InComTech: Jurnal Telekomunikasi dan Komputer
vol.13, no.2, Agustus 2023, 145-154
http://publikasi.mercubuana.ac.id/index.php/Incomtech
P-ISSN: 2085-4811 E-ISSN: 2579-6089

Prediksi Penyakit Jantung Menggunakan *Attribute Weighting k*-Nearest Neighbor

Agustiyar[1*]

[1]*LLDIKTI Wilayah VI Semarang,*
*Email Penulis Koresponden: agustiyar001@gmail.com

Abstrak:
Penyakit kardiovaskular atau lebih dikenal dengan penyakit jantung menjadi salah satu penyebab kematian tertinggi di Indonesia dan di tingkat global. Selain pola hidup sehat untuk mencegah penyakit tersebut, deteksi dini terhadap resiko penyakit jantung dapat dilakukan dengan data mining atau machine learning salah satunya *k*-NN. *k*-NN adalah salah satu metode data mining paling sederhana dan kuat dalam konsistensi hasil klasifikasi, akan tetapi memiliki kekurangan yaitu memberikan bobot yang sama kepada semua atribut. Penelitian ini mengusulkan pembobotan pada atribut untuk mengatasi kelemahan tersebut. Prediksi penyakit jantung digunakan untuk menggambarkan kinerja metode usulan. Pada penelitian ini menggunakan dataset Heart Disease, sebuah dataset publik dari University of California Irvine. Dengan menggunakan nilai k 3, 5, 7, 9 diperoleh rata-rata kinerja metode usulan sebesar 79,87% lebih baik dibandingkan Chi-Square *k*-NN 79,08% dan *k*-NN klasik 65,89%. Penelitian ini menyimpulkan bahwa metode pembobotan atribut berhasil mengatasi kekurangan *k*-NN, jadi metode usulan cocok untuk prediksi penyakit jantung.

Kata Kunci:
Prediksi Penyakit Jantung;
k-Nearest Neighbor;
Pembobotan Atribut;
Weighted Euclidean Distance

Riwayat Artikel:
Diserahkan 23 November 2022
Direvisi 30 Juli 2023
Diterima 31 Juli 2023
Dipublikasi 15 Agustus 2023

DOI:
10.22441/incomtech.v13i2.17883

1. PENDAHULUAN

Penyakit kardiovaskular atau lebih dikenal dengan penyakit jantung merupakan penyebab utama kematian di tingkat global [1]. Pada tahun 2019 terdapat sekitar 17,9 juta jiwa meninggal akibat penyakit jantung [2]. Di Indonesia penyakit jantung menjadi penyebab kematian terbanyak kedua setelah Stroke [3]. Jadi, penyakit jantung menjadi penyebab kematian tertinggi baik di Indonesia maupun di tingkat global. Penyakit jantung merupakan penyakit tidak menular yang kejadiannya bisa dicegah dengan pola hidup sehat. Pola hidup sehat tersebut dapat dilakukan dengan Gerakan Masyarakat Hidup Sehat (Germas) antara lain aktivitas fisik minimal 30 menit sehari, makan buah dan sayur, dan cek kesehatan secara berkala [3]. Metode

pencegahan yang lain adalah dengan memanfaatkan data mining atau machine learning [4] untuk deteksi dini terhadap resiko penyakit jantung.

Saat ini telah banyak penelitian tentang prediksi penyakit jantung menggunakan data mining atau machine learning, diantaranya penelitian Krishnaiah, Narsimha, dan Chandra, 2015 [4], penelitian R dan Thomas, 2017 [5], penelitian Mirza *et al.,* 2019 [6], penelitian 2016 Enriko, Suryanegara and Gunawan, 2016 [7], penelitian Singh dan Kumar, 2020 [8]. *k*-NN merupakan algoritma klasifikasi yang populer, mengklasifikasi data tidak berlabel ke dalam data berlabel yang terdekat. Sejumlah *k* data berlabel yang dekat dengan data tidak berlabel dipilih, kemudian label yang sering muncul dari sejumlah *k* data berlabel tersebut dipilih menjadi label data yang semula tidak berlabel [9]. *k*-NN merupakan metode klasifikasi yang sederhana, bagus untuk klasifikasi data yang besar [10]. Menurut R dan Thomas [5] *k*-NN merupakan metode klasifikasi yang paling sederhana dan kuat dalam konsistensi hasil klasifikasi. Menurut Nayak et al. *k*-NN merupakan metode klasifikasi yang sederhana, akan tetapi hasil klasifikasi sangat dipengaruhi oleh irrelevant attributes [10]. Menurut Liu, Zhu dan Qin [11] *k*-NN merupakan metode data mining yang paling sederhana, akan tetapi memiliki kelemahan, salah satunya adalah memberikan bobot yang sama kepada semua atribut. Menurut Dialameh dan Jahromi, 2016 [12] pembobotan atribut dapat meningkatkan kinerja machine learning dengan mengatasi masalah yang diakibatkan irrelevant, redundant dan noisy atribut.

