ANALISIS PERFORMA ALGORITMA C.45 DAN KLASIFIKASI *DECISION-TREE* DALAM MEMPREDIKSI PENYAKIT DIABETES

Syenira Sheila[1)], Arya Anandya Diphayana[2)]
[1, 2] Fakultas Teknik/Sistem dan Teknologi Infomasi/Universitas Negeri Jakarta
email: syenira19sye@gmail.com, diphayana11@gmail.com

Abstract

This research aims to analyze the performance of the C.45 algorithm with the Decision Tree classification in predicting diabetes. Diabetes occurs when the body cannot use the insulin it produces effectively, resulting in an increase in the concentration of glucose in the blood. People with diabetes have increased every year. In this research, the diabetes dataset was collected using secondary data collection methods in which the dataset was obtained from the Kaggle dataset repository site in .csv format published by Alex Teboul. The dataset is then processed in the pre-processing stage to produce three dataset scenarios which are then used in testing the C.45 algorithm with the Decision Tree classification in the Rapidminer software. The result of this research indicates that the performance of each dataset scenarios using the C.45 algorithm results in the form of dataset scenario 3 having the best accuracy results of 86.05% among the three dataset scenarios, while dataset scenario 2 obtain the best precision and recall results with the result of 69.82% and 80.44% respectively. By using the AUC accuracy value, the performance of the C.45 algorithm with the Decision Tree classification in predicting diabetes categorize as good classification. Among the three dataset scenarios, dataset scenario 3 has the best AUC accuracy value with a score of 0.776.

Keywords: *diabetes, C.45 algorithm, decision-tree classification, rapidminer.*

Abstrak

Penelitian ini bertujuan untuk menganalisis performa algoritma C.45 dengan klasifikasi *Decision-Tree* dalam memprediksi penyakit diabetes. Penyakit diabetes terjadi pada saat tubuh tidak dapat menggunakan insulin yang diproduksi secara efektif, sehingga terjadi peningkatan konsentrasi glukosa dalam darah. Penderita penyakit diabetes mengalami peningkatan setiap tahunnya. Dalam penelitian ini, dataset diabetes dikumpulkan dengan metode pengumpulan data sekunder di mana dataset diperoleh dari situs situs Kaggle *dataset repository* dalam format .csv yang diterbitkan oleh Alex Teboul. Dataset kemudian diproses dalam tahap *pre-processing* sehingga menghasilkan tiga skenario dataset yang kemudian digunakan dalam pengujian algoritma C4.5 dengan klasifikasi *Decision Tree* pada software Rapidminer. Hasil penelitian ini menunjukkan bahwa performa setiap skenario dataset yang menggunakan algoritma C4.5 diperoleh hasil bahwa skenario dataset 3 memiliki hasil *accuracy* terbaik sebesar 86.05% di antara ketiga skenario dataset, sedangkan skenario dataset 2 memperoleh hasil *precision* dan *recall* terbaik dengan masing-masing hasil sebesar 69.82% dan 80.44%. Dengan menggunakan nilai akurasi AUC, diperoleh bahwa performa algoritma C4.5 dengan klasifikasi *Decision-Tree* termasuk ke dalam kategori klasifikasi yang baik. Di antara ketiga skenario dataset, skenario dataset 3 memiliki nilai akurasi AUC yang terbaik dengan perolehan nilai yaitu 0.776.

Kata Kunci: diabetes, algoritma C.45, klasifikasi *decision-tree,* rapidminer.

1. PENDAHULUAN

Diabetes adalah penyakit gangguan metabolik yang mana konsentrasi glukosa dalam darah meningkat tinggi (*Hiperglikemia*). Hal ini disebabkan pankreas tidak memproduksi cukup insulin atau insulin yang diproduksi tidak digunakan secara efektif oleh tubuh. Diabetes atau sering dikenal sebagai kencing manis terbagi menjadi dua tipe, yaitu tipe 1 yang dikarenakan produksi insulin yang tidak memadai oleh pankreas dan diabetes tipe 2 yang disebabkan kegagalan sel dalam respon efektif terhadap insulin yang diproduksi oleh pankreas. Penderita penyakit diabetes mengalami peningkatan setiap tahunnya[1]. Dilaporkan oleh *World Health Organization* (WHO) bahwa kurang lebih 350 juta orang merupakan penderita penyakit diabetes. Kematian yang disebabkan oleh diabetes telah memakan hampir 1,5 juta nyawa pada tahun 2012 dan mayoritas terjadi di negara-negara berkembang. Pada tahun 2030, WHO memprediksi bahwa penyakit diabetes menjadi satu dari tujuh faktor penyebab utama kematian di dunia[2]. Dalam kurun waktu 8 tahun yang akan datang, penderita diabetes di Indonesia akan mengalami kenaikan tiga kali lipat yang akan menjadikan Indonesia sebagai negara urutan keempat dunia dalam masalah penyakit diabetes setelah Amerika Serikat, China, dan India[3].

