Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.2023106510
Vol. 10, No. 3, Juni 2023, hlm. 503-512 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

OPTIMASI ALGORITMA *NAÏVE BAYES* DENGAN DISKRITISASI *K-MEANS* PADA DIAGNOSIS PENYAKIT JANTUNG

Nafa Fajriati[*1], Budi Prasetiyo[2]

[1,2]Universitas Negeri Semarang, Semarang
Email: [1]nafafajriati@students.unnes.ac.id, [2]bprasetiyo@mail.unnes.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 01 September 2022, diterima untuk diterbitkan: 19 Juni 2023)

Abstrak

Penyakit jantung iskemik adalah salah satu jenis penyakit kardiovaskular dengan jumlah penderita yang besar dan menjadi penyebab utama kematian di dunia. Disamping itu, penyakit jantung juga menyebabkan kerugian ekonomi. Diagnosis penyakit jantung pada tahap awal dapat membantu mengurangi risiko kematian dan tingginya biaya perawatan akibat penyakit jantung. Diagnosis penyakit merupakan proses penting yang harus dilakukan secara akurat agar tidak terjadi kesalahan diagnosis. *Data mining* dapat diterapkan untuk meningkatkan akurasi dan mengurangi jumlah kesalahan diagnosis. Salah satu teknik *data mining* adalah klasifikasi. *Naïve Bayes* merupakan algoritma klasifikasi yang memiliki kemampuan yang cukup baik untuk membangun model pengklasifikasi. Pada penelitian ini, dilakukan klasifikasi penyakit jantung menggunakan algoritma *Naïve Bayes*. *Dataset* yang digunakan yaitu *Cleveland heart disease dataset* dari *UCI Machine Learning Repository*. Untuk meningkatkan akurasi klasifikasi menggunakan algoritma *Naive Bayes*, atribut kontinu pada *dataset* diubah menjadi atribut diskrit dengan diskritisasi *K-means*. Diskritisasi *K-means* mengubah nilai setiap atribut kontinu menjadi kategori-kategori diskrit berupa *cluster* sejumlah k yang terbentuk dari proses algoritma *K-means*. Hal tersebut dilakukan karena algoritma *Naïve Bayes* menunjukkan kemampuan klasifikasi yang lebih baik apabila menggunakan data masukan berupa diskrit dibanding kontinu. Hasil akurasi yang diperoleh dari algoritma *Naïve Bayes* tanpa menerapkan diskritisasi *K-means* pada *Cleveland heart disease dataset* adalah 86,89%, sedangkan hasil akurasi yang diperoleh dari algoritma *Naïve Bayes* dengan menerapkan diskritisasi *K-means* pada *Cleveland heart disease dataset* adalah 88,52%. Berdasarkan perbandingan akurasi yang dihasilkan, dapat diketahui adanya peningkatan akurasi sebesar 1,63%. Hal tersebut menunjukkan bahwa diskritisasi *K-means* berperan dalam mengoptimalkan kinerja algoritma *Naïve Bayes* sehingga menghasilkan akurasi yang lebih baik.

Kata kunci: *penyakit jantung, klasifikasi, naïve bayes, diskritisasi, k-means.*

OPTIMIZATION OF NAÏVE BAYES ALGORITHM USING K-MEANS DISCRETIZATION IN HEART DISEASE DIAGNOSIS

Abstract

Ischemic heart disease is a type of cardiovascular disease with a large number of sufferers and is the leading cause of death in the world. In addition, heart disease also causes economic losses. Diagnosing heart disease early can help reduce the risk of death and the high costs of treatment for heart disease. Diagnosis of the disease is an important process that must be carried out accurately to avoid misdiagnosis. Data mining can be applied to improve accuracy and reduce the number of misdiagnoses. One of the data mining techniques is classification. Naïve Bayes is a classification algorithm that has a fairly good ability to build a classifier model. In this study, heart disease was classified using the Naïve Bayes algorithm. The dataset used is the Cleveland heart disease dataset from the UCI Machine Learning Repository. To improve classification accuracy using the Naive Bayes algorithm, continuous attributes in the dataset are changed to discrete attributes using K-means discretization. K-means discretization changes the value of each continuous attribute into discrete categories in the form of k clusters formed from the K-means algorithm process. This is done because the Naïve Bayes algorithm shows a better classification ability when it uses discrete rather than continuous input data. The accuracy results obtained from the Naïve Bayes algorithm without applying the K-means discretization to the Cleveland heart disease dataset are 86.89%, while the accuracy results obtained from the Nave Bayes algorithm by applying the K-means discretization to the Cleveland heart disease dataset are 88.52%. . Based on the comparison of the resulting accuracy, it can be seen that there is an increase in accuracy of 1.63%. This shows

that K-means discretization plays a role in optimizing the performance of the Naïve Bayes algorithm to produce better accuracy.

Keywords: *heart disease, classification, naïve bayes, discretization, k-means.*

1. PENDAHULUAN

Penyakit jantung iskemik adalah salah satu jenis penyakit kardiovaskular dengan jumlah penderita yang besar dan penyebab utama kematian di dunia selain *stroke*. Pada tahun 2019, jumlah kasus penyakit jantung iskemik di dunia mencapai 197 juta kasus dengan jumlah kematian sebanyak 9,74 juta kematian (Roth *et al*., 2020). Di Indonesia sendiri berdasarkan data *Institute for Health Metrics and Evaluation* (IHME) ada 245.343 kematian per tahun akibat penyakit jantung iskemik (*GBD* 2019 *Diseases and Injuries Collaborators*, 2020). Selain menyebabkan kematian, penyakit jantung juga menyebabkan kerugian ekonomi. Data Badan Penyelenggaraan Jaminan Sosial (BPJS) menunjukkan bahwa hingga akhir 2020 penyakit jantung menjadi penyakit yang membutuhkan biaya tertinggi dalam pelayanan Jaminan Kesehatan Nasional (JKN), yaitu hampir sebesar 8,3 triliun rupiah (Kementerian Kesehatan RI, 2021). Diagnosis penyakit jantung pada tahap awal dapat membantu mengurangi risiko kematian dan tingginya biaya perawatan (Reddy *et al*., 2020). Diagnosis penyakit merupakan proses penting yang harus dilakukan secara akurat agar tidak terjadi kesalahan diagnosis. *Data mining* dapat diterapkan untuk meningkatkan akurasi dan mengurangi jumlah kesalahan diagnosis *(*Abdar *et al.*, 2019).

