PENERAPAN METODE *SUPPORT VECTOR MACHINE* MENGGUNAKAN OPTIMASI *GENETIC ALGORITHM* UNTUK PREDIKSI PENYAKIT DIABETES

Frisma Handayanna

***Abstract*—Currently diabetes progressively increasing number of sufferers. Diabetesis a disease that can lead to complication scan even causedeath. Many studies using theme thod of Support Vector Machine predicting diabetes butthe accuracy ofthe resulting valueis stillless accurate. Inthis stud ycreate da mode lof Support Vector Machine algorithm toget the rulein predicting diabetes and provide a more accuratevalue of accuracy. After testing the model of Support Vector Machine and Support Vector Machine, the results obtained are thusobtained testing algorithm using Support Vector Machine where the value obtained accuracy is 74.21% and the AUC value is 0.758 and testing by using Support Vector Machine based Genetic Algorithm accuracy 75.26% values obtained with precision value 76.25 % and the AUC value is 0.771 .**

***Intisari*— Saat ini penyakit diabetes semakin lama semakin meningkat jumlah penderitanya. Penyakit diabetes adalah salah satu penyakit yang dapat menyebabkan komplikasi bahkan dapat menyebabkan kematian. Banyak penelitian yang menggunakan metode *Support Vector Machine* dalam memprediksi penyakit diabetes tetapi nilai akurasi yang dihasilkan masih kurang akurat. Dalam penelitian ini dibuatkan model algoritma *Support Vector Machine* untuk mendapatkan *rule* dalam memprediksi penyakit diabetes dan memberikan nilai akurasi yang lebih akurat. Setelah dilakukan pengujian dengan model *Support Vector Machine* dan *Support Vector Machine* maka hasil yang didapat adalah algoritm asehingga didapat pengujian dengan menggunakan *Support Vector Machine* dimana didapat nilai *accuracy* adalah 74.21 % dan nilai AUC adalah 0.758 dan pengujian dengan mengunakan *Support Vector Machine* berbasis *Genetic Algorithm* didapatkan nilai *accuracy* 75.26% dengan nilai *precision* 76.25% dan nilai AUC adalah 0.771.**

***Kata Kunci*— Diabetes, *Genetic Algorithm*, *Support Vector Machine*.**

I. PENDAHULUAN

Perkiraan terakhir populasi penderita penyakit diabetes menunjukkan 171 juta orang di dunia pada tahun 2000 dan diperkirakan akan meningkat menjadi 366 juta pada 2030 [17].

Program Studi Teknik Informatika STMIK Nusamandiri Jakarta, Jl. Damai No. 8 Warung Jati Barat (Margasatwa) Jakarta Selatan. Telp. (021) 78839513 Fax. (021) 78839421, email: frisma.fha@nusamandiri.ac.id

Penyakit diabetes disebabkan oleh peningkatan kadar glukosa dalam darah, apabila kadar glukosa darah meningkat dalam jangka waktu yang lama maka akan menyebabkan komplikasi seperti gagal ginjal, kebutaan dan serangan jantung [9].

Penyakit diabetes merupakan salah satu penyakit yang mematikan, faktor resiko tinggi dalam keluarga yang menyebabkan penyakit diabetes antara lain dikarenakan orang gemuk yang tidak melakukan latihan fisik, dan orang-orang yang memiliki gaya hidup yang tidak sehat dan makanan yang belebih dari apa yang dibutuhkan oleh tubuh[16].

Penelitian terlebih dahulu mengenai penyakit diabetes telah banyak dilakukan seperti penelitian yang dilakuan oleh T.Jayalskshmi dan A.Santhakumaran yang bejudul *Impact of Preprocessing for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks* menggunakan *Neural Network* untuk diagnosa penyakit diabetes mellitus dengan menggunakan metode prosedur pra-pengolahan dan nilai-nilai yang hilang mempengaruhi set data selama klasifikasi dan hasilnya dampak dari nilai yang hilang teknik dan pra-pengolahan teknik yang membuktikan bahwa beberapa kombinasi nilai-nilai yang hilang dan pra-pengolahan akurasi yang sangat meningkat [9].

Penelitian selanjutnya dilakukan oleh Jianchao Han, Juan C. Rodriguze, Mohsen Beheshti dengan judul Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner menggunakan model ID3 dan *Decison Tree*, dengan metode *Decision tree* yang dilakukan adalah memisahkan sumber data set ke dalam subset berdasarkan uji nilai atribut dan dengan ID3 yaitu dengan pendekatan di mana pohon keputusan yang secara rekursif dibangun dengan top-down dan dimulai dengan dengan memulai serangkaian pelatihan tuple dan label kelas yang terkait, dimana hasilnya akurasi ID3 lebih tinggi dari pada *Decison Tree* [5].

Neural Network mempunyai kelebihan dalam hal kemampuan generalisasi tergantung pada seberapa baik *Neural Network* meminimalkan resiko empiris namun *Neural Network* mempunyai kelemahan dimana menggunakan data pelatihan cukup besar [20]. *Decison tree* dan ID3 mempunyai kelebihan untuk keputusan pengklasifikasi memiliki akurasi yang baik namun memiliki kelemahan karena perlu mengumpulkan lebih banyak data [5]. SVM adalah kasus khusus dari keluarga algoritma yang kita sebut sebagai *regularized* metode klasifikasi linier dan metode yang kuat untuk minimalisasi resiko [21]. Dan kelebihan SVM lainnya adalah dapat meminimalkan kesalahan melalui memaksimalkan margin dengan misahkan antara *hyper-lane* dan satu set data bahkan dengan jumlah sample yang kecil [2].

Namun demikian masalah aplikasi tertentu, tidak semua fitur ini sama-sama penting dan kinerja yang lebih baik dapat dicapai dengan membuang beberapa fitur dengan begitu fitur dalam SVM memiliki pengaruh penting dalam akurasi klasifikasi [23]. Dataset yang tidak penting, fitur yang banyak atau sangat berhubungan secara signifikan akan mengurangi tingkat akurasi klasifikasi dengan menghapus fitur ini, dengan begitu tingkat akurasi efisiensi dan klasifikasi dapat diperoleh [11].