Pentingnya prediksi awal penyakit diabetes salah satunya dikarenakan penyakit diabetes merupakan faktor yang dapat mempercepat terjadinya berbagai penyakit kardiovaskular. Salah satu faktor yang memicu peningkatan gula darah adalah gaya hidup yang kurang sehat seperti konsumsi rokok, konsumsi alkohol, kurangnya latihan jasmani dan kurangnya konsumsi sayur dan buah[3]. Dalam penelitian terdahulu, telah banyak dilakukan penelitian terkait algoritma C4.5 seperti algoritma *Decision-Tree* C4.5 yang digunakan untuk mendiagnosis penyakit [4]. Dalam jurnal tersebut, Sigit Abdillah sebagai penulis mengklasifikasikan stroke atau non-stroke menggunakan teknik klasifikasi data mining dengan algoritma C4.5 dan *Decision-Tree*. Dari hasil penelitian tersebut, didapatkan hasil akurasi pada data *training* yaitu 82.31% dan pada data *testing* yaitu 76.92% menggunakan perhitungan *confusion matrix*. Penerapan algoritma C4.5 dengan klasifikasi *Decision-Tree* dalam penelitian yang dilakukan oleh B A K Permana, et al[5] mampu menghasilkan hasil akurasi performa mencapai 90.38, sehingga dapat dinyatakan bahwa model algoritma yang digunakan dalam penelitian tersebut termasuk dalam kategori sangat baik. Dalam dataset penelitian tersebut digunakan 13 atribut fitur dan didapatkan atribut fitur polydipsia memiliki nilai *gain* tertinggi yaitu 0.440. Penelitian tersebut juga menyatakan bahwa *Decision-Tree* merupakan metode yang stabil untuk klasifikasi dikarenakan waktu konstruksi yang cepat dan interpretasi yang baik, dengan algoritma C4.5 yang memberikan hasil lebih baik dibandingkan algoritma *Iterative Dichotomised 3* (ID3).

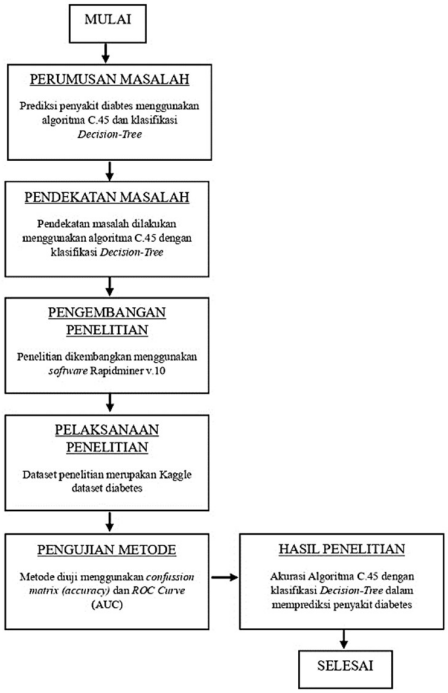
Berdasarkan beberapa penelitian diatas, algoritma C4.5 dengan klasifikasi *Decision-Tree* memiliki hasil yang baik. Penelitian ini bertujuan untuk menganalisis performa algoritma C.45 dengan klasifikasi *Decision-Tree* dalam memprediksi penyakit diabetes dengan 3 skenario dataset dan hasil klasifikasi dilihat menggunakan *confusion matrix* dan kurva ROC.

2. METODE PENELITIAN

Tahapan dalam penelitian ini dimulai dari perumusan masalah yaitu melakukan perbandingan dengan penelitian yang telah dilakukan sebelumnya mengenai prediksi penyakit diabetes menggunakan algoritma yang sejenis yaitu algoritma 4.5 dengan klasifikasi *Decision-Tree.*

Berdasarkan hal-hal yang disebutkan diatas tersebut, penelitian dilakukan menggunakan klasifikasi *Decision-Tree* dan algoritma C4.5 dalam memprediksi penyakit diabetes. Penelitian ini menggunakan *software* Rapidminer studio v. 10 dan pengujian metode dilakukan dengan melihat hasil klasifikasi menggunakan *confusion matrix* dan kurva ROC. Berikut bagan alur penelitian yang dilakukan.

