

Penerapan Algoritma C4.5 Dalam Prediksi Resiko Diabetes Tahap Awal (Early Stage Diabetes)

Wina Yusnaeni¹, Widiarina²

^{1,2}Universitas Bina Sarana Informatika

¹e-mail: wina.wyi@bsi.ac.id

²e-mail: widiarina.wda@bsi.ac.id

Diterima	Direvisi	Disetujui
23-10-2021	30-12-2021	12-01-2022

Abstrak - Penyakit diabetes penyakit yang sering di jumpai di semua kalangan masyarakat. Dari data yang di dapat bahkan setiap tahunnya terjadi peningkatan penderita diabetes di dunia. Banyaknya gejala tahap awal yang bisa dijadikan acuan seseorang di prediksi terkena gejala diabetes atau tidak. Oleh karena itu pentingnya pencegahan sehingga peningkatan akan kesehatan dan bisa terhindar dari penyakit diabetes ini. Berdasarkan alasan tersebut disini peneliti melakukan penelitian dengan menggunakan metode decision tree (C4.5) dengan bantuan aplikasi rapid miner. Data yang didapat merupakan data klasifikasi tahap awal penderita diabetes. Dengan melakukan uji training dan uji coba dengan perbandingan 90:10 menggunakan split data pada aplikasi rapid miner dan mengolah dengan metode decision tree selain itu menambahkan performance untuk menghitung akurasi. Hasilnya di dapat berupa pohon keputusan yang bisa di jadikan role dalam uji testing dataset. Selain itu hasil evaluasi data di dapat tingkat akurasi sebesar 88.46 % dimana jumlah presentase ini dikatakan dalam kategori good classification.

Kata Kunci: diabetes , decision tree, akurasi

Abstract - Diabetes is a disease that is often encountered in all circles of society. From the data in the can even occur annually an increase in people with diabetes in the world. Many of the symptoms of the early stages that can be used as a reference to a person in the prediction of symptoms of diabetes or not. Hence the importance of prevention so that the increase going to health and could be spared from the disease diabetes. Based on those reasons here. researchers conducted the study by using the method of decision tree (C4.5) with the aid of rapid miner. The Data obtained is the data classification of the early stages of diabetic patients. To test the training and trials with a ratio of 90:10 using the split data on the application of rapid miner and process the decision tree method in addition add performance to calculate the accuracy. The results can be in the form of a decision tree that can be made a role in the testing dataset. In addition, the results of the evaluation data in the can the level of accuracy of 88.46 % where the amount of the percentage is said to be in the category of good classification.

Keywords: Diabetes, Decision tree. accuracy

PENDAHULUAN

Salah satu penyakit yang bisa menyebabkan komplikasi penyakit berat lainnya seperti darah tinggi, jantung dll adalah Penyakit Diabetes. Hal ini dikarenakan adanya gangguan yang terjadi pada metabolisme dimana tidak terproduksinya insulin oleh pancreas. Penyakit diabetes ada di seluruh dunia tetapi peningkatan jumlah penderitanya berbeda di setiap negara (Fiarni et al., 2019). Di Indonesia prevalence akan diabetes mencapai 10,9 % , dan mengalami peningkatan (Fiarni et al., 2019). Pola hidup yang tidak sehat yang dianggap sebagai

penyumbang besar yang menyebabkan terkenanya penyakit ini , bahkan penyakit ini bukan hanya menyerang kalangan orang miskin saja tetapi juga semua kalangan masyarakat (Aris, 2019). Oleh karena itu perlunya pengetahuan akan tanda-tanda awal terjangkitnya penyakit diabetes dengan pengetahuan akan beberapa dari gejala-gejala awal yang bisa dijadikan sebagai acuan untuk bisa melakukan pencegahan sebelum terlambat. Umumnya gejala awal terjadi bagi penderita tahap awal diabetes diantaranya sering merasa haus, sering merasa lapar, gatal-gatal dan juga lemas (dr. Apriliana Adyaksari, Sp. PD, 2020).



