Indonesian Journal on Data Science
ISSN 2987-7423 - Vol. 1, No. 2, November 2023, hlm 56-65

Analisis Perbandingan Pengukuran Jarak Algoritma K-Nearest Neighbor Dengan Menggunakan Data *Breast Cancer* Dan Data *Heart Disease*

Herdiesel Santoso [a,1,*], Linda Pratiwi [b,2]

[a] Program Studi Sistem Informasi, STMIK EL-Rahma Yogyakarta
[b] Program Studi Informatika, STMIK EL-Rahma Yogyakarta
[1] herdiesel.santoso@stmikelrahma.ac.id*; [2] pratiwilinda032@gmail.com

* corresponding author

ABSTRACT

Breast Cancer is a cancerous condition that appears in the breast area. This type of cancer is often experienced by women with a characteristic feature of *Breast Cancer*, namely the appearance of unusual lumps in the breast area. *Heart* or *Heart Disease* is a type of Non-Communicable *Disease* (PTM): which results in a fairly high mortality rate. *Heart Disease* is caused by several risk factors including smoking, an unhealthy lifestyle, high cholesterol, hypertension, and diabetes.

Based on these facts, an appropriate algorithm is needed to classify Breast Cancer and *Heart Disease* as an effort to prevent an increase in mortality rates due to *Breast Cancer* and *Heart Disease*. And the algorithm that will be used is the K-Nearest Neighbor algorithm with 3 distance measurement methods, namely *Euclidean distance*, *Manhattan distance*, and *Minkowsky distance*.

From the stages that have been carried out, the final results of the *Euclidean distance* method obtained an *Accuracy* value of 80.88% *Breast Cancer* data at K = 11, and 78.69% *heart Disease* data at K = 11. The *Manhattan distance* method obtained an *Accuracy* value of 89.71% of *Breast Cancer* data on K=11, and 78.69% of *Heart Disease* data on K=20.The *Minkowsky distance* method obtained an *Accuracy* value of 98.53% of *Breast Cancer* data on K=11, and 79.41% of *Heart Disease* data on K=11. This shows that the *Minkowsky distance* method works more optimally than the *Euclidean distance* and *Manhattan distance* methods.

ARTICLE INFO

Article history
Received: 15 Oktober 2023
Revised: 10 November 2023
Accepted: 22 November 2023

Keywords
Breast Cancer
Data Mining
Heart Disease
K-Nearest Neighbor
.

1. Introduction

Breast Cancer (kanker payudara) merupakan kondisi kanker yang muncul di daerah payudara. Kanker jenis ini sering dialami oleh Wanita dengan perkiraan 1.67 juta kasus *Breast Cancer* baru yang didiagnosis pada tahun 2012 (25% dari semua kanker) dengan ciri khas dari *Breast Cancer* yaitu munculnya benjolan yang tidak biasa di area payudara [1]. *Heart* atau *Heart Disease* (penyakit jantung) merupakan salah satu jenis Penyakit Tidak Menular (PTM) tetapi mengakibatkan tingkat kematian yang cukup tinggi. Penyakit jantung disebabkan oleh beberapa faktor resiko diantaranya

merokok, gaya hidup yang tidak sehat, tingginya kolesterol, hipertensi, dan diabetes[2]. Jumlah kasus kematian yang disebabkan oleh penyakit jantung meningkat tiap harinya. Mengutip dari World Health Organization (WHO) saat ini telah lebih dari 17 juta jiwa kehilangan nyawa akibat penyakit jantung. Angka tersebut diprediksi akan terus mengalami peningkatan hingga mencapai 23,3 juta jiwa pada tahun 2030 [3].

Sebagai upaya mencegah peningkatan angka kematian akibat penyakit kanker payudara dapat dilakukan dengan klasifikasi. Banyak teknologi yang bisa digunakan dalam teknik klasifikasi untuk mengelola data yang bisa membantu menentukan seseorang memiliki risiko *Breast Cancer* dan *Heart Disease* atau tidak, salah satu teknologi yang paling banyak digunakan digunakan adalah *data mining* [4]–[8]. Penelitian ini akan mengambil dua data tersebut yaitu *Breast Cancer* dan *Heart Disease*, menggunakan perhitungan algoritma *k-Nearest Neighbor* (k-NN). Teknik klasifikasi yang merupakan salah satu fungsi utama data mining, dapat digunakan untuk proses pengelompokkan data dari data yang telah ada dengan menggunakan data berlabel atau data *supervised*. Salah satu algoritma yang banyak digunakan untuk klasifikasi adalah *k-Nearest Neighbor*. Algoritma *k-Nearest Neighbor* menggunakan pendekatan *supervised learning* dimana data yang digunakan merupakan data berlabel [9]. Selain itu, algoritma ini sederhana dan mudah diintreprestasikan. Meskipun sederhana, algoritma ini telah diuji dibeberapa kasus dan menghsilkan performa yang cukup tinggi [7].