Gambar 1. Alur Penelitian



Dua pendekatan utama yaitu pendekatan kualitatif dan pendekatan kuantitatif, digunakan dalam penelitian ini agar diperoleh pemahaman yang lebih mendalam dan hasil pengujian yang maksimal. Pendekatan kualitatif digunakan dalam "*grounded theory research*" yang mana pengertian dan konsep yang ada dikembangkan menjadi sebuah teori. Data yang digunakan bersifat deskriptif yang berupa gejala-gejala yang ditafsir menurut budaya yang bersangkutan dengan mencari makna semantis universal dari gejala yang sedang diteliti. Sedangkan, pendekatan kuantitatif digunakan dalam pengujian teori, memberikan deskripsi statistik, serta menunjukkan fakta dan hubungan antar variabel. Penekanan dalam pendekatan kuantitatif ini merupakan hal-hal yang bersifat kongkret, uji empiris dan fakta yang nyata dalam menaksir hasil pengujian[6].

Metode yang digunakan dalam penelitian ini dalam pengumpulan data yaitu menggunakan metode pengumpulan data sekunder. Dataset penelitian diperoleh dari situs Kaggle *dataset repository* dalam format .csv yang diterbitkan oleh Alex Teboul. Dataset tersebut merupakan hasil survey indikator penyakit diabetes yang dilakukan melalui telepon. Survey tersebut bernama *The Behavioral Risk Factor Surveillance System* (BRFSS) yang diadakan setiap tahun oleh *Centers for Disease* Control (CDC). Dataset mentah merupakan respon dari 441.455 orang dan memiliki 330 fitur. Selanjutnya dataset tersebut diproses dalam tahap pre-processing sehingga didapatkan tiga skenario dataset bersih yang berbeda.

Skenario dataset 1:
Dataset yang telah dibersihkan terdiri dari 253.680 data hasil survey dengan atribut label 3 kelas yaitu 0 jika tidak menderita diabetes atau menderita diabetes hanya selama mengandung, 1 jika prediabetes dan 2 jika menderita diabetes; dan 21 kelas atribut fitur. Dataset ini memiliki ketidakseimbangan kelas.

Skenario dataset 2:
Dataset yang telah dibersihkan terdiri dari 70.692 data hasil survey dengan atribut label 2 kelas yaitu 0 jika tidak menderita diabetes dan 1 jika menderita diabetes; dan 21 kelas atribut fitur. Dataset ini memiliki perbandingan yang seimbang antara responden yang tidak memiliki diabetes dan yang memiliki diabetes atau pre-diabetes.

Skenario dataset 3:
Dataset yang telah dibersihkan terdiri dari 253.680 data hasil survey dengan atribut label label 2 kelas yaitu 0 jika tidak menderita diabetes dan 1 jika menderita diabetes; dan 21 kelas atribut fitur. Dataset ini memiliki ketidakseimbangan kelas.

Tabel 1. Atribut, Tipe Data dan Nilai Kategori Dalam Indikator Penyakit Diabetes

Atribut Label Skenario Dataset 1		
Atribut	**Tipe Data**	**Nilai**
Diabetes_012	Polynomial	0
		1
		2

Atribut Label Skenario Dataset 2		
Atribut	**Tipe Data**	**Nilai**
Diabetes_binary	Binomial	0
		1

Atribut Label Skenario Dataset 3		
Atribut	**Tipe Data**	**Nilai**
Diabetes_binary	Binomial	0
		1

Atribut Fitur		
Atribut	**Tipe Data**	**Nilai**
HighBP	Binomial	0
		1
HighChol	Binomial	0
		1
CholCheck	Binomial	0
		1
BMI	Real	12-98
Smoker	Binomial	0
		1
Stroke	Binomial	0
		1
HeartDiseaseorAttack	Binomial	0
		1
PhyActivity	Binomial	0
		1
Fruits	Binomial	0
		1
Veggies	Binomial	0
		1
HvyAlcoholConsump	Binomial	0
		1
AnyHealthcare	Binomial	0
		1
NoDocbcCost	Binomial	0
		1
GenHlth	Integer	1-5
MentHlth	Integer	1-30
PhysHlth	Integer	1-30
DiffWalk	Binomial	0
		1
Sex	Binomial	0
		1
Age	Polynomial	1-13
Education	Polynomial	1-6
Income	Polynomial	1-8

Skenario dataset 1, 2 dan 3 memiliki atribut label yang berbeda namun memiliki atribut fitur yang sama.

Berikut ini merupakan penjelasan dari atribut Tabel 1.