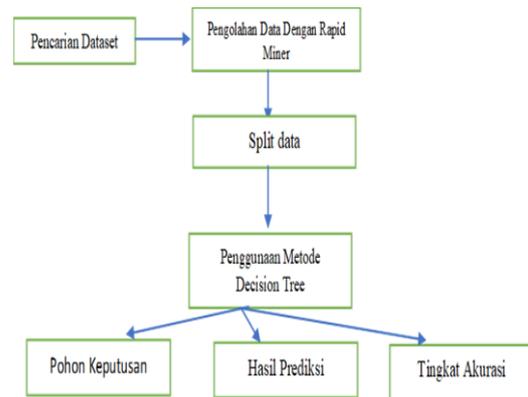
Berdasarkan tingginya tingkat penderita penyakit ini maka pencegahan dan prediksi menjadi factor penting dalam penelitian saat ini (Fiarni et al., 2019). Metode -metode yang digunakan dalam melakukan prediksi yaitu metode data mining. Data mining serangkaian Tindakan untuk menemukan hubungan dari pola dan kecendrungan dari data yang disimpan (Putri et al., 2021) . Data mining di dasarkan pada kelebihan data atau informasi pada suatu perusahaan atau institusi yang di simpan selama bertahun-tahun(Darmawan, 2018). Beberapa metode yang di lakukan untuk melakukan prediksi seperti Algoritma C45, SVM ,Naïve Bayes, Random Forest dll. Dari beberapa metode data mining, algoritma C45 merupakan algoritma yang relative mudah di pahami dimana keputusan di gambarkan dalam bentuk pohon keputusan (Fauzi et al., 2020). Beberapa literatur yang relevan berdasarkan metode algoritma C45 atau studi kasus penyakit diabetes diantaranya: (Aris, 2019) menggunakan metode C.45 Identifikasi Penyakit Diabetes Melitus menghasilkan suatu pohon keputusan yang mencantumkan atribut-atribut seperti jenis kelamin, gula darah , usia, proko dan tipe gula darah lainnya. Penelitian (Noviandi, 2018) penggunaan metode C45 dalam prediksi penyakit diabetes dengan tingkat akurasi 70.32%. Penelitian (Fauzi et al., 2020) penggunaan metode C45 dan K-Modes dalam prediksi pemilihan jurusan dengan tingkat akurasi 87,50%. Penelitian (Fiarni et al., 2019) Penggunaan data mining Algorithm dalam analisis dan prediksi penyakit diabetes komplikasi menghasilkan akurasi rata-rata 68% yang dianggap bisa membantu dalam memprediksi penyakit diabetes tahap awal.

Dari penjabaran diatas peneliti disini melakukan penelitian dengan menggunakan algoritma C45 dalam prediksi tahap awal penyakit diabetes dimana hasil yang dihasilkan adalah pohon keputusan sehingga bisa dilakukan uji coba kesesuaian hasil dengan class berdasarkan data uji. Selain itu penggunaan rapid miner dalam melakukan perhitungan hasil serta tingkat akurasi yang akan menentukan kelayakan prediksi yang dihasilkan.

METODOLOGI PENELITIAN

1. Tahapan Penelitian

Tahapan Penelitian yang dilakukan di gambarkan pada gambar di bawah ini:



Sumber:(Yusnaeni & Widiarina, 2021)

Gambar 1. Tahapan Penelitian dengan menggunakan Rapid Miner

Pada gambar diatas merupakan tahapan penelitian yang dilakukan dengan menggunakan Rapid Miner dimana pertama yang dilakukan adalah pengumpulan data. Data yang digunakan adalah data sekunder yang di dapat dari web <https://archive.ics.uci.edu/ml/datasets.php> . Dari data yang di dapat dilakukan perhitungan dengan menggunakan rapid miner dengan membagi data menjadi 90 persen data Training dan 10 persen data testing , selanjutnya metode yang digunakan dalam perhitungan adalah Algoritma C.45 . yang nantinya akan menghasil hasil prediksi , pohon keputusan dan tingkat akurasi dari data yang diolah.

