Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma C4.5

Alif Abqori Robbani
Universitas Buana Perjuangan
Karawang, Indonesia
if17.alifrobbani@mhs.ubpkarawang.ac.id

Amril Mutoi Siregar
Universitas Buana Perjuangan
Karawang, Indonesia
amrilmutoi@ubpkarawang.ac.id

Dwi Sulistya Kusumaningrum
Universitas Buana Perjuangan
Karawang, Indonesia
dwi.sulistya@ubpkarawang.ac.id

***Abstrak*—**
Diabetes merupakan suatu kondisi kronis bisa berlangsung seumur hidup dan akan mempengaruhi kemampuan tubuh dalam menggunakan energi makanan yang telah dicerna. Berdasarkan WHO atau organisasi kesehatan dunia memprediksi pengidap penyakit diabetes miletus di Republik Indonesia dari 8,4 juta jiwa di tahun 2020, dan pada tahun 2030 naik sampai 21,3 juta jiwa. Sedangakan menurut *International Diabetes Federation* (IDF) juga pada tahun 2009 pengidap penyakit diabetes 7,0 juta jiwa sampai 12,0 juta jiwa tahun 2030. Pada tahun 2030 menurut prediksi WHO dan IDF penderita penyakit diabetes miletus di Indonesia naik 2-3 kali lipat. Pendekatan data mining menjadi sangat penting dalam bidang kesehatan untuk mengambil keputusan berdasarakan data klinis yang besar. Teknik klasifikasi termasuk kedalam bagian metode *supervised learning* yaitu diperlukannya data latihan dalam membangun pola untuk model klasifikasinya. Algoritma C4.5 termasuk dalam algoritma klasifiksasi yang menghasilkan pohon keputusan dan bisa diolah dengan data diskrit dan numerik, selain itu algoritma C4.5 dapat mengasilkan cara yang mudah untuk diinterpretasikan pada penelitian akurasi dari algoritma C4.5 sebesar 74.08%.

***Kata kunci*—** algoritma C4.5, data mining, diabetes

I. PENDAHULUAN

Kesehatan yaitu peranannya sangat penting untuk menunjang kehidupan manusia, dengan memiliki kesehatan yang baik, manusia dapat melakukan aktifitas dengan produktif dalam sosialisasi atau ekonomi untuk mecapai tujuan hidup. Salah satu penyakit yang dapat mengakibatkan komplikasi bahkan kematian adalah penyakit diabetes. Diabetes bukan hanya penyebab dari kematian prematur di dunia, penyakit ini bisa menyebabkan kebutaan, gagal ginjal, dan bisa juga menyebabkan penyakit jantung [1]. Menurut International Diabetes Federetion (IDF) kusus pasien pengidap diabetes miletus di dunia meningkat setiap tahunnya. Tahun 2011 terdapat 366 juta jiwa, tahun 2013 terdapat 382 juta jiwa, tahun 2015 yaitu 415 juta jiwa, sedangkan tahun 2017 naik hingga 425 juta jiwa, dan pada tahun 2019 terdapat 463 juta jiwa. Diperkirakan 2045 pengidap penyakit diabetes terus mengalami kenaikan menjadi 700 juta jiwa yang terdiagnosa [2] (Alisa et al, 2020). Berdasarkan WHO atau organisasi kesehatan dunia memprediksi pengidap penyakit diabetes miletus di Republik Indonesia dari 8,4 juta jiwa di tahun 2020, dan pada tahun 2030 naik sampai 21,3 juta jiwa. Sedangakan menurut International Diabetes Federation (IDF) juga pada tahun 2009 pengidap penyakit diabetes 7,0 juta jiwa sampai 12,0 juta jiwa tahun 2030. Pada tahun 2030 menurut prediksi WHO dan IDF penderita penyakit diabetes miletus di Indonesia naik 2-3 kali lipat [3].