II. KAJIAN LITERATUR

a. Diabetes

Diabetes adalah epidemi yang paling cepat berkembang di Barat dunia. Satu dari tiga anak-anak Amerika akan tumbuh dan terkena diabetes, 24% orang dewasa Amerika resisten insulin dan 45% orang dewasa di atas usia 60 resisten insulin (Mason, 2005). Diabetes adalah salah satu penyebab utama kematian di banyak negara dan penyebab utama kebutaan, gagal ginjal, dan *nontraumatic* amputasi [18].

Faktor penyebab diabetes adalah [15]:

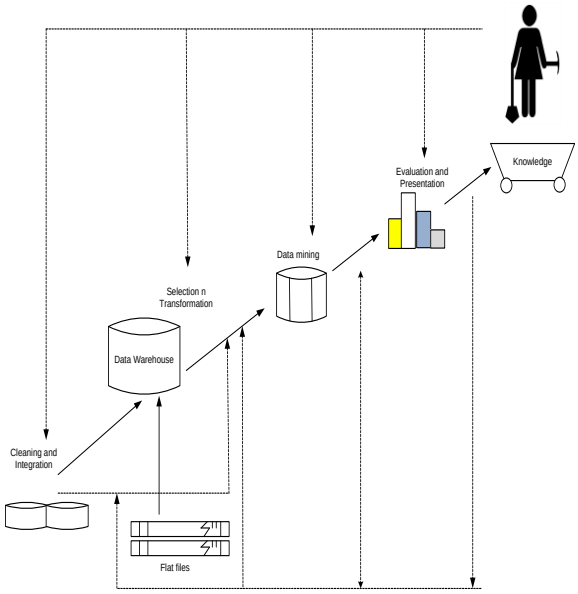
1. Gen diabetes dalam keluarga
   Gen merupakan sel pembawa sifat yang dapat diwariskan orang tua kepada keturunannya, dan diabetes merupakan penyakit yang bisa diwariskan.
2. Insulin dan gula darah
   Insulin adalah karena ketidakmampuan beta sel-sel di pankreas untuk memproduksi insulin [14]. Produksi ini disebabkan oleh tingginya kadar gula dalam darah, sehingga menyebakan insulin diproduksi semakin tinggi.
3. Kegemukan (Obesitas)
   Pada kegemukan atau obesitas sel-sel lemak yang menggemuk yang jumlahnya lebih banyak dari pada keadaan tidak gemuk, sehingga menyebabkan reistensi terhadap insulin dimana gula darah sulit masuk kedalam sel, sehingga gula darah tetap tinggi (hiperglikemi) sehingga terjadilah diabetes, khususnya terjadi pada diabetes tipe2.
4. Asma
   Penderita yang mengalami penyakit asma diharuskan untuk mengkonsumsi obat asma, sehingga memicu terjadinya diabetes, dikarenakan hormon yang digunakan pada obat asma tersebut adalah steroid yang berkerja berlawanan dengan insulin yang menaikkan gula darah.
5. KB
   Pil kontrasepsi merupakan salah satu obat yang mengandung hormon streoid dengan anti insulin rendah.

Ada tiga tipe utama diabetes: [14]

1. Tipe 1 (tergantung insulin) adalah karena ketidakmampuan beta sel-sel di pankreas untuk memproduksi insulin. Hal ini biasanya terjadi di masa kecil atau masa remaja dimana pasien-pasien harus menyuntikkan insulin, karena mereka tidak bisa memproduksinya secara alami.
2. Penderita diabetes tipe 2 (non-insulin dependent) memproduksi insulin, tetapi sel-sel hanya tidak bereaksi dengan baik, pankreas tidak hanya menghasilkan insulin tetapi biasanya overproduces, karena efektivitas sangat berkurang.
3. Jenis ketiga disebut "diabetes gestational", karena hanya mempengaruhi ibu hamil, untuk beberapa alasan wanita hamil lebih rentan terhadap diabetes dibandingkan orang lain.

b. Data Mining

Data Mining merupakan teknologi baru yang sangat berguna untuk membantu perusahaan-perusahaan menemukan informasi yang sangat penting dari gudang data mereka. Beberapa aplikasi data mining fokus pada prediksi, mereka meramalkan apa yang akan terjadi dalam situasi baru dari data yang menggambarkan apa yang terjadi di masa lalu [22]. Secara khusus, koleksi metode yang dikenal sebagai 'data mining' menawarkan metodologi dan solusi teknis untuk mengatasi analisis data medis dan konstruksi prediksi model [1]. Model data mining memberikan contoh penerapannya pada berbagai algoritma dan pada data set yang besar [12].



Sumber : Han dan Kamber (2007)
Gambar 1. Data mining

Tahapan data mining dalam proses penemuan pengetahuan [6]:

1. Pembersihan data (untuk menghilangkan noise dan data tidak konsisten)
2. Integrasi data (di mana beberapa sumber data dapat dikombinasikan)

c. Data seleksi (di mana data yang relevan dengan tugas analisis basis data yang akan diambil)
d. Data transformasi (dimana data diubah atau dikonsolidasikan ke dalam bentuk yang sesuai untuk pertambangan dengan melakukan operasi ringkasan atau agregasi)
e. Data mining (proses esensial dimana metode cerdas diaplikasikan untuk mengekstrak pola data)

6. Pola evaluasi (untuk mengidentifikasi pola yang benar-benar menarik yang mewakili pengetahuan didasarkan pada beberapa langkah-langkah interestingness)
7. Pengetahuan presentasi (dimana visualisasi dan teknik representasi pengetahuan digunakan untuk menyajikan pengetahuan ditambang kepada pengguna)

Terdapat empat penglompokan dalam data mining yaitu klasifikasi, asosiasi, *clustering* dan prediksi [22]:

1. Klasifikasi
   Proses klasifikasi didasarkan: Kelas-variabel dependen dari model-yang merupakan variabel kategori mewakili yang 'label' memakai objek setelah klasifikasin, contohnya loyalitas pelanggan, kelas bintang (galaksi), kelas gempa bumi (badai) [4].
2. Asosiasi
   Setiap asosiasi antara fitur-fitur yang dicari, bukan hanya satu yang memprediksi nilai kelas tertentu [19]. Pada prinsipnya, penemuan aturan asosiasi/asosiasi mempelajari aturan bagaiman kita memahami proses mengidentifikasi aturan antara ketergantungan yang berbeda dari fenomena kelompok. Dengan demikian, mari kita perkirakan kumpulan set yang kita punya masing-masing berisi sejumlah objek/benda-benda. Jadi tujuan kita untuk mencari peraturan yang menghubungkan (asosiasi), obyek ini berdasarkan peraturan ini, untuk dapat memprediksi terjadinya objek/item, berdasarkan kejadian lain [4].
3. *Clustering*
   Cluster adalah menemukan kelompok (kelompok) objek, berdasarkan kemiripan (semacam kemiripan), sehingga dalam setiap kelompok ada kemiripan yang besar, sementara kelompok cukup berbeda dari satu sama lain [4].
4. Prediksi
   Prediksi/perkiraan model yang berkaitan dengan kemampuan untuk memprediksi tanggapan terbaik (output), yang paling dekat ke kenyataan, berdasarkan input data [4].

f. Support Vector Machine

SVM adalah sebuah metode seleksi yang membandingkan parameter standarseperangkatnilai diskrit yang disebut kandidat set, dan mengambil salah satu yang memilikiakurasiklasifikasiterbaik [3]. *Support Vector Machine* (SVM) adalah seperangkat metode yang terkait untuk suatu metode pembelajaran, untuk kedua masalah klasifikasi dan regresi [13]. Dengan berorientasi pada tugas, kuat, sifat komputasi mudah dikerjakan, SVM telah mencapai sukses besar dan dianggap sebagai *state-of-the-art classifier* saat ini [7].

Data yang tersedia dinotasikan sebagai $\overrightarrow{x_i} \in \Re^d$ sedangkan label masing-masing dinotasikan $y_i \in \{-1+1\}$ untuk i = 1,2,....,1 yang mana l adalah banyaknya data. Diasumsikan kedua class –1 dan +1 dapat terpisah secara sempurna oleh hyperplane berdimensi d, yang didefinisikan:

Diasumsikan kedua class –1 dan +1 dapat terpisah secara sempurna oleh *hyperplane* berdimensi d, yang didefinisikan:

$$\vec{w}.\vec{x} + \mathrm{b} = 0 \qquad (1)$$

Sebuah pattern xi yang termasuk class –1 (sampel negatif) dapat dirumuskan sebagai pattern yang memenuhi pertidaksamaan:

$$\vec{w}.\vec{x} + \mathrm{b} = -1 \qquad (2)$$

sedangkan pattern yang termasuk class +1 (sampel positif):

$$\vec{w}.\vec{x} + \mathrm{b} = +1 \qquad (3)$$

Margin terbesar dapat ditemukan dengan memaksimalkan nilai jarak antara hyperplane dan titik terdekatnya, yaitu 1/|| $\vec{w}$ ||.Hal ini dapat dirumuskan sebagai Quadratic Programming (QP) problem, yaitu mencari titik minimal persamaan (4), dengan memperhatikan constraint persamaan (5)

$$\min_{\vec{w}} \tau(\mathrm{w}) = \frac{1}{2}||\vec{w}||^2 \qquad (4)$$

$$y_i(\overrightarrow{x_i}.\vec{w} + \mathrm{b}) - 1 \geq 0, \ \forall i \qquad (5)$$

Problem ini dapat dipecahkan dengan berbagai teknik komputasi, diantaranya *Lagrange Multiplier* sebagaimana ditunjukkan pada persamaan (6)

$$L(w,b,a)\frac{1}{2}||\vec{w}||^2 w \sum_{i=1} a_i \ (y_i((\overrightarrow{x_i}.\vec{w} + \mathrm{b}))$$

(i=1,2,...,l) (6)

α_i adalah Lagrange multipliers, yang bernilai nol atau positif ($\alpha_i \geq 0$). Nilai optimal dari persamaan (6) dapat dihitung dengan meminimalkan L terhadap $\vec{w}$ dan b, dan memaksimalkan L terhadap α_i. Dengan memperhatikan sifat bahwa pada titik optimal gradient L=0, persamaan langkah (6)dapat dimodifikasi sebagai maksimalisasi problem yang hanya mengandung α_i saja, sebagaimana persamaan (7).

Maximize:

$$\sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i,j=1}^{l} a_i a_j y_i y_j x_i x_j \qquad (7)$$

Subject to:

$a_i \geq 0(i = 1,2,\dots,l \sum_i^l a_i \ y_i=0$ (8)

Dari hasil dari perhitungan ini diperoleh α_i yang kebanyakan bernilai positif. Data yang berkorelasi dengan α_i yang positif inilah yang disebut sebagai *support vector*.

g. Genetic Algorithm

Optimasi adalah proses menyesuaikan kepada masukan atau karakteristik perangkat, proses matematis, atau percobaan untuk menemukan output minimum atau maksimum atau hasil. Input terdiri dari variabel, proses atau fungsi dikenal sebagai fungsi biaya, fungsi tujuan, atau kemampuan fungsi, dan output adalah biaya atau tujuan, jika proses adalah sebuah percobaan, kemudian variabel adalah masukan fisik untuk percobaan [7]. *Genetic Algorithm* diterapkan pada masalah optimasi, di mana kita dituntut untuk mengoptimalkan suatu tujuan fungsi dengan memvariasikan beberapa variabel atau parameter. *Genetic Algorithm*mudah mempararelkan dan telah digunakan untuk diklasifikasikan sebagai serta masalah optimasi lainnya, dalam data mining, mereka dapat digunakan untuk mengevaluasi [6]. *Genetic Algorithm* dimulai dengan satu set solusi atau populasi, setiap individu pada set populasi merupakan solusi untuk masalah ini, sifat dari solusi bervariasi, beberapa solusi mungkin sangat baik, sementara yang lain sangat buruk namun kebaikan solusi bervariasi dari individu ke individu dalam populasi dan untuk mengukur kebaikan dari solusi, kita menggunakan fungsi yang disebut fungsi fitness, yang harus dibuat sesuai dengan logika program [19].

