# PENGGABUNGAN ALGORITMA *BACKWARD ELIMINATION* DAN *K-NEAREST NEIGHBOR* UNTUK MENDIAGNOSIS PENYAKIT KANKER PAYUDARA

**Laily Hermawanti**
Program Studi Teknik Informatika Fakultas Teknik Universitas Sultan Fatah (UNISFAT)
Jl. Diponegoro No. 1B Jogoloyo Demak Telp (0291) 686227

**Abstrak :** *K-Nearest Neighbor* merupakan salah satu algoritma yang diusulkan oleh para peneliti *data mining* di bidang kesehatan misalnya penyakit kanker payudara. Penyakit kanker payudara merupakan salah satu penyakit berbahaya dan penyebab kematian di seluruh dunia. Maka dari itu, penyakit kanker payudara perlu didiagnosis. Algoritma yang digunakan dalam penelitian ini adalah penggabungan algoritma *Backward Elimination* dan *K-Nearest Neighbor* (KNN) untuk meningkatkan akurasi dalam diagnosis penyakit kanker payudara. Penelitian ini menggunakan *dataset* kanker payudara yang diperoleh dari *Wisconsin Breast Cancer* (WBC) UCI *Dataset Machine Learning Repository*. Parameter-parameter yang digunakan pada *Data set Wisconsin Breast Cancer* (WBC) adalah *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*. Hasil penelitian ini, pada dataset kanker payudara, algoritma *K-Nearest Neighbor* memiliki nilai akurasi sebesar 90.14% +/- 2.17% dan nilai *Area Under Curve* (AUC) sebesar 0,900 yang termasuk dalam kategori klasifikasi sangat baik (excellent classification). Pada dataset kanker payudara, penggabungan algoritma *Backward Elimination* dan K-Nearest Neighbor memiliki nilai akurasi sebesar 97.28% +/- 1.78% dan nilai *Area Under Curve* (AUC) sebesar 0,991 yang termasuk dalam kategori klasifikasi sangat baik (*excellent classification*). Penggabungan algoritma *Backward Elimination* dan *K-Nearest Neighbor* (KNN) tingkat akurasinya lebih tinggi dari pada algoritma *K-Nearest Neighbor* (KNN) dalam mendiagnosis penyakit kanker payudara.

**Kata Kunci**: *Backward Elimination, K-Nearest Neighbor*, penyakit kanker payudara

## PENDAHULUAN

Kanker payudara adalah kanker yang paling umum pada wanita dan penyebab utama kematian kanker di seluruh dunia (E. Technical and P. Series, 2006). Meskipun etiologi kanker payudara tidak diketahui, faktor risiko berbagai kemungkinan mempengaruhi perkembangan penyakit ini termasuk faktor genetik, hormonal, dan lingkungan. Selama beberapa dekade terakhir, risiko kanker payudara meningkat di negara-negara industri dan berkembang sebesar 1% -2% per tahun, tingkat kematian akibat kanker payudara menurun sedikit (E. Technical and P. Series, 2006). Parameter-parameter kanker payudara terdiri dari (E. Technical and P. Series, 2006) :

- *Clump thickness*: sel *benign* cenderung dikelompokkan dalam monolayers, sementara sel-sel kanker sering dikelompokkan dalam multilayers.
- *Uniformity of cell size*: sel-sel kanker mempunyai ukuran bervariasi.
- *Uniformity of cell shape*: sel-sel kanker mempunyai bentuk bervariasi.
- *Marginal adhesion*: sel-sel normal cenderung tetap bersama-sama.
- *Single epithelial cell size*: sel-sel epitel yang signifikan diperbesar menjadi sel *malignant.*

- *Bare nuclei*: adalah istilah yang digunakan untuk inti (*nuclei*) yang tidak dikelilingi oleh cytoplasm (seluruh sel). Biasanya terlihat di benign.
- *Bland Chromatin*: inti "tekstur" seragam yang dilihat dalam sel *benign*. Dalam sel-sel kanker chromatin cenderung lebih kasar.
- *Normal nucleoli*: *nucleoli* adalah struktur kecil yang terlihat dalam inti atom. Pada sel-sel normal nucleolus biasanya sangat kecil jika terlihat sama sekali. Dalam sel-sel kanker nucleoli menjadi lebih menonjol.
- *Mitoses* : pembelahan satu sel menjadi dua sel.
- *Class* : kelas.

*Data mining* dapat diaplikasikan di bidang kesehatan misalnya mendiagnosis penyakit kanker payudara, penyakit jantung, penyakit diabetes dan lain-lain (D. T. Larose, 2005). Terdapat beberapa metode dalam mendiagnosis penyakit kanker payudara misalnya *K-Nearest Neighbor* (H. A. Fayed and A. F. Atiya, 2009), *Naïve Bayes* (J. Wu dan Z. Cai, 2011) dan lain-lain.