Kualitas hasil pengelompokkan algoritma *k-Nearest Neighbor* sangat bergantung pada jarak kedekatan antar objek dan nilai k yang ditetapkan. Penelitian [10] menggunakan algoritma k-Nearest Neighbor dengan membandingkan beberapa fungsi pengukuran jarak antara *Euclidean distance*, *Manhattan distance*, *tchebychev distance*, *cosine distance* dan *correlation distance*. Penelitian tersebut memberikan hasil akurasi terbaik pada dua metode yaitu *Euclidean distance* dan *Manhattan distance*, dimana pengukuran dengan metode tersebut berhasil memberikan tingkat akurasi sebesar 98,70% pada k=1. Dikarenakan metode k-Nearest Neighbor sangat bergantung pada hasil perhitungan jarak antar objek, maka pemilihan metode untuk perhitungan jarak sangat menentukan hasil pengelompokkan [6]. Berdasarkan hal tersebut, maka penelitian ini akan menggunakan dua jenis data yaitu *Breast Cancer* dan *Heart Disease* dan membandingan metode pengukuran jarak *Euclidean distance*, *Manhattan distance*, dan *Minkowski distance* untuk mengetahui metode mana yang memiliki nilai akurasi tertinggi.

2. Method

Pendekatan data mining yang digunakan dalam melakukan tahapan penelitian ini yaitu CRISP-DM. CRISP-DM (*Cross Industry Standard Process for Data Mining*) suatu standarisasi pemrosesan data mining yang telah dikembangkan dimana data yang ada akan melewati setiap fase terstruktur dan terdefinisi dengan jelas dan efisien [11]. Gambar 1 menunjukan 6 (enam) tahapan dalam metodologi CRIPS-DM, yaitu *business understanding, data understanding, data preparation, modeling, evaluation, dan deployment*.

Fig. 1. Proses Tahapan CRISP-DM

2.1. Business Understanding

Tahapan pertama adalah memahami tujuan dan kebutuhan dari sudut pandang bisnis, kemudian menerjemahkan pengetahuan ini ke dalam pendefinisian masalah dalam data mining [12]. Tujuan penelitian ini adalah melakukan analisis pengaruh metode pengukuran jarak pada algoritma klasifikasi k-NN. Data yang digunakan merupakan data kesehatan yaitu data penyakit *Breast Cancer* dan data *heart*. Dataset *Breast Cancer* payudara memiliki 569 kasus dan data *Heart Disease* memiliki 302 kasus. Metode data ini menggunakan metode pengukuran jarak algoritma *k-Nearest Neighbour* yaitu *Euclidean distance*, *Manhattan distance*, dan *Minkowski distance*.

2.2. Data Understanding

Tahap ini dimulai dengan pengumpulan data yang kemudian alam dilanjutkan dengan proses untuk mendapatkan pemahaman yang mendalam tentang data, mengidentifikasi masalah kualitas data, atau untuk mendeteksi adanya bagian yang menarik hipotesa untuk informasi yang tersembunyi [13]. Keseluruhan data yang akan digunakan dalam penelitian ini adalah sebanyak 569 kasus data *Breast Cancer* dengan 12 atribut, dan 302 kasus data *Heart Disease* dengan 14 atribut. Atribut yang akan digunakan mempunyai tipikal *categorical* dan *numerik*. Atribut data *Breast Cancer* dapat dilihat pada Tabel 1 dan data *Heart Disease* pada Tabel 2.

Table 1. Atribut data *Breast Cancer*

No.	Atribut	Keterangan
1.	Nomor kode sampel	Nomor id pasien
2.	Radius	Data Hasil FNA
3.	Texture	Data hasil FNA
4.	Perimeter	Data hasil FNA
5.	Area	Data hasil FNA
6.	Smoothness	Data hasil FNA
7.	Compactness	Data hasil FNA
8.	Concavity	Data hasil FNA
9.	Concave Point	Data hasil FNA
10.	Symmetry	Data hasil FNA
11.	Fractal Dimension	Data hasil FNA
12	Diagnosis	B untuk benign M untuk malignant

