PREDIKSI PENYAKIT DIABETES MELLITUS MENGGUNAKAN ALGORITMA C4.5

Rini Andanika Siallagan[1], Fitriyani[2]

[1]Universitas Adhirajasa Reswara Sanjaya
e-mail: rinisiallagan97@gmail.com

[2]Universitas Adhirajasa Reswara Sanjaya
e-mail: fitriyani@ars.ac.id

Abstrak

Diabetes mellitus merupakan penyakit kronis yang disebabkan oleh gagalnya organ pankreas memproduksi jumlah hormon insulin secara memadai sehingga menyebabkan peningkatan kadar glukosa dalam darah. Diabetes Mellitus merupakan salah satu penyakit keturunan, penyakit ini bisa di turunkan orang tua kepada anaknya, dan sangat disayangkan bila diusia yang masih muda sudah mengalami diabetes. Pemeriksaan pada penyakit Diabetes Mellitus dalam dunia medis dapat dilakukan dengan cara pendiagnosaan penyakit berdasarkan gejala-gejala yang didertita oleh penderita yang dapat menghasilkan data hasil uji laboratorium dan rekam medis gejala sakit. Untuk meminimalisir angka kematian dari penyakit Diabetes Mellitus ini, para pakar kesehatan harus melakukan pendiagnosaan penyakit sedini mungkin. Klasifikasi dapat dijadiin salah satu penanganan dini dari penyakit ini, klasifikasi ini menggunakan metode algoritma C4.5. Penelitian ini bertujuan untuk membantu para medis untuk mengklasifikasi para pasien yang memilik gejala-gejala penyakit diabetes. Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan (*Decision Tree*). Pohon keputusan merupakan metode klasifikasi dan prediksi yang terkenal. Algoritma C4.5 merupakan metode yang dapat digunakan untuk memprediksi dan mengetahui nilai akurasi pada pasien dengan gejala-gejala penyakit diabetes apakah pasien tersebut berpotensi mengidap diabetes mellitus atau tidak. Berdasarkan hasil pengujian dengan metode cross validation pada aplikasi RapidMiner menghasilkan nilai akurasi sebesar 91,82%.

Kata Kunci: Diabetes Mellitus, *Data Mining*, Algoritma C4.5, RapidMiner

Abstract

Diabetes mellitus is a chronic disease caused by the failure of the pancreas to produce adequate amounts of the hormone insulin, causing an increase in glucose levels in the blood. Diabetes Mellitus is a hereditary disease, this disease can be passed on from parents to their children, and it is very unfortunate if at a young age you have diabetes. Examination of Diabetes Mellitus in the medical world can be carried out by diagnosing the disease based on the symptoms suffered by the patient which can produce laboratory test data and medical records of symptoms of illness. To minimize the death rate from Diabetes Mellitus, health experts must carry out a diagnosis of the disease as early as possible. Classification can be an early treatment of this disease, this classification uses the C4.5 algorithm method. This study aims to help medical professionals classify patients who have diabetes symptoms. The C4.5 algorithm is an algorithm used to form a decision tree (Decision Tree). The decision tree is a well-known method of classification and prediction. The C4.5 algorithm is a method that can be used to predict and determine the accuracy value in patients with symptoms of diabetes whether these patients have the potential to develop diabetes mellitus or not. Based on the results of testing with the cross validation method on the Rapid Miner application, the accuracy value is 91.82%.

Keywords: *Diabetes Mellitus, Data Mining, C4.5 Algorithm, Rapid Miner.*

Naskah diterima 30 Januari 2021; direvisi 7 Februari 2021; diterbitkan 28 Februari 2021