Data mining adalah proses analisis data untuk mendapatkan informasi penting yang dapat berguna dalam pengambilan keputusan (Rino, 2021). Ada berbagai teknik dalam *data mining*. Masing-masing teknik memiliki aturan dan caranya tersendiri yang menentukan jenis permasalahan yang diselesaikan. Klasifikasi merupakan salah satu teknik *data mining* dengan cara memprediksi nilai suatu atribut berdasarkan nilai atribut-atribut lain (Arhami dan Nasir, 2020). Klasifikasi sering digunakan untuk menyelesaikan permasalahan di bidang kesehatan salah satunya yaitu mendiagnosis penyakit (Pandey, 2016). Untuk melakukan klasifikasi perlu dibangun model pengklasifikasi menggunakan algoritma klasifikasi dengan menganalisis data latih yang telah memiliki kelas (Han, Kamber dan Pei, 2012). Jiang *et al*. (2019) dalam penelitiannya mengatakan bahwa algoritma *Naïve Bayes* memiliki kemampuan cukup baik untuk membangun model pengklasifikasi.

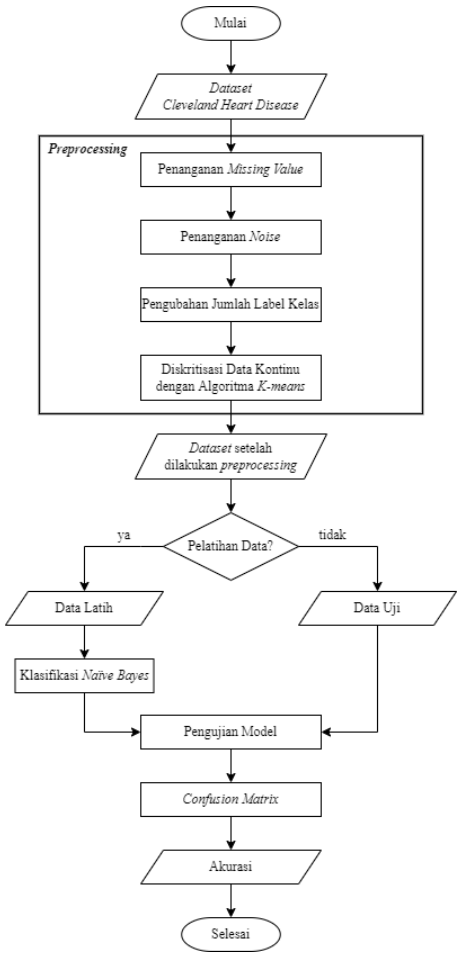
Pada klasifikasi, akurasi menunjukkan seberapa akurat model pengklasifikasi yang dibuat. Semakin tinggi tingkat akurasi maka semakin baik model pengklasifikasi dalam mengklasifikasi kelas data yang belum diketahui. Oleh karena itu, menghasilkan akurasi yang tinggi merupakan sesuatu yang penting. Salah satu yang mempengaruhi akurasi adalah diskritisasi atribut yang bernilai kontinu (Tang *et al*., 2020). Diskritisasi dilakukan pada saat *preprocessing* data yang penting dilakukan karena data masukan berpengaruh pada hasil klasifikasi, seperti algoritma *Naïve Bayes* yang menunjukkan kemampuan klasifikasi yang lebih baik apabila menggunakan data masukan berupa data diskrit dibanding data kontinu (Zhou *et al*., 2021). Pada algoritma *Naïve Bayes* yang berdasar pada pendekatan probabilistik, atribut kontinu memiliki jumlah kemungkinan nilai yang tidak terbatas sehingga tidak mungkin untuk menentukan probabilitas bersyarat secara jelas untuk setiap nilai (Russell dan Norvig, 2010). Dengan diskritisasi maka atribut kontinu diubah menjadi diskrit sehingga nilainya terbatas pada rentang tertentu (Rosenfeld *et al*., 2018). Salah satu penelitian yang menerapkan diskritisasi untuk mengoptimasi algoritma *Naïve Bayes* adalah penelitian oleh Saleh dan Nasari (2018). Penelitian tersebut menggunakan metode *equal width interval discretization* dan membuktikan bahwa penerapan diskritisasi dapat mengoptimalkan akurasi algoritma *Naïve Bayes* yang sebelumnya 90% mengalami peningkatan menjadi 92,8%.

Ada beberapa metode untuk diskritisasi, dari yang sederhana seperti *equal width interval discretization* dan *equal frequency interval discretization* hingga metode yang lebih kompleks seperti *clustering* dengan *K-means* (Saleh dan Nasari, 2018). Ketiga metode tersebut jika dibandingkan, diskritisasi dengan metode *clustering K-means* adalah yang terbaik dalam menangani nilai batas (Setyawan dan Fathicah, 2018). Selain itu, *K-means* adalah algoritma yang sederhana, efisien, dan memiliki kinerja yang stabil pada berbagai masalah (Zhao, Deng dan Ngo, 2018).

Berdasarkan uraian di atas, penelitian ini akan berfokus pada penerapan algoritma *K-means* untuk menangani atribut kontinu pada *Cleveland heart disease dataset* guna meningkatkan akurasi algoritma *Naïve Bayes* dalam mendiagnosis penyakit jantung. *Dataset* tersebut memiliki sebanyak 13 atribut yang terdiri dari atribut yang bernilai nominal dan kontinu. Atribut kontinu adalah atribut yang memiliki jumlah kemungkinan nilai yang tidak terbatas. Atribut kontinu mempengaruhi proses *learning* menggunakan algoritma *Naïve Bayes*, yaitu ketika perhitungan probabilitas bersyarat untuk setiap nilai yang tidak dapat dihitung secara jelas. Oleh karena itu, digunakan diskritisasi dengan algoritma *K-means* untuk mengatasi atribut kontinu.

2. METODE PENELITIAN

Pada penelitian ini, kombinasi algoritma *K-means* dan algoritma *Naïve Bayes* digunakan untuk diagnosis penyakit jantung. Algoritma *K-means* diterapkan untuk diskritisasi atribut kontinu pada *dataset*. Kemudian algoritma *Naïve Bayes* digunakan untuk klasifikasi *dataset* yang telah melalui proses diskritisasi. *Flowchart* metode yang digunakan pada penelitian ini ditunjukkan pada Gambar 1.



Gambar 1. *Flowchart* Klasifikasi Menggunakan algoritma *Naïve Bayes* dengan Menerapkan Diskritisasi *K-means*.

2.1 Pengambilan Data

Penelitian ini menggunakan *Cleveland heart disease dataset* yang didapatkan dari *UCI Machine Learning Repository*. Terdapat 13 atribut dan 1 label dengan 303 sampel data. Tabel 1 menunjukkan deskripsi dari *Cleveland heart disease dataset*.

Tabel 1. Deskripsi dari *Cleveland heart disease dataset*.