Beberapa keuntungan dari *Genetic Algorithm*:

1. Mengoptimalkan dengan variabel kontinyu atau diskrit
2. Tidak membutuhkan informasi derivatif
3. Bersamaan pencarian dari sampling luas permukaan biaya
4. Berkaitan dengan sejumlah besar variabel
5. Apakah cocok untuk komputer paralel
6. Mengoptimalkan variabel dengan permukaan biaya yang sangat kompleks (mereka bisa melompat dari minimum lokal)
7. Menyediakan daftar variabel yang optimal, bukan hanya solusi tunggal
8. Dapat menyandikan variabel sehingga optimasi dilakukan dengan dikodekan variabel, dan
9. Bekerja dengan data numerik yang dihasilkan, data percobaan, atau analisis fungsi.

Komponen *Genetic Algorithm* adalah [4]:

1. Representasi (definisi individu)
2. Evaluasi fungsi (fitness)
3. Populasi
4. Induk pemilihan mekanisme
5. Survivor pemilihan mekanisme
6. Variasi operator (rekombinasi dan mutasi)
7. Parameter yang digunakan genetika algoritma (ukuran populasi, kemungkinan menerapkan operator variasi, dll)

Pada *Genetic Algorithm*, teknik pencarian dilakukan sekaligus atas sejumlah solusi yang mungkin dikenal dengan istilah populasi. Individu yang terdapat dalam satu populasi disebut dengan istilah kromosom.Kromosom ini merupakan suatu solusi yang masih berbentuk simbol.Populasi awal dibangun secara acak, sedangkan populasi berikutnya merupakan hasil evolusi kromosom-kromosom melalui iterasi yang disebut dengan istilah generasi. Pada setiap generasi, kromosom akan melalui proses evaluasi dengan menggunakan alat ukur yang disebut dengan fungsi fitness. Nilai fitness dari suatu kromosom akan menunjukkan kualitas kromosom dalam populasi tersebut. Generasi berikutnya dikenal dengan istilah anak (off-spring) terbentuk dari gabungan 2 kromosom generasi sekarang yang bertindak sebagai induk (*parent*) dengan menggunakan operator penyilangan (*crossover*).Selain operator penyilangan, suatu kromosom dapat juga dimodifikasi dengan menggunakan operator mutasi. Populasi generasi yang baru dibentuk dengan cara menyeleksi nilai fitness dari kromosom anak *(offspring)*,serta menolak

kromosom-kromosom yang lainnya sehingga ukuran populasi (jumlah kromosom dalam suatu populasi) konstan. Setelah melakukan berbagai generasi, maka algoritma ini akan konvergen ke kromosom yang terbaik [10].

III. METODE PENELITIAN

Dalam penelitian ini dilakukan beberapa langkah yang dilakukan dalam proses penelitian.

1. Pengumpulan data
   Pada tahap ini ditentukan data yang akan diproses. Mencari data yang tersedia, memperoleh data tambahan yang dibutuhkan, mengintegrasikan semua data kedalam data set, termasuk variabel yang diperlukan dalam proses.
2. Pengolahan data awal
   Ditahap ini dilakukan penyeleksian data, data dibersihkan dan ditransformasikan kebentuk yang diinginkan sehingga dapat dilakukan persiapan dalam pembuatan model.
3. Metode yang diusulkan
   Pada tahap ini data dianalisis, dikelompokan variabel mana yang berhubungan dengan satu sama lainnya. Setelah data dianalisis lalu diterapkan model-model yang sesuai dengan jenis data.Pembagian data kedalam data latihan (*training data*) dan data uji (*testing data*) juga diperlukan untuk pembuatan model.
4. Eksperimen dan pengujian metode
   Pada tahap ini model yang diusulkan akan diuji untuk melihat hasil berupa *rule* yang akan dimanfaatkan dalam pengambilan keputusan.
5. Evaluasi dan validasi
   Pada tahap ini dilakukan evaluasi terhadap model yang ditetapkan untuk mengetahui tingkat keakurasian model.

Teknik pengumpulan data

Teknik pengumpulan datayang diperoleh adalah data sekunder karena diperoleh dari Pima Indian diabetes database dalam UCI (singkatan dari Pima Diabetes). Masalah yang harus dipecahkan di sini adalah prediksi terjadinya diabetes melitus dalam waktu 5 tahun dengan menggunakan Pima yang berisi 786 orang yang diperiksa dan sebanyak 500 pasien tidak terdeteksi terkena penyakit diabetes, sehingga 268 pasien terdeteksi penyakit diabetes. Dengan atribut dari penyakit diabetes adalah berapa kali hamil, konsentrasi glukosa, tekanan darah, ketebalan lipatan kulit, serum insulin, indeks massa tubuh, diabetes silsilah fungsi dan umur dan kelas sebagai label yang terdiri atas ya dan tidak. Data pasien penyakit diabetes bisa di lihat pada Tabel 1. berikut.

