ISBN : 978-979-587-846-9
Vol.5 No.1

Analisis Penyakit Jantung Menggunakan Metode KNN Dan Random Forest

Lia andiani[1]
Jurusan Ilmu Komputer,
Universitas Sriwijaya
Palembang, Indonesia
Liaandiani1995@gmail.com

Sukemi[2]
Jurusan Ilmu Komputer,
Universitas Sriwijaya
Palembang, Indonesia
Sukemi66@unsri.ac.id

Dian Palupi Rini[3]
Jurusan Ilmu Komputer,
Universitas Sriwijaya
Palembang, Indonesia
Dianpalupi@unsri.ac.id

Abstrak—Jantung merupakan organ tubuh manusia yang mempunyai peran penting dalam kehidupan manusia dan pastinya sangat berbahaya. *Angina stable* yaitu nyeri atau rasa tidak nyaman pada dada yang terjadi saat aktivitas atau stres. Rasa nyeri atau tidak nyaman ini dipicu oleh tingkat aktivitas atau stres yang relatif sama. *Angina unstable* Angin unstable yaitu angina yang pola gejalanya dapat berubah-ubah. Metode klasifikasi K-Nearest Neighbor (KNN) berdasarkan pada perhitungan kedekatan atau K. dan rasa sakit dada (angina) pada pasien. Metode K-Nearest Neighbor (KNN) dan Random forest digunakan ntuk menguji tingkat keberhasilan, penulis menggunakan model perhitungan *Precission, Recall, F1-Score,* dan Akurasi. Berdasarkan hasil uji coba yang telah dilakukan pada penelitian ini, terbukti bahwa metode K-Nearest Neighbor dapat digunakan untuk mengkalsifikasi data penyakit jantung. Dengan akurasi metode KNN 93% dan metode Random Forest 72%.

Kata Kunci—*Jantung, Klasifikasi, KNN, Random forest, Tipe Angina.*

Abstract—The heart is an organ of the human body that has an important role in human life and is certainly very dangerous. Angina stable is pain or discomfort in the chest that occurs during activity or stress. This pain or discomfort is triggered by a relatively equal level of activity or stress. Unstable angina Unstable winds, i.e. angina whose patterns can change. The K-Nearest Neighbor (KNN) classification method is based on calculating proximity or K. and chest pain (angina) in patients. The K-Nearest Neighbor (KNN) method and the random forest were used to approve the success rate, the authors used the Precission, Recall, F1-Score, and Accuracy calculation models. Based on the results of trials that have been carried out in this study, it is proven that the K-Nearest Neighbor method can be used to calcify heart disease data. With the accuracy of the KNN 93% method and the 72% Random Forest method.

Keywords—*Heart, classification, KNN, random forest, typical of Angina.*

I. PENDAHULUAN

Penyakit Jantung telah menjadi penyebab kematian utama di Indonesia. Di Provinsi Jawa Tengah berdasarkan laporan dari Rumah Sakit dan Puskesmas tahun 2006, kasus Penyakit jantung sebesar 26,38 per 1000 penduduk [1]. Penyakit Jantung mempunyai faktor risiko yang bisa diubah. Banyak orang terkena serangan jantung tanpa ada gejala apapun sebelumnya. Selama 50 tahun terakhir, semakin banyak orang terkena penyakit jantung koroner, dan beberapa faktor penyebab utamanya telah diketahui. Sebelumnya , sakit jantung di prediksi 30 % penyebab kematian manusia . diperkirakan WHO pada tahun 2005 ,kematian di sebabkan penyakit jantung ,penyakit jantung rematik meningkat menjadi 17,5 juta 14,4 juta di tahun 1990 .dari jumlah tersebut banyak yang di kaitkan dengan penyakit jantung angina [2].

Jantung merupakan organ tubuh manusia yang mempunyai peran penting dalam kehidupan manusia dan pastinya sangat berbahaya. Jika jantung kita mempunyai masalah akan bermasalah pula pada organ kita yang lainnya. mengingat bahwa banyak kematian yang disebabkan oleh penyakit jantung. Tapi dengan pengetahuan dan informasi yang minim, mustahil untuk dapat menjaga kesehatan jantung. Oleh karena itu dibutuhkan analisis penyakit jantung untuk membantu para ahli penyakit jantung melihat peluang penyakit jantung pada pasien berdasarkan umur atau type rasa sakit di dada .