Penelitian ini mengusulkan Attribute Weighting *k*-Nearest Neighbor, dengan menerapkan pembobotan pada atribut untuk mengatasi kelemahan *k*-NN yang memberikan bobot yang sama kepada semua atribut. Metode usulan akan dibandingkan dengan penelitian sebelumnya.

2. METODE

2.1 Metode *k*-NN

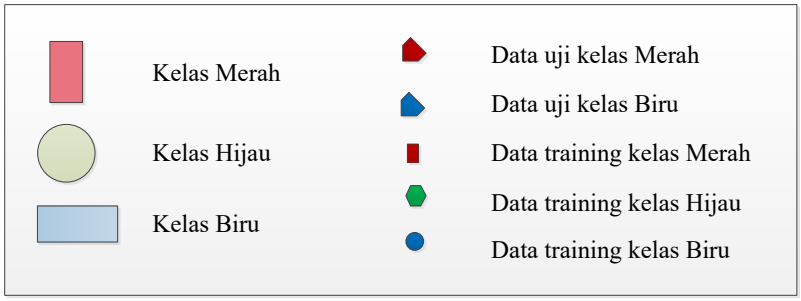
k-NN dikenal sebagai algoritma yang populer, simpel dan kuat dalam konsistensi hasil klasifikasi dan telah banyak digunakan dalam berbagai keperluan prediksi klasifikasi, diantaranya klasifikasi penyakit jantung. Penelitian [7] *k*-NN dengan pembobotan atribut Chi-Square digunakan untuk prediksi penyakit jantung. Penelitian [13] k-NN dikombinasikan dengan teknik Re-Sampling untuk prediksi penyakit jantung. Penelitian [14] *k*-NN dipadukan dengan data pre-processing untuk prediksi penyakit jantung, hasilnya kinerja *k*-NN lebih unggul dari Naïve Bayes, Decision Tree, dan Random Forest.

k-NN adalah metode klasifikasi yang paling sederhana, akan tetapi memiliki kelemahan, di mana semua atribut dianggap memiliki bobot yang sama [11] sedangkan pada kenyataannya tidak semua atribut memiliki pengaruh yang sama terhadap kinerja algoritma [15]. Menurut [12] [16] pemberian bobot pada fitur atau atribut dapat meningkatkan kinerja algoritma.

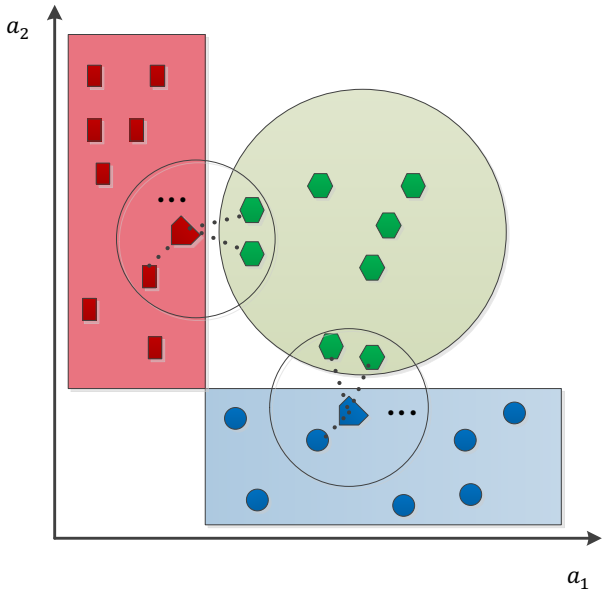
Penelitian ini mengusulkan *Attribute Weighting k*-Nearest Neighbor, di mana pembobotan atribut diterapkan guna mengatasi kelemahan *k*-NN yang memberikan bobot yang sama kepada semua atribut. Metode usulan dibandingkan dengan penelitian sebelumnya.