Tabel 1. Atribut dan data penyakit diabetes

No	Berapa Kali Hamil	Konsentrasi Glukosa	Tekanan darah	Lipatan kulit	Serum Insulin	Masssa Tubuh	Diabetes Silsilah Fungsi	Umur	Kelas
1	6	14	148	35	0	33,6	0,627	50	Ya
2	1	85	66	29	0	26,6	0,351	31	Tidak
3	8	18	64	0	0	23,3	0,672	32	Ya
4	1	89	66	23	94	28,1	0,167	21	Tidak
5	0	13	40	35	168	43,1	2,288	33	Ya
6	5	11	74	0	0	25,6	0,201	30	Tidak
7	3	78	50	32	88	31	0,248	26	Ya
8	10	115	115	0	0	35,3	0,134	29	Tidak
9	2	19	70	45	543	30,5	0,158	53	Ya
10	8	12	96	0	0	0	0,232	54	Ya
11	4	11	92	0	0	37,6	0,191	30	Tidak
12	10	168	74	0	0	38	0,537	34	Ya
13	10	139	80	0	0	27,1	1,441	57	Tidak
14	1	18	60	23	846	30,1	0,398	59	Ya
15	5	16	73	19	175	25,8	0,587	51	Ya

Sumber: UCI Repository (2012)

Pengolahan data awal

Jumlah data awal yang diperoleh dari pengumpulan data yaitu sebanyak 768 data, namun tidak semua data dapat digunakan dan tidak semua atribut digunakan karena harus melalui beberapa tahap pengolahan awal data (*preparation data*)**.**

Tabel 2. Tabel Atribut yang digunakan

No	Atribut	Nilai
1	Berapa Kali Hamil	Berapa kali wanita hamil
2	Konsentrasi Glukosa	Konsentrasi glukosa plasma 2 jam dalam tes toleransi glukosa oral
3	Tekanan Darah	Tekanan Darah diastolik (mmHg)
4	Lipatan Kulit	Triceps ketebalan lipatan kulit (mm)
5	Serum Insulin	2-Jam serum insulin (mu U / ml)
6	Masssa Tubuh	Indeks massa tubuh (berat dalam kg / (tinggi dalam m) ^ 2)
7	Diabetes Silsilah Fungsi	Diabetes silsilah fungsi
8	Umur	Umur (tahun)

Sumber: UCI Repository (2012)

Metode yang disulkan

Pada tahap modeling ini dilakukan pemprosesan data traning sehingga akan membahas metode algoritma yang diuji dengan memasukan data penyakit diabetes.

1. Eksperimen dan Pengujian Metode

Tahap modeling untuk menyelesaikan prediksi penyakit diabetes dengan menggunakan dua metode yaitu algoritma *Support Vector Machine. Support Vector Machine* yaitu suatu metode sebuah metode seleksi fitur, dan mengambil salah satu yang memiliki akurasi klasifikasi terbaik.

2. Evaluasi dan Validasi Hasil

Model yang diusulkan pada penelitian tentang prediksi penyakit diabetes adalah dengan menerapkan *Support Vector Machine* dan *Support Vector Machine* berbasis *Genetic Algorithm*. Penerapan algoritma *Support Vector Machine* dengan menentukan nilai *weight* terlebih dahulu. Setelah didapatkan nilai akurasi dan AUC terbesar, nilai *weight* tersebut akan dijadikan nilai yang akan digunakan untuk mencari nilai akurasi dan AUC tertinggi.

IV. HASIL DAN PEMBAHASAN

Hasil Eksperimen dan Metode

Nilai *training cycles* dalam penelitian ini ditentukan dengan cara melakukan uji coba memasukkan C, epsilon. Berikut ini adalah hasil dari percobaan yang telah dilakukan untuk penentuan nilai *training cycles*:

Tabel 3. Eksperiment penentuan nilai *training cycle SVM*

C	epsiolon	SVM	
		accury	AUC
0.0	0.0	74,21%	0,753
1.0	1.0	65%	0,500
1.0	1.0	65%	0,500
1.0	0.0	74,21%	0,758
1.0	1.0	65%	0,500
1.0	1.0	65%	0,500
1.0	1.0	65%	0,500
1.0	0.0	74,21%	0,758
1.0	0.0	74,21%	0,758

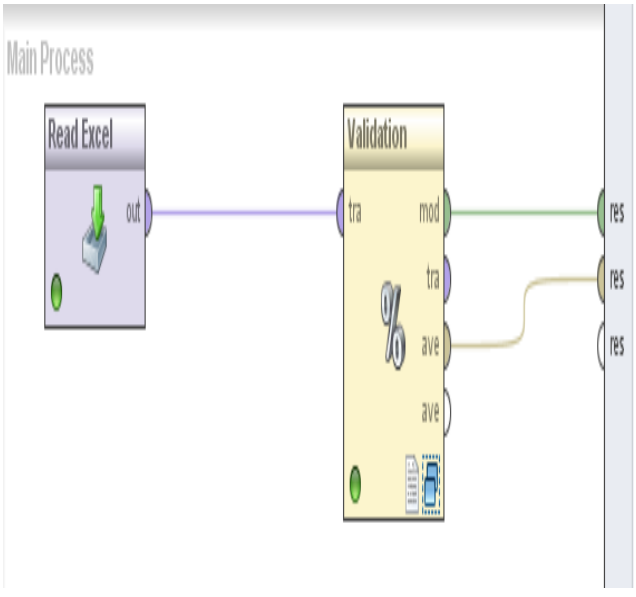
Sumber: Hasil Penelitian (2012)

Hasil terbaik pada *eksperiment* SVM diatas adalah dengan C=0.0 dan Epsilon=0.0 dihasilkan accuracy 74,21% dan AUCnya 0.753 untuk SVM dengan C=1.0 dan Epsilon=0.0 dihasilkan *accuracy* 74,21% dan AUCnya 0.758.

