 ISSN : 2580-1503

PENERAPAN OPTIMASI BERBASIS PARTICLE SWARM OPTIMIZATION (PSO) ALGORITMA NAÏVE BAYES DAN K-NEAREST NEIGHBOR SEBAGAI PERBANDINGAN UNTUK MENCARI KINERJA TERBAIK DALAM MENDETEKSI KANKER PAYUDARA

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

[1)]*Magister Teknik Informatika Univ. Amikom,* [2)] *Fak. Ilmu Komputer & Teknologi Informasi Univ. Gunadarma*
[1)]*Jl. Ring Road Utara, Condong Catur, Sleman 55283* [2)] *Jl. Margonda Raya 100, Depok 16424*
Email : taghfirul.yoga@yahoo.co.id[1)], pri@staff.gunadarma.ac.id[2]

Abstrak

Saat ini kanker payudara menjadi jenis kanker yang sangat menakutkan bagi perempuan diseluruh dunia, hal ini juga berlaku di Indonesia. Salah satu pemanfaatan teknologi informasi dalam bidang kesehatan adalah disiplin ilmu yang berkembang pesat dewasa ini yaitu Data Mining. Dibutuhkan salah satu teknik data optimasi yang bertujuan untuk meningkatkan kinerja metode klasifikasi data mining konvensional yang sudah dipilih dalam penelitian ini. Salah satu algoritma optimasi yang cukup popular adalah Particle Swarm Optimization (PSO). Penelitian ini bertujuan menerapkan dan mengevaluasi perbandingan kinerja terbaik metode klasifikasi data mining algoritma Naïve Bayes dan K-Nearest Neighbor berbasis PSO untuk mendeteksi kanker payudara. Hasil Penelitian ini menjelaskan bahwa penerapan Particle Swarm Optimization (PSO) menghasilkan hasil yang signifikan dalam memberikan peningkatan kinerja (optimasi) pada algoritma Naïve Bayes dan K-Nearest Neighbor. Berdasarkan uji beda menggunakan T-Test didapatkan algoritma Naïve Bayes berbasis Particle Swarm Optimization (PSO) memiliki nilai tertinggi dibanding K-Nearest Neighbor dengan nilai perolehan sebesar 0,978.

Kata kunci: *Data Mining, Naive Bayes, K-Nearest Neighbor, Particle Swarm Optimization (PSO), Klasifikasi.*

1. Pendahuluan

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.

Salah satu pemanfaatan teknologi informasi dalam bidang kesehatan adalah disiplin ilmu yang berkembang pesat dewasa ini yaitu Data Mining. Data mining memiliki banyak sekali manfaat dalam berbagai masalah pengolahan data, sehingga data – data yang tidak memiliki informasi penting dapat digali dan dianalisa. Data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui [1].

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].

2. Metode Penelitian

2.1 Studi Pustaka

2.1.1 Kanker Payudara

Kanker payudara adalah suatu penyakit dimana terjadi pertumbuhan berlebihan atau perkembangan tidak terkontrol dari sel-sel (jaringan) payudara. Kanker bisa mulai tumbuh di dalam kelenjar susu, saluran susu, jaringan lemak maupun jaringan ikat pada payudara [23].

2.1.2 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]. *Ouput* atau keluaran dari *Data mining* ini bisa dipakai untuk memperbaiki pengambilan keputusan di masa depan. Sehingga istilah *pattern recognition* sekarang jarang digunakan karena sudah termasuk bagian dari *Data mining*.

2.1.3 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 :

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

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

2.1.2 Algoritma K-Nearest Neighbor

Algoritma *K-Nearest Neighbor* (k-NN atau KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut, Ketepatan algoritma k-NN ini sangat dipengaruhi oleh ada atau tidaknya fitur-fitur yang tidak relevan, atau jika bobot fitur tersebut tidak setara dengan relevansinya terhadap klasifikasi [17].