2.2 Metode Pembobotan Atribut

Pada Gambar 2, Gambar 3 dicontohkan permasalahan klasifikasi *k*-NN pada data dengan dua atribut (a_1 dan a_2) dan tiga kelas/label, ketiga kelas tersebut adalah Merah (M), Biru (B), dan Hijau (H). Sebagai contoh diambil dua data uji, satu dari kelas M dan satu dari kelas B.



Gambar 1. Legenda Gambar 2, Gambar 3

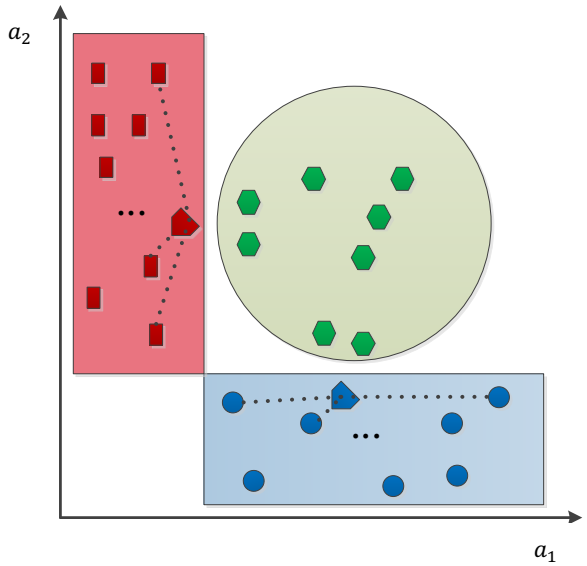


Gambar 2. Klasifikasi *k*-NN Tanpa Pembobotan Atribut

Pada Gambar 2, klasifikasi *k*-NN dengan *k*=3, fungsi jarak Euclidean distance Fungsi

(2) tanpa pembobotan atribut, hasilnya dua data uji tersebut salah diklasifikasikan ke dalam kelas H. Hal ini terjadi karena atribut a_1 dan a_2 diperlakukan sama pentingnya oleh Euclidean distance. Akan tetapi, jika Gambar 2 diperhatikan lebih dalam, pada kelas B sebaran data lebih banyak di a_1 sehingga dapat dikatakan a_1 berkontribusi lebih besar dibandingkan a_2 dalam menentukan anggota kelas B. Sementara pada kelas M, sebaran data lebih banyak di a_2 sehingga dapat dikatakan a_2 berkontribusi lebih besar dibandingkan a_1 dalam menentukan anggota kelas M. Sedangkan pada kelas G, a_1 dan a_2 tersebar secara merata dan dalam rentang yang

sama sehingga dapat dikatakan a_1 dan a_2 memiliki kontribusi yang sama dalam menentukan anggota kelas G.



Gambar 3. Klasifikasi k-NN Menggunakan Pembobotan Atribut

Untuk mengatasi kelemahan pada Gambar 2, maka pada Gambar 3 Euclidean distance ditambah bobot *W* sehingga disebut weighted Euclidean distance Fungsi