No	Atribut	Deskripsi	Tipe
1	*age*	Usia	Numerik
2	*sex*	Jenis kelamin	Nominal
3	*cp*	Nyeri dada	Nominal
4	*trestbps*	Tekanan darah	Numerik
5	*chol*	Kolesterol serum	Numerik
6	*fbs*	Gula darah	Nominal
7	*restecg*	Hasil elektrokardiografi	Nominal
8	*thalach*	Detak jantung maksimum	Numerik
9	*exang*	Angina yang disebabkan olahraga	Nominal
10	*oldpeak*	ST depresi yang disebabkan oleh olahraga dibandingkan dengan keadaan istirahat	Numerik
11	*slope*	Kemiringan ST segmen selama puncak latihan	Nominal
12	*ca*	Jumlah pembuluh darah besar yang diwarnai dengan fluoroskopi	Nominal
13	*thal*	Status jantung	Nominal
14	*num*	Diagnosis penyakit jantung	Nominal

2.2 Penanganan *Missing Value*

Cleveland heart disease dataset yang digunakan pada penelitian ini mengandung beberapa *missing value*. *Missing value* yang ada pada *dataset* berpotensi mengandung informasi penting yang tidak dapat diabaikan (Lin dan Tsai, 2020). Pada penelitian ini, *missing value* diatasi dengan mengisi *missing value* (nilai yang hilang) dengan nilai yang paling banyak muncul (*most frequent*) dalam suatu atribut.

2.3 Penanganan Data *Noise*

Setelah penanganan *missing value*, dilakukan penanganan *noise* pada data. *Noise* pada data seperti *outliers* dapat secara signifikan mempengaruhi hasil *clustering* dengan algoritma *K-means* sehingga harus ditangani (Zhang *et al*., 2021). Pada penelitian ini, *noise* diatasi dengan cara mengidentifikasi *outliers* pada data menggunakan metode *Interquartile Range* (IQR) kemudian nilai *outliers* tersebut diganti dengan median.

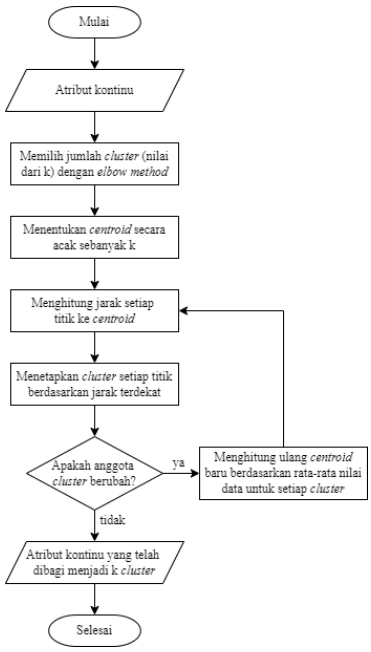
2.4 Pengubahan Jumlah Label Kelas

Setelah dilakukan penanganan *missing value* dan penanganan *noise* pada data, langkah selanjutnya adalah pengubahan jumlah label kelas yang berjumlah lima menjadi dua kelas. Label kelas pada *Cleveland heart disease dataset* terdiri dari nilai 0, 1, 2, 3, dan 4. Nilai 0 menyatakan tidak adanya penyakit jantung dan nilai 1, 2, 3, dan 4 menyatakan adanya penyakit jantung dengan tingkat keparahan tertentu (4 menjadi yang paling tinggi) (Amin, Chiam dan Varathan, 2019). Penelitian-penelitian yang dilakukan dengan *Cleveland heart disease dataset* telah berfokus pada membedakan adanya penyakit jantung (nilai 1, 2, 3, 4) dengan tidak adanya penyakit jantung (nilai 0) (Janosi *et al*., 1988). Pada penelitian ini akan dilakukan prediksi ada atau tidaknya penyakit jantung pada seseorang sehingga nilai 1,2,3, dan 4 dikelompokkan menjadi satu kategori yaitu 1 karena semua menyatakan adanya penyakit jantung.

2.5 Diskritisasi Data dengan Algoritma *K-means*

Atribut yang memiliki sifat kontinu (bertipe numerik) antara lain *age*, *trestbps*, *chol*, *thalach*, dan *oldpeak* dilakukan diskritisasi untuk mengubahnya menjadi atribut bertipe nominal. Metode yang digunakan adalah dengan diskritisasi *K-means*. Diskritisasi *K-means* dilakukan pada masing-masing

atribut dengan membagi data ke dalam beberapa *cluster* menggunakan algoritma *K-means*. Tahapan algoritma *K-means* dapat dilihat pada Gambar 2.



Gambar 2. *Flowchart* Algoritma *K-means*.

Adapun penjelasan tahapan algoritma *K-means* berdasarkan *flowchart* pada Gambar 2 adalah sebagai berikut:

1. Memilih jumlah *cluster* (nilai dari k) dengan *elbow method* dengan cara membandingkan nilai *Sum of Square Error* (SSE) dari masing-masing nilai *cluster*. Dimulai dari k = 2 kemudian terus meningkatkan nilai k hingga k = 10. Jumlah *cluster* terbaik dipilih dari nilai k yang SSEnya mengalami penurunan drastis sebelum mencapai k. Adapun SSE dihitung menggunakan Persamaan 1.

   $$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2 \quad (1)$$

   Keterangan:
   k : jumlah cluster
   C_i : cluster ke-i
   c_i : centroid cluster ke-i
   $dist(p, c_i)$: jarak antara tiap titik p yang ada pada C_i dengan c_i
2. Menentukan *centroid* secara acak sebanyak k.
3. Menghitung jarak setiap titik ke *centroid* menggunakan fungsi *euclidean distance* pada Persamaan 2.

   $$d_{ik} = \sqrt{\sum_{j=1}^{m} (x_{ij} - c_{kj})^2} \quad (2)$$

   Keterangan:
   d_{ik} : *euclidean distance* antara objek i dan k
   m : jumlah atribut
   x_{ij} : nilai data i pada atribut ke-j
   c_{kj} : *centroid* k pada atribut ke-j
4. Menetapkan *cluster* setiap titik berdasarkan jarak terdekat dengan *centroid*.

   Jika anggota *cluster* tidak berubah, *cluster* tersebut adalah hasil akhir *clustering* dan proses selesai. Namun, jika anggota *cluster* berubah, dilakukan langkah selanjutnya yaitu menghitung ulang *centroid* baru kemudian mengulangi langkah ke-3 dan ke-4.
5. Menghitung ulang *centroid* baru berdasarkan rata-rata nilai data untuk setiap *cluster* yang dibentuk menggunakan Persamaan 3.

   $$c_{kj} = \frac{\sum_{i=1}^{p} x_{ij}}{p}; x_{ij} \in cluster\ ke-k \quad (3)$$

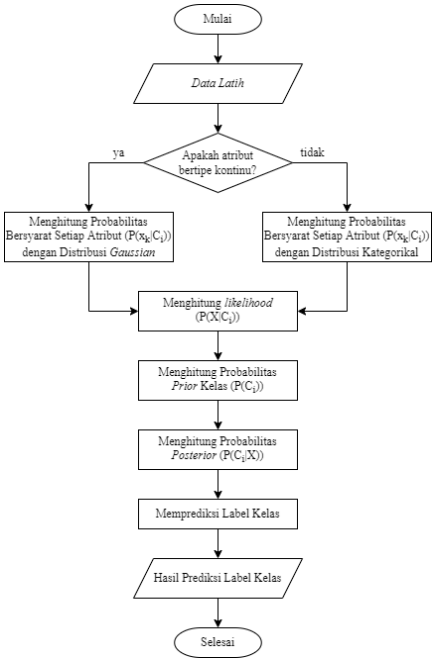
   Keterangan:
   c_{kj} : *centroid* baru
   x_{ij} : nilai data ke-i pada atribut j
   p : jumlah data pada *cluster* ke-k

2.6 Pembagian Data Latih dan Data Uji

Metode *stratified sampling* digunakan untuk membagi data latih dan data uji dengan proporsi data latih sebesar 80% dan data uji sebesar 20%. *Stratified sampling* memastikan label kelas dalam *dataset* memiliki representasi yang tepat pada data latih.