Langkah – langkah algoritma kNN ditunjukkan sebagai berikut [1]:

1. Tentukan nilai latih k, yaitu jumlah tetangga terdekat
2. Menghitung kuadrat jarak euclidian (euclidean distance) masing-masing objek terhadap data sampel yang diberikan

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Dimana
d : jarak kedekatan
x : data training
y : data testing
n : jumlah attribut individu antara 1 sampai n
f : fungsi atribut I antara kasus x & kasus y
Wi : bobot yang diberika pada atribut ke-i

ISSN : 2580-1503

Jarak antara objek x dan y didefinisikan sebagai D_{xy}, dimana xi merupakan *record* yang akan diprediksi dan yi merupakan record data pola sedangkan nilai n didefinisikan sebagai jumlah atribut dan nilai I merujuk pada record ke-i

3. Mengurutkan objek-objek tersebut ke dalam kelompok yang mempunyai jarak euclid terkecil
4. Mengumpulkan kategori Y (Klasifikasi nearest neighbor)
5. Dengan menggunakan kategori mayoritas, maka dapat diprediksikan nilai query instance yang telah dihitung.

2.1.3 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 ($V_{i,m}$), 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
$v_{i,m}$: kecepatan partikel ke-i pada iterasi ke-i
w : faktor bobot inersia
c_1, c_2 : konstanta akeselerasi (*learning rate)*
R : bilangan random (0-1)
$x_{i,d}$: 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

2.1.4 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.

 ISSN : 2580-1503

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*

2.2 Metode Analisis

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 Naïve Baye dan K-Nearest Neighbor.

Gambar 1 CRISP-DM

b) Data Understanding.

Dataset kanker payudara yang diambil pada http://archive.ics.uci.edu, data 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).*

a) 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 Naïve Bayes + PSO dan K-Nearest Neighbor + PSO

b) Evaluation

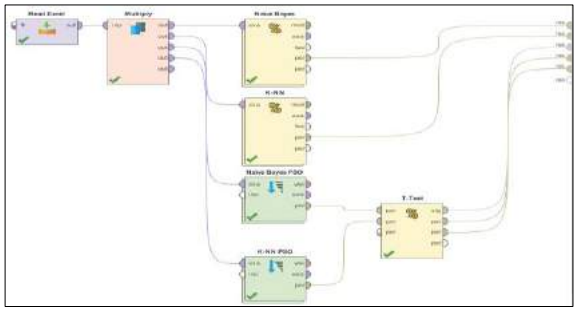
Pada phase ini akan dilakukan proses evaluasi dari phase sebelumnya. Phase evaluasi ini akan dilakukan perbandingan quantitatif dengan mempertimbangkan nilai komparasi *confusion matrix* dengan pengukuran berupa *Accuracy, Precision* dan *Recall*.

c) *Deployment*.

Tahapan penentuan model klasifikasi yang memiliki nilai kinerja terbaik dari hasil uji T-Test hasil komparasi model data mining Naïve Bayes + PSO dan K-Nearest Neighbor + PSO Kemudian dibuat rekomendasi model mana yang terbaik yang diterapkan pendeteksian kanker payudara.

3. Pembahasan

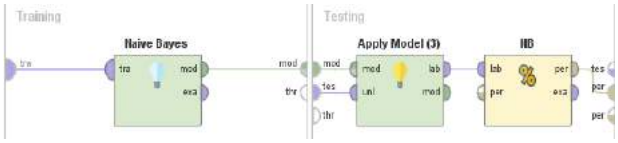
Hasil konfigurasi model pada Rapidminer versi 9 dengan perbandingan metode klasifikasi data mining yaitu Naïve Bayes dan *K-Nearest Neighbor* menggunakan uji beda T-Test untuk mencari kinerja terbaik.