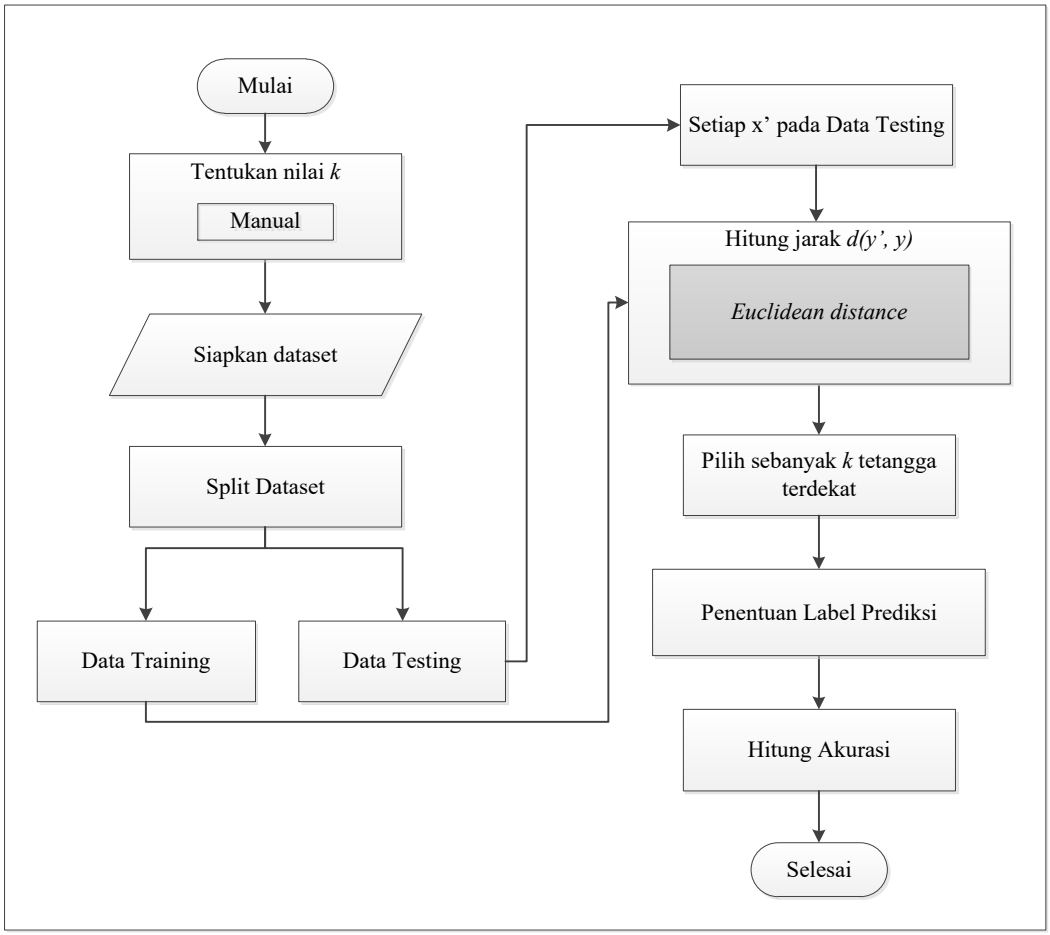
(3). Setiap atribut (a_1 dan a_2) pada kelas M, B, dan H diberikan bobot w_M, w_B, dan w_H dapat dilihat pada Fungsi (1).

$$W = \begin{bmatrix} w_M \\ w_B \\ w_H \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \quad (1)$$

Dengan pemberian bobot W seperti itu kedua data uji dapat diklasifikasikan dengan benar oleh *k*-NN sebagaimana Gambar 3.