. Dengan diterapkannya data mining diharapkan dapat menjadi suatu informasi untuk penyakit diabetes di Indonesia ataupun dunia, sehingga angka penderita penyakit diabetes dapat menurun. Teknik klasifikasi termasuk kedalam bagian metode supervised learning yaitu diperlukannya data latihan dalam membangun pola untuk model klasifikasinya. Teknik klasifikasi memiliki beberapa pilihan algoritma, algoritma yang masuk kedalam metode supervised learning yaitu k-Nearest-Negihbor (k-NN), Naïve Bayes, Support Vector Machine, ID3, dan algoritma C4.5 [4]. Algoritma C4.5 termasuk dalam algoritma klasifiksasi yang menghasilkan pohon keputusan dan bisa diolah dengan data diskrit dan numerik, selain itu algoritma C4.5 dapat mengasilkan cara yang mudah untuk diinterpretasikan. C4.5 telah dicoba oleh diberbagai kasus klasifikasi seperti bidang kepegawaian, perdagangan, medis dan bidang lainnya [5].

Pada penelitian Rahman analisa pasien penyakit liver menggunakan metode Naïve Bayes dan decision tree C4.5 menyimpulkan metode decision tree C4.5 memiliki akurasi yang paling besar 70.29% sedangkan algoritma Naïve Bayes menghasilkan akurasi 67.05% [6]. Penelitian lain oleh Leidiyana & Permana untuk meningkatkan kualitas pemilihan calon karyawan C4.5 menggunakan confusion matrix mendapatkan hasil akurasi sebesar 87.5% [7]. Penelitian lainnya yang menggunakan perbandingan C4.5 dan Naïve Bayes untuk menentukan dosen tetap yang dilakukan oleh Sadikin, C4.5 menghasilkan akurasi tertinggi 91.89% sementara itu Naïve Bayes menghasilkan nilai akurasi sebesar 83.78% [8].

II. DATA DAN METODE

A. Bahan dan Peralatan Penelitian

Pada penelitian ini menggunakan data yang membahas penderita penyakit diabetes, memiliki *record* data 768 data dengan 8 atribut dan 1 sebagai label, data ini diambil dari https://www.kaggle.com/jamaltariqcheema/pima-indians-diabetes-dataset.

Tabel 1 Data Penelitian

No	*Pregnen*	*Glucose*	*Blood P*	*Skin T*	Insulin	BMI	DPF	*Age*	*Outcome*
1	6	148	72	35	169.5	33.6	0.627	50	1
2	1	85	66	29	102.5	26.6	0.351	31	0
3	8	183	64	32	169.5	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	27	102.5	25.6	0.201	30	0
...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
768	1	93	70	31	102.5	30.4	0.315	23	0

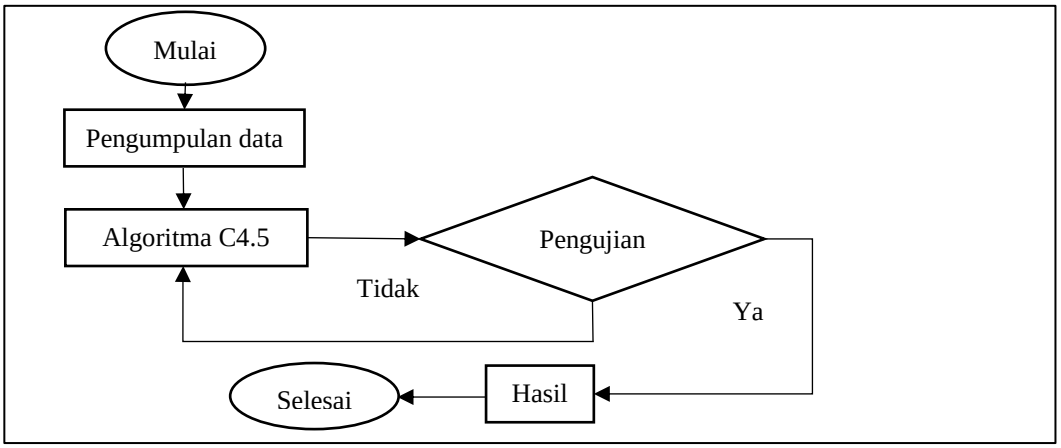
Kemudian, untuk memenuhi kebutuhan penelitian membutuhkan alat seperti perangkat keras dan perangkat lunak sebagai berikut :

