YAYASAN AKRAB PEKANBARU
Jurnal AKRAB JUARA
Volume 6 Nomor 5 Edisi Desember 2021 (221-229)

39

OPTIMASI ALGORITMA GENETIKA PADA ALGORITMA C4.5 UNTUK DETEKSI DINI PENYAKIT DIABETES

Wisti Dwi Septiani, Untung Rohwadi
Universitas Bina Sarana Informatika
(Naskah diterima: 20 November 2021, disetujui: 28 Desember 2021)

Abstract

Diabetes mellitus is included in the top 3 most deadly diseases in Indonesia. Based on WHO data in 2013, diabetes contributed 6.5% to the death of the Indonesian population. Diabetes is a chronic disease characterized by high blood sugar (glucose) levels that exceed normal limits. Previous research used the C4.5 algorithm data mining classification method and showed an accuracy rate o(Septiani & Marlina, 2021)*f 95.96%. The purpose of this study is to increase the accuracy value of Algortima C4.5 by optimizing the addition of the selection feature of the Genetic Algorithm. The result of this further research is a decision tree and an increase in the accuracy value from 95.96% to 96.54% for prediction of early detection of diabetes disease.*
Keyword: *algorithm C4.5, data mining, hepatitis, genetic algorithm*

Abstrak

Diabetes melitus termasuk ke dalam 3 besar penyakit yang paling mematikan di Indonesia. Berdasarkan data WHO pada tahun 2013, diabetes menyumbang sebesar 6,5% pada kematian penduduk Indonesia. Diabetes merupakan penyakit kronis yang ditandai dengan tingginya kadar gula (glukosa) dalam darah yang melebihi batas normal. Penelitian sebelumnya menggunakan metode klasifikasi data mining Algoritma C4.5 dan menunjukkan tingkat akurasi 95,96%. Tujuan dari penelitian ini adalah untuk meningkatkan nilai akurasi Algortima C4.5 dengan melakukan optimasi penambahan fitur seleksi Algoritma Genetika.Hasl dari penelitian lanjutan ini adalah pohon keputusan dan terjadi peningkatan nilai akurasi dari 95,96% menjadi 96,54% untuk prediksi deteksi dini penyakit diabetes.
Keyword: algoritma C4.5, algoritma genetika, data mining

I. PENDAHULUAN

Diabetes adalah salah satu penyakit kronis yang mengancam jiwa dengan pertumbuhan tercepat yang telah mempengaruhi 422 juta orang di seluruh dunia menurut laporan Organisasi Kesehatan Dunia (WHO), pada tahun 2018 (Apriliah, Kurniawan, Baydhowi, & Haryati, 2021). Keterlambatan diagnosis penyakit diabetes menjadi salah satu penyebab meningkatnya

jumlah penderita diabetes (Putri, Irawan, & Rizky, 2021). Juga, peningkatan jumlah penderita diabetes dikarenakan diabetes dikenal sebagai *silent killer*. Hal ini mengacu pada banyaknya yang tidak menyadari bahwa dirinya terkena penyakit diabetes. Penderita biasanya diketahui terjangkit penyakit ini ketika sudah terjadi komplikasi tanpa adanya penanganan di awal. (Efendi & Wibawa, 2018). Oleh karena itu diperlukan kesadaran dan upaya untuk melakukan deteksi dini terhadap penyakit diabetes dengan mengenali gejala yang terjadi. Catatan data medis dalam mendiagnosa suatu penyakit dapat digunakan sebagai bahan untuk penggalian data dan menghasilkan informasi untuk memprediksi gejala penyakit.

International Diabetes Federation (IDF) pada tahun 2013 membuat estimasi bahwa jumlah pengidap diabetes di dunia mencapai 382 juta orang. Diperkirakan dari 382 juta orang tersebut, sekitar 175 juta diantaranya belum terdiagnosa, sehingga terancam berkembang tanpa disadari dan tanpa pencegahan. Di Indonesia sendiri jumlah penderita diabetes cukup tinggi, yaitu sekitar 12 juta orang pada tahun 2013. Jumlah tersebut ternyata meningkat daripada tahun-tahun sebelumnya (Efendi & Wibawa, 2018).