Jantung merupakan pusat kehidupan. Banyak faktor yang mempengaruhi penyakit jantung seperti faktor fikiran dan faktor makanan . ciri-ciri penyakit jantung jika napas terasa berat , pada dada terasa sakit , sakit punggung, pingsan , gemetar, dada terasa panas . Angina adalah kondisi yang biasanya terjadi pada tengah malam dan subuh [1]. Angina ada bermacam-macam. *Angina stable* yaitu nyeri atau rasa tidak nyaman pada dada yang terjadi saat aktivitas atau stres. Rasa nyeri atau tidak nyaman ini dipicu oleh tingkat aktivitas atau stres yang relatif sama. *Angina unstable* yaitu angina yang pola gejalanya dapat berubah-ubah. Biasanya terjadi pada malam hari, saat tidur. Angina ini dapat terjadi lebih sering dan lebih berat dibanding *angina stable*. Angina variant sangat jarang. Kekakuam arteri koronaria menyebabkan angina jenis ini. Angina variant biasanya terjadi saat Anda sedang beristirahat dan nyeri dapat terasa lebih berat. Angina ini dapat diredakan dengan obat-obatan.

Analisis dilakukan untuk mempermudah ahli jantung menyusun data atau menggolongkannya ke dalam pola atau tema. Kemudian memberikan makna terhadap analisis yang dilakukannya, menjelaskan kategori atau pola didalamnya, serta mencari hubungan antara berbagai data .

Penggunaan metode K-Nearest Neighbor (KNN) dan random forest digunakan untuk menguji tingkat keberhasilan, penulis menggunakan model perhitungan *Precission, Recall, F1-Score,* dan Akurasi .

ISBN : 978-979-587-846-9
Vol.5 No.1

II. METODOLOGI

A. Data Persiapan

Data diambil dari kaggle, berupa file CSV dengan total data 1025 buah. Dengan jumlah 526 pasien yang memiliki penyakit jantung dan 499 pasien yang tidak memiliki penyakit jantung. Data ini diambil di tahun 1988 dari sebuah rumah sakit dan terdiri dari empat database: Cleveland, Hongaria, Swiss, dan Long Beach V. Informasi data yang digunakan adalah dataset dari cleveland yang berisi 76 atribut, termasuk atribut yang diprediksi, tetapi semua hasil yang pernah diterbitkan sebelumnya merujuk pada penggunaan subset 14 diantaranya. Ada bidang tambahan yaitu bidang "target" mengacu pada adanya penyakit jantung pada pasien. Berisi bilangan bulat bernilai 0 = tidak ada penyakit dan 1 = penyakit [3].

B. Informasi atribut

Informasi Atribut pada data penyakit jantung dibutuhkan untuk dapat mengelompokkan data dan mempermudah analisis data dalam menentukan penyakit jantung berdasarkan kategorinya [3]. Atribut yang digunakan ditunjukkan pada TABEL 1.

TABEL 1 TABEL ATRIBUT

No	Attribute Name	Description	Range of valeus
1	age	Age of the person in years	29 to 79
2	sex	Gender of the person [1:Male,0: Female]	0,1
3	cp	Chest pain type [1-Typical Type 1 Angina 2-AtypicalTypeAngina 3-Non-angina pain 4-Asymptomatic)	1,2,3,4
4	trestbps	Resting Blood Pressure in mm Hg	94 to 200
5	chol	Serum cholesterol in mg/dl	126 to 564
6	fbs	Fasting Blood Sugar in mg/dl	0,1
7	restecg	Resting Electrocardiographic Results	0,1,2
8	thalach	Maximum Heart Rate Achieved	71 to 202
9	exang	Exercise Induced Angina	0,1
10	Oldpeak	ST depression	1 to 3
11	slope	Slope of the Peak Exercise ST segment	1,2,3
12	ca	Number of major vessels colored by fluoroscopy	0 to 3
13	thal	3–Normal,6–FixedDefect,7–ReversibleDefect	3,6,7
14	num	Class Attribute	0 or 1

C. Analisis

Analisis mempunyai fungsi untuk mengumpulkan data-data yang terdapat pada suatu lingkungan tertentu [4] . Analisis dapat diterapkan diberbagai jenis lingkungan dan keadaan. Menyusun data atau menggolongkannya ke dalam pola atau tema. Kemudian memberikan makna terhadap analisis, menjelaskan kategori atau pola, serta mencari hubungan antara berbagai data.