1. Diabetes_02
   Diabetes_02 merupakan atribut label dari skenario dataset 1 yang melingkupi responden yang memiliki diabetes dengan kategori 0 (tidak menderita diabetes atau hanya selama mengandung), 1 (prediabetes) dan 2 (menderita diabetes).
2. Diabetes_binary
   Diabetes_binary merupakan atribut label dari skenario dataset 2 dan 3 yang melingkupi responden yang memiliki diabetes dengan kategori 0 (tidak menderita diabetes) dan 1 (menderita diabetes).
3. *HighBP* (Tekanan Darah Tinggi)
   HighBP merupakan atribut yang melingkupi responden memiliki tekanan darah tinggi atau tidak.
4. *HighChol* (Kolesterol Tinggi)
   HighChol merupakan atribut yang melingkupi responden memiliki kolestrol darah tinggi atau tidak.
5. *CholCheck* (Cek Kolesterol)
   CholCheck ialah atribut yang mengindikasikan apakah responden pernah atau tidak mengecek kadar kolestrol dalam 5 tahun terakhir.
6. BMI (*Body Mass Index*)
   BMI ialah atribut yang didapatkan dengan cara membagi berat bedan dengan tinggi badan.

$$BMI = \frac{Berat\ Badan\ (kg)}{(Tinggi\ badan)^2\ (m)} \quad \text{.....(1)}$$

7. *Smoker* (Perokok)
   Smoker ialah atribut yang mengindikasikan apakah responden pernah atau tidak merokok 100 batang rokok (5 bungkus rokok) seumur hidupnya.
8. *Stroke*
   Stroke ialah atribut yang mengindikasikan apakah responden memiliki stroke (pernah) atau tidak.
9. *HeartDiseaseorAttack* (Penyakit Jantung atau Serangan Jantung)
   HeartDiseaseorAttack merupakan atribut yang mengindikasikan apakah responden memiliki *Coronary Heart Disease* (CHD)*/ Myocardial Infarction* (MI) atau tidak.
10. *PhysActivity* (Aktivitas Fisik)
    PhysActivity merupakan atribut yang mengindikasikan apakah responden melakukan aktivitas fisik yang dilakukan dalam 30 hari terakhir (tidak termasuk bekerja) atau tidak.
11. *Fruits* (Konsumsi Buah-buahan)
    Fruits merupakan atribut yang mengindikasikan apakah responden mengonsumsi buah-buahan minimal sekali dalam sehari atau tidak.
12. *Veggies* (Konsumsi Sayuran)

Veggies merupakan atribut yang mengindikasikan apakah responden mengonsumsi sayur-sayuran minimal sekali dalam sehari atau tidak.

13. *HvyAlcoholConsump* (Konsumsi Alkohol Berat)
    HvyAlcoholConsump merupakan atribut yang mengindikasikan apakah responden peminum alkohol berat (untuk pria dewasa meminum 14 botol alkohol per minggu dan untuk wanita dewasa meminum 7 botol alkohol per minggu) atau tidak.
14. *AnyHealthcare* (Perawatan Kesehatan)
    AnyHealthcare merupakan atribut yang mengindikasikan apakah responden memiliki *health care* coverage/asuransi kesehatan/HMO atau tidak.
15. *NoDocbcCost* (Tidak Dapat Memeriksa Diri ke Dokter)
    NoDocbcCost merupakan atribut yang mengindikasikan apakah dalam rentang waktu 12 bulan responden dapat memeriksakan ke dokter atau tidak dapat memeriksakan diri dikarenakan biaya.
16. *GenHlth* (Kesehatan General)
    GenHlth merupakan atribut yang mengindikasikan tingkat kesehatan general responden, 1 jika sangat sehat sekali, 2 jika sangat sehat, 3 jika sehat, 4 jika cukup sehat, 5 jika tidak sehat.
17. *MentHlth* (Kesehatan Mental)
    MentHlth merupakan atribut yang mengindikasikan berapa banyak hari responden merasa kesehatan mentalnya tidak baik, termasuk didalamnya stress, depresi, permasalahan emosi dalam 30 hari terakhir.
18. *PhysHlth* (Kesehatan Fisik)
    PhysHlth merupakan atribut yang mengindikasikan berapa banyak hari responden merasa kesehatan fisiknya tidak baik, termasuk didalamnya sakit fisik dan cedera dalam 30 hari terakhir.
19. *DiffWalk* (Kesulitan Berjalan)
    DiffWalk merupakan atribut yang mengindikasikan apakah responden memiliki kesulitan berjalan/menaiki tangga atau tidak
20. *Sex* (Jenis Kelamin)
    Sex ialah atribut yang melingkupi jenis kelamin responden dengan kategori 0 jika wanita dan 1 jika pria.
21. *Age* (Umur)
    Age merupakan atribut yang melingkupi umur responden yang dikategorikan dalam 13 kategori umur.
22. *Education* (Pendidikan)
    Education merupakan atribut yang melingkupi tingkat pendidikan responden yang dikategorikan dalam 6 tingkat pendidikan.
23. *Income* (Pendapatan)
    Income merupakan atribut yang melingkupi tingkat pendapatan responden yang dikategorikan dalam 8 tingkat pendapatan.

