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

**Laily Hermawanti*, Sucianna Ghadati Rabiha**
Jurusan Teknik Informatika, Fakultas Teknik, Universitas Sultan Fatah
Jl. Diponegoro 1A, Jogoloyo - Demak.
*Email: lailyhermawanti@gmail.com

**Abstrak**

*K-Nearest Neighbor merupakan salah satu algoritma yang diusulkan oleh para peneliti data mining di bidang kesehatan misalnya penyakit jantung. Penyakit jantung adalah salah satu penyakit berbahaya dan penyebab kematian di seluruh dunia. Maka dari itu, penyakit jantung perlu didiagnosis. Algoritma yang digunakan dalam penelitian ini adalah penggabungan algoritma Backward Elimination dan K-Nearest Neighbor (KNN) untuk meningkatkan akurasi dalam diagnosis penyakit jantung. Penelitian ini menggunakan dataset jantung yang diperoleh dari UCI Dataset Machine Learning Repository. Hasil penelitian ini, pada dataset jantung, algoritma K-Nearest Neighbor memiliki nilai akurasi sebesar 88.62% +/- 0.09% dan nilai Area Under Curve (AUC) sebesar 0.958 +/- 0.000 yang termasuk dalam kategori klasifikasi sangat baik (excellent classification). Pada dataset jantung, penggabungan algoritma Backward Elimination dan dan K-Nearest Neighbor memiliki nilai akurasi sebesar 89.55% +/- 6.01% dan nilai Area Under Curve (AUC) sebesar 0.966 +/- 0.056 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 jantung.*

***Kata kunci** : Backward Elimination, K-Nearest Neighbor, penyakit jantung*

## 1. PENDAHULUAN

Penyakit jantung adalah salah satu penyakit yang menyebabkan kematian di Amerika (Promotion, 2009). Lebih dari 600.000 orang-orang Amerika meninggal setiap tahun disebabkan penyakit jantung. Istilah "penyakit jantung" menjelaskan beberapa tipe kondisi jantung (Promotion,2009). Salah satu tipe penyakit jantung adalah *coronary artery disease*, yang menyebabkan serangan jantung. Jenis penyakit jantung yang lain termasuk katup jantung atau jantung yang tidak terpompa dengan baik dan menyebabkan gangguan jantung. Beberapa orang meninggal karena penyakit jantung (Promotion,2009). Maka dari itu penyakit jantung perlu didiagnosis.

*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 jantung misalnya *K-Nearest Neighbor (*M. Moradian dan A. Baraani, 2009), *Neural Network (*M. Moradian and A. Baraani, 2009), *Naïve Bayes* (Wu dan Cai, 2011) dan lain-lain.

Menurut Moradian dan Barani (2009), algoritma K-Nearest Neighbor (KNN) adalah salah satu algoritma klasifikasi yang digunakan sebagian besar dalam aplikasi yang berbeda. Salah satu masalah dari algoritma KNN adalah semua atribut dalam menghitung jarak antara record baru dan record yang tersedia dalam dataset training. Hal ini menyebabkan proses klasifikasi yang tidak baik dan menurunkan akurasi algoritma klasifikasi. Pendekatan utama untuk menangani masalah ini adalah untuk atribut-atribut bobot yang berbeda ketika menghitung jarak antara dua record. Dalam pembahasan ini, menggunakan aturan-aturan asosiasi untuk atribut-atribut bobot dan menyarankan algoritma klasifikasi baru K-Nearest Neighbor Based Association (KNNBA) yang dapat meningkatkan akurasi algoritma KNN. Pengujian ini menggunakan 15 UCI data set, dan dibandingkan dengan yang algoritma klasifikasi lain Naïve Bayes (NB), Neural Network (NN), J4.8, NBTREE, VFI, LWL dan IBK. Algoritma Naïve Bayes (NB) menggunakan dataset jantung menghasilkan akurasi sebesar 83.707%. Algoritma Neural Network (NN) menggunakan dataset jantung menghasilkan akurasi sebesar 78.15%. Algoritma J48 menggunakan dataset jantung menghasilkan akurasi sebesar 76.67%. Algoritma Naïve Bayes Tree (NBTree) menggunakan dataset jantung menghasilkan akurasi sebesar 80.37%. Algoritma VFI menggunakan dataset

jantung menghasilkan akurasi sebesar 80%. Algoritma LWL menggunakan dataset jantung menghasilkan akurasi sebesar 71.85%. Algoritma IBK menggunakan dataset jantung menghasilkan akurasi sebesar 81.48%. Algoritma KNNBA menggunakan dataset jantung menghasilkan akurasi sebesar 81.487%.

