Terakreditasi SINTA Peringkat 2
Surat Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti No. 10/E/KPT/2019
masa berlaku mulai Vol. 1 No. 1 tahun 2017 s.d Vol. 5 No. 3 tahun 2021

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vl. 4 No. 3 (2020) 569 - 575 ISSN Media Elektronik: 2580-0760

Seleksi Fitur pada Klasifikasi Penyakit Gula Darah Menggunakan *Particle Swarm Optimization* (PSO) pada Algoritma C4.5

Dwi Meylitasari Br.Tarigan[1], Dian Palupi Rini[2], Samsuryadi[3]
[1,2,3]Magister Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Sriwijaya
[1]dwimeylitasaritarigan@gmail.com, [2]dian.palupi.rini@gmail.com, [3]samsuryadi@gmail.com

Abstract

Diabetes Mellitus (DM) is a disease caused by blood sugar level increased were higher than the maximum limit. Food consumed tends to contain uncontrolled sugar which could cause the drastic increase of blood sugar level. It is necessary to efforts, to increasing the public awareness to controlling blood sugar and the risks of increasing blood sugar level so as to determine of preventive and early detection measures One of used of data mining technique is information technology in the health sector which used a lot as a decision maker to predicting and diagnosing a several disease. This research aims to optimizing the features on classification of the data mining with the C4.5 algorithm using Particle Swarm Optimization (PSO) to detect the blood sugar level in patient. The dataset used is the effect of physical activity to the Blood Sugar Level at H. Abdul Manan Simatupang Kisaran Regional Public Hospital. The amount of dataset used is 42 record with 10 attributes. The result of this research obtained that the Particle Swarm Optimization (PSO) may increasing the accuracy performance of C4.5 from 86% to 95%. Whereas the evaluation result of the AUC Value increasing from 0,917 to 0,950. From those 10 attributes which are then selection with using PSO into 7 attributes used to determine the prediction of sugar level. Therefore the Algorithm C4.5 using the Particle Swarm Optimization (PSO) may provide the best solution to the accuracy of detection blood sugar levels.

Keywords: Data Mining , Algorithm C4.5, Particle Swarm Optimization, PSO, classification

Abstrak

Diabetes Melitus (DM) merupakan penyakit yang disebabkan oleh meningkatnya kadar gula darah yang lebih tinggi dari batas maksimum. Makanan yang dikonsumsi kecendrungan memiliki kandungan gula yang tidak terkontrol yang dapat menyebabkan kadar gula darah meningkat secara drastis. Hal tersebut perlu dilakukan upaya untuk meningkatkan kesadaran kepada masyarakat dalam mengontrol gula darah dan risiko meningkatnya kadar gula darah sehingga dapat menentukan langkah-langkah pencegahan dan deteksi dini yang tepat. Salah satu penggunaan teknik data mining merupakan teknologi informasi di bidang kesehatan yang banyak digunakan sebagai sistem pengambil keputusan untuk memprediksi dan mendiagnosa beberapa penyakit. Penelitian ini bertujuan untuk mengoptimasi fitur pada klasifikasi data mining dengan algoritma C4.5 menggunakan *Particle Swarm Optimization* (PSO) untuk mendeteksi kadar gula darah pada pasien. Dataset yang digunakan adalah Pengaruh Aktifitas Fisik Terhadap Kadar Gula Darah di RSUD H.Abdul Manan Simatupang Kisaran. Dataset yang digunakan berjumlah 42 record dengan 10 atribut. Hasil penelitian ini didapatkan bahwa *Particle Swarm Optimization* (PSO) dapat meningkatkan kinerja akurasi C4.5 dari 86% menjadi 95%. Sedangkan hasil evaluasi pada nilai AUC meningkat dari 0,917 menjadi 0,950. Dari 10 atribut tersebut yang kemudian diseleksi menggunakan PSO menjadi 7 atribut yang digunakan dalam menentukan prediksi kadar gula. Dengan demikian algoritma C4.5 menggunakan *Particle Swarm Optimization* (PSO) dapat memberikan solusi terbaik terhadap akurasi pendeteksi kadar gula darah.

Kata kunci: *data mining*, algoritma C4.5, *Particle Swarm Optimization*, PSO, klasifikasi.

1. Pendahuluan

Dari data WHO, prevelensi diabates dunia dari tahun 1980 meningkat menjadi 2 kali lipat, meningkat dari 4,7 persen menjadi 8,5 persen, diduduki oleh penderita yang memiliki usia lanjut [1]. Pada tahun 2025 Indonesia di prediksi meningkat menjadi 5 besar dengan jumlah diabetes sebanyak 12,4 juta jiwa [2].