Literatur mengenai data mining dalam hal prediksi diabetes telah dilakukan dengan beberapa metode yaitu prediksi penyakit diabetes menggunakan Algoritma Multilayer Percepton dengan akurasi 77,7% (Setiadi, 2012), menggunakan Algoritma Decision Tree dengan akurasi 84,9% (Andriani, 2013), selanjutnya prediksi penyakit diabetes menggunakan Naive Bayes dengan akurasi 72% dan peningkatan menjadi 74,74% penambahan seleksi fitur algoritma genetika (Handayanna, Rinawati, Arisawati, & Dewi, 2017), menggunakan algoritma ID3 dengan pemilihan atribut terbaik menghasilkan akurasi 84,77% (Efendi & Wibawa, 2018), dan prediksi diabetes pada tahap awal menggunakan algoritma klasifikasi Random Forest (Apriliah et al., 2021). Belum banyak studi tentang perbandingan algoritma klasifikasi.

Pada penelitian ini dilakukan optimasi terhadap Algoritma C4.5 menggunakan algoritma Genetika. Optimasi adalah proses menyelesaikan suatu masalah tertentu supaya berada pada kondisi yang paling menguntungkan dari suatu sudut pandang, yaitu berhubungan dengan pencarian nilai minimum atau nilai maksimum (Buani, 2016). Pada penerapan optimasi menggunakan algoritma genetika terdapat peningkatan akurasi dari

97,66% menjadi 99,33% pada agoritma Naïve Bayes untuk prediksi *fertility* (Buani, 2016)dan peningkatan akurasi dari 83,81% menjadi 86,47% pada algoritma C4.5 untuk prediksi *phising website* (Sunge, 2018).

Decision Tree merupakan salah satu algoritma klasifikasi data mining. Menurut Gorunescu dalam (Sunge, 2018) Algortima dalam kasifikasi yang banyak digunakan ialah Decision Tree. Dikarenakan sangat mudah dimengerti dan dijabarkan oleh banyak pengguna juga mudah dipahami dimana cabang pohon disimpulkan dalam bentuk klasifikasi. Tujuan dari penelitian ini adalah melakukan optimasi algoritma C4.5 menggunakan algoritma genetika untuk meningkatkan nilai akurasinya sehingga prediksi yang dihasilkan lebih baik dan akurat.

II. KAJIAN TEORI

2.1 Data Mining

Data mining yang sering juga disebut *knowledge discovery in database* (KDD) adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menentukan pola keteraturan, pola hubungan dalam set data berukuran besar (Santosa, 2007).

Berdasarkan tugasnya, *data mining* dikelompokkan menjadi 6 yaitu deskripsi, estimasi, prediksi, klasifikasi, clustering, dan asosiasi (Larose, 2005). Klasifikasi (taksonomi) adalah proses menempatkan objek tertentu (konsep) dalam satu set kategori, berdasarkan masing-masing objek (konsep) *property* (Gorunescu, 2011). Proses klasifikasi didasarkan pada empat komponen mendasar yaitu kelas, prediktor, *training set*, dan pengujian *dataset*.

Diantara model klasifikasi yang paling popular adalah *Decision/Classification Trees, Bayesian Classifiers/Naïve Bayes Classifiers, Neural Networks, Statistical Analysis, Genetic Algorithms, Rough Sets, K-Nearest Neighbor Classifier, Rule-based Methods, Memory Based Reasoning, Support Vector Machines* (Gorunescu, 2011).