2. Algoritma C.45

Salah satu data metode data mining Sering digunakan dalam prediksi adalah Algoritma C4.5 . Pohon keputusan dapat digunakan untuk mewakili dan membuat keputusan, dapat dilihat sebagai serangkaian titik yang terdiri dari node tunggal atau menyebar ke bagian daun(Gifu, 2021). Algoritma C4.5 sudah dikenal banyak prang dalam hal klasifikasi berupa angka atau kategori (Elisa, 2017). Tahapan dalam perhitungan dengan menggunakan metode algoritma C4.5 yaitu (Much Aziz Muslim, Budi Prasetyo, Eva Lalily Hamum M, Anisa Juli H, Mirqotussa'adah, Siti Hardiyanti R, 2019): pemilihan atribut yang di gunakan sebagai akar, selanjutnya pembuatan cabang -cabang dari setiap akar (node) , membagi kasus dalam setiap node dan proses diulangi sampai cabang memiliki kelas yang sama. Perhitungan dalam algoritma C4.5 dilakukan dengan menghitung nilai Entropy, dan gain ratio untuk penentuan akar, node dan daun.

HASIL DAN PEMBAHASAN

1. Dataset (Pengumpulan Data)

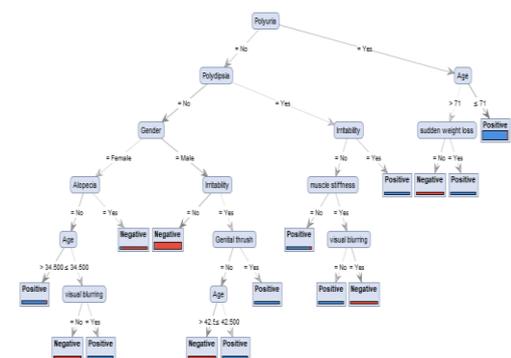
Dataset yang digunakan adalah dataset yang di ambil dari dataset uci dengan dataset di tahun 2020 . data yang digunakan terdiri dari 520 data dengan hasil clas berupa positive dan negatif . Atribut -atribut yang digunakan dalam dataset :

Tabel 1. Atribut Yang digunakan dalam Dataset

No	Nama Atribut	Atribut Information
1	Usia (Age)	16-90 Tahun
2	JK(Gender)	Male/Female
3	Sering Buang Air Kecil (Polyuria)	Yes / No
4	Sering Haus (Polydipsia)	Yes / No
5	Cepat kehilangan Berat Badan (sudden weight loss)	Yes / No
6	Kelelahan (weakness)	Yes / No
7	Banyak Makan (Polyphagia)	Yes / No
8	Infeksi Jamur (Genital thrush)	Yes / No
9	Penglihatan Kabur(visual blurring)	Yes / No
10	Gatal-Gatal (Itching)	Yes / No
11	Mudah marah (Irritability)	Yes / No
12	Lama Penyembuhan Luka (delayed healing)	Yes / No
13	Kelumpuhan(partial paresis)	Yes / No
14	Kekakuan Otot (muscle stiffness)	Yes / No
15	Kerontokan Rambut (Alopecia)	Yes / No
16	Kegemukan (Obesity)	Yes / No
17	Class	Positif/Negative

Sumber : (Yusnaeni & Widiarina, 2021)

- Pengolahan Dengan Menggunakan Rapid Miner Dataset yang ada di simpan dalam bentuk excel atau .xls. selanjutnya dataset akan diolah dengan rapid miner .
- Pembagian Data uji dan Data testing Dalam pengujiannya Data dibagi menjadi 90 persen Data Training dan 10 Persen data Testing.
- Penggunaan Algoritma C4.5 Didalam Aplikasi Rapid Miner setelah melakukan read dataset, Split data selanjutnya penggunaan metode yang ada disini adalah Decision Tree.
- Pohon Keputusan Hasil dari Perhitungan bisa dalam bentuk Pohon Keputusan seperti tergambar di bawah ini:



Sumber : (Yusnaeni & Widiarina, 2021)

Gambar 2. Pohon Keputusan hasil Perhitungan Dataset

Berdasarkan gambar Pohon keputusan diatas akar atau root di mulai dari Sering Buang Air Kecil (Polyuria). Contoh aturan atau rule yang terbentuk yaitu:

- Jika Polyuria =Yes, Age >71, sudden weight loss=yes maka Hasil Class adalah Positif
- Jika Polyuria =Yes, Age <=71, Maka Hasil Class=Positif.

