Analisis Penerapan Optimasi Perbandingan Kinerja Algoritma C4.5 dan Naïve Bayes Berbasis Particle Swarm Optimization (PSO) Untuk Mendeteksi Kanker Payudara

Taghfirul Azhima Yoga Siswa[1], Prihandoko[2]

[1]taghfirul.yoga@yahoo.co.id

[2]pri@staff.gunadarma.ac.id

[1] *Magister Teknik Informatika Universitas Amikom Yogyakarta, Jl. Ring Road Utara, Condong Catur, Sleman 55283, Indonesia*

[2] *Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Gunadarma, Jl. Margonda Raya 100, Depok 1642, Indonesia*

Intisari— **Perlu dilakukan upaya pencegahan untuk meningkatkan kesadaran masyarakat dalam mengenali gejala dan risiko penyakit kanker payudara sehingga dapat menentukan langkah-langkah pencegahan dan deteksi dini yang tepat. Sejalan dengan hal itu data mining merupakan salah satu pemanfaatan teknologi informasi dalam bidang kesehatan yang banyak digunakan sebagai sistem pendukung keputusan klinis dalam memprediksi dan mendiagnosa berbagai penyakit dengan akurasi data yang sangat baik. Penelitian ini bertujuan mengevaluasi perbandingan penerapan optimasi kinerja terbaik metode klasifikasi data mining algoritma C4.5 dan Naïve Bayes berbasis Particle Swarm Optimization (PSO) untuk mendeteksi kanker payudara menggunakan pengukuran confusion matrix, AUC dan T-Test. Dataset kanker payudara yang digunakan berjumlah 699 record dengan 11 parameter indikator yang terdiri dari Code Number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses, dan Class yang diolah menggunakan software RapidMiner Versi 9. Hasil penelitian ini didapatkan bahwa optimasi Particle Swarm Optimization (PSO) dapat meningkatkan kinerja akurasi C4.5 dari 90,19% menjadi 94,29% dan Naïve Bayes 97,65% menjadi 97,96%. Hasil kinerja terbaik yang diuji menggunakan T-Test adalah algoritma Naïve Bayes (PSO) memiliki nilai tertinggi sebesar 0,980. Dengan demikian algoritma Naïve Bayes berbasis Particle Swarm Optimization (PSO) dapat memberikan solusi terbaik terhadap akurasi pendeteksian penyakit kanker payudara.**

Kata kunci— Data Mining, Klasifikasi, C.45, Naïve Bayes, K-Nearest Neighbor, Paticle Swarm Optimization.

Abstract— **Breast cancer is a very frightening cancer for the world around the world, this also applies in Indonesia. In line with this, data mining has been one of the uses of information technology in the health sector which is a supporter of results in predicting and diagnosing various diseases with excellent data accuracy. This study aims to do the work of data mining algorithm C4.5 and Naïve Bayes based on Particle Swarm Optimization (PSO) to detect cancer using a confusion matrix, AUC and T-Test. The breast cancer dataset is 699 records with 11 indicator parameters consisting of Code Number, Clot Thickness, Cell Size Uniformity, Cell Shape Uniformity, Marginal Adhesion, Single Epithelial Cell Size, Nuclear Bare, Bland Chromatin, Normal Nucleolus, Mitosis, and Classes processed using RapidMiner Version 9 software. The results of this study found that Particle Swarm Optimization (PSO) Optimization can improve C4.5 accuracy performance from 90.19% to 94.29% and Naïve Bayes 97.65% to 97.96% Results the best performance using the T-Test is the Naïve Bayes (PSO) algorithm which has the highest value of 0.980. Thus the Naïve Bayes algorithm based on Particle Swarm Optimization (PSO) can provide the best solution for detecting breast cancer.**

Keywords— Data Mining, Klasifikasi, C.45, Naïve Bayes, K-Nearest Neighbor, Paticle Swarm Optimization

I. PENDAHULUAN

Kanker yang diketahui sejak dini memiliki kemungkinan untuk mendapatkan penanganan lebih baik. Oleh karena itu, perlu dilakukan upaya pencegahan untuk meningkatkan kesadaran masyarakat dalam mengenali gejala dan risiko khususnya penyakit kanker payudara sehingga dapat menentukan langkah-langkah pencegahan dan deteksi dini yang tepat. Saat ini kanker payudara menjadi jenis kanker yang sangat menakutkan bagi perempuan diseluruh dunia, hal ini juga berlaku di Indonesia. Kanker payudara adalah tumor ganas yeng terbentuk dari sel - sel payudara yang tumbuh dan berkembang tanpa terkendali sehingga dapat menyebar di antara jaringan atau organ di deket payudara atau ke bagian tubuh lainya.

Pendekatan data mining menjadi sangat penting dalam industri kesehatan dalam mengambil keputusan berdasarkan analisis dari data klinis yang besar. Teknik data mining yang digunakan sebagai sistem pendukung keputusan klinis dalam memprediksi dan mendiagnosa berbagai penyakit dengan akurasi data yang sangat baik [1]. Salah satu

penggunaan data mining dalam proses eksplorasi data adalah teknik pengklasifikasian data.

Beberapa penelitian 3 tahun terakhir terkait data mining dengan teknik klasifikasi yang membahas tentang diagnosa penyakit diantaranya antara lain : Klasifikasi Jenis Kanker Berdasarkan Struktur Protein Menggunakan Algoritma Naive Bayes [2], Sistem Pakar Diagnosis Penyakit Demam: DBD, Malaria dan Tifoid Menggunakan Metode K-Nearest Neighbor – Certainty Factor [3], Sistem Klasifikasi Penyakit Asma Menggunakan Algoritma Naïve Bayes (Studi Kasus : Puskesmas Sungai Salak) Menggunakan Algoritma Naive Bayes [4], Penerapan Algoritma C4.5 Berbasis Adaboost Untuk Prediksi Penyakit Jantung [5], Klasifikasi Penyakit Stroke Menggunakan Metode Naive Bayes Classifier (Studi Kasus Pada Rumah Sakit Umum Daerah Undata Palu) [6] dan Model Data Mining sebagai Prediksi Penyakit Hipertensi Kehamilan dengan Teknik Decision Tree [7].

