# PENINGKATAN METODE NAIVE BAYES CLASSIFICATION UNTUK PENENTUAN TINGKAT KEGANASAN KANKER PAYUDARA MENGGUNAKAN PARTICLE SWARM OPTIMIZATION

**Imma Rizki Fitriani**

*Universitas Dian Nuswantoro*
*Email : fitriani.imma@gmail.com*

## ABSTRAK

Kanker payudara merupakan salah satu jenis kanker yang sering ditemukan pada kebanyakan wanita. Pada umumnya pendeteksian tingkat keganasan kanker payudara dilakukan secara prognosis, yaitu "tebakan terbaik atau prediksi" tim medis dalam menentukan sembuh atau tidaknya pasien dari kanker payudara. Penelitian tentang *breast cancer* telah banyak dilakukan untuk mengetahui tingkat keganasan *breast cancer*, dimana secara umum tingkat keganasan kanker payudara diukur dengan memperhatikan stadium penderita kanker payudara yaitu stadium I, II, III, dan IV. Penelitian ini menganalisis tentang pengelompokan data kanker payudara untuk mengetahui kanker tersebut termasuk kanker jinak atau kanker ganas. Untuk mengklasifikasi tingkat keganasan dapat dilakukan dengan pemanfaatan *bioinformatic* menggunakan teknik *data mining* salah satuya dengan algoritma klasifikasi *Naive Bayes Classifier* (NBC). NBC dapat bekerja lebih efektif jika dikombinasikan dengan beberapa prosedur pemilihan atribut seperti *Particle Swarm Optimization* (PSO) untuk membobot atribut. Desain penelitian menggunakan model proses CRISP-DM karena penyelesaian masalah dalam penelitian ini mengarah pada masalah strategi bisnis. Penelitian ini menggunakan *data set* publik *Breast Cancer Wisconsin* (WBC). Dari hasil pengujian dengan *tenfold cross validation* dan *confusion matrix* diketahui bahwa Naive Bayes Classifier (NBC) dalam PSO terbukti memiliki akurasi 96,86%, sedangkan algoritma NBC memiliki akurasi 95,85%. Hasil penelitian ini terbukti bahwa PSO dapat meningkatkan akurasi algoritma NBC.

**Kata kunci :** kanker payudara, klasifikasi , *data mining*, Naive Bayes Classifier, Particle Swarm Optimization

## 1. PENDAHULUAN

Kanker payudara atau *Breast Cancer* merupakan salah satu jenis kanker yang sering ditemukan pada kebanyakan wanita [1]. Kanker payudara terjadi karena pertumbuhan berlebihan atau perkembangan yang tidak terkendali dari sel-sel jaringan payudara [2]. Berdasarkan data Sistem Informasi Rumah Sakit (SIRS) tahun 2007, kanker payudara menempati urutan pertama pada pasien rawat inap di seluruh rumah sakit di Indonesia, yaitu 16,85% [3]. Menurut profil kesehatan Departemen Kesehatan

Republik Indonesia, tahun 2007 kanker yang diderita oleh wanita Indonesia adalah kanker payudara atau *Breast Cancer* dengan angka kejadian 26 per 100.000 perempuan [3]. Pada tahun 2011, *World Health Organization* (WHO) memperkirakan bahwa lebih dari 508.000 wanita diseluruh dunia meninggal karena *breast cancer* [1].

Beberapa peneliti telah menganalisa tingkat keganasan kanker payudara dengan metode klasifikasi menggunakan *data mining*, diantaranya yang dilakukan oleh Bellaachia,dkk [4] menggunakan *Naive Bayes Classifier* (NBC), C4.5 (information gain), dan Artificial Neural Network (ANN) untuk memprediksi kanker payudara.

Dari hasil penelitian Bellaachia,dkk [4] algoritma NBC untuk penentuan tingkat keganasan kanker payudara hasil akurasinya masih kurang dibanding menggunakan algoritma C4.5. Namun, NBC mempunyai akurasi dan kecepatan yang tinggi saat diterapkan pada data yang besar [5]. NBC dapat menangani data yang tidak lengkap (*missing value)* serta kuat terhadap atribut yang tidak relevan dan *noise* pada data [6]. NBC akan bekerja lebih efektif jika dikombinasikan dengan beberapa prosedur pemilihan atribut [7].

## 2. NAIVE BAYES

*Naive Bayes Classifier* disebut juga *Bayesian Classification* merupakan metode pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan dari suatu *class.* NBC didasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi serupa *decision tree* dan *neural network*. Selain itu, NBC terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam *database* yang besar [5].