Diabetes Mellitus (DM) adalah penyakit yang tidak menular yang menjadi salah satu masalah serius bagi dunia kesehatan di Indonesia. [3]. Penderita DM harus

Diterima Redaksi : 12-04-2020 | Selesai Revisi : 30-05-2020 | Diterbitkan Online : 20-06-2020

memperhatikan pola makan yang meliputi jadwal, jumlah dan jenis makanan serta mengatur jadwal pemeriksaan kadar gula rutin ke dokter. Kadar gula darah meningkat dratis setelah mengkonsumsi makanan tertentu karena kecenderungan makanan yang dikonsumsi memiliki kandungan gula darah yang tidak terkontrol [4].

Faktor risiko penderita diabetes adalah gaya hidup pasien yang kurang sehat seperti, aktivitas fisik, diet yang tidak sehat dan tidak seimbang. Oleh karena itu, untuk penanggulangan diabetes melitus adalah mengendalikan faktor risiko [5]. Telah banyak instansi kesehatan rumah sakit, puskesmas, klinik yang mengatasi berbagai pasien DM, dari beberapa instansi tersebut masih banyak yang belum memberikan data yang cepat dan akurat. Perlu secara efektif dalam mengelola informasi dengan proses data mining, dari data yang besar akan menghasilkan data yang baru dan dapat memberikan informasi secara cepat dan aktual.

Data mining adalah sebuah langkah penting dalam proses menemukan pengetahuan dan informasi [6]. Dalam data mining prediksi dan klasifikasi banyak digunakan untuk menganalisis suatu data yang dapat menggambarkan kelas data atau untuk memprediksi data di masa depan.Dalam melakukan prediksi terhadap penyakit, telah banyak dilakukan penelitian khususnya di bidang computer science, dengan teknik data mining untuk memprediksi penyakit menggunakan berbagai algoritma seperti Naive Bayes, SVM, ID3, C45 dan lain-lain.

Penelitian yang dilakukan oleh Wu.H, dengan melakukan prediksi penyakit diabetes menggunakan algoritma Regresi Logistik, pada model yang dilakukan menghasilkan akurasi prediksi sebesar 3,04%. Dalam penelitiannya Wu H, memprediksi dengan beberapa pengukuran dengan menggunakan variabel yaitu waktu kelahiran, glukosa plasma, tekanan darah, insulin, BMI, riwayat diabetes dan umur [7].

Pada penelitian lain yang dilakukan oleh Sisodia dengan menggunakan algoritma Naive Bayes menghasilkan tingkat akurasi sebesar 76,03%. Data yang digunakan untuk mengukur tingkat akurasi ini menggunakan data Pima Indians Diabetes Databases (PIDD) [8].

Penelitian lain pada prediksi Kelahiran Bayi Secara Prematur yang dilakukan oleh Ari dengan menggunakan Algoritma C4.5 Berbasis Particle Swarm Optimization. Penelitian ini menggunakan data record sebanyak 250 record dan menghasilkan tingkat akurasi sebesar 93,60% dengan algoritma C4.5, sedangkan C4.5 berbasis PSO menghasilkan akurasi 96.00%. Optimization of C4.5. Algoritma C4.5 berbasis Praticle Swarm Optimization (PSO) memiliki tingkat akurasi tertinggi yaitu 96% dibandingkan dengan dua algoritma lainnya [9].

Tejas mengusulkan dengan membandingkan beberapa teknik klasifikasi data mining yaitu Naïve Bayes, Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), dan Decision Tree. Dari perbandingan algoritma tersebut menghasilkan 97,33% tingkat akurasi tertinggi pada algoritma Artificial Neural Network (ANN). Tingkat akurasi ini membuktikan bahwa algoritma machine learning memiliki potensi secara signifikan untuk meningkatkan lebih dari metode klasifikasi konvensional [10].

Selain itu penelitian yang dilakukan oleh Meng-Chai, dengan menggunakan dataset dari UCI Machine Learning database dalam meyeleksi fitur. Meng-Chai membandingkan 4 teknik algoritma dan menghasilkan akurasi yang berbeda. Pada algoritma Regresi Logistik sebesar 83,59%, SVM sebesar 86,51% pada C4.5 menghasikan akurasi sebesar 82,56% , dan pada C4.5 + PSO menghasilkan akurasi 90,77%. Hasil dari perbandingan kedua algoritma tersebut, PSO pada algoritma C4.5 dapat memberikan tingkat akurasi tertinggi dibanding algoritma yang lainnya [11].

Algoritma C4.5 merupakan algoritma yang memiliki kemampuan dalam mengolah dataset seperti klasifikasi, pada setiap atribut bersifat diskrit, binari dan *continue* [12]. Sedangkan pada PSO dinilai mampu untuk meningkatkan kinerja klasifikasi, karena pada penelitian ini PSO digunakan sebagai seleksi fitur yang efektif , karena menggunakan beberapa parameter sehingga waktu komputasi akan cepat [11].

