NASKAH PUBLIKASI

ANALISIS PERBANDINGAN PENGUKURAN JARAK ALGORITMA K-NEAREST NEIGHBOR DENGAN MENGGUNAKAN DATA BREAST CANCER DAN DATA HEART

Disusun Oleh

Nama	: Linda Pratiwi
Nomor Mahasiswa	: 12171564
Program Studi	: Informatika
Jenjang	: Sarjana

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN ILMU KOMPUTER
EL RAHMA
YOGYAKARTA
2022

HALAMAN PERSETUJUAN

Setelah melakukan bimbingan, telaah, arahan dan koreksi terhadap penulisan skripsi saudara :

Linda Pratiwi, NIM : 12171564 yang berjudul :

ANALISIS PERBANDINGAN PENGUKURAN JARAK ALGORITMA *K-NEAREST NEIGHBOR* DENGAN MENGGUNAKAN DATA BREAST CANCER DAN DATA *HEART*

Pembimbing berpendapat bahwa skripsi tersebut di atas sudah dapat diajukan dalam sidang ujian skripsi

Yogyakarta, 24 Januari 2022
Pembimbing

Herdiesel Santoso, S.T.,S.Kom.,M.Cs
NPP. 201640066

ANALISIS PERBANDINGAN PENGUKURAN JARAK ALGORITMA K-NEAREST NEIGHBOR DENGAN MENGGUNAKAN DATA BREAST CANCER DAN DATA HEART

Linda Pratiwi[1], Herdiesel Santoso
Program Studi Informatika, STMIK EL-Rahma Yogyakarta
E-mail: pratiwilinda032@gmail.com

Intisari

Breast Cancer (kanker payudara) merupakan kondisi kanker yang muncul di daerah payudara. Kanker jenis ini sering dialami oleh Wanita dengan ciri khas dari kanker payudara yaitu munculnya benjolan yang tidak biasa di area payudara. Heart atau Heart Disease (Penyakit Jantung) merupakan salah satu jenis Penyakit Tidak Menular (PTM): yang mengakibatkan tingkat kematian yang cukup tinggi. Penyakit jantung dosebabkan oleh beberapa factor resiko diantaranya merokok, gaya hidup yang tidak sehat, tingginya kolesterol, hipertensi, dan diabetes.

Berdasarkan fakta tersebut diperlukan algoritma yang tepat untuk mengklasifikasi Breast Caner (kanker payudara) dan Heart Disease (penyakit jantng) sebagai salah satu upaya mencegah peningkatan angka kematian akibat Breast Cancer dan Heart Disease. Dan algoritma yang akan digunakan yaitu algoritma K-Nearest Neighbor dengan 3 metode pengukuran jarak yaitu Euclidean distance, Manhattan distance, dan Minkowsky distance.

Dari tahapan-tahapan yang telah dilakukan diperoleh hasil akhir metode Euclidean distance memperoleh nilai akurasi 80.88% data Breast Cancer pada K=11, dan 78.69% data Heart Disease pada K=11. Metode Manhattan distance memperoleh nilai akurasi 89.71% data Breast Cancer pada K=11, dan 78.69% data Heart Disease pada K=20. Metode Minkowsky distance memperoleh nilai akurasi 98.53% data Breast Cancer pada K=11, dan 79.41% data Heart Disease pada K=11. Hal tersebut menunjukkan bahwa metode Minkowsky distance berkerja lebih maksimal dibandingkan metode Euclidean distance dan Manhattan distance.

Abstract

Breast Cancer (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 Caner (breast cancer) and Heart Disease (jantng 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.

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

PENDAHULUAN

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 kanker baru yang didiagnosis pada tahun 2012 (25% dari semua kanker) dengan ciri khas dari kanker payudara yaitu munculnya benjolan yang tidak biasa di area payudara. Heart atau Heart Disease (penyakit jantung) merupakan salah satu jenis Penyakit Tidak Menular (PTM): yang mengakibatkan tingkat kematian yang cukup tinggi. Penyakit jantung dosebabkan oleh beberapa factor resiko diantaranya merokok, gaya hidup yang tidak sehat, tingginya kolesterol, hipertensi, dan diabetes. Jumlah kasus kematian yang disebabkan oleh Heart Disease meningkat tiap harinya. Mengutip dari World Health Organization (WHO) saat ini telah lebih dari 17 juta jiwa kehilangan nyawa akibat Heart Disease. Angka tersebut diprediksi akan terus mengalami peningkatan hingga mencapai 23.3 juta jiwa pada tahun 2030.

