Jurnal E-KOMTEK (Elektro-Komputer-Teknik)
Vol. 4 No. 1 (2020) pp. 30-47
https://jurnal.politeknik-kebumen.ac.id/index.php/E-KOMTEK
p-ISSN : 2580-3719 e-ISSN : 2622-3066

Evaluasi Performa Algoritma C4.5 dan C4.5 Berbasis PSO untuk Memprediksi Penyakit Diabetes

Nurahman
Sistem Informasi, Universitas Darwan Ali, Kota Waringin Timur, Indonesia, 13210

E-mail : nurrahman.ikhtiar@gmail.com
Doi : https://doi.org/10.37339/e-komtek.v4i1.230

Diterbitkan oleh Politeknik Dharma Patria Kebumen

Info Artikel
Diterima :
29-05-2020
Diperbaiki :
15-06-2020
Disetujui :
17-06-2020

ABSTRAK

Tubuh manusia terdiri dari berbagai organ yang setiap saat diperlukan oleh manusia untuk beraktivitas. Aktivitas manusia dapat dilakukan jika kesehatan tubuh dalam keadaan baik. Salah satu penyakit yang berakibat komplikasi bahkan berujung kematian adalah diabetes. Penderita penyakit diabetes dari tahun ke tahun meningkat. Hal ini disampaikan oleh artikel Atlas Diabetes yang diterbitkan pada edisi ke-7 tahun 2015 dari IDF menyebutkan ditahun 2015 penderita penyakit diabetes akan mencapai 415 juta pasien dari 220 negara kemudian penderita diabetes akan meningkat menjadi 642 juta pasien di tahun 2040. Bidang keilmuan data mining ikut melakukan riset. Data mining salah satu bidang ilmu yang melakukan pengolahan terhadap data untuk mengetahui pengetahuan baru terhadap suatu kasus. Paper ini dilakukan suatu pemodelan algoritma klasifikasi data mining. Penerapan pemodelan dilakukan dengan menggunakan algoritma C4.5 dan C4.5 berbasis PSO. Penerapan pemodelan akan dilakukan peninjauan berdasarkan nilai performa akurasi dan AUC. Setelah dilakukan peninjauan terhadap kedua pemodelan tersebut, diperoleh hasil bahwa C4.5 berbasis PSO memiliki performa terbaik sehingga masuk pada kategori good classification.
Kata Kunci: Data mining; Algoritma C4.5; Particle Swarm Optimization

ABSTRACT

The human body is made up of various organs which are needed by humans for activities. Human activities can be done if the body's health is in good condition. One disease that results in complications and even death is diabetes. Diabetics increase from year to year. This was conveyed by the Atlas Diabetes article published in the 7th edition of 2015 from IDF mentioning that in 2015 diabetics will reach 415 million patients from 220 countries then diabetics will increase to 642 million patients in 2040. Data mining scientific fields participate do research. Data mining is one of the fields of science that conducts data processing to find out new knowledge in a case. This paper is carried out a modeling of data mining classification algorithms. Application of modeling is done by using the C4.5 and C4.5 algorithm based on PSO. The modeling application will be reviewed based on the accuracy and AUC performance values. After reviewing the two models, the results show that PS4 based C4.5 has the best performance so that it falls into the category of good classification.
Keywords: *Data mining; Algoritma C4.5; Particle Swarm Optimization*

Alamat Korespondensi : Jl. Letnan Jenderal Suprapto No.73 Kebumen, Jawa Tengah, Indonesia 55431

1. PENDAHULUAN

Organ tubuh manusia sangatlah kompleks dan masing-masing dari organ tubuh memiliki peran saling membutuhkan. Jika satu dari organ tersebut mengalami masalah maka akan mempengaruhi organ-organ yang lain. Bahkan setiap manusia mempercayai bahwa kehidupan akan selalu berjalan dengan baik ketika memiliki badan yang sehat dan terhindar dari segala penyakit. Berbagai jenis penyakit akan selalu menyerang manusia kapanpun dan dimanapun. Bahkan beberapa penyakit juga menyebabkan kematian dikarenakan menurunnya fungsi organ tubuh pada manusia. Penyakit yang terdengar menakutkan karena dapat membawa manusia berujung pada kematian diantaranya penyakit jantung, paru-paru, ginjal, stroke, diabetes dan juga masih terdapat penyakit lainnya.

Penyakit diabetes mellitus merupakan sebuah penyakit metabolisme yang sering kali ditandai adanya peningkatan terhadap kadar gula darah atau disebut glukosa seseorang melebihi batas normal (*hyperglycemia*) kebutuhan tubuh **[1]**. Penyakit ini tergolong dengan penyakit yang menakutkan dengan penderita yang sangat banyak di berbagai belahan dunia internasional. Artikel internasional terbitan edisi ke-7 pada Atlas diabetes terdapat penjelasan IDF mengenai perkembangan penderita diabataes. Berdasarkan data-data yang dihimpun IDF tahun 2015 penderita penyakit diabetes mencapai 415 juta orang dari 220 negara di dunia internasional. Bahkan pada artikel tersebut juga menjelaskan bahwa penderita penyakit diabetes akan mengalami lonjakan mencapai 642 juta penderita diabetes ditahun 2040 **[2]**. Diketaui pula bahwa penyebab kematian terbesar didunia internasional yaitu jantung dan pembuluh darah (kardiovaskular) yang 50% diantaranya berkaitan langsung dengan penyakit diabetes. Sedangkan dalam paper lain juga menjelaskan mengenai penyakit diabetes dibedakan dalam beberapa kategori. Perkembangan penyakit diabetes di Indonesia penderita terabanyak terdapat pada usia antara 55 sampai 64 tahun. Bahkan penyakit tersebut juga masih menyerang manusia di usia 65 sampai 74 tahun. Kemudian dari kategori jenis kelamin diketahui penderita diabetes jenis kelamin perempuan adalah 1,8% daripada laki-laki 1,2%. Selanjutnya dilihat dari kategori domisili diketahui bahwa penderita diabetes melitus lebih banyak pada daerah perkotaan, diperkirakan lingkungan kota mencaapai 1,9% dibandingkan perdesaan yaitu 1,0% **[3]**.

