E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1265 - 1270

KLASIFIKASI PENYAKIT DIABETES MELLITUS BERDASARKAN FAKTOR-FAKTOR PENYEBAB DIABETES MENGGUNAKAN ALGORITMA C4.5

Ronna Putri Fadhillah[1)], Raisya Rahma[2)], Arni Sepharni[3)], Ratna Mufidah[4)], Betha Nurina Sari[5)], Agung Pangestu[6)]

[1, 2,3,4,5)]Teknik Informatika, Universitas Singaperbangsa Karawang
Jl. HS.Ronggo Waluyo, Puseurjaya, Telukjame Timur, Karawang Kode Pos 41361
[6)]Teknik Elektro, Universitas Global Jakarta
Jl. Boulevard Grand Depok City, Tirtajaya, Sukmajaya, Kota Depok Kode Pos 16412
e-mail: ronna.fadhillah18111@student.unsika.ac.id[1)], raisya.rahma18190@student.unsika.ac.id[2)], arni.sepharni18105@student.unsika.ac.id[3)], ratna.mufidah@cs.unsika.ac.id[4)], betha.nurina@staff.unsika.ac.id[5)], agungp@jgu.ac.id[6)]

ABSTRAK

Diabetes merupakan penyakit yang disebabkan oleh tingginya gula darah pada seseorang. Terdapat banyak faktor yang menjadi penyebab terjadinya diabetes, faktor-faktor tersebut diantaranya seperti faktor keturunan, gula darah yang tinggi, berat badan, usia, dan faktor lainnya. Angka kematian yang disebabkan oleh penyakit diabetes ini semakin banyak dan setiap tahunnya diperkirakan akan terus meningkat angka kasus kematiannya. Diagnonisis dini dan menerapkan pola hidup sehat merupakan dua langkah awal dalam mencegah terjadinya penyakit diabetes mellitus. Dataset yang digunakan pada penelitian ini merupakan data yang didapatkan dari data open source Kaggle, yaitu data Pima Indians Diabetes. Metode klasifikasi yang digunakan yaitu degan menerapkan algoritma C4.5 yang mampu menghasilkan tingkat akurasi yang baik. dilakukan seleksi fitur terhadap dataset dengan menggunakan heatmap yang menghasilkan fitur Pregnancies, Glucose, BMI, Age, dan outcome. Hasil dari penelitian ini didapati nilai akurasi sebesar 76%. Hasil ini lebih baik dibanding dengan penelitian sebelumnya yang menggunakan algoritma SVM yang hanya mampu menghasilkan tingkat akurasi sebesar 70%.

***Kata Kunci**: Algoritma Decision Tree, Data Mining, Fitur Heatmap, Sistem Pakar.*

ABSTRACT

Diabetes is a disease caused by high blood sugar in a person. There are many factors that cause diabetes, these factors include heredity, high blood sugar, weight, age, and other factors. The number of deaths caused by diabetes is increasing and every year the number of cases of death will continue to increase. Early diagnosis and implementing a healthy lifestyle are the first two steps in preventing diabetes in a person. Therefore, this study classified the factors that cause diabetes mellitus. The dataset used in this study is data obtained from open source Kaggle, namely Pima Indians Diabetes data. The classification method used is by applying the C4.5 algorithm which is able to produce a good level of accuracy. Feature selection is carried out on the dataset using a heatmap which produces Pregnancies, Glucose, BMI, Age, and Outcome features. The results of this study found an accuracy value of 76%. This result is better than previous research using the SVM algorithm which is only able to produce an accuracy rate of 70%.

***Keywords**: Decision Tree Algorithm, Data Mining, Heatmap Feature, Expert System.*

I. PENDAHULUAN

DIABETES Melitus (DM) adalah salah satu sindrom yang ditandai dengan kelainan pada metabolik yang terganggu serta pada kenaikan konsentrasi gula darah yang abnormal yang disebabkan oleh defisiensi insulin, ataupun sensitivitas insulin yang rendah dari jaringan, maupun keduanya [1]. Gula darah begitu penting untuk kesehatan, sebab kesehatan adalah sumber energi yang berperan penting untuk sel-sel serta jaringan [2]. Pada umumnya di usia muda kandungan glukosa darah akan bertambah secara ringan, namun usia 50 tahun atau lebih kandungan glukosa ini berkemungkinan akan alami kenaikan secara progresif. Orang dengan gaya hidup pasif, jarang atau bahkan tidak pernah melakukan kegiatan akan lebih terasa pada saat mengalami peningkatan glukosa [3].