Penelitian yang dilakukan oleh H. A. Fayed dan A. F. Atiya menggunakan metode *K-Nearest Neighbor* (KNN). Penelitian ini menggunakan *data set* dari *UCI machine learning depository*. Permasalahan KNN adalah komputasi dan penyimpanan. KNN memerlukan penyimpanan seluruh *training set* yang menjadi penyimpanan untuk *data set* yang besar dan waktu komputasi yang lama pada tahap klasifikasi. Penelitian ini mengusulkan algoritma kondensasi baru. Algoritma baru tersebut adalah *Template Reduction for* KNN (TRKNN). Pendekatan TRKNN yang diusulkan untuk mengurangi ukuran *template set*. Metode TRKNN mempunyai kelebihan yaitu implementasinya sederhana dan komputasinya cepat. Algoritma TRKNN dengan menghasilkan akurasi sebesar 95% (H. A. Fayed and A. F. Atiya, 2009).

Penelitian yang dilakukan oleh J. Wu dan Z. Cai menggunakan banyak metode efektif untuk meningkatkan performa *Naive Bayes.* Pembahasan ini mengevaluasi performa konfigurasi baru (DE-WNB) pada 36 UCI seluruh standar *data set* dalam sistem Weka. Hasil eksperimen menunjukkan akurasi klasifikasi algoritma baru DE-WNB lebih tinggi dari algoritma lain yang digunakan untuk membandingkan. Algoritma *Differential Evolution Weighted Naïve Bayes* (DE-WNB) menghasilkan keakuratan sebesar 73.09%±7.51% untuk *dataset* kanker payudara *wisconsin*. Algoritma *Naïve Bayes* (NB) menghasilkan keakuratan sebesar 72.94%±7.71%. Algoritma *Gain Ratio-*

*Weighted Naïve Bayes* (GR-WNB) menghasilkan keakuratan sebesar 70.30%±1.37%. Algoritma *Correlation-based Feature Selection-Weighted Naïve Bayes* (CFS-WNB) menghasilkan keakuratan sebesar 71.73%±7.40%. Algoritma *Mutual Information-Weighted Naïve Bayes* (MI-WNB) menghasilkan keakuratan sebesar 70.30%±1.37%. Algoritma *Tree-Weighted Naïve Bayes* (*Tree*-WNB) menghasilkan keakuratan sebesar 72.39%±7.47% (J. Wu dan Z. Cai, 2011).

Dari penelitian-penelitian yang pernah dilakukan tentang diagnosis penyakit kanker payudara terutama yang menggunakan algoritma *K-Nearest Neighbor*, akurasinya belum tinggi. Kelebihan-kelebihan spesifik model penggabungan algoritma *K-Nearest Neighbor* dan *Backward Elimination* pada penyakit kanker payudara yang akan diteliti dibanding teknik-teknik diagnosis lain yaitu *Backward Elimination* dapat mereduksi ukuran *data set* sehingga dapat meningkatkan akurasi pada *Backward Elimination* (J. Han and M. Kamber, 2006). Maka dari itu, penelitian ini menggunakan penggabungan algoritma *Backward Elimination* dan *K-Nearest Neighbor* untuk mendiagnosis penyakit kanker payudara sehingga dapat meningkatkan akurasi dibandingkan dengan penelitian-penelitian sebelumnya.

## METODOLOGI

Penelitian ini menggunakan proses *Cross-Standard Industry-Data Mining* (CRISP-DM) dengan tahap-tahap penelitian meliputi pemahaman bisnis, pemahaman data, pengolahan data, pemodelan dan evaluasi (Larose, 2005).

***Tahap Pemahaman Bisnis***

Penelitian ini dilakukan untuk menerapkan penggabungan algoritma *Backward Elimination* dan *K-Nearest Neighbor* untuk meningkatkan akurasi dalam mendiagnosis penyakit kanker payudara.

***Tahap Pemahaman Data***

Penelitian ini mengambil *dataset* kanker payudara dari *UCI Machine Learning* (Frank dan Asuncion, 2010).