2. MATERIAL DAN METODE

2.1 Material

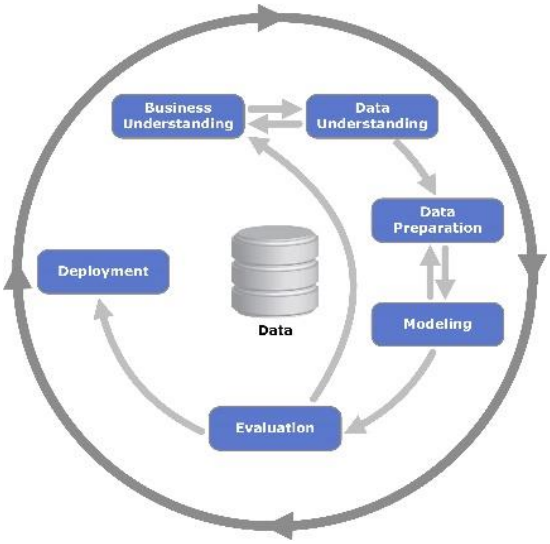
Perkembangan mengenai penyakit diabetes yang semakin meningkat banyak para ilmuan melakukan berbagai pengembangan pengetahuan terutama bidang kesehatan. Selain itu juga tidak ketinggalan pula para pakar ilman bidang teknologi informasi. Seperti yang telah kita ketahui bahwa banyaknya data mengenai kasus diabetes membuat para ahli teknologi berusaha memanfaatkan data yang ada. Sekumpulan data akan menjadi sampah yang tidak bermanfaat didalam database jika data tersebut tidak diolah dengan baik. Ilmu yang mempelajari tantang data saat ini dikenal dengan nama penambangan data (*data mining*). Penambangan data (*data mining*) merupakan ilmu tentang pembelajaran model sekumpulan data untuk menemukan pengetahuan yang sebelumnya tidak diketahui. *Data mining* juga dapat diartikan sebagai serangkaian proses untuk menemukan pola dari dataset baik dilakukan secara otomatis maupun semi otomatis dengan menghasilkan pola yang dapat bermanfaat **[4]**. Penerapan *data mining* memiliki tugas yang berbeda-beda, sehingga penerapannya data mining dibagi menjadi beberapa kolompok, yaitu: *Classification, Prediction, Clustering, Description, Estimation,* dan *Association*.

Penelitian mengenai *data mining* banyak diminati oleh para pakar ilmu teknologi dan informasi. Penelitian tersebut dilakukan dengan memanfaatkan sekumpulan data untuk dapat diketahui pola maupun pengetahuan baru dari data tersebut. Data yang semula tertumpuk pada sebuah database mencapai ratusan, ribuan bahkan milyaran data maka akan disebut sebagai sampah database. Hal ini tentunya akan membuat komputer menumpuk banyaknya sampah. Perkembangan pengetahuan mengenai data mining maka akan membuat data-data tersebut menjadi bermanfaat. Salah satu kelompok data mining yang diminati oleh para peneliti yaitu Classification. Metode Classification diterapkan dalam berbagai bidang aktivitas kehidupan manusia salah satunya mengklasifikasi berbagai kesehatan dan penyakit. Penelitian terkait klasifikasi terhadap penyakit diabetes diantaranya **[5]**, **[6]**, dan **[7]**. Penelitian tentang Algoritma decision tree C4.5 **[8]**, **[9]**, **[10]**, **[11]**, dan **[12]**. Pada penelitian **[13]** menilai hasil dari evaluasi kinerja beberapa algoritma klasifikasi Decision Tree, dari hasil tersebut menunjukan *performa decision tree* C4.5 memiliki performa yang terbaik dibandingkan dengan algoritma *decision tree* yang lainnya.

Peneliti-peneliti sebelumnya menjelaskan tentang teknik-teknik pada metode klasifikasi. Klasifikasi menggunakan salah satu Algoritma data mining seperti C4.5 dapat diterapkan dengan berbagai model atau pola yang berbeda-beda. Pada penelitian ini melihat kasus mengenai penyakit diabetes yang telah di uraikan semakin meningkat. Kemudian penelitian-penelitian sebelumnya dapat dijadian sebagai acuan maupun referensi. Dari pertimbangan kasus penderita diabetes dan artikel-artikel paper peneliti sebelumnya maka paper ini akan membahas mengenai pola dari suatu algoritma dengan melihat nilai performa. Performa Algoritma untuk mengklasifikasi data dapat dijadikan sebagai nilai keakuratan terhadap suatu keputusan. Semakin tinggi nilai performa klasifikasi algoritma terhadap suatu data maka akan sebakin baik untuk menentukan suatu keputusan. Utuk itu, penelitian ini juga akan melakukan penerapan model maupun pola pada algortima C4.5 hingga dapat menunjukkan nilai performa yang baik. Penerapan pola tidak hanya dilakukan pada algoritma C4.5 saja, melainkan C4.5 akan dikolaborasi dengan *Particle Swarm Optimization* (PSO). Sehingga dikenal dengan Algoritma C4.5 berbasis *Particle Swarm Optimization*. Hasil penelitian nantinya diharapkan dapat memberikan pengetahuan model dari metode klasifikasi data mining dan menjadi rekomendasi dalam melakukan analisis terhadap penyakit diabetes.

2.2 Metode

Metodologi yang digunakan dalam penelitian ini yaitu Metodologi *Cross Industry standard Process For Data Mining* atau disebut CRISP-DM. Metodologi CRISP-DM memiliki 6 tahapan dalam menyelesaikan kasus penelitian, seperti yang disajikan pada **Gambar 1 [14] [15]**.



Gambar 1. Tahapan metodologi CRIPS-DM

a. *Business Understanding Phase* (Pemahaman Bisnis)

Tahapan ini yang dilakukan peneliti: (1) menentukan tema dan berbagai kebutuhan yang tepat pada penelitian secara keseluruhan, (2) menentukan tujuan, batasan dan formula dalam suatu penelitian, dan (3) menetukan strategi-strategi penelitian untuk mencapai suatu tujuan penelitian.

b. *Data Understanding Phase* (Pemahaman Data)

Tahapan ini yang dilakukan peneliti: (1) mencari dan mengumpulkan data penelitian, (2) mendeskripsikan data (mengetahui dan menjelaskan mengenai data, atribut, kelas dan tipe data)

c. *Data Preparation Phase* (Pengolahan Data) atau *Fase Pre-Processing Data*

Tahapan ini yang dilakukan peneliti: (1) melakukan identifikasi terhadap data, (2) memilih kasus dan variabel untuk dilakukan analisis, (3) menetukan format data sesuai kebutuhan software analisis, dan (4) mencari dan pembersihan terhadap data mising

d. *Modelling Phase* (Permodelan)

Tahapan ini yang dilakukan peneliti: (1) melakukan pemilihan terhadap pemodelan yang akan diterapkan pada penelitian, dan (2) melakukan langkah-langkah untuk mengoptimalkan hasil.

e. *Evaluation Phase* (Evaluasi)

Tahapan ini yang dilakukan peneliti: (1) melakukan evaluasi terhadapat model yang diterapkan pada algoritma, (2) mengaplikasikan model yang sudah disiapkan pada fase pemodelan, (3) menentukan nilai performa apa yang diukur untuk dibahas, dan (4) Memutuskan model yang digunakan yaitu hasil dari proses data mining.

f. *Deployment Phase* (Penyebaran)

Tahapan terakhir dalam CRIPS-DM setelah model dievaluasi dan dipilih yaitu tahapan penyebaran. Tahapan penyeberan yang paling sederhana adalah pembuatan laporan. Model yang memiliki nilai performa terbaik dapat dibahas dan direkomendasikan untuk digunakan dalam pengambilan keputusan.

3. HASIL DAN PEMBAHASAN

3.1 *Business Understanding Phase*

Penelitian ini fokus pada bidang data mining dengan membahas topik mengenai penyakit diabetes mellitus. Pada paper ini ingin mengidentifikasi mengenai apakah seseorang terdiagnosis penyakit *diabetes* atau tidak dengan penerapkan algoritma pada data mining. Setiap algoritma diaplikasikan dengan model kemudian ditinjau nilai performa terbaik untuk dapat direkomendasikan dalam mendiagnosis penyakit *diabetes*.

