IMPLEMENTASI ALGORITMA *CART* DALAM KLASIFIKASI PENYAKIT DIABETES

Fida Maisa Hana[a]*, Widya Cholid Wahyudin [b], Saiful Ulya[c], Deka Setia Negara[d]
[abcd]Universitas Muhammadiyah Kudus. Jalan Ganesha No.I Kudus. Indonesia
Email : fidamaisa@umkudus.ac.id

Abstrak

Diabetes adalah suatu penyakit metabolik yang disebabkan oleh kurangnya produksi insulin pada pankreas, hal ini mengakibatkan ketidakseimbangan gula dalam darah sehingga konsentrasi kadar gula darah meningkat. Penderita penyakit diabetes dari tahun ke tahun semakin meningkat. Estimasi dari International Diabetes Federation (IDF), terdapat 382 juta orang yang menderita penyakit diabetes pada tahun 2012. Diperkirakan pada tahun 2035 jumlahnya meningkat menjadi 592 juta orang. Pencatatan terhadap penyakit ini perlu dilakukan agar dapat dilakukan pencegahan. Salah satu pencatatan yang bisa dilakukan adalah dengan memanfaatkan teknik klasifikasi data mining. Penelitian ini melakukan implementasi algoritma *CART (Classification And Regression Trees)* dalam klasifikasi penyakit diabetes. Hasil akurasi tertinggi didapatkan pada saat klasifikasi menggunakan algoritma *CART* tanpa *pruning* dan *prepruning* yaitu sebesar 100%. Sedangkan jika dengan *pruning* dan *prepruning* menghasilkan akurasi sebesar 96.15%.

Kata Kunci : data mining, klasifikasi, diabetes, *CART*.

Abstract

Diabetes is a metabolic disease caused by a lack of insulin production in the pancreas, this results in an imbalance of sugar in the blood so that the concentration of blood sugar levels increases. Patients with diabetes from year to year are increasing. Estimates from the International Diabetes Federation (IDF), there are 382 million people suffering from diabetes in 2012. It is estimated that by 2035 the number will increase to 592 million people. Recording of this disease needs to be done so that prevention can be done. One of the records that can be done is by utilizing data mining classification techniques. This study implements the CART (Classification And Regression Trees) algorithm in the classification of diabetes. The highest accuracy results were obtained when classification using the CART algorithm without pruning and prepruning was 100%. Meanwhile, pruning and prepruning produce an accuracy of 96.15%.

***Keywords** : data mining, classification, diabetes, CART.*

I. PENDAHULUAN

Diabetes merupakan satu dari banyak penyakit yang dikarenakan oleh pola makan dan gaya hidup yang tidak baik. Diabetes adalah suatu penyakit metabolik yang disebabkan oleh kurangnya produksi insulin pada pankreas, hal ini mengakibatkan ketidakseimbangan gula dalam darah sehingga konsentrasi kadar gula darah meningkat (Kemenkes, 2014). Menurut data diketahuai bahwa 1 dari 2 orang penderita diabetes tidak sadar jika dirinya mengidap penyakit diabetes, seharusnya 80% kejadian diabetes dapat dicegah. Jika dilakukan kontrol yang baik dan penderita penyakit diabetes menjalani pola hidup sehat, maka penderita diabetes dapat berumur panjang(IDF, 2015).

Penderita penyakit diabetes dari tahun ke tahun semakin meningkat. Estimasi dari International Diabetes Federation (IDF), terdapat 382 juta orang yang menderita penyakit diabetes pada tahun 2012. Diprediksi pada tahun 2035 jumlahnya bertambah menjadi 592 juta orang. Diprediksi dari 382 juta orang tersebut, 175 juta di antaranya belum terdiagnosis, sehingga terancam berkembang progresif menjadi komplikasi

tanpa disadari dan tanpa pencegahan (Kemenkes, 2014).

Dengan adanya masalah ini, perlu disikapi dengan adanya pendeteksian sejak dini penyakit diabetes. Deteksi penyakit diabetes sejak dini diharapkan dapat menurunkan resiko komplikasi pada pasien diabetes diwaktu mendatang. Guna mengetahui pasien penderita penyakit diabetes sejak dini, Pencatatan terhadap penyakit ini perlu dilakukan agar dapat dilakukan pencegahan. Salah satu pencatatan yang bisa dilakukan adalah dengan memanfaatkan teknik klasifikasi dengan data mining.