2.7 Tahapan *Naïve Bayes*

Tahapan algoritma *Naïve Bayes* pada data latih dalam melakukan klasifikasi dapat dilihat pada Gambar 3.



Gambar 3. *Flowchart* Algoritma *Naïve Bayes*.

Adapun penjelasan tahapan algoritma *Naïve Bayes* berdasarkan *flowchart* pada Gambar 3 adalah sebagai berikut:

1. Membaca data latih.
2. Mengidentifikasi tipe atribut termasuk atribut kategorikal atau atribut kontinu.
   a. Jika atribut kategorikal, probabilitas bersyarat $P(x_k|C_i)$ dihitung dengan distribusi kategorikal menggunakan Persamaan 4.

$$P(x_k|C_i) = \frac{jumlah\ x_k\ pada\ A_k\ dengan\ label\ kelas\ C_i}{jumlah\ label\ kelas\ C_i} \quad (4)$$

b. Jika atribut bernilai kontinu, atribut diasumsikan memiliki distribusi *Gaussian* dan probabilitas bersyarat $P(x_k|C_i)$ dihitung dengan distribusi *Gaussian* menggunakan Persamaan 5.

$$P(A_k = x_k|Y = C_i) = \frac{1}{\sqrt{2\pi}\sigma_{C_i}} e^{-\frac{(x_k-\mu_{C_i})^2}{2\sigma_{C_i}^2}} \quad (5)$$

Keterangan:
A_k : atribut ke-k
x_k : nilai dari atribut A_k
Y : kelas yang dicari
C_i : label kelas ke-i
μ_{C_i}: rata-rata hitung (*mean*) atribut A_k dengan label kelas C_i
σ_{C_i}: standar deviasi atribut A_k dengan label kelas C_i
Adapun persamaan untuk menghitung rata-rata hitung (*mean*) dapat dilihat pada Persamaan 6.

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \quad (6)$$

Keterangan:
μ : rata-rata hitung (*mean*)
x_i : nilai sampel ke-i
n : jumlah sampel
Adapun persamaan untuk menghitung standar deviasi dapat dilihat pada Persamaan 7.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1}} \quad (7)$$

Keterangan:
σ : standar deviasi
x_i : nilai sampel ke-i
μ : rata-rata hitung (*mean*)
n : jumlah sampel

3. Menghitung *likelihood* $P(X|C_i)$ dengan asumsi *naïve* ketidakbergantungan antara atribut menggunakan Persamaan 8.

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \quad (8)$$
$$P(X|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)$$

Keterangan:
X : kriteria suatu kasus berdasarkan masukan
C_i : label kelas ke-i, i adalah jumlah label kelas
x_k : nilai dari atribut A_k untuk tupel X
$P(X|C_i)$: probabilitas kriteria masukan X dengan label kelas C_i
$P(x_k|C_i)$: probabilitas nilai x_k dengan label kelas C_i

4. Menghitung probabilitas *prior* kelas menggunakan Persamaan 9.

$$P(C_i) = \frac{jumlah\ label\ kelas\ C_i}{jumlah\ sampel} \quad (9)$$

Keterangan:
C_i : label kelas ke-i, i adalah jumlah label kelas
$P(C_i)$: probabilitas label kelas C_i

5. Menghitung probabilitas *posterior* $P(C_i|X)$ menggunakan Persamaan 10. (6)

$$P(C_i|X) = P(X|C_i)P(C_i) \quad (10)$$

Keterangan:
$P(C_i|X)$: probabilitas kemunculan label kelas C_i dengan kriteria masukan X

6. Memprediksi label kelas untuk X berdasarkan probabilitas *posterior* $P(C_i|X)$. X termasuk ke dalam label kelas C_i jika dan hanya jika (7) $P(C_i|X) > P(C_j|X)$ untuk $1 \leq j \leq m, j \neq 0$ di mana m adalah jumlah label kelas.

2.8 Tahap Evaluasi

Pada penelitian ini, evaluasi dilakukan dengan perhitungan akurasi menggunakan *confusion matrix*. Akurasi merupakan persentase data uji yang diklasifikasikan dengan benar oleh pengklasifikasi. Adapun langkah perhitungan akurasi adalah sebagai berikut.

1. Masukkan hasil pengujian klasifikasi pada tabel *confusion matrix* seperti pada Tabel 2.

Tabel 2. Hasil Pengujian Klasifikasi pada Tabel Confusion Matrix.

Klasifikasi		**Kelas prediksi**		
		Ya	Tidak	Jumlah
Kelas aktual	Ya	TP	FN	P
	Tidak	FP	TN	N

Keterangan:
True Positive (TP): jumlah data positif yang diklasifikasi dengan benar
True Negative (TN): jumlah data negatif yang diklasifikasi dengan benar
False Positive (FP): jumlah data negatif yang diklasifikasi sebagai data positif
False Negative (FN): jumlah data positif yang diklasifikasi sebagai data negatif
Positive (P): jumlah data positif
Negative (N): jumlah data negatif

2. Hitung akurasi menggunakan Persamaan 11.

$$Akurasi = \frac{TP+TN}{P+N} \times 100\% \quad (11)$$

3. Simpulkan hasil akurasi yang diperoleh.

3. HASIL DAN PEMBAHASAN

Diskritisasi *K-means* diterapkan untuk menangani atribut yang bertipe kontinu pada *dataset*. Terdapat 8 atribut kontinu yang harus melalui proses diskritisasi *K-means* yaitu *age*, *trestbps*, *chol*, *thalach*, dan *oldpeak*. Masing-masing atribut kontinu didiskritisasi dengan cara membagi data ke dalam beberapa *cluster* menggunakan algoritma *K-means* pada *flowchart* Gambar 2. Pemilihan jumlah *cluster* (k) dilakukan dengan metode *elbow* berdasarkan nilai SSE. Dari metode *elbow*, didapatkan bahwa masing-masing atribut kontinu akan didiskritisasi menggunakan algoritma *K-means* menjadi 3 *cluster*.