1. Evaluasi dan Validasi Hasil

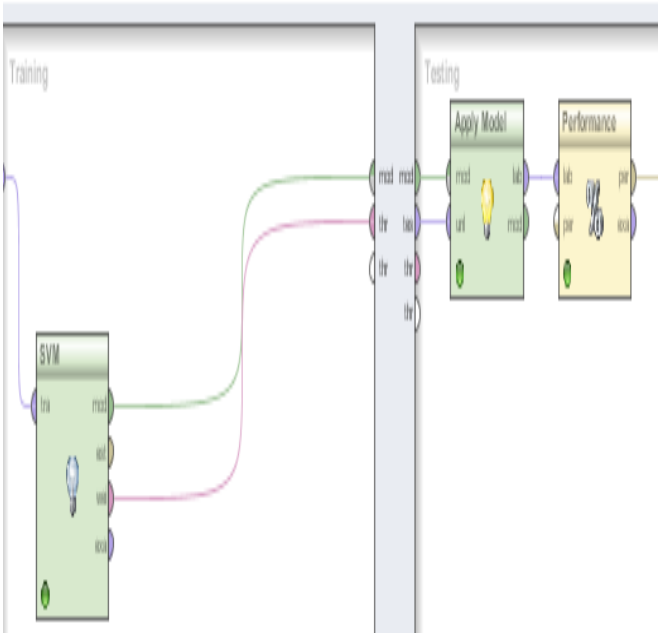
a. Hasil Pengujian Model *Support Vector Machine*

Hasil dari pengujian model yang dilakukan adalah memprediksi penyakit diabetes dengan *Support Vector Machine* untuk menentukan nilai *accuracy* dan *AUC.* Dalam menentukan nilai tingkat keakurasian dalam model *Support Vector Machine.* Metode pengujiannya menggunakan *cross validation* dengan desain modelnya sebagai berikut.



Sumber: Hasil Penelitian (2012)
Gambar 2. Desain model Validasi

Pada penelitian penentuan hasil penyakit diabetes menggunakan algoritma *Support Vector Machine* berbasis pada framework RapidMiner sebagai berikut:



Sumber: Hasil Penelitian (2012)
Gambar 3. Model pengujian validasi *Support Vector Machine*

Nilai *accuracy*, *precision*, dan *recall* dari *data training* dapat dihitung dengan menggunakan Rapid Miner. Hasil pengujian dengan menggunakan model *Support Vector Machine* didapatkan hasil pada Tabel 4.

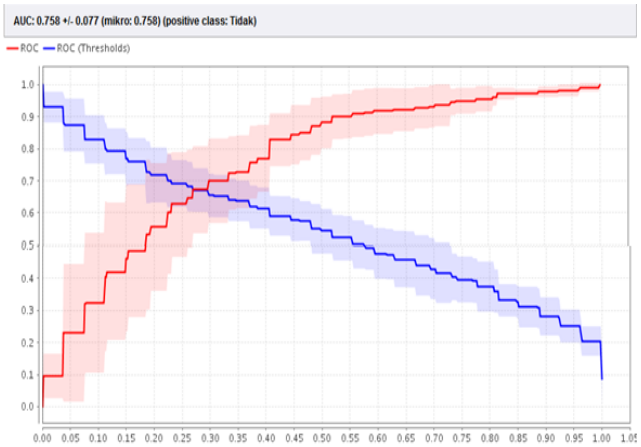
Tabel 4 diketahui dari 768 data, 118 diklasifikasikan *ya* sesuai dengan prediksi yang dilakukan dengan metode SVM, lalu 48 data diprediksi ya tetapi ternyata hasilnya prediksi tidak, 452 data *class* tidak diprediksi sesuai, dan 150 data diprediksi tidak ternyata hasil prediksinya ya.

Tabel 4. Model *Confusion Matrix* untuk Metode *Support Vector Machine*

accuracy:74.21% +/-5.79% (mikro: 74.22%)			
	True Ya	True Tidak	Class precission
pred. Ya	118	48	71.08%
pred. Tidak	150	452	75.08%
class recall	44.03%	90.40%	

Sumber: Hasil Penelitian (2012)

Hasil perhitungan divisualisasikan dengan kurva ROC. Perbandingan kedua metode komparasi bisa dilihat pada Gambar 4 yang merupakan kurva ROC untuk algoritma *Support Vector Machinenes* Kurva ROC pada gambar 4 mengekspresikan *confusion matrix* dari Tabel 4. Garis horizontal adalah *false positives* dan garis vertikal *true positives*.



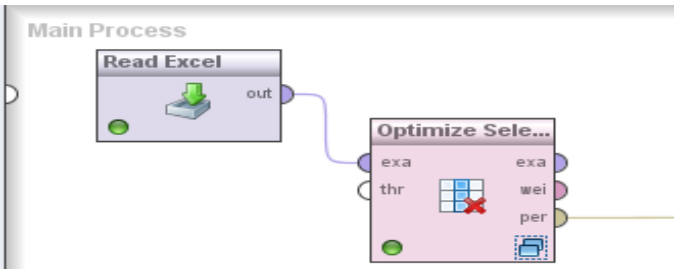
Sumber: Hasil Penelitian (2012)
Gambar 4.Kurva ROC dengan Metode *Support Vector Machine*

Dari Gambar 4 terdapat grafik ROC dengan nilai AUC (*Area Under Curve*) sebesar 0.758 dimana diagnosa hasilnya *Fair classification.*

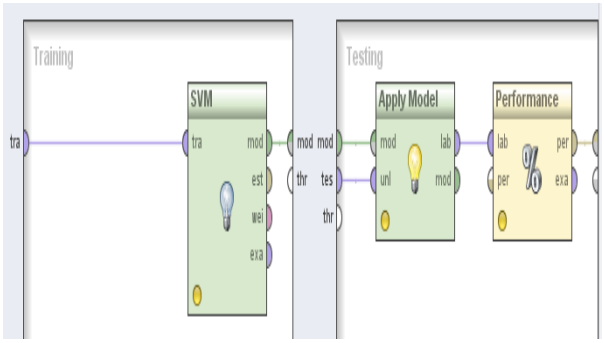
Berdasarkan hasil *eksperiment* yang dilakukan untuk memecahkan masalah prediksi hasil prediksi penyakit diabetes, dapat disimpulkan bahwa hasil *eksperiment* menggunakan metode *Support Vector Machine* mempunyai tingkat akurasi sebesar 74.21 % dan mempunyai nilai AUC sebesar 0.753. Setelah dilakukan penyesuaian pada parameter C dan epsilon didapat nilai akurasi terbaik untuk algoritma *Support Vector Machine* yaitu mempunyai akurasi sebesar 74.21 % dan nilai AUCnya sebesar 0.758.