Beberapa penelitian yang berkaitan dengan perbandingan kinerja algoritma diantaranya Analisis kinerja teknik klasifikasi data mining dengan melakukan perbandingan hasil eksperimen untuk berbagai teknik klasifikasi data mining Naïve Bayes, Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), dan Decision Tree menggunakan WEKA. Menghasilkan Algoritma terbaik Multilayer Perceptron classifier (ANN) dengan akurasi 97,33%. Hasil ini membuktikan bahwa algoritma learning machine memiliki potensi secara signifikan meningkatkan lebih dari metode klasifikasi konvensional [8].

Perbandingan klasifikasi data mining algoritma (Naïve Bayes dan C4.5) dalam mengelola data transaksi penjualan POS (Pont Of Sales) [9]. Hasil penelitian ini algoritma C4.5 bekerja mengelompokkan beberapa data sampel pelatihan yang akan menghasilkan pohon keputusan berdasarkan fakta pada data pelatihan. Sedangkan Bayes, keputusan diperoleh berdasarkan pengalaman yang ada pada peristiwa sebelumnya. Bayes menghitung kejadian yang terjadi dalam data menjadi sampel untuk menentukan keputusan tentang masalah yang dihadapi.

Perbandingkan dua metode dalam data mining, yaitu metode Logistic Regresi dan metode Bayesian, untuk memprediksi tingkat risiko diabetes dengan aplikasi berbasis web dan sembilan atribut data pasien [10]. Hasil penelitian Logistic Regresi dan Bayesian, memiliki kelebihan skor kinerja yang berbeda dan baik pada keduanya. Dari pengukuran akurasi tertinggi dan ROC menggunakan dataset yang sama, di mana kelebihan Bayesian memiliki akurasi tertinggi dengan skor 0,91. Selain itu skor ROC metode Regression Logistic memiliki akurasi tertinggi dengan skor 0,988, sedangkan pada Bayesian 0,964.

Komparasi optimasi algoritma klasifikasi data mining C4.5 dan Naïve Bayes berbasis Particle Swarm Optimization Penentuan Resiko Kredit [11]. Berdasarkan hasil pengujian bahwa nilai akurasi algoritma C4.5 sebesar 85.40% dan nilai akurasi algoritma Naïve Bayes sebesar 85,09%. Dari kedua algoritma tersebut kemudian dilakukan kombinasi dengan optimasi Particle Swarm Optimization, dengan hasil algoritma C4.5+PSO memiliki nilai tertinggi berdasarkan nilai accuracy sebesar 87.61%, AUC sebesar 0.860 dan precision sebesar 88.96% sedangkan nilai recall tertinggi diperoleh oleh algoritma Naïve Bayes+PSO sebesar 96.75%.

Perbandingan tiga tradisional model algoritma klasifikasi seperti Naive Bayes, k-NN (lazy classifiers) dan Decision Tree berdasarkan nilai performa akurasi dan time execution pada dataset kanker leukimia yang datasetnya terdiri dari 7.130 atribut dan 72 records [12]. Penelitian ini membuktikan pada algoritma Naive Bayes memiliki nilai performa akurasi yang terbaik yaitu 91,17% daripada model algoritma klasifikasi lainnya yaitu Decision Tree dan K-NN.

Hasil klasifikasi dari masing-masing algoritma dalam penelitian ini nantinya akan dibandingkan untuk mendapatkan evaluasi kinerja terbaik dalam pendeteksian kanker payudara. Dengan demikian, dibutuhkan salah satu teknik data optimasi yang bertujuan untuk meningkatkan kinerja metode klasifikasi data mining konvensional yang sudah dipilih. Salah satu algoritma optimasi yang cukup popular adalah Particle Swarm Optimization (PSO). Particle Swarm Optimization (PSO) telah banyak memecahkan masalah optimasi algoritma [13], [14], [15].

Kanker Payudara

Kanker payudara adalah suatu penyakit dimana terjadi pertumbuhan berlebihan atau perkembangan tidak terkontrol dari sel-sel (jaringan) payudara [18]. Kanker bisa mulai tumbuh di dalam kelenjar susu,

saluran susu, jaringan lemak maupun jaringan ikat pada payudara.

Data Mining

Data mining, sering disebut juga sebagai Knowledge Discovery in Database (KDD), adalah kegiatan yang meliputi pengumpulan, pemakaian data-data yang berukuran besar [19].

Algoritma C4.5

Algoritma C4.5 diperkenalkan oleh J. Ross Quinlan diakhir tahun 1970 hingga awal tahun 1980-an. J. Ross Quinlan seorang peneliti dibidang mesin pembelajaran yang merupakan pengembangan dari algoritma ID3 (Interative Dichotomiser), algoritma tersebut digunakan untuk membentuk pohon keputusan [16].

Ada beberapa tahap dalam membuat sebuah pohon keputusan dengan algoritma C4.5 [17], yaitu:

1. Menyiapkan data training. Data training biasanya diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokkan ke dalam kelas-kelas tertentu.
2. Menentukan akar dari pohon. Akar akan diambil dari atribut yang terpilih dengan cara menghitung nilai Gain dari masing-masing atribut, nilai Gain yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung nilai Gain dari atribut, hitung dahulu nilai entropy yaitu :

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

Keterangan :
S : himpunan kasus
A : atribut
n : jumlah partisi S
Pi : proporsi dari Si terhadap S

3. Kemudian hitung nilai Gain dengan metode information gain :

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

Keterangan :
S : himpunan kasus
A : atribut
n : jumlah partisi atribut A
|Si| : jumlah kasus pada partisi ke-i
|S| : jumlah kasus dalam S

4. Ulangi langkah ke-2 hingga semua tupel terpartisi.
5. Proses partisi pohon keputusan akan berhenti saat :
   a. Semua tupel dalam node N mendapat kelas yang sama.
   b. Tidak ada atribut di dalam tupel yang dipartisi lagi.
   c. Tidak ada tupel di dalam cabang yang kosong.

