

Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma *Decision Tree* C4.5

Fida Maisa Hana

Universitas Muhammadiyah Kudus, Jalan Ganesha 1 Purwosari Kudus

fidamaisa@umkudus.ac.id

Diterima : 31 Agustus 2020

Disetujui : 28 September 2020

Abstrak— Diabetes Melitus (DM) merupakan penyakit kronis yang banyak diderita oleh penduduk Indonesia, penyakit ini disebabkan karena kadar glukosa dalam darah di atas nilai normal. Penyakit ini termasuk penyakit yang rumit dan mematikan, oleh karena itu dibutuhkan perawatan medis yang kontinu agar resiko terjadinya komplikasi bisa dihindari. Dari tahun ke tahun jumlah pengidap penyakit diabetes semakin bertambah. Merujuk dari sumber data Federasi Diabetes Internasional, pengidap penyakit diabetes sebanyak 10 juta jiwa pada tahun 2015 di Indonesia, ditahun 2040 diprediksi jumlah warga Indonesia yang terjangkit penyakit diabetes mengalami peningkatan sebesar 16.2 juta jiwa penduduk Indonesia. Oleh sebab itu deteksi penyakit diabetes sangat diperlukan guna merendahkan komplikasi penyakit diabetes di waktu yang akan datang. Algoritma C4.5 adalah salah satu metode yang bisa dimanfaatkan untuk meramal penyakit diabetes. C4.5 Decision Tree adalah algoritma yang bisa menghasilkan keputusan dengan cara membentuk pohon keputusan. Dari hasil Pengujian menghasilkan akurasi yang cukup besar yaitu 97,12 % Precision sebesar 93,02% %, dan Recall sebesar 100,00%.

Keywords : *Data Mining, Klasifikasi, C4.5 Decision Tree, Diabetes Mellitus*

I. PENDAHULUAN

Diabetes mellitus adalah penyakit metabolis yang kronis yang mana pasien penyakit diabetes tidak menghasilkan jumlah insulin yang cukup atau bisa dikatakan tubuh pasien tidak sanggup memanfaatkan insulin dengan baik sehingga menyebabkan gula darah di dalam tubuh mengalami jumlah yang berlebihan, kondisi ini sering kali dirasakan setelah komplikasi terjadi pada organ tubuh [1]. Pasien didiagnosa menderita penyakit diabetes pada saat kadar glukosa darahnya melebihi nilai normal [2]. Penyakit diabetes melitus adalah penyakit yang memiliki kompleksitas tinggi, perawatan medis yang berkelanjutan sangat dibutuhkan guna menurunkan dampak komplikasi dengan pengecekan glikemik [3].

Banyaknya penderita diabetes dari tahun ke tahun semakin bertambah. Pasien diabetes di Indonesia sebesar 10 juta jiwa di tahun 2015. Merujuk pada data Federasi Diabetes Internasional, diprediksi penderita penyakit diabetes di Indonesia akan bertambah menjadi 16.2 pada tahun 2040 [4]. Guna menyikapi masalah ini, perlu adanya pendeteksian sejak dini penyakit diabetes. Deteksi sejak dini diharapkan dapat menurunkan resiko komplikasi pada pasien diabetes diwaktu mendatang. Guna menganalisa pasien pengidap penyakit Diabetes sejak dini, Pencatatan terhadap penyakit ini banyak dilakukan agar dapat dilakukan pencegahan. Salah satu yang pencatatan yang bisa dilakukan adalah dengan memanfaatkan teknik klasifikasi dengan data mining.

Data mining adalah sebuah metode untuk melakukan akuisisi pengetahuan. Dengan data mining, informasi-informasi implisit dan berharga dari sebuah data dapat diekstrak [5]. Menurut [6] data mining yaitu suatu prosedur menciptakan ikatan dimana memiliki arti, pola, dan kecondongan dengan mengamati kelompok data besar yang berada dalam *storage* dengan menggunakan teknik identifikasi pola. Adapun metode yang biasanya operasikan pada data mining antara lain: deskripsi atau penggambaran, prediksi atau ramalan, clustering, klasifikasi dan asosiasi, dan estimasi. [6].

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 [7]. Pada penelitian ini, teknik klasifikasi dimanfaatkan untuk meramal orang mana yang terjangkit penyakit diabetes dan tidak terjangkit. Beberapa algoritma dapat digunakan untuk perhitungan proses klasifikasi. Algoritma klasifikasi diantaranya adalah Decision Tree C4.5, Naive Bayes, dan k-nearest neighbor (KNN) [8].