Data mining merupakan sebuah proses penemuan relasi yang berarti, pola, dan kecondongan dengan mengamati sekumpulan data dalam jumlah besar yang terdapat dalam media penyimpanan dengan memanfaatkan metode pengenalan pola seperti metode statistik dan matematika (Larose & Larose, 2014). Data mining adalah proses pencarian informasi atau pola unik dan menarik dalam data yang terpilih dengan memanfaatkan metode atau algoritma tertentu. Data mining memiliki variasi metode atau algoritma yang banyak. Penetapan algoritma yang tepat sangat bergantung pada tujuan dan proses Knowledge Discovery in Database (KDD)(Muzakir & Wulandari, 2016). Adapun metode yang biasanya operasikan pada data mining antara lain: deskripsi atau penggambaran, prediksi atau ramalan, clustering, klasifikasi dan asosiasi, dan estimasi.

Klasifikasi merupakan sebuah proses untuk menciptakan fungsi atau model menjelaskan kelas pada data atau konsep guna untuk memprediksi kelas dari sebuah objek yang labelnya belum didapatkan (Putra & Chan, 2018). Konsep klasifikasi merupakan satu bagian dari metode data mining yang mempunyai pekerjaan inti menjalankan analisis prediksi (Praningki & Budi, 2018). Pada penelitian ini, teknik klasifikasi dimanfaatkan untuk meramal orang mana yang mengidap penyakit diabetes dan tidak. Beberapa algoritma dapat digunakan untuk perhitungan proses klasifikasi. Beberapa algoritma klasifikasi diantaranya adalah decission tree C4.5, naive bayes, dan k-nearest neighbor (KNN)(Kurniawan & Ivandari, 2017), *decision tree* ID3(Yusa et al., 2016), *decision tree CART* (Alverina et al., 2018).

Decision tree atau pohon keputusan adalah pendekatan yang sangat populer dan umum digunakan untuk klasifikasi. Pengklasifikasi pohon keputusan ini dimulai dengan set pelatihan dengan label kelasnya yang terkait. Node root adalah fitur utama. Setiap simpul internal merepresentasikan atribut tes, setiap cabang mencerminkan hasil tes dan setiap simpul daun mencermikan label kelas. Untuk mengidentifikasi label kelas untuk sampel yang tidak diketahui, Pengklasifikasi pohon keputusan akan melacak jalur dari akar ke simpul daun, yang menyimpan label kelas untuk sampel tersebut (Jain et al., 2015). Algortima *decision tree* yang sering digunakan adalah ID3, C4.5 & *CART* (Bashir et al., 2015).

CART atau kepanjangan dari classification dan regression tree) adalah sebuah algoritma decion tree yang mengukur tingkat ketidakmurnian untuk data yang diberikan dan membangun pohon biner dimana setiap node internal mengeluarkan tepat dua kelas untuk atribut tertentu (Bashir et al., 2015). Dalam penelitian yang dilakukan oleh (Hariati et al., 2018) dengan judul penelitian "Penerapan Algoritma *CART Decision tree* pada Penerima Program Bantuan Pemerintah Daerah Kabupaten Kutai Kartanegara" menghasilkan akurasi sebesar 98,18%.

Merujuk pada penjelasan di atas maka penelitian ini akan melakukan implementasi algoritma *CART* dalam klasifikasi penyakit diabetes.

II. Landasan Teori

A. Data Mining

Data Mining mewakili kompleks teknologi yang digunakan oleh banyak disiplin ilmu: matematika, statistik, ilmu komputer, fisika, teknik, biologi, dan banyak disiplin ilmu lainnya. Data mining bermanfaat bagi berbagai aplikasi dalam berbagai ranah ilmu yang berbeda, yaitu: bisnis, sains, kesehatan, industri, dan teknik. Data mining adalah ilmu mengeksplorasi kumpulan data besar untuk

mengekstraksi informasi implisit, yang sebelumnya tidak diketahui dan berpotensi memiliki manfaat. Sejak tahun 1990, sudah mulai dikenal istilah data mining hal ini karena kebutuhan mengolah data adalah hal yang bermanfaat dan sangat perlu dilakukan (Gorunescu, 2011)

Ilmu Data mining adalah perpaduan ilmu dari *artificial intelligence*, statistik, dan penelitian basis data yang selalu meningkat. Menurut Larose metode data mining merupakan sebuah proses menentukan ikatan yang mengandung arti, pola, dan keterkaitan dengan mengolah kelompok data. Dalam data mining terdapat enam metode yang biasa di jalankan yaitu ramalan atau prediksi, penggambaran atau deskripsi, klasifikasi, estimasi, asosiasi dan clustering (Larose & Larose, 2014).

