Optimasi Algoritma Naïve Bayes dengan Menggunakan Algoritma Genetika untuk Prediksi Kesuburan *(Fertility)*

Duwi Cahya Putri Buani
Program Studi Teknik Informatika, STMIK Nusa Mandiri Jakarta
dputribuani8@gmail.com

Abstract *- Level fertility (Fertility) in two decades has decreased, in several studies that have been published stating that the cause of the decline in fertility (fertility) is the environmental and lifestyle factors such as alcohol and cigarettes affect the level of quality seperma. This study aimed to test the ability of Naïve Bayes in making predictions. Naïve Bayes has some weaknesses, these weaknesses can be eliminated by performing optimization using Genetic Algorithms. Previous research using Naïve Bayes showed an accuracy rate of 97.66% after optimization by using the same data to optimize Naïve Bayes with Genetic Algorithm result increased to 99.33% accuracy.*
Keywords: Algorithm Genetic, Fertility, Naïve Bayes.

Abstrak - Tingkat kesuburan (fertilitas) dalam dua dekade mengalami penurunan, dalam beberapa studi yang telah dipublikasikan menyatakan bahwa penyebab penurunan fertilitas (kesuburan) adalah faktor lingkungan dan gaya hidup seperti alkohol dan rokok mempengaruhi tingkat kualitas sperma. Penelitian ini bertujuan untuk menguji kemampuan Naïve Bayes dalam membuat prediksi. Naïve Bayes memiliki beberapa kelemahan, kelemahan ini dapat dihilangkan dengan melakukan optimasi menggunakan Algoritma Genetika. Penelitian sebelumnya menggunakan Naïve Bayes menunjukkan tingkat akurasi 97,66% setelah optimasi dengan menggunakan data yang sama untuk mengoptimalkan Naïve Bayes dengan Algoritma Genetika result meningkat menjadi akurasi 99,33%.
Kata Kunci: Algoritma Genetika, Kesuburan, Naïve Bayes.

A. PENDAHULUAN

Salah satu penurunan kesehatan dalam dua dekade adalah penurunan tingkat kesuburan (*fertility*). Yang paling parah adalah penurunan tingkat kesuburan pada laki-laki. Penelitian telah menunjukan bahwa faktor lingkungan dan gaya hidup seperti mengkonsumsi alkohol dan rokok mempengaruhi tingkat kwalitas sperma (*semen*). Analisis tingkat kesuburan sperma sangat penting untuk evaluasi potensi kesuburan pada laki-laki.

Beberapa publikasi penelitian yang telah melakukan analisis tingkat kesuburan (fertility), menunjukan perbedaan tentang penyebab terjadinya penurunan kwalitas sperma seperti menurut (Irvin Ds,1989) menyatakan bahwa berbagai kelainan mulai dari gangguan hormonal, masalah fisik hingga masalah psikologis diketahui bisa menyebabkan *infertilitas* pada pria. Meskipun banyak pilihan pengobatan namun banyak kasus tidak dapat diatasi. Kebanyakan kasus *infertilitas* pria disebabkan oleh kerusakan testis yang berujung pada ketidak mampuan testis untuk memproduksi sperma. Sekali rusak, testis tidak akan dapat mengembalikan kemampuannya untuk memproduksi sperma. Menurut (Irvine, 2000, Slingart et al, 2001) Selain itu juga ada beberapa faktor yang mempengaruhi potensi kesuburan pria, dimana terjadi peningkatan kejadian penyakit reproduksi, peneliti yang lain juga menyebutkan bahwa faktor lingkungan dan pekerjaan juga merupakan penyebab tingginya gangguan pada fungsi reproduksi (Giwercman, 2001; wong Zielihuis, Thomas, Merkus & Steegers-Theunissen 2003), dan hasil penelitian lain menyatakan bahwa gaya hidup juga mempengaruhi tingkat kesuburan (Martini et al, 2004; Agarwal, Desai, Ruffoli dan Carpi, 2008). Untuk mengatasi permasalahan-permasalahan di atas maka Dokter dengan menggunakan data yang diperoleh dari analisa sperma melakukan evaluasi potensi kesuburan pria dan membandingkan hasilnya dengan nilai referensi yang ditentukan dan ditetapkan oleh Organisasi Kesehatan Dunia (WHO,1999). (Rowe dkk, 2000) merekomendasikan bahwa penafsiran hasil harus dilakukan dengan mempertimbangkan faktor-faktor tertentu yang dapat mengubah air mani parameter, seperti demam, paparan racun, masa kanak-kanak penyakit, dll.

Machin Learning telah diterapkan dalam berbagai bidang, mulai dari disiplin ilmu teknik dan ilmu biomedis. Dalam bidang kesehatan, pakar dan sistem pendukung keputusan telah dikembangkan untuk meningkatkan efesiensi.

Penelitian ini dilakukan untuk mengoptimasi algoritma naive bayes dengan menggunakan algoritma genetika agar prediksi yang dihasilkan lebih akurat.

Menurut(Kusrini dan E. T. Luthfi, 2009) Naïve Bayes adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu *class*. Naïve

Bayes didasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi serupa dengan *decision tree* dan neural network. Naïve Bayes terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar. Menurut (Wahyuni, 2014) *Naïve bayes* merupakan algoritma klasifikasi data mining yang menganggap masing-masing atribut tidak saling berhubungan. Maka dari itu digunakan algoritma genetika untuk membantu naïve bayes dalam menentukan atribut-atribut yang harus digunakan sehingga dapat meningkatkan akurasi.