Table 2. Atribut data *Heart Disease*

No.	Atribut	Keterangan
1	Age	29.0 – 77.0
2	Sex	Male, Female
3	CP	Abnang, Angina, Asympt, Notang
4	Trestbps	94.0 – 200.0
5	Chol	126.0 – 569.0
6	Fbs	True, False
7	Restecg	Norm, Hyp, Abn
8	Thalach	99.0 – 103.0
9	Exang	True, False
10	Oldpeak	0.0 – 6.2
11	Slope	Down, Flat, Up
12	CA	0.0 – 3.0
13	Thal	Normal, Rever, Fixed
14	Target	Healthy, Sick

Penyusunan basis data dilakukan dengan memasukkan data-data riil penelitian yang diperoleh data data *Breast Cancer* dan data *heart*. Data kasus pada basis kasus akan dijadikan sebagai acuan untuk menghasilkan solusi bagi kasus dijadikan untuk penghasil solusi bagi kasus baru. Jumlah data latih

yang digunakan dalam penelitian ini sebanyak 455 data training dan 114 data testing kasus *Breast Cancer* payudara, sebanyak 242 data training dan 60 data testing kasus *Heart Disease*.

2.3. Data Preparation

Dalam tahap ini meliputi semua kegiatan untuk membangun dataset akhir (data yang akan diproses pada tahap pemodelan (*modeling*) dari data mentah. Pada tahapan ini dilakukan pengecekan atau pencarian apakah terdapat data yang hilang (*missing value*) atau tidak, pemilihan , *record*, dan atribut-atribut data, termasuk proses pembersihan dan transfarmasi data.Transformasi data dilakukan untuk mengubah data menjadi nilai dengan format tertentu. Seperti dalam atribut data yang *categorical* akan diubah menjadi data *numeric*al. Untuk atribut data yang *categorical* akan di konversi ke numerik. Untuk perhitungan menggunakan 80% data testing dan 20% data testing.

2.4. Modeling

Dalam tahap ini melakukan pemilihan dan penerapan berbagai pemodelan data beberapa parameternya akan disesuaikan untuk mendapatkan nilai yang optimal. Algoritma *k-Nearest Neighbor* (k-NN) merupakan sebuah metode untuk melakukan klasifikasi terhadap obyek baru berdasarkan (K) tetangga terdekatnya [6]. Algoritma *k-Nearest Neighbor* termasuk algoritma *supervised learning*, yang mana hasil dari *query instance baru*, diklasifikasikan berdasarkan mayoritas dari kategori pada k-NN. Kelas yang paling banyak muncul, yang akan menjadi kelas hasil klasifikasi. Algoritma k-NN digunakan untuk mengklasifikasi objek oleh mayoritas suara k benda referensi terdekat. Jadi, k-NN terdiri dari dua proses yang memakan waktu: komputasi jarak dan peringkat jarak [7]. Ada beberapa metode pengukuran jarak pada algoritma k-NN yang akan digunakan untuk penelitian [10].

2.4.1. Euclidean distance

Euclidean distance merupakan salah satu metode perhitungan jarak dari dua buah titik dalam *Euclidean space* (meliputi bidang *euclidean* dua dimensi, tiga dimensi, atau bahkan lebih). Untuk mengukur tingkat kemiripan data dengan rumus *Euclidean distance* digunakan persamaan 1.

$$d(x,y)^{=}|x-y| = \sqrt{\sum_{i=-1}^{n}(xi-yi)^2} \quad (1)$$

dimana d = jarak antara x dan y; x dan y = dua objek data yang memiliki n atribut bernilai numerik, ; i = setiap data; n = jumlah data.

2.4.2. Manhattan distance

Manhattan distance digunakan untuk menghitung perbedaan absolut (mutlak) antara koordinat sepasang objek. Untuk mengukur tingkat kemiripan data dengan rumus *Manhattan distance* digunakan persamaan 2.

$$d(\mathrm{x},\mathrm{y}) = \sum_{i=1}^{n}|xi - yi| \quad (2)$$

dimana d = jarak antara x dan y; x dan y = dua objek data yang memiliki n atribut bernilai numerik, ; i = setiap data; n = jumlah data.

2.4.3. Minkowski distance

Minkowski distance merupakan sebuah metrik dalam ruang vektor dimana suatu norma didefinisikan (*normed vector space*) sekaligus dianggap sebagai generalisasi dari *Euclidean distance* dan *Manhattan distance*. Dalam pengukuran jarak objek menggunakan *Minkowski distance* biasanya digunakan niali p adalah 3 atau ∞. Untuk mengukur tingkat kemiripan data dengan rumus *Manhattan distance* digunakan persamaan 3.

$$d(\mathrm{x},\mathrm{y}) = (\sum_{i=1}^{n}|xi-yi|p)1/p \quad (3)$$

dimana d = jarak antara x dan y; x dan y = dua objek data yang memiliki n atribut bernilai numerik, ; i = setiap data; n = jumlah data, p = power.