1. Pendahuluan

Diabetes Mellitus merupakan salah satu penyakit kronis yang mematikan. Menurut (Hairani et al, 2018) Penyakit diabetes merupakan salah satu penyakit paling banyak diderita oleh manusia seluruh dunia Penyakit ini juga merupakan jenis penyakit yang banyak diamati dibanyak negara saat ini. Penyakit ini terus dan menjadi semakin meningkat pada tingkat yang sangat mengkhawatirkan. Angka kejadian Diabetes mellitus meningkat dalam beberapa dekade. Secara umum diperkirakan sebanyak 422 juta jiwa orang dewasa terdiagnosis Diabetes mellitus pada tahun 2014, jumlah ini lebih banyak dibandingkan dengan tahun 1980 (sebanyak 108 juta jiwa). Hal ini disertai dengan peningkatan faktor risiko seperti obesitas dan kebiasaan-kebiasaan dalam kehidupan seseorang yang tidak banyak melakukan aktifitas fisik atau tidak banyak melakukan gerakan. Di Indonesia sebanyak 2,1 % terdiagnosis Diabetes Mellitus (RISKESDAS 2013) dalam (Efendi et al, 2018) dengan prevalensi usia paling banyak terdiagnosis pada usia 55 – 64 tahun.

Menurut Report WHO (Report of a WHO / IDF Consultation 2006) saat ini ada 246 juta penderita diabetes diseluruh dunia, dan jumlah ini diperkirakan akan meningkat menjadi 380 juta pada tahun 2025. Diabetes menyebabkan penyakit atau komplikasi lain yang setiap tahunya mengakibatkan kematian 3,8 juta jiwa. Komplikasi lain yang lebih sering terjadi dan mematikan akibat diabetes adalah serangan jantung dan stroke. Sebagian besar kematian terjadi karena kenaikan kadar glukosa secara terus menerus sehingga mengakibatkan rusaknya pembuluh darah, saraf dan struktur internal lainya.

Pada penelitian ini menggunakan dataset publik, dataset *Early stage diabetes risk prediction dataset* dari UCI *Repository*. Menurut Wahono dalam (Fitriyani, 2018). Penelitian menggunakan dataset publik sangat dianjurkan karena sebanyak 64.79% penelitian menggunakan dataset publik dan sebanyak 35.21% penelitian menggunakan dataset privat. Selain itu, penggunaan dataset publik dapat membuat penelitian berulang, terbantahkan dan diverifikasi (Catal & Diri, 2009) dalam (Fitriyani, 2018).

Dalam penelitian ini, algoritma yang digunakan adalah algoritma C4.5. Algoritma data mining C4.5 merupakan salah satu algoritma yang digunakan untuk melakukan klasifikasi atau pengelompokan dan bersifat prediktif. Kelebihan dari algortima ini mampu menangani atribut dengan tipe diskrit atau kontinu. Menurut Larose dalam (Novandya, 2017), algoritma C4.5 adalah ekstensi Quinlan untuk algoritma ID3 untuk menghasilkan pohon keputusan, algoritma C4.5 rekursif mengunjungi setiap node keputusan, memilih split optimal sampai tidak ada perpecahan lanjut yang memungkinkan. Pada dasarnya konsep dari algoritma C4.5 adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan (rule). Menurut Xindong dalam (Novandya, 2017) mengemukakan bahwa C4.5 adalah algoritma yang cocok untuk masalah klasifikasi prediksi dan data mining. C4.5 memetakan nilai atribut menjadi kelas yang dapat diterapkan untuk klasifikasi baru.

Adapun tujuan dari penelitian ini, yaitu untuk mengimplementasikan algoritma C4.5 untuk menghasilkan nilai akurasi dalam memprediksi penyakit diabetes mellitus, serta menerapkan algoritma C4.5 dalam prediksi penyakit diabetes mellitus.

Diabetes Mellitus

Diabetes melitus merupakan penyakit kronis yang ditandai dengan tingginya kadar gula (glukosa) di dalam darah. Diabetes adalah penyakit gangguan metabolik menahun akibat pankreas memproduksi sedikit insulin, diabetes ini juga dapat diakibatkan tidak efektifnya tubuh dalam menggunakan insulin yang diproduksi. Insulin adalah hormon yang mengatur keseimbangan kadar gula darah. Akibatnya terjadi peningkatan konsentrasi glukosa didalam darah (hiperglikemia). Menurut Jayalakshmi & Santhakumaran dalam (Fatmawati, 2016). Penyakit diabetes disebabkan oleh peningkatan kadar glukosa dalam darah, apabila kadar glukosa darah meningkat dalam jangka waktu yang lama maka akan menyebabkan komplikasi seperti gagal ginjal, kebutaan dan serangan jantung.