Gambar 2 Main Model Penelitian

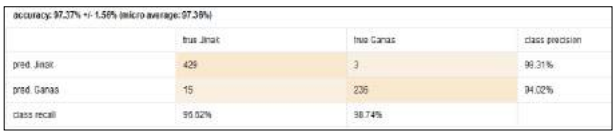
3.1 Algoritma Naïve Bayes

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



Gambar 3 Konfigurasi Model Algoritma *Naïve Bayes*

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



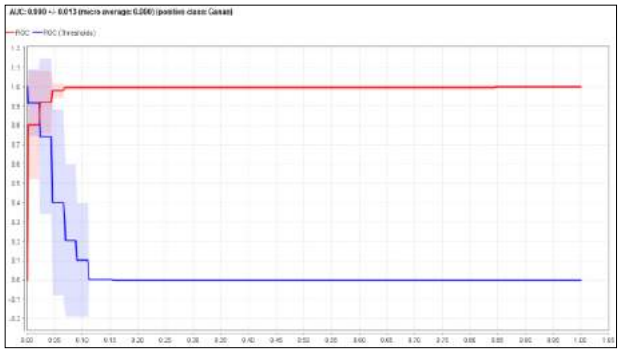
	true Jinak	true Ganas	class precision
pred. Jinak	429	3	99.31%
pred. Ganas	15	236	94.02%
class recall	96.62%	98.74%	

Gambar 4 Hasil *Confusion Matrix Naïve Bayes*

ISSN : 2580-1503

```
PerformanceVector:
accuracy: 97.37% +/- 1.56% (micro average: 97.36%)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  429     3
Ganas:  15      236
precision: 94.31% +/- 4.32% (micro average: 94.02%) (positive class: Ganas)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  429     3
Ganas:  15      236
recall: 98.71% +/- 2.76% (micro average: 98.74%) (positive class: Ganas)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  429     3
Ganas:  15      236
AUC (optimistic): 0.991 +/- 0.012 (micro average: 0.991) (positive class: Ganas)
AUC: 0.990 +/- 0.013 (micro average: 0.990) (positive class: Ganas)
AUC (pessimistic): 0.990 +/- 0.013 (micro average: 0.990) (positive class: Ganas)
```

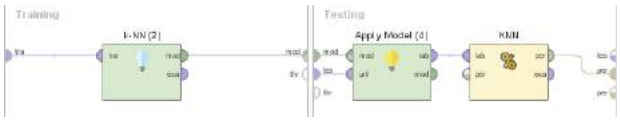
Gambar 5 Hasil Performance *Naïve Bayes*



Gambar 6 Kurva ROC Algoritma *Naïve Bayes*

3.2 Algoritma K-Nearest Neighbor

Hasil model konfigurasi Algoritma *K-Nearest Neighbor* pada Rapidminer versi 9 dengan *performance* 10-*fold Cross Validation*.



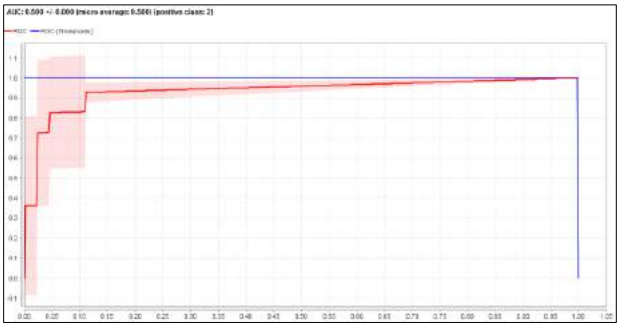
Gambar 7 Konfigurasi Model Algoritma *K-Nearest Neighbor*