2.5. Evaluation

Pada Tahap ini dilakukan evaluasi terhadap keefektifan dan kualitas model sebelum digunakan dan menentukan apakah model dapat mencapai tujuan yang ditetapkan pada fase awal (*Business Understanding*). Model Evaluation yang digunakan adalah *K-fold Cross Validation* menggunakan *Widget Cross Validation RapidMinner*. Nilai yang dievaluasi adalah *Accuracy,Precission* dan *Recall* dari algoritma yang digunakan. Proses evaluasi mengikuti langkah pada Gambar 2.

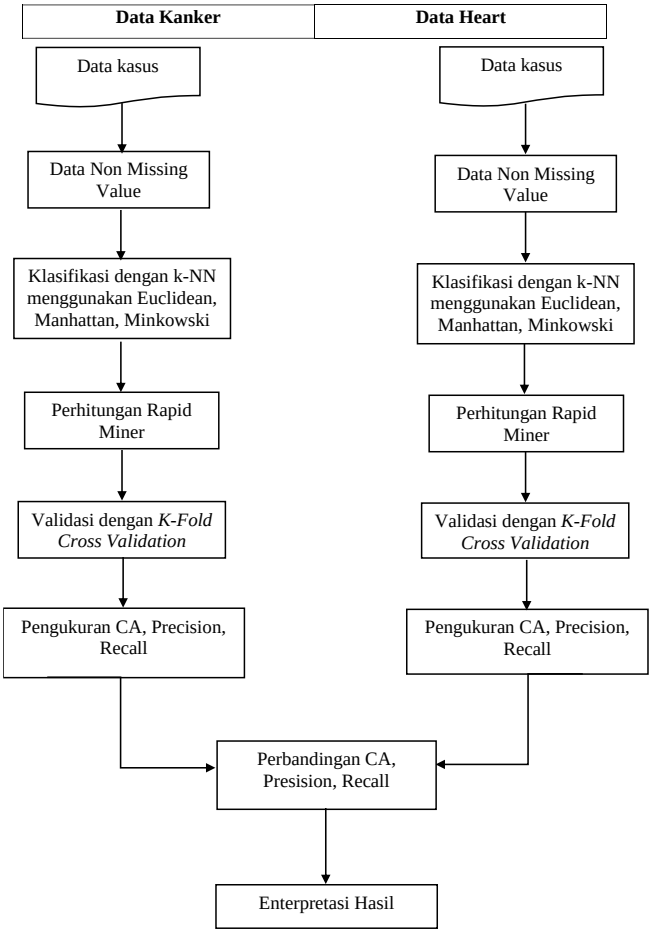

Fig. 2. Langkah evaluasi penelitian

Perhitunag *Accuracy,Precission* dan *Recall*, dilakukan dengan melihat tabel *confusion matrix* memberikan penilaian performance klasifikasi berdasarkan objek dengan benar atau salah [14]. *Confusion matrix* berisi informasi aktual (*actual*) dan prediksi (*Predicted*) pada *system* klasifikasi, seperti pada Tabel 3.

Table 3. *Confusion Matrix*

Nilai Prediksi	Nilai Aktual	
	TP	FN
	FP	TN

Keterangan :
TP = tupel positif yang diklasifikasikan positif; TN = tupel positif yang diklasifikasikan negatif; FP = tupel negatif yang diklasifikasikan positif; FN = tupel negatif yang diklasifikasikan negatif.

Untuk menghitung tingkat akurasi pada matriks digunakan persamaan 4.

$$Akurasi = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (4)$$

Untuk menghitung tingkat *precision* pada matriks digunakan persamaan 5.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

Untuk menghitung tingkat *recall* pada matriks digunakan persamaan 6.

$$Recall = \frac{TN}{TN + FN} \quad (6)$$

2.6. Deployment

Pada tahap terakhir informasi yang diperoleh akan diatur dan dipresentasikan dalam bentuk khusus sehingga dapat digunakan oleh pengguna. Pada tahap ini berupa laporan sederhana tentang data mining dalam perusahaan secara berulang. Dengan uji coba menggunakan banyak model yang telah dibuat.

3. Results and Discussion

Metode k-NN digunakan untuk menghitung tetangga terdekat dari data testing ke seluruh data training. Langkah perhitunganya adalah sebagai berikut.