Berdasarkan penelitian [9] yang melakukan analisa komparasi pembelajaran mesin pada teknik klasifikasi ramalan tenaga kerja non-aktif menghasilkan akurasi paling tinggi pada skenario supply training dan Cross Validation didapatkan dengan menggunakan teknik klasifikasi C4.5 daripada menggunakan algoritma K-Nearest Neighbors (KNN) maupun Naive Bayes (NB). Algoritma C4.5 digunakan untuk memprediksi keputusan dengan menghasilkan pohon keputusan dapat digunakan untuk klasifikasi. Algoritma C4.5 adalah pengembangan dari algoritma ID3, Ross Quinlan adalah sosok yang mengembangkan algoritma C4.5 ini [9]. Kelebihan yang dimiliki oleh algoritma ini adalah dapat dipahami dengan mudah, lentur, dan mengagumkan karena dapat digambarkan dalam bentuk pohon keputusan [10]. Merujuk pada penjelasan di atas maka penelitian ini akan melakukan klasifikasi Penyakit Diabetes dengan memanfaatkan algoritma Decision Tree C4.5.

II. LANDASAN TEORI

1. Data Mining

Metode yang dipakai untuk menggali informasi yang belum ditemukan dengan cara manual dari suatu kumpulan data disebut dengan data mining [11]. Sejak tahun 1990, sudah mulai dikenal istilah data mining hal ini karena kebutuhan mengolah data adalah hal yang bermanfaat dan sangat perlu dilakukan dari bermacam bidang ilmu, dalam bidang kesehatan, akademik, bahkan industri [10]. Ilmu Data mining adalah perpaduan ilmu dari *artificial intelligence*, statistik, dan penelitian basis data yang selalu meningkat. Menurut artikel [6] metode data mining merupakan sebuah proses menentukan ikatan yang mengandung arti, pola, dan keterkaitan dengan mengolah kelompok data. Dalam data mining terdapat 6 metode yang biasa di jalankan yaitu ramalan atau prediksi, penggambaran atau deskripsi, klasifikasi, estimasi, asosiasi dan clustering [6].

a. Deskripsi

Tujuan dari operasi adalah untuk mengenal pola yang terbentuk berulang-ulang pada sekelompok data, selanjutnya menjadikan pola data tersebut membentuk kriteria dan aturan yang gampang dipahami.

b. Prediksi atau ramalan

Ramalan mempunyai kemiripan dengan teknik klasifikasi, tapi di sini data bagi kelas sesuai dengan perilaku atau nilai yang diprediksi pada waktu mendatang.

c. Klasifikasi

Klasifikasi atau pengelompokan adalah teknik mengenali suatu karakteristik dari data dan data tersebut dikelompokkan dalam suatu kelas-kelas.

d. Estimasi

Estimasi atau bisa disebut menerka, teknik ini memiliki kesamaan dengan prediksi, tapi di sini yang diestimasi berupa bilangan numerik.

e. Clustering

Clustering adalah klasterisasi atau pengelompokan data dengan kelas data yang mempunyai karakteristik sama dijadikan satu kelompok dan yang berbeda dikelompokkan yang lain.

f. Asosiasi

Teknik asosiasi adalah mencari atribut yang nampak pada kondisi tertentu atau menghasilkan aturan assosiatif antara suatu kombinasi item.

2. 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 dimasukkan ke dalam salah satu kelas yang sudah diartikan terlebih dahulu [6]. 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 klasifikasi paling sederhana adalah klasifikasi biner [7]. Beberapa algoritma dapat digunakan untuk perhitungan proses klasifikasi. Algoritma klasifikasi diantaranya adalah Decision Tree C4.5, Naive Bayes, dan k-nearest neighbor (KNN) [8].

3. Algoritma C4.5

Algoritma ID3 dikembangkan lagi menjadi Algoritma C4.5 [6] Quinlan adalah sosok yang mengembangkan Algoritma C4.5 ini [10], algoritma C4.5 merupakan teknik klasifikasi yang banyak digunakan oleh peneliti. Sebuah pohon keputusan atau decision tree adalah hasil dari perhitungan algoritma C4.5 [8].

