JNATIA Volume 1, Nomor 1, November 2022
Jurnal Nasional Teknologi Informasi dan Aplikasinya

Penerapan Algoritma *K-Nearest Neighbor* dalam Klasifikasi Penyakit Gagal Jantung

Ni Ketut Intan Setiawati[a1], I Gede Arta Wibawa[a2]

[a]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Udayana
Bali, Indonesia
[1]intansetiawati154@gmail.com
[2]gede.arta@unud.ac.id

Abstract

Heart failure is a clinical syndrome caused by an abnormality in the structure of the heart's function that causes the heart to be unable to pump adequate amounts of blood so that the body does not receive enough nutrients for metabolic needs. The high prevalence of heart failure needs serious attention so that it can be treated early and reduce the risk of complications. Along with the development of technology, various classification methods have been used by other researchers to determine whether a person has heart failure or not. In this study, the classification of heart failure will use the K-Nearest Neighbor algorithm. K-Nearest Neighbor (KNN) is a classification algorithm based on the proximity of data to other data. The results of this study are the best accuracy values obtained with k = 7, where after going through an evaluation with the confusion matrix, the accuracy of the classification of heart failure with KNN is 91%.

Keywords: *K-Nearest Neighbor, Classification, Heart Failure, Accuracy, Confusion Matrix*

1. Introduction

Dalam bidang kesehatan, berbagai sindroma klinik masih membutuhkan perhatian serius salah satunya adalah gagal jantung. Gagal jantung adalah kondisi dimana fungsi jantung tidak bekerja semestinya. Jantung yang seharusnya berperan dalam pemenuhan kebutuhan metabolisme tubuh, seperti oksigen dan nutrisi menjadi tidak maksimal. Hal ini dikarenakan jantung tidak dapat memompa darah dalam jumlah yang cukup karena terjadi abnormalitas pada bagian struktur atau fungsi jantung[1]. Tingginya angka prevalensi gagal jantung mendorong beberapa peneliti untuk melakukan penelitian mengenai gagal jantung yang bertujuan untuk melakukan penanganan lebih dini dan mengurangi risiko komplikasi.

Pada era digital ini, mulai berkembang teknologi yang dapat memudahkan tenaga medis untuk mengetahui apakah seseorang mengidap gagal jantung atau tidak. Salah satu bentuk dari metode yang dapat dilakukan adalah dengan melakukan klasifikasi. Klasifikasi merupakan teknik yang dapat digunakan untuk membantu menyelesaikan masalah yang berkaitan dengan kumpulan objek yang akan dikelompokkan. Terdapat banyak metode yang bisa digunakan sebagai solusi permasalahan klasifikasi, salah satunya adalah *K-Nearest Neighbor* (KNN). *K-Nearest Neighbor* adalah algoritma yang digunakan dalam klasifikasi dengan konsep dasar menggunakan jarak antara data yang digunakan dengan data lain yang berdekatan[2]. KNN umumnya digunakan karena tergolong sederhana sehingga proses implementasi menjadi lebih mudah dibandingkan algoritma klasifikasi lain dan tetap menghasilkan keluaran yang cukup akurat.

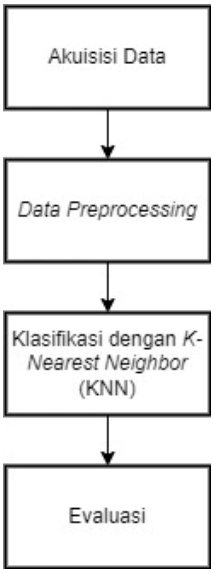
Penelitian menggunakan KNN sudah pernah dilakukan dalam beberapa tahun terakhir untuk menyelesaikan masalah klasifikasi. Seperti penelitian [3] *K-Nearest Neighbor* digunakan untuk melakukan klasifikasi penyakit kanker payudara, dengan hasil yang diperoleh adalah akurasi tertinggi sebesar 0,93, presisi bernilai 0,97, *recall* dengan nilai 0,98, dan nilai *F-measure* sebesar 0.94. Penelitian terdahulu menggunakan KNN juga dilakukan pada penelitian [4], dimana peneliti menggunakan metode *K-Nearest Neighbor* untuk mengklasifikasikan seseorang mengidap penyakit ginjal kronis atau tidak. Pada penelitian tersebut dihasilkan nilai akurasi sebesar 85,83%