B. Klasifikasi

Proses mencari sebuah karakteristik data dan dipetakan dalam kelas-kelas sesuai dengan karakteristiknya masing-masing disebut dengan klasifikasi. Pada klasifikasi proses mencari karakteristik sebuah objek dilakukan, selanjutnya objek dengan karakteristik yang sama dimasukan ke dalam salah satu kelas yang sudah diartikan terlebih dahulu(Larose & Larose, 2014). Proses klasifikasi adalah proses menghitung data yang ada sebelumnya atau disebut juga data training dengan data baru atau data testing. Proses ini akan menghasilkan kemungkinan dalam data testing.

Dalam klasifikasi dataset yang digunakan harus memiliki label atau atribut tujuan. Meramal objek kelas pada setiap persoalan dalam data adalah tujuan dari klasifikasi. Dimulai dengan satu set data di mana kelas dikenal adalah sebuah tugas klasifikasi. Adapun Jenis masalah klasfikasi paling sederhana adalah klasifikasi biner (Putra & Chan, 2018). Beberapa algoritma dapat digunakan untuk perhitungan proses klasifikasi. Algoritma klasifikasi diantaranya adalah Decission Tree C4.5, Naive Bayes, dan k-nearest neighbor (KNN) (Kurniawan & Ivandari, 2017).

C. Algoritma *CART*

CART (Classification And Regression Trees) merupakan algoritma yang dikembangkan oleh Leo Breiman, Jerome H. Fridman, Richard S. Olshen dan Charles J pada tahun 1984. *CART* merupakan salah satu metode atau algoritma teknik eksplorasi data decision tree. (Timofeev, 2004). Metode klasifikasi *CART* merupakan metode klasifikasi data mining yang nonparametrik yang bermanfaat untuk memperoleh sekelompok data yang akurat sebagai penciri dari suatu pengklasifikasian.

Metode *CART* ini terdiri dari dua metode yaitu metode pohon klasifikasi dan pohon regresi. *CART* akan menghasilkan pohon klasifikasi (classification trees) Jika variabel dependen yang dimiliki bertipe kategorik. Sedangkan *CART* akan menghasilkan pohon regresi (regression trees) jika variabel dependen yang dimiliki bertipe kontinu atau numerik.(Prabawati et al., 2019)

Dalam pembentukan algoritma *CART* terdapat tiga proses, yaitu pembentukan pohon klasifikasi, pemotongan pohon klasifikasi dan penentuan pohon klasifikasi optimum. Rincian tahapan pembentukan algoritma *CART* adalah sebagai berikut(Prabawati et al., 2019):

1) Pembentukan pohon klasifikasi

Tahap pertama yang harus dikerjakan adalah penentuan pemilih tiap simpul dengan cara menentukan variabel dan threshold. Proses pembentukan pohon klasifikasi terdiri dari:

a) Penentuan Pemilah

Guna mencari himpunan data yang lebih homogen, maka sampel data learning akan dipilah. Indeks gini akan digunakan untuk memilih data karena indeks gini berfungsi untuk melepaskan kelas dengan anggota paling besar atau kelas terpenting dalam simpul terlebih dahulu. Adapun persamaan indeks gini adalah:

$$i(t) = \sum_{i,j-1} p(j|t)p(i|t), i \neq j$$

Dimana $p(j|t)$ merupakan proporsi kelas j pada simpul t. Pemilahan yang terpilih akan membangun sebuah himpunan kelas yang disebut dengan simpul. Simpul tersebut akan melakukan pemilahan secara rekursif sampai diperoleh terminal nodes. Tahap selanjtunya adalah memilih kriteria *goodness of split*

untuk mengevaluasi pemilah dari pemilah s pada simpul t dengan persamaan:

$$\phi(s,t) = \Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R)$$

t_L= cabang kiri dari nokhtah keputusan t

t_R= cabang kanan dari nokhtah keputusan t

Pemilah yang menghasilkan $\phi(s,t)$ lebih tinggi merupakan pemilah terbaik.