Tahap menghasilkan pohon keputusan menggunakan algoritma C4.5 adalah sebagai berikut:

1. Tahap awal membuat pohon keputusan adalah dengan membentuk akar pohon, selanjutnya data dibedakan sesuai dengan atribut yang serasi untuk dibentuk daun.
2. Tree pruning yaitu proses pemangkasan cabang pohon yang tidak diperlukan oleh pohon yang sudah terbentuk atau dengan kata lain dilakukan penyederhanaan ukuran pohon karena pohon keputusan yang dibentuk biasanya bentuknya besar. Selain itu pemangkasan juga dilakukan dengan tujuan untuk memangkas jumlah eror pada hasil prediksi. Terdapat dua tahap dalam proses pruning yaitu pre-pruning dan post-pruning.

a. Pre-pruning

Pada pre-pruning, pembentukan suatu subtree dihentikan lebih awal (dengan tujuan partisi data training tidak lebih jauh). node berubah menjadi node akhir pada saat berhenti.

- b) Post-pruning, dalam tahap ini dilakukan penyederhanaan pohon dengan cara memangkas cabang-cabang subtree setelah tree selesai dibentuk. Node akhir diperoleh dari node yang jarang dipotong.

3. Pembuatan aturan keputusan

Pohon yang telah dibentuk dibuat aturan keputusan. Dengan tindakan penelusuran dari akar sampai ke daun, aturan tersebut diturunkan dari pohon keputusan.

Tahapan pohon keputusan (decision tree) dibangun adalah sebagai berikut [11]:

1. Pertama adalah memilih atribut sebagai Akar
2. Kedua pembuatan cabang pada setiap nilai
3. Selanjutnya pembagian kasus dalam cabang
4. Pada Setiap Cabang, prosesnya diulangi Sampai semua kasus pada setiap cabang mempunyai kelas yang sama untuk menentukan atribut sebagai akar, disesuaikan pada nilai gain paling tinggi dari atribut-atribut yang ada. Cara mencari nilai gain adalah dengan persamaan dibawah ini:

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

(1)

defini:

S menyatakan himpunan kasus

A menyatakan Atribut

n menyatakan jumlah partisi atribut A

|S_i| menyatakan Jumlah Kasus pada Partisi ke-i

|S| menyatakan jumlah kasus dalam S

sedangkan untuk menghasilkan nilai Entropy adalah dengan rumus sebagai berikut:

$$Entropy(S) = \sum_{i=1}^k -P_i \log_2 P_i \quad (2)$$

definis:

S menyatakan himpunan kasus

k menyatakan jumlah partisi S

p_i menyatakan probabilitas yang didapat dari jumlah (ya/tidak) dibagi total kasus 3.

4. Confussion Matrix

Confussion Matrix merupakan sebuah hasil evaluasi dari sebuah klasifikasi data mining yang diwujudkan dalam sebuah tabel [10]. *Confussion matrix* adalah metode yang banyak dipakai untuk menghitung nilai akurasi. Pengukuran kinerja menggunakan *confussion matrix* memiliki 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 *Confussion Matrix* secara umum dapat dicermati pada tabel dibawah ini:

Tabel 1. Tabel *Confussion Matrix* [10]

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

Untuk menghitung akurasi digunakan rumus sebagai berikut:

$$Akurasi = \frac{TP + TN}{TP + FN + FP + TN} \times 100\% \quad (3)$$

Algoritma klasifikasi pastinya berusaha untuk menghasilkan model yang menghasilkan akurasi yang baik. Kinerja model dari algoritma klasifikasi ditentukan pada saat model dihadapkan pada data *testing*, karena rata-rata model yang dipakai dapat memprediksi dengan benar pada semua data yang menjadi data *trainingnya*, [12].

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

$$Sensitivitas = \frac{TP}{TP+FN} \quad (4)$$

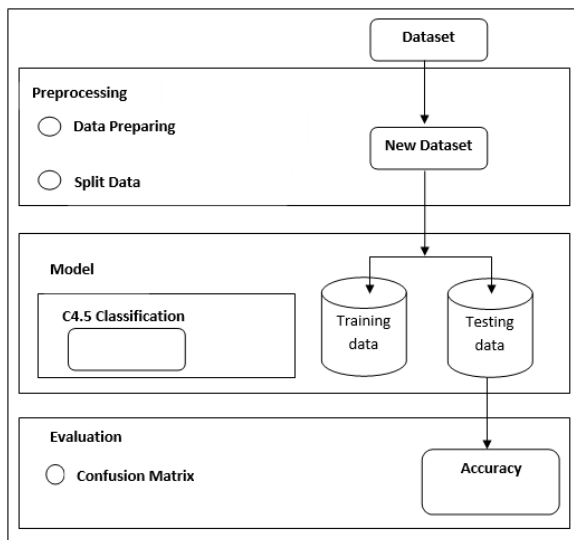
Sedangkan precision adalah rasio ramalan benar positif dipadukan dengan semua hasil yang diprediksi positif. *Precision* menggambarkan matrik untuk menghitung kemampuan sistem dalam menghasilkan data yang penting. *Precision*