Sebagai upaya mencegah peningkatan angka kematian akibat Breast Heart Disease dapat dilakukan prediksi Breasrt Cancer dan Heart Disease yang terdapat pada manusia. Banyak teknologi yang bisa digunakan dalam Teknik prediksi untuk mengelola data yang bisa membantu menentukan seseorang memiliki risiko Breast Cancer dan Heart Disease atau tidak. Data mining adalah salah satu teknologi yang paling banyak digunakan.

Kesimpulannya pada penelitian ini akan mengambil dari dua data tersebut yaitu Breast Cancer dan Heart Disease. Dan diharapkan pada penelitian berikutnya akan ada aplikasi yang dapat mendiagnosis penyakit Breast Cancer dan Heart Disease. Pada penelitian ini akan menggunakan perhitungan algoritma k-NN dengan metode pengukuran jarak *Eucludean Distance*, *Manhattam Distance*, dan *Minkowski Distance*.

Teknik klasifikasi yang merupakan salah satu fungsi utama data mining, dapat digunakan untuk proses pengelompokkan data dari data yang telah ada dengan enggunakan data berlabel atau data *supervised*. Salah satu algoritma yang banyak digunakan untuk klasifikasi adalah *k-Nearest Neighbor* atau kNN. Algoritma k-

Nearest Neighbor menggunakan pendekatan supervised learning dimana data yang digunakan merupakan data berlabel. Selain itu, algoritma ini sederhana dan mudah diintreprestasikan. Meskipun sederhana, algoritma ini telah diuji dibeberapa kasus dan menghsilkan performa yang cukup tinggi. (Atmaja, 2019)

Kualitas hasil pengelompokkan algoritma *k-Nearest Neighbor* sangat bergantung pada jarak kedekatan antar objek dan nilai k yang ditetapkan (AhmedMedjahed et al., 2013) melakukan penelitian pada algoritma *k-Nearest Neighbor* dengan membandingkan beberapa fungsi pengukuran jarak. Pada penelitian tersebut, dilakukan perbandingan metode pengukuran jarak antara *Euclidean distance*, *Manhattan distance, tchebychev distance, cosine distance* dan *correlation distance*. Penelitian ini memberikan hasil akurasi terbaik pada dua (2) 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. Berdasarkan hal tersebut, maka penelitian ini akan menggunakan dua jenis data untuk membandingan metode pengukuran jarak *Eucludean distance*, *Manhattan distance*, dan *Minkowski distance* untuk mengetahui metode mana yang memiliki nilai akurasi tertinggi.

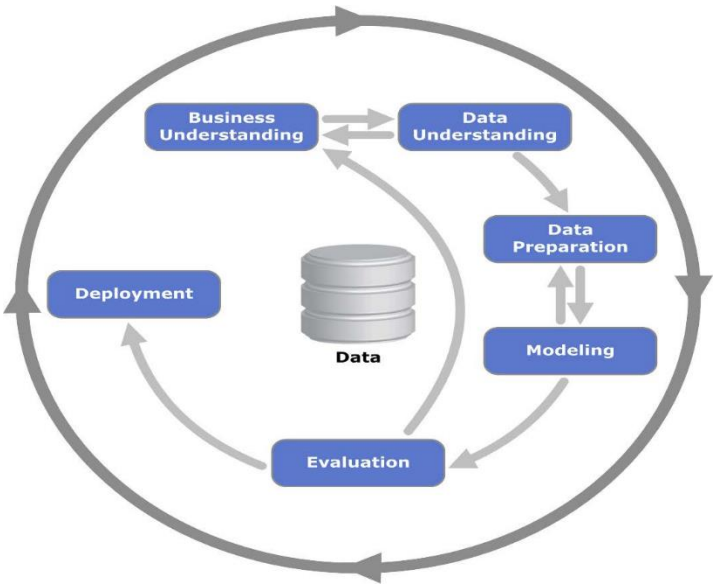
LANDASAN TEORI

1. Data Mining

Data mining adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam *database*, data *warehouse*, atau penyimpanan informasi lainnya. Data mining berkaitan dengan bidang ilmu-ilmu lain, seperti *database system*, data *warehousing*, statistik, *machine learning*, *information retrieval*, dan komputasi tingkat tinggi. Selain itu, data mining didukung oleh ilmu lain seperti *neural network*, pengenalan pola, spatial data *analysis*, *image database*, *signal processing*.