D. Klasifikasi

Klasifikasi adalah suatu proses pengkategorian atau prosedur pembelajaran terawasi yang digunakan untuk memprediksi hasil dari data yang ada. Klasifikasi ini menjadi suatu pendekatan untuk diagnosis penyakit jantung dengan menggunakan algoritma klasifikasi, dan untuk meningkatkan klasifikasi pengklasifikasian sebagai pelengkap dari pengklasifikasi.

E. K- Nearest Neighbor (KNN)

Pengklasifikasian menggunakan fungsi jarak dari data baru ke data training (kedekatan). Prinsip kerja K-Nearest Neighbor (KNN) adalah mencari jarak terdekat antara data yang akan dievaluasi dengan K tetangga (*neighbor*) terdekatnya dalam data pelatihan [5]. Data pelatihan diproyeksikan ke ruang berdimensi banyak, dimana masing-masing dimensi merepresentasikan fitur dari data. Ruang ini dibagi menjadi bagian-bagian berdasarkan klasifikasi data pelatihan. Algoritma di perlihatkan pada Gambar 1.

Rumus K-NN ditunjukkan pada persamaan (1)

$$d_1 = \sqrt{\sum_{i=1}^{p}(X_{2i} - X_{2i})^2} \quad \text{...(1)}$$

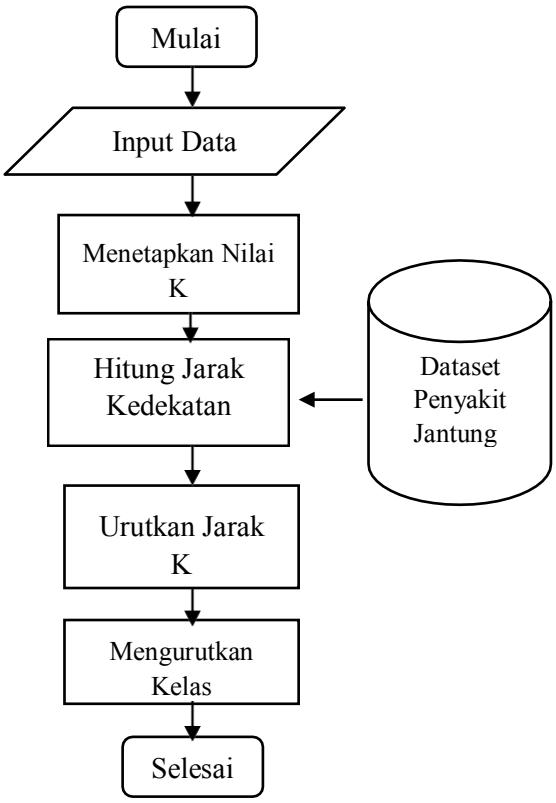
Keterangan:

X_1 = Sampel Data
X_2 = Data Uji / Testing
i = Variabel Data
d = Jarak
p = Dimensi Data

Klasifikasi menggunakan voting terbanyak diantara klasifikasi dari K obyek. Algoritma KNN menggunakan

ISBN : 978-979-587-846-9
Vol.5 No.1

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.