Menurut Wu dan Cai (2011), mengusulkan banyak metode efektif untuk meningkatkan kinerja *Naïve Bayes* dengan menggabungkan metode-metode seperti *backwards sequential elimination, lazy elimination* dan lain-lain. Mengujikan algoritma baru menggunakan 36 data set dari UCI Repository diseleksi dengan perangkat lunak Weka dan dibandingkan dengan algoritma *Differential Evolution Weighted Naïve Bayes* (DE-WNB), *Naïve Bayes* (NB), *Gain Ratio-Weighted Naïve Bayes* (GR-WNB), MI-WNB, *Correlation-based Feature Selection-Weighted Naïve Bayes* (CFS-WNB) dan *Tree-Weighted Naïve Bayes* (*Tree*-WNB). Algoritma *Differential Evolution Weighted Naïve Bayes* (DE-WNB) menggunakan *dataset* jantung menghasilkan keakuratan sebesar 83.44±5.51%. Algoritma *Naïve Bayes* (NB) menggunakan *dataset* jantung menghasilkan keakuratan sebesar 83.78±5.41%. Algoritma *Gain Ratio-Weighted Naïve Bayes* (GR-WNB) menggunakan *dataset* jantung menghasilkan keakuratan sebesar 81.63±6.23%. Algoritma *Correlation-based Feature Selection-Weighted Naïve Bayes* (CFS-WNB) menggunakan *dataset* jantung menghasilkan keakuratan sebesar 84.22±5.99%. Algoritma *Mutual Information-Weighted Naïve Bayes* (MI-WNB) menggunakan *dataset* jantung menghasilkan keakuratan sebesar 82.93±6.14%. Algoritma *Tree-Weighted Naïve Bayes* (*Tree*-WNB) menggunakan *dataset* jantung menghasilkan keakuratan sebesar 84.04±5.90%.

Dari penelitian-penelitian yang pernah dilakukan tentang diagnosis penyakit jantung terutama yang menggunakan algoritma *K-Nearest Neighbor*, akurasinya belum tinggi. Kelebihan-kelebihan spesifik model penggabungan algoritma *K-Nearest Neighbor* dan *Backward Elimination* pada penyakit jantung 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 jantung sehingga dapat meningkatkan akurasi dibandingkan dengan penelitian-penelitian sebelumnya.

## 2. 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).

### 2.1 Tahap Pemahaman Bisnis

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

### 2.2 Tahap Pemahaman Data

Penelitian ini menggunakan *dataset* jantung yang diperoleh dari *UCI Machine Learning* (Frank dan Asuncion, 2010). Data set jantung terdiri dari 123 record. dengan parameter-parameter sebagai berikut: *age, sex, chest pain type (cp), resting blood pressure (tresbpd), serum cholestoral in mg/dl (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved (thalach), exercise induced angina (exang), oldpeak, slope, number of major vessels colored by flourosopy (ca), thal, diagnosis of heart disease (num).*

### 2.3 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.

ISBN 978-602-99334-3-7

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*.

### 2.4 Tahap Pemodelan

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

#### 2.4.1 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.

#### 2.4.2 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. *Contohnya F(Tresbps|Num, Fbs, Exang, Oldpeak, Age, Ca), F(Ca| Num, Fbs, Exang, Oldpeak, Age, Tresbps), F(Slope| Num, Fbs, Exang, Oldpeak, Age, Restecg), F(Restecg| Num, Fbs, Exang, Oldpeak, Age, Slope), F(Cp| Num, Fbs, Exang, Oldpeak, Age, Thal),* dan *F(Thal| Num, Fbs, Exang, Oldpeak, Age, Cp)*
2. Menentukan variabel dengan F-statistik parsial terkecil dan menguji Fmin, dalam kasus ini *Ca*.
3. Jika Fmin tidak signifikan, dalam kasus ini, variabel dihilangkan dari model.
4. Variabel dengan F-statistik parsial adalah *cups*, karena *cups* tidak signifikan. Kemudian *Tresbps, Slope, Restecg, Cp,* dan *Thal* juga dihilangkan dari model.
5. Pada sisi lain, variabel dengan F-statistik terkecil adalah variabel indikator *Num.* 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 :
   $y = \beta 0 + \beta 1(Age) + \beta 2(Oldpeak) + \beta 3(Exang) + \beta 4(Fbs) + \beta 5(Num)\ \varepsilon$
6. Menghitung F-test parsial.