Algoritma Naïve Bayes

Klasifikasi Bayesian adalah klasifikasi statistik yang bisa memprediksi probabilitas sebuah class. Klasifikasi Bayesian ini dihitung berdasarkan Teorema Bayes. Teorema Bayes adalah perhitungan statistik dengan menghitung probabilitas kemiripan kasus lama yang ada dibasis kasus dengan kasus baru. Teorema Bayes memiliki tingkat akurasi yang tinggi dan kecepatan yang baik ketika diterapkan pada database yang besar [16].

Persamaan dari teorema Bayes adalah sebagai berikut:

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)} (1)$$

Keterangan :
X : Kriteria suatu kasus berdasarkan masukan
Ci : Kelas solusi pola ke-i, dimana i adalah jumlah label kelas
P(Ci|X) : Probabilitas kemunculan label kelas Ci dengan kriteria masukan X
P(X|Ci) : Probabilitas kriteria masukan X dengan label kelas Ci
P(Ci) : Probabilitas label kelas Ci

Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) dikembangkan oleh Kennedy dan Eberhart (1995) sebagai algoritma optimasi yang bersifat stokastik dan berdasarkan pada model simulasi sosial. Secara umum PSO memiliki karakteristik yaitu konsepnya sederhana, mudah implementasinya, efisien dalam komputasi. Modifikasi kecepatan dan posisi tiap partikel dapat dihitung menggunakan kecepatan saat ini dan jarak pbesti,d ke gbestd seperti ditunjukan persamaan berikut

$$v_{i,m} = w.v_{i,m} + c_1 * R * (pbest_{i,m} - x_{i,m}) + c_2 * R * (gbest_m - x_{i,m})$$

Menghitung kecepatan baru untuk tiap partikel (solusi potensial) berdasarkan pada kecepatan sebelumnya (Vi,m), lokasi partikel dimana nilai

fitness terbaik telah dicapai (pbest), dan lokasi populasi global (gbest untuk versi global, lbest untuk versi local) atau local neighborhood pada algoritma versi local dimana nilai fitness terbaik telah dicapai.

$$x_{id} = x_{i,m} + v_{i,m}$$

Memperbaharui posisi tiap partikel pada ruang solusi. Dua bilangan acak c1 dan c2 dibangkitkan sendiri. Penggunaan berat inersia w telah memberikan performa yang meningkat pada sejumlah aplikasi [21]. Hasil dari perhitungan partikel yaitu kecepatan partikel diantara interval [0,1].

Dimana:

n : jumlah partikel dalam kelompok
d : dimensi
vi,m : kecepatan partikel ke-i pada iterasi ke-i
w : faktor bobot inersia
c1, c2 : konstanta akeselerasi (learning rate)
R : bilangan random (0-1)
xi,d : posisi saat ini dari partikel ke-i pada iterasi ke-i
pbest : posisi terbaik sebelumnya dari partikel ke-i
gbest : partikel terbaik diantara semua partikel dalam satu kelompok atau populasi

Pengujian dan Evaluasi

Model validasi yang digunakan pada penelitian ini adalah 10 fold cross validation. 10 fold cross validation digunakan untuk mengukur kinerja model prediksi. Setiap dataset secara acak dibagi menjadi 10 bagian dengan ukuran yang sama. Selama 10 kali, 9 bagian untuk melatih model (data training) dan 1 bagian digunakan untuk menguji (data testing) yang lainya setiap kali dilakukan pengujian. Pegukuran pada evaluasi kinerja klasifikasi bertujuan untuk mengetahui seberapa akurat model klasifikasi dalam prediksi kelas dari suatu baris data [19].

Tabel 1 Confusion Matrix

Class		Actual	
		True	False
Predic	True	True Positif(TP)	False Negative (FN)
	False	False Positive (FP)	True negative (TN)

True positive (tp) merupakan jumlah record positif dalam data set yang diklasifikasikan positive. True negative (tn) merupakan jumlah record negative dalam data set yang diklasifikasikan negative. False positive (fp) merupakan jumlah record negatif dalam data set yang diklasifikasikan potitif. False negative (fn) merupakan jumlah record positive dalam data set yang diklasifikasikan negative.

Metode confusion matrix merepresentasikan hasil evaluasi model dengan menggunakan tabel matriks, jika dataset terdiri dari dua kelas, kelas pertama dianggap positif, dan kelas kedua dianggap negative [20]. Evaluasi menggunakan confusion matrix menghasilkan nilai akurasi, presisi, recall. Akurasi dalam klasifikasi merupakan presentase ketepatan record data yang diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi. Precision atau confidence merupakan proporsi kasus yang diprediksi positif yang juga positif benar pada data yang sebenarnya. Recall atau sensitivity merupakan proporsi kasus positif yang sebenarnya yang diprediksi positif secara benar.