Berikut penjelasan mengenai metode NBC [8]:

1) Setiap data dipresentasikan sebagai vector berdimensi-n yaitu $X=(x_1,x_2,x_3,.....,x_n)$, dimana n adalah gambaran dari ukuran yang dibuat di test dari n atribut yaitu $A_1$,-$A_2$,$A_3$,.....$A_n$.
2) m adalah kumpulan kategori yaitu $C_1,C_2,C_3,.....,C_m$. Diberikan data test $X$ yang tidak diketahui kategorinya, maka *classifier* akan memprediksi bahwa $X$ adalah milik kategori dengan posterior probability tertinggi berdasarkan kondisi $X$. Oleh karena itu, NBC menandai bahwa test $X$ yang tidak diketahui tadi ke kategori $C_i$ jika dan hanya jika :

$$P(Ci|X) > P(Cj|X) \text{ untuk } 1 \leq j \leq m, j \neq 1$$

Kemudian memaksimalkan $P(C_i \mid X)$. Class $C_i$ dari $P(C_i \mid X)$ yang dimaksimalkan biasa disebut *maximum posteriori hypothesis.*

$$P(X|Ci) = \frac{P(X|Ci)\,P(Ci)}{P(X)}$$

*3)* $P(X)$ adalah konstan untuk semua kategori, hanya $P(X \mid C_i)$. $P(C_i)$ yang perlu dimaskimalkan. Jika *class prior probability* tidak diketahui, maka akan diasumsikan sama dengang hasil dari kategori-kategori yang lain seperti $P(C_1)=P(C_2)=.....$ $P(C_m)$ dan oleh karena itu kita akan memaksimalkan $P(X|C_i).P(C_i)$. Perlu dicatat bahwa *class prior probability* mungkin diperkirakan dengan perhitungan $P(C_i) = s_i$ dimana $s_i$ adalah jumlah dari data training $s$

dari kategori $C_i$ dan *s* adalah jumlah total data training.

4) Diberikan data dengan banyak atribut, ini akan menjadi komputasi yang kompleks untuk mengomputasi $P(X \mid C_i)$. Untuk mengurangi komputasi pada saat mengevaluasi $P(X \mid C_i)$, maka dapat dihitung menggunakan perhitungan :

$$P(\boldsymbol{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$
$$= P(x_1|C_i) \times P(x_2|C_i) \times \cdots \times P(x_n|C_i).$$

Dimana $x_k$ adalah nilai-nilai atribut dalam sampel *X* dan probabilitas $P(x_1 \mid C_i)$, $P(x_2 \mid C_i)$,....., $P(x_n \mid C_i)$ dapat diperkirakan dari data training. Jika P(X|Ci) sama dengan nol, maka menggunakan pendekatan estimasi sebagai berikut [9]:

$$P(X|Ci) = \frac{n_c + n_{equiv}\, p}{n + n_{equiv}}$$

Dimana n merupakan total dari jumlah *record* dari kelas *Ci*, $n_c$ adalah jumlah contoh training dari kelas X yang menerima nilai *Ci*, $n_{equiv}$ adalah nilai konstan dari ukuran sampel yang equivalen. *P* adalah peluang estimasi prior, *P=1/k* dimana *k* adalah jumlah kelas dalam variabel target.

## 3. PARTICLE SWARM OPTIMIZATION

Particle Swarm Optimization (PSO) merupakan algoritma pencarian berbasis populasi yang diinisialisasi dengan populasi solusi acak, dan digunakan untuk memecahkan masalah optimasi [10]. PSO diperkenalkan oleh Kennedy dan Eberhart pada tahun 1995 berdasarkan penelitian terhadap perilaku kawanan burung dan ikan. Setiap partikel dalam PSO juga dikaitkan dengan kecepatan partikel terbang melalui ruang pencarian dengan kecepatan yang dinamis disesuaikan untuk perilaku historis mereka. Oleh karena itu, partikel memiliki kecenderungan untuk terbang menuju daerah pencarian yang lebih baik dan lebih baik selama proses pencarian [10].

Rumus untuk menghitung perpindahan posisi dan kecepatan partikel yaitu [11]:

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

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

Dimana :
$V_i(t)$= Kecepatan partikel *i* saat iterasi *t*
$X_i(t)$= posisi partikel *i* saat iterasi *t*
$c_1$ dan $c_2$= *learning rates* untuk kemaampuan individu (*cognitive*) dan pengaruh sosial (*group*)
$r_1$ dan $r_2$= bilangan random yang berdistribusi uniformal dalam interval 0 dan 1
$Xpbest_i$=posisi terbaik partikel *i*
XGbest= posisi terbaik global

## 4. NAIVE BAYES DALAM PARTICLE SWARM OPTIMIZATION

PSO diterapkan pada pembobotan atribut seperti algoritma dibawah ini :