Untuk melihat apakah dari hasil pohon keputusan tersebut sesuai maka bisa dilakukan data testing sebagai berikut:

Tabel 2. Table Testing

Atribut	Atribut Information	
Age	42	55
Gender	Male	Female
Polyuria	No	Yes
Polydipsia	No	Yes
sudden weight loss	No	Yes
weakness	No	Yes
Polyphagia	No	No
Genital thrush	No	No
visual blurring	No	Yes
Itching	No	No
Irritability	No	No
delayed healing	No	No
partial paresis	No	Yes
muscle stiffness	No	No
Alopecia	Yes	No
Obesity	No	Yes
Class	Negative	Positive

Berdasarkan data diatas untuk dilakukan testing dengan pohon keputusan yaitu :

- Jika Polyuria=No, gender =Male, Irritability=No Maka Class= negative
- Jika Polyuria=Yes, Age <=71 Maka Class= Negatif

Berdasarkan 2 Data testing yang diuji coba diatas di dapat bahwa pohon keputusan memberikan jawaban yang sesuai dengan pohon keputusan. Dimana testing pertama menghasilkan class negative dan hasil kedua menghasilkan positif.

6. Hasil prediksi

Berdasarkan data Testing di dapat hasil prediksi

dengan dominan sama antara class dengan hasil Prediksi:

Row No.	class	prediction	confidence	confidence	Age	Gender	Polyuria	Polydipsia	weight loss	weakness	Fatigable
1	Positive	Positive	0.973	0.027	58	Male	Yes	Yes	No	Yes	Yes
2	Positive	Positive	0.973	0.027	43	Male	Yes	Yes	Yes	Yes	No
3	Positive	Positive	0.900	0.100	38	Male	No	Yes	No	No	No
4	Positive	Positive	0.958	0.042	65	Female	No	No	No	No	No
5	Positive	Positive	0.973	0.027	38	Female	Yes	Yes	Yes	Yes	Yes
6	Positive	Positive	1	0	28	Female	No	No	No	No	No
7	Positive	Negative	0	1	90	Female	No	Yes	Yes	No	No
8	Positive	Positive	0.973	0.027	25	Male	Yes	Yes	No	No	Yes
9	Positive	Positive	0.973	0.027	45	Female	Yes	Yes	Yes	Yes	Yes
10	Positive	Positive	0.973	0.027	55	Male	Yes	Yes	Yes	Yes	Yes
11	Positive	Positive	1	0	54	Male	No	No	No	No	No
12	Positive	Positive	0.973	0.027	61	Male	Yes	No	No	Yes	Yes
13	Positive	Positive	0.973	0.027	56	Male	Yes	No	Yes	Yes	No
14	Positive	Positive	0.973	0.027	60	Male	Yes	No	No	No	No
15	Positive	Positive	0.973	0.027	39	Female	Yes	Yes	No	No	Yes
16	Positive	Negative	0	1	90	Female	No	Yes	Yes	No	No
17	Negative	Positive	0.973	0.027	67	Male	Yes	No	No	Yes	Yes
18	Negative	Positive	1	0	40	Male	No	No	No	Yes	No
19	Negative	Negative	0.943	0.057	37	Male	No	No	No	No	No
20	Negative	Negative	0.943	0.057	35	Male	No	No	Yes	Yes	No

Sumber:(Yusnaeni & Widiarina, 2021)

Gambar 3. Hasil Prediksi dengan rapid Miner

7. Evaluasi hasil dengan Tingkat Akurasi

Dari Hasil Model yang ada akan di uji tingkat akurasi setelah dimasukkan data testing dari split data yang dilakukan di aplikasi rapid minor. Dan hasil nya bisa di lihat pada table di bawah ini:

Tabel 3. Confusion Matrik Decision tree Algoritma C4.5

	true Positive	true Negative	class precision
pred. Positive	27	4	87.10%
pred. Negative	2	19	90.48%
class recall	93.10%	82.61%	

Di Tabel 3 menunjukkan nilai akurasi sebesar 88.46%. dimana Jumlah Data benar Pada Prediksi Positive Sebanyak 27 Data dan 4 data tidak sesuai masuk ke prediksi negative . sedangkan prediksi negative yang sesuai ada 2 data sedangkan tidak sesuai dengan prediksi negative berjumlah 19. Dengan akurasi 88.46 % . Menurut (Gorunescu, 2011) Tingkat Akurasi dengan Range Accuracy bernilai 0.80 – 0.90 = good classification.