Data Mining

Data mining digunakan untuk memperoleh pengetahuan tersembunyi dalam suatu database. Data mining juga merupakan sebuah proses mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar dengan memanfaatkan teknik statistik, matematika, kecerdasan buatan dan pembelajaran mesin[7]. Istilah lain dari data mining yaitu *Knowledge Discovery in Database* (KDD), yang mana merupakan proses pengumpulan, pemakaian data historis yang digunakan dalam menemukan keteraturan, pola atau hubungan yang cenderung tidak disadari keberadaannya dalam suatu dataset berukuran besar. Hasil dari data mining dapat digunakan untuk memperbaiki pengambilan keputusan di masa yang akan datang[8].

Rapidminer

Rapidminer merupakan *software* yang bersifat terbuka (*open source*). *Software* yang sebelumnya bernama YALE ini digunakan dalam menganalisis data mining, text mining dan menganalisis prediksi. Rapidminer membuat keputusan yang paling baik dikarenakan menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna. Rapidminer memiliki kurang lebih 500 operator data mining, termasuk operator untuk input, output, data preprocessing dan visualisasi. Rapidminer merupakan *software* yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang dapat diintegrasikan pada produknya sendiri. Rapidminer dapat bekerja di semua sistem operasi sebab *software* tersebut ditulis menggunakan bahasa java[9].

Algoritma Decision Tree C45

Decision Tree merupakan model prediksi menggunakan struktur berhierarki atau struktur pohon. *Decision-Tree* adalah struktur flowchart yang menyerupai Tree (pohon), dimana setiap simpul internal menandakan suatu tes pada

atribut, setiap cabang merepresentasikan hasil tes, dan simpul daun merepresentasikan kelas atau distribusi kelas[10].

Algoritma C.45 merupakan pengembangan dari algoritma ID3 yang mana digunakan untuk mengklasifikasikan berupa pohon keputusan. Pengembangan dilakukan dalam mengatasi *missing* data, data *continue* dan *pruning*[11]. Input algoritma ini yaitu berupa *training samples dan samples. Training samples* merupakan data yang dijadikan contoh dan digunakan untuk membangun sebuah "pohon" yang kebenarannya telah teruji, sedangkan *samples* digunakan sebagai parameter dalam melakukan klasifikasi yang mana *samples* berbentuk dalam *field-field* data[10]. Algoritma C4.5 merupakan salah satu solusi pemecahan kasus yang sering digunakan dalam pemecahan masalah pada teknik klasifikasi. Pemilihan atribut sebagai simpul, baik simpul akar (*root*) atau simpul internal didasarkan pada nilai *Gain* tertinggi dari atribut-atribut yang ada. Penghitungan nilai *Gain* digunakan rumus Persamaan 2.

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

….(2)

S : Himpunan kasus
A : Atribut
n : Jumlah partisi himpunan atribut A
|| : Jumlah kasus pada partisi ke- i
| S | : Jumlah kasus dalam S

Untuk menghitung nilai Entropy dapat dilihat pada Persamaan 3.

$$Entropy\ (S) = \sum_{i=1}^{n} - pi * \log_2 (pi) \quad \text{……(3)}$$

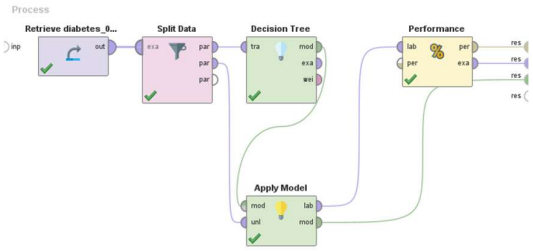
n : Jumlah partisi S
pi : Proporsi dari terhadap S

3. HASIL DAN PEMBAHASAN

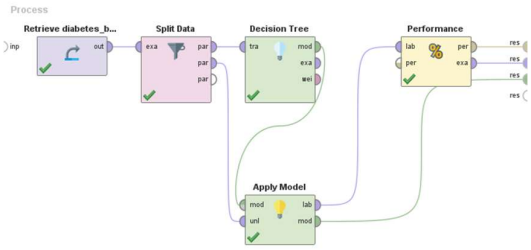
Pengujian Algoritma C4.5

Pengujian algoritma C4.5 dengan klasifikasi *Decision-Tree* menggunakan dataset dengan tiga skenario yang berbeda pada software Rapidminer dapat dilihat pada Gambar 2, Gambar 3 dan Gambar 4. Ketiga gambar tersebut menunjukkan proses perhitungan menggunakan model algoritma C4.5 menggunakan *software* Rapidminer. Pada proses pertama dilakukan penginputan dataset, kemudian dilakukan *split data* menjadi data training yaitu sebesar 80% dan data testing yaitu sebesar 20%. Selanjutnya data training masuk ke model fit yang diwakili oleh *Decision Tree* yang menggunakan kriteria *gain_ratio* yaitu algoritma C4.5 dan data testing masuk ke model predict yang diwakili oleh *Apply Model*. Pada *Apply Model* dilakukan pengujian model C4.5 pada data testing. Proses terakhir yaitu performa untuk mengetahui hasil pengujian menggunakan algoritma C4.5.