1. Menentukan K dengan Parameter (Tetangga Terdekat)

Nilai K didapatkan dengan membandingkan bilangan ganjil antara 1-20. Maka K yang akan digunakan dalam perhitungan antara lain K=1, K=3, K=5, K=7, K=9, K=11, K=13, K=15, K=17 dan K=19. Nilai K akan digunakan dalam pehitungan jarak *Eucledian distance*, *Manhattan distance*, dan *Minkowski distance*.

2. Menghtiung Jarak antara data training Tabel 3 dengan data testing Tabel 4.

Table 4. Contoh data training

id	Radius_ mean	Texture_ mean	Perimeter_ mean	Area_ mean	Smoothness_ mean	Compactness_ mean	Concavity_ mean	Concave point_mean	Symmetry_ mean	Fractal_ dimension_ mean
842302	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871
842517	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667
84300903	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999

Table 5. Contoh data testing

id	radius_ mean	texture_ mean	perimeter_mean	area_mean	smoothness_ mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean
91485	20.59	21.24	137.8	1320	0.1085	0.1644	0.2188	0.1121	0.1848	0.06222

Perhitungan *Euclidean distance* seluruh data training 1 dengan data testing 1 menggunakan persamaan 1.

$$\sqrt{\begin{matrix}(17{,}99-20{,}59)^2+(10{,}38-21{,}24)^2+(122{,}8-137{,}8)^2+\\(1001-1320)^2+(0{,}1184-0{,}1085)^2+(0{,}2776-0{,}1644)^2+\\(0{,}3001-0{,}2188)^2+(0{,}1471-0{,}1121)^2+(0{,}2419-0{,}1848)^2\\+(0{,}07871-0{,}06222)^2\end{matrix}}^2 = 319{,}5482006$$

Perhitungan *Manhattan distance* seluruh data training 1 dengan data testing 1 menggunakan persamaan 2.

$|17{,}99 - 20{,}59| + |10{,}38 - 21{,}24| + |122{,}8 - 137{,}8| + |1001 - 1320| + |0{,}08474 - 0{,}1085| + |0{,}07864 - 0{,}1644| + |0{,}3001 - 0{,}2188| + |0{,}1471 - 0{,}1121| + |0{,}2419 - 0{,}1848| + |0{,}07871 - 0{,}06222|$ = 347,77299

Perhitungan *Minskowski distance* seluruh data training 1 dengan data training 1 menggunakan persamaan 3.

$$\sqrt[3]{\begin{matrix}(17{,}99 - 20{,}59)^3 + (10{,}38 - 21{,}24)^3 + (122{,}8 - 137{,}8)^3 + \\ (1001 - 1320)^3 + (0{,}1184 - 0{,}1085)^3 + (0{,}2776 - 0{,}1644)^3 + \\ (0{,}3001 - 0{,}2188)^3 + (0{,}1471 - 0{,}1121)^3 + (0{,}2419 - 0{,}1848)^3 \\ + (0{,}07871 - 0{,}06222)^3\end{matrix}} = 2588841425$$

3. Melakukan evaluasi menggunakan *K-fold Cross Validation*

Selanjutnya dilakukan analisis metode pengujian terhadapt model-model yang bertujun untuk mendapatkan model yang paling akurat. Metode yang digunakan untuk evaluasi adalah *K-fold Cross Validation* dengan jumlah data 512 data *Breast Cancer* payudara dan 274 data *Heart disease*.

3.1. Evaluasi Nilai K

Analisis pertama dilakukan dengan mencari nilai K yang menghasilkan akurasi tertinggi dan metode jarak yang memiliki *Accuracy*,*Precission* dan *Recall* terbaik. Perbandingan nilai K dapat dilihat pada Tabel 6.

Table 6. Nilai k yang memiliki akurasi terbaik

K-Terdekat	Nilai K	Akurasi Terbaik
Euclidean Distance	**11**	**79.69%**
Manhattan distance	**11**	**85%**
Minskowky Distance	**11**	**97.65%**

Dapat dilihat bahwa metode perhitungan jarak *Minkowski distance* memiliki nilai akurasi terbaik dibanding *Manhattan distance* dan *Euclidean distance*, dengan nilai 97.65%. Sedangkan pada manhattan memiliki akurasi nilai terbaik yaitu 85% dan *Euclidean distance* memiliki akurasi terbaik sebesar 79.69%.