I. Perangkat Keras
- Laptop yang digunakan yaitu. Processor (intel core i3) RAM 8GB dengan sistem operasi *windows* 10.
- *Flashdisk* 8GB
- *Smartphone*

II. Perangkat Lunak
- Python
- Sistem operesai *windows* 10
- Mircosoft *office*
- RapidMiner
- Google chrome

B. Prosedur Penelitian

Prosedur penelitian ini sebagai berikut :



Gambar 1 Prosedur Penelitian

1. Pengumpulan Data

Pada tahapan pengumpulan data ini diambil dari salah satun *online* repository yang memiliki jumlah data 768 dan 8 atribut.

2. Tahapan Algoritma C4.5

Berikut tahapan-tahapan *decision tree* algoritma C4.5 [9]:

1) Menyiapkan data latih atau uji.
2) Menghitung *entropy*(S) adalah parameter yang digunakan untuk informasi keberagaman setiap nilai atribut kategori atau kriteria terhadap atribut keputusan dalam sebuah dataset. untuk menentukan *entropy* dengan rumus:

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

3) Menghitung nilai *gain* (S, A) adalah untuk mengukur efektivitas masing masing atribut pada *node* tertentu untuk mengklasifikasikan data, nilai terbesar akan menjadi akar pohon utama dengan rumus:

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

4) Mengulangi Langkah ke 2 sampai seemua *record* data terpartisi.
5) Partisi akan berhenti apabila:
   a) Semua *record* yang ada pada simpul N mendapat kelas yang sama.
   b) Tidak adanya atribut pada *record* yang dipartisi.
   c) Tidak adanya *record* pada cabang yang kosong.

3. Pengujian

Pada tahapan pengujian untuk menentukan tingkat akurasi dan pohon keputusan algoritma C4.5. Dalam pengujian penelitian ini dilakukan beberapa proses yaitu:

a) Perhitungan secara manual untuk menghasilkan pohon keputusan.
b) Pengujian menggunakan *tool* RapidMiner untuk menghasilkan akurasi.

III. HASIL DAN PEMBAHASAN

A. Hasil Analisis Data

Setelah melewati beberapa tahapan pada prosedur percobaan, maka untuk menemukan hasil dari data di perlukannya seleksi data dengan tujuan untuk pemilihan atribut untuk memfokuskan data dan atribut-atribut yang digunakan hanya yang penting saja, atribut yang tidak perlukan berarti tidak digunakan. Atribut yang digunakan adalah:

- Glokosa
- *BloodPressure*
- BMI
- *Age*

Atribut yang digunakan selanjutnya akan dihitung manual menggunakan algoritma C4.5, penulis menggunakan atribut tersebut karena menurut dr. Alifian semuanya berhubungan dengan penyakit diabetes.

Atribut yang tidak digunakan menurut dr. Alifian, (*pregnancies*) jumlah kehamilan tidak berpengaruh terhadap penyakit diabetes, (insulin) insulin adalah obat untuk diabetes, (*skinthickness*) adalah ketebalan kulit untuk pengobatan pada luka diabetes, dan dalam DPF di dalam kesulurahan datanya mengandung *noise*.

Tabel 2 setelah tahap seleksi

No	*Glucose*	*Blood P*	BMI	*Age*	*Outcome*
1	148	72	33.6	50	1
2	85	66	26.6	31	0
3	183	64	23.3	32	1
4	89	66	28.1	21	0
5	137	40	43.1	33	1
6	116	74	25.6	30	0
...	...	...	...	...	...
...	...	...	...	...	...
768	93	70	30.4	23	0

Perhitungan manual algoritma C4.5 perlu merubah data yang bertipe numerik harus dikategorikan terlebih dahulu, untuk mempermudah proses perhitungan.

