakukan perhitungan untuk memprediksi apakah *user* positif/negatif diabetes.



Gambar 11 Halaman Lihat Data

Halaman terakhir yang penting adalah halaman lihat data. Pada halaman ini *user* dapat melihat riwayat dari *input data* yang telah dilakukan sebelumnya atau *user* dapat memasukan nama dan menu-*update* atau *delete* data mereka.

III. PEMBAHASAN

Tabel 4. Data Set

K	GD	TD	KK	I	BMI	R	U	H
2	102	86	36	120	45.5	0.127	23	P
0	95	85	25	36	37.40	0.247	24	P

1	81	72	18	40	26.6	0.283	24	N
1	90	62	18	59	25.1	1.268	25	N
3	102	44	20	94	30.8	0.4	26	N
3	163	70	18	105	31.6	0.268	28	P
7	124	70	33	215	25.5	0.161	37	N
5	96	74	18	67	33.6	0.997	43	N
4	125	70	18	122	28.9	1.114	45	P
11	120	80	37	150	42.3	0.785	48	P

Keterangan data:
K = Kehamilan
GD = Gula Darah
TD = Tekanan Darah
KK = Ketebalan Kulit
I = Insulin
BMI = *Body Mass Index*
R = Riwayat Diabetes
U = Umur
H = Hasil (Positif / Negatif)

Sebelumnya dilakukan perhitungan manual dengan menggunakan 10 data yang terdiri dari 5 data positif dan 5 data negatif dengan 8 atribut dan 1 label. Berikut merupakan perhitungan manual untuk algoritma *C4.5* dan *Naïve Bayes.* Algoritma *C4.5.* Tahap pertama dalam algoritma ini adalah mencari *entrophy* dari total data yang digunakan adalah 1 dengan cara sebagai berikut:

$$Entrophy\ (\text{total}) = \left(-\frac{5}{10} * LOG_2\left(\frac{5}{10}\right)\right) + \left(-\frac{5}{10} * LOG_2\left(\frac{5}{10}\right)\right) = 1 \quad (7)$$

Setelah mendapatkan *entrophy* total tahap selanjutnya adalah mencari *gain* untuk setiap atrbibut, sebelumnya telah ditemukan *entrophy* untuk masing-masing atribut. Dalam perhitungan manual akan digunakan contoh atribut dengan *gain* terbesar yaitu tekanan darah, tekanan darah dengan nilai <=77 dan >77. <=77 memiliki jumlah 7 data dan >77 dengan jumlah 3 data dan ditemukan nilai *entrophy* berupa 0.86312 untuk <=77 dan >77 dengan nilai 0. Setelah ditemukan nilai *entrophy* didapatkan nilai *gain* dengan nilai 0.395816 yang didapatkan dengan rumus sebagai berikut:

$$Gain\ =\ 1 - \left(\frac{7}{10} * 0.86312\right) + \left(\frac{3}{10} * 0\right) = 0.395816 \quad (8)$$

Dikarenakan atribut tekanan darah memiliki nilai *gain* terbesar dibandingkan dengan atribut lain maka pada iterasi pertama, atribut tekanan darah akan menjadi *node* atau daun pertama pada pohon keputusan. Setelah dilakukan iterasi pertama tahap selanjutnya dilakukan iterasi kedua. Ulangi proses pencarian *entrophy* total namun bukan total data melainkan *entrophy* dari <=77 yang terdiri dari 2 positif dan 5 negatif dan didapatkan hasil sebagai berikut:

Tahap selanjutnya adalah mengulangi proses pencarian *gain* terhadap data yang <=77 terhadap 8 atribut. Setelah dilakukan perhitungan maka didapatkan atribut dengan *gain* terbesar yaitu gula darah dengan nilai <=124,5 yang terdiri dari 5 data (0 positif dan 5 negatif) dan >124,5 yang terdiri dari 2 data (2 positif dan 0 negatif) dan didapatkan hasil *gain* berupa 0.863121 dengan rumus sebagai berikut:

$$Gain\ =\ 0.863121 - \left(\frac{5}{7} * 0\right) + \left(\frac{2}{7} * 0\right) = 0.863121 \quad (9)$$

Setelah didapatkan *gain* terbesar pada iterasi kedua maka dapat diposisikan bahwa gula darah akan menjadi daun ke2 atau node kedua dari pohon keputusan yang dibuat. Dikarenakan data pada atribut tidak dapat dibagi kembali maka perhitungan manual untuk algoritma *C4.5* akan berhenti di iterasi kedua dengan *rule* sebagai berikut:

1. Jika Blood Pressure >77 maka positif diabetes.
2. Jika Blood Pressure <= 77 dan glucose >124.500 maka positif diabetes.