Dari beberapa uraian diatas, algoritma C4.5 berbasis PSO dinilai mampu menghasilkan tingkat akurasi yang tinggi dalam membangun model klasifikasi. Penelitian ini bertujuan untuk mengoptimasi fitur pada klasifikasi data mining dengan algoritma C4.5 menggunakan Particle Swarm Optimization (PSO) untuk mendeteksi kadar gula darah pada pasien. Dataset yang digunakan adalah Pengaruh Aktifitas Fisik Terhadap Kadar Gula Darah di RSUD H.Abdul Manan Simatupang Kisaran.

2. Metode Penelitian

Pada metode penelitian ini menjelaskan tentang metode penelitian, dataset yang digunakan, teori algoritma C4.5, algoritma *Particle Swarm Optimization* (PSO), validasi pengujian menggunakan K-Fold Cross Validation, pengukuran performa menggunakan metode evaluasi confusion matrix.

2.1. Pengumpulan Data

Pada tahap pengumpulan data ini adalah teknik atau cara yang akan dipakai untuk mengumpulkan data. Data yang kita cari harus sesuai dengan tujuan penelitian. Dalam pengumpulan data terdapat sumber data yaitu data primer dan data sekunder. Data primer adalah data yang hanya dapat kita peroleh dari sumber asli atau pertama, sedangkan data sekunder merupakan data yang sudah tersedia sehingga kita tinggal mencari dan mengumpulkan. Data sekunder dapat diperoleh dengan lebih mudah dan cepat karena sudah tersedia, misalnya

di perpustakaan, perusahaan, instansi, biro pusat statistik, dan kantor pemerintahan.

Pada penelitian ini, peneliti menggunakan data sekunder yang telah diambil pada RSUD H.Abdul Manan Simatupang Kisaran. Atribut yang terdapat pada didalamnya merupakan faktor-faktor yang mempengaruhi aktifitas terhadap Kadar Gula Darah pasien Diabetes Mellitus.

2.2. *Dataset*

Penelitian ini menggunakan *dataset* yang diambil dari RSUD H.Abdul Manan Simatupang Kisaran, data Pasien dengan Pengaruh Aktifitas Fisik Terhadap Kadar Gula Darah Diabetes Melitus. Dari sampel ini, yang nantinya akan dapat diberlakukan poupulasi. Sampel yang diambil dari populasi harus betul-betul representatif (mewakili). Pada data ini memiliki 42 record dan 10 Atribut. Atribut yang digunakan adalah, Jenis Kelamin (Laki-laki dan Perempuan), Umur (dibawah 45 tahun dan diatas 45 tahun) , Pendidikan (SD/SMP/ SMA dan Perguruan Tinggi), IMT (Obesitas dan Normal), Riwayat Ayah (Tidak Ada dan Ada), Riwayat Ibu (Tidak Ada dan Ada), Pengetahuan (Tidak Baik dan Baik), Aktifitas (Tidak Baik dan Baik), Diet (Tidak Baik dan Baik) dan Obat (Tidak Baik dan Baik). Sedangkan yang menjadi target atau kelas dalam prediksi adalah Kadar Gula Tinggi dan Kadar Gula Normal.

Tabel 1. Total Sampel dengan Kelas Prediksi

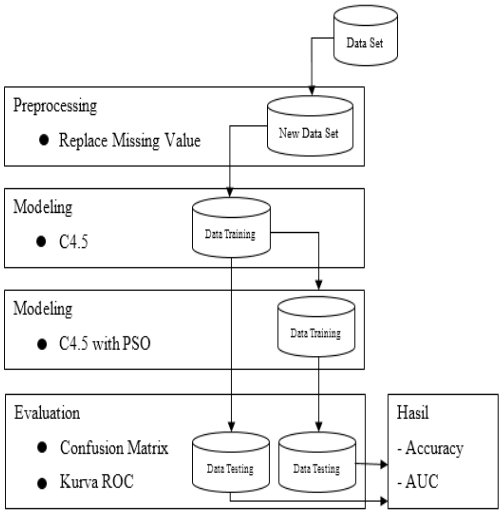
Tinggi (0)	Normal (1)	Total Sampel
31	11	42

Berikut adalah dataset pasien penderita Kadar Gula Darah dengan beberapa variabel penyebabnya.