2.2 Algoritma C4.5

Algoritma C4.5 menggunakan konsep *information gain* atau *entropy reduction* untuk memilih pembagian yang optimal (Larose, 2005). Tahapan dalam membuat pohon keputusan dengan algoritma C4.5 (Gorunescu, 2011) yaitu:

1. Mempersiapkan data *training,* dapat diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokkan dalam kelas-kelas tertentu.
2. Menentukan akar dari pohon dengan menghitung nilai *gain* yang tertinggi dari ma-

sing-masng atribut atau berdasarkan nilai *index entropy* terendah. Sebelumnya dihitung terlebih dahulu nilai *index entropy*, dengan rumus:

$$Entropy\ (i) = -\sum_{j=1}^{m} f(i,j).log\ 2\ f[(i,j)]$$

3. Hitung nilai gain dengan rumus:

$$Entropy\ split = \sum_{i=1}^{p} \binom{n1}{n}.\ IE(i)$$

4. Ulangi langkah ke-2 hingga semua record terpartisi. Proses partisi pohon keputusan akan berhenti disaat:
   a. Semua tupel dalam *record* dalam simpul N mendapat kelas yang sama.
   b. Tidak ada atribut dalam *record* yang dipartisi lagi.
   c. Tidak ada *record* di dalam cabang yang koson

2.3 Algoritma Genetika

Pada tahun 1970 Algoritma Genetica (GA) diperkenalkan oleh John Holland di Universitas Michigan, bahwa dari bagian masalah merupakan bentuk dari adaptasi dari alam maupun buatan yang dapat diformulakan menjadi bagian genetika (Sunge, 2018). Menurut Desiani dan Muhammad dalam (Buani, 2018) Algoritma genetika merupakan suatu algoritma pencarian berdasarkan pada mekanisme seleksi alam dan genetika alam. Algoritma genetika dimulai dengan sekumpulan solusi awal (individu) yang disebut populasi. Satu hal yang sangat penting adalah bahwa satu individu menyatakan satu solusi. Populasi awal akan berevolusi menjadi populasi baru melalui serangkaian iterasi (generasi). Pada akhir iterasi, algoritma genetika mengembalikan satu anggota populasi yang terbaik sebagai solusi untuk masalah yang dihadapi. Pada setiap iterasi, proses evolusi yang terjadi adalah sebahai berikut:

a) Dua individu dipilih sebagai orang tua (*parent*) berdasarkan mekanisme tertentu. Kedua parent ini kemudian dikawinkan melalui operator crossover (kawin silang) untuk menghasilkan dua individu anak atau offspring.
b) Dengan probabilitas tertentu, dua individu anak ini mungkin mengalami perubahan gen melalui operator mutation.
c) Suatu skema penggantian *(replacement scheme)* tertentu diterapkan sehingga menghasilkan populasi baru.

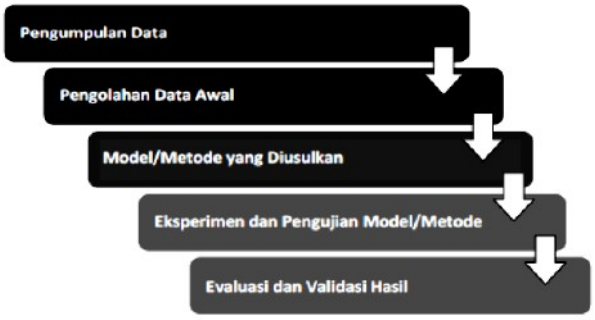
Proses ini terus berulang sampai kondisi berhenti *(stopping condition)* tertentu. Kondisi berhenti bisa berupa jumlah iterasi tertentu, waktu tertentu, atau ketika variansi individu-individu dalam populasi tersebut

sudah lebih kecil dari suatu nilai tertentu yang diinginkan

III. METODE PENELITIAN

Penelitian dilakukan dengan cara melakukan eksperimen dalam bentuk sistem penunjang keputusan untuk prediksi deteksi dini penyakit diabetes.

Untuk menyelesaikan penelitian dibuat desain penelitian yang dirancang sebagai acuan atau pedoman penelitian dan dapat digambarkan sebagai berikut:



Sumber: (Septiani & Rohwadi, 2021)

Gambar 1. Kerangka Penelitian

1. Pengumpulan Data

   Data yang digunakan pada penelitian adalah data riwayat pasien penyakit rumah sakit di Sylhet, Bangladesh dalam bentuk dataset diabetes dari *Machine Learning Repository* UCI (Universitas California Invine) dengan alamat web: http://archives.uci.edu/ml/. Data yang terkumpul sebanyak 520 data dengan atribut: *age, gender, polyuria, polydipsia, sudden_weight_loss, weakness, polyphagia, genital_thrush, visual_blurring, itching, irritability, delayed_healing, partial_paresis, muscle_stiffness, alopecia, obesity,* dan *class* (atribut hasil prediksi).