3.2. Perhitungan *Accuracy, Precission* dan *Recall*

Setelah didapatkan nilai K terbaik, langkah selanjutnya adalah membuat tabel *confusion matrix* dengan bantuan RapidMiner. Tabel 7 merupakan *confusion matrix* data *Breast cancer* dan Tabel 8 merupakan *confusion matrix* data *Heart disease*. Tabel *confusion matrix* digunakan untuk membantu perhitungan adalah *Accuracy*,*Precission* dan *Recall*.

Table 7. *Confusion matrix* data *Breast cancer*

	Euclidean distance			Manhattan distance			Minkowski distance		
	True M	**True B**	**Class precision**	**True M**	**True B**	**Class precision**	**True M**	**True B**	**Class precision**
Pred M	4	0	100%	8	0	100%	16	4	80%
Pred B	13	51	79.69%	9	51	85%	1	47	97.92%
Class recall	23.53%	100%		47.06%	100%		94.12%	92%	

Table 8. *Confusion matrix* data *Heart disease*

	Euclidean distance			Manhattan distance			Minkowski distance		
	True Sick	True Health	Class precision	True Sick	True Health	Class precision	True Sick	True Health	Class precision
Pred Sick	25	13	65,79%	24	11	69%	24	17	59%
Pred Health	11	12	52,17%	12	14	54%	12	8	40%
Class recall	69,44%	48%		66,67%	56%		66,67%	32%	

Perhitungan *Accuracy* menggunakan persaman 4, *Precission* menggunakan persamaan 5 dan *Recall* menggunakan persamaan 6. Hasil *Cross Validation* perbandingan nilai *Accuracy*,*Precission* dan *Recall* dapat dilihat pada Tabel 9.

Table 9. Perbandingan nilai *Accuracy*, *Precission* dan Recall

Parameter	*Breast Cancer*			*Heart Disease*		
	Manhattan	*Euclidean*	*Minkowski*	*Manhattan*	*Euclidean*	*Minkowski*
CA	85%	79,69%	97,92%	53,85%	52,17%	52,46%
Precision	100%	100%	80%	68,57%	65,79	40%
Recall	47,06%	23,53%	94,12%	66,67%	69,44%	66,67%

Hasil evaluasi dengan menggunakan *K-fold Cross Validation* menghasilkan nilai CA (*Clasification Accuracy*) metode *Euclidean distance* memiliki akurasi 79,69% untuk data *Breast Cancer* dan 52,17% untuk data *Heart Disease*. Nilai *Accuracy* dengan metode *Manhattan distance* yaitu 85% untuk data *Breast Cancer* dan 53,85% untuk data *heart*. Nilai *Accuracy* dengan metode *Minkowsky distance* yaitu 97,92% untuk data *Breast Cancer* dan 52,46% untuk data *Heart Disease.*

Nilai *Precission* metode *Euclidean distance* memiliki akurasi 100% untuk data *Breast Cancer*, dan 65,79% untuk data *Heart Disease*. Nilai *Precission* dengan metode *Manhattan distance* yaitu 100% untuk data *Breast Cancer* dan 68,57% untuk data *Heart Disease*. Nilai *Precission* dengan metode *Minkowsky distance* yaitu 80% untuk data *Breast Cancer* dan 40% untuk data *Heart Disease.* Nilai *Recall* metode *Euclidean distance* memiliki akurasi 23,53% untuk data *Breast Cancer*, dan 69,44% untuk data *Heart Disease.* Nilai *recall* dengan metode *Manhattan distance* yaitu 47,06% untuk data *Breast Cancer* dan 66,67% untuk data *Heart Disease*. Nilai *Recall* dengan metode *Minkowsky distance* yaitu 94,12% untuk data *Breast Cancer* dan 66,67% untuk data *Heart Disease*.

3.3. *Correlation Matrix pada Breast Cancer dan Heart Disease*

3.3.1 Gejala pada Data *Breast Cancer*

Gejala yang paling dominan penyebab timbulnya penyakit breast cancer dapat dilihat pada Gambar 3. Pada Gambar 3 dapat diketahui bahwa nilai korelasi 0.998 dan 0.987 itu berada pada rentang nilai koefisien korelasi 0.80 – 1.00 dengan demikian diketahui bahwa hubungan antara *radius_mean* dengan *perimeter_mean* dan *radius_mean* dengan *area_mean* adalah "Sangat Kuat" dan untuk gejala tersebut sangat berpengaruh pada penyakit Breast Cancer. Dan untuk gejala – gejala yang memiliki nilai negatif merupakan gejala tersebut tidak berpengaruh pada penyakit *Breast Cancer*.