Gambar 1. Algoritma K - NN

Keterangan algoritma K-Nearest Neighbor:

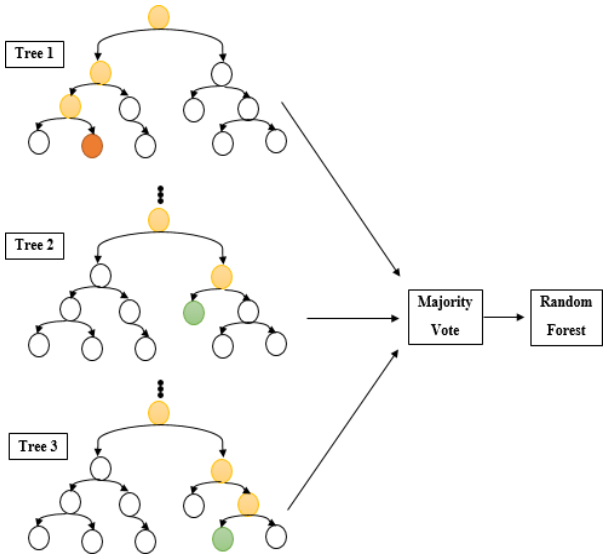
1. Tentukan parameter K
2. Hitung jarak antara data yang akan dievaluasi dengan semua data latih.
3. Urutkan jarak data
4. Tentukan jarak terdekat sampai urutan K
5. Pasangkan kelas yang bersesuaian
6. Cari jumlah kelas dari tetangga yang terdekat dan tetapkan kelas tersebut sebagai kelas data yang akan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

F. *Random Forest*

Random forest (RF) adalah suatu algoritma yang digunakan pada klasifikasi data dalam jumlah yang besar. Klasifikasi *random forest* dilakukan melalui penggabungan pohon (*tree*) dengan melakukan *training* pada sampel data yang dimiliki. Penggunaan pohon (tree) yang semakin banyak akan mempengaruhi akurasi yang akan didapatkan menjadi lebih baik. Penentuan klasifikasi dengan *random forest* diambil berdasarkan hasil *voting* dari *tree* yang terbentuk. Pemenang dari *tree* yang terbentuk ditentukan dengan *vote* terbanyak.

[6]. Pohon keputusan dimulai dengan cara menghitung nilai entropy sebagai penentu tingkat ketidakmurnian atribut. Untuk menghitung nilai entropy digunakan rumus seperti pada persamaan (2).

$$\text{Entropy (Y)} = -\sum i \; p(c|Y) \; log_2 \; p(c|Y) \quad \text{...(2}$$



Gambar 3. Rumus Entropy Random Forest

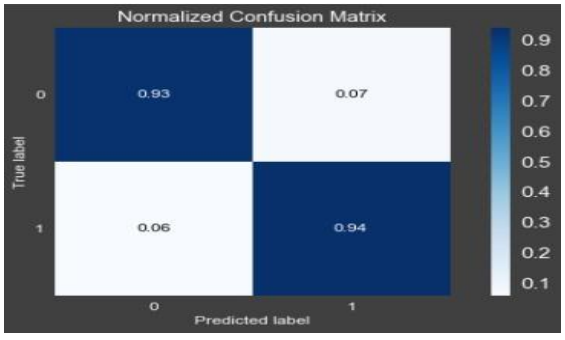
G. *Confusion matrix*

Metode ini menggunakan tabel matriks yang digunakan untuk membandingkan jumlah TP terhadap jumlah record yang positif dengan jumlah TN terhadap jumlah record yang negatif. Sebuah tabel klasifikasi binary untuk menentukan data yang digunakan bersifat *True Positives (TP), False Positives (FP), False Negatives (FN),* dan *True Negatives (TN).* Penentuan dilihat dari *Gambar 3*. Untuk menghitung *precision*, *Recall*, dan *F1-score* digunakan persamaan (3), (4), dan (5) di bawah ini:

$$Precision = \frac{TP}{TP+FP} \quad \text{...(3)}$$

$$Recall = \frac{TP}{TP+FN} \quad \text{...(4)}$$

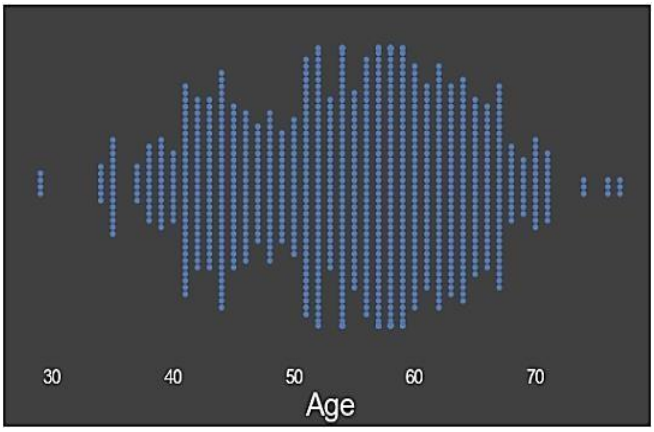
$$\text{F1} = \left(\frac{2}{recall^{-1} + precision^{-1}}\right) \quad \text{...(5)}$$