2. Pengolahan Data Awal

   Dilakukan *data validation, data integration and transformation,* dan *data size and dicritization*. Pada dataset diabetes ini semua atribut digunakan dan tidak ada data kosong.

3. Model/Metode yang Diusulkan

   Model yang diusulkan adalah Algoritma Decision Tree dengan seleksi fitur Algoritma Genetika.

4. Eksperimen dan Pengujian Model/Metode

   Eksperimen dilakukan dengan mengolah dataset diabetes menggunakan Algoritma Decision Tree dengan seleksi fitur Algoritma Genetika. *Tools* yang digunakan adalah Rapidminer.

5. Evaluasi dan Validasi Hasil

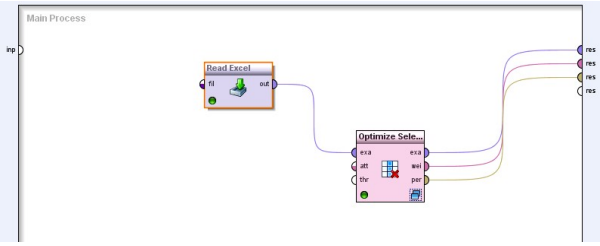
   Evaluasi dan validasi diuji tingkat akurasinya dengan Confusion Matrix.

IV. HASIL PENELITIAN

Pengujian dengan tools RapidMiner untuk mengolah data dengan tahapan sebagai berikut:

1. Pengujian menggunakan Algoritma C4.5

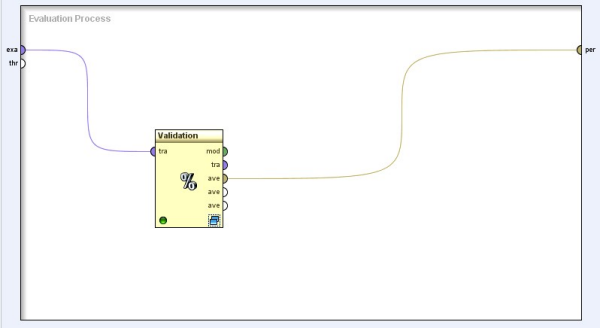
dengan fitur seleksi Algoritma Genetika.



Sumber: (Septiani & Rohwadi, 2021)

Gambar 2. *Optimize Selection (Evolutionary*

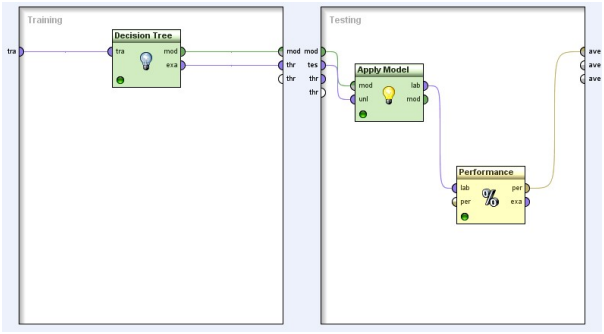
Berdasarkan gambar 2, database Diabetes dihubungankan dengan *Feature Optimize Selection* (*Evolutionary*) dan didalamnya diberikan *Cross Validation*.



Sumber: (Septiani & Rohwadi, 2021)

Gambar 3. *Cross Validation*

Berdasarkan gambar 3, penggunaan *Cross Validation* dalam prediksi hepatitis terdiri dari *10-fold validation* Pada *Cross Validation* terdapat tahap dalam penggunaan algoritma *Decision Tree*.



Sumber: (Septiani & Rohwadi, 2021)

Gambar 4. *Model Decision Tree*

Pada gambar 4 setelah model *Decision Tree* maka tahap terakhir dilakukan proses terhadap model tersebut untuk menampilkan hasil berupa tingkat akurasi.