Proses pengolahan diskritisasi *K-means* disajikan di bawah ini untuk atribut *age* sebagai sampel data. Pertama, jumlah *cluster* dipilih dengan metode *elbow* berdasarkan selisih SSE terbesar antara jumlah *cluster* k-1 dan jumlah *cluster* k. Tabel hasil perhitungan metode *elbow* untuk atribut *age* dapat dilihat pada Tabel 3.

Tabel 3. Perhitungan Metode *Elbow* Atribut *Age*.

Jumlah *Cluster*	Nilai SSE	Selisih
2	7766,85	-
3	3637,13	4129,72
4	2228,77	1408,36
5	1596,13	632,64
6	1127,29	468,84
7	861,65	265,64
8	667,99	193,66
9	554,48	113,51
10	443,46	111,02

Berdasarkan Tabel 3, diperoleh selisih nilai SSE terbesar adalah antara jumlah *cluster* 2 dan jumlah *cluster* 3 yaitu 4129,72, sehingga jumlah *cluster* terbaik yang dipilih untuk nantinya digunakan dalam proses *clustering* atribut *age* menggunakan algoritma *K-means* adalah 3.

Kedua, menentukan *centroid* secara acak sebanyak k. Sebagai sampel untuk contoh perhitungan manual algoritma *K-means*, digunakan atribut *age*. Data yang digunakan berjumlah 15. Sampel data dapat dilihat pada Tabel 4.

Tabel 4. Sampel Data untuk Diskritisasi.

Data ke-	*age*
1	63
2	67
3	67
4	37
5	41
6	56
7	62
8	57
9	63
10	53
11	57
12	56
13	56
14	44
15	52

Centroid dipilih secara acak sebanyak tiga *centroid* sesuai dengan jumlah *cluster* (k) untuk atribut *age* yang sudah ditentukan menggunakan metode *elbow*. *Centroid* yang dipilih dari sampel data untuk contoh perhitungan manual ini yaitu:

- Data ke-5 = $c_1 = 41$
- Data ke-6 = $c_2 = 56$
- Data ke-9 = $c_3 = 63$

Ketiga, menghitung jarak setiap titik ke *centroid* menggunakan fungsi *euclidean distance* pada Persamaan 2.

Sebagai contoh akan dihitung jarak titik 53 dengan setiap *centroid*.

$$d_{ik} = \sqrt{\sum_{j=1}^{m}\left(x_{ij} - c_{kj}\right)^2}$$

$$d_{ik} = \sqrt{(53-41)^2}$$
$$d_{ik} = \sqrt{(12)^2}$$
$$d_{ik} = 12$$

Jarak titik 53 dengan c_1 adalah 12.

$$d_{ik} = \sqrt{\sum_{j=1}^{m}\left(x_{ij} - c_{kj}\right)^2}$$

$$d_{ik} = \sqrt{(53-56)^2}$$
$$d_{ik} = \sqrt{(-3)^2}$$
$$d_{ik} = 3$$

Jarak titik 53 dengan c_2 adalah 3.

$$d_{ik} = \sqrt{\sum_{j=1}^{m}\left(x_{ij} - c_{kj}\right)^2}$$

$$d_{ik} = \sqrt{(53-63)^2}$$
$$d_{ik} = \sqrt{(-10)^2}$$
$$d_{ik} = 10$$

Jarak titik 53 dengan c_3 adalah 10.

Adapun semua hasil perhitungan jarak setiap titik ke *centroid* dapat dilihat pada Tabel 5.

Tabel 5. Hasil Perhitungan Jarak Setiap Titik Sampel Data ke Centroid.

Data ke-	*age*	Jarak dengan			Jarak terdekat
		c_1	c_2	c_3	
1	63	22	7	0	0
2	67	26	11	4	4
3	67	26	11	4	4
4	37	4	19	26	4
5	41	0	15	22	0
6	56	15	0	7	0
7	62	21	6	1	1
8	57	16	1	6	1
9	63	22	7	0	0
10	53	12	3	10	3
11	57	16	1	6	1
12	56	15	0	7	0
13	56	15	0	7	0
14	44	3	12	19	3
15	52	11	4	11	4

Keempat, menetapkan *cluster* setiap titik berdasarkan jarak terdekat dengan *centroid*. Hasil pengelompokan *cluster* dapat dilihat pada Tabel 6.

Tabel 6. Hasil Pengelompokan Cluster pada Sampel Data.

Data ke-	*age*	Termasuk *cluster*
1	63	3
2	67	3
3	67	3
4	37	1
5	41	1
6	56	2
7	62	3
8	57	2
9	63	3
10	53	2
11	57	2
12	56	2
13	56	2
14	44	1
15	52	2

Pada tahap ini, jika anggota *cluster* tidak berubah dari iterasi sebelumnya, *cluster* tersebut adalah hasil akhir *clustering*, Namun, jika anggota *cluster* berubah, dilakukan langkah selanjutnya yaitu

menghitung ulang *centroid* baru kemudian mengulangi langkah ke-3 dan ke-4.

Kelima, menghitung ulang *centroid* baru berdasarkan rata-rata nilai data untuk setiap *cluster* yang dibentuk menggunakan Persamaan 3.

Perhitungan c_1 baru adalah sebagai berikut:

$$c_{kj} = \frac{\sum_{i=1}^{p} x_{ij}}{p}; x_{ij} \in cluster\ ke-k$$

$$c_1 = \frac{\sum_{i=1}^{3} x_{ij}}{3}$$

$$c_1 = \frac{37+41+44}{3}$$

$$c_1 = \frac{122}{3}$$

$$c_1 = 40{,}67$$

Jadi c_1 baru yaitu 40,67.

Perhitungan c_2 baru adalah sebagai berikut:

$$c_{kj} = \frac{\sum_{i=1}^{p} x_{ij}}{p}; x_{ij} \in cluster\ ke-k$$

$$c_2 = \frac{\sum_{i=1}^{7} x_{ij}}{7}$$

$$c_2 = \frac{56+57+53+57+56+56+52}{7}$$

$$c_2 = \frac{387}{7}$$

$$c_2 = 55{,}28$$

Jadi c_2 baru yaitu 55,28.

Perhitungan c_3 baru adalah sebagai berikut:

$$c_{kj} = \frac{\sum_{i=1}^{p} x_{ij}}{p}; x_{ij} \in cluster\ ke-k$$

$$c_3 = \frac{\sum_{i=1}^{5} x_{ij}}{5}$$

$$c_3 = \frac{63+67+67+62+63}{5}$$

$$c_3 = \frac{322}{5}$$

$$c_3 = 64{,}4$$

Jadi c_3 baru yaitu 64,4.