pada data mining adalah hasil jumlah data yang true positive dibagi dengan jumlah data yang dikenali sebagai positif. Untuk menghitung precision menggunakan persamaan dibawah ini:

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

III. METODOLOGI PENELITIAN

Pada penelitian ini akan menggunakan metode eksperimen dalam penyelesaian penelitian ini. Secara garis besar penelitian ini akan melakukan perhitungan terhadap dataset dengan menggunakan algoritma klasifikasi C4.5. Adapun kerangka penelitian dari penelitian dapat kita cermati pada gambar dibawah ini :



Gambar 1. Kerangka Penelitian

Tahap pertama adalah menyiapkan dataset, selanjutnya di spilt datanya dibagi menjadi data *testing* dan data *training* dengan presentase 80% untuk data *training* dan 20% untuk data *testing*. Data *training* bertindak sebagai pembentuk pola atau model sedangkan data *testing* bertindak sebagai penguji model. Model yang digunakan adalah klasifikasi dengan algoritmya C4.5. selanjutnya dievaluasi menggunakan confusion matrix dan menghasilkan akurasi.

A. Sumber 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

Pada tahap ini, 520 data diabetes dibagi menjadi dua, yaitu data *training* dan data *testing* dengan prosentase 80% untuk data *training* dan 20% untuk data *testing*. Data *training* bertindak sebagai pembentuk pola atau model dan data *testing* sebagai penguji model.

C. Metode yang Diusulkan

Penelitian ini menggunakan metode klasifikasi data mining dengan algoritma C4.5. Tahap untuk membuat pohon keputusan dengan algoritma C4.5 secara garis besar yaitu:

1. Tahap awal membentuk pohon keputusan adalah dengan membentuk akar (berada paling atas). Selanjutnya pembagian data sesuai dengan atribut-atribut yang serasi untuk dibentuk daun.
2. Tahap selanjutnya adalah pemangkasan pohon yang sudah terbentuk atau disebut juga tree pruning, pengenalan dan memangkas cabang yang tidak digunakan pada pohon yang sudah terbentuk. Selain untuk mengurangi ukuran pohon, pemangkasan pohon juga dilakukan dengan tujuan untuk memangkas tingkat eror pada prediksi kasus baru dari hasil pemecahan yang dilakukan dengan divide and conquer. Dua pendekatan dalam Pruning yaitu Pre-pruning dan Post-pruning.

3. Tahap berikutnya adalah membentuk aturan keputusan dari pohon yang telah terbentuk. Aturan diturunkan dari pohon keputusan dengan cara menelusuri mulai dari hingga ke daun.

IV. HASIL DAN PEMBAHASAN

Variabel data penelitian yang digunakan pada penelitian ini dapat dilihat pada tabel 2 yakni sebagai berikut.

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	<i>Polyuria</i>	1.Ya, 2.Tidak
4	<i>Polydipsia</i>	1.Ya, 2.Tidak
5	<i>Suddenweight loss</i>	1.Ya, 2.Tidak
6	<i>Weakness</i>	1.Ya, 2.Tidak
7	<i>Polyphagia</i>	1.Ya, 2.Tidak
8	<i>Genital thrush</i>	1.Ya, 2.Tidak
9	<i>Visual blurring</i>	1.Ya, 2.Tidak
10	<i>Itching</i>	1.Ya, 2.Tidak
11	<i>Irritability</i>	1.Ya, 2.Tidak
12	<i>Delayed healing</i>	1.Ya, 2.Tidak
13	<i>Partial paresis</i>	1.Ya, 2.Tidak
14	<i>Muscle stiffness</i>	1.Ya, 2.Tidak
15	<i>Alopecia</i>	1.Ya, 2.Tidak
16	Obesitas	1.Ya, 2.Tidak
17	Kelas	1.Positif 2.Negatif

Dari Tabel 2 Ada 16 variabel dataset gejala dan 1 variabel class penentu klasifikasi.