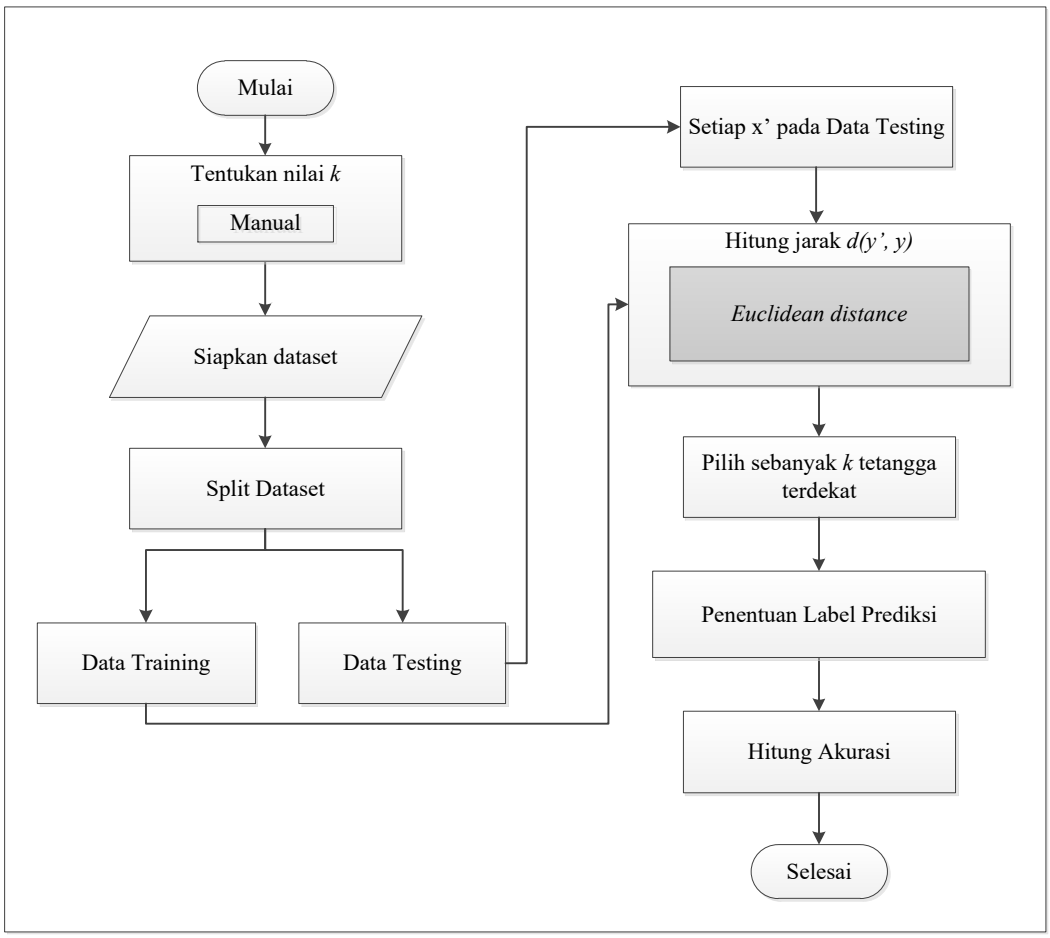
2.3 Metode yang Diusulkan

Pada penelitian ini mengusulkan *Attribute Weighting k*-Nearest Neighbor (AWKNN). Perbedaan metode usulan dengan *k*-NN klasik terletak pada fungsi jarak yang digunakan, pada metode usulan menggunakan *weighted Euclidean distance* sebagai fungsi jaraknya sedangkan pada *k*-NN klasik menggunakan *Euclidean distance* tanpa tambahan bobot.