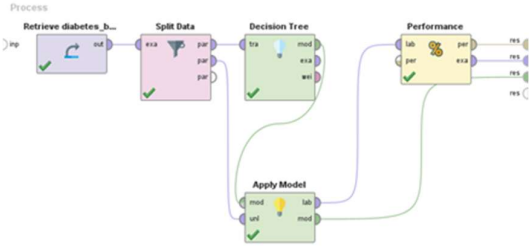
Gambar 2. Pengujian Algoritma C4.5 Skenario Dataset 1



Gambar 3. Pengujian Algoritma C4.5 Skenario Dataset 2



Gambar 4. Pengujian Algoritma C4.5 Skenario Dataset 3



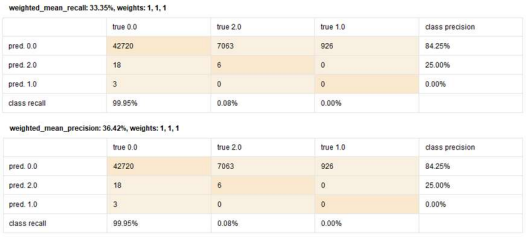
Hasil Pengujian Algoritma C4.5

Hasil performa algoritma C4.5 dengan tiga skenario dataset berbeda menggunakan software Rapidminer ditunjukkan pada Gambar 5, Gambar 6 dan Gambar 7.

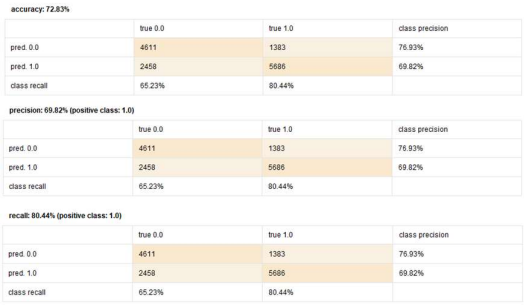
Gambar 5. Hasil Pengujian Algoritma C4.5 Skenario Dataset 1

accuracy: 84.21%

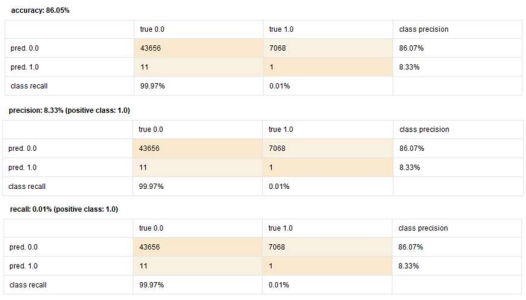
	true 0.0	true 2.0	true 1.0	class precision
pred. 0.0	42720	7063	926	84.25%
pred. 2.0	18	6	0	25.00%
pred. 1.0	3	0	0	0.00%
class recall	99.95%	0.08%	0.00%	



Gambar 6. Hasil Pengujian Algoritma C4.5 Skenario Dataset 2



Gambar 7. Hasil Pengujian Algoritma C4.5 Skenario Dataset 3



Pada Tabel 2 kesimpulan performa setiap skenario dataset menggunakan algoritma C4.5 dapat dilihat dan diperoleh hasil bahwa skenario dataset 3 memiliki hasil *accuracy* terbaik di antara ketiga skenario dataset tersebut, sedangkan hasil *precision* dan *recall* terbaik diperoleh pada skenario dataset 2.

Tabel 2. Performa 3 Skenario Dataset Klasifikasi Penyakit Diabetes Menggunakan Algoritma C4.5

	Accuracy	**Precision**	**Recall**
Skenario Dataset 1	84.21%	33.35%	36.42%
Skenario Dataset 2	72.83%	69.82%	80.44%
Skenario Dataset 3	86.05%	8.33%	0.01%

Dari ketiga skenario dataset tersebut, dapat dilihat bahwa ketidakseimbangan dataset juga dapat berpengaruh pada hasil klasifikasi. Nilai *precision* dan *recall* yang terlalu kecil menjadi salah satu kemungkinan yang dihasilkan akibat ketidakseimbangan data. Perbedaan yang terlalu signifikan antara kelas mayoritas dan minoritas dapat menyebabkan terjadinya *imbalance ratio* atau rasio tidak seimbang. Kesalahan klasifikasi kelas minoritas dapat terjadi akibat kelas minoritas dianggap sebagai kelas mayoritas[12].