Ada beberapa gejala diabetes mellitus yang perlu kita waspadai, yaitu:

a. Polydipsia (cepat haus)
b. Polyuria (banyak buang air kecil)
c. Polyphagia (cepat lapar)
d. Sudden Weight Loss (penurunan berat badan)
e. Weakness (rasa lelah dam lemah yang tidak biasa)
f. Visual Blurring (pandangan kabur)

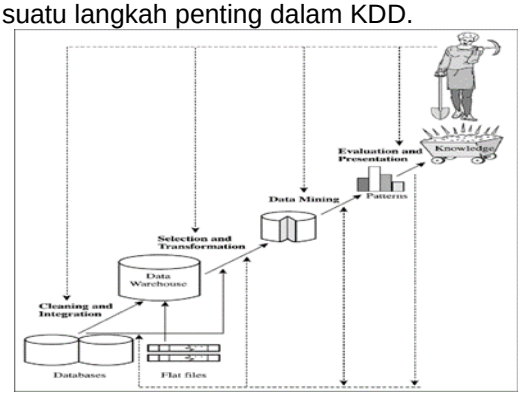
g. Delayed Healing (pemulihan luka yang lama atau sering infeksi)
h. Acanthosis Nigricans (warna kulit gelap)

Menurut (Fatimah, 2015) ada beberapa faktor resiko yang dialami penderita Diabetes Mellitus, yaitu:

a. Obesitas (kegemukan)
b. Hipertensi
c. Riwayat Keluarga Diabetes Mellitus
d. Dislipedimia
e. Umur
f. Riwayat persalinan
g. Faktor Genetik
h. Alkohol dan Rokok

Data Mining

Menurut Tan dalam (Haryati et al., 2015) data mining merupakan proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar. Menurut (Muzakir 2016) dalam (Mirqotussa'adah et al., 2017) Pada bidang kesehatan, *data mining* dapat dimanfaatkan untuk memprediksi suatu penyakit dari data rekam medis pasien. Dengan metode klasifikasi pada *data mining*, data seperti umur, jenis kelamin, tekanan darah dan atribut lainnya, dapat digunakan menjadi faktor pendukung dalam memprediksi kemungkinan pasien terkena suatu penyakit. Menurut Han, Kamber, dan Pei (2011) dalam (Haryati et al., 2015) Istilah data mining kadang disebut juga Knowledge Discovery in Database (KDD). Istilah data mining sering dipakai, mungkin karena istilah ini lebih pendek dari Knowledge Discovery in Database. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi saling berkaitan satu sama lain. Data mining dianggap hanya sebagai suatu langkah penting dalam KDD.



Gambar 1. Tahapan Proses KDD

Menurut (Haryati et al., 2015) Proses KDD secara garis besar dapat dijelaskan sebagai berikut:

a. *Data Cleansing*

Pembersihan data, untuk menghilangkan noise dan data yang tidak konsisten.

b. *Data Integration*

Integrasii data, di mana beberapa sumber data dapat dikombinasikan.

c. *Selection*

Seleksi data, di mana data yang relevan dengan tugas analisis yang diambil dari database.

d. *Data Transformation*

Data transformasi (dimana data diubah dan digabung ke dalam bentuk yang sesuai untuk pertambangan dengan melakukan ringkasan atau agregasi operasi)

e. *Data Mining*

Data mining, merupakan proses esensial dimana metode cerdas diaplikasikan untuk mengekstrak data pola.

f. *Pattern Evolution*

Evaluasi Pola, untuk mengidentifikasi pola yang benar-benar menarik yang mewakili pengetahuan.

g. *Knowledge Presentation*

Presentasi pengetahuan, dimana visualisasi dan teknik representasi pengetahuan digunakan untuk menyajikan pengetahuan hasil data mining kepada pengguna.

Algoritma C4.5

Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan (Decision Tree). Pohon keputusan merupakan metode klasifikasi dan prediksi yang terkenal. Pohon keputusan berguna untuk mengekspolari data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target. Banyak algoritma yang dapat dipakai dalam pembentukan pohon keputusan, antara lain : ID3, CART, dan C4.5. Algoritma C4.5 merupakan pengembangan dari algoritma ID3, Proses pada pohon keputusan adalah mengubah bentuk data (tabel) menjadi model pohon, mengubah model pohon menjadi rule, dan menyederhanakan rule.