***Tahap Pengolahan Data***

Teknik-teknik pengolahan data awal (data *pre-processing*) yang digunakan pada penelitian ini adalah (Han dan Kamber, 2006) :

1. Data *cleaning* dapat digunakan untuk data yang *missing value*. Karena ditemukan adanya data yang terlewat tidak terisi (*missing value*) pada data. Pengolahan data awal dilakukan untuk mengisi nilai yang *missing value* dengan pekerjaan *replace missing value* dilakukan.
2. Data *reduction* digunakan untuk menghasilkan *data set* yang volumenya lebih kecil. Salah satu

strategi *data reduction* yang digunakan pada penelitian ini adalah *attribute subset selection. Attribute subset selection* digunakan untuk mereduksi ukuran *data set* dengan menghilangkan atribut-atribut yang tidak relevan atau *redudant*. Salah satu teknik *attribute subset selection* yang digunakan pada penelitian ini adalah *Backward Elimination.*

***Tahap Pemodelan***

Model yang digunakan dalam tahap ini menggunakan penggabungan algoritma *Backward Elimination* dan *K-Nearest Neighbor*.

## KAJIAN PUSTAKA

### Algoritma *K-Nearest Neighbor*

Algoritma *K-Nearest Neighbor* merupakan salah satu algoritma yang digunakan untuk klasifikasi, meskipun juga dapat digunakan untuk estimasi dan prediksi (Larose, 2005). *K-Nearest Neighbor* adalah contoh algoritma berbasis pembelajaran, di mana *data set* pelatihan (*training*) disimpan, sehingga klasifikasi untuk *record* baru yang tidak diklasifikasi didapatkan dengan membandingkannya dengan *record* yang paling mirip dengan *training set* (Larose, 2005). Langkah-langkah algoritma *K-Nearest Neighbor* adalah (Larose, 2005):

1. Menentukan parameter k, misal k = 5.
2. Menghitung jarak (*similarity*) di antara semua *training records* dan objek baru.
3. Pengurutan data berdasarkan nilai jarak dari nilai yang terkecil sampai terbesar.
4. Pengambilan data sejumlah nilai k (misal k=5).
5. Menentukan label yang frekuensinya paling sering di antara *k training records* yang paling dekat dengan objek.

### Algoritma *Backward Elimination*

*Backward Elimination* menghilangkan atribut-atribut yang tidak relevan (Han dan Kamber, 2006). Algoritma *Backward Elimination* didasarkan pada model regresi linear (Noori, dkk., 2011). Langkah-langkah *Backward Elimination* adalah :

1. Mulai semua variabel pada model F-statistik parsial dihitung setiap variabel pada model.
2. Menentukan variabel dengan F-statistik parsial terkecil dan menguji Fmin.
3. Jika Fmin tidak signifikan, dalam kasus ini, variabel dihilangkan dari model.
4. Menentukan variabel dengan F-statistik parsial .
5. Pada sisi lain, variabel dengan F-statistik terkecil adalah variabel

indicator. Bagaimanapun, p-value diasosiasikan dengan Fmin tidak cukup membenarkan model yang tidak inklusi (*noninclusion*) menurut kriteria (lebih dari bit). Maka dari itu, prosedur menghasilkan dan melaporkan model sebagai berikut :

$y = \beta_0 + \beta_1$(*single epithelial cell size*) + $\beta_2$(*normal nucleoli*) + $\beta_3$(*marginal adhesion*) + $\varepsilon$

6. Menghitung F-test parsial.

**Algoritma *Backward Elimination* – *K-Nearest Neighbor***

Langkah-langkah algoritma *Backward Elimination* - *K-Nearest Neighbor* adalah sebagai berikut :

1. Mulai semua variabel pada model F-statistik parsial dihitung setiap variabel pada model.
2. Menentukan variabel dengan F-statistik parsial terkecil dan menguji Fmin.
3. Jika Fmin tidak signifikan, dalam kasus ini, variabel dihilangkan dari model.
4. Menentukan variabel dengan F-statistik parsial.
5. Menentukan variabel dengan F-statistik terkecil sebagai variabel indikator. Bagaimanapun, p-value dan Fmin tidak hanya menggunakan model yang tidak inklusi *(noninclusion*) menurut kriteria (lebih dari bit). Maka dari itu, prosedur menghasilkan dan melaporkan model sebagai berikut :

   $y = \beta_0 + \beta_1$(*marginal adhesion*) + $\beta_2$(*single epithelial cell size*) + $\beta_3$(*normal nucleoli*) + $\varepsilon$

6. Menentukan atribut-atribut yang dipilih oleh *Backward Elimination.*
7. Menentukan parameter k, misal k = 5.
8. Menghitung jarak (*similarity*) di antara semua *training records* dan objek baru.
9. Pengurutan data berdasarkan nilai jarak dari nilai yang terkecil sampai terbesar.
10. Pengambilan data sejumlah nilai k (misal k=5).
11. Menentukan label yang frekuensinya paling sering di antara *k training records* yang paling dekat dengan objek.