Berikut adalah persamaan model confusion matrix:

a. Nilai akurasi (acc) adalah proporsi jumlah prediksi yang benar. Dapat dihitung dengan menggunakan persamaan:

$$akurasi = \frac{tp + tn}{tp + tn + fp + fn}$$

b. Sensitivity atau recall digunakan untuk membandingkan proporsi tp terhadap tupel yang positif, yang dihitung dengan menggunakan persamaan:

$$Sensitivity = \frac{tp}{tp + fn}$$

c. PPV (positive predictive value) atau precision adalah proporsi kasus dengan hasil diagnosa positif, yang dihitung dengan menggunakan persamaan:

$$PPV = \frac{tp}{tp + fp}$$

Hasil akurasi juga dapat dilihat dengan melakukan perbandingan klasifikasi menggunaakan curva Receiver Operating Characteristic (ROC) dari hasil confusion matrix. ROC menghasilkan dua garis dengan bentuk true positif yang ditandai dengan garis vertical dan false positive yang ditandai dengan garis horiozontal. ROC adalah grafik antara sensitivitas true positive rate pada sumbu X dan sumbu Y. Kurva ROC menunjukkan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan confusion matrix.

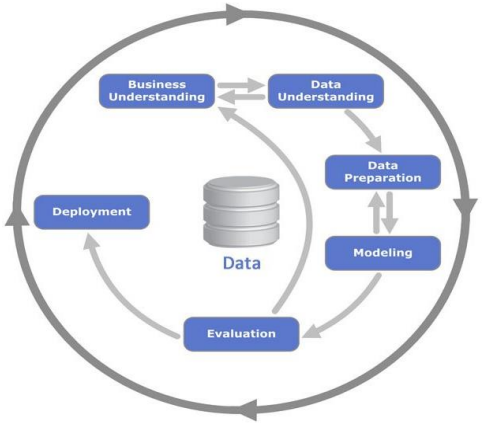
Tingkat akurasi dapat di diagnosa sebagai berikut [21]

a. Akurasi 0.90 – 1.00 = Excellent classification
b. Akurasi 0.80 – 0.90 = Good classification
c. Akurasi 0.70 – 0.80 = Fair classification
d. Akurasi 0.60 – 0.70 = Poor classification
e. Akurasi 0.50 – 0.60 = Failure classification

II. METODOLOGI PENELITIAN

Metode analisis dalam penelitian ini mengacu pada tahapan proses CRISP-DM. CRISP-DM (CRoss-Industry Standard Process for Data Mining) merupakan suatu konsorsium perusahaan yang didirikan oleh Komisi Eropa pada tahun 1996 dan telah ditetapkan sebagai proses standar dalam data mining yang dapat diaplikasikan di berbagai sektor industri.

a. Business Understanding.
   Dalam penelitian ini fokus pada pendeteksian kanker payudara dengan menggunakan perbandingan 2 algoritma klasifikasi data mining yaitu Algoritma C4.5 dan Naïve Bayes.

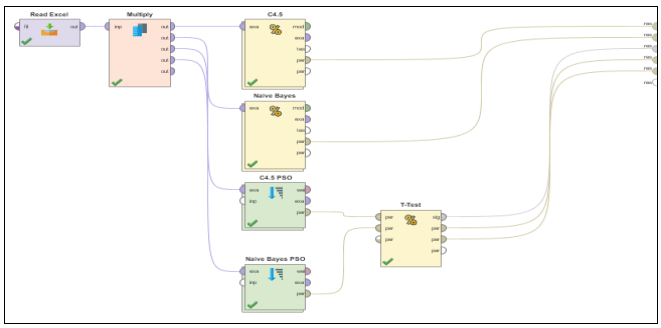


Gambar 1. CRISP-DM

b. Data Understanding.
   Dataset kanker payudara diambil pada http://archive.ics.uci.edu, Breast Cancer dari Dr. William H. Woldberg (1989-1991) University of Wisconsin Hospital, Madison, USA.
c. Data Preparation.
   Dalam penelitian ini dilakukan pemilihan data seluruh indikator dalam membentuk dataset kanker payudara. Selanjutnya dilakukan pembobotan nilai yang secara default sudah ada http://archive.ics.uci.edu. Dataset kanker payudara ini berjumlah 699 dengan 11 parameter indikator yang akan diuji antara lain: Sample Code Number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses, dan Class (atribut hasil prediksi).
d. Modelling.
   Model klasifikasi dimulai dari dataset akan dilakukan pemodelan dengan algoritma klasifikasi sehingga dihasilkan model klasifikasi dan memunculkan parameter evauasi. Model yang ada dalam penelitian ini adalah perbandingan optimasi Algoritma C4.5 + PSO dan Naïve Bayes + PSO
e. Evaluation
   Pada phase ini akan dilakukan proses evaluasi dari phase sebelumnya. Phase evaluasi ini akan dilakukan perbandingan quantitatif dengan mempertimbangkan nilai komparasi confusion matrix dengan pengukuran berupa Accuracy, Precision dan Recall.
f. Deployment.
   Tahapan penentuan model klasifikasi yang memiliki nilai kinerja terbaik dari hasil uji T-Test hasil komparasi model data mining Algoritma C4.5 + PSO dan Naïve Bayes + PSO. Kemudian dibuat rekomendasi model mana yang terbaik yang diterapkan pendeteksian kanker payudara.

III. HASIL DAN PEMBAHASAN

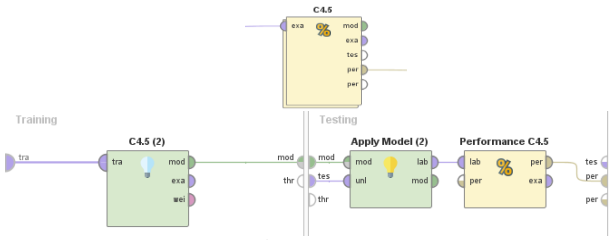
Hasil konfigurasi model pada Rapidminer versi 9 dengan perbandingan metode klasifikasi data mining yaitu Algoritma *C4.5* dan Naïve Baye menggunakan uji beda T-Test untuk mencari kinerja terbaik.



Gambar 2 Main Model Penelitian

Algoritma C4.5

Hasil model konfigurasi Algoritma *C4.5* pada Rapidminer versi 9 dengan *performance* 10-*fold Cross Validation*.