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

Gambar 8 Hasil *Confusion Matrix K-Nearest Neighbor*

```
PerformanceVector:
accuracy: 95.45% +/- 2.59% (micro average: 95.46%)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  433     20
Ganas:  11      219
precision: 95.30% +/- 3.30% (micro average: 95.22%) (positive class: Ganas)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  433     20
Ganas:  11      219
recall: 91.65% +/- 6.17% (micro average: 91.63%) (positive class: Ganas)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  433     20
Ganas:  11      219
AUC (optimistic): 0.998 +/- 0.002 (micro average: 0.998) (positive class: Ganas)
AUC: 0.500 +/- 0.000 (micro average: 0.500) (positive class: Ganas)
AUC (pessimistic): 0.894 +/- 0.064 (micro average: 0.894) (positive class: Ganas)
```

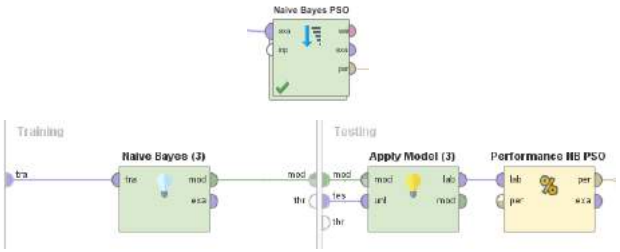
Gambar 9 Hasil *Performance* Algoritma *K-Nearest Neighbor*



Gambar 10 Kurva *ROC* Algoritma *K-Nearest Neighbor*

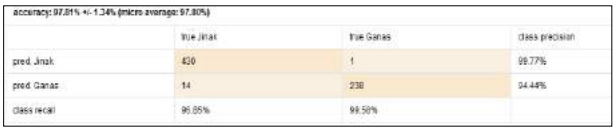
3.3 Algoritma Naïve Bayes + *Particle Swarm Optimization (PSO)*

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



Gambar 11 Konfigurasi Model Algoritma *Naïve Bayes + PSO*

Hasil pengujian dan validasi melalui *confusion matrix* Algoritma Naïve Bayes + PSO pada Rapidminer versi 9 tervisualisasi pada gambar 12. Hasil performance dengan pengukuran akurasi, *precision*, dan *recall* digambarkan pada gambar 13 sedangkan pada gambar 14 adalah hasil kurva ROC Algoritma *Naïve Bayes + PSO*.

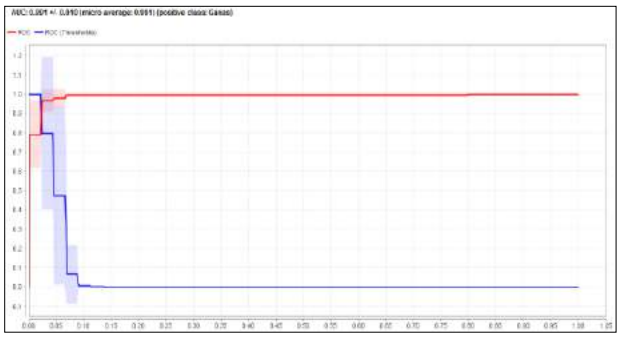


Gambar 12 Hasil *Confusion Matrix Naïve Bayes + PSO*

ISSN : 2580-1503

```
PerformanceVector:
accuracy: 97.81% +/- 1.34% (micro average: 97.80%)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  430     1
Ganas:  14      238
precision: 94.60% +/- 3.43% (micro average: 94.44%) (positive class: Ganas)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  430     1
Ganas:  14      238
recall: 99.58% +/- 1.25% (micro average: 99.58%) (positive class: Ganas)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  430     1
Ganas:  14      238
AUC (optimistic): 0.991 +/- 0.010 (micro average: 0.991) (positive class: Ganas)
AUC: 0.991 +/- 0.010 (micro average: 0.991) (positive class: Ganas)
AUC (pessimistic): 0.991 +/- 0.010 (micro average: 0.991) (positive class: Ganas)
```

