Faktor Exacta 7(4): 366-371, 2014
ISSN: 1979-276X

PENERAPAN ALGORITMA KLASIFIKASI NEAREST NEIGHBOR (K-NN) UNTUK MENDETEKSI PENYAKIT JANTUNG

MEI LESTARI
mei_6s@yahoo.co.id
Program Studi Teknik Informatika, Fakultas Teknik, Matematika dan IPA
Universitas Indraprasta
Jln. Jl. Nangka No. 58 Tanjung Barat Jagakarsa, Jakarta Selatan

Abstrak. Data WHO menyatakan bahwa sebanyak 7,3 juta penduduk dunia meninggal dikarenakan penyakit jantung. Meskipun penyakit jantung merupakan penyakit yang tidak menular, penyakit ini merupakan jenis penyakit yang mematikan nomor satu di dunia. Penyakit jantung disebut juga dengan penyakit jantung koroner, penyakit ini terjadi bila darah ke otot jantung terhenti/tersumbat, sehingga mengakibatkan kerusakan berat pada jantung (Rajkumar & Reena, September 2010). Penerapan Algoritma Klasifikasi Nearest Neighbor (K-NN) diharapkan dapat mengatasi masalah efektifitas dan akurasi dalam mendeteksi penyakit jantung. Pada penelitian ini digunakan algoritma K-NN dengan k = 9 pada 100 data pasien penyakit jantung. Hasil penelitian diperoleh nilai akurasi sebesar 70% serta nilai AUC sebesar 0.875 yang masuk kedalam klasifikasi baik, sehingga algoritma K-NN dapat digunakan dan diterapkan untuk mendeteksi penyakit jantung.

Abstract. Data from WHO states that 7,3 million people die because of Heart Disease. Although Heart Disease is not contagious, this kind of disease is the number 1 disease causing death. Heart disease is also called as coroner disease. It happens when the blood drifting to the heart muscles stops so that it causes heart disorder (Rajkumar & Reena, September 2010). The application of Nearest Neighbor classification (K-NN) algorithm is expected to overcome the problems on the affectivity and accuracy in detecting heart disease. In this research, the K-NN with K = 9 is used on 100 patients of heart disease. The result revealed is that the accuracy is 70% and the AUC is 0.875 which belong to the good classification. Hence, K-NN algorithm can be used and applied in detecting heart disease.

Keywords: Nearest Neighbor, Jantung, Algoritma, Akurasi.

PENDAHULUAN

Jantung merupakan organ manusia yang berperan dalam sistem peredaran darah. Penyakit jantung adalah sebuah kondisi dimana jantung tidak dapat melaksanakan tugasnya dengan baik. Data WHO menyatakan bahwa sebanyak 7,3 juta penduduk dunia meninggal dikarenakan penyakit jantung. Meskipun penyakit jantung merupakan penyakit yang tidak menular, penyakit ini merupakan jenis penyakit yang mematikan nomor satu di dunia.

Penyakit jantung disebut juga dengan penyakit jantung koroner, penyakit ini terjadi bila darah ke otot jantung terhenti/tersumbat, sehingga mengakibatkan kerusakan berat pada jantung (Rajkumar & Reena, September 2010). Penyebab utama penyakit jantung adalah penggunaan tembakau, fisik tidak aktif, diet yang tidak sehat dan penggunaan alkohol, resiko penyakit jantung bertambah dengan meningkatnya usia, tekanan darah tinggi, mempunyai kolesterol tinggi, dan kelebihan berat badan.

Algoritma *Nearest Neighbor* (K-NN) merupakan algoritma klasifikasi berdasarkan kedekatan jarak suatu data dengan data yang lain. Pada algoritma K-NN, data berdimensi

q,jarak dari data tersebut ke data yang lain dapat dihitung. Nilai jarak inilah yang digunakan sebagai nilai kedekatan/kemiripan antara data uji dengan data latih. Nilai K pada K-NN berarti K-data terdekat dari data uji.

Untuk menangani masalah efektifitas dan akurasi dalam mendeteksi penyakit jantung maka dibuatlah sistem pendeteksi penyakit jantung menggunakan algoritma klasifikasi *nearest neighbor* (K-NN).

Dalam penelitian ini, pertanyaan yang diajukan adalah "Apakah algoritma *nearest neighbor* dapat diterapkan pada otomatisasi sistem pendeteksi penyakit jantung?", "Bagaimana tingkat akurasi pendeteksian penyakit jantung menggunakan algoritma *nearest neighbor*?".