**Tahap Evaluasi**

Evaluasi dan validasi pada penelitian ini menggunakan *confusion matrix* (*accuracy*) dan ROC *Curve.*

## HASIL DAN PEMBAHASAN

Akurasi *dataset* kanker payudara dapat dilihat pada Tabel 1. *Area Under Curve* (AUC) *dataset* kanker payudara dapat dilihat pada Tabel 2. Pada tabel 1, algoritma *K-Nearest Neighbor*

menggunakan *dataset* kanker payudara menghasilkan akurasi sebesar 90.14% +/- 2.17%, sedangkan algoritma *Backward Elimination-K-Nearest Neighbor* menghasilkan akurasi sebesar 97.28% +/- 1.78% sehingga mengalami peningkatan akurasi. Hasil pada tabel 1, akurasi metode *Backward Elimination-K-Nearest Neighbor* lebih tinggi dari algoritma *K-Nearest Neighbor*.

**Tabel 1. Akurasi *Dataset* Kanker Payudara**

| Algoritma | Akurasi (%) |
|---|---|
| *K-Nearest Neighbor* | 90.14% +/- 2.17% |
| *Backward Elimination – K-Nearest Neighbor* | 97.28% +/- 1.78% |

**Tabel 2. *Area Under Curve* (AUC) *Dataset* Kanker Payudara**

| Algoritma | *Area Under Curve* (AUC) |
|---|---|
| *K-Nearest Neighbor* | 0.900 |
| *Backward Elimination – K-Nearest Neighbor* | 0.991 |

*Area Under Curve* (AUC) *dataset* kanker payudara dapat dilihat pada Tabel 2. Pada tabel 2, algoritma *K-Nearest Neighbor* menggunakan *dataset* kanker payudara menghasilkan *Area Under Curve* (AUC) sebesar 0.900 yang termasuk dalam kategori klasifikasi sangat baik *(excellent classification)*. Algoritma *Backward Elimination-K-Nearest Neighbor* menghasilkan AUC sebesar 0,991 yang termasuk dalam kategori "klasifikasi sangat baik (*excellent classification*)".

Hasil menunjukkan metode *Backward Elimination-K-Nearest Neighbor* dapat mencapai akurasi yang tinggi dalam mendiagnosis penyakit kanker payudara. Percobaan ini dilakukan untuk menunjukkan peningkatan akurasi dari algoritma *K-Nearest Neighbor* menjadi *Backward Elimination-K-Nearest Neighbor*.

## KESIMPULAN

Penggabungan algoritma *Backward Elimination* dan *K-Nearest Neighbor* tingkat akurasinya lebih tinggi dari pada algoritma *K-Nearest Neighbor* dalam mendiagnosis penyakit kanker payudara. Penelitian ini menunjukkan penggabungan algoritma *Backward Elimination* dan *K-Nearest Neighbor* merupakan salah satu algoritma yang tepat dalam mendiagnosis penyakit kanker payudara.

## DAFTAR PUSTAKA

E. Technical and P. Series, *Guidelines for management of breast cancer*. World Health Organization, 2006.

Larose, D.T., (2005), *Discovering Knowledge in Data: An Introduction to Data Mining*. United States of America: John Wiley & Sons, Inc.

H. A. Fayed and A. F. Atiya, (2009), *"A Novel Template Reduction Approach for the -Nearest Neighbor Method,"* vol. 20, no. 5, pp. 890–896, 2009.

Wu, J. and Cai, Z. (2011), "*Attribute Weighting via Differential Evolution Algorithm for Attribute Weighted Naive Bayes ( WNB )*," vol. 5, pp. 1672–1679.

Han, J. and Kamber, M. (2006), *Data Mining Concept dan Techniques*, 2nd ed. United States of America: Diane Cerra.

Witten, I.H., Frank, E., and Hall, M.A., (2011), *Data mining: Practical Machine Learning Tools and Techniques*, 3rd ed. USA: Kauffmann, Morgan.

Gorunesco, F., 2011, *Data Mining Concept Model Technique*. Romania: Springer.

Noori, R., Karbassi, A.R., A. Moghaddamnia, Han, D., Zokaei-ashtiani, M.H., and Farokhnia, A., (2011), "Assessment of input variables determination on the SVM model performance using PCA , Gamma test , and forward selection techniques for monthly stream flow prediction," *Journal of Hydrology*, vol. 401, no. 3–4, pp. 177–189.

Larose, D.T., (2007), *Data Mining Methods and Models*. New Jersey, Canada: John Wiley & Sons, Inc.

Frank, A. and Asuncion, A., (2010), "UCI Machine Learning Repository" http://archive.ics.uci.edu/ml/datasets.html, Irvine, CA: University of California, School of Information and Computer Science.