Peneliti melakukan pembagian terhadap penyakit diabetes menjadi 3, yaitu diabetes tipe 1, diabetes tipe 2, dan diabetes *gestasional*. Diabetes *gestasional* merupakan salah satu jenis diabetes yang terjadi diakibatkan oleh perubahan hormonal pada saat kehamilan [4] dan pada umumnya dibetes yang tidak terkontrol pada masa kehamilan dapat meningkatkan resiko kematian janin. Diabetes dapat menimbulkan komplikasi di beberapa anggota tubuh selain itu juga bisa meningkatkan resiko kematian dini secara keseluruhan. Kemungkinan

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1265 - 1270

komplikasi yang akan terjadi yaitu diantaranya amputasi kaki, kerusakan saraf, kehilangan peglihatan, serta gagal ginjal.

Menurut *International Diabetes Federation* yang mengidap penyakit diabetes pada tahun 2015 sebanyak 415 juta jiwa, dan diperkirakan meningkat sebanyak 227 juta jiwa atau menjadi 642 juta jiwa pada tahun 2040 yang mengidap penyakit tersebut. Di setiap negara jumlah Diabetes Melitus mengalami pengingkatan dan kasus terbanyak orang yang mengalami Diabetes Melitus berada di usia antara 40-59 tahun [5]. Dilihat dari angka kematian yang tinggi yang diakibatkan oleh diabetes, diagnosis dini begitu penting dilakukan untuk menekan angka kematian.

Selain itu diagnosis dini juga merupakan titik awal penderita untuk mencegah terjadinya diabetes lebih parah dengan melakukan pola hidup sehat agar tidak mengalami komplikasi [6]. Sedangkan seseorang yang sudah lama mengalami diabetes yang tidak terdiagnosis dan tidak diobati, berkemungkinan besar kesehatan tubuhnya akan semakin buruk. Dilakukan penelitian pada awal 2022 yaitu dengan dilakukan edukasi atau pendidikan kesehatan dan juga untuk mendeskripsikan *self care* pada anak penderita diabetes mellitus tipe 1 dengan pendekatan Teori Dheotera Orem yang manghasilkan bahwa *self care* penderita diabetes mellitus tipe 1 mengalami perbaikan yang signifikan setelah dilakukan tindakan edukasi selama 3 siklus [7].

Telah dilakukan penelitian sebelumnya yang menerapkan algoritma C4.5 untuk prediksi penyakit diabetes dilihat dari faktor-faktor penyebab diabetes seperti jumlah wanita melahirkan, kadar gula darah, tekanan darah, insulin, *Body Massa Indeks* (BMI), usia, dan hasil kelas menghasilkan akurasi sebesar 70.32% [8]. Penelitian lain dilakukan klasifikasi terhadap *dataset* penderita penyakit diabetes dengan menggunakan metode KNN yang berjudul "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes" yang menghasilkan tingkat akurasi tertinggi 39% pada K=3, presisi tertinggi 65% pada K=3 dan K=5, *recall* tertinggi 36% pada K=3, dan *F-Measure* tertinggi pada K=3 [9].

Penelitian ini memiliki kelebihan dibanding dengan penelitian sebelumnya, yaitu data yang digunakan merupakan *dataset* sampel wanita dengan spesifikasi khusus yaitu usia minumal 21 tahun. Penggunaan algoritma C4.5 dalam penelitian ini diharapkan dapat menghasilkan tingkat akurasi yang lebih tinggi lagi dari penelitian-penelitian sebelumnya dengan menggunakan atribut-atribut dari faktor penyebab penyakit diabetes seperti kehamilan, glukosa, BMI, usia, dan hasil dalam mengklasifikasikan seseorang mengidap penyakit diabetes atau tidak dilihat dari faktor-faktor penyebabnya.

II. METODE PENELITIAN

A. Alur Penelitian

Berikut merupakan alur dari penelitian yag dilakukan oleh peneliti pada penelitian ini dapat dilihat pada Gambar 1.

Gambar 1. Alur Penelitian

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1265 - 1270

B. Pengumpulan Data

Penelitian ini menggunakan *dataset* asli dari *National Insttute of Diabetes and Digestive and Kidney Diseases* yang berisi data *Pima Indians Diabetes* yang didapatkan dari situs *open source Kaggle*(www.kaggle.com) yang terdiri dari 769 baris dan 9 kolom, termasuk data demografi, beberapa faktor, dan hasil. 769 baris tersebut adalah sampel wanita berusia minimal 21 tahun dari suku Indian Pima. *Datasate* terdiri dari 9 atribut, yaitu kehamilan, glukosa, tekanan darah, ketebalan kulit, insulin, BMI, fungsi silsilah diabetes, usia, dan *outcome.* 9 atribut tersebut kemudian dilakukan seleksi fitur guna untuk memudahkan dalam proses klasifikasi.