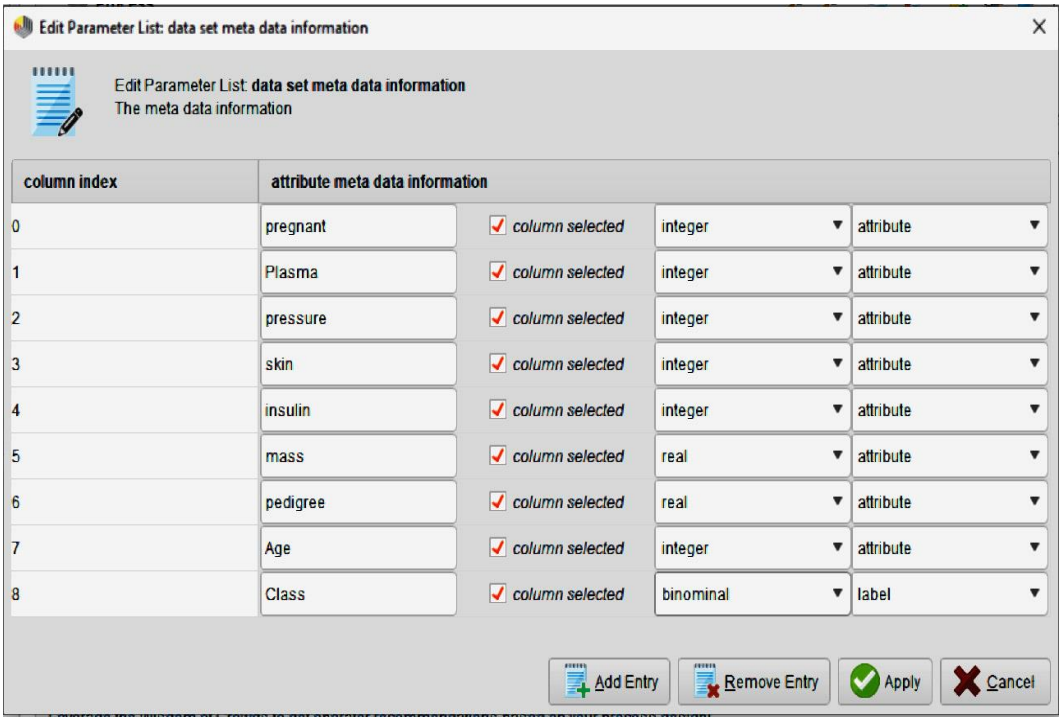
3.2 *Data Understanding Phase*

Dataset yang digunakan memiliki jumlah sebanyak 768 data pengamatan terhadap pasien. Dengan 8 atribut dan 1 class yang berisikan data numerik dengan beberapa type data. Kemudian Parameter indikator yang akan diuji pada dataset *Diabetes Milletus* yang diambil adalah :

a. *Number of times pregnant* adalah berapa kali hamil dengan *Type data Integer*
b. *Plasma glucose concentration 2 hours in an oral glucose tolerance test* adalah kadar gula plasma 2 jam hasil tes toleransi glukosa oral dengan *Type data Integer*
c. *Diastolic blood pressure (mm Hg)* adalah tekanan darah diastolik (mm Hg) dengan *Type data Integer*
d. *Triceps skin fold thickness (mm)* adalah pengukuran lemak di dalam ubuh atau tebalnya lipatan dari kulit (mm) dengan *Type data Integer*
e. *2-Hour serum insulin (mu U/ml)* adalah insulin serum 2-Jam (mu U / ml) dengan *Type data Integer*
f. *Body mass index (weight in kg/(height in m)^2)* adalah ketepatan antara berat badan (kg) pasien dengan tinggi badan (m^2) pasien, *Type data* yang digunakan adalah *Real*
g. *Diabetes pedigree function* adalah riwayat penyakit diabetes pada diri sendiri maupun keluarga terdekat (yang memiliki hubungan ginetik dengan pasien) dengan *Type data Real*
h. *Age (years)* adalah umur (tahun) dengan *Type data Integer*
i. *Class* adalah sebagai kolom label dengan *Type data Binominal*

Sebagaimana tujuan dari pemrosesan data mining yaitu agar pengetahuan yang diperoleh membantu para pemangku kebijakan. Untuk itu, penentuan tipe data perlu dilakukan dengan tepat sehingga metode dan kondisi data dapat menghasilkan informasi yang lebih akurat. Penentuan tipe data dalam kasus ini dilakukan dengan Software Rapid Miner

Studio 9.6. Tahapan dalam menentukan tipe data pada kasus penelitian ini disajikan pada **Gambar 2**.



Gambar 2. Penentuan Type Data pada Dataset Diabetes

3.3 *Data Preparation Phase*

Paper ini membahas pengolahan data mengenai jenis Klasifikasi data. Pendekatan yang dilakukan adalah bersifat kuantitatif, yaitu dengan memahami cara kerja Algoritma-algoritma Data Mining yang akan digunakan dalam mengelola data gejala-gejala penyakit Diabetes *Mellitus*. Hasil diagnosa akan membantu para bidang kesehatan untuk memutuskan terhadap penderita penyakit *Diabetes Mellitus*.

Dataset *Diabetes* memiliki jumlah *instance* sebanyak 768, dengan *type data* yang sudah ditentukan. Pada dataset awal nama parameter yang akan diuji masih berupa sebuah kalimat. Kemudian Langkah untuk memudahkan penelitian dalam mendeskripsikan atribut-atribut yang ada pada dataset maka dilakukan pemberian nama kolom pada dataset tersebut. Setiap Atribut akan diberi kode/nama kolom pengganti agar penulisan kolom lebih mudah. Pembuatan kolom ditentukan berdasarkan nama parameter indikator yang akan diuji. Pengkodean antribut dataset di penelitian ini disajikan pada **Tabel 1**.

Tabel 1. Penamaan/kode pada atribut

No	Indikator	Keterangan	Kolom
1	*Number of times pregnant*	Berapa kali hamil	*Pregnan*
2	*Plasma glucose concentration a 2 hours in an oral glucose tolerance test*	Kadar gula plasma 2 jam dalam tes toleransi glukosa oral	*Plasma*
3	*Diastolic blood pressure (mm Hg)*	Tekanan darah pada arteri saat kondisi jantung rileks atau disebut diastolik (mm Hg)	*Pressure*
4	*Triceps skin fold thickness (mm)*	Pengukuran lemak di dalam tubuh atau disebut tebalnya lipatan pada kulit (mm)	*Skin*
5	*2-Hour serum insulin (mu U/ml)*	Insulin serum 2-Jam (mu U / ml)	*Insulin*
6	*Body mass index (weight in kg/(height in m)^2)*	Keseimbangan antara berat badan dengan satuan kg dan tinggi badan (m^2)	*Mass*
7	*Diabetes pedigree function*	Mengetahui riwayat penyakit diabetes yang ada pada keluarga dalam satu genetic	*Pedigree*
8	*Age (years)*	Usia dari pasien (tahun)	*Age*
9	*Class*	Kelas / Label	*Class*

Seperti yang telah diketahui sebelumnya bahwa dataset pada penelitian ini adalah numerik. Untuk itu, table dibawah ini memperlihatkan mengenai range nilai pada setiap atribut-atribut dari Dataset penyakit *diabetes*. Range nilai data numerik dari dataset disajikan pada **Tabel 2**.