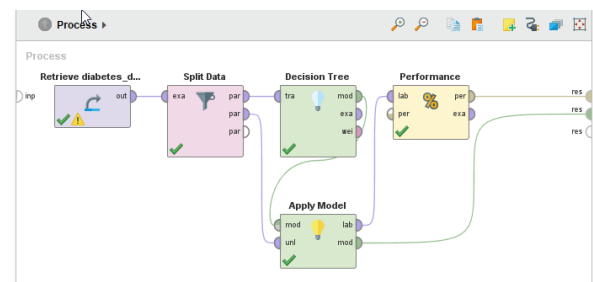
Selanjutnya dari jumlah data 520, data dibagi menjadi dua dengan presentase 80% data *training* dan 20% data *testing*. 80% dari jumlah data 520 didapatkan data *training* dengan jumlah 416 dan 20% dari jumlah data 520 didapatkan data *testing* dengan jumlah 104. Dari data yang peroleh selanjutnya dibentuk pohon keputusan menggunakan algoritma C4.5 guna menghasikan alur keputusan. Tahap awal adalah memilih

atribut sebagai akar dengan melakukan perhitungan gain information.

1. Tahap selanjutnya setelah mencari entropy total adalah mencari nilai entropy untuk masing-masing atribut dengan rumus yang sama. Menggunakan rumus persamaan (2).
2. Tahap selanjutnya adalah menghitung nilai gain dari masing- masing atribut dengan rumus persamaan (1) seperti landasan teori di atas.

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

Atribut dengan nilai information gain paling tinggi dipakai sebagai akar untuk membangun pohon keputusan. Penerapan dan pengujian dataset dilakukan pada rapidminer. Berikut ini merupakan proses klasifikasi penderita biabetes menggunakan *RapidMiner 9.7*:



Gambar 1. Proses Klasifikasi Menggunakan Aplikasi *Rapidminer 9.7*.

Berikut merupakan analisis hasil pengujian dari 104 data *testing* dengan 416 data *training* dengan setiap data memiliki 16 variabel dataset gejala dan 1 variabel kelas.

Accuracy: 97.12%			
	True Positive	True Negative	Class Precision
Pred. Positive	61	0	100.00 %
Pred. Negative	3	40	93.02 %
Class recall	95.31%	100.00%	

Gambar 3. Hasil Akurasi

Adapun untuk *Performance Vektornya* adalah sebagai berikut:

PerformanceVector

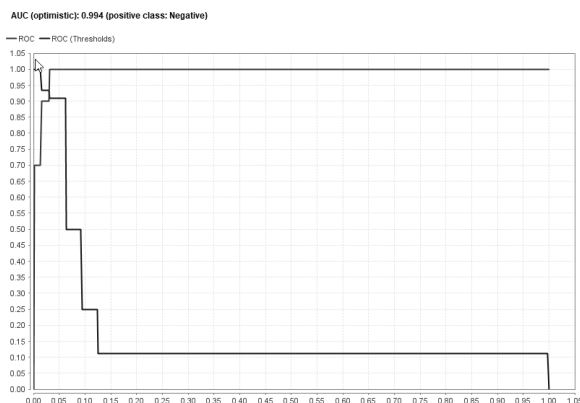
```

PerformanceVector:
accuracy: 97.12%
ConfusionMatrix:
True:  Positive      Negative
Positive:    61       0
Negative:    3       40
precision: 93.02% (positive class: Negative)
ConfusionMatrix:
True:  Positive      Negative
Positive:    61       0
Negative:    3       40
recall: 100.00% (positive class: Negative)
ConfusionMatrix:
True:  Positive      Negative
Positive:    61       0
Negative:    3       40
AUC (optimistic): 0.994 (positive class: Negative)
AUC: 0.982 (positive class: Negative)
AUC (pessimistic): 0.981 (positive class: Negative)
    
```

Gambar 4. Hasil *Performance Vector*

Pengujian menghasilkan akurasi yang cukup besar yaitu 97,12 % Precision sebesar 93,02% %, dan Recall sebesar 100,00%

Pengujian ini juga menghasilkan ROC (*Reciver Operating Characteristic*) seperti gambar 5 dibawah ini



Gambar 5. Hasil Kurva *ROC*

Kurva ROC (*Reciver Operating Characteristic*) diatas menunjukkan algoritma C4.5 memiliki nilai AUC sebesar 0.994 yang artinya Excellent Classification (Sangat Bagus) Dalam hasil penelitian, menunjukkan bahwa

penggunaan Algoritma C4.5 dalam klasifikasi penyakit diabetes memberikan akurasi yang tinggi.