C. Seleksi Fitur

Seleksi fitur bertujuan guna mengurangi atribut, sebab banyaknya fitur pada *dataset* yang digunakan dapat mengakibatkan *overfitting* dan perlu menentukan fitur yang dibutuhkan dalam pengolahan data. Seleksi fitur menggunakan *heatmap* untuk menyeleksi fitur dari *dataset* dilihat dari nilai hubungan antar fitur-fitur tersebut terhadap variabel *outcome* sebagai *target value* yang akan diprediksi. Fitur yang terpilih merupakan fitur yang akan dipergunakan karena mempunyai hubungan tertinggi terhadap variabel *outcome*. Didapati hasil seleksi fitur yang akan digunakan adalah fitur kehamilan, glukosa, BMI, usia, dan *outcome*.

D. Klasifikasi dengan Algoritma C4.5

Klasifikasi merupakan salah satu teknik dalam pengolahan data yang bekerja dengan cara objek yang dipergunakan dibagi menjadi kelas-kelas dengan jumlah kelas sesuai dengan yang diinginkan. Klasifikasi dapat menciptakan suatu pola yang dapat memisahkan tiap-tiap kelas data yang bertujuan guna menentukan objek yang tergolong ke dalam kategori tertentu dilihat dari perilaku serta atribut dari kelompok yang telah didefinisikan [10]. Klasifikasi yang dilakukan dalam penelitian ini bertujuan untuk menggolongkan data termasuk ke dalam kelas *Ya* atau 1 positif diabetes dan *Tidak* atau 0 negatif diabetes dilihat dari beberapa faktor penyebab penyakit diabetes.

Ros Quinlan menemukan sebuah algoritma bernama Algoritma C4.5, algoritma ini adalah algoritma yang dikembangkan dari Algoritma ID3, tujuan digunakan algoritma C4.5 adalah guna melakukan klasifikasi terhadap data yang memiliki atribut berupa numerik maupun kategorial, kemudian setelah proses klasifikasi menghasilkan beberapa ketentuan yang bisa dipergunakan untuk melakukan prediksi nilai atribut yang memiliki tipe diskrit dari *record* yang baru [11]. Selain mudah dimengerti dan fleksibel, algoritma C4.5 juga memiliki kelebihan yaitu menarik karena mampu divisualisasikan ke dalam bentuk gambar.Berikut merupakan tahapan dari algoritma C4.5:

1. Menyiapkan data latih (biasanya data latih diambil dari rat-rata yang telah terjadi sebelumnya dan sudah dikelompokkan ke dalam beberapa kelas tertentu.
2. Tentukan atribut sebagai akar.
3. Membuat cabang untuk masing-masing nilai.
4. Membagi kasus ke dalam cabang yang sudah dibuat.
5. Ulangi ke tahap awal bagi setiap cabang sampai semua kasus mempunyai kelas yang sama.

Proses ini akan berakhir ketika semua *record* mendapatkan kelas yang sama, dan ketika tidak ada atribut *record* yang dipartisi lagi dan tidak terdapat dalam cabang yang kosong.

E. Evaluasi

Tujuan digunakannya evaluasi dalam penelitian ini yaitu untuk mengetahui hasil akhir dari model yang telah dibuat sebelumnya oleh peneliti. Evaluasi dalam penelitian ini hanya melihat pada tingkat akurasi yang dihasilkan. Akurasi adalah nilai akhir yang dihasilkan oleh model, yaitu dengan cara merepresentasikan dari keseluruhan jumlah *dataset* yang benar dikenali. Nilai akurasi dapat dihitung dengan cara membagi total *dataset* benar dikenali dengan total *dataset* dan data uji. Berikut merupakan rumus untuk menghitung nilai akurasi (*acc*):

$$acc = \frac{TP + TN}{TP + FP + FN + TN}$$

Keterangan:
TP= Data positif benar dikenali
TN = Data negatif benar dikenali
FP = Data positif salah dikenali
FN = Data negatif salah dikenali