Tabel 2. Range nilai yang terdapat pada atribut dataset

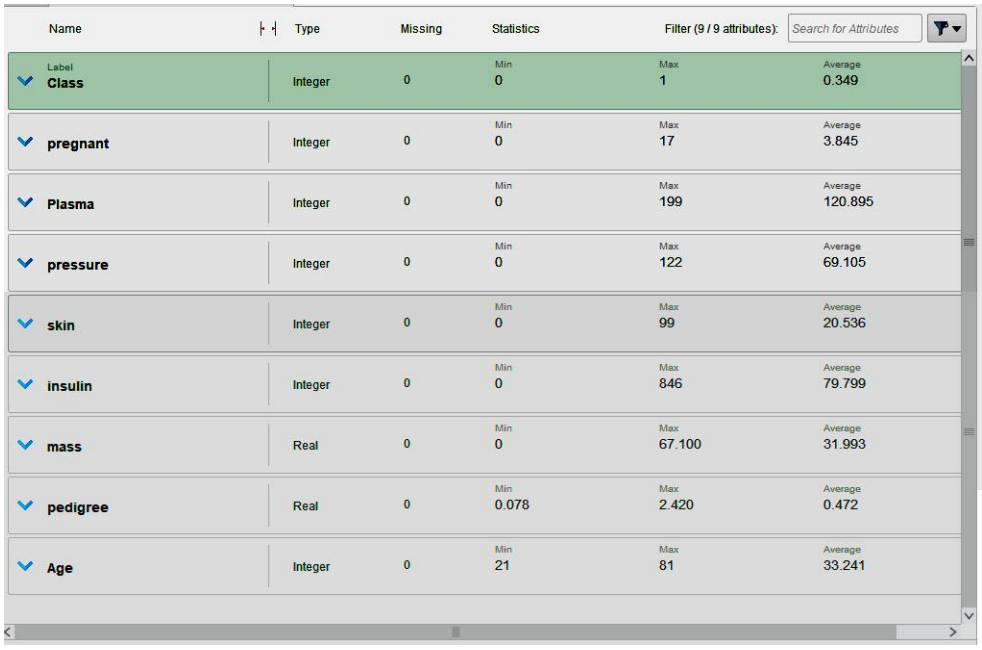
Atribut	Nilai
Pregnan	0 – 17
Plasma	0 – 199
Pressure	0 – 122
Skin	0 – 99
Insulin	0 – 846
Mass	0 – 67.1
Predigree	0.078 – 2.420
Age	21 – 81
Class	*Tested_Positif, Tested_Negative*

Nilai dari setiap atribut akan berperan dalam menentukan hasil diagnosis yang dilakukan dalam penelitian ini. Nilai atribut dari Tabel 2 juga dapat dijelaskan sebagai berikut:

a. atribut *pregnan* yaitu atribut untuk mengetahui berapa kali kehamilan penderita diabetes, pada *Dataset Pima Indians Diabetes* memiliki jumlah kehamilan antara 0 sampai 17 kali kehamilan.
b. Atribut *plasma* yaitu antribut untuk mengetahui kadar gula plasma pada pasiean hasil dari test toleransi glukosa oral 2 jam setelah makan. Pada *Dataset Pima Indians Diabetes* diketahui memiliki hasil nilai antara 0 sampai 199 Mg/dL.
c. Atribut *Pressure* yaitu atribut Tekanan darah *diastolik* (mm Hg) yang dimiliki pasien. Pada dataset yang digunakan diketahui tekanan darah *diastolik* pasien antara 0 sampai 122 mm Hg.
d. Atribut *Skin* yaitu atribut mengenai ketebalan lipatan kulit atau lemak didalam tubuh pasien. Ketebalan lipatan kulit pasien pada Dataset yang digunakan berkisar antara 0 sampai 99 mm.
e. Atribut *Insulin* yaitu atribut menunjukkan nilai hormon yang dihasilkan oleh pankreas untuk mengolah gula dalam tubuh atau disebut insulin pasien. *Insulin* diketahui dari *dataset* yang digunakan berkisar antara 0 sampai 846 mu U/ml
f. Atribut *Mass* yaitu atribut mengenai berat badan pasien. Pada dataset yang digunakan berat badan penderita berkisar antara 0 sampai 67.1 kg/m2.
g. Atribut *Predigree* yaitu atribut mengenai riwayat *diabetes* yang ada didalam keluarga pasien. Pada dataset yang digunakan Riwayat *diabetes* dalam keluarga pasien berkisar antara 0.078 sampai 2.420.
h. Atribut *Age* yaitu atribut umur penderita penyakit *diabetes*. Pada dataset yang digunakan umur penderita penyakit *diabetes* antara 21 sampai 81 tahun.
i. Atribut *Class* digunakan untuk menentukan *class positive* dan *class negative* pada dataset yang digunakan dipenelitian ini.

Langkah berikutnya yang dilakukan yaitu mengamati dataset secara teliti. Hal ini dilakukan untuk mengetahui apakah dataset sudah dalam kondisi valid atau masih terdapat *missing*. Jika ditemukan *missing* data maka yang dilakukan adalah melakukan cleaning terhadapat data yang missing tersebut. *Cleaning* data merupakan bagian dari tahapan

preprocessing. Langkah *prepocessing* terhadap data set sangat penting dilakukan agar analisis dapat menghasilkan performa yang lebih baik. Untuk memudahkan mengecakkan terhadap *missing* data maka dalam penelitian ini dibantu dengan menggunakan *software Rapid Miner Studio Versi 9.6* dengan *type* file dataset csv. Hasil dari *cleaning* data disajikan pada **Gambar 3**.



Name	Type	Missing	Statistics			Filter (9 / 9 attributes): Search for Attributes
Label **Class**	Integer	0	Min 0	Max 1	Average 0.349	
pregnant	Integer	0	Min 0	Max 17	Average 3.845	
Plasma	Integer	0	Min 0	Max 199	Average 120.895	
pressure	Integer	0	Min 0	Max 122	Average 69.105	
skin	Integer	0	Min 0	Max 99	Average 20.536	
insulin	Integer	0	Min 0	Max 846	Average 79.799	
mass	Real	0	Min 0	Max 67.100	Average 31.993	
pedigree	Real	0	Min 0.078	Max 2.420	Average 0.472	
Age	Integer	0	Min 21	Max 81	Average 33.241	

Gambar 3. Tidak ada data missing pada dataset

3.4 *Modelling Phase*

Tahapan *Business Understanding* tadi dijelaskan bahwa penelitian fokus dalam penerapan model untuk identifikasi terhadap pasien *terdiagnosis* penyakit *diabetes* atau tidak dengan menerapkan algoritma data mining. Peninjauan performa algoritma dilakukan dengan membandingkan nilai dari Akurasi dan nilai *AUC*. Berdasarkan *type data* yang telah disampaikan pada tahapan *Data Understanding* diketahui memiliki *type data integer* dan *real*, sedangkan pada kolom class adalah *binominal*. Karena fitur setiap dataset memiliki data bersifat numerik yaitu dengan *type data integer* dan *real* maka peran data ini dapat dilakukan dengan clasifikasi menggunakan aloritma *C4.5.* Selain itu, paper ini juga membahas tentang penerapan metode klasifikasi dengan model algoritma C4.5 *berbasis* PSO. Dari kedua model tersebut akan ditinjau mengenai performa yang dihasilkan yaitu dengan memperhatikan akurasi dan *AUC.* Berikut adalah pemodelan yang akan diterapkan dalam penelitian ini.

a. Model Algoritma C4.5

Algoritma *C4.5* ini merupakan algoritma berbasis pohon keputusan. Algoritma ini menggunakan nilai dari Gain Ratio sebagai node akar. Untuk mendapatkan nilai Gain Ratio memerlukan beberapa tahapan diantaranya yaitu *entropy, information gain, split info*, dan *gain ratio*. Sehingga rumus tahapan dari algoritma *C4.5* dapat diterapkan sebagai berikut.