Gambar 3 Konfigurasi Model Algoritma *C4.5*

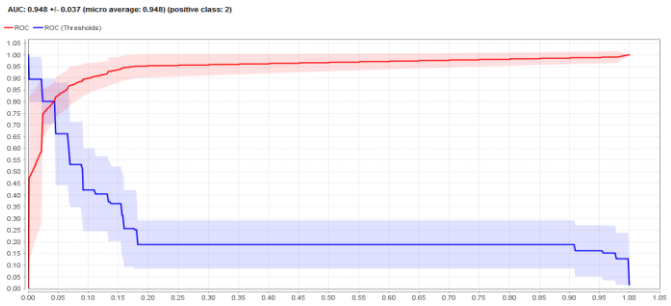
Hasil pengujian dan validasi melalui *confusion matrix* Algoritma *C4.5* pada Rapidminer versi 9 tervisualisasi pada hasil *performance* dengan pengukuran akurasi, *precision*, dan *recall* digambarkan pada gambar 4 sedangkan pada gambar 6 adalah hasil kurva ROC.

	true 1	true 2	class precision
pred. 1	421	44	90.54%
pred. 2	23	195	89.45%
class recall	94.82%	81.59%	

Gambar 4 Hasil *Confusion Matrix C4.5*

```
PerformanceVector:
accuracy: 90.19% +/- 2.81% (micro average: 90.19%)
ConfusionMatrix:
True:   1       2
1:      421     44
2:      23      195
precision: 90.16% +/- 7.00% (micro average: 89.45%) (positive class: 2)
ConfusionMatrix:
True:   1       2
1:      421     44
2:      23      195
recall: 81.49% +/- 6.80% (micro average: 81.59%) (positive class: 2)
ConfusionMatrix:
True:   1       2
1:      421     44
2:      23      195
AUC (optimistic): 0.974 +/- 0.028 (micro average: 0.974) (positive class: 2)
AUC: 0.948 +/- 0.037 (micro average: 0.948) (positive class: 2)
AUC (pessimistic): 0.923 +/- 0.052 (micro average: 0.923) (positive class: 2)
```

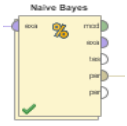
Gambar 5 Hasil *Performance* Algoritma *C4.5*

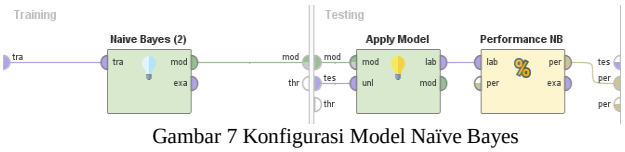


Gambar 6 Kurva ROC Algoritma *C4.5*

Algoritma Naïve Bayes

Hasil model konfigurasi Algoritma Naïve Bayes pada Rapidminer versi 9 dengan *performance* 10-*fold Cross Validation*.





Gambar 7 Konfigurasi Model Naïve Bayes

Hasil pengujian dan validasi melalui *confusion matrix* Algoritma *Naive Bayes* pada Rapidminer versi 9 tervisualisasi pada hasil *performance* dengan pengukuran akurasi, *precision*, dan *recall* digambarkan pada gambar 8 sedangkan pada gambar 10 adalah hasil kurva ROC.

	true 1	true 2	class precision
pred. 1	430	2	99.54%
pred. 2	14	237	94.42%
class recall	96.85%	99.16%	

Gambar 8 Hasil *Confusion Matrix Naïve Bayes*

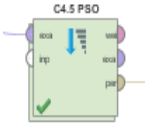
```
PerformanceVector:
accuracy: 97.65% +/- 1.88% (micro average: 97.66%)
ConfusionMatrix:
True:   1       2
1:      430     2
2:      14      237
precision: 94.60% +/- 4.75% (micro average: 94.42%) (positive class: 2)
ConfusionMatrix:
True:   1       2
1:      430     2
2:      14      237
recall: 99.17% +/- 1.67% (micro average: 99.16%) (positive class: 2)
ConfusionMatrix:
True:   1       2
1:      430     2
2:      14      237
AUC (optimistic): 0.991 +/- 0.012 (micro average: 0.991) (positive class: 2)
AUC: 0.990 +/- 0.014 (micro average: 0.990) (positive class: 2)
AUC (pessimistic): 0.990 +/- 0.014 (micro average: 0.990) (positive class: 2)
```

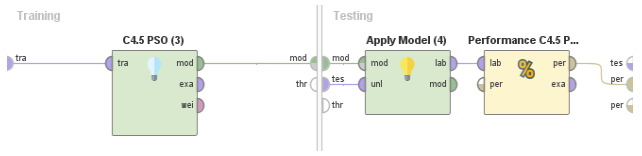
Gambar 9 Hasil Performance *Naïve Bayes*

Gambar 10 Kurva ROC Algoritma *Naïve Bayes*

Algoritma C4.5 + *Particle Swarm Optimization (PSO)*

Hasil model konfigurasi Algoritma *C4.5* berbasis *Particle Swarm Optimization (PSO)* pada Rapidminer versi 9 dengan *performance* 10-*fold Cross Validation*.