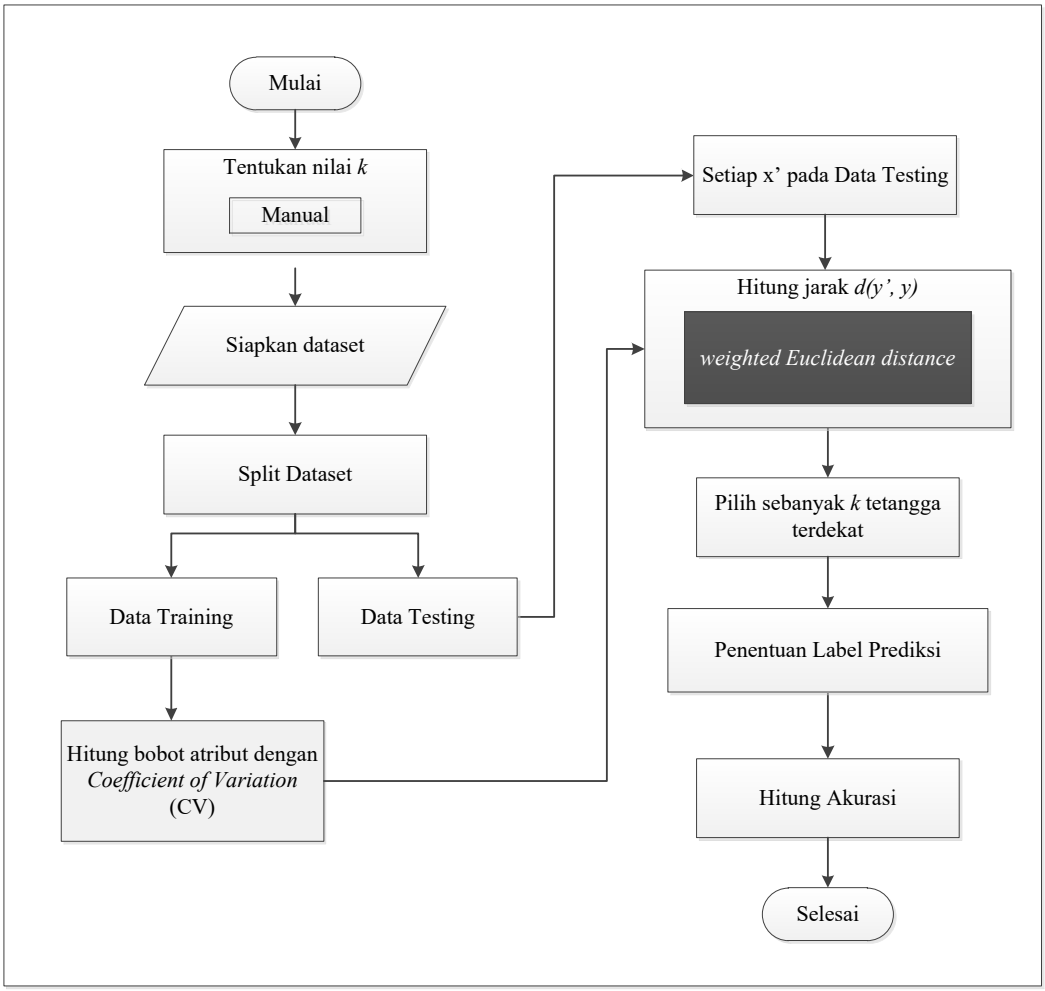


Gambar 4. Metode *k*-NN Klasik

Pada



Gambar 4 ditunjukkan diagram aktifitas algoritma *k*-NN klasik. Dataset dipecah menjadi data training dan testing, kemudian setiap data testing diukur jaraknya terhadap data training dengan *Euclidean distance* Fungsi (2) tanpa diberikan bobot. Sejumlah *k* tetangga terdekat dipilih, label prediksi ditentukan berdasar mayoritas label dari *k* tetangga terdekat yang telah terpilih, kemudian kinerja *k*-NN dihitung menggunakan akurasi.



Gambar 5. Metode Usulan AWKNN

Pada Gambar 5 ditunjukkan diagram aktifitas metode usulan, di mana w*eighted Euclidean distance* dipakai untuk menangani kelemahan *k*-NN yang memberikan bobot yang sama kepada semua atribut. Nilai *k* ditentukan secara manual menggunakan nilai *k* ganjil [17] supaya tidak ambigu dalam penentuan label prediksi. Dataset dibagi menjadi data *training* dan data *testing*. Pada data *training*, setiap atribut dihitung bobotnya dengan *coefficient of variation* Fungsi (5). Selanjutnya setiap data *testing* dihitung jaraknya terhadap data *training* menggunakan *weighted Euclidean distance*. Kemudian mengambil sebanyak *k* tetangga untuk dijadikan kandidat hasil klasifikasi, penentuan label hasil klasifikasi menggunakan Fungsi

(8). Terakhir, mengukur kinerja metode usulan dengan menghitung akurasi.

Weighted Euclidean distance digunakan dalam mengukur jarak data *training* dengan data *testing* sebagaimana Fungsi

(3) d: distance, y': data *testing*, y: data *training*, a: atribut, n: jumlah atribut, w adalah bobot yang dihitung menggunakan Fungsi (4) a: atribut, n: jumlah atribut, cv_a: *coefficient of variation* atribut ke-a.

$$d(y',y) = \sqrt{\sum_{a=1}^{n} \left(y'_a - y_a\right)^2} \tag{2}$$

$$d(y',y) = \sqrt{\sum_{a=1}^{n} w\left(y'_a - y_a\right)^2} \tag{3}$$

$$w_a = \frac{cv_a}{\sum_{a=1}^{n} cv_a} \tag{4}$$

$$cv_a = \frac{\delta_a}{\mu_a} \tag{5}$$

di mana δ_a: standar deviasi atribut ke-a, μ_a: nilai rata-rata atribut ke-a.

$$\delta_a = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ia} - \mu_a)^2} \tag{6}$$

di mana a: atribut, m: jumlah data, x_{ia}: data ke-i atribut ke-a, μ_a: nilai rata-rata atribut ke-a.

$$\mu_a = \frac{1}{n}\sum_{i=1}^{n} x_{ia} \tag{7}$$

di mana a: atribut, m: jumlah data, x_{ia}: data ke-i atribut ke-a. Pada metode usulan penentuan label hasil klasifikasi menggunakan Fungsi (8).

$$l' = arg\,max_l \sum_{i=1}^{k} l = l_i \tag{8}$$

di mana l': label hasil klasifikasi, l: label, l_i: label ke-i dari k tetangga terdekatnya, k: jumlah tetangga terdekat.