KESIMPULAN

Dari penjelasan diatas dapat disimpulkan penggunaan metode klasifikasi algoritma C4.5 dikatakan dapat menjadi alternatif dalam membantu memprediksi tahap awal penyakit diabetes. Model pohon keputusan yang di dapat dan telah dilakukan uji data testing menghasilkan tingkat akurasi cukup tinggi yaitu 88.46% dimana hal ini masuk ke dalam kategori klasifikasi baik (good clasification). Untuk penelitian selanjutnya bisa di tingkatkan kembali prediksi menjadi lebih detail sehingga bisa

meramalkan dengan atribut yang ada maka akan bisa di prediksi kedepannya berapa lama dari kebiasaan yang dilakukan akan terkena penyakit diabetes atau tidak.

REFERENSI

Aris, F. (2019). *Penerapan Data Mining untuk Identifikasi Penyakit Diabetes Melitus dengan Menggunakan Metode Klasifikasi*. 1(1), 1–6. Darmawan, E. (2018). C4.5 Algorithm Application for Prediction of Self Candidate New Students in Higher Education. *Jurnal Online Informatika*, 3(1), 22. <https://doi.org/10.15575/join.v3i1.171>

dr. Apriliana Adyaksari, Sp. PD, M. K. (2020). Waspada! Inilah Tanda-Tanda Gejala Diabetes. <https://www.Emc.Id/Id/Care-plus/Waspada-Inilah-Tanda-Tanda-Gejala-Diabetes,1>.

Elisa, E. (2017). Analisa dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT.Arupadhatu Adisesanti. *Jurnal Online Informatika*, 2(1), 36. <https://doi.org/10.15575/join.v2i1.71>

Fauzi, A., Saraswati, N. M., Cipta, R., & Hariyono, S. (2020). *PENERAPAN ALGORITMA K-MODES DAN C4.5 UNTUK PREDIKSI PEMILIHAN JURUSAN DI UNIVERSITAS PERADABAN PADA SISWA SMA (Studi Kasus: SMA Islam Ta'allumul Huda Bumiayu)*. 1(2), 57–64.

Fiarni, C., Sipayung, E. M., & Maemunah, S. (2019). Analysis and prediction of diabetes complication disease using data mining algorithm. *Procedia Computer Science*, 161, 449–457. <https://doi.org/10.1016/j.procs.2019.11.144>

Gifu, D. (2021). The Use of Decision Trees for Analysis of the Epilepsy. *Procedia Computer Science*, 192, 2844–2853. <https://doi.org/10.1016/j.procs.2021.09.055>

Gorunescu, F. (2011). *Data Mining Concepts, Models and Techniques*. Springer Berlin Heidelberg. <https://doi.org/DOI.10.1007>

Much Aziz Muslim, Budi Prasetyo, Eva Lalily Hamum M, Anisa Juli H, Mirqotussa'adah, Siti Hardiyanti R, A. N. Z. (2019). Data Mining Algoritma C4.5. In N. C. Eka Listiana (Ed.), *Data Mining Algoritma C4.5*. CV. Harian Jateng Network.

Noviandi. (2018). Implementasi Algoritma Decision Tree C4.5 Untuk Prediksi Penyakit Diabetes. *Jurnal INOHIM*, 6(1), 1–5. <https://inohim.esaunggul.ac.id/index.php/INO/>

- article/view/142
- Putri, sanni ucha, Irawan, E., & Rizky, F. (2021). Implementasi Data Mining Untuk Prediksi Penyakit Diabetes. *KESATRIA(Jurnal Penerapan Sistem Informasi Dan Manajemen,* 2(1), 39–46.
- Yusnaeni, W., & Widiarina. (2021). *laporan Akhir: Penerapan Algoritma C4.5 Dalam Prediksi Resiko Diabetes Tahap Awal (Early Stage Diabetes).*