Tabel 2. Dataset Pasien penderita Kadar Gula Darah

Jenis Kelamin	Umur	Pendidikan	IMT	Riwayat Ayah	Riwayat Ibu	Pengetahuan	Aktifitas	Diet	Obat	Kadar Gula
2	0	1	0	0	0	0	0	1	1	0
2	0	2	1	1	0	0	1	0	0	0
2	0	1	0	0	0	0	0	0	1	0
1	1	1	1	0	0	1	0	1	0	1
2	0	1	0	0	0	0	0	0	1	0
1	0	2	1	1	0	0	1	0	0	0
2	0	1	1	0	0	1	1	1	0	1
2	0	1	0	0	0	0	0	0	1	0
2	0	1	0	0	0	0	1	0	1	0
1	0	2	1	0	0	0	1	1	0	1
2	1	1	1	1	0	0	0	0	0	0
2	0	2	1	0	0	0	1	0	1	1
1	0	1	0	0	0	0	0	1	1	0
2	0	1	0	0	0	0	0	0	0	0
1	0	2	1	0	0	1	1	1	0	1
2	1	2	1	1	1	0	0	0	0	0
1	0	1	1	0	0	1	1	0	1	1
2	0	1	0	0	0	0	0	0	0	0
1	1	1	0	0	1	0	1	0	0	0
2	1	2	0	1	0	0	0	0	0	0
2	0	2	1	0	0	0	1	1	0	1
2	0	2	1	0	0	0	0	0	1	0
2	0	1	1	0	0	1	0	0	1	0
1	1	1	0	0	0	0	0	0	0	0
2	0	1	1	0	1	1	0	0	1	0
1	0	1	1	0	0	1	1	1	1	1
2	0	2	0	0	1	0	1	0	0	0
1	0	2	1	1	0	0	0	0	1	0
2	1	2	1	0	0	0	0	0	0	0
1	0	2	0	0	0	0	0	0	1	0
2	1	2	0	1	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	1	0	0
1	0	1	1	0	0	1	1	1	0	1

2.3. Metode Penelitian



Gambar 1. Metode Penelitian

Pada bagian ini menjelaskan metode penelitian yang akan digunakan. Berdasarkan model algoritma yang akan dipakai pada penelitian ini adalah : Algoritma C4.5 adalah model klasifikasi yang mengklasifikasikan dengan atribut yang dipakai dan akan membentuk pohon keputusan dengan aturan (rules) *Particle Swarm Optimization* (PSO) yaitu pencarian solusi yang terbaik dengan menyeleksi fitur dengan meningkatkan bobot atribut *(attribute weight*).

2.4. *Algoritma C4.5*

Algoritma C4.5 merupakan kelompok algoritma pohon Keputusan (decision tree). Algoritma ini mempunyai input berupa training samples data dan samples. Training samples data contoh yang akan digunakan untuk membangun sebuah tree yang telah diuji kebenarannya. Sedangkan samples merupakan field-field data yang nantinya akan digunakan sebagai parameter dalam melakukan klasifikasi data [14]. Algoritma C4.5 memiliki kelebihan yaitu mudah dimengerti, fleksibel, dan menarik karena dapat divisualisasikan dalam bentuk gambar (pohon keputusan) [13]. Algoritma C4.5 digunakan untuk membangun sebuah pohon keputusan dengan langkah-langkah sebagai berikut [15] :

Langkah 1. Proses awal yang dilakukan adalah menyiapkan data training yang akan digunakan pada pengujian algoritma c4.5 dengan PSO

Langkah 2. Memilih atribut yang akan dijadikan sebagai akar

Langkah 3. selanjutnya membuat cabang pada tiap-tiap nilai

Langkah 4. Membagi kasus dalam cabang berdasarkan nilai *entrophy* yang memilil nilai 0. Jika tidak terdapat lagi nilai 0 maka porses pencarian akan berhenti.

Langkah 5. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Untuk memilih atribut yang akan dijadikan sebagai akar, didasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada. Sebelum menghitung *gain* dari atribut, hitung dahulu nilai entopi yaitu :

$$Entrophy\ (S) = \sum_{i=1}^{n} pi\ log_2\ pi \qquad (1)$$

Keterangan :
- S : Himpunan (dataset)
- n : Banyaknya record
- Pi : probability yang didapat dari jumlah ya atau tidak dibagi keseluruhan total kasus

Untuk menghitung gain digunakan rumus:

$$Gain\ (A) = Entrophy\ (S) - \sum_{i}^{n} \frac{|Si|}{|S|}\ x\ Entrophy\ (Si) \quad (2)$$

Keterangan :
- S : himpunan (dataset)
- A : atribut yang akan dipakai
- n : jumlah partisi atribut A
- |Si| : jumlah kasus pada partisi ke-i
- |S| : jumlah kasus dalam S

2.5. *Particle Swarm Optimization* (PSO)

Particle Swarm Optimization (PSO) adalah algoritma pencarian berbasis populasi yang digunakan untuk mencari solusi acak dan salah satu algoritma optimasi yang dapat digunakan untuk pengambilan keputusan. PSO adalah metode optimasi heuristic global yang diperkenalkan oleh Dr. Kennedy dan Eberhart pada tahun 1995 berdasarkan penelitian perilaku kawanan burung dan ikan [16].