b) Penentuan Simpul Terminal

Pada saat jumlah pengamatan pada simpul kurang dari standar yang telah ditentukan (N_{min}) atau ketika sudah mencapai batasan tingkat kedalaman (*depth*) dalam pohon maksimal makan pengembangan pohon akan berhenti.

c) Penandaan Label Kelas

Tahap selanjutnya adalah penandaan lebel kelas. Label kelas pada simpul terminal disesuaikan menurut aturan jumlah terbanyak, yaitu jika:

$$p(j_0|t) = max_j \frac{N_j(t)}{N(t)}$$

Label kelas untuk simpul terminal t adalah j_0 yang menghasilkan nilai perkiraan kesalahan pengklasifikasian pada simpul t yang paling kecil sebesar:

$$r(t) = 1 - max_i p(j|t)$$

Pemangkasan pohon klasifikasi guna mencegah terjadinya Overfitting , maka dilakukan Pemangkasan pohon. Overfitting yaitu sebuah situasi saat algoritma menghasilkan pengklasifikasian yang sangat sesuai dengan data training namun kehilangan kemampuan untuk mengeneralisasikan kejadian yang tidak direpresentasikan selama pelatihan. Pemangkasan pohon dapat dilakukan dengan pengukuran cost complexity minimum dengan rumus:

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|$$

R_a(T) adalah kombinasi linear biaya dan kompleksitas pohon yang dibentuk dengan menambahkan cost penalty bagi kompleksitas terhadap biaya kesalahan klasifikasi pohon. Selanjutnya, dilakukan pencarian pohon bagian T*(a)*<T*max* yang meminimumkan R_a(T) yaitu:

$$R_a(T(a)) = min_{T<T_{max}} R_a(t)$$

2) Penentuan pohon klasifikasi optimum

Pohon klasifikasi yang terbentuk bisa saja akan memiliki tingkat kompleksitas yang tinggi. Oleh sebab itu, pengoptimalan terlebih dahulu dianggap perlu sebelum digunakan untuk mengklasifikasikan data baru. Optimasi pohon akan memilih ukuran pohon yang tepat dan memotong nodes yang tidak signifikan.

D. Confussion Matrix

Confusion Matrix merupakan sebuah hasil evaluasi dari sebuah klasifikasi data mining yang diwujudkan dalam sebuah tabel (Gorunescu, 2011). *Confusion matrix* adalah metode yang banyak dipakai untuk menghitung nilai akurasi. Pengukuran kinerja menggunakan *confusion matrix* melimiki empat istilah sebagai gambaran dari hasil klasifikasi. Adapun keempat istilah tersebut yaitu :

1. *False Positive* (FP), yaitu data negatif tapi terprediksi sebagai data positif.
2. *False Negative* (FN), yaitu data positif yang terprediksi sebagai data negatif.
3. *True Positive* (TP), yaitu data positif yang terprediksi benar.
4. *True Negative* (TN), yaitu data negatif yang terprediksi dengan benar.

dengan klasifikasi yang sebenarnya. Bentuk *Confusion Matrix* secara umum dapat dicermati pada tabel dibawah ini:

Tabel 1. Confussion Matrix

Classification		***Predicted class***	
		Class : Yes	***Class : No***
Observed Class	*ClassYes*	***A(True Positive)***	***B(False Negative)***
	Class No	***C(False Positive)***	***D(True Negative)***

Untuk menghitung akurasi digunakan rumus sebagai berikut:

$$Akurasi = \frac{TP + TN}{TP + FN + FP + TN} \times 100\%$$

Algoritma klasifikasi pastinya berusaha untuk menghasilkan model yang menghasilkan akurasi yang baik. Kinerja model dari algoritma klasifikasi ditentukan pada saat model didahapkan pada data *testing*, karena rata-rata model yang dipakai dapat memprediksi dengan benar pada semua data yang menjadi data *training*nya (Istiawan & Khikmah, 2019).