3. Jika Blood Pessure <=77 dan glucose <=124.500 maka negatif diabetes.

Setelah mendapatkan hasil dari algoritma *C4.5,* dilakukan perhitungan kembali menggunakan algoritma *Naïve Bayes* dengan data dan jumlah data yang sama. Pada tahap pertama algoritma ini hal yang harus dilakukan adalah menghitung *average* untuk setiap hasil positif dan negatif pada setiap atribut. Contoh data set yang digunakan berisikan 10 data yang terdiri dari data positif (2, 0, 7, 4, 11) dan data negatif (1, 1, 3, 7, 5) maka dalam mencari average kehamilan untuk positif adalah $(2 + 0 + 3 + 4 + 11)/5 = 4$. Contoh lainnya adalah average untuk tekanan darah negatif. Tekanan darah negatif terdiri dari (72, 62, 44, 70, 74) maka $(72 + 52 + 44 + 70 + 74)/5 = 64.4$. Hasil dari perhitungan *average* untuk setiap atribut akan ditunjukan pada tabel dibawah ini.

Tabel 5. Tabel *Average*

AVG		H	GD	TD	KK	I	BMI	R	U
	P	4	121	78.2	26.8	106.6	37.14	0.5142	33.6
	N	3.4	98.6	64.4	21.4	95	28.32	0.6128	31

Setelah mendapatkan nilai *average*, tahap selanjutnya adlah mencari *variance* dengan rumus sebagai berikut:

$$S^2 = \frac{\sum(x_i - x)^2}{n - 1} \qquad 10$$

Keterangan:

S^2 = Sampel *variance.*

x_i = nilai dari satu pengamatan.

x = nilai rata-rata dari semua pengamatan.

n = jumlah pengamatan.

Setelah menerapkan rumus *variance* didapatkan tabel hasil sebagai berikut:

Tabel 6. Tabel *Variance*

VARIANCE		H	GD	TD	KK	I	BMI	R	U
	P	17.5	704.5	61.2	86.7	1821.8	48.793	0.187979	143.3
	N	6.8	261.8	150.8	42.8	4876.5	13.827	0.233667	72.5

Setelah mendapatkan hasil *variance* tahap selanjutnya adalah mencari hasil standar deviasi, namun bukan berdasarkan data pada tabel namun data pada tabel *variance,* berikut merupakan rumusnya:

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}} \qquad 11$$

Setelah menerapkan rumus diatas, didapatkan hasil tabel seperti berikut.

Tabel 7. Tabel Standar Deviasi

STD		H	GD	TD	KK	I	BMI	R	U
	P	4.1833	26.5424	7.823044	9.311283478	42.68255	6.98552	0.433565	11.97088
	N	2.607688	16.1802	12.2801	6.542170894	69.83197	3.718477	0.483391	8.514699

Setelah mendapatkan semua rumus yang diperlukan, algoritma ini membutuhkan soal atau data yang perlu di uji hasilnya. Contoh soalnya adalah kehamilan 2, gula darah 80, tekanan darah 80, ketebalan kulit 23, insulin 140, BMI 31.2, riwayat diabetes 0.4 dan umur 24 akan menghasilan positif atau negatif. Setelah menentukan data soal yang akan di uji, tahap selanjutnya adalah menghitung probabilitas dari data soal tersebut. Contoh probabilitas untuk kehamilan adalah (1/ ((4.1833) ∗ SQRT(2 ∗ 3.14))) ∗ EXP(−((2 − 4)^2)/(2 ∗

4.1833^2))). 4.1833 menunjukan kepada standar deviasi yang didapatkan untuk atribut kehamilan (positif), 2 adalah soal/data uji yang digunakan dan 4 adalah *average* dari atribut kehamilan pada hasil positif. Berdasarkan rumus diatas didapatkan hasil probabilitas kehamilan (positif) 0.085. Berikut merupakan hasil probabilitas untuk data soal yang digunakan.

Tabel 8. Tabel Probabilitas

Probabilitas	P	0.085	0.005	0.05	0.0394316	0.007	0.04	0.889	0.024
	N	0.132	0.013	0.015	0.0591984	0.005	0.08	0.743	0.033

Setelah mendapatkan nilai probabiltias, tahap selanjutnya adalah mengkalikan semua angka positif dan negatif dan membandingkannya , untuk mengetahui soal yang digunakan termasuk kedalam positif atau negatif. Sebelumnya telah didapatkan 0.5 yang didapatkan dari menghitung berapa kali nilai positif atau negatif muncul pada kelas hasil.

$$Positif = 0.085 * 0.005 * 0.0394 * 0.007 * 0.04 * 0.889 * 0.024 * 0.5 = 2.23635E-12 \quad (12)$$

$$Negatif = 0.132 * 0.013 * 0.015 * 0.0591984 * 0.005 * 0.743 * 0.033 * 0.5 = 6.64067E-12 \quad (13)$$

Setelah mendapatkan nilai untuk HASIL, tahap terakhir dalam perhitungan ini adalah menghitung tingkat *confidence* dari kedua hasil diatas yang hasilnya ditunjukan dibawah ini.

$$Positif = \frac{2.23635E-12}{(2.23635E-12 + 6.64067E-12)} = 0.251925845 \quad (14)$$

$$Negatif = \frac{6.64067E-12}{(6.64067E-12 + 2.23635E-12)} = 0.748074155 \quad (15)$$

Berdasarkan hasil *confidence* diatas, dapat diputuskan bahwa soal uji yang terdiri dari kehamilan 2, gula darah 80, tekanan darah 80, ketebalan kulit 23, insulin 140, BMI 31.2, riwayat diabetes 0.4 dan umur 24, menghasilkan keputusan berupa negatif diabetes, dikarenakan *confidence* negatif lebih besar dibandingkan *confidence* positif.