TINJAUAN PUSTAKA

Algoritma *Nearest Neighbor*

External variable secara langsung akan mempengaruhi persepsi manfaat dan prsepsi kemudahan dari pengguna. Persepsi kemudahan penggunaan dipengaruhi oleh variabel eksternal yang berkenaan dengan karakteristik sistem yang meningkatkan penggunaan dari teknologi, seperti *mouse, touch screen, menu* dan selain itu, pelatihan individu juga akan mempengaruhi kemudahan penggunaan. Semakin banyak pelatihan yang diterima individu, semakin besar tingkat kemudahan dalam penggunaan.

Tujuan algoritma KNN adalah mengklasifikasikan obyek baru berdasarkan atribut dan *training sample. Clasifier* tidak menggunakan model apapun untuk dicocokkan dan hanya berdasarkan pada memori. Diberikan titik *query*, akan ditemukan sejumlah *k* obyek atau (titik training) yang paling dekat dengan titik query. Klasifikasi menggunakan voting terbanyak diantara klasifikasi dari *k* obyek. Algoritma KNN menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari *query instance* yang baru. Algoritma metode KNN sangatlah sederhana, bekerja berdasarkan jarak terpendek dari *query instance* ke *training sample* untuk menentukan KNN-nya.

Nilai *k* yang terbaik untuk algoritma ini tergantung pada data. Secara umum, nilai *k* yang tinggi akan mengurangi efek noise pada klasifikasi, tetapi membuat batasan antara setiap klasifikasi menjadi semakin kabur. Nilai *k* yang bagus dapat dipilih dengan optimasi parameter, misalnya dengan menggunakan *cross-validation*. Kasus khusus dimana klasifikasi diprekdisikan berdasarkan training data yang paling dekat (dengan kata lain, k=1) disebut algoritma Nearest Neighbor.

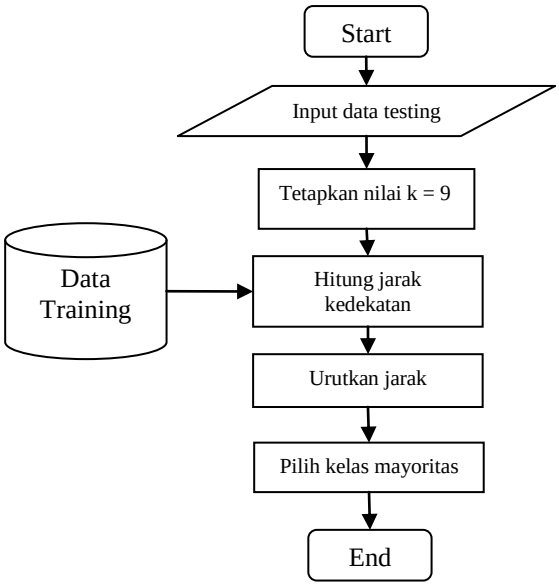
Kelebihan KNN *(K-Nearest Neighbor):*

1. Tangguh terhadap *training* data yang memiliki banyak *noise*.
2. Efektif apabila *training* datanya besar.

Kelemahan KNN *(K-Nearest Neighbor):*

1. KNN perlu menentukan nilai dari parameter *k* (jumlah dari tetangga terdekat).
2. *Training* berdasarkan jarak tidak jelas mengenai jenis jarak apa yang harus digunakan.
3. Atribut mana yang harus digunakan untuk mendapatkan hasil terbaik.
4. Biaya komputasi cukup tinggi karena diperlukan perhitungan jarak dari tiap *query instance* pada keseluruhan *training sample.*

Untuk lebih jelasnya algoritma K-NN dapat dilihat pada gambar 1.



Gambar 1. algoritma K-NN

1. Tentukan parameter *K*
2. Hitung jarak antara data yang akan dievaluasi dengan semua pelatihan
3. Urutkan jarak yang terbentuk (urut naik)
4. Tentukan jarak terdekat sampai urutan *K*
5. Pasangkan kelas yang bersesuaian
6. Cari jumlah kelas dari tetangga yang terdekat dan tetapkan kelas tersebut sebagai kelas data yang akan dievaluasi

Rumus KNN:

$$d_i = \sqrt{\sum_{i=1}^{p} (x_{2i} - x_{1i})^2} \quad \ldots(1)$$

Keterangan:

x_1 = Sampel Data

x_2 = Data Uji / Testing

i = Variabel Data

d = Jarak

p = Dimensi Data

METODE

Data utama diperoleh dari *University of California Irvine Machine learning data repository*. Sedangkan data pendukung diambil dari buku, jurnal dan publikasi lainnya. Sampel dari penelitian ini adalah data profile penderita penyakit jantung, data tersebut bersifat publik yang didapatkan dari *University of California Irvine machine learning data repository*.