Sensitivitas atau Recall adalah rasio prediksi benar positif dipadukan dengan keseluruhan data yang benar positif atau menghitung proporsi positif asli yang diramal secara benar sebagai positif. Dalam sensitivitas berkaitan dengan kecakapan eksperimen untuk mengenali hasil yang positif dari sejumlah data yang seharusnya positif. Untuk mengukur sensitivitas atau recall menggunakan persamaan dibawah ini:

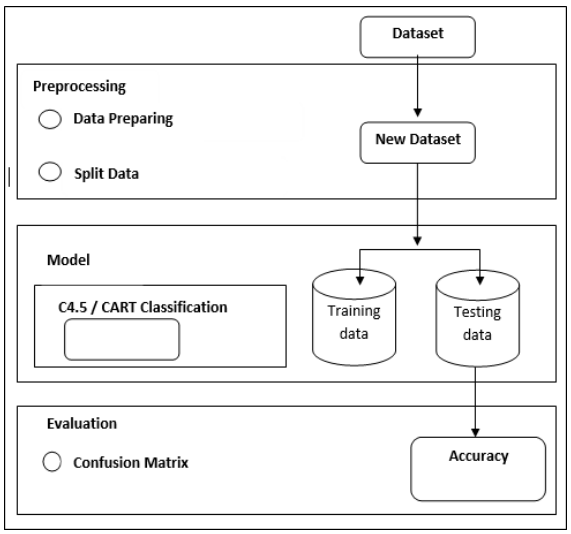
$$Sensitivitas = \frac{TP}{TP+FN}$$

Sedangkan precision adalah rasio ramalan benar positif dipadukan dengan semua hasil yang diprediksi positif. *Precision* menggambarkan matrik untuk menghitung kemampuan sistem dalam menghasilkan data yang penting. *Precison* pada data mining yaitu hasil jumlah data yang true positive dibagi dengan jumlah data yang dikenali sebagai positif. Untuk mengukur *precision* menggunakan persamaan dibawah ini:

$$Precision = \frac{TP}{TP + FP}$$

III. METODE PENELITIAN

Metode penelitian yang akan digunakan dalam penyelesaian penelitian ini adalah metode eksperimen. Secara garis besar penelitian ini akan melakukan perhitungan terhadap dataset dengan menggunakan algoritma klasifikasi *CART*. Adapun kerangka penelitian dari penelitian bisa kita lihat pada Gambar dibawah:



Gambar 1. Kerangka Penelitian

Tahap awal dalam penelitian ini adalah adalah menyiapkan dataset, selanjtunya dilakukan pembagian data uji dan data latih dengan split data, lalu masuk ke tahap klasifikasi dengan algoritma *CART*, selanjutnya dilakukan perbandingan dengan melakukan *pruning* atau tidak, dan yang terakhir hitung akurasi dengan confusion matrix.

A. Menyiapkan Data

Pada penelitian ini menggunakan data dari sumber dataset UCI Machine Learning Repository yang bisa ditemukan di alamat web https://archive.ics.uci.edu/ml/machine-learning-databases/00529/. Dataset yang digunakan adalah *Early stage diabetes risk prediction dataset* dimana file tersebut bernama diabetes_data_upload.csv. Variabel yang dipakai pada penelitian ini adalah sebanyak 17 variabel dengan jumlah data sebanyak 520. Ini termasuk data tentang orang-orang termasuk gejala yang dapat menyebabkan diabetes. Kumpulan data ini dibuat dari kuesioner langsung kepada orang-orang yang baru saja menjadi penderita diabetes, atau yang masih nondiabetes tetapi memiliki sedikit atau lebih gejala. Data dikumpulkan dari pasien dengan menggunakan kuesioner langsung dari Sylhet Diabetes Hospital of Sylhet, Bangladesh.

B. Split Data Otomatis

Setelah menyiapkan 520 data diabetes, tahap selanjutnya adalah membagi data menjadi dua, yaitu data *training* dan data *testing* dengan prosentase 70% untuk data *training* dan 30% untuk data *testing*. Data

training bertindak sebagai pembentuk pola atau model dan data *testing* sebagai penguji model.

