

Performa Algoritma K-Nearest Neighbour dalam Memprediksi Penyakit Jantung

Hanna Willa Dhany

Universitas Pembangunan Pancabudi Medan

Email : hdhany@dosen.pancabudi.ac.id

Abstrak

Penyakit jantung telah terbukti sebagai salah satu penyebab utama kematian di seluruh dunia, inilah alasan utama mengapa prediksi Penyakit jantung akurat yang memiliki risiko sangat penting untuk mencegah dan mengobatinya. Diagnosis CHF (Gagal Jantung Kongesif) yang tepat waktu sangat penting untuk menghindari bahaya yang membahayakan nyawa acara. Terlepas dari ketepatan waktu, akurasi memainkan peran yang luar biasa peran penting dalam domain medis karena terkait dengan kehidupan seseorang.. Penyakit Jantung telah terbukti sebagai salah satu penyebab utama kematian, itulah sebabnya prediksi risiko penyakit jantung yang akurat dan tepat waktu sangat penting. Metode klinis, misalnya angiografi cara terbaik dan paling efektif untuk mendiagnosis penyakit jantung, penelitian menunjukkan bahwa itu tidak hanya mahal tetapi juga memiliki efek samping. Maka penulis melakukan penelitian terhadap pengujian data dari dataset Heart Disease dengan melakukan pengujian terhadap akurasi data confusion matrix dengan menggunakan metode K-nearest Neighbour yang menghasilkan akurasi sebesar 81.31% dengan Classification Error sebesar 18.68%,

Kata Kunci: Penyakit Jantung, Akurasi, Prediksi, Klasifikasi

Abstract

Heart disease has been proven to be one of the leading causes of death worldwide, this is the main reason why accurate heart disease prediction at risk is so important to prevent and treat it. A timely diagnosis of CHF (Congesive Heart Failure) is essential to avoid the life-threatening hazards of the event. Apart from punctuality, accuracy plays an extraordinary role in the medical domain as it is linked to a person's life. Heart disease has been shown to be one of the leading causes of death, which is why accurate and timely prediction of heart disease risk is so important. Clinical methods, for example angiography are the best and most effective way to diagnose heart disease, research shows that it is not only expensive but also has side effects. So the authors conducted research on testing data from the Heart Disease dataset by testing the accuracy of the confusion matrix data using the K-nearest neighbor method which resulted in an accuracy of 81.31% with a Classification Error of 18.68%,

Keywords: Heart Disease, Accuracy, Prediction, Classification

1. Pendahuluan

Klasifikasi berfungsi untuk melakukan prediksi kelas dari suatu objek yang kelaskelasnya belum diketahui. Klasifikasi adalah pola yang dilakukan secara terarah [1]. Metode klasifikasi yang umum digunakan yaitu *Decision Tree*, *KNearest Neighbor*, *Naïve Bayes*, *Neural Network* dan *Support Vector Machine* [9].

Salah satu kelompok yang terdapat pada pembelajaran berbasis *instance* adalah metode *K-Nearest Neighbor* (KNN). Pencarian kelompok yang dilakukan KNN kepada k objek dalam pengujian yang paling dekat dengan objek data yang baru atau data uji. KNN berguna dalam klasifikasi yang dilakukan kepada objek yang didasari terhadap data pembelajaran yang memiliki jarak yang dekat dengan objek yang telah diuji. Pendekatan yang dilakukan pada pencarian kasus perhitungan pendekatan antara masalah baru dengan masalah sebelumnya dengan melakukan penyetaraan bobot dari penjumlahan fitur yang ada.

Gagal jantung telah terbukti sebagai salah satu penyebab utama kematian di seluruh dunia, inilah alasan utama mengapa prediksi Gagal jantung akurat yang memiliki risiko sangat penting untuk mencegah dan mengobatinya. Diagnosis CHF (Gagal Jantung Kongesif) yang tepat waktu sangat penting untuk menghindari bahaya yang membahayakan nyawa acara. Terlepas dari ketepatan waktu, akurasi memainkan peran yang luar biasa peran penting dalam domain medis karena terkait dengan kehidupan seseorang. Sejumlah besar penelitian telah dilakukan telah dilakukan klasifikasi dan prediksi penyakit menggunakan ML teknik. Sebaliknya, kesepakatan belum dibuat teknik atau pengklasifikasi mana yang paling cocok untuk kumpulan data tertentu. Namun, hal itu terbukti dari pemilihan dan pengurangan fitur teknik meningkatkan akurasi dan keandalan pengklasifikasi. Selain itu, juga jelas bahwa ansambel pengklasifikasi telah ada terbukti meningkatkan akurasi klasifikasi.