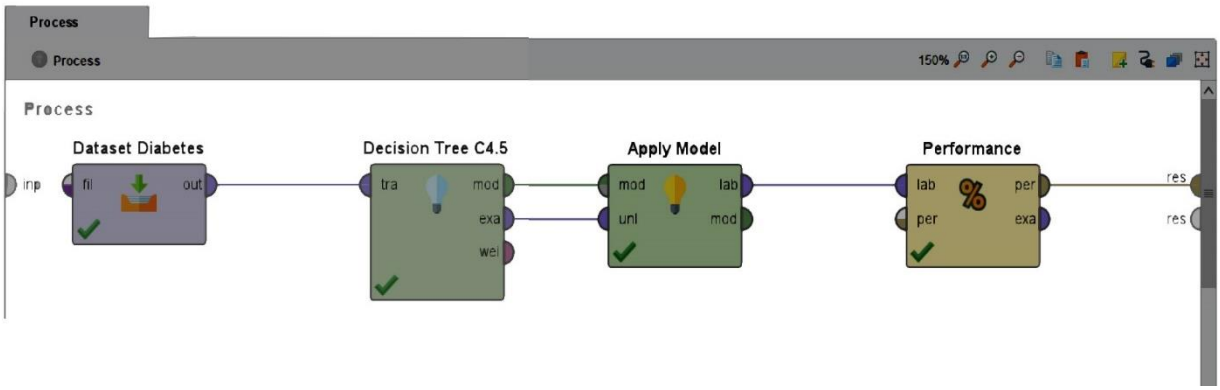
$$(1) Entropy = \sum_{i=i}^{n} -p_i * log_2 p_i$$

$$(2)\ Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} -\frac{|S_i|}{S} * Entropy(S_i)$$

$$(3)\ SplitInfo_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * Log_2\left(\frac{|D_j|}{|D|}\right)$$

$$(4)\ GainRatio = \frac{Gain(S,A)}{SplitInfo_A(D)}$$

Hasil persamaan di atas dilakukan pemodelan dengan algoritma *C4.5* dengan menggunakan *Rapid Miner Studio 9.6.* Hasil algoritma C4.5 disajikan pada **Gambar 4**.



Gambar 4. *Model Algoritma C4.5*

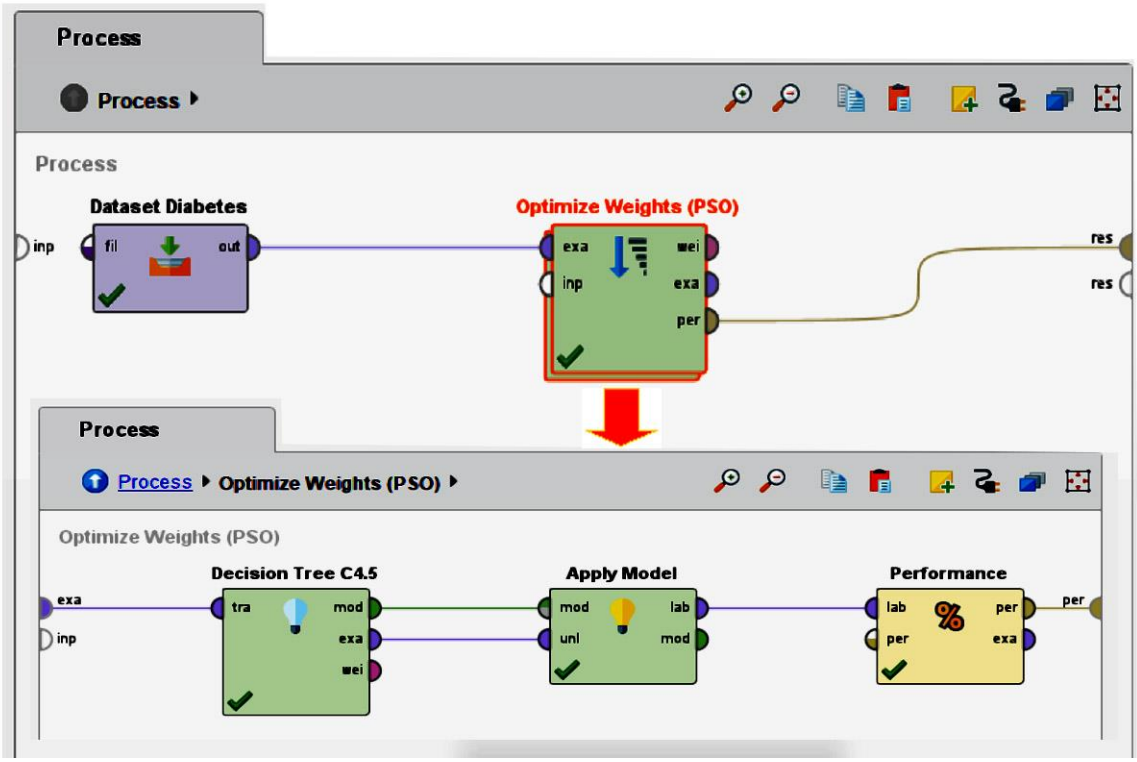
b. Model Algoritma C4.5 Berbasis Particle Swarm Optimizatio

Algoritma *Particle Swarm Optimization(PSO)* ini merupakan algoritma yang berbasis populasi dengan mengesplotasikan individu kedalam pencarian. Populasi dalam *PSO* disebut *swarm* sedangkan individu disebut sebagai *particle*. Setiap *particle* akan berpindah dengan kecepatan yang diadaptasi dari daerah pencarian dan menyimpan sebagai posisi terbaik yang pernah dicapai. Selain itu *particle* dalam *PSO* juga selalu dikaitkan dengan kecepatan *particle* terbang melalui ruang pencarian yang memiliki kecepatan dinamis berdasarkan perilaku historis. Sehingga partikel dalam algoritma *PSO* akan selalu cenderung terbang menuju pencarian yang lebih baik saat proses pencarian. Persamaan dari Algoritma *Particle Swarm Optimization (PSO)* adalah sebagai berikut.

$(1) V_i(t) = V_i(t-1) + c_1 r_1 [X_{pbest\ i} - X_i(t)] + c_2 r_2 [X_{Gbest} - X_i(t)]$

$(2)\ X_i(t) = X_i(t-1) + V_i(t)$

Hasil dari persamaan Algoritma *PSO* diatas dilakukan pemodelan dengan Algoritma *C4.5* sehingga pemodelan dikenal dengan nama pemodelan Algoritma *C4.5 berbasis PSO*. Pemodelan *C4.5 berbasis PSO* disajikan pada **Gambar 5**.