Manfaat dari penelitian ini adalah:

1. Manfaat praktis, dimana diharapkan dari hasil penelitian ini dapat dijadikan sebagai acuan dan rekomendasi bagi para dokter dalam melakukan diagnosa tingkat kesuburan semen (*fertility*) dengan hasil yang lebih baik.
2. Manfaat kebijakan, dimana hasil penelitian ini bisa dijadikan sebagai bahan pertimbangan dan rujukan untuk penentuan diagnosa tingkat kesuburan (*fertility*).
3. Manfaat Teoritis, hasil dari penelitian ini adalah dapat membantu dalam pengembangan teori tentang *Klasifikasi naïve bayes* dan algoritma genetika, untuk kepentingan selanjutnya.

.Sedangkan tujuan dari penelitian ini diharapkan dapat :

1. Meningkatkan hasil prediksi tingkat kesuburan dengan menggunakan algoritma Naive Bayes.
2. Menerapkan algoritma genetika untuk meningkatkan hasil prediksi tingkat keseuburan.

Ruang lingkup penelitian dalam prediksi tingkat kesuburan adalah melakukan penolahan data dengan menggunakan rapidminer dan membuat aplikasi untuk melakukan prediksi dengan menggunakan Visual Basic 6.0.

B. TINJAUAN PUSTAKA

1. Data Mining.

Menurut (Witten, 2011) Data mining merupakan perpaduan dari ilmu statistik, kecerdasan buatan (sitem pakar) dan penelitian dalam bidang database, untuk itu diperlukan penyaringan melalui sejumlah besar material data atau melakukan penyelidikan dengan cerdas tentang keberadaan suatu data yang memiliki nilai Daryl Pregibons. Menurut (Gorunescu, 2011) Data mining juga dapat didefenisikan sebagai sebuah proses untuk menemukan pola data.

Menurut (Han, 2006) *Knowledge discovery from data* (KDD) juga merupakan bagian dari proses data mining, dimana dalam proses penjelajahan pengetahuan dimulai dari beberapa database dengan melakukan proses cleaning dan integration sehingga menghasilkan data warehouse. Selanjutnya melakukan proses *selection* dan transformation kemudian sebut sebagai data mining untuk menemukan pola dan mendapatkan pengetahuan dari data.

2. Algoritma *Naïve Bayes*

Algoritma Naïve Bayes merupakan suatu bentuk klasifikasi data dengan menggunakan metode *probabilitas* dan *statistik.* Metode ini pertama kali dikenalkan oleh ilmuwan Inggris Thomas Bayes, yaitu digunakan untuk memprediksi peluang yang terjadi di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai *teorema Bayes*. Menurut (Bramer, 2007) Metode *Teorema bayes* kemudian dikombinasikan dengan *naive* yang diasumsikan dengan kondisi antar atribut yang saling bebas. *Algoritma Naive Bayes* dapat diartikan sebagai sebuah metode yang tidak memiliki aturan, *Naive Bayes* menggunakan cabang matematika yang dikenal dengan teori probabilitas untuk mencari peluang terbesar dari kemungkinan klasifikasi, dengan cara melihat frekuensi tiap klasifikasi pada data training. Menurut (Kusrini & Luthfi, 2009) *Naive Bayes* juga termasuk metode klasifikasi yang sangat populer dan masuk dalam sepuluh algoritma terbaik dalam data mining, algoritma ini juga dikenal dengan nama *Idiot's Bayes, Simple Bayes*, dan *Independence Bayes*. *Klasifikasi bayesian* memiliki kemampuan klasifikasi serupa dengan *decision tree* dan *neural network*. Klasifikasi *Naive Bayes* adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class.

Bayes rule digunakan untuk menghitung probabilitas suatu class. *Algoritma Naive Bayes* memberikan suatu cara mengkombinasikan peluang terdahulu dengan syarat kemungkinan menjadi sebuah formula yang dapat digunakan untuk menghitung peluang dari tiap kemungkinan yang terjadi.

Berikut adalah bentuk umum dari *teorema bayes*:

$$P(H|X) = \frac{p(X|H|P(H)}{P(X)}$$

Keterangan:

X = Data dengan *class* yang belum diketahui.

H = *Hipotesis* data X merupakan suatu *class spesifik*.

P(H|X) = *Probabilitas hipotesis* H berdasarkan kondisi X (*posteriori probability*).
P(H) = *Probabilitas Hipotesis* H (*prior probability*).
P(X|H) = *Probabilitas* X berdasar kondisi pada *Hipotesis* H

P(X) = Probabilitas dari X.

Metode algoritna Naïve bayes merupakan penyederhanaan *metode bayes*. Untuk mempermudah pemahaman, maka *Teorema Bayes* disederhanakan menjadi:

P(H|X) = P(X|H) P(X)
Metode Bayes rule digunakan dan diterapkan untuk melakukan penghitungan terhadap *posterior* dan *probabilitas* dari data sebelumnya. Dalam analisis *bayesian*, fungsi klasifikasi akhir dihasilkan dengan menggabungkan kedua sumber informasi (*prior dan posterior*) untuk menghasilkan probabilitas menggunakan aturan *bayes*.
Menurut(N.A. Zaidi, dkk, 2013) Naïve Bayes memiliki kelemahan yaitu attribut atau fitur independen sering salah dan hasil estimasi probabilitas tidak dapat berjalan optimal. Untuk mengatasi kelemahan tersebut salah satu caranya dengan metode pembobotan attribut untuk meningkatkan akurasi dari Naïve Bayes tersebut.