Untuk hasil *clustering* secara lengkap pada atribut *age*, *trestbps*, *chol*, *thalach*, dan *oldpeak* dapat dilihat pada Tabel 7.

Tabel 7. Hasil Clustering pada Atribut *Age*, *Trestbps*, *Chol*, *Thalach*, dan *Oldpeak*.

Atribut	*Cluster*	Nilai	*Centroid*
age	1	29, 34, 35, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49	42,93
	2	50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60	55,46
	3	61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 74, 76, 77	65,32
trestbps	1	94, 100, 101, 102, 104, 105, 106, 108, 110, 112, 114, 115, 117, 118, 120, 122, 123	113.89
	2	124, 125, 126, 128, 129, 130, 132, 134, 135, 136, 138, 140, 142	132.77
	3	144, 145, 146, 148, 150, 152, 154, 155, 156, 158, 160, 164, 165, 170	153.75
chol	1	126, 131, 141, 149, 157, 160, 164, 166, 167, 168, 169, 172, 174, 175, 176, 177, 178, 180, 182, 183, 184, 185, 186, 187, 188, 192, 193, 195, 196, 197, 198, 199, 200, 201, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221	195.32
	2	222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 273, 274, 275, 276	247.19
	3	277, 278, 281, 282, 283, 284, 286, 288, 289, 290, 293, 294, 295, 298, 299, 300, 302, 303, 304, 305, 306, 307, 308, 309, 311, 313, 315, 318, 319, 321, 322, 325, 326, 327, 330, 335, 340, 341, 342, 353, 354, 360	305.86
thalach	1	88, 90, 95, 96, 97, 99, 103, 105, 106, 108, 109, 111, 112, 113, 114, 115, 116, 117, 118, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133	118.42
	2	134, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160	149.29
	3	161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 177, 178, 179, 180, 181, 182, 184, 185, 186, 187, 188, 190, 192, 194, 195, 202	171.94
oldpeak	1	0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7	0.12
	2	0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.8, 1.9, 2.0, 2.1	1.33
	3	2.2, 2.3, 2.4, 2.5, 2.6, 2.8, 2.9, 3.0, 3.1, 3.2, 3.4, 3.5, 3.6, 3.8, 4.0	2.93

Tahap diskritisasi ini mengubah nilai setiap atribut kontinu menjadi kategori-kategori berupa *cluster* sejumlah k yang terbentuk dari proses algoritma *K-means*. Hasil *preprocessing* data dengan menerapkan diskritisasi *K-means* ditunjukkan pada Tabel 8.

Tabel 8. Hasil *Preprocessing* Data dengan Menerapkan Diskritisasi *K-Means*.

age	…	trestbps	chol	…	thalach	exang	oldpeak	…	thal	num
3	…	3	2	…	2	0	3	…	6	0
3	…	3	3	…	1	1	2	…	3	2
3	…	1	2	…	1	1	3	…	7	1
1	…	2	2	…	3	0	3	…	3	0
1	…	2	1	…	3	0	2	…	3	0
…	…	…	…	…	…	…	…	…	…	…
1	…	1	2	…	1	0	2	…	7	1
3	…	3	1	…	2	0	3	…	7	2
2	…	2	1	…	1	1	2	…	7	3
2	…	2	2	…	3	0	1	…	3	1
1	…	2	1	…	3	0	1	…	3	0

Cleveland heart disease dataset yang telah melalui proses diskritisasi *K-means* kemudian dibagi menjadi data latih dan data uji dengan proporsi data latih sebesar 80% dan data uji sebesar 20%. Dengan jumlah data sebanyak 303 data, maka jumlah data latih adalah 242 data sedangkan jumlah data uji adalah 61 data. Proses klasifikasi dimulai dengan menggunakan data latih sebagai data *learning*/pembelajaran menggunakan algoritma *Naïve Bayes* sehingga terbentuk model pengklasifikasi. Pada penerapan algoritma *Naïve Bayes* tersebut, seluruh atribut diperlakukan dengan

sama pada setiap tahap algoritma *Naïve Bayes*, utamanya pada perhitungan probabilitas bersyarat yang menggunakan distribusi kategorikal. Setelah model pengklasifikasi dibuat, selanjutnya dilakukan pengujian terhadap model pengklasifikasi menggunakan data uji. Evaluasi kinerja algoritma *Naïve Bayes* dihitung menggunakan *confusion matrix*. Hasil klasifikasi yang dihitung menggunakan *confusion matrix* ditunjukkan pada Tabel 9.

Tabel 9. *Confusion Matrix* Hasil Klasifikasi Menggunakan Algoritma *Naive Bayes* dengan Menerapkan *Diskritisasi K-Means*.

Klasifikasi		**Kelas prediksi**		
		0	1	Jumlah
Kelas aktual	0	29	4	33
	1	3	25	28

Untuk menghitung persentase akurasi yang dihasilkan, dapat menggunakan Persamaan 11.

$$Akurasi = \frac{29+25}{33+28} \times 100\%$$

$$Akurasi = 88{,}52\%$$

Hasil akurasi yang didapatkan dari penerapan klasifikasi menggunakan algoritma *Naïve Bayes* dengan menerapkan diskritisasi *K-means* akan dibandingkan dengan hasil akurasi klasifikasi menggunakan algoritma *Naïve Bayes* tanpa menerapkan diskritisasi *K-means* untuk mengetahui ada atau tidaknya peningkatan akurasi algoritma *Naïve Bayes* yang disebabkan oleh penerapan diskritisasi *K-means*. Tanpa menerapkan diskritisasi *K-means* maka atribut kontinu pada *dataset* tidak ditangani sehingga dataset hasil *preprocessing* data seperti ditunjukkan pada Tabel 10.

Tabel 10. Hasil *Preprocessing* Data Tanpa Menerapkan Diskritisasi *K-Means*.

age	…	trestbps	chol	…	thalach	exang	oldpeak	…	thal	num
63	…	145	233	…	150	0	2.3	…	6	0
67	…	160	286	…	108	1	1.5	…	3	2
67	…	120	229	…	129	1	2.6	…	7	1
37	…	130	250	…	187	0	3.5	…	3	0
41	…	130	204	…	172	0	1.4	…	3	0
…	…	…	…	…	…	…	…	…	…	…
45	…	110	264	…	132	0	1.2	…	7	1
68	…	144	193	…	141	0	3.4	…	7	2
57	…	130	131	…	115	1	1.2	…	7	3
57	…	130	236	…	174	0	0.0	…	3	1
38	…	138	175	…	173	0	0.0	…	3	0

Dengan pembagian data latih dan data uji yang juga dengan proporsi data latih sebesar 80% dan data uji sebesar 20%, proses klasifikasi *dataset* tanpa diskritisasi *K-means* dimulai dengan menggunakan data latih sebagai data *learning*/pembelajaran menggunakan algoritma *Naïve Bayes* sehingga terbentuk model pengklasifikasi. Pada penerapan algoritma *Naïve Bayes* tersebut, perlakuan terhadap atribut nominal dan atribut kontinu memiliki perbedaan. Perbedaan terletak pada perhitungan probabilitas bersyarat. Probabilitas bersyarat untuk atribut nominal dihitung dengan distribusi kategorikal, sedangkan probabilitas bersyarat untuk atribut kontinu dihitung dengan distribusi *Gaussian*. Setelah model pengklasifikasi dibuat, selanjutnya dilakukan pengujian terhadap model pengklasifikasi menggunakan data uji. Evaluasi kinerja algoritma *Naïve Bayes* dihitung menggunakan *confusion matrix*.