Gambar 5. *Model C4.5 Berbasis PSO*

3.5 *Evaluation Phase*

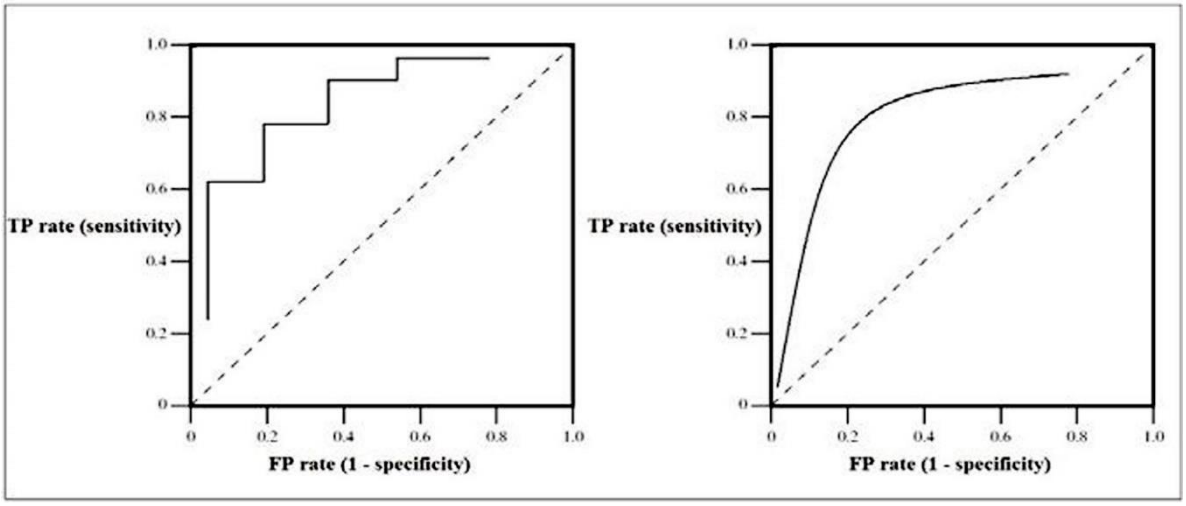
Tahapan evaluasi dilakukan pengukuran terhadap nilai performa akurasi. Perhitungan Nilai akurasi akan terbentuk berdasarkan confusion matrix. Dimana confusion matrix ini memperhatikan nilai *TP (True Positive)*, *TN (True Negative)*, *FP (False Positive)*, dan *FN (False Negative)*. *Confusion matrix* penelitian secara umum dapat disajikan pada **Tabel 3**.

Tabel 3. Contoh confusion matrix

	Act. True	Act. False
Pred. True	TP	FP
Pred. False	FN	TN

Selain memperhatikan nilai dari akurasi, pada penelitian ini juga meninjau performa nilai *AUC* yang dihasilkan dari kurva *ROC*. Kurva *ROC* adalah suatu grafik yang terdiri dari

garis vertical dan garis horizontal. Kurva *ROC* memiliki 2 dimensi yaitu dimensi *false positive* terdapat pada garis horizontal yaitu sumbu X sedangkan dimensi *true positive* terdapat pada garis *vertical* yaitu sumbu Y **[16]**. Dalam implementasi *True Positive* dikenal sebagai *True Positive Rate* atau *TP Rate* sedangkan *false positive* dikenal sebagai *false positive rate* atau *TP Rate.* Selain itu pada grafik *ROC* juga menggambarkan sebuah *trade off* antara manfaat dan biaya. Misalnya manfaat terdapat pada sumbu Y (*true positive*) sedangkan biaya terdapat pada sumbu X (*false positive*). Tampilan dua jenis kurva *ROC* (*discrete dan continuous*) disjaikan pada **Gambar 6.**



Gambar 6. Kurva *ROC*

Kurva *ROC* menunjukkan nilai *true positive* dan nilai *false positive.* Titik 0,0 terdapat pada pojok kiri bawah yaitu antara nilai TP dan nilai FP. Kemudian titik 1,1 menunjukan bahwa terdapat nilai klasifikasi positif sedangkan titik 0,1 merupakan klasifikasi sempurna. Klasifikasi sempurna yang terjadi pada kurva *ROC* karena tidak adanya FN dan tidak ada FP. Pegacakan data yang sempurna akan memberikan titik pada sepanjang garis diagonal dari kiri bawah dan sudut kanan atas sehingga garis tersebut akan membagi ruang *ROC* menjadi 2 bagian yaitu ruang di atas garis *diagonal* (nilai klasifikasi baik) dan ruang di bawah garis *diagonal* (nilai klasifikasi buruk). Pada kurva *ROC* akan menghasilkan nilai *AUC* yang akan dibagi menjadi beberapa bagian tingkat diagnosis yaitu **[17]:**

a. *excellent classification* yaitu nilai akurasi antara 0,90 sampai 1,00
b. *good classification* yaitu nilai akurasi antara 0,80 sampai 0,90
c. *fair classification* yaitu nilai akurasi antara 0,70 sampai 0,80
d. *poor classification* yaitu nilai akurasi antara 0,60 sampai 0,70
e. *failure classification* yaitu nilai akurasi antara 0,50 sampai 0,60

a. Hasil Pengujian dan Evaluasi Performa Algoritma C4.5

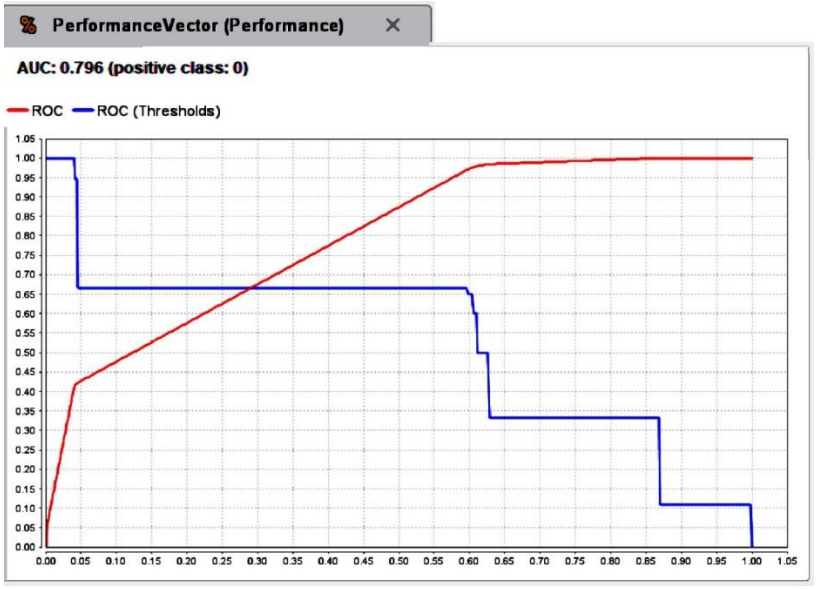
Proses pengujian dan evaluasi algoritma *C4.5* dilakukan perhitungan menggunakan (*confussion matrix*). Model confusion matrix akan membentuk *matrix* yang terdiri dari nilai true positif dan true negatif, kemudian disajikan nilai *accuracy*, dan *AUC*. Nilai *accuracy* dapat diperoleh dari persamaan $Accuracy = \frac{TN+TP}{TN+FP+FN+TP}$. TN adalah nilai *true negative rate*, TP adalah nilai *true positive rate, FN* adalah *False negative rate* dan FP *False positive rate.* Dengan dasar persamaan nilai *accuracy* tersebut penelitian ini menampilkan nilai predikasi menjadi 2 bagian yaitu pred 0 dan pred 1. Setiap nilai prediksi memiliki 2 nilai kebenaran yaitu *true 0* dan *true 1*. Hasil evaluasi nilai akurasi algoritma *C4.5* disjaikan pada **Gambar 7**.