Pengolahan data dalam penelitian ini adalah proses pengelompokan data-data yang telah dikumpulkan sebelumnya dengan tujuan untuk menentukan variabel-variabel yang akan digunakan beserta himpunan-himpunan yang termasuk kedalam variabel-variabel yang digunakan. Terdapat 13 atribut yang digunakan dalam mendiagnosa penyakit

jantung yaitu *Age, Sex, Chest Pain Type, Tresting Blood Pressure, Serum cholestoral dalam mg/dl, Fasting blood sugar>120 mg/dl, Resting electrocardiographic result, The Slope of the peak exercise ST segment, Excercise Induced Angina, Old Peak, CA (Number of Major Vessels), Maximum Heart Rate, Achieved (Thalac), Thal*

Terdapat 100 data pasien penyakit jantung yang diolah menggunakan algoritma Klasifikasi Nearest Neighbor (KNN) dengan k = 9.

HASIL DAN PEMBAHASAN

Pada penelitian ini digunakan 110 *records* pasien. 100 *records* digunakan sebagai data latih (*training data*) dan 10 *records* digunakan sebagai data uji (*testing data*). Untuk menentukan apakah seorang pasien terkena penyakit jantung digunakan 9 data terdekat atau K= 9. Klasifikasi dilakukan dengan menggunakan mayoritas suara diantara klasifikasi dari K objek. Algoritma KNN menggunakan klasifikasi ketetanggaan sebagai prediksi terhadap data baru. Dari 9 data tersebut diperoleh kelas mayoritas, maka data uji tersebut masuk kedalam kelas mayoritas. Contoh perhitungan kedekatan atau jarak antar atribut adalah sebagai berikut:

Tabel 1. Contoh Data Latih

AGE	SEX	CHEST PAIN TYPE	RESTING BLOOD PREASURE	SERUM CHOLESTORAL	FASTING BLOOD SUGAR	RESTING ELECTROCARDIOGRAPHIC	MAXIMUM HEART RATE	EXERCISE	OLDPEAK	SLOPE	CA	THAL	CLASS
70	1	4	130	322	0	2	109	0	2,4	2	3	3	2
67	0	3	115	564	0	2	160	0	1,6	2	0	7	1
57	1	2	124	261	0	0	141	0	0,3	1	0	7	2

Tabel 2. Contoh Data Uji

AGE	SEX	CHEST PAIN TYPE	RESTING BLOOD PREASURE	SERUM CHOLESTORAL	FASTING BLOOD SUGAR	RESTING ELECTROCARDIOGRAPHIC	MAXIMUM HEART RATE	EXERCISE	OLDPEAK	SLOPE	CA	THAL	CLASS
44	0	3	108	141	0	0	175	0	0,6	2	0	3	1

Perhitungan kedekatan kasus antara data training dengan data uji adalah sebagai berikut:
Kuadrat jarak data training baris ke-1 dengan data uji =
$(70-44)^2+(1-0)^2+(4-3)^2+(130-108)^2+(322-141)^2+(0-0)^2+(2-0)^2+(109-175)^2+(0-0)^2+(2,4-0,6)^2+(2-2)^2+(3-0)^2+(3-3)^2 = 38295,24$

Evaluasi dan Validasi

Tabel 3. Hasil Evaluasi dan Validasi

NO	AGE	SEX	CHEST PAIN TYPE	RESTING BLOOD PREASURE	SERUM CHOLESTORAL	FASTING BLOOD SUGAR	RESTING ELECTROCARDIOGRAPHIC	MAXIMUM HEART RATE	EXERCISE	OLDPEAK	SLOPE	CA	THAL	KLASIFIKASI	DIKLASIFIKASIKAN SEBAGAI
1	44	0	3	108	141	0	0	175	0	0,6	2	0	3	1	1
2	67	1	4	120	237	0	0	71	0	1	2	0	3	2	2
3	49	0	4	130	269	0	0	163	0	0	1	0	3	1	1
4	57	1	4	165	289	1	2	124	0	1	2	3	7	2	2
5	63	1	4	130	254	0	2	147	0	1,4	2	1	7	2	2
6	48	1	4	124	274	0	2	166	0	0,5	2	0	7	2	1
7	51	1	3	100	222	0	0	143	1	1,2	2	0	3	1	1
8	60	0	4	150	258	0	2	157	0	2,6	2	2	7	2	1
9	59	1	4	140	177	0	0	162	1	0	1	1	7	2	1
10	45	0	2	112	160	0	0	138	0	0	2	0	3	1	1