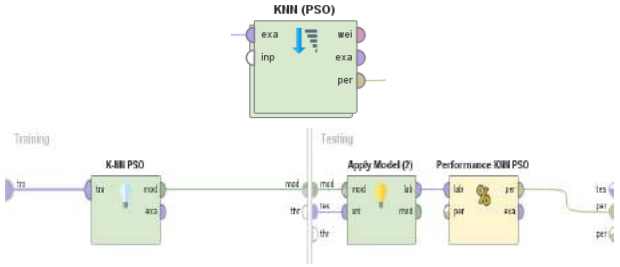
Gambar 13 Hasil *Performance Naïve Bayes + PSO*



Gambar 14 Kurva *ROC Algoritma Naïve Bayes + PSO*

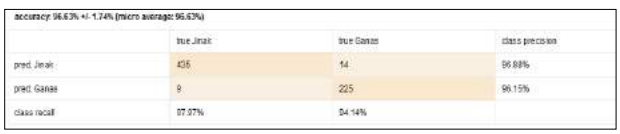
3.4 Algoritma K-Nearest Neighbor + *Particle Swarm Optimization (PSO)*

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



Gambar 15 Konfigurasi Model Algoritma *K-Nearest Neighbor* + PSO

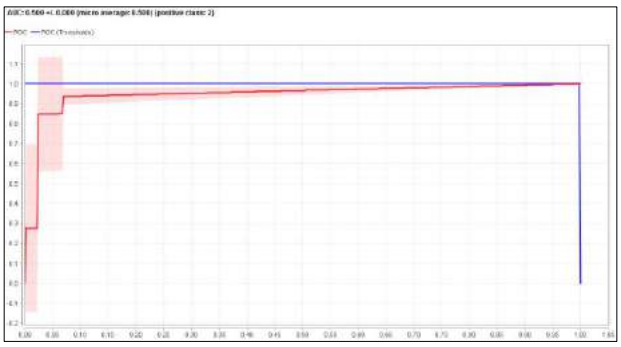
Hasil pengujian dan validasi melalui *confusion matrix* Algoritma *K-Nearest Neighbor + PSO* pada Rapidminer versi 9 tervisualisasi pada gambar 16. Hasil *performance* dengan pengukuran akurasi, *precision*, dan *recall* digambarkan pada gambar 17 sedangkan pada gambar 18 adalah hasil kurva *ROC* Algoritma *K-Nearest Neighbor + PSO*



Gambar 16 Hasil Confusion Matrix Naïve Bayes + PSO

```
PerformanceVector:
accuracy: 96.63% +/- 1.74% (micro average: 96.63%)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  435     14
Ganas:  9       225
precision: 96.35% +/- 3.66% (micro average: 96.15%) (positive class: Ganas)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  435     14
Ganas:  9       225
recall: 94.15% +/- 4.24% (micro average: 94.14%) (positive class: Ganas)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  435     14
Ganas:  9       225
AUC (optimistic): 0.999 +/- 0.001 (micro average: 0.999) (positive class: Ganas)
AUC: 0.500 +/- 0.000 (micro average: 0.500) (positive class: Ganas)
AUC (pessimistic): 0.922 +/- 0.041 (micro average: 0.922) (positive class: Ganas)
```

Gambar 17 Hasil *Performance K-Nearest Neighbor + PSO*



Gambar 18 Kurva ROC *Algoritma K-Nearest Neighbor + PSO*

3.5 Hasil Perbandingan Kinerja Algoritma

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	*Naïve Bayes*	*Naïve Bayes (PSO)*	*KNN*	*KNN (PSO)*
Accuracy	97.37%	97.81%	95.45%	96.63%
Precision	94.31%	94.60%	95.30%	96.35%
Recall	98.71%	99.58%	91.65%	94.15%
AUC	0.990	0.991	0.500	0.500

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 *Naive Bayes* dan *K-Nearest Neighbor*. Pada indikator akurasi berdasarkan evaluasi yang dilakukan secara *confusion matrix* ternyata terbukti bahwa hasil akurasi tertinggi pada perbandingan hasil kinerja klasifikasi tertinggi didapatkan algoritma *Naïve Bayes* berbasis *PSO* sebesar 97.81% disusul algoritma *K-Nearest Neighbor* berbasis *Particle Swarm Optimization (PSO)* sebesar 96,63%.