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1265 - 1270

III. HASIL DAN PEMBAHASAN

A. *Dataset*

Penelitian ini menggunakan data berjumlah 768 data dengan 9 atribut. Melalui proses seleksi fitur untuk menentukan fitur yang diperlukan dan mencapai angka akurasi yang optimal maka dari proses tersebut atribut yang akan digunakan hanya 5 atribut siantaranya 4 sebagai prediktor dan 1 sebagai *target value*. *Dataset* yang terdapat pada Tabel 1 merupakan ringkasan dari data yang diperoleh tidak diikuti dengan penjelasan secara rinci yang menyebutkan maksud dari masing-masing atribut. Hal ini bisa dijadikan sebagai acuan langkah awal untuk menganalisis maksud dari data dengan pencarian informasi. Informasi yang diperoleh dapat dilihat pada Tabel I.

TABEL I
DATASET DIABETES

Pregnancies	*Glucose*	*Blood Preasure*	*Skin Thickness*	*Insulin*	BMI	*Diabetes Pedigree Function*	*Age*	*Outcome*
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1
3	126	88	41	235	39.3	0.704	27	0
8	99	84	0	0	35.4	0.388	50	0
7	196	90	0	0	39.8	0.451	41	1
9	119	80	35	0	29	0.263	29	1
11	143	94	35	146	36.6	0.254	51	1

B. *Preprocessing*

Pada tahap ini dilakukan untuk mempersiapkan data agar dapat diolah dengan baik. Tahap ini meliputi *target value* yaitu atribut *outcome* , penentuan prediktor yaitu atribut yang sudah melakukan proses seleksi fitur diantaranya. *Pregnancies, Glucose,* BMI, *Age,* serta tahapan *train test split data*.*Dateset* yang telah melalui tahap *preprocessing* dapat dilihat pada Tabel 2.

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1265 - 1270

TABEL II
HASIL *PREPROCESSING*

Pregnancies	*Glucose*	BMI	*Age*	*Outcome*
6	148	33.6	50	1
1	85	26.6	31	0
8	183	23.3	32	1
1	89	28.1	21	0
0	137	43.1	33	1
5	116	25.6	30	0
3	78	31	26	1
10	115	35.3	29	0
2	197	30.5	53	1
8	125	0	54	1
4	110	37.6	30	0
10	168	38	34	1
10	139	27.1	57	0
1	189	30.1	59	1
5	166	25.8	51	1
7	100	30	32	1
0	118	45.8	31	1
7	107	29.6	31	1
1	103	43.3	33	0
1	115	34.6	32	1
3	126	39.3	27	0
8	99	35.4	50	0
7	196	39.8	41	1
9	119	29	29	1
11	143	36.6	51	1

C. Data Processing

Dilakukan pembagian anatara *data traning* dan *data testing* pada penelitian ini, metode *Grid search cross validation* digunakan untuk membagi data tersebut. *Grid search cross validation* yaitu suatu metode guna melakukan validasi terhadap model yang lebih dari satu dan *hyperparameter* masing-masing dilakukan secara otomatis dan sistematis. Pada proses ini *cross validation* diatur 10 dimana di setiap kombinasi anatara model dan parameter dilakukan validasi sebanyak 10 kali dengan cara data dibagi menjadi 10 bagian sama besar secara acak (9 bagian *training* serta 1 bagian *testing*).

Proses pengklasifikasian terhadap *dataset* yang terlah melalui tahap *preprocessing* dilakukan dengan menerapkan algoritma C4.5. *Dataset* diolah sehingga dapat menghasilkan nilai akurasi dengan menggunakan bahasa pemrograman python yang dibuat oleh peneliti dengan bantuan *tools* Google Colab.

D. Data Mining

Data mining yang dilakukan dalam penelitian ini yaitu dengan menggunakan algoritma C4.5. Hasil dari tahap klasifikasi dengan menggunakan algoritma C4.5 ditunjukkan dengan *Coonfusion Matrix* yang dapat dilihat pada Tabel 3.

TABEL III
HASIL *CONFUSION MATRIX*

Akurasi	***Predicted Negative***	***Predicted Positive***
Actual Negative	37	69
Actual Positive	30	18

E. Evaluasi

Akurasi diukur dari *data training* dan *data testing* pada Algoritma C4.5 dengan *cross validation* 100 bagian menghasilkan tingkat akurasi dengan rata-rata prediksi yaitu 76% (Tabel II), ini artinya didapati model yang cukup bagus dilihat dari tingkat akurasi yang tinggi (Tabel III).