yang dapat digolongkan cukup tinggi. Berdasarkan uraian di atas, penulis melakukan penelitian mengenai klasifikasi menggunakan *K-Nearest Neighbor* sebagai lanjutan dari penelitian sebelumnya. Penelitian ini akan dilakukan melalui beberapa tahapan mulai dari akuisisi data, *data preprocessing*, klasifikasi menggunakan KNN, dan melakukan evaluasi. Melalui penelitian ini diharapkan dapat memudahkan proses klasifikasi pada data gagal jantung dan mengetahui bagaimana tingkat akurasi yang akan dihasilkan sehingga dapat menjadi pembanding untuk penelitian lainnya.

2. Research Methods

2.1. Desain Penelitian

Berikut adalah tahapan – tahapan dari penelitian yang dilakukan:



Gambar 1. Desain Penelitian

2.2. Akuisisi Data

Pada tahap akuisisi, ditentukan data apa yang akan digunakan dalam penelitian. Dimana pada penelitian kali ini data yang digunakan diperoleh dari *website Kaggle* yaitu *Heart Failure Prediction Dataset*. *Dataset* ini berjumlah 918 data dengan 12 atribut. Berikut adalah keterangan dari atribut *dataset* yang digunakan:

Tabel 1. *Heart Failure Prediction Dataset*

No.	Atribut	Keterangan
1	Age	Umur pasien
2	Sex	Jenis Kelamin
3	ChestPainType	Jenis nyeri dada
4	RestingBP	Tekanan darah dalam kondisi istirahat
5	Cholesterol	Kadar kolesterol
6	FastingBS	Kadar gula darah
7	RestingECG	Kondisi ECG pasien dalam keadaan istirahat
8	MaxHR	Detak jantung maksimum
9	ExerciseAngina	Nyeri dada saat berolahraga
10	Oldpeak	Penurunan ST setelah olahraga
11	ST_Slope	Kemiringan segmen ST untuk latihan maksimum
12	HeartDisease	Kelas *output* (target)

2.3. Data Preprocessing

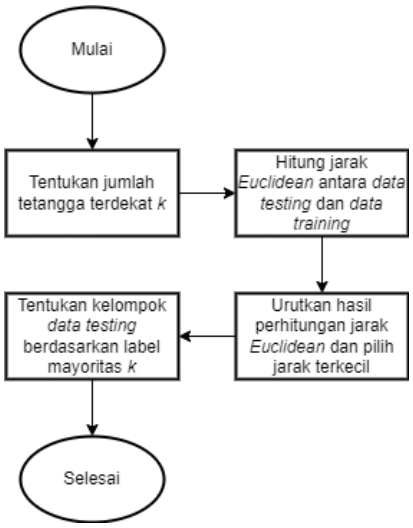
Pada tahap *data preprocessing* ini data akan diolah sehingga lebih terstruktur sebelum dilakukan pemodelan pada tahap klasifikasi. Pada tahap *preprocessing* ini, akan dilakukan beberapa proses pengolahan data mulai dari pengecekan apakah pada *dataset* terdapat *missing value* atau tidak, melakukan *feature scaling* sehingga kualitas data dapat ditingkatkan karena memiliki kesamaan pada skalanya, serta mengubah *categorical data* menjadi *numerical data*.

2.4. Klasifikasi dengan *K-Nearest Neighbor* (KNN)

K-Nearest Neighbor (KNN) adalah salah satu dari sekian banyak metode klasifikasi yang menggunakan pembelajaran data pada klasifikasi sebelumnya untuk mengklasifikasikan data lainnya. KKN yang tidak membangun sebuah model melainkan hanya mempelajari data yang sebelumnya telah diklasifikasikan ini membuat KNN kerap disebut sebagai teknik *lazy learning*. Algoritma *K-Nearest Neighbor* yang termasuk dalam *supervised learning* ini akan mengklasifikasikan hasil *query instance* yang baru, dimana mayoritas kedekatan jarak dari kategori yang ada dalam KNN akan dijadikan dasar dalam klasifikasi[3]. Pada KNN grup kelas yang jarak vektornya paling dekat akan menjadi dasar dalam memilih kelas baru bagi suatu data[5]. Pada KNN nilai *k* memiliki arti *k*-data terdekat dari data uji. Dengan kata lain *k* ini adalah jumlah data atau tetangga yang jaraknya paling dekat dengan suatu objek[6]. Terdapat beberapa cara untuk menentukan dekat atau jauhnya jarak antara *data training* dan data yang baru, salah satunya adalah dengan menggunakan *Euclidean Distance*. Jarak *Euclidean* dapat dihitung menggunakan persamaan (1)[2].