Tabel 1 Perubahan Tipe Data

Atribut	Nilai	Kategori	Referensi
Glukosa	<140	NORMAL	dr. Karlina Lestari
	141-199	PREDIABETES	https://www.alodokter.com/komunitas/topic/diabetes-107

Atribut	Nilai	Kategori	Referensi
	>200	DIABETES	
BloodPressure	<80	NORMAL	dr. Meva Nareza
	81-89	PRAHIPERTENSI	https://www.alodokter.com/berapa-tekanan-darah-normal-orang-dewasa
	90-99	HIPERTENSI1	
	100-119	HIPERTENSI2	
	>120	KRISIS	
BMI	<18,5	KURANG	dr. Tjin Willy
	18,6-29,9	NORMAL	https://www.alodokter.com/obesitas/diagnosis
	>30	OBESITAS	
Age	21-59	DEWASA	dr. Karlina Lestari
	>60	LANSIA	https://www.sehatq.com/artikel/risiko-penyakit-berdasarkan-klasifikasi-umur-menurut-who

Tabel 4 dataset sesudah perubahan tipe data

No	*Glucose*	*Blood P*	*BMI*	*DPF*	*Age*	*Outcome*
1	PRADIABETES	NORMAL	OBESITAS	SEDANG	DEWASA	YA
2	NORMAL	NORMAL	NORMAL	SEDANG	DEWASA	TIDAK
3	PRADIABETES	NORMAL	NORMAL	SEDANG	DEWASA	YA
4	NORMAL	NORMAL	NORMAL	SEDANG	DEWASA	TIDAK
5	NORMAL	NORMAL	OBESITAS	TINGGI	DEWASA	YA
6	NORMAL	NORMAL	NORMAL	SEDANG	DEWASA	TIDAK
7	NORMAL	NORMAL	OBESITAS	SEDANG	DEWASA	YA
...	...	...	...	...	...	...
...	...	...	...	...	...	...
768	NORMAL	NORMAL	OBESITAS	SEDANG	DEWASA	TIDAK

B. Hasil Perhitungan Manual

Pada tahap pengolahan data dilakukan perhitungan manual dengan menggunakan algoritma C4.5 total keseluruhan data yang digunakan yaitu 768 data, atribut yang digunakan terlampir pada Tabel 4 diatas. Perhitungan manual algoritma C4.5 perlu merubah data yang bertipe numerik harus dikategorikan terlebih dahulu, untuk mempermudah proses perhitungan.

Perhitungan manual algoritma C4.5 dapat dilakukan. Algoritma C4.5 adalah algoritma yang menghasilkan suatu pohon keputusan (decision tree), pertama hal yang harus dilakukan dalam perhitungan algoritma C4.5 adalah menentukan entropy dari setiap atribut dengan rumus sebagai berikut:

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

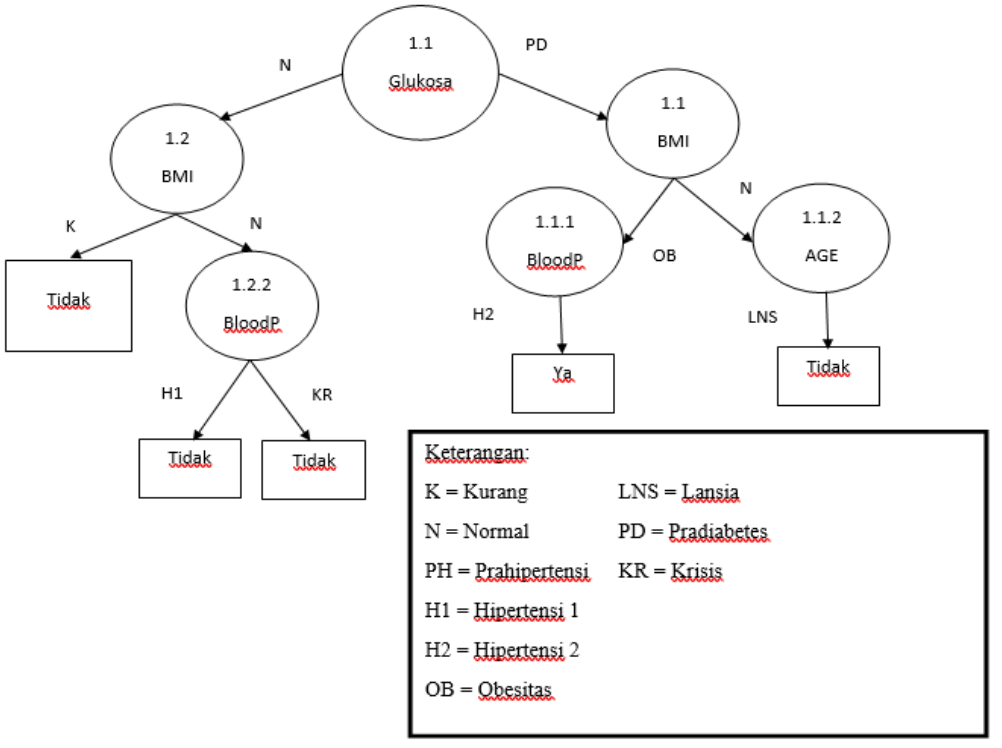
Keterangan:
S = himpunan (dataset) kasus
n = jumlah partisi S
pi = jumlah sampel pada kelas i