ISBN : 978-979-587-846-9
Vol.5 No.1

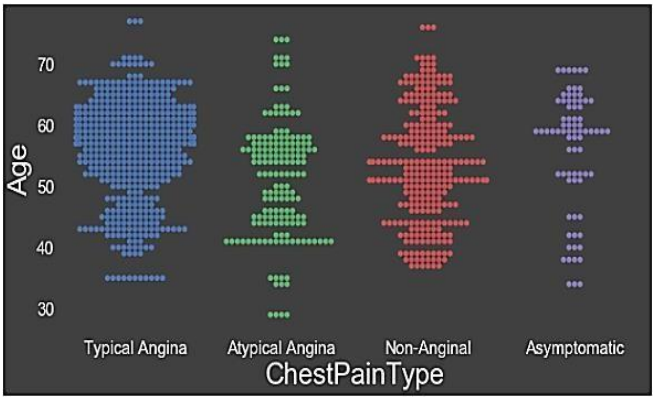
III. ANALISIS DATA

Analisis data dilakukan setelah mendapatkan data jantung yang telah dikelompokan perkategori. Kemudian, tahap selanjutnya adalah proses pengolahan data agar data dapat di kelompokkan atau dikategorikan sehingga, membentuk sebuah pola tertentu dan diberi makna [1] . Pengelompokan data ini di uji dan dihasilkan menghasilkan akurasi menggunakan bahasa phyton dalam program jupyter notebook. Analisis yang pertama dilakukan bersadarkan umur dengan rentang 30-70 tahun yang di tunjuukan pada Gambar 4.

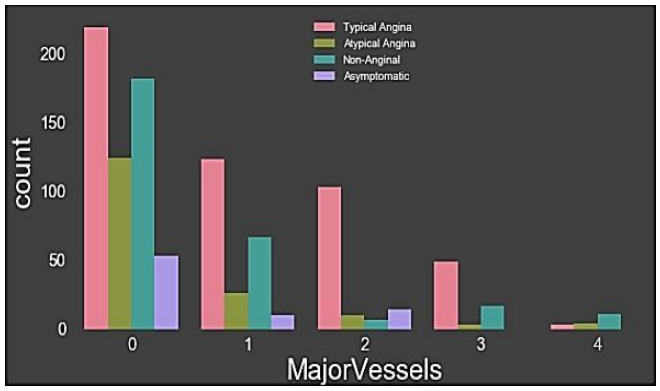


Gambar 5. Pengelompokan data berdasarkan umur

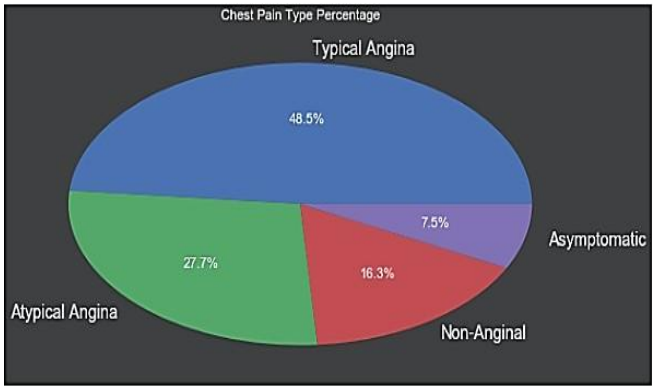
Berdasarkan Gambar 5 dapat dilihat bahwa umur dapat menentukan rasa tidak nyaman atau rasa sakit pada jantung. Meliputi typical angina, atypical angina, non—angina dan asyptomatic. Pada Gambar 6 dapat dilihat relasi antara jumlah pembuluh darah besar (vessel) dengan type rasa sakit dada yang ada pada pasien. Gambar 7 yang menunjukkan presentase distribusi sakit pada dada setiap jenis nyeri dada.