$$d_{Euc}(x,y) = \sqrt{\sum_i^p (x_1 - x_2)^2} \quad (1)$$

Keterangan:
d_{Euc} = jarak antara *data training* dan *data testing*
x_1 = *data training* atau sampel data
x_2 = *data testing* atau data uji
p = dimensi data
i = variabel data
Berikut adalah *flowchart* dari algoritma *K-Nearest Neighbor* yang digunakan:



Gambar 2. *Flowchart* algoritma KNN

2.5. Evaluasi

Evaluasi dilakukan untuk menilai performa dari model klasifikasi, dimana evaluasi ini dapat dilakukan dengan menggunakan *confusion matrix*. *Confusion matrix* adalah metode yang digunakan dalam *data mining* untuk menghitung nilai akurasi[7]. Berikut adalah tabel *confusion matrix* yang menyatakan data uji yang benar dan salah ketika diklasifikasikan.

Tabel 2. *Confusion Matrix*

Kinerja Klasifikasi	Nilai Kelas Prediksi	
Nilai Kelas Aktual	Positif (1)	Negatif (0)
Positif (1)	*True Positive* (TP)	*False Negative* (FN)
Negatif (0)	*False Positive* (FP)	*True Negative* (TN)

Melalui *confusion matrix*, nilai akurasi dapat ditentukan dengan pembagian jumlah data yang telah terklasifikasikan secara benar dengan total sampel data yang diuji. Berikut adalah persamaan untuk menghitung akurasi dengan *confusion matrix*:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (2)$$

Keterangan:
TP = total *record* data positif yang diklasifikasikan benar sebagai nilai positif
FP = total *record* data negatif namun hasil klasifikasinya adalah nilai positif
FN = total *record* data positif namun hasil klasifikasinya adalah nilai negatif
TN = total *record* data negatif yang diklasifikasikan benar sebagai nilai negatif

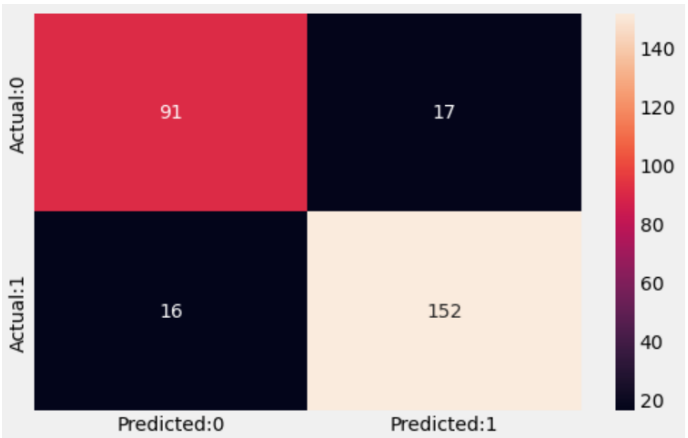
3. Result and Discussion

Pada penelitian ini *dataset* yang digunakan adalah *Heart Failure Prediction Dataset* yang berjumlah 918 data dengan 508 *record* diklasifikasikan menderita gagal jantung dan 410 *record* diklasifikasikan bukan penderita gagal jantung. Berikut adalah beberapa contoh data pada *Heart Failure Prediction Dataset*.