3. *Fitur Selection*

Menurut (Kusuma, 2003) *Feature selection* adalah sebuah proses yang bisa digunakan pada *machine learning* dimana sekumpulan dari *features* yang dimiliki data digunakan untuk pembelajaran algoritma. *Subset* yang baik memiliki sedikitnya dimensi angka yang paling banyak berkontribusi untuk akurasi dan nantinya akan dibuang sisa dari dimensi yang tidak berkepentingan, ini merupakan langkah penting dalam tahap *preprocessing*. Menurut (Kusuma, 2003) *Forward selection* dimulai tanpa variabel dan menambahkan mereka satu persatu, pada setiap langkah ditambahkan variable yang menurunkan *error* paling banyak, sampai semua *error* dihilangkan. Menurut (Kusuma, 2003) *Backward sele ction* dimulai dengan semua variabel dan membuangnya satu persatu, pada setiap langkah membuang variable yang memiliki error paling banyak.

4. Algoritma Genetika

Menurut (Zukhri, 2014) Optimasi adalah proses menyelesaikan suatu masalah tertentu supaya berada pada kondisi yang paling menguntungkan dari suatu sudut pandang. Masalah yang harus diselesaikan berkaitan erat dengan data-data yang dapat dinyatakan dalam satu atau beberapa variabel. Pengertian menguntungkan, biasanya berhubungan dengan pencarian nilai minimum atau pencarian nilai maksimum, bergantung pada sudut pandang yang digunakan. Menurut (Suyanto,2007) Algoritma genetika (AG) diperkenalkan pertama kali oleh John Holland (1975) dari Universitas Michigan, John Holland mengatakan bahwa setiap masalah yang berbentuk adaptasi(alami maupun buatan) dapat diformulasikan ke dalam terminologi genetika. Algoritma genetika merupakan suatu algoritma pencarian berdasarkan pada mekanisme seleksi alam dan genetika alam. Algoritma genetika dimulai dengan sekumpulan solusi awal(individu) yang disebut populasi. Satu hal yang sangat penting adalah bahwa satu individu menyatakan satu solusi. Populasi awal akan berevolusi menjadi populasi baru melalui serangkaian iterasi (generasi). Menurut (Anita, 2006) pada akhir iterasi, algoritma genetika mengembalikan satu anggota populasi yang terbaik sebagai solusi untuk masalah yang dihadapi. Pada setiap iterasi, proses evolusi yang terjadi adalah sebahai berikut:

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

Proses ini terus berulang sampai kondisi berhenti *(stopping condition)* tertentu. Kondisi berhenti bisa berupa jumlah iterasi tertentu, waktu tertentu, atau ketika variansi individu-individu dalam populasi tersebut sudah lebih kecil dari suatu nilai tertentu yang diinginkan.

Berikut ini contoh aplikasi algoritma genetika yang digunakan untuk menyelesaikan masalah kombinasi. Misalkan ada persamaan:

“a+3b+2c=30”

Untuk menyelesaikan permasalahan persamaan diatas dapat dilakukan dengan algoritma genetika, berikut langkah-langkah penyelesaiannya:

a) Langkah 1 adalah menentukan populasi awal

C1 :[15_2_7]
C2 :[20_3_1]
C3 :[10_5_8]
C4 :[24_8_9]
C5 :[17_4_6]

b) Langkah 2 adalah evaluasi nilai fitnes Dengan rumus yang sebelumnya telah ditentukan yaitu a+3b+2c=30, maka akan menghasilkan nilai *fitnes* pada setiap *cromosom,*

Fitnes C1 = 1/(1 + | (a+3b+2c) – 30 |)
= 1/(1 + | ((15)+(3*2)+(2*7)-30|)
= 1/(1 + 8)
= 1/9
= 0,111

Fitnes C2 = 1/(1 + | (a+3b+2c) – 30 |)
= 1/(1 + | ((20)+(3*3)+(2*1)-30|)
= 1/(1 + 1)
= 1/2
= 0,500

Fitnes C3 = 1/(1 + | (a+3b+2c) – 30 |)
= 1/(1 + | ((10)+(3*5)+(2*8)-30|)
= 1/(1 + 11)
= 1/2
= 0,083

Fitnes C4 = 1/(1 + | (a+3b+2c) – 30 |)
= 1/(1 + | ((24)+(3*8)+(2*9)-30|)
= 1/(1 + 36)
= 1/36
= 0,027

Fitnes C5 = 1/(1 + | (a+3b+2c) – 30 |)
= 1/(1 + | ((17)+(3*4)+(2*6)-30|)
= 1/(1 + 10)
= 1/10
= 0,100

Total nilai fitnes adalah 0,111+0,500+0,083+0,027+0,100 = **0,821**

Probabilitas masing-masing cromosom menjadi:
P[1]= 0,111/0,821= 0,135
P[2]= 0,500/0,821= 0,609
P[3]= 0,083/0,821= 0,101
P[4]= 0,027/0,821= 0,032
P[5]= 0,100/0,821= 0,121
Dari hasil probabilitas tertinggi, dihasilkan bahwa *cromosom* 2 mempunyai nilai *fitness* paling tinggi. Maka *cromosom* 2 juga mempunyai kesempatan paling besar dalam proses seleksi selanjutnya dengan *Roullete Wheel*.

c) Penentuan *Cromosom* Induk
Untuk proses seleksi digunakan *Roullete Wheel*, untuk itu diperlukan nilai kumulatif probabilitasnya dari setiap kromosom, yakni sebagai berikut:

C1 = 0,135
= 0,135
C2 = 0,135 + 0,609
= 0,744
C3 = 0,135 + 0,609 + 0,101
= 0,845
C4 = 0,135 + 0,609 + 0,101 + 0,032 = 0,877
C5 = 0,135 + 0,609 + 0,101 + 0,032 + 0,121 = 1