Setiap partikel dalam PSO juga diartikan dengan kecepatan partikel terbang melalui ruang pencarian dengan kecepatan yang dinamis disesuaikan untuk perilaku historis mereka

Oleh karena itu, partikel memiliki kecendruangan untuk terbang menuju daerah pencarian yang lebih baik dan lebih baik selama proses pencarian [16]. Rumus untuk menghitung perpindahan posisi dan kecepatan partikel yaitu :

$$Vi\ (t) = Vi\ (t-1) + c1r1\ [Xpbest\ i - Xi\ (t)] + c2r2\ [XGbest\ i - Xi\ (t) \qquad (3)$$

$$\text{Xi (t)} = \text{Xi (t} - 1) + \text{Vi (t)} \qquad (4)$$

Keterangan :
- Vi (t) : kecepatan partikel i saat iterasi t
- Xi (t) : posisi partikel i saat iterasi t
- c1 dan c2 : learning rates untuk kemampuan individu (cognitive) dan pengaruh sosial (group)
- XPbest i : posisi terbaik partike i
- XGbest i : posisi terbaik global
- r1 dan r2 : bilangan random yang bernilai antara 0 sampai 1

2.6. *K-Fold Cross Validation*

K-Fold Cross Validation merupakan metode statistik untuk menilai dan membandingkan algoritma pembelajaran dengan membagi data menjadi dua yaitu data latih dan data uji [17]. *K-Fold Cross Validation* merupakan bentuk dasar lintas validasi dimana *K-Fold Cross Validation* akan melakukan perulangan sebanyak K validation. Model validasi pada penelitian ini adalah 10 fold cross validation. Model 10 *K-Fold Cross Validation* ini akan membagi data menjadi 10 bagian dan akan melakukan perulangan sebanyak 10 kali dalam melakukan setiap kali pengujian.

2.7. *Confusion Matrix*

Confusion Matrix adalah metode evaluasi atau metode pengujian dari dataset yang mendeskripsikan hasil berdasarkan data testing. Metode evaluasi ini menghitung diantara dua kelas, dimana kelas pertama dianggap positif dan kelas kedua dianggap negatif [18]. Evaluasi *Confusion Matrix* ini akan menghasilkan nilai *accuracy, sensitifity (recall)* dan *precicion*. Metode evaluasi menggunakan matriks akan terdeskripsi seperti tabel dibawah ini :

Tabel 3. Model Evaluasi Confusion Matrix

Klasifikasi yang benar	Diklasifikasikan sebagai	
	+	-
+	*True Positive*	*False Negative*
-	*False Positive*	*True Negative*

Tabel 3 menjelaskan bahwa :
- *True Positive* (TP) : merupakan jumlah kasus bernilai positif diklasifikasikan positif,
- *True Negative* (TN) : merupakan jumlah kasus bernilai negatif diklasifikasikan negatif,
- *False Positive* (FP) : merupakan jumlah kasus bernilai negatif diklasifikasikan positif,

False Negative (FN) : merupakan kasus bernilai positif diklasifikasikan negatif.

Nilai *accuracy* pada metode evaluasi confusion matrix merupakan proporsi atau besarnya jumlah prediksi yang benar, dapat dihitung menggunakan persamaan dibawah ini :

$$accuracy = \frac{tp+tn}{tp+tn+fp+fn} \quad (5)$$

Nilai *sensitifity* atau *recall* pada metode evaluasi confusion matrix digunakan untuk memilih model yang paling efisien dan mengukur kinerja klasifikasi terhadap tupel yang positif yang diidentifikasikan dengan benar, *sensitifity* dapat dihitung menggunakan persamaan dibawah ini :

$$sensitifity = \frac{tp}{tp+fn} \quad (6)$$

Nilai *precicion* pada metode evaluasi *confusion matrix* adalah tingkat akurasi antara data yang diminta dengan hasil prediksi, maka *precision* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. *Precision* dapat dihitung menggunakan persamaan dibawah ini :

$$precision = \frac{tp}{tp+fp} \quad (7)$$

3. Hasil dan Pembahasan

Hasil akhir pada Algoritma C4.5 akan membentuk dan menghasilkan pohon keputusan. Hasil tersbeut akan memberikan informasi terbaru Atribut pada yang akan dijadikan sebagai paramater untuk kriteria dalam pembentuk pohon keputusan, dimana beberapa atribut atau kriteria yang dieliminasi tidak diperlukan. Karena sampel hanya menguji berdasarkan kriteria atau kelas sehingga memudahkan dalam pengambil keputusan dan dapat dikendalikan dalam mengambil tindakan.Kemudian pada *Particle Swarm Optimization* PSO akan menyeleksi fitur yang menjadi atribut pilihan.

3.1. Perhitungan Manual Algoritma

Langkah awal dalam perhitungan Algoritma C4.5 ini dengan membagi data *training* dan *testing*, untuk menghitung entropy dari kelas yang ada, pada penelitian ini ada 2 kelas yang menjadi kelas yaitu Kadar Gula Normal dan Tinggi pada hubungan Aktifitas pasien. Setelah menghitung Entrophy akan dihitung nilai gain pada setiap atribut, dan nilai tertinggi pada gain akan menjadi akar. Dengan menggunakan persamaan 1 akan menghasilkan nilai entropy total sebagai berikut :

Entropy (total) = ((-31/42)* Log2 (31/42) +
(-16/42)*Log2(16/42)))
= 0,829607103

Kemudian untuk mengetahui nilai gain, akan dihitung terlebih dahulu nilai entrophy masing-masing pada setiap atribut. Berikut adalah nilai entrophy yang telah dihitung pada setiap atribut dan nilai gain.