3.3.2 Gejala pada Data *Hearth disease*

Gejala yang paling dominan penyebab timbulnya penyakit heart disease dapat dilihat pada Gambar 4. Pada gambar 4 dapat diketahui bahwa nilai korelasi paling tinggi yaitu 0.296 – 0.387 yang berada pada rentang nilai koefisien korelasi 0.20 – 0.30 dengan demikian diketahui bahwa hubungan antara gejala *thalach* dengan gejala *slope* adalah "Sangat Kuat" dan untuk gejala tersebut sangat berpengaruh pada penyakit *Heart disease*. Dan untuk gejala – gejala yang memiliki nilai negatif merupakan gejala tersebut tidak berpengaruh pada penyakit *Heart disease*.

Attributes	id	radius_m...	texture_m...	perimeter...	area_mean	smoothne...	compactn...	concavity...	concave p...	symmetry...	fractal_di...
id	1	0.075	0.100	0.073	0.097	-0.013	0.000	0.050	0.044	-0.022	-0.053
radius_me...	0.075	1	0.324	0.998	0.987	0.171	0.506	0.677	0.823	0.148	-0.312
texture_me...	0.100	0.324	1	0.330	0.321	-0.023	0.237	0.302	0.293	0.071	-0.076
perimeter_...	0.073	0.998	0.330	1	0.987	0.207	0.557	0.716	0.851	0.183	-0.261
area_mean	0.097	0.987	0.321	0.987	1	0.177	0.499	0.685	0.823	0.151	-0.283
smoothnes...	-0.013	0.171	-0.023	0.207	0.177	1	0.659	0.522	0.554	0.558	0.585
compactne...	0.000	0.506	0.237	0.557	0.499	0.659	1	0.883	0.831	0.603	0.565
concavity_...	0.050	0.677	0.302	0.716	0.685	0.522	0.883	1	0.921	0.501	0.337
concave po...	0.044	0.823	0.293	0.851	0.823	0.554	0.831	0.921	1	0.462	0.167
symmetry_...	-0.022	0.148	0.071	0.183	0.151	0.558	0.603	0.501	0.462	1	0.480
fractal_dim...	-0.053	-0.312	-0.076	-0.261	-0.283	0.585	0.565	0.337	0.167	0.480	1

Fig. 3. Hasil *Correlation Matrix* pada Data *Breast cancer*

Attributes	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
age	1	-0.098	-0.069	0.279	0.214	0.121	-0.116	-0.399	0.097	0.210	-0.169	0.276	0.068
sex	-0.098	1	-0.049	-0.057	-0.198	0.045	-0.058	-0.044	0.142	0.096	-0.031	0.118	0.210
cp	-0.069	-0.049	1	0.048	-0.077	0.094	0.044	0.295	-0.394	-0.149	0.120	-0.181	-0.162
trestbps	0.279	-0.057	0.048	1	0.123	0.178	-0.114	-0.047	0.068	0.193	-0.121	0.101	0.062
chol	0.214	-0.198	-0.077	0.123	1	0.013	-0.151	-0.010	0.067	0.054	-0.004	0.071	0.099
fbs	0.121	0.045	0.094	0.178	0.013	1	-0.084	-0.009	0.026	0.006	-0.060	0.138	-0.032
restecg	-0.116	-0.058	0.044	-0.114	-0.151	-0.084	1	0.044	-0.071	-0.059	0.093	-0.072	-0.012
thalach	-0.399	-0.044	0.295	-0.047	-0.010	-0.009	0.044	1	-0.379	-0.344	0.387	-0.213	-0.096
exang	0.097	0.142	-0.394	0.068	0.067	0.026	-0.071	-0.379	1	0.288	-0.258	0.116	0.207
oldpeak	0.210	0.096	-0.149	0.193	0.054	0.006	-0.059	-0.344	0.288	1	-0.578	0.223	0.210
slope	-0.169	-0.031	0.120	-0.121	-0.004	-0.060	0.093	0.387	-0.258	-0.578	1	-0.080	-0.105
ca	0.276	0.118	-0.181	0.101	0.071	0.138	-0.072	-0.213	0.116	0.223	-0.080	1	0.152
thal	0.068	0.210	-0.162	0.062	0.099	-0.032	-0.012	-0.096	0.207	0.210	-0.105	0.152	1