Ditinjau dari indikator pengukuran *precision* dan *recall* juga mengalami peningkatan yang signifikan diantaranya algoritma *Naïve Bayes* dengan nilai *precision* dan *recall*

sebesar 94.31% dan 98.71% menjadi 94.60% dan 99.58% sedangkan algoritma *K-Nearest Neighbor* dangan nilai *precision* dan *recall* sebesar 95.30% dan 91.65% menjadi 96.35% dan 94.15%.

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 *Naïve Bayes* dengan predikat *Excelent Classification* namun tidak pada algoritma *K-Nearest Neighbor* berdasarkan nilai *AUC* masuk pada predikat *Faillure Classification.*

Dari hasil T-Test pada gambar 19 dapat disimpulkan bahwa algoritma *Naïve Bayes Particle Swarm Optimization (PSO)* memiliki nilai tertinggi sebesar 0,978 disusul algoritma *K-Nearest Neighbor* (PSO) 0,966. Dengan demikian algoritma Naïve Bayes Particle Swarm Optimization (PSO) dapat memberikan solusi terbaik terhadap akurasi pendeteksian penyakit kanker payudara.

A	B	C
	0.978 +/- 0.013	0.966 +/- 0.017
0.978 +/- 0.013		0.108
0.966 +/- 0.017		

Gambar 19 Hasil Uji T-Test

Pairwise t-Test

```
Probabilities for random values with the same result:
-----   0.108
-----   -----
Values smaller than alpha=0.050 indicate a probably significant difference between the mean values!
List of performance values:
0: 0.978 +/- 0.013
1: 0.966 +/- 0.017
```

Gambar 20 Pairwise T-Test

4. Kesimpulan

Kesimpulan

1. Terbukti optimasi *Particle Swarm Optimization* (PSO) dapat meningkatkan kinerja pada akurasi *Naïve Bayes* sebesar 97.37% menjadi 97.81%, dan akurasi *K-Nearest Neighbor* dari 95.45% mengalami peningkatan menjadi 96.63%.
2. Hasil kinerja terbaik yang diuji menggunakan *T-Test* bahwa algoritma *Naïve Bayes* (*PSO*) memiliki nilai tertinggi sebesar 0,978 sedangkan algoritma *K-Nearest Neighbor (PSO)* sebesar 0,966. Dengan demikian algoritma *Naïve Bayes Particle Swarm Optimization (PSO)* dapat memberikan solusi terbaik terhadap akurasi pendeteksian penyakit kanker payudara.

Saran

1. Dibutuhkan jumlah data yang lebih besar, atribut yang lebih kompleks, bahkan menggunakan sampel penyakit lain yang sifatnya memiliki struktur data baru sehingga hasil pengukuran yang dihasilkan akan lebih berguna dan lebih handal akurasinya.
2. Dilakukan metode optimasi untuk meningkatkan kinerja algoritma seperti *Ant Colony Optimization* (ACO), *Genetik Algorithm* (GA), dan lain sebagainya.
3. Melakukan pengujian dan perbandingan pada algoritma lain ataupun menggunakan metode *hybrid* untuk mendapatkan pengetahuan komparasi yang lebih luas.
4. Melakukan pengembangan pada *preprocessing* data dengan menggunakan metode seleksi atribut yang lain seperti *chi-square*, *information index* dan sebagainya untuk ketepatan penyeleksian atribut.

Daftar Pustaka

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

[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

[13] 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.

[14] 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

[15] 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

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

[17] Suyanto, 2017, Data Mining Untuk Klasifikasi Dan Klasterisasi Data, Informatika, Bandung

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

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

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

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

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

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