Langkah selanjutnya adalah dengan menggunakan bilangan acak R antara 0 sampai dengan 1, bilangan acak dipilih sesuai dengan jumlah *cromosom:*
R[1]= 0,234
R[2]= 0,451
R[3]= 0,508
R[4]= 0,134
R[5]= 0,680
Memilih kromosom ke x sebagai *parent* dengan syarat C[x-1] < R < C[x]. Angka acak R[1] < nilai kumulatif dari C2, sehingga C2 nanti akan dilakukan *crossover* dengan C1. Hasil seleksi Roullete Wheel pada populasi ini untuk *crossover* menjadi:
C1 Menjadi C2 = [20_3_1]
C2 Menjadi C2 = [20_3_1]
C3 Menjadi C3 = [10_5_8]
C4 Menjadi C4 = [24_8_9]
C5 Menjadi C5 = [17_4_6]

d) Perkawinan silang atau Crossover
Dalam *crossover* kita menentukan Probability (pr), yaitu sebesar 0.5 atau 50%. Hanya kromosom yang nilai R lebih kecil dari 0.5 yang akan bermutasi. Maka Kromosom ke y akan dipilih menjadi induk jika R[y] < pr, dari bilangan acak R diatas maka yang dijadikan parent adalah C1, C2, dan C4 . Sedangkan C3 dan C5 > 0.5, sehingga tidak dilakukan seleksi. Selanjutnya setelah melakukan pemilihan parent, dilanjutkan menentukan kromosom yang akan di lakukan perkawinan silang dengan pengambilan sejumlah atribut 1-3. Dalam hal ini posisi cut-point (cp) dipilih menggunakan bilangan acak dari 1-3 sesuai banyaknya crossover yang terjadi. Misalnya didapatkan posisi crossover adalah 3, maka kromosom parent akan dipotong mulai gen ke 3

kemudian potongan gen tersebut saling ditukarkan antar parent.

C1 >< C2
cp(C1) = 2
cp(C2) = 3
cp(C4)= 1
nilai Offspring[1] = C1 >< C2 dengan cp(C1)

[20_3_1]><[20_3_1]
[20_3_1]

nilai Offspring[2] = C2 >< C2 dengan cp(C2)

= [20_3_1]>< [20_3_1]
= [20_3_1]

nilai Offspring[4] = C4 >< C4 dengan cp(C4)

= [24_8_9]>< [24_8_9]
= [24_8_9]

Sehingga Populasi baru yang dihasilkan dari *crossover* adalah:

C1 = [20_3_1]
C2 = [20_3_1]
C3 = [10_5_8]
C4 = [24_8_9]
C5 = [17_4_6]

e) Mutasi Kromosom

Jumlah kromosom yang mengalami mutasi dalam satu populasi ditentukan oleh persentase p mutation. Proses mutasi dilakukan dengan cara mengganti satu gen yang terpilih secara acak dengan suatu nilai baru yang didapat secara acak.

Total gen = (gen dalam kromosom) * jumlah kromosom

= 3 * 5
= 15

Tentukan posisi gen yang akan mengalami mutasi dengan menggunakan bilangan acak antara 1 sampai dengan total gen, yaitu antara 1 sampai 15. Misalkan ρm kita tentukan 10% maka jumlah gen yang mengalami mutasi adalah 10% dari 15 yaitu 1,5 atau 1 gen.

Kemudian gunakan bilangan acak dari total gen misalkan yang terpilih adalah posisi gen 6 yang akan mengalami mutasi. Dengan demikian yang akan mengalami mutasi adalah kromosome ke-2 gen nomor 3. Maka nilai gen pada posisi tersebut akan diganti dengan bilangan acak 0-30. Misalkan bilangan acak yang digunakan adalah 3, maka kromosom ke-2 berubah menjadi [20_3_3]Populasi pada generasi pertama menjadi:

C1 = [20_3_1]
C2 = [20_3_3]
C3 = [10_5_8]
C4 = [24_8_9]
C5 = [17_4_6]

Kromosom ke-2 kemudian di uji pada rumus a+3b+2c=30, agar menjadi kromosom yang ingin dicapai. (20)+(3*5)+2*1 <> 30, maka populasi ini belum memiliki kromosom yang ingin dicapai. Kromosom-kromosom pada populasi ini akan mengalami proses yang sama seperti generasi sebelumnya yaitu proses evaluasi, seleksi, *crossover* dan mutasi yang kemudian akan menghasilkan kromosom-kromosom baru untuk generasi yang selanjutnya. Proses ini akan berulang sampai sejumlah generasi yang telah ditetapkan sebelumnya. Yang telah dipaparkan adalah bagaimana algoritma genetika bekerja dalam menseleksi atribut yang akan digunakan oleh algoritma naive bayes dalam memprediksi tingkat kesuburan. Pada penelitian sebelumnya, Hairil Kurniadi dalam publikasi tesisnya pada tahun 2014 menggunakan Metode Naive Bayes untuk memprediksi tingkat kesuburan. Selain itu, Macmillan Simfukwe dkk juga melakukan penelitian untuk mengetahu tingkat kesuburan(fertility) dengan menggunakan Naive Bayes dan Nerual Network.