C. Metode yang Diusulkan

Model yang digunakan adalah klasifikasi penyakit diabetes adalah algoritmya *CART*. Masing-masing akan dilakukan *pruning* dan *prepruning* dan tanpa *pruning* dan *prepruning*. selanjutnya hasil klasifikasi dievaluasi mengguanakn confussion matrix dan menghasilkan akurasi.

IV. Hasil Dan Pembahasan

Data yang digunakan berjumlah 520 sampel. Masing-masing sampel terdapat 16 Variabel variabel dataset gejala penyakit diabetes dan 1 variabel sebagai kelas penentu klasifikasi data penelitian yang dipakai pada penelitian ini dapat dilihat pada tabel 2 dibawah ini.

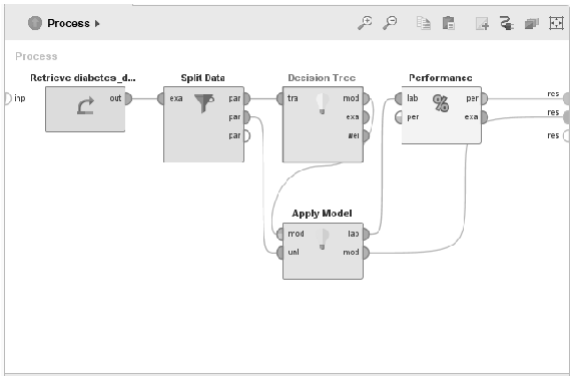
Tabel 2. Variabel data penelitian

No	Atribut	Value
1	Umur	1) 20-35, 2) 36-45, 3) 46-55, 4) 56-65, 5) diatas 65
2	Jenis Kelamin	1.Pria, 2.Wanita
3	*Polyuria*	1.Ya, 2.Tidak
4	*Polydipsia*	1.Ya, 2.Tidak
5	*Suddenweight loss*	1.Ya, 2.Tidak
6	*Weakness*	1.Ya, 2.Tidak
7	*Polyphagia*	1.Ya, 2.Tidak
8	*Genital thrush*	1.Ya, 2.Tidak
9	*Visual blurring*	1.Ya, 2.Tidak
10	*Itching*	1.Ya, 2.Tidak
11	*Irritability*	1.Ya, 2.Tidak
12	*Delayed healing*	1.Ya, 2.Tidak
13	*Partial paresis*	1.Ya, 2.Tidak
14	*Muscle stiffness*	1.Ya, 2.Tidak
15	*Alopecia*	1.Ya, 2.Tidak
16	Obesitas	1.Ya, 2.Tidak
17	Kelas	1.Positif, 2.Negatif

Dari jumlah data 520, dibagi menjadi dua dengan presentase 70% data *training* dan 30% data *testing*, didapatkan data *training* dengan jumlah 364 dan data *testing* dengan jumlah 156. Selanjutnya masuk ke tahap klasifikasi.

1. Klasifikasi dengan Algoritma *CART* dengan *pruning* dan *prepruning.*

Berikut ini merupakan proses klasifikasi dengan algoritma *CART* dengan *pruning* dan *prepruning*.



Gambar 2. Proses Klasifikasi *CART* Menggunakan Aplikasi *Rapidminer* 9.7.

Berikut merupakan analisis hasil pengujian dari 156 data *testing* dengan 364 data *training*. Klasifikasi dengan algoritma *CART* dengan *pruning* dan *prepruning* menghasilkan akurasi sebesar 96,15%.

Tabel 3. Hasil akurasi *CART* dengan *pruning* dan *prepruning*

	True Positive	True Negative	Calss Precision
Pred. Positive	71	6	92,21%
Pred. Negative	0	79	100%
Class recall	100%	92,94%	

Adapun untuk *Performance Vektor* klasifikasi dengan algoritma *CART* dengan *pruning* dan *prepruning* adalah sebagai berikut:

PerformanceVector

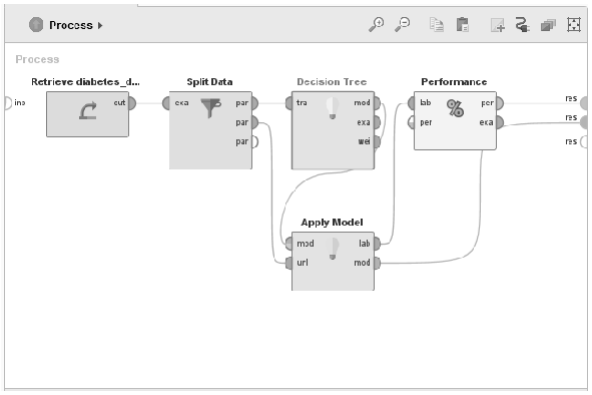
```
PerformanceVector:
accuracy: 96.15%
ConfusionMatrix:
True:   Positive        Negative
Positive:       71      6
Negative:       0       79
precision: 100.00% (positive class: Negative)
ConfusionMatrix:
True:   Positive        Negative
Positive:       71      6
Negative:       0       79
recall: 92.94% (positive class: Negative)
ConfusionMatrix:
True:   Positive        Negative
Positive:       71      6
Negative:       0       79
AUC (optimistic): 0.998 (positive class: Negative)
AUC: 0.985 (positive class: Negative)
AUC (pessimistic): 0.972 (positive class: Negative)
```

Gambar 3. Hasil Performance Vektor CART dengan pruning dan prepruning Aplikasi Rapidminer 9.7.

Pengujian menghasilkan akurasi yang cukup besar yaitu 96.15 % Precision sebesar 100.00%, dan Recall sebesar 92.94%

2. Klasifikasi dengan Algoritma *CART* tanpa *pruning* dan *prepruning*.

Berikut ini merupakan proses klasifikasi dengan Algoritma *CART* tanpa *pruning* dan *prepruning*.



Gambar 4. Proses Klasifikasi CART Menggunakan Aplikasi Rapidminer 9.7.

Sedangkan klasifikasi dengan algoritma *CART* tanpa *pruning* dan *prepruning* menghasilkan akurasi sebesar 100%.

Tabel 4. Hasil akurasi *CART* tanpa *pruning* dan *prepruning*

	True Positive	True Negative	Calss Precision
Pred. Positive	71	0	100%
Pred. Negative	0	85	100%
Class recall	100%	92,94%	

Sedangkan untuk *Performance Vektor* klasifikasi dengan algoritma *CART* tanpa *pruning* dan *prepruning* adalah sebagai berikut:

PerformanceVector

```
PerformanceVector:
accuracy: 100.00%
ConfusionMatrix:
True:   Positive        Negative
Positive:       71      0
Negative:       0       85
precision: 100.00% (positive class: Negative)
ConfusionMatrix:
True:   Positive        Negative
Positive:       71      0
Negative:       0       85
recall: 100.00% (positive class: Negative)
ConfusionMatrix:
True:   Positive        Negative
Positive:       71      0
Negative:       0       85
AUC (optimistic): 1.000 (positive class: Negative)
AUC: 0.500 (positive class: Negative)
AUC (pessimistic): 1.000 (positive class: Negative)
```

Gambar 5. Hasil Performance Vektor CART tanpa pruning dan prepruning pada Aplikasi Rapidminer 9.7.

Pengujian menghasilkan akurasi yang sempurna yaitu sebesar yaitu 100.00 %, Precision sebesar 100.00%, dan Recall sebesar 100.00%

V. KESIMPULAN

Dari dari sumber dataset UCI Machine Learning Repository dengan judul *"Early stage diabetes risk prediction dataset*" yang berjumlah *520* sampel data dengan data training sejumlah 364 dan data testing sejumlah 156 dan 16 variabel dari masing-masing sampel terbukti dapat dijadikan sebagai data untuk klasifikasi penyakit diabates.

Hasil pengujian dengan klasifikasi algoritma *CART* dengan *pruning* dan *prepruning* menghasilkan akurasi yang cukup besar yaitu 96.15 % Precision sebesar 100.00%, dan Recall sebesar 92.94%. sedangkan klasifikasi dengan algoritma *CART* tanpa *pruning* dan *prepruning* menghasilkan akurasi yang sempurna yaitu sebesar yaitu 100.00 %, Precision sebesar 100.00%, dan Recall sebesar 100.00%.

Hasil akurasi tertinggi dan sempurna pada klasifikasi penyakit diabetes didapatkan pada saat klasifikasi menggunakan algoritma *CART* dengan *pruning* dan *prepruning* yaitu sebesar 100%.