Hasil klasifikasi yang dihitung menggunakan *confusion matrix* ditunjukkan pada Tabel 11.

Tabel 11. *Confusion Matrix* Hasil Klasifikasi Menggunakan Algoritma *Naive Bayes* Tanpa Menerapkan Diskritisasi *K-Means*.

Klasifikasi		**Kelas prediksi**		
		0	1	Jumlah
Kelas aktual	0	28	5	33
	1	3	25	28

Untuk menghitung persentase akurasi yang dihasilkan, dapat menggunakan Persamaan 11.

$$Akurasi = \frac{28+25}{33+28} \times 100\%$$

$$Akurasi = 86{,}89\%$$

Dari penelitian ini, klasifikasi menggunakan algoritma *Naïve Bayes* tanpa menerapkan diskritisasi *K-means* memperoleh hasil akurasi sebesar 86,89% sedangkan klasifikasi menggunakan algoritma *Naïve Bayes* dengan menerapkan diskritisasi *K-means* memperoleh hasil akurasi sebesar 88,52%. Berdasarkan hasil tersebut, terjadi peningkatan akurasi klasifikasi menggunakan algoritma *Naïve Bayes* sebesar 1,63% dengan menerapkan diskritisasi *K-means*. Akurasi klasifikasi menggunakan algoritma *Naïve Bayes* yang data kontinu pada *dataset* tidak melalui proses diskritisasi menghasilkan akurasi lebih kecil karena probabilitas bersyarat untuk atribut kontinu dihitung dengan asumsi bahwa data mengikuti distribusi normal (*Gaussian*). Padahal pada kenyataannya, data bisa saja tidak mengikuti distribusi normal. Hal tersebut menyebabkan berkurangnya kinerja klasifikasi. Dengan diskritisasi, data dibagi kedalam beberapa kelompok diskrit tanpa memperhatikan distribusinya. Penerapan diskritisasi *K-means* membuat data menjadi tereduksi karena mengurangi data dari semula banyak variasi nilai kontinu ke pengelompokan nilai diskrit, sehingga data lebih mudah dipahami yang dapat meningkatkan akurasi algoritma klasifikasi.

Hasil penelitian ini dibandingkan dengan hasil penelitian sebelumnya yang menggunakan *dataset* yang sama yaitu *Cleveland heart disease dataset* untuk mengetahui lebih baik atau tidak metode yang digunakan dalam penelitian ini daripada metode yang sudah digunakan pada penelitian sebelumnya. Hasil perbandingan dapat dilihat pada Tabel 12.

Tabel 12. Perbandingan akurasi dengan penelitian sebelumnya.

Penulis	Diskritisasi	Algoritma Klasifikasi	Metode Evaluasi	Akurasi
Putri, Rahmawati dan Azhar (2020)	-	*Naïve Bayes*	*Confusion Matrix*	87%
Penulis	Diskritisasi *K-means*	*Naïve Bayes*	*Confusion Matrix*	88,52%

Berdasarkan perbandingan yang dilakukan diketahui bahwa akurasi yang dihasilkan dalam penelitian ini lebih baik daripada penelitian oleh Putri, Rahmawati dan Azhar (2020) pada *dataset* yang sama. Hal yang membedakan penelitian ini dengan penelitian sebelumnya tersebut adalah penerapan diskritisasi *K-means* untuk menangani atribut kontinu pada *dataset*.

Beberapa penelitian lain telah dilakukan terkait optimasi algoritma *Naïve Bayes* dengan menerapkan diskritisasi. Saleh dan Nasari (2018) pada penelitiannya menggunakan algoritma *Naïve Bayes* untuk klasifikasi jurusan siswa dan menerapkan metode *equal width interval discretization* untuk meningkatkan akurasi klasifikasi menggunakan algoritma *Naïve Bayes*. Hasil akurasi pada penelitian tersebut dibandingkan dengan akurasi penelitian sebelumnya yang menggunakan algoritma *Naïve Bayes* untuk klasifikasi jurusan siswa tanpa menerapkan diskritisasi. Hasilnya membuktikan bahwa penerapan diskritisasi dapat meningkatkan akurasi algoritma *Naïve Bayes* pada penelitian sebelumnya dengan akurasi 90% meningkat menjadi 92,8%. Penelitian lain oleh Nugroho, Prihandoyo dan Somantri (2022) menggunakan algoritma *Naïve Bayes* untuk klasifikasi program studi bagi calon mahasiswa baru dan menerapkan metode *equal width interval discretization* untuk meningkatkan akurasi klasifikasi menggunakan algoritma *Naïve Bayes*. Penelitian tersebut membandingkan akurasi yang diperoleh dengan akurasi pada penelitian sebelumnya yang menggunakan algoritma *Naïve Bayes* untuk klasifikasi program studi bagi calon mahasiswa baru tanpa menerapkan diskritisasi. Hasilnya membuktikan bahwa klasifikasi menggunakan algoritma *Naïve Bayes* yang menerapkan diskritisasi menghasilkan akurasi yang lebih tinggi dengan 97,66% dibanding tanpa menerapkan diskritisasi yang menghasilkan akurasi lebih rendah yaitu 96,68%.

4. KESIMPULAN

Berdasarkan hasil penelitian, diskritisasi *K-means* dapat menangani atribut kontinu pada *dataset* dan meningkatkan akurasi algoritma *Naïve Bayes*. Penerapan algoritma *K-means* sebagai metode diskritisasi menggunakan parameter k=3 yang dipilih dari metode *elbow*. Hasil akurasi yang diperoleh dari klasifikasi menggunakan algoritma *Naïve Bayes* dengan menerapkan diskritisasi *K-means* pada *Cleveland heart disease dataset* adalah 88,52%. Akurasi tersebut lebih besar dibandingkan akurasi dari klasifikasi menggunakan algoritma *Naïve Bayes* tanpa menerapkan diskritisasi *K-means* yang hanya memperoleh akurasi 86,89%. Berdasarkan perbandingan akurasi yang dihasilkan, dapat diketahui adanya peningkatan akurasi sebesar 1,63%. Hal tersebut menunjukkan bahwa diskritisasi *K-means* berperan dalam mengoptimalkan kinerja algoritma *Naïve Bayes* sehingga menghasilkan akurasi yang lebih baik.