5. Evaluasi Penelitian

Evaluasi adalah kunci ketika membuat aplikasi berbasis *data mining.* Ada berbagai macam cara dalam melakukan evaluasi. Jika kita memiliki data yang kita gunakan dalam proses pelatihan, maka tidak serta merta menjadikan data tersebut sebagai indikator keberhasilan aplikasi yang kita buat.oleh karena itu kita membutuhkan metode untuk tertentu guna memprediksi performa berdasarkan eksperimen untuk berbagai macam data selain data *traning* (Widodo, 2013). Dalam penelitian ini evaluasi penelitian dilakukan dengan menggunakan *Confussion matrix* dan Kurva ROC

a) *Confussion Matrix*

Confussion matrix adalah sebuah metode untuk melakukan evaluasi dengan menggunakan tabel matrix. Pada tabel 2 dapat dilihat bahwa jika *dataset* terdiri dari dua *class*, dimana *class* yang satu dianggap sebagai *class positif* dan *class* yang lainnya dianggap sebagai *class negatif* (Bramer, 2007). Evaluasi dengan mengunakan fungsi *confusion matrix* akan mengahsilkan nilai *accuracy, precision,* dan *recall*.

Menurut (Han & kamber, 2006) Nilai *accuracy* adalah *presentase* dari jumlah

record data yang diklasifikasikan secara baik dan benar dengan menggunakan sebuah algoritma dan dapat membuat klasifikasi setelah dilakukan pengujian pada hasil klasifikasi tersebut. Nilai *precison* atau yang juga dikenal dengan nama *confidence value* merupakan *proporsi* dari jumlah kasus yang diprediksi mendapatkan hasil positif dimana nilainya juga positif pada data yang sebenarnya. Menurut (Powers, 2011) nilai dari *Recall* atau *sensitivity value* merupakan *proporsi* dari jumlah kasus yang bernilai positif yang sebenarnya dan diprediksi positif secara benar.

Tabel 1. Model *Confussion Matrix*

Correct Claassification	*Classified as*	
	Positive	*Negative*
Positive	*True Positives*	*False Negatives*
Negative	*False Negatives*	*True Negatives*

Model *Confussion Matrix* pada tabel 1, dapat dijelaskan sebagai berikut, dimana *True Positive* merupakan jumlah *record positive* yang diklasifikasikan sebagai *positive*, *false positive* adalah jumlah *record negative* yang diklasifikasikan sebagai *positive, false negative* adalah jumlah *record positive* yang diklasifikasikan sebagai *negative*, *true negative* adalah jumlah *record negative* yang diklasifikasikan sebagai *negative*, kemudian masukkan data uji. Setelah data uji dimasukkan ke dalam *Confussion Matrix*, hitung nilai-nilai yang telah dimasukkan tersebut untuk dihitung jumlah *sensitivity* (*recall*), *Specifity, Precission*, dan *Accuracy.* Selanjutnya *Sensitivity* digunakan untuk membandingkan jumlah t_pos terhadap jumlah *record* yang *positive* sedangkan *Specifity, Precision* adalah jumlah t_neg terhadap jumlah *record* yang *negative*. Berikut persamaan dari *confussion matrix*.(Han& Kamber).

Sensitivity = $\frac{t_{pos}}{pos}$

Specifity = $\frac{t\_neg}{neg}$

Precision= $\frac{t\_pos}{t_{pos} + f\_pos}$

Accuracy= $Sensitify\ \frac{pos}{pos+neg} + specify\ \frac{neg}{pos+neg}$

Keterangan:

t_pos : jumlah *true positive*
t_neg : jumlah *true negative*
p : jumlah *record positive*
n : jumlah *tupel negative*
f_pos : jumlah *false positive*

b) Kurva ROC

Fungsi Kurva *ROC* adalah untuk menunjukkan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan *Confusion Matrix*, ROC adalah grafik dua dimensi dengan *false positives* sebagai garis *horizontal dan true positive* sebagai garis *vertical* (Vercellis, 2009). Dalam masalah klasifikasi menggunakan dua kelas keputusan (*klasifikasi biner*), masing-masing objek dikelompokkan dalam (P, N), yaitu positif dan negatif. Selain itu ada beberapa model klasifikasi (seperti pohon keputusan) menghasilkan *label class diskrit* (hanya menunjukkan class yang diprediksi oleh objek), kalsifikasi yang lain seperti *Naïve Baiyes* dan *Neural Network* juga menghasilkan output yang berkesinambungan, dimana ambang batas yang berbeda mungkin diterapkan untuk memprediksi keanggotaan *class*. Secara teknis *kurva ROC* juga dikenal sebagai grafik *ROC*, dua dimensi grafik dimana tingkat TP diplot pada sumbu Y dan tingkat FP diplot pada sumbu X (Gorunescu, 2011). Hasil perhitungan dapat divisualisasikan dengan kura ROC (*Receiver Operating Characteristic)* atau AUC (*Area Under Curve*). Berikut tingkat nilai diagnosa dari ROC, yaitu: (Gorunescu, 2011).

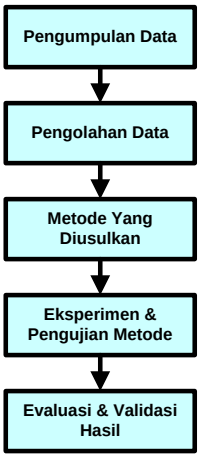
1) Akurasi bernilai 0.90 – 1.00 = *Excellent classification*
2) Akurasi bernilai 0.80 – 0.90
   a. = *Good classification*
3) Akurasi bernilai 0.70 – 0.80
   a. = *Fair classification*
4) Akurasi bernilai 0.60 – 0.70
   a. = *Poor classification*
5) Akurasi bernilai 0.50 – 0.60 = *Failure*

C. METODE PENELITIAN

Jenis penelitian yang dilakukan adalah model eksperimen dalam bentuk sistem penunjang keputusan untuk prediksi tingkat kesuburan menggunakan algoritma Naive Bayes yang dioptimasi dengan algoritma genetika. Jenis data yang digunakan dalam

penelitian ini adalah data primer dan data sekunder.mPenelitian dilakukan dengan menggunakan tools aplikasi rapidminer 5. Langkah-langkah dalam penelitian ini antara lain adalah:

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.
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. Berikut gambar langkah-langkah dalam tahapan penelitian:



Gambar 1. Tahapan Penelitian

D. HASIL DAN PEMBAHASAN

Tahapan analisa pembuatan model algoritma *Naïve Bayes*, langkah awal yang harus dilakukan adalah terlebih dahulu mencari nilai *probabilitas* hipotesis untuk tiap-tiap *class* P(H). Hipotesis dilakukan terhadap data yang terdiri dari 105 *record*. Hipotesis yang didapatkan yaitu pengklasifikasian data menjadi Normall dan Tidak Normal. Eksperimen yang penulis lakukan dalam penelitian ini adalah dengan menghitung *probabilitas Prior* dan *probabilitas posterior* dengan menggunakan data sebanyak 160 *record*.

Penelitian ini menggunakan metode K-fold *Cross Validation* untuk membagi data training dan data testing. Data dibagi menjadi 2 bagian yaitu sebesar 90% dataset untuk metode pelatihan dan 10% dataset akan digunakan untuk metode pengujian.

1. Menghitung Probabilitas Prior.

Setelah mendapatkan jumlah data yang akan diolah, maka selanjutnya adalah menghitung *Probabilitas Prior* dan *Probabilitas posterior*, dalam bentuk persamaan dibawah ini:

Total data yang akan di olah = 160
Data Normal = 140
Data Tidak Normal = 20
P(Normal) = 140/160 = 0,875
P(Tidak Normal)= 20/160 = 0,125

Setelah didapatkan nilai *probabilitas* untuk tiap hipotesis dari *class*, maka langkah selanjutnya adalah melakukan penghitungan terhadap kondisi *probabilitas* tertentu (*Probabilitas X*) dengan menggunakan data berdasarkan *probabilitas* tiap hipotesis (*Probabilitas H*) atau yang dinamakan dengan *probabilitas Prior*. Selanjutnya untuk mengetahui hasil perhitungan dari *Probabilitas Prior*, maka dilakukan penghitungan dengan cara merinci jumlah kasus dari tiap-tiap atribut variabel data, adapun hasil perhitungan *probabilitas prior* dengan menggunakan *Algoritma Naïve Baye*. Setelah melakukan tahapan pertama yaitu menentuka probabilitas prior maka akan di hasilkan *class, class* yang dihasilkan dari data diagnosa tingkat kesburuan adalah class Normal dan class Tidak Normal. Yang selanjutnya class tersebut adalah hasil dari prediksi tingkat kesuburan.

2. Menghitung Probabilitas Postereor.

Tahapan selanjutnya adalah menghitung *Probabilitas Postereor* untuk menentukan *class* terhadap temuan kasus baru, dengan cara terlebih dahulu menghitung *Probabilitas Posterior*nya, hal tersebut dilakukan apabila ditemukan kasus baru dalam pengolahan data. Berikut tabel *probabilitas posterior* untuk menghitung kasus baru yang ditemukan:

Table 2. Perhitungan *Probabilitas Postereor*

Data X		P(X|CI	
Attribut	**Nilai Value**	**Normal**	**Tidak Normal**
Cuaca	Hujan	0,806	0,194
Usia	24-29	0,925	0,075
Sakit Bawaan	Tidak	0,852	0,148
Kecelakaan/ Trauma	Tidak	0,908	0,092
Intervensi Bedah	Tidak	0,989	0,124
Radiasi	Tidak Ada	0,958	0,042
Konsumsi Alkohol	Tidak Pernah	0,960	0,040
Kebiasaan Merokok	Sesekali	0,844	0,156
Lama Duduk	12-16 Kali	0,750	0,250

Setelah mengetahui nilai *probabilitas* dari setiap atribut terhadap *probabilitas* tiap *class* atau yang dirumuskan dalam bentuk persamaan P(X|Ci), maka langkah berikutnya adalah melakukan penghitungan terhadap total keseluruhan *probabilitas* tiap *class*. Berikut persamaan untuk menghitung probabilitas tiap *class*:

P(X|Kelas =Normal)
=0,806+0,925+0,852+0,908+0,989+0,958+0,960+0,844+0,750 = 7,991

P(X|Kelas =Tidak Normal)
=0,194+0,075+0,148+0,092+0,124+0,042+0,040+0,156+0,250 = 1.121

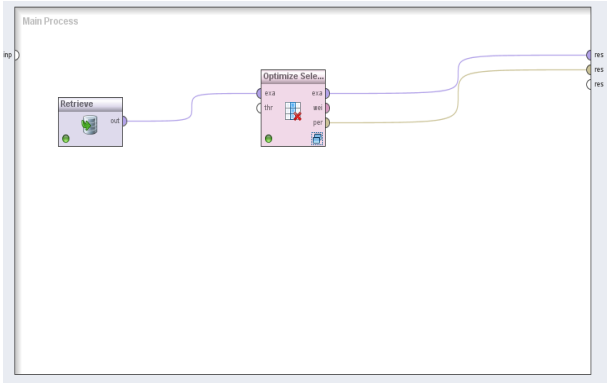
P(X|Kelas= Normal P(Normal)
= 7,991 x 0,875 = 6,992

P(X|Kelas= Tidak Normal) P(Tidak Normal)
= 5,044 x 0,125 = 0,140

Hasil perhitungan terhadap *probabilitas* tiap *class* diatas, diketahui bahwa nilai P(X|Normal) lebih besar daripada nilai P(X|Tidak Normal), sehingga dapat diambil kesimpulan bahwa dalam kasus prediksi tingkat kesuburan tersebut akan masuk kedalam klasifikasi Normal dan tidak termasuk kedalam klasifikasi Tidak Normal.