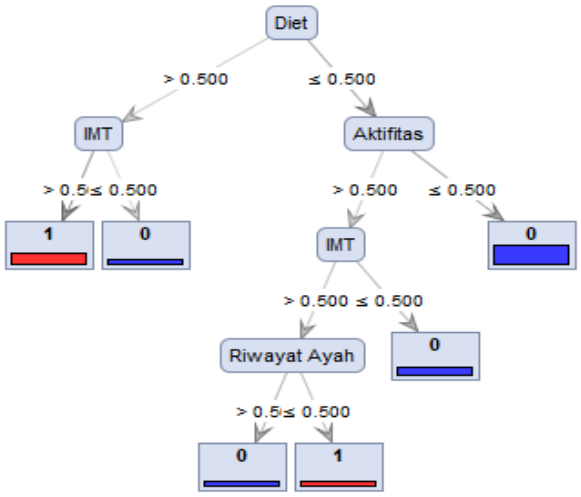
Tabel 4. Model Jumlah Entropy dan Gain

Nilai	Jumh Kasus	Tinggi (0)	Normal (1)	Entropy (E)	Gain
Total Kasus	42	31	11	0,829607103	
Jenis Kelamin					0,150086876
Laki-laki (1)	15	7	8	0,996791632	
Perempuan (2)	27	24	3	0,503258335	
Umur					0,02643746
> 45 Tahun (0)	33	23	10	0,884963636	
<= 45 Tahun (1)	9	8	1	0,503258335	
Pendidikan					0,001811335
SD/SDMP/ SMA (0)	25	18	7	0,855450811	
Perguruan Tinggi (1)	17	13	4	0,787126586	
K_IMT					0,296868169
Obesitas (0)	21	21	0	0	
Normal (1)	22	10	11	1,017047056	
Riwayat Ayah					0,094415108
Tidak (0)	34	23	11	0,908178347	
Ya (1)	8	8	0	0	
Riwayat Ibu					0,094415108
Tidak (0)	34	23	11	0,908178347	
Ya (1)	8	8	0	0	
Pengetahuan					0,247192875
Tidak Baik (0)	33	29	4	0,532835063	
Baik (1)	9	2	7	0,764204507	
Aktifitas					0,289764063
Tidak Baik (0)	25	24	1	0,532835063	
Baik (1)	11	2	9	0,764204507	
Diet					0,395724735
Tidak Baik (0)	31	29	2	0,34511731	
Baik (1)	11	2	9	0,68403844	
Konsumsi Obat					0,002117643
Tidak Baik (0)	21	15	6	0,86312057	
Baik (1)	21	16	5	0,79185835	

Pada tabel 4 dapat dilihat bahwa *entrophy* yang tertinggi akan digunakan pada perhitungan gain, setiap atribut dengan entrophy tertinggi yang dihitung akan menghasilkan masing-masing gain yang berbeda juga. Nilai gain tertinggi akan dijadikan node akar (*root node*), kemudian dilakukan perulangan perhitungan dan akan mendapatkan hasil nilai gain tertinggi.

3.2. Pohon Keputusan

Sesuai perhitungan manual dengan mencari nilai entropy dan gain tertinggi maka pengujian dengan meggunakan rappid minner didapatkan pohon keputusan seperti Gambar 2.



Gambar 2. Pohon Keputusan Perhitungan Algoritma C4.5

Gambar 2 menjelaskan bahwa pada atribut diet dengan nilai gain tertinggi menunjukkan kontruksi pohon yang merupakan pembentukan akar. Atribut yang terpilih kemudian dijadikan fungsi fitness untuk penerapan algortima pada PSO.

3.3. Evaluasi pada Model Confusion Matrix

Pengujian ini menggunakan evaluasi model *Confusion Matrix*. Model ini akan membentuk matrix dari *true positive* atau tupel *positive* dan *true negatif* atau tupel negatif.