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

Gambar 3. *Heart Failure Prediction Dataset*

Setelah melalui proses pengecekan *missing value* dan melakukan *encoding* untuk atribut yang bertipe *categorical* serta melakukan *feature scaling*, pada klasifikasi dengan KNN ini akan dilakukan dengan membagi data latih (*data training*) sebanyak 70% dan 30% digunakan sebagai data uji (*data testing*) terlebih dahulu. Untuk mengetahui apakah seseorang menderita gagal jantung atau tidak, digunakan 3 data tetangga terdekat pada awal klasifikasi, dengan kata lain pada KNN nilai *k* = 3. Dilanjutkan dengan proses evaluasi, dimana hasil pengklasifikasian dapat dilihat menggunakan *confusion matrix* yang akan digunakan untuk menentukan nilai akurasi seperti gambar berikut.



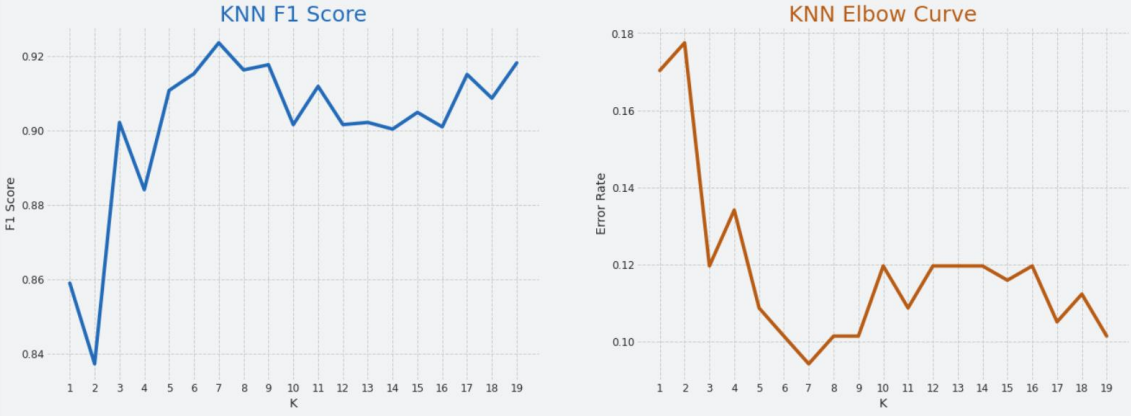
Gambar 4. *Confusion Matrix*

Berdasarkan gambar *confusion matrix* di atas, nilai akurasi dari model klasifikasi pada *Heart Failure Prediction Dataset* dengan menggunakan KNN dengan nilai *k* = 3 adalah 88% yang bisa dilihat pada gambar 5.

	precision	recall	f1-score	support
0	0.85	0.84	0.85	108
1	0.90	0.90	0.90	168
accuracy			0.88	276
macro avg	0.87	0.87	0.87	276
weighted avg	0.88	0.88	0.88	276

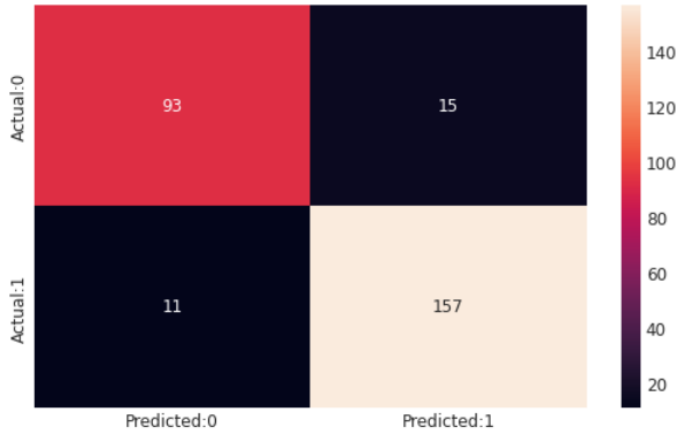
Gambar 5. Hasil Klasifikasi dengan metode KNN

Untuk menemukan hasil akurasi terbaik dalam metode *K-Nearest Neighbor*, nilai *k* memiliki pengaruh yang besar. Dimana hasil klasifikasi akan dipengaruhi oleh nilai *k* yang berbeda. Dengan memilih *k* yang paling optimal, maka dapat memperkecil nilai *error rate* yang dihasilkan. Berikut adalah grafik yang memperlihatkan nilai *k* paling optimal, dimana *error rate* yang dihasilkan bernilai kecil dan F1 *Score* yang tinggi.