2. Hasil akurasi dari pengujian Algoritma C4.5 dengan fitur seleksi Algoritma Genetika.

Tabel 1. Akurasi Algoritma Decision Tree

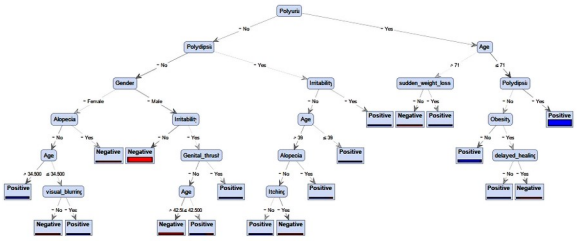
	True Positive	**True Negative**	**Class Precision**
Pred Positive	308	6	99.09%
Pred Negative	12	194	94.17%
Class Recall	96.25%	97.00%	
Accuracy		**96.54%**	

Sumber: (Septiani & Rohwadi, 2021)

Berdasarkan data di tabel 1 dapat diambil kesimpulan bahwa hasil prediksi menggunakan Decision Tree dengan fitur seleksi Algoritma Generika tingkat akurasinya adalah 96,54%. Terjadi peningkatan nilai akurasi dari penelitian sebelumnya yaitu 95,96% (Septiani & Marlina, 2021) dengan menggu-

nakan metode yang sama tetapi tidak menggunakan fitur seleksi.

Hasil dari penggunaan Algoritma C4.5 ini adalah pohon keputusan seperti pada gambar 7 yang menghasilkan 19 *rule*.



Gambar 5. Pohon Keputusan

Sumber: (Septiani & Rohwadi, 2021)

Sebanyak 19 *rules* dihasilkan dengan ketentuan sebagai berikut:

R1: IF Polyuria = NO dan Polydipsia = NO dan Gender = Female dan Alopecia = NO dan Age > 34,5 maka "POSITIVE" ok

R2: IF Polyuria = NO dan Polydipsia = NO dan Gender = Female dan Alopecia = NO dan Age < 34,5 dan Visual_blurring= NO maka "NEGATIVE"

R3: IF Polyuria = NO dan Polydipsia = NO dan Gender = Female dan Alopecia = NO dan Age < 34,5 dan Visual_blurring= YES maka "POSITIVE"

R4: IF Polyuria = NO dan Polydipsia = NO dan Gender = Female dan Alopecia = YES maka "NEGATIVE"

R5: IF Polyuria = NO dan Polydipsia = NO dan Gender = Male dan Irritability = NO maka "NEGATIVE"

R6: IF Polyuria = NO dan Polydipsia = NO dan Gender = Male dan Irritability = YES dan Genital_trush = NO dan Age > 42,5 maka "NEGATIVE"

R7: IF Polyuria = NO dan Polydipsia = NO dan Gender = Male dan Irritability = YES dan Genital_trush = NO dan Age < 42,5 maka "POSITIVE"

R8: IF Polyuria = NO dan Polydipsia = NO dan Gender = Male dan Irritability = YES dan Genital_trush = YES maka "POSITIVE"

R9: IF Polyuria = NO dan Polydipsia = YES dan Irritability = NO dan Age > 39 dan Alopecia = NO dan Itching = NO maka "POSITIVE"

R10: IF Polyuria = NO dan Polydipsia = YES dan Irritability = NO dan Age > 39 dan Alopecia = NO dan Itching = YES maka "NEGATIVE"

R11: IF Polyuria = NO dan Polydipsia = YES dan Irritability = NO dan Age > 39 dan Alopecia = YES maka "POSITIVE"

R12: IF Polyuria = NO dan Polydipsia = YES dan Irritability = NO dan Age < 39 maka "POSITIVE"

R13: IF Polyuria = NO dan Polydipsia = YES dan Irritability = YES maka "POSITIVE"

R14: IF Polyuria = YES dan Age > 71 dan Sudden_weight_loss = NO maka "NEGATIVE"

R15: IF Polyuria = YES dan Age > 71 dan Sudden_weight_loss = YES maka "POSITIVE"

R16: IF Polyuria = YES dan Age < 71 dan Polydispia = NO dan Obesity = NO maka "POSITIVE"

R17: IF Polyuria = YES dan Age < 71 dan Polydispia = NO dan Obesity = YES dan Delayed_healing = NO maka "POSITIVE"