accuracy: 77.34%

	true 1	true 0	class precision
pred. 1	106	12	89.83%
pred. 0	162	488	75.08%
class recall	39.55%	97.60%	

Gambar 7. Performa Akurasi *C4.5*

Berdasarkan hasil perhitungan evaluasi performance dengan confusion matrix yang sudah dilakukan pada gambar 7 maka diperoleh perhitungan tingkat kinerja algoritma C4.5 dengan indikator accuracy sebesar 77,34%. Dari nilai akurasi tersebut terdapat *class predision* pada pred 0 adalah 75,08% dan *class predision 1* bernilai 89,83%. Selanjutnya yaitu dengan memperhatikan performa dari nilai *AUC* pada algoritma *C4.5*. Kurva *ROC* yang menunjukan nilai *AUC* pada algoritma C4.5 disajikan pada **Gambar 8**.



Gambar 8. Hasil Kurva *ROC* Algoritma *C4.5*

Hasil kurva ROC C4.5 pada **Gambar 8** mengekspresikan *confusion matrix* dari **Gambar 7**. Garis *horizontal* adalah *false positives* dan garis *vertikal* adalah true *positives*. Hasil Kurva *ROC* algoritma *C4.5* menghasilkan nilai *AUC (Area Under Curve)* sebesar 0.796. Dengan demikian dapat disimpulkan bahwa nilai akurasi klasifikasi C4.5 masuk diagnosa kategori *fair classification.*

b. Hasil Pengujian dan Evaluasi Performa Algoritma C4.5 Berbasis PSO

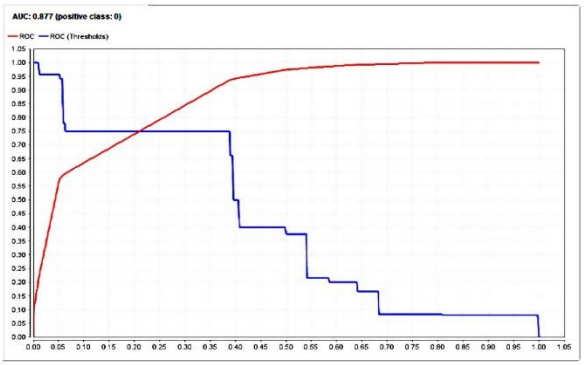
Proses pengujian dan evaluasi algoritma *C4.5 berbasis Particle Swarm Optimization (PSO)* dilakukan perhitungan menggunakan (*confussion matrix*). *Model confusion matrix* akan membentuk *matrix* yang terdiri dari nilai *true positive* dan *true negative*, kemudian disajikan nilai *accuracy*, dan *AUC.* Nilai *accuracy* diperoleh dari proporsi prediksi kebenaran. Proporsi kebenaran akan dibagi menjadi 2 bagian yaitu pred 0 dan pred 1. Kedua nilai prediksi tersebut maka diperolehlah nilai *accuracy.* Nilai akurasi dari algoritma *C4.5 berbasis PSO* disajikan pada **Gambar 9**.

accuracy: 82.29%

	true 1	true 0	class precision
pred. 1	164	32	83.67%
pred. 0	104	468	81.82%
class recall	61.19%	93.60%	

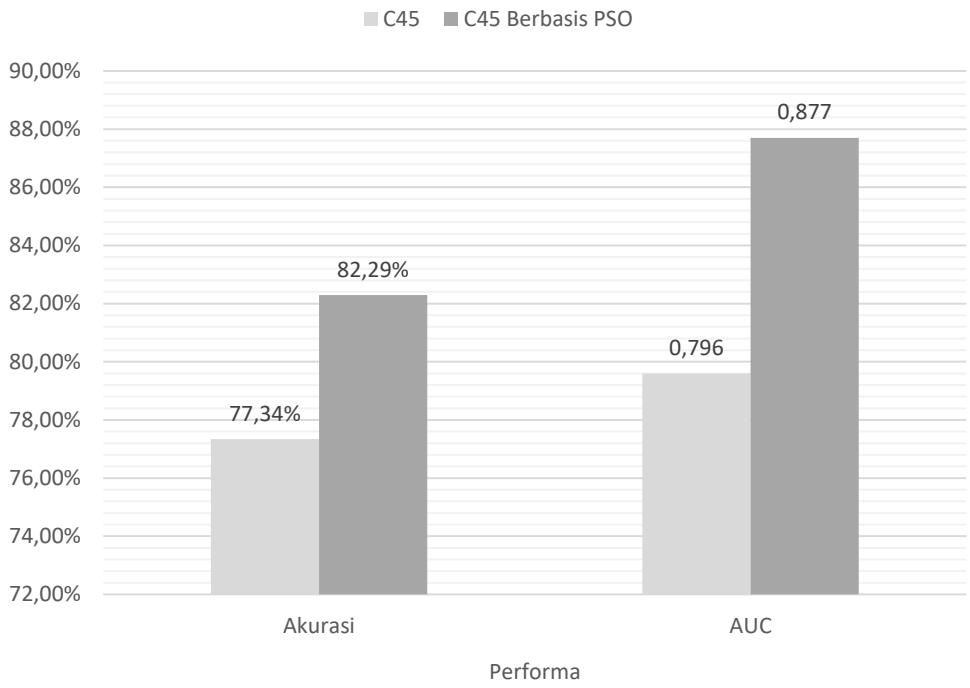
Gambar 9. Performa Akurasi *C4.5 Berbasis PSO*

Berdasarkan Gambar 9, diketahui pred 0 memiliki nilai *Class Precision* 81,82% dan pred 1 memiliki nilai *Class Precision* 83,67%. Kemudian hasil perhitungan evaluasi *performance* dengan *confusion matrix* yang sudah dilakukan maka diperoleh perhitungan tingkat kinerja algoritma *C4.5 berbasis Particle Swarm Optimization (PSO)* dengan indikator *accuracy* sebesar 82,29%. Berikutnya yaitu memperhatikan nilai *AUC* dari algoritma *C4.5 berbasis PSO.* Kurva *ROC* yang menampilkan hasil dari nilai *AUC* Algoritma *C4.5 berbasis PSO* disajikan pada **Gambar 10**.