Gambar 6. Visualisasi tipe sakit dada berdasarkan umur



Gambar 4. Visualisasi jumlah pembuluh darah besar dan tipe sakit dada



Gambar 7. Visualisasi presentase rasa sakit dada

A. Evaluasi

Evaluasi dilakukan secara mendalam dengan tujuan agar hasil pada tahap pemodelan sesuai dengan sasaran yang ingin dicapai. Hasil perhitungan pengujian metode K-Nearest Neighbor (KNN) dan metode Random Forest dengan model *precision*, *recall*, dan *f1-score* dapat di lihat pada TABEL 2 dan TABEL 3 .

TABEL 2 TABEL EVALUASI METODE KNN

No	**Evaluasi**	**Ket**
1	*Precision*	0.93
2	*Recall*	0.92
3	*F1-Score*	0.93

TABEL 3 TABEL EVALUASI METODE RANDOM FOREST

No	**Evaluasi**	**Ket**
1	*Precision*	0.80
2	*Recall*	0.67
3	*F1-Score*	0.73

ISBN : 978-979-587-846-9
Vol.5 No.1

IV. HASIL

Analisis Penelitian ini dilaksanakan di bagian rekam medik sebuah Rumah Sakit di cleveland 1988. Berdasarkan data rekam medik secara retrospektif diperoleh sebanyak 1025 pasien, meliputi 526 penderita penyakit jantung dan 499 tidak penderita penyakit jantung.

Klasifikasi data menggunakan metode K-Nearest Neighbor (KNN) dihasilkan dari perhitungan akurasi persamaan (6) .

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} * 100\% \quad ...(6)$$

Hasil menyatakan bahwa nilai akurasi terhadap klasifikasi metode K-Nearest Neighbor (KNN) sebesar 0,93% dan metode Random Forest sebesar 0,73%.

V. KESIMPULAN

Penelitian ini ingin menganalisa penyakit jantung berdasarkan umur dan rasa sakit dada (angina) pada pasien. Metode K-Nearest Neighbor (KNN) dan random forest digunakan ntuk menguji tingkat keberhasilan, penulis menggunakan model perhitungan *Precission, Recall, F1-Score,* dan Akurasi.

Berdasarkan hasil uji coba yang telah dilakukan pada penelitian ini, terbukti bahwa metode K-Nearest Neighbor dapat digunakan untuk mengklasifikasi data penyakit jantung. Dengan akurasi metode KNN 93% dan metode Random Forest 72%.

REFERENCES

[1] D. Zahrawardani, K. S. Herlambang, and H. D. Anggraheny, "Analisis Faktor Risiko Kejadian Penyakit Jantung Koroner di RSUP Dr Kariadi Semarang," *J. Kedokt. Muhammadiyah*, vol. 1, no. 3, p. 13, 2013.

[2] M. Mawi, "Indeks massa tubuh sebagai determinan penyakit jantung koroner pada orang dewasa berusia di atas 35 tahun," vol. 23, no. 3, pp. 87–92, 2001.

[3] C. B. C. Latha and S. C. Jeeva, "Informatics in Medicine Unlocked Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics Med. Unlocked*, vol. 16, no. November 2018, p. 100203, 2019.

[4] K. V Kiruthikaa, V. F. J, and S. Yuvaraj, "Analysis of Prediction Accuracy of Heart Diseases using Supervised Machine Learning Techniques for Developing Clinical Decision Support Systems," no. 4, pp. 433–437, 2018.

[5] M. E. I. Lestari, "PENERAPAN ALGORITMA KLASIFIKASI NEAREST NEIGHBOR (K-NN) UNTUK MENDETEKSI PENYAKIT JANTUNG," vol. 7, no. September 2010, pp. 366–371, 2014.

[6] Y. S. Nugroho, "Sistem Klasifikasi Variabel Tingkat Penerimaan Konsumen Terhadap Mobil Menggunakan Metode Random Forest," no. June, 2017.