R18: IF Polyuria = YES dan Age < 71 dan Polydispia = NO dan Obesity = YES dan Delayed_healing = YES maka "NEGATIVE"

R19: IF Polyuria = YES dan Age < 71 dan Polydispia = YES maka "POSITIVE"

V. KESIMPULAN

Rules yang dihasilkan dari pengujian penggunaan fitur seleksi Algoritma Genetika yang diterapkan pada Algoritma C4.5 dapat dijadikan kontribusi dalam pengambilan keputusan terhadap penyakit diabetes. Evaluasi dalam pengujian menggunakan Algoritma C4.5 dengan seleksi fitur Algoritma Genetika ini didapatkan nilai akurasi 96.54%. Hasil penelitian ini menunjukkan adanya peningkatan nilai akurasi dari penelitian sebelumnya 95.96 % tanpa fitur seleksi. Sehingga dapat disimpulkan bahwa penggunaan fitur seleksi mampu meningkatkan nilai akurasi. Penelitian ini dapat dijadikan masukkan untuk dilanjutkan kembali dengan metode optimasi lain seperti Adabbost dan PSO. Berdasarkan penelitian ini dapat diberikan saran untuk diadakannya penelitian lebih lanjut dengan melakukan pengujian dengan metode lain seperti SVM, Nural Network, ataupun komparasi dari beberapa metode klasifikasi data mining

DAFTAR PUSTAKA

Andriani, A. (2013). Sistem Prediksi Penyakit Diabetes Berbasis Decision Tree. *Jurnal Bianglala Informatika*, *I*(1), 1–10.

Apriliah, W., Kurniawan, I., Baydhowi, M., & Haryati, T. (2021). Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest. *Sistemasi: Jurnal Sistem Informasi*, *10*(1), 163–171. https://doi.org/10.32520/stmsi.v10i1.1129

Buani, D. C. P. (2018). Prediksi Penyakit Hepatitis Menggunakan Algoritma Naive Bayes Dengan Seleksi Fitur Algoritma Genetika. *Jurnal Evolusi*, *6*(2), 1–5. Retrieved from ejournal.bsi.ac.id

Efendi, M. S., & Wibawa, H. A. (2018).

Prediksi Penyakit Diabetes Menggunakan Algoritma ID3 dengan Pemilihan Atribut Terbaik (Diabetes Prediction using ID3 Algorithm with Best Attribute Selection). *JUITA*, *VI*(1), 29–35.

Gorunescu, F. (2011). *Data Mining: Concepts and Techniques*. Verlag Berlin Heidelberg: Springer.

Handayanna, F., Rinawati, Arisawati, E., & Dewi, L. S. (2017). Prediksi Penyakit Diabetes Menggunakan Naive Bayes dengan Optimasi Parameter Menggunakan Algoritma Genetika. *KNiST (Konferensi Nasional Ilmu Sosial & Teknologi)*, 71–76.

Larose, D. T. (2005). *Discovering Knowledge in Database*. New Jersey: John Willey & Sons Inc.

Putri, S. U., Irawan, E., & Rizky, F. (2021). Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4.5. *KESATRIA Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)*, *2*(1), 39–46.

Santosa, B. (2007). *Data Mining Teknik Pemanfaat Data Untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu.

Septiani, W. D., & Marlina, M. (2021). Comparison of Decision Tree, Naïve Bayes, and Neural Network Algorithm for Early Detection of Diabetes. *Jurnal Pilar Nusa Mandi*ri, 17(1), 73–78. https://doi.org/10.33480/pilar.v17i1.2213

Septiani, W. D., & Rohwadi, U. (2021). *Laporan Akhir Penelitian*. Jakarta.

Setiadi, A. (2012). Penerapan Algoritma Multilayer Perceptron Untuk Deteksi Dini Penyakit Diabetes. *Paradigma*, *14*(1), 46–59.

Sunge, A. S. (2018). Optimasi Algoritma C4.5 Menggunakan Genetic Algoritma Dalam Memprediksi Website Phishing. *Seminar Nasional Inovasi Dan Tren (SNIT)*, 92.