3. HASIL DAN PEMBAHASAN

Intel Core i7-7500U 2,70 GHz, 8 GB, dan Sistem Operasi Windows 10 Home 64-bit digunakan dalam eksperimen. Metode usulan dibangun dengan bahasa pemrograman Python.

Heart Disease dataset (Hungarian) dari UCI Machine Learning Repository digunakan dalam penelitian ini sebagaimana penelitian [7] [13] [14]. Dataset ini terdiri atas 303 data, 14 atribut, dan 2 label 0 (sehat) dan 1 (memiliki penyakit jantung). Pada dataset dilakukan *pre-processing*. Ringkasan Heart Disease dataset ditampilkan pada Tabel 1.

Tabel 1 Ringkasan Heart Disease Dataset

No	Atribut	Keterangan
1	Age	Umur (th)
2	Sex	Kelamin (laki-laki atau perempuan)
3	Cp	Chest pain type/Jenis nyeri dada
4	Thestbps	Resting blood pressure (tekanan darah istirahat)
5	Cho	Serum cholesterol (kolestrol)
6	Restecg	Resting electrographic results (elektrografi istirahat)
7	Fbs	Fasting blood sugar (gula darah puasa)
8	Thalach	Max. heart rate achieved (detak jantung maksimal)
9	Exang	Exercise induced angina
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope of the peak exercise ST segment
12	Ca	No. of major vessels colored
13	Thal	Defect type (jenis cacat)
14	Label	0,1 (Healthy, Have Heart Disease)

Pada penelitian ini digunakan *state-of-the-art stratified 10-fold cross-validation* pada tahap *learning* and *testing*, di mana dataset dibagi menjadi 10 bagian yang sama dan dilakukan proses *learning* sebanyak 10 kali. *10-fold cross-validation* digunakan karena metode ini menjadi metode validasi standar dan *state-of-the-art* dalam penelitian [18]. Pada Tabel 2 ditampilkan metode validasi *Stratified 10 Fold Cross Validation.* Dataset dibagi menjadi sepuluh bagian sama besar, pada validasi pertama, satu bagian yang berwarna hitam menjadi data *testing* sedangkan sembilan bagian yang tidak berwarna menjadi data *training*. Ketika satu bagian menjadi data *testing*, sembilan bagian menjadi data *training*, begitu seterusnya.

Tabel 2 Stratified 10 Fold Cross Validation

n-validasi	Partisi Dataset									
1	■									
2		■								
3			■							
4				■						
5					■					
6						■				
7							■			
8								■		
9									■	
10										■

Eksperimen pada algoritma *k*-NN klasik, dan metode usulan menggunakan nilai *k* (3,5,7,9) [19] kemudian hasilnya dibandingkan dengan penelitian [7]. Pada Tabel 3 ditampilkan hasil eksperimen metode usulan dan perbandingan dengan metode sebelumnya. Pada Tabel 3 ditampilkan akurasi algoritma yang diperbandingkan. Pada *k*-NN klasik akurasi tertinggi 66,77% diperoleh dengan *k=5,* sedangkan rata-ratanya sebesar 65,89%. Pada metode *Chi-Square*+*k*-NN [7] akurasi tertinggi 80,61% diperoleh dengan *k*=9, sedangkan rata-ratanya sebesar 79,08%. Pada metode usulan (AWKNN) akurasi tertinggi 81,32% diperoleh dengan *k*=3, sedangkan rata-ratanya sebesar 79,87%.

Tabel 3 Akurasi Algoritma yang Diperbandingkan

Metode/ Nilai *k*	***3***	***5***	***7***	***9***	***Rata-Rata***
k-NN klasik	65,14%	66,77%	65,35%	66,32%	65,89%
Chi-Square+*k*-NN [7]	78,23%	79,25%	78,23%	80,61%	79,08%
Metode Usulan (AWKNN)	81,32%	80,28%	78,26%	79,62%	**79,87%**

4. KESIMPULAN

Attribute Weighting *k*-Nearest Neighbor (AWKNN) merupakan pengembangan *k*-NN di mana Euclidean Distance dimodifikasi menjadi Weighted Euclidean Distance. Dengan Weighted Euclidean Distance, bobot setiap atribut diperhitungkan dalam mengukur kedekatan antar data. Hasil uji coba AWKNN pada prediksi penyakit jantung dengan dataset publik Heart Disease dataset (Hungarian) diperoleh hasil kinerja AWKNN lebih baik dibandingkan dengan *k*-NN. Dengan pembobotan atribut dapat meningkatkan kinerja machine learning sesuai penelitian [12]. Jadi metode yang diusulkan mampu mengatasi kekurangan algoritma *k*-NN. Pada penelitian ini untuk meningkatkan kinerja *k*-NN dilakukan

pembobotan atribut. Saran untuk penelitian selanjutnya dapat dicoba pembobotan pada neighbors untuk meningkatkan kinerja *k*-NN.