Gambar 11 Konfigurasi Algoritma *C4.5* + PSO

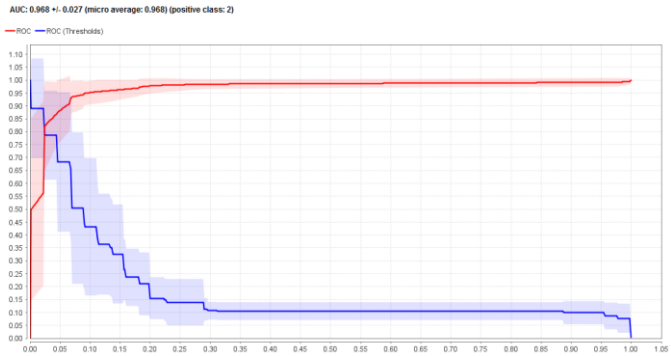
Hasil pengujian dan validasi melalui *confusion matrix* Algoritma *C4.5 + PSO* pada Rapidminer versi 9 tervisualisasi pada hasil *performance* dengan pengukuran akurasi, *precision*, dan *recall* digambarkan pada gambar 12 sedangkan pada gambar 14 adalah hasil kurva ROC.

	true 1	true 2	class precision
pred. 1	424	19	95.71%
pred. 2	20	220	91.67%
class recall	95.50%	92.05%	

Gambar 12 Hasil Confusion Matrix *C4.5* + PSO

```
PerformanceVector:
accuracy: 94.29% +/- 3.11% (micro average: 94.29%)
ConfusionMatrix:
True:   1       2
1:      424     19
2:      20      220
precision: 91.90% +/- 6.36% (micro average: 91.67%) (positive class: 2)
ConfusionMatrix:
True:   1       2
1:      424     19
2:      20      220
recall: 91.97% +/- 4.14% (micro average: 92.05%) (positive class: 2)
ConfusionMatrix:
True:   1       2
1:      424     19
2:      20      220
AUC (optimistic): 0.977 +/- 0.025 (micro average: 0.977) (positive class: 2)
AUC: 0.968 +/- 0.027 (micro average: 0.968) (positive class: 2)
AUC (pessimistic): 0.960 +/- 0.028 (micro average: 0.960) (positive class: 2)
```

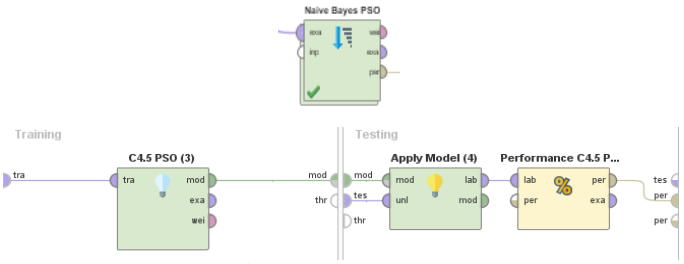
Gambar 13 Hasil *Performance C4.5* + PSO



Gambar 14 Kurva ROC Algoritma *C4.5* + PSO

Algoritma Naïve Bayes + *Particle Swarm Optimization (PSO)*

Hasil model konfigurasi Algoritma *Naïve Bayes* berbasi *Particle Swarm Optimization (PSO)* pada Rapidminer versi 9 dengan *performance* 10-*fold Cross Validation*.



Gambar 15 Konfigurasi Model Algoritma Naïve Bayes PSO

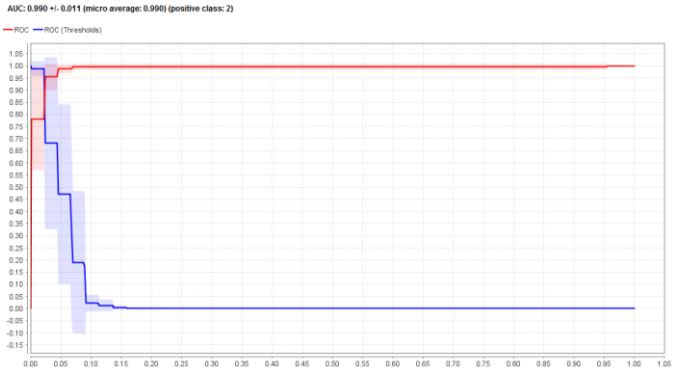
Hasil pengujian dan validasi melalui *confusion matrix* Algoritma *Naive Bayes + PSO* pada Rapidminer versi 9 tervisualisasi pada hasil *performance* dengan pengukuran akurasi, *precision*, dan *recall* digambarkan pada gambar 16 sedangkan pada gambar 18 adalah hasil kurva ROC.

	true 1	true 2	class precision
pred. 1	432	2	99.54%
pred. 2	12	237	95.18%
class recall	97.30%	99.16%	

Gambar 16 Hasil Confusion Matrix Naïve Bayes PSO

```
PerformanceVector:
accuracy: 97.96% +/- 1.33% (micro average: 97.95%)
ConfusionMatrix:
True:   1       2
1:      432     2
2:      12      237
precision: 95.40% +/- 3.58% (micro average: 95.18%) (positive class: 2)
ConfusionMatrix:
True:   1       2
1:      432     2
2:      12      237
recall: 99.17% +/- 1.67% (micro average: 99.16%) (positive class: 2)
ConfusionMatrix:
True:   1       2
1:      432     2
2:      12      237
AUC (optimistic): 0.990 +/- 0.011 (micro average: 0.990) (positive class: 2)
AUC: 0.990 +/- 0.011 (micro average: 0.990) (positive class: 2)
AUC (pessimistic): 0.990 +/- 0.011 (micro average: 0.990) (positive class: 2)
```

Gambar 17 Hasil Performance Naïve Bayes PSO



Gambar 18 Kurva ROC Algoritma Naïve Bayes + PSO

Pembahasan

Berdasarkan hasil pengujian dan evaluasi didapatkan perbandingan hasil performa dari

masing masing algoritma dan penerapan optimasi yang dapat dijabarkan sebagai berikut.