Menurut Kusrini dan Lutfi dalam (Haryati et al., 2015) Ada beberapa tahap dalam membuat sebuah decision tree dengan algoritma C4.5, yaitu:

1. Menyiapkan data training. Data training biasanya dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokkan ke dalam kelas-kelas tertentu.

2. Menentukan akar dari pohon, akar akan diambil dari atibut yang terpilih dengan cara menghitung nilai Gain dari masing-masing atribut, nilai Gain yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung nilai Gain dari atribut, hitung dahulu nilai entropy yaitu:

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

Keterangan:
S : himpunan kasus
N : jumlah partisi S
Pi : proporsi dari Si terhadap S

3. Kemudian hitung nilai Gain dengan metode information gain:

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

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

4. Ulangi langkah ke-2 hingga semua tupel terpartisi.
5. Proses partisi pohon keputusan akan berhenti saat:

a. Semua tupel dalam node N mendapat kelas yang sama.
b. Tidak ada atribut di dalam tupel yang dipartisi lagi.
c. Tidak ada tupel di dalam cabang yang kosong.

2. Metode Penelitian

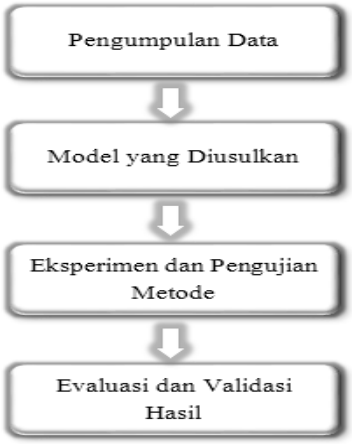
Pada dasarnya semua penelitian akan selalu didahului dengan identifikasi masalah, hal ini berguna untuk peneliti agar fokus pada titik permasalahan. Dalam kasus ini peneliti ingin menganalisa dan mengetahui seberapa besar nilai akurasi yang dihasilkan dalam klasifikasi penyakit diabetes mellitus menggunakan algoritma C4.5. mengukur seberapa besar akurasi yang dihasilkan metode algoritma ini dan bagaimana pola dan cara kerja klasifikasi dalam data mining tersebut.

Analisis Sumber Data

Pada penelitian, peneliti menggunakan dataset publik. Dataset yang digunakan adalah dataset yang diambil dari UCI Repository yang bernama *Early stage diabetes risk prediction dataset*.

Prosedur Penelitian

Prosedur penelitian adalah langkah – langkah yang digunakan sebagai alat untuk mengumpulkan data. Didalam prosedur penelitian ini, membahas tentang metode dan teknik pengumpulan data. Metode penelitian yang digunakan dalam penelitian ini adalah metode eksperimental. Tahapan dalam metode eksperimental yang dilakukan dalah sebagai gambar berikut:



Gambar 2. Prosedur Penelitian

Pengumpulan Data

Tahapan pertama yang dilakukan dalam penelitian ini adalah pengumpulan data. Data yang akan digunakan dalam penelitian ini adalah data *public* yaitu data *Early stage diabetes risk prediction dataset* dari *Uci Machine Learning Repository*. Dataset ini berisikan 520 *record* dengan 17 atribut. Dataset ini dapat diunduh di url: https://archive.ics.uci.edu/ml/machine-learning-databases/00529/ (Islam et al., 2020).

Model yang Diusulkan

Dalam penelitian ini metode yang diusulkan adalah metode klasifikasi prediksi data mining algoritma C4.5 dan Decision Tree terhadap dataset penyakit diabetes mellitus berdasarkan gejala-gejala yang muncul dengan menggunakan RapidMiner. Pengujian model menggunakan Cross Validation, evaluasi dengan Confusion Matrix sehingga dihasilkan akurasi dari metode tersebut.