2. Proses Data Mining

CRISP-DM (*Cross Industry Standart Proses Data Mining*) merupakan suatu konsorsium perusahaan yang didirikan oleh Komisi Eropa pada tahun 1996 dan telah ditetapkan sebagai proses standart dalam data mining yang diaplikasikan berbagai *sector industry*. Berikut ini adalah gambar proses siklus hidup perkembangan CRISP-DM.



Gambar 1 Proses Tahapan CRISP-DM

Keterangan gambar:

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. Selanjutnya akan ditemukan rencana dan strategi untuk mencapai tujuan tersebut.
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.
3. Data *Preparation*
   Dalam tahap ini meliputi semua kegiatan untuk membangun dataset akhir (data yang akan diprosespada tahap pemodelan (*modeling*) dari data mentah. Tahap ini dapat diulang beberapa kali. Pada tahap ini juga mencakup pemilihan *table*, *record*, dan atribut-atribut data, termasuk proses pembersihan dan transfarmasi data untuk kemudian dijadikan masukan dalam tahap pemodelan (*modeling*)
4. *Modeling*
   Dalam tahap ini melakukan pemilihan dan penerapan berbagai pemodelan dna beberapa parameternya akan disesuaikan untuk mendapatkan nilai yang optimal.
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*)

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 bnyak model yang telah dibuat (Larose, 1999)

3. Algoritma K-Nearest Neighbor

Algoritma *K-Nearest Neighbor* (KNN) merupakan sebuah metode untuk melakukan klasifikasi terhadap obyek baru berdasarkan (K) tetangga terdekatya (Grunescu, 2011). KNN termasuk algoritma *supervised learning*, yang mana hasil dari *query instance* baru, diklasifikasikan berdasarkan mayoritas dari kategori pada KNN. Kelas yang paling banyak muncul, yang akan menjadi kelas hasil klasifikasi.

KNN merupakan algoritma klasifikasi pertama diusulkan oleh T.M. Tutup dan P.E. Hart. Mengklasifikasi data karena kesederhanaannya, kemudahan implementasi dan efektivitas merupakan satu dari sepuluh besar penambangan data Algoritma yang sudah berkembang diterapkan berbagai bidang pengenalan pola, diagnosis kanker, klasifikasi teks dll.

Algoritma KNN digunakan untuk mengklasifikasi objek oleh mayoritas suara k benda referensi terdekat. Jadi, KNN terdiri dari dua proses yang memakan waktu: komputasi jarak dan peringkat jarak. Pekerjaan kami berfokus pada dua bagian ini.

Metode Pengukuran Jarak ALgoritma k-NN

Ada bebrapa metode pengukuran jarak pada algoritma k-NN yang akan digunakan untuk penelitian ini adalah(Ulya et al., 2021).

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 rumus berikut:

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

   dimana ,
   d = jarak antara x dan y
   x = data pusat klister
   y = data pada atribut
   i = setiap data
   n = jumlah data
   xi = data pada pusat klaster ke I
   yi = data pada setiap data ke i

2. *Manhattan distance*

Manhattan distance digunakan untuk menghitung perbedaan *absolut* (mutlak) antara koordinat sepasang objek. Rumus yang digunakan sebagai berikut:

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

dimana,
d = jarak antara x dan y
x = data pusat klister
y = data pada atribut
i = setiap data
n = jumlah data
xi = data pada pusat klaster ke i
yi = data pada setiap data ke i