- Identifikasi populasi sampel
- Hitung $P(C_i)$ pada setiap kelas
- Inisialisasi posisi setiap patikel atribut ke-j

- Untuk Setiap Atribut dilakukan
    - Evaluasi nilai fungsi tujuan
    - Cari Pbest dan Gbest
    - Update kecepatan dan posisi particle
    - Gbest = bobot atribut ke-j
- hitung P(X|Ci), i=1,2 untuk setiap kelas atau atribut
- Bandingkan hasil P(X|$C_i$)

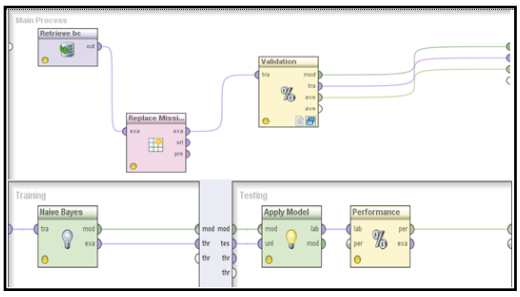
Identifikasi populasi sampel dari *data set Wisconsin Breast Cancer* (WBC). Hitung P($C_i$) untuk setiap kelas, dalam kasus *data set* pada penelitian ini terdiri dari 2 kelas yaitu jinak dan ganas.

Inisialisasi posisi setiap partikel atribut ke-j merupakan awal dari tahap pembobotan atribut dengan PSO. Langkah selanjutnya adalah evaluasi nilai fungsi tujuan dari setiap partikel untuk mendapatkan posisi terbaik (Pbest) dan posisi global terbaik (Gbest), kemudian *update* kecepatan dan posisi partikel. Ulangi langkah evaluasi nilai fungsi tujuan sampai mencapai konvergen, kemudian Gbest = bobot atribut ke-j. Cek apakah nilai j sudah maksimal, jika belum ulangi langkah-langkah dari inisialisasi posisi setiap partikel atribut ke-j sampai menemukan bobot atribut ke-j. Ulangi langkah tersebut sampai nilai j sudah maksimal atau semua atribut sudah terbobot.

Kemudian hitung P(X|Ci), i=1,2 untuk setiap kelas atau atribut. Setelah itu bandingkan, jika P(X|$C_1$) > P(X|$C_2$) maka kesimpulannya adalah $C_1$ atau dalam kasus pada penelitian ini berarti kanker jinak. Jika P(X|$C_1$) < P(X|$C_2$) maka kesimpulannya $C_2$ atau kanker ganas.

## 5. EKPERIMEN

Data yang digunakan pada penelitian ini menggunakan *public data set* berasal dari University of California, Irvine (UCI) Machine Learning dengan judul Wisconsin Breast Cancer (Original). Data ini berjumlah 699 *record* dan terdiri dari 11 atribut, dengan 10 atribut bertipe numerik dan 1 bertipe kategorikal [12],[13]. Dilakukan proses eliminasi pada atribut *sample code number*, sehingga hanya 10 atribut yang digunakan dengan 9 atribut sebagai variabel predictor dan 1 atribut sebagai variabel tujuan/target. Terdapat 16 data missing pada atribut *bare nucleui*. Untuk menangani data missing tersebut, dilakukan proses *replace missing value* dengan model *average* berdasarkan jumlah data.

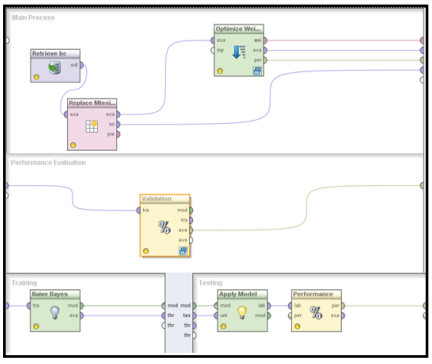


Gambar 1. Desain model NBC

Hasil dari model di atas menghasilkan nilai akurasi confusion matrix sebesar 95,85%.

Pada model NBC-PSO, pertama dilakukan uji coba dengan memberi nilai pada parameter *population size* secara default 5, 10-600 dengan *maximum number of generation* 100 bernilai konstan. *Population size* adalah jumlah individual pada tiap generasi, sedangkan *maximum number of generation* adalah jumlah

generasi maksimum untuk menghentikan jalannya algoritma. Terpilih nilai *population size* terbaik adalah 10 dengan hasil akurasi 96,86 %.



Gambar 2. Desain model NBC-PSO

Selanjutnya dilakukan percobaan dengan population size bernilai tetap 10 dan maximum number of generation bernilai 100-1500. Akurasi tertinggi dan waktu eksekusi terendah terjadi pada saat *maximum number of generation* bernilai 100 dengan nilai akurasi sebesar 96,86%.