Setelah mendapatkan informasi seluruh nilai entropy, selanjutnya menghitung nlai gain dari seluruh atribut. Menghitung Nilai Gain dengan rumus sebagai berikut [10]:

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

Keterangan :
S = himpunan kasus
A = atribut
n = jumlah partisi atribut A
|Si| = jumlah kasus pada partisi ke-i
|S| = jumlah kasus dalam S

C. Hasil *Decision Tree*



Berdasarkan pohon keputusan diatas menghasilkan *rule* sebagai berikut:
- *IF* gluoksa prediabetes *AND* BMI obesitas *AND bloodpressure* hipertensi2 = YA
- *IF glukosa prediabetes AND BMI normal AND age lansia = TIDAK*
- *IF glukosa normal AND BMI kurang = TIDAK*
- *IF glukosa normal AND BMI normal AND bloodpressure hipertensi1 = TIDAK*
- *IF glukosa normal AND BMI normal AND bloodpressure krisis = TIDAK*

D. Hasil *Confusion Matrix*

Tabel 5 *Confusion Matrix*

	Ya	Tidak
Ya	116	47
Tidak	152	453

Accuracy $= \frac{\text{jumlah prediksi benar}}{\text{jumlah total prediksi}} = \frac{116+453}{768} * 100\ \%$

= 74,08 %

Laju *error* $= \frac{\text{jumlah prediksi salah}}{\text{jumlah total prediksi}} = \frac{47+152}{768} * 100\ \%$

= 26 %

Precision $= \frac{\text{TP}}{\text{TP+FP}} = \frac{116}{116+152} * 100\ \%$

= 43,28%

Recall $= \frac{\text{TP}}{\text{TP+FN}} = \frac{116}{116+47} * 100\ \%$

= 71,16%

F-Measure $= 2 * \frac{Precision * Recall}{Precision + Recall} = 2 * \frac{0{,}4328 * 0{,}7116}{0{,}4328 + 0{,}7117} = 2 * \frac{0{,}307}{1{,}14} * 100\ \% = 2 * 0{,}269 * 100$

= 53,8 %

E. Hasil Pengujian RapidMiner dan Python

Tabel 6 Hasil Pengujian RapidMiner dan Python

Hasil Pengujian	**RapidMiner**	**Python**
Akurasi	74,78%	77,05%
Error	25,22%	-
Recall Ya	43,75%	46%
Recall Tidak	91,33%	95%
Precision Ya	72,92%	75%
Precision Tidak	75,27%	85%

Pada Tahapan pengujian RapidMiner dan Python dilakukan pembagian data training 70% dan data testing 30%, yang digunakan pada pengujian RapidMiner dan Python adalah data testing 30%