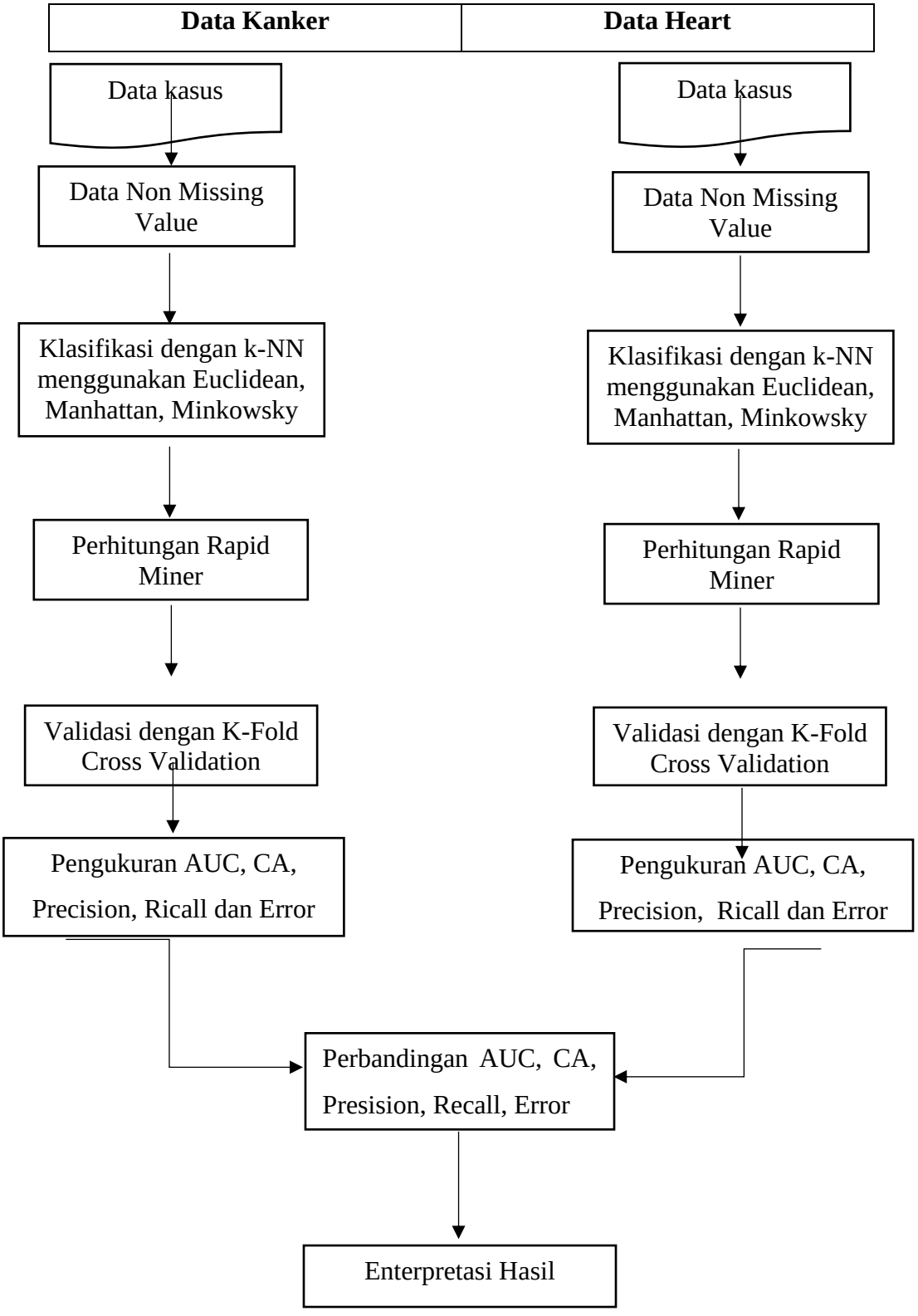
3. *Minkowski distance*

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 ∞. Berikut rumus yang digunakan untuk menghitung jarak dalam metode ini.

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

dimana,
d = jarak antara x dan y
x = data pusat klister
y = data pada atribut
i = setiap data
n = jumlah data
xi = data pada pusat klaster i
yi = data pada setiap data ke i
p = power

1.1 Perancangan Perbandingan Data



RANCANGAN PENELITIAN

Pada bab ini akan dilakukan pembuatan model klasifikasi yang dibutuhkan terkait dengan perancangan dalam penyeleksian perbandingan data yang menggunakan data *missing value* dan *non missing value* pada data kanker dan data *heart* dengan menggunakan *K-Nearest Neighbour*. Sebelum memulai proses klasifikasi, harus ditentukan terlebih dahulu variable data yang dibutuhkan.

Business Understanding

Pengaruh pengukuran jarak pada algortma k-NN. Data yang digunakan dalam penelitian ini merupakan data penyakit kanker dan data *heart*. Dataset kanker payudara memiliki 569 kasus dan data set *heart* memiliki 302 kasus. Data yang akan digunakan adalah *id, age, sex, radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, sp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal* digunakan sebagai atribut kelas. Metode data ini menggunakan metode pengukuran jarak algoritma *K-Nearest Neighbour* yaitu *Euclidean distance*, *Manhattan distance*, dan *Minkowski distance*.