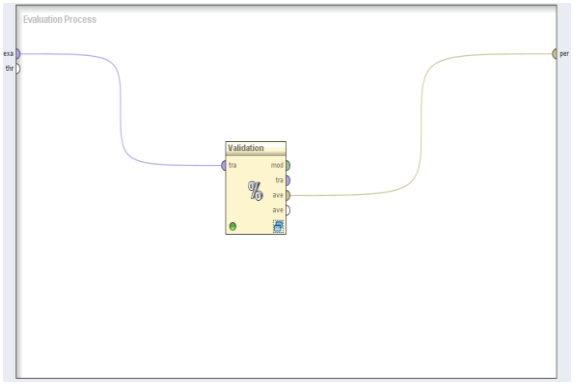
3. Pengujian Menggunakan Naïve Bayes dan dioptimasi dengan Algoritma Genetika.

Berikut adalah gambar K-fold validation untuk model algoritma Naive Bayes berbasis atau dioptimasi dengan menggunakan algoritma genetika.



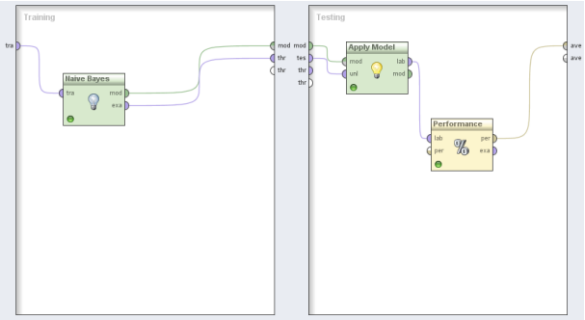
Gambar 2. Koneksi Data dengan *Optimaize Selection*

Database Diagnosa Tingkat Kesuburan dihubungkan dengan operator *optimize selection (evolutionary)* untuk dilakukan pemilihan attribut-attribut yang relevan dengan proses prediksi hasil tingkat kesuburan. Di dalam *optimaize selection (evalutionary)* terdapat proses *cross validation* seperti yang terlihat pada Gambar 3.



Gambar 3. Penggunaan Operator *Cross Validation*

Cross Validation yang digunakan dalam penelitian ini adalah *10-fold validation.* Dataset yang berjumlah 160 data dengan 9 atribut akan dibagi menjadi 10 bagian. Dimana setiap bagian akan dibentuk secara acak. Prinsip adari 10-fold validation adalah 1:9, dimana 1 bagian menjadi data testing dan 9 bagian menjadi data traning, sehingga 10 bagian tersebut dapat berkesempatan menjadi data testing. Setelah dilakukan traning dan testing maka dapat diukut tingkat akurasinya. Di dalam *cross validation* terdapat proses penerapan algoritma naïve bayes seperti yang terdapat pada gambar 4.



Gambar 4. Penggunaan Model Naïve Bayes

4. Hasil akurasi dari pengujian Naïve Bayes yang dioptimasi dengan Algoritma Genetika.

Setelah dilakukan penerapan model algoritma naïve bayes dengan optimasi menggunakan algoritma genetika maka hasil dapat dilihat pada Table 2.

Table 3. *Confussion Matrix*

	True Tidak Normal	**True Normal**	***Class Precision***
Tidak Normal	0	0	0,00
Normal	1	140	99,29 %
Class recall	0,00%	100%	
Accuracy	**99,33 %**		

Akurasi = $\frac{(TN+TP)}{(TN+FN+TP+FP}$ = $\frac{1+140}{0+1+140+0}$ =100 %

Precision = $\frac{(TP)}{(TP+FP)}$ = $\frac{140}{140+1}$

= 0,9929=99,33 %

Recall = $\frac{(TP)}{(TP+FN)}$ = $\frac{140}{140+0}$

= 1= 100%

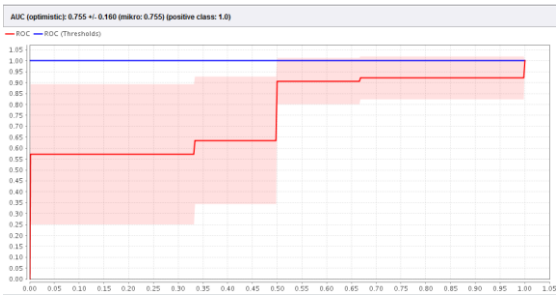
f- Measure = $\frac{2*recall*precision}{(recall+precision)}$ = $\frac{2*1*0,9929}{1+1}$

=99,33 %

Penelitian sebelumnya dilakukan oleh Hairil Kurniadi (2014) menggunakan data yang sama dan menggunakan Naive Bayes hasil akurasinya adalah 97,66% sedangkan setelah dioptimasi dengan menggunakan algoritma genetika hasil akurasinya meningkat menjadi 99,33%.

5. Evaluasi dengan kurva ROC.

Berikut ini adalah kurva AUC dengan menggunakan algoritma Naïve Bayes yang di optimasi dengan algoritma genetika.