Eksperimen dan Pengujian Metode

Dalam tahapan desain dan eksperimen algoritma dilakukan dengan menggunakan aplikasi bantu yaitu rapid miner. Aplikasi rapid miner sengaja digunakan karena dapat digunakan dalam berbagai platform seperti windows dan linux. Selain itu aplikasi ini juga

gratis dan tersedia pembaruan setiap bulannya. Dalam penggunaannya aplikasi ini mudah digunakan. Pengguna cukup mempersiapkan dataset kemudian diaplikasikan dengan cara *drag and drop* pada aplikasi untuk mendesain dan melakukan perhitungan. Beberapa algoritma populer dan terbaik juga telah tersedia dalam aplikasi ini.

Evaluasi dan Validasi Hasil

Validasi merupakan proses pengujian performa algoritma. Pada umunya validasi dilakukan dengan mengulang proses perhitungan sampai beberapa kali. Proses validasi dalam penelitian ini menggunakan *cross validation. Cross validation* adalah proses membagi dataset menjadi dua bagian data *training* dan bagian data *testing*. Validasi dan pengujian dilakukan untuk mengetahui tingkat akurasi. Pengukuran akurasi merupakan tahapan untuk membuktikan tingkat performa suatu algoritma terhadap dataset yang digunakan. Dalam penelitian ini digunakan *confusion matrix* sebagai alat ukur performa algoritma klasifikasi. *Confussion matrix* atau matrik kebingungan merupakan sebuah perhitungan yang membandingkan dataset dengan hasil klasifikasi sesuai dengan data sebenarnya dengan jumlah keseluruhan data. Hasil akhir dari matrik ini adalah tingkat akurasi dengan satuan persen (%). Tingkat akurasi ini yang nantinya dijadikan acuan para peneliti terkait performa algoritma klasifikasi tersebut.

3. Hasil dan Pembahasan

Data yang akan digunakan dalam penelitian ini adalah data *public* yaitu data *Early stage diabetes risk prediction dataset* dari *Uci Machine Learning Repository*. Dataset ini berisikan 520 *record* dengan 17 atribut (Islam et al., 2020). Data yang digunakan sebagai data sampel atau data traning sebesar 110 data dengan 16 atribut, yang mana 15 atribut sebagai regular attribute dan 1 atribut sebagai special atribute. Atribute-atribut yang digunakan terdiri dari:

Tabel 1. Tabel Atribut

No	Atribut	Tipe	Nilai atribut
1	Gender	Binominal	Male, Female
2	Poliuria	Binominal	Yes, No
3	Polidipsia	Binominal	Yes, No
4	Sudden weight loss	Binominal	Yes, No
5	Weakness	Binominal	Yes, No
6	Polyphagia	Binominal	Yes, No
7	Genital thrush	Binominal	Yes, No
8	Visual blurring	Binominal	Yes, No
9	Itching	Binominal	Yes, No
10	Irritability	Binominal	Yes, No
11	Delayed healing	Binominal	Yes, No
12	Partial paresis	Binominal	Yes, No
13	Muscle stiffness	Binominal	Yes, No
14	Alopecia	Binominal	Yes, No
15	Obesity	Binominal	Yes, No
16	Class	Binominal	Positive, Negative

Eksperimen dan Pengujian Metode

Pada tahap ini dilakukan eksperimen dan pengujian metode yang digunakan yaitu mebghitung dan mendapatkan rule-rule yang ada pada algoritma yang diusulkan yaitu Algoritma C4.5. Yang pertama tentukan dahulu akar pohon untuk decision tree. Akar pohon ditentukan dengan cara mencari nilai gain dari masing-masung atribut. Nilai gain atribut yang paling tinggi akan menjadi akar pohon. Namun sebelum mencari nilai gain, cari terlebih dahulu nilai entropy.