REFERENSI

[1] World Health Organization, "The top 10 causes of death," 2020.

[2] World Health Organization, "Cardiovascular diseases (CVDs)," *WHO reports*, no. June, CRC Press, pp. 6–7, 11-Jun-2021.

[3] Kemenkes, "Penyakit jantung penyebab kematian terbanyak ke-2 di indonesia," *Kementrian Kesehat. Republik Indones.*, no. September, pp. 1–2, 2019.

[4] V. Krishnaiah, G. Narsimha, and N. S. Chandra, "Heart Disease Prediction System Using Data Mining Technique by Fuzzy K-NN Approach," vol. 1, no. V, pp. 371–384, 2015.

[5] T. P. R and J. Thomas, "Human Heart Disease Prediction System using Data Mining Techniques," *Proc. IEEE Int. Conf. Circuit, Power Comput. Technol. ICCPCT 2017*, 2017.

[6] I. Mirza, A. Mahapatra, D. Rego, and K. Mascarenhas, "Human Heart Disease Prediction Using Data Mining Techniques," in *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*, 2019, pp. 1–5.

[7] I. K. A. Enriko, M. Suryanegara, and D. Gunawan, "Heart disease prediction system using k-Nearest neighbor algorithm with simplified patient's health parameters," *J. Telecommun. Electron. Comput. Eng.*, vol. 8, no. 12, pp. 59–65, 2016.

[8] A. Singh and R. Kumar, "Heart Disease Prediction using Machine Learning Algorithms," in *2020 International Conference for Emerging Technology (INCET)*, 2020, pp. 1–7.

[9] C. Raju, E. Philipsy, S. Chacko, L. P. Suresh, and D. R. S, "A Survey on Predicting Heart Disease using Data Mining Techniques," no. March, pp. 2–3, 2018.

[10] S. Nayak, M. K. Gourisaria, M. Pandey, and S. S. Rautaray, "Prediction of heart disease by mining frequent items and classification techniques," *2019 Int. Conf. Intell. Comput. Control Syst. ICCS 2019*, no. Iciccs, pp. 607–611, 2019.

[11] S. Liu, P. Zhu, and S. Qin, "An improved weighted knn algorithm for imbalanced data classification," *2018 IEEE 4th Int. Conf. Comput. Commun. ICCC 2018*, pp. 1814–1819, 2018.

[12] M. Dialameh and M. Z. Jahromi, "Proposing a General Feature Weighting Function," *Expert Syst. Appl.*, 2016.

[13] N. Khateeb and M. Usman, "Efficient heart disease prediction system using K-nearest neighbor classification technique," *ACM Int. Conf. Proceeding Ser.*, pp. 21–26, 2017.

[14] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," *SN Comput. Sci.*, vol. 1, no. 6, pp. 1–6, 2020.

[15] K. Siminski, "Fuzzy weighted C-ordered means clustering algorithm," *Fuzzy Sets Syst.*, vol. 1, pp. 1–33, 2017.

[16] J. Wu, S. Pan, X. Zhu, Z. Cai, P. Zhang, and C. Zhang, "Self-adaptive attribute weighting for Naive Bayes classification," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1487–1502, 2015.

[17] V. Kotu and B. Deshpande, *Predictive Analytics using Rapidminer*. Elsevier, 2015.

[18] I. Witten, E. Frank, and M. Hall, *Data Mining Practical Machine Learning Tools and Techniques Third Edition*, vol. 277. Elsevier Inc., 2011.

[19] A. B. Hassanat, M. A. Abbadi, and A. A. Alhasanat, "Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach," *Int. J. Comput. Sci. Inf. Secur.*, vol. 12, no. 8, pp. 33–39, 2014.