IV. KESIMPULAN

Kesimpulan yang didapatkan dari penelitian ini adalah, diketahui ilmu komputer *data mining* khususnya algoritma *C4.5* dan *Naïve Bayes* terbukti dapat mengklasifikasi seseorang terjangkit diabetes dengan menguji *dataset* yang digunakan. Setelah pengujian yang dilakukan dengan perhitungan *manual* dan menggunakan *software rapidminer*, menghasilkan bahwa algoritma *Naïve Bayes* memiliki tingkat akurasi yang lebih tinggi dibandingkan *C4.5*, yang dimana *Naïve Bayes* menghasilkan tingkat akurasi 85.00% dan AUC 0.936 sedangkan *C4.5* dengan nilai 78.65% dengan AUC 0.727.

REFERENSI

[1] W. H. Organization, "Diabetes." https://www.who.int/health-topics/diabetes#tab=tab_1

[2] T. I. D. Federation, *The International Diabetes Federation Diabetes Atlas 10 Edition*, 10th ed. Brussels: International Diabetes Federation, 2021.

[3] E. D. Sikumbang, "Penerapan Data Mining Penjualan Sepatu Menggunakan Metode Algoritma Apriori," vol. 4, no. 1, pp. 156–161, 2018.

[4] A. R. Sukma, R. Halfis, and A. Hermawan, "Klasifikasi Channel Youtube Indonesia Menggunakan Algoritma C4.5," *TEKNIK KOMPUTER*, vol. V, pp. 21–28, 2019, doi: 10.31294/jtk.v4i2.

[5] M. J. Zaki and W. M. Jr, *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 2014.

[6] B. Thuraisingham, *Data Mining: Technologies, Techniques, Tools, and Trends*. CRC Press, 2014.

[7] A. A. Aldino and H. Sulistiani, "DECISION TREE C4.5 ALGORITHM FOR TUITION AID GRANT PROGRAM CLASSIFICATION (CASE STUDY: DEPARTMENT OF INFORMATION SYSTEM, UNIVERSITAS TEKNOKRAT INDONESIA)," 2020.

[8] R. Rino, "The Comparison of Data Mining Methods Using C4.5 Algorithm and Naive Bayes in Predicting Heart Disease," *Tech-E*, vol. 4, no. 2, p. 44, 2021, doi: 10.31253/te.v4i2.543.

[9] P. Mayadewi and E. Rosely, "PREDIKSI NILAI PROYEK AKHIR MAHASISWA MENGGUNAKAN ALGORITMA KLASIFIKASI DATA MINING," 2015. [Online]. Available: https://www.researchgate.net/publication/283570705

[10] S. Vector Machine Untuk Mengklasifikasi Dan Memprediksi Angkutan Udara and S. Fachrurrazi, "PENGGUNAAN METODE SUPPORT VECTOR MACHINE UNTUK MENGKLASIFIKASI DAN MEMPREDIKSI ANGKUTAN UDARA JENIS PENERBANGAN DOMESTIK DAN PENERBANGAN INTERNASIONAL DI BANDA ACEH," *Jurnal Sistem Informasi*, vol. 2, no. 2, pp. 1–10, 2018.

[11] D. Jollyta, M. Siddik, H. Mawengkang, and S. Efendi, *Teknik Evaluasi Cluster Solusi Menggunakan Python Dan Rapidminer*. DeePublish, 2021.

[12] A. M. Siregar and A. Puspabhuana, *DATA MINING: Pengolahan Data Menjadi Informasi dengan RapidMiner*. CV Kekata Group, 2017.

[13] E. Jando and P. N. Andrianus, *Algoritma dan Pemrograman dengan Bahasa Java*. Yogyakarta: Penerbit Andi, 2018.

[14] M. Rusli, I. K. Rinartha, and Y. P. Atmojo, *Belajar Pemrograman Java dengan Netbeans*. Penerbit Andi, 2016.

[15] J. Enterprise, *Mengenal Java dan Database dengan NetBeans*. Jakarta: Elex Media Komputindo, 2015.

BIOGRAPHY

Hansen Lahir di Tangerang pada tanggal 28 Mei 2000. Menyelesaikan pendidikan Strata 1 (Sarjana S1) pada tahun 2022 pada Fakultas Sains dan Teknologi, program studi Teknik Informatika di Universitas Buddhi Dharma.

Susanto Hariyanto, Saat ini bekerja sebagai dosen tetap pada Universitas Buddhi Dharma yang mengajar Program Studi Teknik Informatika.