V. KESIMPULAN

Dari 16 atribut yang terdapat dalam dataset diabetes yaitu umur, *Alopecia*, *Gender*, *Polyuria*, *Polydipsia*, *sudden weight loss*, *weakness*, *Polyphagia*, *Genital thrush*, *Irritability*, *delayed healing*, *partial paresis*, *Itching*, *visual blurring*, *muscle stiffness*, dan Obesitas dapat dijadikan sebagai data untuk klasifikasi penderita penyakit diabetes.

Penelitian ini menggunakan Algoritma C4.5 untuk pengklasifikasian seseorang terkena penyakit diabetes atau tidak. Dari 520 data dibagi menjadi 416 sebagai data *training* dan 104 sebagai data *testing*. Dari hasil Pengujian menghasilkan akurasi yang cukup besar yaitu 97,12 % Precision sebesar 93,02% %, dan Recall sebesar 100,00%

Adapun Kurva ROC (*Reciver Operating Characteristic*) menunjukkan algoritma C4.5 memiliki nilai AUC sebesar 0.994 yang artinya *Excellent Classification*, ini menunjukkan bahwa menggunakan Algoritma C4.5 untuk klasifikasi penderita penyakit diabetes menghasilkan akurasi yang tinggi.

DAFTAR PUSTAKA

- [1] Khairani, "Pengetahuan Diabetes Mellitus Dan Upaya Pencegahan Pada Lansia Di Lam Bheu Aceh Besar," *Pengetah. Diabetes Mellit. Dan Upaya Pencegah. Pada Lansia Di Lam Bheu Aceh Besar*, vol. 3, no. 3, pp. 58–66, 2012.
- [2] Nurlina, "Jurnal Media Keperawatan : Politeknik Kesehatan Makassar Jurnal Media Keperawatan : Politeknik Kesehatan Makassar," *J. Media Keperawatan Politek. Kesehat. Makassar*, vol. 10, no. 01, pp. 59–66, 2019.
- [3] ADA, "Classification and Diagnosis of Diabetes Mellitus," *Stand. Med. Care Diabetes*, vol. 39, no. January, 2016, doi: 10.1016/B978-0-323-18907-1.00038-X.

- [4] Suwarno and AA Abdillah, "Penerapan Algoritma Bayesian Regularization Backpropagation Untuk Memprediksi Penyakit Diabetes," *J. MIPA*, vol. 39, no. 2, pp. 150–158, 2017.
- [5] H. Rahmawan and A. SN2, "Penentuan Rekomendasi Petrainingan Pengembangan Diri Bagi Pegawai Negeri Sipil Menggunakan Algoritma C4 . 5 Dengan Principal Component Analysis," *J. TEKNO KOMPAK*, vol. 14, no. 1, pp. 5–10, 2020.
- [6] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition*, vol. 9780470908. 2014.
- [7] P. P. Putra and A. S. Chan, "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 Inform.*, vol. 3, no. 1, p. 24, 2018, doi: 10.35314/isi.v3i1.296.
- [8] F. Kurniawan and Ivandari, "Komparasi Algoritma Data Mining Untuk Klasifikasi Penyakit Kanker Payudara," *J. Stmik*, vol. XII, no. 1, pp. 1–8, 2017.
- [9] N. S. B. Kusrorong, D. R. Sina, N. D. Rumlaklak, J. I. Komputer, and U. N. Cendana, "Kajian Machine Learning Dengan Komparasi Klasifikasi Prediksi Dataset Tenaga Kerja Non-Aktif," vol. 7, no. 1, pp. 37–49, 2019.
- [10] F. Gorunescu, *Data Mining: Concepts, Models, and Techniques*. Springer, 2011.
- [11] R. Dita Merawati, "Penerapan Data Mining Penentu Minat Dan Bakat Siswa Smk Dengan Metode C4 . 5," *J. Algor*, vol. 1, no. 1, pp. 28–37, 2019.
- [12] D. Istiawan and L. Khikmah, "Implementation of C4.5 Algorithm for Critical Land Prediction in Agricultural Cultivation Areas in Pemali Jratun Watershed," *Indones. J. Artif. Intell. Data Min.*, vol. 2, no. 2, p. 67, 2019, doi: 10.24014/ijaidm.v2i2.7569.