b. Hasil Pengujian Model *Support Vector Machine* berbasis *Genetic Algorithm*

Pada penelitian penentuan hasil penyakit diabetes menggunakan algoritma *Support Vector Machine* berbasis *Genetic Algorithm* (GA) pada framework RapidMiner sebagai berikut



Sumber: Hasil Penelitian (2015)
Gambar 5. Model pengujian validasi *support vector machine*



Sumber: Hasil Penelitian (2015)

Gambar 5. Model pengujian validasi *Support Vector Machine* Berbasis *Genetic Algorithm* (GA)

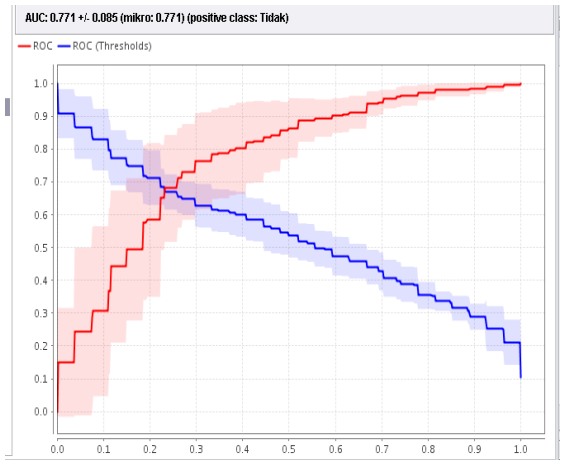
Nilai *accuracy*, *precision*, dan *recall* dari *data training* dapat dihitung dengan menggunakan Rapid Miner. Hasil pengujian dengan menggunakan model *Support Vector Machine* Berbasis *Genetic Algorithm* didapatkan hasil pada Tabel 5.
Tabel 5 diketahui dari 768 data, 126 diklasifikasikan *ya* sesuai dengan prediksi yang dilakukan dengan metode SVM berbasis GA, lalu 48 data diprediksi ya tetapi ternyata hasilnya prediksi tidak, 452 data *class* tidak diprediksi sesuai, dan 142 data diprediksi tidak ternyata hasil prediksinya ya.

Tabel 5. Model *Confusion Matrix* untuk Metode *Support Vector Machine* berbasis *Genetic Algorithm*

accuracy:75.26% +/-4.91% (mikro: 77.34%)			
	True Ya	True Tidak	Class precission
pred. Ya	126	48	72,41%
pred. Tidak	142	452	76.09%
class recall	47.01%	90.40%	

Sumber: Hasil Penelitian (2015)

Berdasarkan Tabel 5.tersebut menunjukan bahwa, tingkat akurasi dengan menggunakan algoritma SVM berbasis *Genetic Algorithm* (GA) adalah sebesar 75,26%.
Hasil perhitungan divisualisasikan dengan kurva ROC. Perbandingan kedua metode komparasi bisa dilihat pada Gambar 6 yang merupakan kurva ROC untuk algoritma *Support Vector Machine* berbasis *Genetic Algorithm* (GA).



Sumber: Hasil Penelitian (2015)

Gambar 6. Kurva ROC dengan Metode *Support Vector Machine berbasis Genetic Algorithm*(GA)

Dari hasil pengujian tersebut, baik evaluasi menggunakan *counfusion matrix* maupun ROC *curve* terbukti bahwa hasil pengujian algoritma SVM berbasis GA memiliki nilai akurasi yang lebih tinggi dibandingkan dengan algoritma SVM Nilai akurasi untuk model algoritma SVM sebesar 74.21% dan nilai akurasi untuk model algoritma SVM berbasis GA sebesar 75.26 % dengan selisih akurasi 1,05%, dapat dilihat pada Tabel 4.5 dibawah ini:

Tabel 6. Pengujian algoritma SVM dan SVM berbasis GA

	Accuracy	***AUC***
SVM	74.21%	0.753
SVM berbasis GA	75.26 %	0.771

Sumber: Hasil Penelitian (2015)

Untuk evaluasi menggunakan ROC *curve* sehingga menghasilkan nilai AUC (*Area Under Curve*) untuk model algoritma *Support Vector Machine* mengasilkan nilai 0.753 dengan nilai diagnosa *Fair Classification*, sedangkan untuk algoritma *Support Vector Machine* (SVM) berbasis *Genetic Algorithm* (GA) menghasilkan nilai 0.771 dengan nilai diagnose *Fair Classification*, dan selisih nilai keduanya sebesar 0.018. Dengan demikian algoritma *Support Vector Machine* (SVM) berbasis *Genetic Algorithm* (GA) dapat memberikan solusi untuk permasalahan dalam prediksi hasil prediksi penyakit diabetes.

V. KESIMPULAN

Berikut ini kesimpulan yang penulis ambil setelah melakukan penelitian

1. Pengujian model dengan menggunakan *Support Vector Machine* dan *Support Vector Machine* berbasis *Genetic Algorithm* dengan menggunakan data penyakit diabetes yang terkena penyakit atau tidak. Model yang dihasilkan diuji untukmendapatkan nilai *accuracy*, *precision* dan AUC dari setiap algoritmasehingga didapat pengujian dengan menggunakan *Support Vector Machine*didapat

nilai *accuracy* adalah 74.21 % dengan nilai *precision* 74.75 % dan nilai AUC adalah 0.753

2. Pengujian dengan mengunakan *Support Vector Machine* berbasis *Genetic Algorithm* didapatkan nilai *accuracy* 75.26% dengan nilai *precision* 76.25% dan nilai AUC adalah 0.771. Maka dapat disimpulan pengujian pengujian data diabetes UCI data set menggunakan *Support Vector Machine* berbasis *Genetic Algorithm* lebih baik dari pada *Support Vector Machine* sendiri. Dengan demikian dari hasil pengujian model diatas dapat disimpulkan bawa *Support Vector Machine* berbasis *Genetic Algorithm* memberikan pemecahan untuk permasalahan prediksi penyakit diabetes lebih baik.