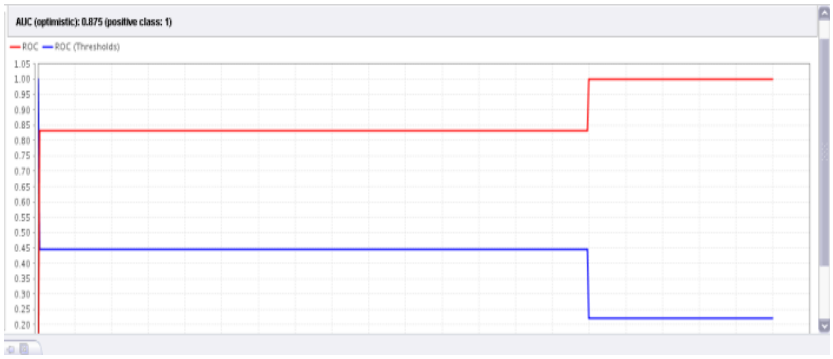
Untuk mengukur tingkat akurasi klasifikasi metode yang digunakan antara lain *confusion matrix*. Perhitungan kedekatan kasus pada data training dengan kasus pada data testing, diketahui dari 10 data 4 data termasuk kedalah kelas "1" atau terkena penyakit jantung dan 6 data termasuk kedalam kelas "2" yaitu normal atau tidak terkena penyakit jantung.

Perhitungan kedekatan data training dengan kasus pada data testing 3 data diprediksikan masuk kedalam kelas "1" tetapi ternyata termasuk kedalam kelas "2". Dari 100 data training dan 10 data testing dan menggunakan metode K-NN dengan nilai K = 9 diperoleh tingkat akurasi sebesar 70%. Tabel 4 berikut merupakan tabel coffusion matrix untuk metode K-NN.

Tabel 4. Cofussion Matrix

Kelas Prediksi	Kelas sebenarnya	
	1	2
1	4	3
2	0	3

Kurva berikut merupakan kurva ROC untuk perhitungan cofussion matrix. Nilai AUC pada data mining dibagi menjadi beberapa kelompok yaitu sangat baik untuk nilai 0.90-1.00, klasifikasi baik untuk nilai 0.80-0.90, klasifikasi cukup untuk nilai 0.70-0.80, klasifikasi buruk untuk nilai 0.60-0.70 dan klasifikasi salah untuk nilai 0.50-0.60 (Gorunescu, 2011). Pada kurva terlihat nilai AUC adalah 0.875 sehingga dapat disimpulkan metode KNN termasuk kedalam klasifikasi baik.



Gambar 2. Kurva ROC dengan Metode KNN

PENUTUP

Dalam penelitian ini dilakukan penerapan algoritma K-NN dengan k = 9 pada data pasien untuk mendeteksi penyakit jantung. Kedekatan antara kasus pada data training dan data uji dilakukan untuk menentukan kelas data uji tersebut. Untuk mengukur kinerja algoritma tersebut dilakukan dengan menggunakan *cofusion matrix* dan kurva ROC, diperoleh nilai akurasi 70% dan termasuk klasifikasi baik karena memiliki nilai AUC 0.875.

Terdapat saran yang dapat diterapkan guna penelitian selanjutnya, yaitu, dilakukan komparasi terhadap algoritma atau metode data mining lainnya dalam mendeteksi penyakit jantung, untuk mengetahui algoritma mana yang lebih akurat dan efisien, sehingga dapat ditentukan algoritma yang tepat untuk mendeteksi penyakit jantung.

DAFTAR PUSTAKA

Bramer, Max. 2007. **Principles of Data Mining**. London: Springer

Gorunescu, Florin. 2011. **Data Mining: Concepts, Models, and Techniques**. Verlag Berlin Heidelberg: Springer

Han, J.,&Kamber, M. 2006. **Data Mining Concept and Tehniques**. San Fransisco: Morgan Kauffman.

Kusrini & Luthfi, E. T. 2009. **Algoritma Data Mining**. Yogyakarta: Andi Publishing.

Sumathi, & S., Sivanandam, S.N. 2006. **Introduction to Data Mining and its Applications**. Berlin Heidelberg New York: Springer.