Gambar 6. Grafik F1 *Score* dan *Error Rate*

Berdasarkan gambar di atas, ditentukan bahwa nilai *k* paling optimal adalah *k* = 7. Selanjutnya proses klasifikasi akan dilakukan kembali sesuai proses sebelumnya, namun dengan menggunakan *k* bernilai 7. Pada tahap evaluasi, kembali menggunakan *confusion matrix* yang menampilkan hasil klasifikasi seperti berikut.



Gambar 7. *Confusion Matrix* Hasil Klasifikasi

Hasil performa klasifikasi dengan metode *K-Nearest Neighbor* pada penyakit gagal jantung dapat dilihat pada gambar 8.

	precision	recall	f1-score	support
0	0.89	0.86	0.88	108
1	0.91	0.93	0.92	168
accuracy			0.91	276
macro avg	0.90	0.90	0.90	276
weighted avg	0.91	0.91	0.91	276

Gambar 8. Hasil Klasifikasi Terbaik dengan KNN

Pada gambar hasil klasifikasi di atas, diperoleh nilai akurasi terbaik yaitu 91%. Dimana nilai akurasi ini dihasilkan dengan menggunakan nilai *k* = 7. Nilai akurasi ini lebih tinggi dibandingkan nilai akurasi sebelumnya menggunakan *k* = 3 yaitu sebesar 88%.

4. Conclusion

Berdasarkan penelitian yang telah dilakukan, model klasifikasi gagal jantung menggunakan metode *K-Nearest Neighbor* menghasilkan nilai akurasi tertinggi, yaitu sebesar 91%. Nilai akurasi ini diperoleh ketika klasifikasi dilakukan dengan *k* = 7. Hal ini dapat membuktikan bahwa klasifikasi dengan menggunakan KNN menghasilkan performa yang baik dan dapat mengklasifikasikan secara akurat. Diharapkan penelitian ini dapat membantu tenaga medis dalam mengklasifikasikan penderita gagal jantung. Untuk mendapatkan hasil yang lebih baik, disarankan untuk menggunakan algoritma klasifikasi lainnya sehingga dapat diketahui algoritma mana yang paling efektif digunakan dalam proses klasifikasi penyakit gagal jantung.

References

[1] Nurkhalis and R. J. Adista, “Manifestasi Klinis dan Tatalaksana Gagal Jantung,” *Jurnal Kedokteran Nanggroe Medika*, vol. 3, no. 3, pp. 36–46, 2020.

[2] D. A. M. Reza, A. M. Siregar, and Rahmat, “Penerapan Algoritma K-Nearest Neighbord Untuk Prediksi Kematian Akibat Penyakit Gagal Jantung,” *Scientific Student Journal for Information, Technology and Science*, vol. III, no. 1, pp. 105–112, 2022.

[3] D. Cahyanti, A. Rahmayani, and S. A. Husniar, “Analisis Performa Metode KNN pada Dataset Pasien Pengidap Kanker Payudara,” *Indonesian Journal of Data and Science*, vol. 1, no. 2, pp. 39–43, 2020.

[4] A. Ariani and Samsuryadi, “Klasifikasi Penyakit Ginjal Kronis menggunakan K-Nearest Neighbor,” *Prosiding Annual Research Seminar 2019*, vol. 5, no. 1, pp. 148–151, 2019.

[5] R. Amilia and E. Prasetyo, “Klasifikasi Diagnosa Penyakit Demam Berdarah Dengue pada Anak Menggunakan Metode K-Nearest Neighbor Studi Kasus Rumah Sakit PKU Muhammadiyah Ujung Pangkah Gresik,” *Indexia: Informatic and Computational Intelegent Journal*, vol. 2, no. 2, pp. 1–10, 2020.

[6] I. A. A. Angreni, S. A. Adisasmita, M. I. Ramli, and S. Hamid, “Pengaruh Nilai K pada Metode K-Nearest Neighbor (KNN) Terhadap Tingkat Akurasi Identifikasi Kerusakan Jalan,” *Rekayasa Sipil*, vol. 7, no. 2, pp. 63–70, 2018.

[7] M. F. Rahman, M. I. Darmawidjadja, and D. Alamsah, “Klasifikasi Untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (RBNN),” *Jurnal Informatika*, vol. 11, no. 1, pp. 36–45, 2017.