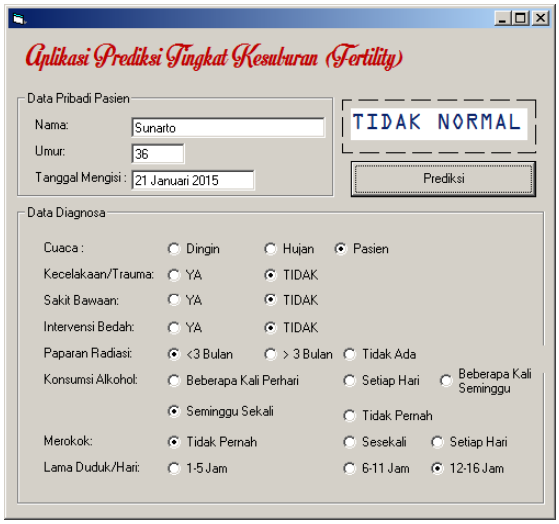


Gambar 5. Kurva AUC

Pada Gambar 3. menunjukkan grafik ROC dengan nilai AUC (*Area Under Curve*) sebesar 0.75 dengan tingkat diagnosa *Fair Classification* untuk untuk model naive bayes berbasis algoritma genetika.

6. Aplikasi Prediksi Tinggkat Kesuburan

Berikut adalah applikasi yang diperoleh dari hasil algoritma Naïve Bayes dan ALgoritma Genetika untuk memprediksi Tingkat Kesubutan.



Gambar 6. Aplikasi Prediksi Tingkat Kesuburan *(Fertility)*

E. KESIMPULAN

Penelitian ini membuktikan bahwa gaya hidup dapat mempengaruhi tingkat kesuburan*(Fertility)*. Menambahkan Algoritma Genetika untuk meningkatkan hasil prediksi yang dilakukan oleh Naïve Bayes terbukti dapat meningkatkan prediksi dimana penelitian sebelumnya yang dilakukan oleh Hairl Kurniadai (2014) dengan tingkat akurasi prediksi kesuburan(Fertility) adalah 97,66% menjadi 99,33%.

DAFTAR PUSTAKA

[1] Irvine DS.(1998). Epidemiology and aetiology of male infertility. Hum. Reprod.;Vol 13(1):33-44.

[2] Irvine, D. S. (2000). *Male reproductive health: Cause for concern?* Andrologia, 32(4–5), 195–208.

[3] Carlsen, E., Giwercman, A., Keiding, N., & Skakkebaek, N. E. (1992). *Evidence for decreasing quality of semen during past 50 years. BMJ*, 305(6854), 609–613.

[4] Martini, A. C., Molina, R. I., Estofan, D., Senestrari, D., Fiol de Cuneo, M., & Ruiz, R. D. (2004). *Effects of alcohol and cigarette consumption on human seminal quality. Fertility and Sterility*, 82(2), 374–377.

[5] WHO (1999). WHO laboratory manual for the examination of human semen and sperm-cervical mucus interaction (4th ed.). Published on behalf of the World Health Organization by Cambridge University Press, Cambridge, UK.

[6] Rowe P.J., Comhaire F.H. (2000). WHO manual for the standardized investigation, diagnosis and management of the infertile male, Cambridge University Press

[7] Kusrini., E. T. Luthfi. (2009). Algoritma Data Mining 1st ed. Yogyakarta, Indonesia: Andi.

[8] Wahyuni, Diana Tri, Sutojo, T,Luthfiarta, Ardytha. (2014). Prediksi Hasil Pemilu Legislatif DKI Jakarta Menggunakan Naïve Bayes dengan Algoritma sebagai Fitur Seleksi.

[9] Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning and Tools.* Burlington: Morgan Kaufmann Publisher.

[10] Gorunescu, F. (2011). *Data Mining Concepts, Model and Technique*. Berlin: Springer.

[11] Han, J., & Kamber, M. (2006). *Data Mining Concepts and technique*. San Francisco: Diane Cerra.

[12] Bramer, Max. (2007). *Principles of Data Mining*.London: Springer. ISBN-10: 1-84628-765-0, ISBN-13: 978-1-84628-765-7.

[13] N. A. Zaidi, J. Cerquides, M. J. Carman, and G.I.Webb.(2013). "Alleviating Naive Bayes Attribute Independence Assumption byAttributeWeighting," *Journal of Machine Learning Research*, no. 14, pp. 1947-1988.

[14] Kusuma dewi, Sri. (2003). Artificial Intelligent.Yogyakarta : Graha Ilmu.

[15] Zukhri, Zainudin.(2014). Algoritma Genetika Metode Komputasi untuk Menyelesaikan Maslah Optimasi. Yogyakarta: Andi Offset.

[16] Suryanto, MT, Msc. (2007). Artificial Intelligent, Searching, Reasoning Planning dan Learning. Bandung: Informatika Bandung.

[17] Anita, Desiani, Arhami Muhammad. (2006). Konsep Kecerdasan Buatan. Yogyakarta: Cv. Andi Offset.

[18] Kurniadi, Hairil.(2014). Prediksi Tingkat Kesuburan*(Fertility):*STMIK Nusa Mandiri

[19] Gourunescu dalam Widodo, Prabowo Pudjo. (2013). Penerapan *Data Mining* dengan Matlab.Bandung: Rekayasa Sains.

[20] Bramer, Max. (2007). *Principles of Data Mining*.London: Springer. ISBN-10: 1-84628-765-0, ISBN-13: 978-1-84628-765-7.

[21] Powers, D.M.W. (2011). *Evaluation: From Precision, Recall and F-Measure to ROC*, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies, ISSN: 2229-3981 & ISSN: 2229-399X, Volume 2, Issue 1, 2011, pp-37-63.

[22] Vercellis, Carlo. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making.* United Kingdom: John Willey & Son

[23] *Anik Andriani,* Sistem Prediksi Penyakit Diabetes Berbasis Decision Tree, Vol 1, No 1 (2013): Bianglala 2013