Survei ini bertujuan untuk menggali, meringkas, dan menganalisis secara kritis makalah penelitian terbaru dan mutakhir untuk menemukan kesenjangan penelitian untuk studi masa depan. Penelitian ini telah dilakukan sistematis, untuk membantu pembaca mendapatkan pengetahuan sebelumnya penelitian yang dilakukan dalam penfeteksian gagal jantung dan risikonya.

Penulis bermaksud untuk menerapkan berbagai teknik klasifikasi pada dataset jantung dan membandingkan serta menilai dengan kinerja dalam hal akurasi, spesifisitas, dan sensitivitas. Disini penulis menerapkan algoritma K-Nearest Neighbour dalam melakukan analisa terhadap dataset jantung, yang nantinya akan dapat dilihat dimana data yang mengalami gagal jantung dan resiko.

2. Metode Penelitian

Pada penelitian ini, untuk mengetahui kinerja dari metode *K-Nearest Neighbour* maka digunakan data observasi berupa dataset *Kaggle Heart Disease* yang berasal dari (<https://www.kaggle.com/>). Data kanker serviks ini melibatkan 303 sampel dengan 13 atribut serta dua label kelas meliputi: 0 = lebih sedikit kemungkinan serangan jantung 1 = lebih banyak kemungkinan serangan jantung. Adapun deskripsi dari dataset *Heart Disease* dapat dilihat pada tabel 3.1 sebagai berikut:

Tabel 1. *Dataset UCI Heart Disease*

No.	Atribut		Detail
1	age	age	usia
2	sex	sex	seks
3	cp	chest pain type	jenis nyeri dada
4	trestbps	resting blood pressure	tekanan darah istirahat
5	chol	serum cholestoral	serum kolestoral
6	fbs	fasting blood sugar	gula darah puasa
7	restecg	resting electrocardiographic result	hasil elektrokardiografi istirahat
8	thalach	maximum heart rate achieved	detak jantung maksimum tercapai
9	exang	exercise induced angina	latihan angina yang diinduksi
10	oldpeak	old peak	puncak tua
11	slope	the slope of the peak exercise	kemiringan latihan puncak
12	ca	number of major vessels	jumlah kapal utama
13	thal	thal	thal

Sumber : (<https://www.kaggle.com/>).

Berdasarkan tabel diatas, Masing-masing atribut memiliki sejumlah *tuple* yang proporsional yang terdiri dari 13 indikator sebagai atribut dan empat label kelas yang meliputi: lebih sedikit kemungkinan serangan jantung dan lebih banyak kemungkinan serangan jantung. Dataset *Heart Disease* tersebut juga mempunyai karakteristik *missing value* dan data yang duplikat sehingga terlebih dahulu dilakukan data *preprocessing* guna meningkatkan tingkat akurasi pengklasifikasiannya melalui partisi data yang akan dibagi menjadi 90% sebagai data *training* dan 10% sebagai data *testing*.

Untuk melihat apakah model penelitian yang diusulkan mampu meningkatkan performa pengklasifikasian K-Nearest Neighbour (K-NN), maka akan dilakukan berdasarkan hasil tabulasi *Confusion Matrix (two-class prediction)* yang kemudian hasil tersebut akan mendapatkan hasil yang diinginkan.

Berikut adalah rancangan dan prosedur penelitian dari pengklasifikasian *K-Nearest Neighbour*,.



3. Hasil dan Pembahasan

Berikut adalah hasil akurasi model klasifikasi *K-Nearest Neighbors* menggunakan bahasa pemrograman *Python* versi 3.7.3, Saat nilai $K=2$, sebanyak 303 *instances*. Agar hasil akurasi yang diharapkan lebih akurat maka dilakukan partisi data menggunakan metode *K-Fold Cross Validation*.*K-fold* adalah salah satu metode *Cross Validation* yang populer dengan melipat data sebanyak K dan mengulangi (men-iterasi) experimennya sebanyak K juga. Kemudian untuk melihat rasio error dari masing – masing nilai K , maka dilakukan iterasi untuk menghitung nilai error (*error-rate*) sehingga menghasilkan nilai K yang optimal saat dilakukan uji coba terhadap model klasifikasi *K-Nearest Neighbors* Konvensional. Berikut adalah hasil output dari pengujian:

Tabel 2 . *Confusion Matrix* Kinerja Klasifikasi

Kinerja Klasifikasi	Predicted Class	
Actual Class	Predicted. Class 1	Predicted. Class 0
Actual. Class 1	107 (True Positive)	33 (False Negative)
Actual. Class 0	4 (False Positive)	54 (True Negative)

Berdasarkan tabel diatas, maka dilanjutkan dengan menghitung nilai *Accuracy* dari model klasifikasi *KNN* dari dataset *Heart Disease Dataset*. Berikut hasil perhitungannya:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{107+54}{107+54+4+33} = \frac{161}{198} = 0,8131*100\% = 81.31\%$$

$$Classification\ Error = \frac{FP+FN}{TP+TN+FP+FN} = \frac{4+33}{107+54+4+33} = \frac{37}{198} = 0.1868 * 100\% = 18.68\%$$

Maka penelitian ini menghasilkan nilai akurasi pada algoritma *K-NN* pada dataset *Heart Disease* sebesar 81.31% dengan *Classification Error* sebesar 18.68%.

4. Kesimpulan

Penyakit Jantung telah terbukti sebagai salah satu penyebab utama kematian, itulah sebabnya prediksi risiko penyakit jantung yang akurat dan tepat waktu sangat penting. Metode klinis, misalnya angiografi cara terbaik dan paling efektif untuk mendiagnosis penyakit jantung, penelitian menunjukkan bahwa itu tidak hanya mahal tetapi juga memiliki efek samping. Maka penulis melakukan penelitian terhadap pengujian data dari dataset *Heart Disease* dengan melakukan pengujian terhadap akurasi data confusion matrix dengan menggunakan metode *K-nearest Neighbour* yang menghasilkan akurasi sebesar 81.31% dengan *Classification Error* sebesar 18.68%, yang berarti data yang telah tersedia dapat digunakan sebagai acuan terhadap penelitian selanjutnya untuk mencari nilai akurasi terhadap data penyakit jantung. Saya harap peneliti selanjutnya dapat menganalisa akurasi data dengan menggunakan metode yang lain yang performa akurasinya lebih tinggi dari yang saya teliti ini.

Daftar Pustaka

- [1] Boon, K. et al. 2018. Paroxysmal atrial fibrillation prediction based on HRV analysis and non-dominated sorting genetic algorithm III. *Computer Methods and Programs in Biomedicine*. 153, (2018), 171-184.
- [2] Gou, J., Yi, Z., Du, L. & Xiong, T. January 2012. A Local Mean-Based k-Nearest Centroid Neighbor Classifier. *The Computer Journal* 55(6) : pp. 1058-1071.
- [3] J. Wassan, H. Wang and H. Zheng. 2018. Machine Learning in Bioinformatics. *Encyclopedia of Bioinformatics and Computational Biology*, vol. 1, pp. 300-308, 2018.
- [4] Pan, Z., Wang, Y. & Ku, W. September 2016. A New K-Harmonic Nearest Neighbor Classifier Based On The Multi-Local Means. *Expert Systems With Applications* 67: 115-125.
- [5] Mahajan, R. et al. 2017. Improved detection of congestive heart failure via probabilistic symbolic pattern recognition and heart rate variability metrics. *International Journal of Medical Informatics*. 108, (2017), 55-63.
- [6] Puspito, Toto, Andri. 2017. *Menentukan Pilihan Sekolah didalam Penerimaan Peserta Didik Baru Dengan Menggunakan Metode Naive Bayes dan K-Nearest Neighbor* (Studi Kasus: PPDB Online Jenjang SMP Kota Metro). *Jurnal Informatika*, Volume. 17, No.1.pp.49-59.
- [7] Purushottam et al. 2016. Efficient Heart Disease Prediction System. *Procedia Computer Science*. 85, (2016), 962-969.
- [8] Raviya, K.H. & Gajjar, B., 2013. Performance Evaluation of Different Data Mining Classification Algorithm Using WEKA. *Indian Journal of Research*, II(1).
- [9] Sahu, H., Shrma, S. & Gondhalakar, S., 2011. A Brief Overview on Data Mining Survey. *International Journal of Computer Technology and Electronics Engineering*, 1(3), pp.114-21.
- [10] T. Setiyori, "Penerapan Information Gain pada K-Nearest Neighbor untuk Klasifikasi Tingkat Kognitif Soal pada Taksonomi Bloom," *J. Sist. Inf.*, vol. 6, Feb. 2017.