Tabel 2 Hasil Komparasi Kinerja Algoritma

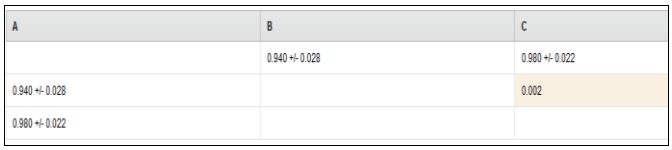
Parameter Kinerja	***C4.5***	***C4.5 (PSO)***	***Naïve Bayes***	***Naïve Bayes (PSO)***
Accuracy	90,19%	94,29%	97,65%	97,96%
Precision	90,16%	91,90%	94,60%	95,40%
Recall	81,49%	91,97%	99,17%	99,17%
AUC	0,948	0,968	0,990	0,990

Berdasarkan hasil penerapan optimasi *Particle Swarm Optimization (PSO)* pada tabel 2 didapatkan bahwa terbukti optimasi *Particle Swarm Optimization (PSO)* dapat meningkatkan performa atau kinerja algoritma algoritma *C4.5* dan *Naive Bayes*. Pada indikator akurasi algoritma *C4.5* mengalami peningkatan dari 90,19% menjadi 94,29% dan akurasi *Naïve Bayes* 97,65% menjadi 97,96%. Selain itu berdasarkan evaluasi yang dilakukan secara *confusion matrix* ternyata terbukti bahwa hasil akurasi tertinggi pada perbandingan hasil kinerja klasifikasi didapatkan oleh algoritma *Naïve Bayes* berbasis *PSO* sebesar 97.96% disusul algoritma C4.5 berbasis *Particle Swarm Optimization (PSO)* sebesar 94.29%.

Ditinjau dari indikator pengukuran *precision* dan *recall* juga mengalami peningkatan yang signifikan diantaranya algoritma *C4.5* dengan nilai *precision* dan *recall* sebesar 90,16% dan 81,49% meningkat menjadi 91,90% dan 91,97% dan algoritma *Naïve Bayes* dengan nilai *precision* dan *recall* sebesar 94,69% dan 99,17% menjadi 95,40% dan 99,17%.

Analisis yang berbeda ditinjau dari pengukuran *AUC* bahwa penerapan optimasi berbasis *Particle Swarm Optimization (PSO)* tidak terlalu berpengaruh pada pingkatan kinerja algoritma secara keseluruhan. Terjadi peningkatan pada algoritma *C4.5* namun tidak pada algoritma *Naïve Bayes.* Secara keseluruhan algoritma *C4.5* dan *Naïve Bayes* berdasarkan nilai *AUC* masuk pada predikat *Excelent Classification.*

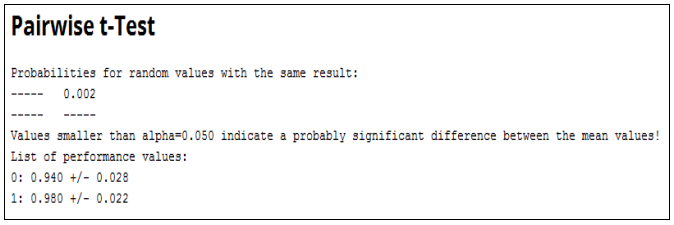
Dari hasil T-Test pada gambar 19 dapat disimpulkan bahwa algoritma Naïve Bayes Particle Swarm Optimization (PSO) memiliki nilai tertinggi sebesar 0,980 disusul algoritma C4.5 (PSO) 0,940. Dengan demikian algoritma Naïve Bayes Particle Swarm Optimization (PSO) dapat memberikan solusi terbaik terhadap akurasi pendeteksian penyakit kanker payudara.



A	B	C
	0.940 +/- 0.028	0.980 +/- 0.022
0.940 +/- 0.028		0.002
0.980 +/- 0.022		

Gambar 19 Hasil Uji T-Test

Standar nilai ambang batas untuk menentukan signifikansi sesuai petunjuk pada *Rapidminer* adalah di bawah 0.05. Sehingga dari nilai gambar 19 menunjukkan bahwa nilai tingkat signifikansi model algoritma *Naïve Bayes* yang dioptimasi *Particle Swarm Optimization* (*PSO*) menghasilkan nilai 0.002, lebih kecil dari nilai ambang batas. Sehingga hal tersebut berarti dapat dipastikan kinerja dari model algoritma *Naïve Bayes* berbasis PSO secara nyata akan lebih baik dari model algoritma *C4.5 (PSO).* Dari beberapa pengujian yang telah dilakukan, keseluruhan data yang telah dihasilkan dapat dijadikan sebagai dasar atau pendukung pengambilan keputusan dalam membuat kesimpulan hasil penelitian.



Pairwise t-Test

```
Probabilities for random values with the same result:
-----   0.002
-----   -----
Values smaller than alpha=0.050 indicate a probably significant difference between the mean values!
List of performance values:
0: 0.940 +/- 0.028
1: 0.980 +/- 0.022
```

Gambar 20 Pairwise T-Test

IV. PENUTUP

KESIMPULAN

a. Terbukti optimasi *Particle Swarm Optimization (PSO)* dapat meningkatkan kinerja akurasi algoritma *C4.5* dan *Naive Bayes.* Pada akurasi algoritma *C4.5* mengalami peningkatan dari 90,19% menjadi 94,29%, akurasi *Naïve Bayes* dengan nilai 97,65% menjadi 97,96%.
b. Hasil kinerja terbaik yang diuji menggunakan *T-Test* pada algoritma *C4.5* dan *Naïve Bayes berbasis Particle Swarm Optimization (PSO)* dapat dihasilkan bahwa algoritma Naïve Bayes (PSO) memiliki nilai tertinggi sebesar 0,980 dilanjutkan algoritma *C4.5 (PSO)* sebesar 0,943. Dengan demikian algoritma *Naïve Bayes Particle Swarm Optimization* (PSO) dapat

memberikan solusi terbaik terhadap akurasi pendeteksian penyakit kanker payudara.