## 6. HASIL

Berdasarkan hasil percobaan, diperoleh akurasi NBC-PSO paling tinggi terjadi pada saat *population size* bernilai 10 dan *maximum number of generation* bernilai 100. Akurasi NBC-PSO 96,86%, sedangkan akurasi NBC 95,85%.

Tabel 1. Komparasi akurasi NBC dan NBC-PSO

| Perbandingan | NBC | NBC-PSO |
|---|---|---|
| Akurasi confusion matrix (%) | 95,85 | 96,86 |

Tabel 1. Hasil Pembobotan Atribut PSO

| Atribut | Bobot |
|---|---|
| Clump Thickness | 1 |
| Uniformity of Cell Size | 1 |
| Uniformity of Cell Shape | 1 |
| Marginal Adhesion | 0 |
| Single Epithelial Cell Size | 1 |
| Bare Nuclei | 1 |
| Bland Chromatin | 0 |
| Normal Nucleoli | 0 |
| Mitoses | 0 |

Hasil pembobotan atribut yaitu 4 atribut mempunyai bobot 0, 5 atribut mempunyai bobot 1. Sehingga atribut yang berbobot 0 dapat dihilangkan karena tidak mempunyai pengaruh pada akurasi penentuan tingkat keganasan kanker payudara.

## 7. KESIMPULAN

Algoritma naive bayes classifier (NBC) dengan PSO dan algoritma NBC tanpa PSO, dapat diterapkan untuk penentuan tingkat keganasan kanker payudara. Hasil penelitian menunjukan bahwa model NBC-PSO memiliki akurasi yang lebih baik dengan 96,86 % dibandingkan model NBC dengan akurasi 95,85%. Dari 9 atribut terdapat 4 atribut mempunyai bobot

0, serta 5 atribut mempunyai bobot 1. Sehingga atribut yang berbobot 0 dapat dihilangkan karena tidak mempunyai pengaruh pada akurasi penentuan tingkat keganasan kanker payudara.

Dengan demikian terbukti bahwa penerapan PSO pada pembobotan atribut NBC dapat meningkatkan nilai akurasi. Hal ini menjadikan NBC-PSO memberikan solusi pemecahan malasah dalam penentuan tingkat keganasan kanker payudara.

## 8. DAFTAR PUSTAKA


[1]WHO,"[Online].Available:http://www.who.int/cancer/detection/braestcancer/en/index1.html. [Accessed 10 Januari 2014].

[2][Online].Available:http://www.breastcancer.org/symptoms/understand_bc/what_is_bc. [Accessed 10 Januari 2014].

[3]"DinasKesehatanNasional,"[Online].Available:http://www.depkes.go.id/index.php/berita/press-release/1060-jikatidak-dikendalikan-26-juta-orang-di-dunia-menderita-kanker-.html. [Accessed 10 Januari 2014].

[4] Abdelghani Bellaachia, Erhan Guven. Predicting Breast Cancer Survivability Using Data Mining Techniques. 2006

[5] Kusrini dan E. T. Luthfi, Algoritma Data Mining, Yogyakarta: ANDI, 2009.

[6] Gurunescu,F.(2011). Data mining concept, models, and techniques. Verlag berlin Heidelberg : Springer

[7] Witten,I.H., Frank, E, and Hall, M.A.(2011). Data Mining Practical Machine Learning Tools And Techniques. Burlington, USA: Morgan Kaufmann Publishers.

[8]J. Han dan M. Kamber, Data Mining : Concepts and Techniques, Third Edition, San Fransisco: Morgan Kaufmann, 2012.

[9]D. T. Larose, Data Mining Method And Models, Hoboken, New Jersey: John Wiley & Sons, Inc., 2006.

[10] A. Abraham, C. Grosan and V. Ramos, Swarm Intelligence In Data Mining, Verlag Berlin Heidelberg: Springer, 2006.

[11]J. Lin and J. Yu, "Weighted Naive Bayes Classification Algorithm Based On Particle Swarm Optimization," *Communication Software and Networks (ICCSN), IEEE 3rd International Conference on* , pp. 444-447, 2011.

[12] "UCI Machine Learning repository Breast Cancer Wisconsin," University CaliforniaIrvine,[Online].Available:http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28-Original%29. [Accessed 20 November 2013].

[13] W. H. Wolberg and O. L. Mangasarian, "Multi Surface method of Pattern Separation For Medical Diagnosis Applied To Breast Cytology," *Proc. Natl. Acad. Sci. USA,* vol. 87, pp. 9193-9196, 1990.