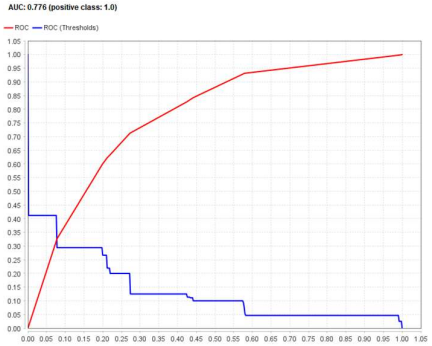
Hasil Visualisasi Kurva ROC (AUC)

Hasil dari prediksi penyakit diabetes menggunakan tiga skenario dataset berbeda dapat dilihat dalam kurva ROC (*Receive Operating Characteristic*). Skenario dataset yang dapat ditampilkan dalam bentuk kurva ROC hanya skenario dataset 2 dan 3 saja, sebab dalam skenario dataset 1 terdapat 3 kelas atribut label yang digunakan dalam dataset tersebut. Berikut hasil kurva ROC(AUC) yang dapat dilihat pada Gambar 8 dan Gambar 9.

Gambar 8. Hasil Visualisasi ROC *Curve* (AUC) Skenario Dataset 2

Pada Gambar 8 dapat dijelaskan bahwa nilai AUC yaitu 0.788 (*positive class*: 1).

Gambar 9. Hasil Visualisasi ROC *Curve* (AUC) Skenario Dataset 3



Pada Gambar 9 dapat dijelaskan bahwa nilai AUC yaitu 0.776 (*positive class*: 1).

Dalam tesisnya, Yasilnacar melakukan penyajian terhadap penilaian AUC untuk mengategorikan model prediksi yang dihasilkan sebagai berikut dalam Tabel 3 [13].

Tabel 3. Klasifikasi Nilai AUC

Nilai AUC	Keterangan
>0.9 – 1	Luar biasa
>0.8 – 0.9	Sangat baik
>0.7 – 0.8	Baik
>0.6 – 0.7	Cukup baik
0.5 – 0.6	Tidak baik

Berdasarkan kedua akurasi nilai AUC tersebut, dapat disimpulkan bahwa algoritma C4.5 dengan klasifikasi *Decision-Tree* yang digunakan untuk menganalisis dan memprediksi penyakit diabetes menggunakan dataset diatas termasuk dalam kategori klasifikasi yang baik. Dari hasil akurasi Roc diatas, diperoleh nilai akurasi AUC skenario dataset 3 lebih baik dibandingkan nilai akurasi AUC skenario dataset 2.

4. PENUTUP

Berdasarkan pembahasan diatas, dapat dinyatakan bahwa kesimpulan penelitian ini menggunakan algoritma C4.5 dan klasifikasi *Decision-Tree* dalam memprediksi penyakit yaitu sebagai berikut.

Kesimpulan

Dalam penelitian ini dibandingkan tiga skenario dataset untuk menguji performa algoritma C4.5 dengan klasifikasi *Decision-Tree* menggunakan *software* Rapidminer dalam memprediksi penyakit diabetes. Dataset diperoleh dari situs Kaggle. Ketiga skenario dataset yang diuji memiliki atribut label yang berbeda yaitu skenario dataset 1 yang memiliki atribut label 3 kelas, skenario dataset 2 dengan atribut label 2 kelas, dan skenario dataset 3 dengan atribut label 2 kelas; dan atribut fitur yang sama yaitu sebanyak 21 kelas atribut. Hasil pengujian diperoleh menggunakan *confusion matrix* (*accuracy*) dan kurva ROC. Dari hasil pengujian performa algoritma C4.5 dengan klasifikasi *Decision-Tree* dapat diambil kesimpulan bahwa algoritma C4.5 memiliki nilai performa yang baik atau dapat dikatakan termasuk kategori klasifikasi yang baik dalam memprediksi penyakit diabetes. Hal ini didasarkan dari hasil kurva ROC yang diperoleh yaitu >0.7–0.8. Di antara ketiga skenario dataset tersebut, diperoleh hasil bahwa skenario dataset 3 memiliki hasil *accuracy* terbaik sebesar 86.05% sedangkan skenario dataset 2 memperoleh hasil *precision* dan *recall* terbaik dengan masing-masing hasil sebesar 69.82% dan 80.44%. Sedangkan dari hasil akurasi kurva ROC diperoleh skenario dataset 3 memiliki nilai akurasi AUC terbaik yaitu sebesar 0.776. Dari ketiga skenario dataset tersebut, dapat disimpulkan pula bahwa keseimbangan dataset juga berperan penting dalam hasil klasifikasi. Nilai *precision* dan *recall* yang terlalu kecil menjadi salah satu kemungkinan yang dihasilkan akibat ketidakseimbangan data.