SARAN

a. Dibutuhkan jumlah data yang lebih besar, atribut yang lebih kompleks, bahkan menggunakan sampel penyakit lain yang sifatnya baru sehingga kedepanya hasil pengukuran yang dihasilkan akan lebih berguna dan lebih handal akurasinya.
b. Menggunakan metode optimasi lain seperti Ant Colony Optimization (ACO), Genetik Algorithm (GA) dan lain sebagainya.
c. Melakukan pengujian dan perbandingan pada algoritma lain ataupun menggunakan metode hybrid untuk mendapatkan pengetahuan komparasi yang lebih luas.
d. Melakukan pengembangan pada tahap preprocessing data dengan menggunakan metode seleksi atribut yang lain seperti chi-square, information index dan lain sebagainya untuk ketepatan penyeleksian atribut.

REFERENSI

[1] Saiyed, Sohana. 2016, A Survey on Naïve Bayes Based Prediction of Heart Disease Using Risk Factors, India: CSPIT Changa, Vol 3, Issued 25 Maret 2016.

[2] Tawang Wulandari, Marji & Lailil Muflikhah, 2018, Klasifikasi Jenis Kanker Berdasarkan Struktur Protein Menggunakan Algoritma Naive Bayes, Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, e-ISSN: 2548-964X, Vol. 2, No. 10 Oktober 2018

[3] Elsa Nuramilus Shofia, Rekyan Regasari Mardi Putri, Achmad Arwan, 2017, Sistem Pakar Diagnosis Penyakit Demam: DBD, Malaria dan Tifoid Menggunakan Metode K-Nearest Neighbor – Certainty Factor, Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, e-ISSN: 2548-964X, Vol. 1, No. 5, Mei 2017

[4] Muhdi, Abdullah, Usman, 2017, Sistem Klasifikasi Penyakit Asma Menggunakan Algoritma Naïve Bayes, Jurnal SISTEMASI, ISSN:2302-8149, Volume 6, Nomor 3, September 2017 : 33 – 39

[5] Abdul Rohman, Vincent Suhartono & Catur Supriyanto, 2017, Penerapan Algoritma C4.5 Berbasis Adaboost Untuk Prediksi Penyakit Jantung, Jurnal Teknologi Informasi, ISSN 1907-3380, Volume 13 Nomor 1, Januari 2017

[6] Deny Wiria Nugraha, A.Y. Erwin Dodu & Novilia Chandra, 2017, Klasifikasi Penyakit Stroke Menggunakan Metode Naive Bayes Classifier, SemanTIK, ISSN : 2502-8928, Vol.3, No.2, Jul-Des 2017, pp. 13-22

[7] Ari Muzakir, Rika Anisa Wulandari, 2016, Model Data Mining sebagai Prediksi Penyakit Hipertensi Kehamilan dengan Teknik Decision Tree, Scientific Journal of Informatics, p-ISSN 2407-7658, Vol. 3, No. 1, Mei 2016

[8] Tejas Mehta, Dhaval Kathiriya, 2016, Performance Analysis of Data Mining Classification Techniques, International Journal of Innovative Research in Science, Engineering and Technology, ISSN : 2319-8753. Vol. 5, Issue 3

[9] Leni Marlina, Muslim, Andysah Putera Utama Siahaan, Data Mining Classification Comparison (Naïve Bayes and C4.5 Algorithms), International Journal of Engineering Trends and Technology (IJETT) – Volume 38 Number 7, ISSN: 2231-5381, 2016

[10] Andri Permana Wicaksono, Tessy Badriyah, Achmad Basuki, 2016, Comparison of The Data-Mining Methods in Predicting The Risk Level of Diabetes, International Journal of Engineering Technology (E MITTER), Vol. 4, No. 1, ISSN: 2443-1168

[11] Achmad Rifai, Rizki Aulianita, 2018, Komparasi Algoritma Klasifikasi C4.5 dan Naïve Bayes Berbasis Particle Swarm Optimization Untuk Penentuan Resiko Kredit, Journal Speed – Sentra Penelitian Engineering dan Edukasi, ISSN : 1979-9330, Volume 10 No 2 – 2018.

[12] Durairaj, M., & Deepika, R, 2015, Comparative Analysis of Classification Algorithms for the Prediction of Leukemia Cancer. International Journal of Advanced Research in Computer Science and Software Engineering; Volume 5, No. 8, pp. 787-791

[12] Saprudin, 2017, Penerapan Particle Swarm Optimization (PSO) Untuk Klasifikasi dan Analisis Kredit Dengan Menggunakan Algoritma C4.5, Jurnal Informatika Universitas Pamulang, ISSN 2541-1004, Vol 2, No.4 Desember 2017.

[13] Husin Muhamad, Cahyo Adi Prasojo, Nur Afifah Sugianto, Listiya Surtiningsih, Imam Cholissodin, 2017, Optimasi Naïve Bayes Classifier Dengan Menggunakan Particle Swarm Optimization Pada Data Iris, Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), p-ISSN: 2355-7699, Vol. 4, No. 3, September 2017, hlm. 180-184

[14] Mirza Yogy Kurniawan, Muhammad Edya Rosadi, 2017, Optimasi Decision Tree Menggunakan Particle Swarm Optimization Pada Data Siswa Putus Sekolah, JTIULM - Volume 2, Nomor 1, Juni 2017: 15 - 22

[15] Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques.

[16] Kusrini dan Emha Taufiq Lutfi, 2009, Algorima Data Mining, Andi Offset, Yogyakarta.

[17] Rahayu-Tjioe, A, 1991, Kanker payudara. Yayasan Kanker Wisnuwardhana, Surabaya

[18] Santosa, B. 2007. Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis. Graha Ilmu, Yogyakarta.

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

[20] Gorunescu, F. (2011). Data Mining Concepts, Models and Techniques, Springer, Verlag Berlin Heidelrbeg

[21] Wu, X., Shi, Y., & Eberhart, R. (2004). Recent Advances in Particle Swarm. IEEE, 90-97