IV. KESIMPULAN DAN SARAN

Hasil klasifikasi dari penderita penyakit diabetes dengan data yang diambil dari situs https://www.kaggle.com/jamaltariqcheema/pima-indians-diabetes-dataset total data 768 dengan menggunakan algoritma C4.5 dilakukan dengan tahapan pengumpulan data, pengolahan data dan pengujian dengan diterapkannya algoritma C4.5. Untuk perhitungan manual terdapat beberapa tahapan pada algoritma C4.5 yakni mencari nilai entropy, kemudian setelah mencari nilai entropy selanjutnya mencari nilai gain, setelah nilai gain didapatkan mencari nilai gain tertinggi untuk dijadikan node akar. Lakukan perhitungan berulangkali sampai hasilnya telah memiliki keputusan semuanya. Perhitungan manual menggunakan algoritma C4.5 membentuk decision tree yang memiliki 5 rule diharapkan menjadi suatu informasi tentang penyakit diabetes. Evaluasi dari penelitian ini diukur dengan akurasi, laju *error*, *precision*, *recall*, dan *f-measure*. Dengan 768 data memiliki akurasi 74,08%, laju *error* 26%, *precision* 43,28%, *recall* 71,16%, dan *f-measure* 53,8%.

Saran yang dapat diberikan berdasarkan penelilitian ini untuk menggunakan dataset dari rumah sakit yang berada di sekitar Indonesia. Semakin banyak data semakin akurat, dan menggunakan algoritma yang berbeda dengan tujuan untuk mengetahui perbandingannya.

PENGAKUAN

Naskah ilmiah ini adalah sebagian dari penelitian Tugas Akhir milik Alif Abqori Robbani Dengan Judul Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma C4.5 yang dibimbing Amril Mutoi Siregar, M.Kom dan Dwi Sulistya Kusumaningrum, M.Pd.

DAFTAR PUSTAKA

[1] Infodatin 2020 Diabetes Melitus (pp. 1–10). (2020). Kementrian Kesehatan RI.

[2] Alisa, F., Amelia, W., Sastra, L., & Despitasari, L. (2020). *Edukasi Online Pelaksanaan Aktifitas Fisik Pada Pasien Diabetes*.*2*, 53–57.

[3] Ente, D. R., Thamrin, S. A., Arifin, S., Kuswanto, H., & Andreza, A. (2020). Klasifikasi Faktor-Faktor Penyebab Penyakit Diabetes Melitus Di Rumah Sakit Unhas Menggunakan Algoritma C4.5. *Indonesian Journal of Statistics and Its Applications*, *4*(1), 80–88. https://doi.org/10.29244/ijsa.v4i1.330

[4] Khotimah, N., & Istiawan, D. (2018). Perbandingan Algoritma C4.5, Naïve Bayes dan K-Nearest Neighbour untuk Prediksi Lahan Kritis di Kabupaten Pemalang. *Urecol*, *7*(1), 41–50.

[5] Yuningsih, L., Setiawan, I., & Sunarto, A. (2020). Rancangan Aplikasi Prediksi Kelulusan Siswa. *Jurnal Ilmiah Komputer*, *16*(2), 121–132.

[6] Rahman, N. T. (2020). *Analisa Algoritma Decision Treedan Naïve Bayespada Pasien Penyakit Liver*. *10*(2), 144–151.

[7] Leidiyana, H., & Permana, A. A. (2020). *Pemodelan Klasifikasi Dalam Meningkatkan Proses*.

[8] Sadikin, M., Rosnelly, R., & Gunawan, T. S. (2020). *Perbandingan Tingkat Akurasi Klasifikasi Penerimaan Dosen Tetap Menggunakan Metode Naive Bayes Classifier dan C4 . 5*. *4*, 1100–1109. https://doi.org/10.30865/mib.v4i4.2434

[9] Ayudhitama, A. P., & Pujianto, U. (2020). Analisa 4 Algoritma Dalam Klasifikasi Penyakit Liver Menggunakan. *Jurnal Informatika Polinema*, *6*, 1–9.

[10] Setiawan, R. (2020). Analisis Kelayakan Pemberian Kredit Nasabah Koperasi Menggunakan Algoritma C4.5. *Techno Xplore : Jurnal Ilmu Komputer Dan Teknologi Informasi*, *5*(2), 74–78. https://doi.org/10.36805/technoxplore.v5i2.1175