Saran

Penulis mengharapkan model yang dibangun dapat disempurnakan di masa yang akan datang dengan mengombinasikan dengan model lainnya, memasukkan lebih banyak data dari sumber lain dan mempertimbangkan indikator-indikator kesehatan lainnya seperti pandangan kabur, kekakuan otot, dan lain sebagainya.

Ucapan terima kasih

Penulis mengucapkan terima kasih kepada Ibu Murien Nugraheni, S.T., M. Cs. selaku Dosen Mata Kuliah Data Mining yang telah membimbing penulis dalam penulisan jurnal ini.

5. REFERENSI

[1] K. K. Sajida Perveen, Muhammad Shahbaz, Aziz Guergachi, “Performance Analysis of Data Mining Classification Techniques to Predict Diabetes,” in *Procedia Computer Science*, 2016, pp. 115–121, doi: https://doi.org/10.1016/j.procs.2016.04.016.

[2] J. I. Marzuki, K. Mataram, and N. T. Bar, “KOMPARASI AKURASI METODE CORRELATED NAIVE BAYES CLASSIFIER DAN NAIVE BAYES CLASSIFIER UNTUK DIAGNOSIS PENYAKIT DIABETES Hairani , Gibran Satya Nugraha , Mokhammad Nurkholis Abdillah , Muhammad Innuddin InfoTekJar (Jurnal Nasional Informatika dan Teknolog,” *InfoTekJar (Jurnal Nas. Inform. dan Teknol. Jaringan)*, vol. 3, no. 1, pp. 6–11, 2018.

[3] S. N. R. Toharin, S. CAHYATI, W. H. M

Kes, and Z. M. H. Kes, "Hubungan Modifikasi Gaya Hidup Dan Kepatuhan Konsumsi Obat Antidiabetik Dengan Kadar Gula Darah Pada Penderita Diabetes Melitus Tipe 2 Di Rs Qim Batang Tahun 2013," *Unnes J. Public Heal.*, vol. 4, no. 2, pp. 153–161, 2015.

[4] A. Sigit, "PENERAPAN ALGORITMA DECISION TREE C4.5 UNTUK DIAGNOSA PENYAKIT STROKE DENGAN KLASIFIKASI DATA MINING PADA RUMAH SAKIT SANTA MARIA PEMALANG," Universitas Dian Nuswantoro, 2015.

[5] B. A. C. Permana, R. Ahmad, H. Bahtiar, A. Sudianto, and I. Gunawan, "Classification of diabetes disease using decision tree algorithm (C4.5)," *J. Phys. Conf. Ser.*, vol. 1869, no. 1, 2021, doi: 10.1088/1742-6596/1869/1/012082.

[6] Jonathan Sarwono, "Memadu Pendekatan Kuantitatif dan Kualitatif," *J. Ilm. Manaj. Bisnis*, vol. 9, no. 2, pp. 119–132, 2010, [Online]. Available: www.jonathansarwono.info.

[7] E. et al Turban, "Decision Support Systems and Intelligent Systems (Sistem Pendukung Keputusan dan Sistem Cerdas," *Andi Offset*, 2005.

[8] B. Santosa, *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu, 2007.

[9] M. H. Luluk Elvitaria, "MEMPREDIKSI TINGKAT PEMINAT EKSTRAKURIKULER PADA SISWA SMK ANALISIS KESEHATAN ABDURRAB MENGGUNAKAN ALGORITMA C4.5 (STUDI KASUS: SMK ANALIS KESEHATAN ABDURRAB)," *RABIT (Jurnal Teknol. dan Sist. Inf. Univrab)*, vol. 2, no. 2, pp. 220–233, 2017.

[10] I. Br Ginting, Selvia Lorena., Zarman, Wendi., Hamidah, "Analisis Dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Memprediksi Masa Studi Mahasiswa Berdasarkan Data Nilai Akademik," 2014.

[11] N. Shita, Rizky Tahara., & Marliani, "Aplikasi Data Mining Dengan Metode Classification Berbasis Algoritma C4.5.," in *Seminar Nasional Sistem Informasi Indonesia,* p. 201.

[12] & M. S. M. Gagah Gumelar, Norlaila2, Quratul Ain, Riza Marsuciati, Silvi Agustanti Bambang, Andi Sunyoto, "Kombinasi Algoritma Sampling dengan Algoritma Klasifikasi untuk Meningkatkan Performa Klasifikasi Dataset Imbalance," in *Prosiding SISFOTEK*, 2021, pp. 5 (1), 250–255.

[13] E. K. Yasilnacar, "The Application of Computational Intelligence to Landslide Suspectibility Mapping in Turkey," University of Melbourne, 2005.