Untuk yang pertama cari dahulu nilai dari entropy totalnya. Dari 110 data di dapat jumlah positif sebesar 58 dan negatif sebesar 52, maka:

Entropy (Total) = ((-58/110)*log2(58/110)+
(-52/110)*log2(52/110))
= 0,997852777

Kemudian hitung nilai entropy dan gain dari masing-masing atribut, sebagai berikut:

Gender (Male) =((24/74)*log2(24/74) +
(-50/74)*log2(50/74))

= 0,909022156

Gender (Female) =((34/36)*log2(34/36) + (-2/36)*log2(2/36))
= 0,309543429

Gain= 0,997852777– ((74/110)*0,909022156) – ((36/110)* 0,309543429)
= 0,285023658

Lakukan perhitungan yang sama untuk mencari nilai entrophy dan nilai gain dari atribut lainnya. Setelah dicari nilai entropy dan gain untuk keseluruhan atribut, maka didapat nilai-nilai seperti tabel berikut:

Tabel 2. Tabel Perhitungan Nilai Entropy dan Gain

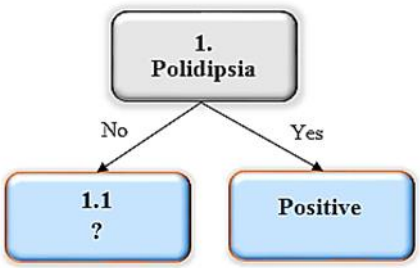
		JUMLAH	POSITIVE	NEGATIVE	ENTHROPY	GAIN
TOTAL		110	58	52	0,997852777	
Gender						0,285023658
	Male	74	24	50	0,909022156	
	Female	36	34	2	0,309543429	
Poliuria						0,41631709
	Yes	48	45	3	0,337290067	
	No	62	14	48	0,770629069	
Polidipsia						0,550542071
	Yes	44	44	0	0	
	no	66	14	52	0,745517643	
Sudden Weight Loss						0,31484342
	Yes	44	40	4	0,439496997	
	No	66	18	48	0,845350937	
Weakness						0,095461814
	Yes	63	43	20	0,901598235	
	No	47	15	32	0,903453555	
Polyphagia						0,100971163
	Yes	51	37	14	0,847861745	
	No	59	21	38	0,939254721	
Genital Thrush						0,000613041
	Yes	16	9	7	0,988699408	
	No	94	49	45	0,998693408	
Visual Blurring						0,073432039
	Yes	50	35	15	0,881290899	
	No	60	23	37	0,96036227	
Itching						0,000763391
	Yes	51	26	25	0,999722648	
	No	59	32	27	0,994813175	
Irritability						0,081784617
	Yes	21	18	3	0,591672779	
	No	89	40	49	0,99261089	
Delayed Healing						0,000464366
	Yes	48	26	22	0,994984828	
	No	62	32	30	0,999249248	
Partial Paresis						0,210779952
	Yes	47	39	8	0,658191266	
	No	63	19	44	0,883222559	
Muscle Stiffness						0,006679702
	yes	37	22	15	0,974024864	
	No	73	36	37	0,999864833	
Alopecia						0,097868501
	Yes	39	28	11	0,858230793	
	No	71	47	24	0,922919288	
Obesity						0,017398947
	Yes	17	12	5	0,873981048	
	No	93	46	47	0,999916596	

Berdasarkan hasil perhitungan nilai entropy dan gain di atas, dapat dilihat bahwa atribut polidipsia memiliki nilai gain terbesar yakni sebesar 0,550542071, maka atribut polidipsia digunakan sebagai akar pohon atau root. Kemudian setelah root ditentukan, buat cabang sesuai partisi. Atribut polidipsia memiliki 2 cabang yaitu “Yes” dan “No” sehingga dari root polidipsia akan dibentuk 2 cabang, sebagai berikut:

Tabel 3. Tabel Polidipsia

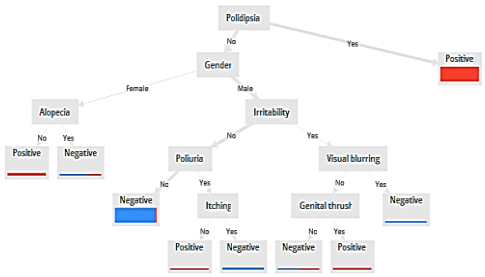
Polidipsia	Jumlah	Positive	Negative
Yes	44	44	0
No	66	14	52

Dari data di atas, diketahui bahwa partisi “Yes” memiliki nilai 44 pada class positive dan memiliki nilai 0 pada class negative. Partisi “No” memiliki nilai 14 pada class positive dan nilai 52 pada class negative. Sehingga partisi “Yes” menjadi daun pohon atau leaf yang berada pada class positive. Dengan demikian maka dapat dibuat pohon keputusan sebagai berikut:



Gambar 3. Decison Tree Polidipsia

Kemudian untuk mencari class dari partisi cabang-cabang selanjutnya, hitung kembali nilai-nilai entropy dan gain dari setiap artibut. Dan atribut dengan nilai gain terbesar akan menjadi daun pohon atau leaf selanjutnya.
Setelah dilakukan perhitungan secara keseluruhan, maka diperoleh pohon keputusan seperti gambar berikut ini:



Gambar 4. Pohon keputusan (*Decision Tree*)

Decision Tree di atas dapat dibaca dengan rule seperti berikut:

IF Polidipsia = No
AND Gender = Female
AND Alopecia = No
THEN Positive

IF Polidipsia = No
AND Gender = Female
AND Alopecia = Yes
THEN Negative

IF Polidipsia = No
AND Gender = Male
AND Irritability = No
AND Poliuria = No
THEN Negative

IF Polidipsia = No
AND Gender = Male
AND Irritability = No
AND Poliuria = Yes
AND Itching = No
THEN Positive

IF Polidipsia = No
AND Gender = Male
AND Irritability = No
AND Poliuria = Yes
AND Itching = Yes
THEN Negative

IF Polidipsia = No
AND Gender = Male
AND Irritability = No
AND Visual blurring = No
AND Genital thrush = No
THEN Negative

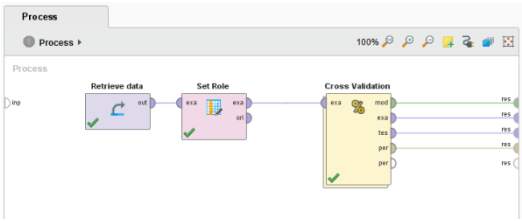
IF Polidipsia = No
AND Gender = Male
AND Irritability = No
AND Visual blurring = No
AND Genital thrush = Yes
THEN Positive

IF Polidipsia = No
AND Gender = Male
AND Irritability = No
AND Visual blurring = Yes
THEN Positive

IF Polidipsia = Yes
THEN Positive

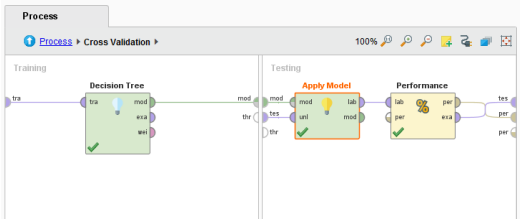
Pengujian Metode Algoritma C4.5

Berikut ini pengujian menggunakan cross validation pada aplikasi RapidMiner



Gambar 5. Pengujian *Cross Validation*

Model yang telah dibentuk, diuji tingkat akurasinya dengan memasukkan atau uji yang berasal dari data training dengan modeling “Tree-Decision Tree”, lalu testing dengan Apply Model & Validation “Performance – Predictive % Performance (Clasification) pada aplikasi RapidMiner untuk menguji tingkat akurasi.



Gambar 6. Validation Model Algoritma C4.5

3.3. Evaluasi dan Validasi Hasil

Setelah data diolah maka dapat diuji tingkat akurasinya untuk melihat kinerja dari metode C4.5. penelitian ini bertujuan untuk melihat akurasi analisis data penyakit diabetes mellitus dan memprediksi apakah gejala-gejala yang dialami penderita dapat memprediksi apakah penderita tersebut positive atau negative penyakit diabetes mellitus. Pengujian tingkat akurasi ini dilakukan dengan menggunakan confussion matrix.

accuracy: 91.82% +/- 9.04% (micro average: 91.82%)

	true Negative	true Positive	class precision
pred. Negative	49	6	89.09%
pred. Positive	3	52	94.55%
class recall	94.23%	89.66%	

Gambar 7. Nilai Akurasi dari Confusion Matrix Algoritma C4.5

Berdasarkan gambar di atas dapat dilihat bahwa class negative menghasilkan class recall sebesar 94,23% dan class presision sebesar 89,09%, sedangkan class positive menghasilkan class recall sebesar 89,65% dan class presision sebesar 94,55%.