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 04, Desember 2022 : 1265 - 1270

TABEL II
HASIL TINGKAT AKURASI

Akurasi	Nilai	Persentase
Fold_Best_Score	0.7556319407720783	≈76%

Setelah dilakukan proses seleksi atribut dalam algoritma C4.5 dilakukan pengukuran akurasi antara data latih dan data uji dengan perbandingan 80:20. Rentang yang dimiliki nilai akurasi berada antara 50% hingga 100% dengan tingkat akurasi rata-rata prediksi yang didapat dalam penelitian ini sebesar 76%. Dapat diartikan bahwa telah didapati model yang sangat bagus dengan tingkat akurasi cukup tinggi. Penelitian lain mengenai klasifikasi penyakit diabetes namun dengan menggunakan algoritma yang berbeda yaitu menggunakan SVM (*Support Vector Machine*) dengan menggunakan 9 faktor utama antara lain kehamilan, gluksa, tekanan darah, ketebalan kulit, insulin, BMI, fungsi silsilah diabetes, usia, dan hasil, menghasilkan tingkat akurasi hanya sebesar 704%[12].

IV. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan menegnai faktor-faktor penyakit terjadinya diabetes melitus menggunakan algoritma C4.5 dengan melalui proses seleksi fitur yang dilakukan dengan melihat nilai korelasi fitur tertinggi terhadap variabel *outcome* sebagai *target value.* Pembagian data pada penelitian ini menggunakan *grid search cross validation,* di mana pada proses tersebut data dibagi dengan *cross validation* 10 yang menghasilkan akurasi sebesar 76%.

DAFTAR PUSTAKA

[1] M. S. Kadhm, "*An Accurate Diabetes Prediction System Based on K-means Clustering and Proposed Classification Approach*," *Int. J. Appl. Eng. Res.*, vol. 13, no. 6, pp. 4038–4041, 2018.
[2] A. M. Argina, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes,"*Indones. J. Data Sci.*, vol. 1, no. 2, pp. 29–33, 2020.
[3] Indriyanti, D. Sugianti, and M. A. Al Karomi, "Peningkatan Akurasi Algoritma KNN dengan Seleksi Fitur G ain Ratio untuk Klasifikasi Penyakit Diabetes Mellitus," *IC-Tech*, vol. 7, no. 2, pp. 1–6, 2017.
[4] A. Ridwan, "Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 4, no. 1, pp. 15–21, 2020, doi: 10.47970/siskom-kb.v4i1.169.
[5] F. Aris and B. Benyamin, "Penerapan Data Mining untuk Identifikasi Penyakit Diabetes Melitus dengan Menggunakan Metode Klasifikasi," *Router Res.*, vol. 1, no. 1, pp. 1–6, 2019.
[6] H. Hendri, "Implementasi Data Mining Dengan Metode C4.5 Untuk Prediksi Mahasiswa Penerima Beasiswa," *Indones. J. Comput. Sci.*, vol. 10, no. 2, pp. 312–321, 2021, Accessed: Jan. 30, 2022. [Online]. Tersedia: http://ijcs.stmikindonesia.ac.id/index.php/ijcs/article/view/396.
[7] A.F. Nusantara and A. Kusyairi, "Aplikasi Teori Orem pada Perkembangan Perilaku *Self Care* Pasien Diabetes Mellitus Tipe 1," *Jurnal Penelitian Keperwatan*, vol. 8, no. 1, 2022.
[8] Noviandi, "Implementasi Algoritma Decision Tree C4.5 Untuk Prediksi Penyakit Diabetes," *J. INOHIM*, vol. 6, no. 1, pp. 1–5, 2018.
[9] A.M. Argina, "penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," *Indoneisan Journal of Data and Science*, vol. 1, no. 2, 2020.
[10] I. Romli and A. T. Zy, "Penentuan Jadwal Overtime Dengan Klasifikasi Data Karyawan Menggunakan Algoritma C4.5," *J. Sains Komput. Inform. (J-SAKTI)*, vol. 4, no. 2, pp. 694–702, 2020.
[11] M. Muhamad, A. P. Windarto, and S. Suhada, "Penerapan Algoritma C4.5 Pada Klasifikasi Potensi Siswa Drop Out," *KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer)*, vol. 3, no. 1, pp. 1–8, doi: 10.30865/komik.v3i1.1688, 2019.
[12] S. You and M. Kang, "*A Study on Methos to Prevent Pima Indians Diabetes using SVM*," *Korean Journal of Artificial Intelligence*, vo. 8, no. 2, pp. 7-10, 2020