Pada bagian ini, penulis memberikan saran-saran berdasarkan permasalahan serta kesimpulan yang penulis dapat selama penelitian, yaitu :

1. Penelitian ini diharapkan dapat digunakan sebagai bahan pertimbangan memprediksi penyakit diabetes oleh pihak medis, sehingga dapat meningkatkan akurasi dalam prediksi prediksi penyakit diabetes.
2. Penelitian ini dapat dikembangkan dengan metode optimasi lainnya seperti *Ant Colony Optimization* (ACO), *Particle Swarm Optimization* (PSO), dan lainnya.
3. Penelitian ini dapat dikembangkan dengan metode klasifikasi data mining lainnya seperti *Neural Network*, *Naive Bayes*, *KNN* dan lainnya untuk melakukan perbandingan.

REFERENSI

[1] Bellazzi, R., & Zupanb, B. Predictive Data Mining In Clinical Medicine: Current Issues And And Guidelines. International Journal Of Medical Informatics 7 7 , 81–97. 2008

[2] Chunjiang, He. Cuilian, Zhang. Yan, Zhao. A New Svm Merged Into Data Information. IEEE Asia-Pacific Conference, 14-17. 2009.

[3] Dong, Y., Xia, Z., Tu, M., & Xing, G. An Optimization Method For Selecting Parameters In Support Vector Machine. Sixth International Conference On Machine Learning And Applications , 1. 2007.

[4] Gorunescu, F. Data Mining Concepts,Models And Techniques. Verlag Berlin Heidelberg: Springer. 2011.

[5] Han, J., Rodriguze, J. C., & Beheshti, M. Diabetes Data Analysis And Prediction Model Discovery Using Rapidminer. Second International Conference On Future Generation Communication And Networking. 96-99. 2008

[6] Han, J., dan Kamber, M. Data Mining Concepts And Techniques. San Francisco: Morgan Kaufmann Publisher. 2007.

[7] Haupt, R. L., dan Haupt, S. E. Practical Genetic Algorithms. Untied States Of America: A John Wiley & Sons Inc Publication. 2004.

[8] Huang, K., Yang, H., King, I., & Lyu, M. Machine Learning Modeling Data Locally And Globally. Berlin Heidelberg: Zhejiang University Press, Hangzhou And Springer-Verlag Gmbh. 2008.

[9] Jayalskshmi dan Santhakumaran. Impact of Preprocessing for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks. IEEE International Conference , 109-112. 2012.

[10] Kusumadewi, S., dan Purnomo, H. Penyelesaian Masalah Optimasi Dengan Teknik-Teknik Heuristik. Yogyakarta: Graha Ilmu. 2015.

[11] Lin, S.-W., Shiue, Y.-R., Chen, S.-C., & Cheng, H.-M, Applying enhanced data mining approaches in predicting bank performance: A case of Taiwanese commercial banks. Expert Systems with Applications' 11543–11551. 2009.

[12] Larose, D. T. Data Mining Methods And Models. New Jersey: A John Wiley & Sons. 2007.

[13] Maimon, O. Data Mining And Knowledge Discovery Handbook. New York Dordrecht Heidelberg London: Springer. 2012.

[14] Mason, R. The Natural Diabetes Cure.Usa: 4th Printing Spring 2012. 2005.

[15] Nurrahmani, U. Stop!Diabetes Mellitus. Yogyakarta: Familia. Moertini, V. S. (2002). Data Mining Sebagai Solusi Bisnis. Integral, Vol. 7 No. 1, April 2002., 2012.

[16] Nuwangi, S., Oruthotaarachchi, C. R., Tilakaratna, J., & Caldera, H. A. Utilization Of Data Mining Techniques In Knowledge Extraction For Diminution Of Diabetes. 2012 Second Vaagdevi International Conference On Information Technology For Real World Problems, 3-8. 2012.

[17] Report Who. Definition And Diagnosis Of Diabetes Mellitus And Intermediate Hyperglycemia. Switzerland: Who Document Production Services. 2006.

[18] Robert, F. G.,Zgonis, T., & Driver, V. R. Diabetic Foot Disorders: A Clinical Practice Guideline (2006 Revision). The Journal Of Foot & Ankle Surgery , 3. 2006.

[19] Shukla, A., Tiwari, R., & Kala, R. Real Life Applications Of Soft Computing. New York: Taylor & Francis Group. 2012.

[20] Vapnik, V. N.. An Overview Of Statistical Learning Theory. Ieee Transactions On Neural Networks. IEEE Transactions On Neural Networks, Vol. 10, No. 5, September 988-999. 1999.

[21] Weiss, S. M., Indurkhya, N., & Zhang, T. Fundamentals Of Predictive Text Mining. London: Springer. 2012.

[22] Witten, I. H., Frank, E., & Hall, M. A. Data Mining Practical Machine Learning Tools And Techniques. Burlington, Usa: Morgan Kaufmann Publishers. 2011.

[23] Zhao, M., Fu, C., Ji, L., Tang, K., & Zhou, M.. Feature Selection And Parameter Optimization For Support Vector Machine: A New Approach Based On Genetic Algorithm With Feature Chromosomes. School Of Computer Science And Technology, 5197–5204. 2011.

Frisma Handayanna, M.Kom. Tahun 2012 lulus dari Program Strata Satu (S1) Program Studi Sistem Informasi STMIK Nusa Mandiri Jakarta. Tahun 2012. lulus dari Program Strata Dua (S2) prodiImu Komputer STMIK Nusa Mandiri Jakarta Aktif mengajar di STMIK Nusa Mandiri Jakarta. Telah melakukan penulisan paper di Jurnal STMIK Antarbangsa no Jurnal Sistem Informasi no ISSN 2089-8711 Vol. IV No.1 Februari 2015.