4. Kesimpulan

Berdasarkan penelitian dan pengujian yang dilakukan pada penerapan decision tree dengan menggunakan algoritma C4.5, maka didapat kesimpulan sebagai berikut:

a. Prediksi penyakit diabetes mellitus dengan class negative dan positive dapat dilakukan dengan metode decision tree algoritma C4.5. Penerapan metode algoritma C4.5 pada prediksi penyakit diabetes mellitus menghasilkan nilai akurasi sebesar 91,82%
b. Penerapan algoritma C4.5 pada prediksi penyakit diabetes mellitus merupakan salah satu solusi yang baik. Algoritma C4.5 juga sangat efektif dan fleksibel jika digunakan dala proses pengklasifikasian. Dengan metode ini, maka dapat mempermudah para medis untuk mengklasifikasi penyakit diabetes mellitus berdasarkan gejala-gejala yang dialami oleh penderita.

Peneliti mengusulkan beberapa saran kepada pengembang pendidikan dan peneliti selanjutnya, yaitu:

a. Melakukan evaluasi lebih lanjut terhadap data-data mengenai penyakit diabetes mellitus dan membandingkan dengan beberapa algoritma lainnya seperti CART, Linear Discriminant Analysis sehingga dapat mendapatkan model terbaru dan rule terbaru serta nilai akurasi yang lebih baik lagi.
b. Melakukan *feature engineering* atau rekayasa fitur terhadap fitur-fitur data yang digunakan untuk mendapatkan akurasi yang lebih maksimal. Seperti membuat web atau aplikasi supaya lebih tersistem dalam mendapatkan nilai akurasnya.

Referensi

Catal, C., & Diri, B. (2009). A systematic review of software fault prediction studies. *Expert Systems with Applications*, *36*(4), 7346–7354. https://doi.org/10.1016/j.eswa.2008.10.027

Fatmawati, F. (2016). Perbandingan Algoritma Klasifikasi Data Mining Model C4.5 Dan Naive Bayes Untuk Prediksi Penyakit Diabetes. *None*, *13*(1), 50–59.

Fitriyani, F. (2018). Metode Bagging Untuk Imbalance Class Pada Bedah Toraks Menggunakan Naive Bayes. *Jurnal Kajian Ilmiah*, *18*(3), 278. https://doi.org/10.31599/jki.v18i3.281

Haryati, S., Sudarsono, A., & Suryana, E. (2015). *IMPLEMENTASI DATA MINING UNTUK MEMPREDIKSI MASA STUDI MAHASISWA MENGGUNAKAN ALGORITMA C4 . 5*. *11*(2), 130–138.

Islam, M. M. F., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2020). Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. In M. Gupta, D. Konar, S. Bhattacharyya, & S. Biswas (Eds.), *Computer Vision and Machine Intelligence in Medical Image Analysis* (pp. 113–125). Springer Singapore.

Novandya, A. (2017). Penerapan Algoritma Klasifikasi Data Mining Dalam. *KNiST*, *XIV*(2), 120–129.

Fatimah, R. N. (2015). Diabetes Mellitus Tipe 2. 4(5), 1-9.

Efendi, M. S., Wibawa, H. A. (2018). Prediksi Penyakit Diabetes Menggunakan Algoritma ID3 dengan Pemilihan Atribut Terbaik *(Diabetes Prediction using ID3 Algorithm with Best Attribute Selection). VI(1), 29-35.*

Hairani, Nugraha, G. S., Abdillah, M. N., & Innuddin, M. (2018). Komparasi Akurasi Metode *Correlated Naive Bayes Classifier* Dan *Naive Bayes Classifier* Untuk Diagnosis Penyakit Diabetes. 3(1), 6-11.

Mirqotussa'adah, M., Muslim, M. A., Sugiharti, E., Prasetiyo, B., & Alimah, S. (2017). Penerapan Dizcretization dan Teknik Bagging Untuk Meningkatkan Akurasi Klasifikasi Berbasis Ensemble pada Algoritma C4.5 dalam Mendiagnosa Diabetes. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, *8*(2), 135.