DAFTAR PUSTAKA

ABDAR, M., KSIĄŻEK, W., ACHARYA, U.R., TAN, R.-S., MAKARENKOV, V. dan PŁAWIAK, P., 2019. A new machine learning technique for an accurate diagnosis of coronary artery disease. *Computer Methods dan Programs in Biomedicine*, 179. https://doi.org/10.1016/j.cmpb.2019.104992.

AMIN, M.S., CHIAM, Y.K. dan VARATHAN, K.D., 2019. Identification of significant features dan data mining techniques in predicting heart disease. *Telematics dan Informatics*, 36(Agustus 2018), pp.82–93. https://doi.org/10.1016/j.tele.2018.11.007.

ARHAMI, M. dan NASIR, M., 2020. *Data Mining - Algoritma dan Implementasi*. [ebook] Penerbit Andi.

GBD 2019 DISEASES AND INJURIES COLLABORATORS, 2020. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, 396(10258), pp.1204–1222. https://doi.org/10.1016/S0140-6736(20)30925-9.

HAN, J., KAMBER, M. dan PEI, J., 2012. *Data Mining: Concepts dan Techniques*. 3rd ed. Waltham: Elsevier. https://doi.org/10.1016/C2009-0-61819-5.

JANOSI, A., STEINBRUNN, W., PFISTERER, M. dan DETRANO, R., 1988. *UCI Machine Learning Repository: Heart Disease Data Set*. *Uci*, Tersedia melalui: <https://archive.ics.uci.edu/ml/datasets/heart+disease> [Diakses 25 Mei 2022].

JIANG, L., ZHANG, L., YU, L. dan WANG, D., 2019. Class-specific attribute weighted naive Bayes. *Pattern Recognition*, 88, pp.321–330. https://doi.org/10.1016/j.patcog.2018.11.032.

KEMENTERIAN KESEHATAN RI, 2021. *Profil Kesehatan Indonesia 2020*. Jakarta: Kementerian Kesehatan Republik Indonesia.

LIN, W.-C. dan TSAI, C.-F., 2020. Missing value imputation: a review dan analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53(2), pp.1487–1509. https://doi.org/10.1007/s10462-019-09709-4.

NUGROHO, W.E., PRIHANDOYO, T. dan SOMANTRI, O., 2022. Optimalisasi Metode Naive Bayes untuk Menentukan Program Studi bagi Calon Mahasiswa Baru dengan Pendekatan Unsupervised Discretization. *Infotekmesin*, 13(1), pp.161–167. https://doi.org/10.35970/infotekmesin.v13i1.1048.

PANDEY, S.C., 2016. Data mining techniques for medical data: A review. *International Conference on Signal Processing, Communication, Power and Embedded System, SCOPES 2016 - Proceedings*, pp.972–982. https://doi.org/10.1109/SCOPES.2016.7955586.

PUTRI, I.E., RAHMAWATI, D. dan AZHAR, Y., 2020. Comparison of Data Mining Classification Methods To Detect Heart Disease. *Jurnal Pilar Nusa Mandiri*, 16(2), pp.213–218. https://doi.org/10.33480/pilar.v16i2.1481.

REDDY, G.T., REDDY, M.P.K., LAKSHMANNA, K., RAJPUT, D.S., KALURI, R. dan SRIVASTAVA, G., 2020. Hybrid genetic algorithm dan a fuzzy logic classifier for heart disease diagnosis. *Evolutionary Intelligence*, 13(2), pp.185–196. https://doi.org/10.1007/s12065-019-00327-1.

RINO, R., 2021. The Comparison of Data Mining Methods Using C4.5 Algorithm dan Naive Bayes in Predicting Heart Disease. *Tech-E*, 4(2), pp.44–51. https://doi.org/10.31253/te.v4i2.543.

ROSENFELD, A., ILLUZ, R., GOTTESMAN, D. dan LAST, M., 2018. Using discretization for extending the set of predictive features. *Eurasip Journal on Advances in Signal Processing*, (1), pp.1–11. https://doi.org/10.1186/s13634-018-0528-x.

ROTH, G. A., MENSAH, G. A., JOHNSON, C. O., ADDOLORATO, G., AMMIRATI, E., BADDOUR, L. M., BARENGO, N. C., BEATON, A., BENJAMIN, E. J., BENZIGER, C. P., BONNY, A., BRAUER, M., BRODMANN, M., CAHILL, T. J., CARAPETIS, J. R., CATAPANO, A. L., CHUGH, S., COOPER, L. T., CORESH, J., … FUSTER, V., 2020. Global Burden of Cardiovascular Diseases dan Risk Factors, 1990-2019: Update From the GBD 2019 Study. *Journal of the American College of Cardiology*, 76(25), pp.2982–3021. https://doi.org/10.1016/j.jacc.2020.11.010.

RUSSELL, S.J. dan NORVIG, P., 2010. *Artificial Intelligence A Modern Approach Third Edition*. 3rd ed. New Jersey: Pearson.

SALEH, A. dan NASARI, F., 2018. Penggunaan Teknik Unsupervised Discretization pada Metode Naive Bayes dalam Menentukan Jurusan Siswa Madrasah Aliyah. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 5(3), pp.353--360. https://doi.org/10.25126/jtiik.201853705.

SETYAWAN, D.A. dan FATHICAH, C., 2018. Pengembangan Metode Decision Tree dengan Diskritisasi Data dan Splitting Atribut Menggunakan Hierarchical Clustering dan Dispersion Ratio. *JUTI: Jurnal Ilmiah Teknologi Informasi*, 18(2), pp.179–187. https://doi.org/10.12962/j24068535.v18i2.a1005.

TANG, X., LI, J., LIU, M., LIU, W. dan HONG, H., 2020. Flood susceptibility assessment based on a novel random Naïve Bayes method: A comparison between different factor discretization methods. *Catena*, 190(February), p.104536. https://doi.org/10.1016/j.catena.2020.104536.

ZHANG, Z., FENG, Q., HUANG, J., GUO, Y., XU, J. dan Wang, J., 2021. A local search algorithm for k-means with outliers. *Neurocomputing*, 450, pp.230–241. https://doi.org/10.1016/j.neucom.2021.04.028.

ZHAO, W.L., DENG, C.H. dan NGO, C.W., 2018. k-means: A revisit. *Neurocomputing*, 291, pp.195–206. https://doi.org/10.1016/j.neucom.2018.02.072.

ZHOU, Y., KANG, J., KWONG, S., WANG, X. dan ZHANG, Q., 2021. An evolutionary multi-objective optimization framework of discretization-based feature selection for classification. *Swarm dan Evolutionary Computation*, 60(February 2020), p.100770. https://doi.org/10.1016/j.swevo.2020.100770.