Tabel 5. Evaluasi pada *Confusion Matrix*

Klasifikasi	*True* (Tinggi)	*True* (Normal)
Pred (Tinggi)	29	4
Pred (Normal)	2	7

Berdasarkan tabel 5 menghasilkan rincian *True Positive* (TP) 29, *False Negative* (FN) 2, *False Positive* (FP) 4, *True Negative* (TN) 7. Dari hasil tersebut maka dapat dihitung nilai *accuracy*, *sensitivity* (*recall*), specifity dan *precision*. Dapat dilihat pada tabel dibawah nilai *accuracy*, *sensitivity* (*recall*), specifity dan *precision* :

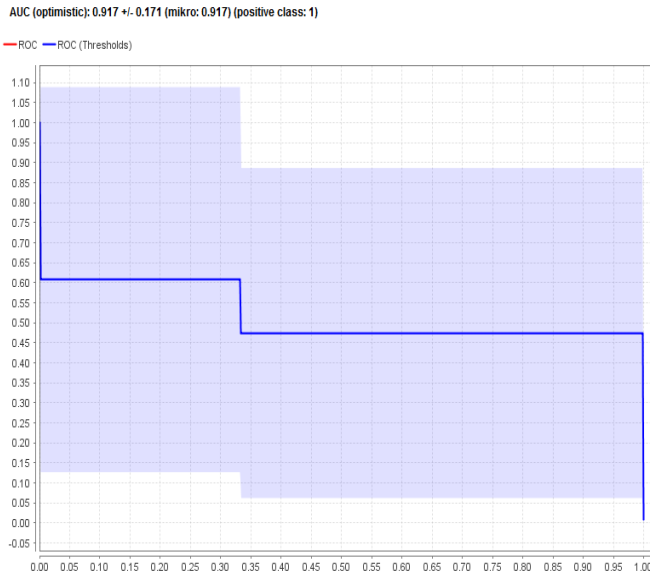
Tabel 6. Hasil *Confusion Matrix* pada Algoritma C4.

Accuracy	86%
Sensitivity (*Recall*)	88%
Specificity	78%
Precision	94%

Pada tabel 6 menjelaskan bahwa pada tingkat akurasi prediksi menggunakan algoritma C4.5 sebesar 86%, *recall* 88%, *specificity* 78% dan *precision* 94%. Dalam mengevaluasi performance pada *Machine Learning*, digunakan confusion matrix. Confusion Matrix mempresentasikan prediksi dan kondisi sebenarnya (aktual) dari data yang dihasilkan.

3.4. Evaluasi pada Kurva ROC

Pengujian dengan menggunakan evaluasi kurva ROC pada Algoritma C4.5 dapat dilihat pada kurva dibawah ini. Kurva grafik dibawah menunjukkan nilai AUC sebesar 0,917. Berikut Kurva ROC.



Gambar 3. Performa AUC pada Algoritma C4.5

3.5 Evaluasi dengan Model *Confusion Matrix* pada Algoritma C4.5 berbasis *Particle Swarm Optimization* (PSO)

Pengujian ini menggunakan evaluasi model dengan *Confusion Matrix*. Model ini akan membentuk matrix dari *true* positif atau tupel positif dan *true* negatif atau tupel negatif.

Tabel 7. Model Evaluasi *Confusion Matrix* pada Algoritma C4.5 berbasis *Particle Swarm Optimization* (PSO)

Klasifikasi	*True* (Tinggi)	*True* (Normal)
Pred (Tinggi)	31	2
Pred (Normal)	0	9

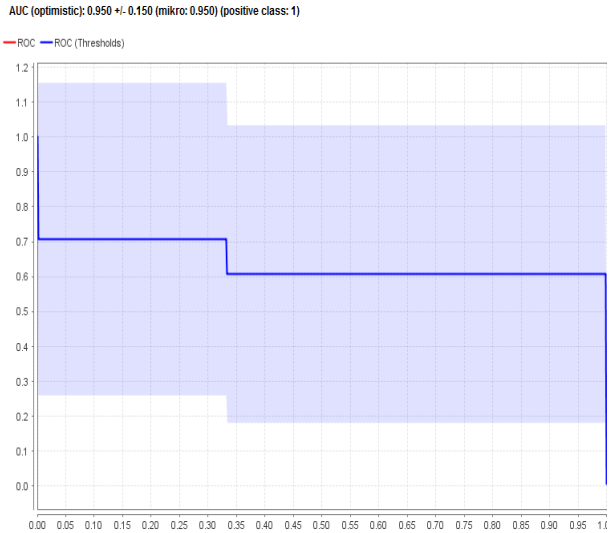
Berdasarkan tabel 7 hasil menggunakan data testing dengan rincian *True Positive (TP)* 29, *False Negative (FN)* 2, *False Positive (FP)* 4, *True Negative (TN)* 7. Dari hasil tersebut maka akan dapat dihitung nilai *accuracy, sensitivity (recall), specifity dan precision.* Dapat dilihat pada tabel dibawah nilai *accuracy, sensitivity (recall), specifity dan precision* :

Tabel 8. Hasil Confusion Matrix Algoritma C4.5 dan PSO

Accuracy	95%
Sensitivity (*Recall*)	94%
Specificity	100%
Precision	100%

3.6. Evaluasi Kurva ROC pada Algortitma C4.5 berbasis *Particle Swarm Optimization* (PSO)

Pengujian dengan Algoritma C4.5 menggunakan PSO dapat dilihat pada gambar 4. Kurva grafik pada gambar 4 menunjukkan nilai AUC sebesar 0,950. Nilai pada auc naik sebesar 0,033.