VI. DAFTAR PUSTAKA

Alverina, D., Chrismanto, A. R., & Santosa, R. G. (2018). Perbandingan Algoritma C4.5 dan CART dalam Memprediksi Kategori

Indeks Prestasi Mahasiswa. *Jurnal Teknologi Dan Sistem Komputer*, *6*(2), 76–83. https://doi.org/10.14710/jtsiskom.6.2.2018. 76-83

Bashir, S., Qamar, U., Khan, F. H., & Javed, M. Y. (2015). An Efficient Rule-Based Classification of Diabetes Using ID3, C4.5, & CART Ensembles. *Proceedings - 12th International Conference on Frontiers of Information Technology, FIT 2014*, 226–231. https://doi.org/10.1109/FIT.2014.50

Hariati, Wati, M., & Cahyono, B. (2018). Penerapan Algoritma CART Decision Tree pada Penentuan Penerima Program Bantuan Pemerintah Daerah Kabupaten Kutai Kartanegara. *Jurti*, *2*(1), 27–36.

IDF. (2015). IDF Diabetes Atlas. In *Offshore* (Vol.72, Issue 11).

Istiawan, D., & Khikmah, L. (2019). Implementation of C4.5 Algorithm for Critical Land Prediction in Agricultural Cultivation Areas in Pemali Jratun Watershed. *Indonesian Journal of Artificial Intelligence and Data Mining*, *2*(2), 67. https://doi.org/10.24014/ijaidm.v2i2.7569

Jain, L. C., Behera, H. S., Mandal, J. K., & Mohapatra, D. P. (2015). Computational Intelligence in Data Mining - Volume 1: Proceedings of the International Conference on CIDM, 20-21 December 2014. *Smart Innovation, Systems and Technologies*, *31*, 549–562. https://doi.org/10.1007/978-81-322-2205-7

Kemenkes. (2014). Situasi dan Analisis Diabetes. In *American Journal of Medical Genetics, Part A* (Vol. 161, Issue 5, pp. 1058–1063). https://doi.org/10.1002/ajmg.a.35913

Kurniawan, F., & Ivandari. (2017). Komparasi Algoritma Data Mining Untuk Klasifikasi Penyakit Kanker Payudara. *Jurnal Stmik*, *XII*(1), 1–8.

Larose, D. T., & Larose, C. D. (2014). Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition. In *Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition* (Vol. 9780470908).

https://doi.org/10.1002/9781118874059

Muzakir,A.,& Wulandari, R. A. (2016). Model

Data Mining sebagai Prediksi Penyakit Hipertensi Kehamilan dengan Teknik Decision Tree. *Scientific Journal of Informatics*, *3*(1), 19–26. https://doi.org/10.15294/sji.v3i1.4610

Prabawati, N. I., Widodo, & Duskarnaen, M. F. (2019). Kinerja Algoritma Classification a nd Regression Tree (Cart) da lam Mengklasifikasikan Lama Masa Studi Mahasiswa y ang Mengikuti Organisasi d i Universitas Negeri Jakarta Avalaiable at : Avalaiable at : *Jurnal Pinter*, *3*(2), 139–145.

Praningki, T., & Budi, I. (2018). Sistem Prediksi Penyakit Kanker Serviks Menggunakan CART, Naive Bayes, dan k-NN. *Creative Information Technology Journal*, *4*(2), 83. https://doi.org/10.24076/citec.2017v4i2.100

Putra, P. P., & Chan, A. S. (2018). Pengembangan Aplikasi Perhitungan Prediksi Stock Motor Menggunakan Algoritma C 4.5 Sebagai Bagian dari Sistem Pengambilan Keputusan (Studi Kasus di Saudara Motor). *INOVTEK Polbeng – Seri Informatika*, *3*(1), 24. https://doi.org/10.35314/isi.v3i1.296

Timofeev, R. (2004). Classification and Regression Trees (CART) Theory and Applications. *Journal of Gastroenterology and Hepatology (Australia)*, *13*(1), 81–87. https://doi.org/10.1111/j.1440-1746.1998.tb00550.x

Yusa, M., Utami, E., & Luthfi. Taufiq, E. (2016). Evaluasi Performa Algoritma Klasifikasi Decision Tree Id3 ,. *InfoSys Journal*, *4*(1), 23–34. http://e-journal.potensi-utama.ac.id/ojs/index.php/INFOSYS/article/view/136