Gambar 10. Hasil Kurva *ROC* Algoritma *C4.5 berbasis PSO*

Hasil Kurva *ROC* yang ditunjukkan gambar 10 tersebut merupakan hasil dari *confusion matrix* pada gambar 9. Garis *horizontal* adalah *false positives* dan garis *vertikal* adalah *true positives*. Hasil Kurva *ROC C4.5 berbasis PSO* menghasilkan nilai *AUC (Area Under Curve)* sebesar 0.877. Dengan demikian dapat disimpulkan bahwa nilai akurasi klasifikasi *C4.5 berbasis PSO* masuk diagnosa kategori *good classification*. Hasil dari performa algoritma dapat disajikan kedalam sebuah diagram. Diagram akan menampilkan nilai dari akurasi dan juga nilai dari *AUC*. Diagram dari performa kinerja Algoritma data mining disajikan pada **Gambar 11**.



Gambar 11. Perbandingan Nilai Akurasi dan *AUC*

3.6 *Deployment Phase*

Hasil dari pemodelan pada paper ini adalah analisis dengan meninjau performa kinerja algoritma untuk membantu dalam mendukung keputusan. Pemilihan pemodelan dilihat dari hasil nilai performa akurasi dan *AUC* terbaik. Diketahui hasil dari perbandingan Algoritma *C4.5* dan *C4.5 berbasis PSO* bahwa performa akurasi dan *AUC* terbaik terdapat pada *C4.5 berbasis PSO*. Sehingga dapat disimpulkan bahwa *C4.5 berbasis PSO* direkomendasikan sebagai algoritma untuk mendiagnosis penyakit diabetes dengan performa akuarasi = 82,29% dan *AUC* = 0,877. Dari performa yang dihasilkan tersebut maka dapat dikategorikan menjadi *good classification*. Kemudian pemodelan dapat di jadikan sebagai dasar untuk merancang dan membangun sebuah *system* keputusan terhadap diagnosis penyakit diabetes.

4. KESIMPULAN

Penderita Penyakit *Diabetes* yang semakin meningkat diberbagai dunia internasional membuat para pakar keilmuan untuk selalu berusaha meningkatkan pengetahuannya. Langkah yang dilakukan salah satunya yaitu dengan melakukan berbagai riset. Riset bidang teknologi informasi juga dapat membantu dalam pengolahan data agar data dapat dijadikan sebagai pendukung keputusan atau memberikan pengetahuan yang sebelumnya tidak diketahui.

Metode klasifikasi data mining pada penelitian ini dilakukan dengan dua pemodelan. Pemodelan pertama dengan menggunakan algoritma C4.5 kemudian pemodelan kedua adalah *C4.5* berbasis *PSO*. Penerapan pada kedua model tersebut dilakukan perbandingan terhadap nilai performa dari kinerja algoritma. Nilai performa yang ditinjau dalam penelitian ini yaitu performa akurasi dan *AUC*.

Algoritma *C4.5* menunjukkan nilai performa akurasi sebesar 77,34% dan nilai *AUC* sebesar 0.796. Dari performa yang dihasilkan *C4.5* nilai akurasi masuk pada kategori *fair classification*. Sedangkan penerapan dua kombinasi algoritma yaitu *C4.5 berbasis PSO* menunjukan nilai performa akurasi 82,29% dan menghasilkan nilai *AUC* 0,877. Dari performa yang dihasilkan *C4.5 berbasis PSO* menunjukkan kategori *good classification*. Sehingga dapat disimpulkan bahwa perbandingan nilai akurasi dan *AUC* terbaik *C4.5 berbasis PSO*. [3]

REFERENSI

[1] L. W. Marewa, Kencing Manis (Diabetes Mellitus) Di Sulawesi Selatan. Yayasan Pustaka Obor Indonesia, 2015.

[2] H. Tandra, Segala Sesuatu yang Harus Anda Ketahui Tentang Diabetes. Gramedia Pustaka Utama, 2017.

[3] Khairani, INFODATIN : Hari Diabetes Sedunia 2018". Kementerian Kesehatan RI Pusat Data dan Informasi, 2019.

[4] D. S. Purnia and A. I. Warnilah, "Implementasi Data Mining Pada Penjualan Kacamata Menggunakan Algoritma Apriori", *IJCIT Indones. J. Comput. Inf. Technol.*, vol. 2, no. 2, p. 9, 2017.

[5] A. Andriani, "Sistem Prediksi Penyakit Diabetes Berbasis Decision Tree", *J. Bianglala Inform.*, vol. 1, no. 1, p. 10, 2013.

[6] F. Fatmawati, "Perbandingan Algoritma Klasifikasi Data Mining Model C4.5 Dan Naive Bayes Untuk Prediksi Penyakit Diabetes", *J. Techno Nusa Mandiri*, vol. 13, no. 1, pp. 50–59, Mar. 2016.

[7] J. J. Pangaribuan, "Mendiagnosis Penyakit Diabetes Melitus Dengan Menggunakan Metode Extreme Learning Machine", *J. Inf. Syst. Dev. ISD*, vol. 1, no. 2, p. 9, 2016.

[8] Y. Yolanda and F. Firdaus, "Penerapan Data Mining Dalam Klasifikasi Hubungan Antara Kejadian Katarak dengan Diabetes Mellitus Menggunakan Algoritma C4.5", undergraduate, Sriwijaya University, 2015.

[9] M. F. Salim and S. Sugeng, "Analisis Rekam Medis Pasien Diabetes Mellitus Melalui Implementasi Teknik Data Mining di RSUP Dr. Sardjito Yogyakarta", *J. Kesehat. Vokasional*, vol. 2, no. 2, pp. 167–174, May 2018, doi: 10.22146/jkesvo.30331.

[10] P. S. Kumar and V. Uma Tejaswi, "Diagnosing Diabetes using Data Mining Techniques", *Int. J. Sci. Res. Publ.*, vol. 7, no. 6, p. 5, 2017.

[11] S. Bashir, U. Qamar, F. H. Khan, and M. Y. Javed, "An Efficient Rule-Based Classification of Diabetes Using ID3, C4.5, & CART Ensembles", in *2014 12th International Conference on Frontiers of Information Technology*, Islamabad, Pakistan, Dec. 2014, pp. 226–231, doi: 10.1109/FIT.2014.50.

[12] Giat, "Analisis Teknik Data Mining Algoritma C4.5 Dan K-Nearest Neighbor" Untuk Mendiagnosa Penyakit Diabetes Mellitus", *SNTIBD*, vol. 1, no. 1, pp. 77–82, 2016.

[13] M. Yusa, E. Utami, and E. T. Luthfi, "Evaluasi Performa Algoritma Klasifikasi Decision Tree ID3, C4.5, dan CART Pada Dataset Readmisi Pasien Diabetes", pp. 12, 2018.

[14] N. Yuda, "Data Mining Menggunakan Algoritma Naïve Bayes Untuk Klasifikasi Kelulusan Mahasiswa Universitas Dian Nuswantoro. (Studi Kasus: Fakultas Ilmu Komputer Angkatan 2009)", *Comput. Sci.*, 2014.

[15] S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, "DMME: Data Mining Methodology for Engineering Applications – A Holistic Extension to the CRISP-DM model", *Procedia CIRP*, vol. 79, pp. 403–408, Jan. 2019, doi: 10.1016/j.procir.2019.02.106.

[16] C. Vercellis, Business Intelligence: Data Mining and Optimization for Decision Making, 1 edition. Chichester, U.K: Wiley, 2009.

[17] F. Gorunescu, Data Mining: Concepts, Models and Techniques. Berlin Heidelberg: Springer-Verlag, 2011.