#### 2.4.3 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. *Contohnya F(Tresbps|Num, Fbs, Exang, Oldpeak, Age, Ca), F(Ca| Num, Fbs, Exang, Oldpeak, Age, Tresbps), F(Slope| Num, Fbs, Exang, Oldpeak, Age, Restecg), F(Restecg| Num, Fbs, Exang, Oldpeak, Age, Slope), F(Cp| Num, Fbs, Exang, Oldpeak, Age, Thal),* dan *F(Thal| Num, Fbs, Exang, Oldpeak, Age, Cp)*

2. Menentukan variabel dengan F-statistik parsial terkecil dan menguji Fmin, dalam kasus ini *Ca*.
3. Jika Fmin tidak signifikan, dalam kasus ini, variabel dihilangkan dari model.
4. Variabel dengan F-statistik parsial adalah *cups*, karena *cups* tidak signifikan. Kemudian *Tresbps, Slope, Restecg, Cp,* dan *Thal* juga dihilangkan dari model.
5. Pada sisi lain, variabel dengan F-statistik terkecil adalah variabel indikator *Num.* 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:

   $y = \beta 0 + \beta 1(Age) + \beta 2(Oldpeak) + \beta 3(Exang) + \beta 4(Fbs) + \beta 5(Num)\ \varepsilon$

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.

### 2.5 Tahap Evaluasi

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

## 3. HASIL DAN PEMBAHASAN

Akurasi *dataset* jantung dapat dilihat pada Tabel 1. *Area Under Curve* (AUC) *dataset* jantung dapat dilihat pada Tabel 2. Pada tabel 1, algoritma *K-Nearest Neighbor* menggunakan *dataset* jantung menghasilkan akurasi sebesar 88.62% +/- 0.09%, sedangkan algoritma *Backward Elimination-K-Nearest Neighbor* menghasilkan akurasi sebesar 89.55% +/- 6.01% 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 Jantung**

| Algoritma | Akurasi (%) |
|---|---|
| *K-Nearest Neighbor* | 88.62% +/- 0.09% |
| *Backward Elimination – K-Nearest Neighbor* | 89.55% +/- 6.01% |

**Tabel 2. *Area Under Curve* (AUC) *Dataset* Jantung**

| Algoritma | *Area Under Curve* (AUC) |
|---|---|
| *K-Nearest Neighbor* | 0.958 +/- 0.000 |
| *Backward Elimination – K-Nearest Neighbor* | 0.966 +/- 0.056 |

*Area Under Curve* (AUC) *dataset* jantung dapat dilihat pada Tabel 2. Pada tabel 2, algoritma *K-Nearest Neighbor* menggunakan *dataset* jantung menghasilkan *Area Under Curve* (AUC) sebesar 0.958 +/- 0.000 yang termasuk dalam kategori klasifikasi sangat baik *(excellent classification)*. Algoritma *Backward Elimination-K-Nearest Neighbor* menghasilkan AUC sebesar 0.966 +/- 0.056 yang termasuk dalam kategori klasifikasi sangat baik *(excellent classification).* Hasil menunjukkan metode *Backward Elimination-K-Nearest Neighbor* dapat mencapai akurasi yang tinggi dalam

ISBN 978-602-99334-3-7

mendiagnosis penyakit jantung. Percobaan ini dilakukan untuk menunjukkan peningkatan akurasi dari algoritma *K-Nearest Neighbor* menjadi *Backward Elimination-K-Nearest Neighbor*.

## 4. KESIMPULAN

Dari penelitian-penelitian yang pernah dilakukan tentang diagnosis penyakit jantung terutama yang menggunakan algoritma *K-Nearest Neighbor*, akurasinya belum tinggi. Hasil penelitian ini, pada *dataset* jantung, algoritma *K-Nearest Neighbor* memiliki nilai akurasi sebesar 88.62% +/- 0.09% dan nilai *Area Under Curve* (AUC) sebesar 0.958 +/- 0.000 yang termasuk dalam kategori klasifikasi sangat baik (*excellent classification*). Pada *dataset* jantung, penggabungan algoritma *Backward Elimination* dan *K-Nearest Neighbor* memiliki nilai akurasi sebesar 89.55% +/- 6.01% dan nilai *Area Under Curve* (AUC) sebesar 0.966 +/- 0.056 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 jantung.

.

## DAFTAR PUSTAKA


N. C. for C. D. P. and H. Promotion, (2009), "Heart Disease," *National Center for Chronic Disease Prevention and Health Promotion*, pp.

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

Moradian, M. and Baraani, A. (2009), "KNNBA: K-Nearest-Neighbor Based Association Algorithm," *Journal of Theoretical and Applied Information Technology*.

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.