Fig. 4. Hasil *Correlation Matrix* pada Data *Heart disease*

4.Conclusion

Penggunaan data mining dengan data *Breast Cancer* untuk penelitian sebanyak 512 data training dan 57 data testing, dan data *Heart Disease* untuk penelitian sebanyak 274 data training dan 30 data testing. Berdasarkan perhitungan *Accuracy,Precission* dan *Recall* menggunakan metode *K-fold Cross Validation*, metode *Minkowsky distance* memiliki kinerja akurasi yang paling bagus. Selanjutnya antara data *Breast cancer* dan *Heart disease*, yang memiliki data *Breast Cancer* dapat digunakan untuk penelitian selanjunya, karena memiliki *Accuracy,Precission* dan *Recall* yang lebih baik. Pada penelitian ini didapatkan nilai K yang paling optimal yaitu K=11, dimana nilai K=11 memiliki tingkat akurasi yang tinggi dibandingkan dengan nila K yang lain. Dan pada penelitian ini di dapatkan gejala yang paling dominan atau yang paling berpengaruh pada penyakit *Breast Cancer* yaitu *radius_mean* dengan *perimeter_mean* dan pada penyakit *Heart Disease* yaitu gejala *thalach* dengan gejala *slope.*

References

[1] H. Kurniasih, Sumiyati, S. P. Winarso, and Z. Fitria, "The Level of Knowledge, Attitudes, Behaviour of Women in Reproductive Age (WRA) with Online Class BSE," *J. Kebidanan*, vol. 12, no. 2, pp. 112–118, 2022, doi: https://doi.org/10.31983/jkb.v12i2.6906.

[2] D. D. Anggraini and A. C. Hidajah, "Hubungan antara Paparan Asap Rokok dan Pola Makan dengan Kejadian Penyakit Jantung Koroner pada Perempuan Usia Produktif," *Amerta Nutr.*, vol. 2, no. 1, p. 10, 2018, doi: 10.20473/amnt.v2i1.2018.10-16.

[3] WHO, "World Health Statistics 2023." [Online]. Available:

https://iris.who.int/bitstream/handle/10665/367912/9789240074323-eng.pdf?sequence=1

[4] W. Nugraha, "Prediksi Penyakit Jantung Cardiovascular Menggunakan Model Algoritma Klasifikasi," *J. Manag. dan Inform.*, vol. 9, no. 2, pp. 3–8, 2021.

[5] A. Alhamad, A. I. S. Azis, B. Santoso, and S. Taliki, "Prediksi Penyakit Jantung Menggunakan Metode-Metode Machine Learning Berbasis Ensemble – Weighted Vote," *J. Edukasi dan Penelit. Inform.*, vol. 5, no. 3, p. 352, 2019, doi: 10.26418/jp.v5i3.37188.

[6] S. W. Binabar and Ivandari, "Optimasi Parameter K pada Algoritma KNN untuk Deteksi Penyakit Kanker Payudara," *IC-Tech*, vol. XII, no. 2, pp. 11–18, 2017.

[7] I. Handayani and I. Ikrimach, "Comparison of K-Nearest Neighbor and Naïve Bayes for Breast Cancer Classification Using Python," *IJISCS (International J. Inf. Syst. Comput. Sci.*, vol. 5, no. 1, p. 1, 2021, doi: 10.56327/ijiscs.v5i1.953.

[8] M. M. Ahsan, L. S. Akter, and S. Zahed, "Machine-Learning-Based Disease Diagnosis : A Comprehensive Review," *Healthcare*, vol. 10, no. 3, pp. 1–30, 2022.

[9] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition*, vol. 9780470908. 2014. doi: 10.1002/9781118874059.

[10] S. Ahmed Medjahed, T. Ait Saadi, and A. Benyettou, "Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules," *Int. J. Comput. Appl.*, vol. 62, no. 1, pp. 1–5, 2013, doi: 10.5120/10041-4635.

[11] M. A. Hasanah, S. Soim, and A. S. Handayani, "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir," *J. Appl. Informatics Comput.*, vol. 5, no. 2, pp. 103–108, 2021, doi: 10.30871/jaic.v5i2.3200.

[12] O. Niakšu, "CRISP Data Mining Methodology Extension for Medical Domain," *Balt. J. Mod. Comput.*, vol. 3, no. 2, pp. 92–109, 2015.

[13] H. Seetha, M. N. Murty, and B. K. Tripathy, *Modern Technologies for Big Data Classification and Clustering*. Hershey PA: IGI Global, 2018. doi: 10.4018/978-1-5225-2805-0.

[14] H. Said, N. H. Matondang, and H. N. Irmanda, "Penerapan Algoritma K-Nearest Neighbor Untuk Memprediksi Kualitas Air Yang Dapat Dikonsumsi," *Techno.Com*, vol. 21, no. 2, pp. 256–267, 2022, doi: 10.33633/tc.v21i2.5901.