Gambar 4. Performa AUC pada Algoritma C4.5 *Particle Swarm Optimization* (PSO)

4. Kesimpulan

Dari penelitian yang telah dilakukan menggunakan algoritma C4.5 dengan Particle Swarm Optimazation (PSO) pada dataset Pasien dengan Pengaruh Aktifitas Fisik Terhadap Kadar Gula Darah di RSUD H.Abdul Manan Simatupang Kisaran ini menghasilkan nilai akurasi yang berbeda. Nilai akurasi pada pengujian yang dilakukan dengan algoritma C4.5 sebesar 86%, sedangkan nilai akurasi pada algoritma C4.5 menggunakan PSO sebesar 95%, sehingga dapat disimpulkan bahwa penggunaan PSO dapat meningkatkan nilai akurasi. Sedangkan evaluasi pada Kurva ROC menunjukan selisih 0,033.

Dari 10 atribut pada dataset, yang diseleksi menggunakan PSO menjadi 7 atribut yang digunakan dalam menentukan prediksi Kadar Gula Darah. Atribut yang digunakan tersebut adalah : Umur, Pendidikan, IMT, Riwayat Ayah, Riwayat Ibu, Diet. Dapat disimpulkan bahwa penggunaann algoritma *Particle Swarm Optimization* (PSO) mampu menyeleksi atribut pada C4.5, sehingga menghasilkan tingkat akurasi yang lebih tinggi.

Daftar Rujukan

[1] Garnita, Dita., 2012. Faktor Risiko Diabetes Melitus di Indonesia (Analisis Data Sakerti.2007), Depok : FKM UI.
[2] Arisman., 2011. Obesitas, Diabetes Melitus, dan Dislipidemia. Jakarta: EGC.
[3] Krisnatuti & Yehrina., 2008. Diet Sehat untuk Penderita Diabetes Mellitus. Jakarta: Penebar Swadaya.
[4] Tandra., 2009. Segala Sesuatu Yang Harus Anda Ketahui Tentang Diabetes. Jakarta: Kompas Gramedia.
[5] Anani, S., Udiyono, A., Ginanjar, P., 2012. Hubungan antara Perilaku Pengendalian Diabetes dan Kadar Gula Darah Pasien Rawat Jalan Diabetes Melitus (Studi Kasus di RSUD Arjawinangun Kabupaten Cirebon). Jurnal Kesehatan Masyarakat, vol.1, pp.466-478.
[6] Lakshmi, B.N., Raghunandhan, G.H., 2011. A conceptual overview of data mining. Proceedings of the National Conference on Innovations in Emerging Technology, pp. 27-32.
[7] Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. Informatics Med Unlocked [Internet]. 2018; 10 (Februari 2020) : 100–7. Available from: https://doi.org/10.1016/j.imu.2017.12.006
[8] Sisodia, DS. Prediction of Diabetes using Classification Algorithms. Procedia Comput Sci [Internet]. 2018. 132 (Iccids) : 1578–85. Available from: https://doi.org/10.1016/j.procs. 2018.05.122
[9] Puspita Ari., 2016. Prediksi Kelahiran Bayi Secara Prematur dengan Menggunakan Algoritma C4.5 Berbasis *Particle Swarm Optimization*. Jurnal Teknik Informatika STMIK Antar Bangsa.Vol. II, pp.11-16.
[10] Tejas Mehta, Dhaval Kathiriya., 2016. Performance Analysis of Data Mining Classification Techniques, International Journal of Innovative Research in Science, Engineering and Technology, ISSN : 2319-8753. Vol. 5, Issue 3,pp.3116-3122.
[11] .Meng-Chang Tsai, Kun-Huang Chen, Chao-Ton Su, and Hung-Chun Lin. 2012 "An Application of PSO Algorithm and Decision Tree for Medical Problem," 2nd International Conference on Intelligent Computational Systems, pp. 124-126.
[12] Kotsiantis, S. B., 2007. "Supervised Machine Learning: A Review of Classification Techniques," Department of Computer Science and Technology, pp. 249-268.
[13] Daniel T.Larose., 2005. Discovering in Data Mining, An Introduction to Data Mining. Willey Interscience.
[14] Rusda Wajhillah., 2014. Optimasi Algoritma Klasifikasi C4.5 berbasis *Particle Swarm Optimization* untuk Prediksi Penyakit Jantung. SWABUMI, Vol.1.
[15] Sunjana., 2010. Klasifikasi Data Nasabah Sebuah Asuransi Menggunakan Algoritma C4.5, Seminar Nasional Aplikasi Teknologi Informasi, pp. D31-D34.
[16] A. Abraham, C.Grosan and V.Ramos., 2006. Swarm Intelligence in Data Mining. Verlag Berlin Heidelberg: Springer.
[17] Gorunescu, F., 2011. Data Mining Concepts, Models and Techniques. Verlag Berlin Heidelrbeg : Springer.