Data Understanding

Keseluruhan data yang akan digunakan dalam penelitian ini adalah sebanyak 569 kasus data kanker dengan 12 atribut, dan 302 kasus data heart dengan 14 atribut. Atribut yang akan digunakan mempunyai tipikal *categorical* dan *numerik*.

Penyusunan basis data dilakukan dengan memasukkan data-data *riil* penelitian yang diperoleh data data kanker 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 kanker payudara, sebanyak 242 data training dan 60 data testing kasus *heart*.

2.1 Data *Preparation* atau Persiapan Data

Pada tahapan ini dilakukan pengecekan atau pencarian apakah terdapat data yang hilang (*missing value*) atau tidak.

Transformasi data dilakukan untuk mengubah data menjadi nilai dengan format tertenru. Seperti dalam atribut data yang *categorical* akan diubah menjadi data *numeric*. Untuk mendapatkan data yang berkualitas, ada beberapa tehnik *prepocesing* yang digunakan, yaitu data *validation*. Untuk atribut data yang *categorical* akan di konversi ke *numeric*. Untuk perhitungan menggunakan 80% data testing dan 20% data testing.

HASIL DAN PEMBAHASAN

Untuk menghitung tetangga terdekat dari data set diatas maka kita perlu melakukan, perhtiungan sebagai berikut:

1. Menentukan K dengan Parameter (Tetangga Terdekat)

Pada penelitian ini kita akan membandingkan K sebagai tetangga terdekat dengan K benilai ganjil. 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 dengan metode *Euclidean distance*

Tabel 4.1 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

Tabel 4.2 Dataset testing 1:

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 set 1:

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

$$= \mathbf{319{,}5482006}$$

Perhitungan *Manhattan distance* seluruh data training 1 dengan data set 1:

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

$$= \mathbf{347{,}77299}$$

Perhitungan *Minskowsky distance* seluruh data training 1 dengan data set 1:

$$\sqrt[4]{\begin{matrix}(17{,}99-20{,}59)^4+(10{,}38-21{,}24)^4+(122{,}8-137{,}8)^4+\\(1001-1320)^4+(0{,}1184-0{,}1085)^4+(0{,}2776-0{,}1644)^4+\\(0{,}3001-0{,}2188)^4+(0{,}1471-0{,}1121)^4+(0{,}2419-0{,}1848)^4\\+(0{,}07871-0{,}06222)^4\end{matrix}}$$

$$= \mathbf{2588841425}$$

Evaluation

Pada tahapan business understanding dijelaskan bahwa tujuan dari penelitian ini adalah mengimplementasikan metode pengukuran jarak algoritma KNN untuk

memprediksi penyakit kanker dengan perbandingan dua data yang berbeda. Selanjutnya dilakukan analisis metode pengujian terhadapt model-model yang bertujun untuk mendapatkan model yang paling akurat. Algoritma dengan hasil terbaik akan digunakan untuk proses selanjutnya.

Evaluasi Model

Model Evaluation yang digunakan adalah *K-fold Cross Validation* menggunakan *Widget Cross Validation RapidMinner*. Nilai yang dievaluasi adalah *AUC,Accurasy,Precission* dan *Recall* dari algoritma yang digunakan. Hasil *Cross Validation* paaada algoritma KNN. Data yang digunakan untuk metode *K-fold Cross Validation* berjumlah 512 data kanker payudara dan 274 data *heart*. Metode *K-fold Cross Validation* membagi data menjadi sejumlah segmen berdasarkan nilai *fold*.

Hasil dan Analisis

Analisis yang dilakukan terhadapat Perbandingan *Euclidean distance, Manhattan distance,* dan *Minskowky distance* memprediksi penyakit kanker payudara yang digunakan untuk model klasifikasi. Model klasifikasi yang paling bagus dari hasil klasifikasi yang dihasilkan dapat dilihat pada Tabel 5.8 berikut :

Tabel 5.3 Nilai K Paling Optimal

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

Pada tabel dapat kita lihat 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%.

2.2 Kesimpulan

Berdasarkan penelitian dan hasil pembahasan yang telah dilakukan, maka dapat disimpulkan bahwa:

Penggunaan data mining dengan data kanker untuk penelitian sebanyak 512 data training dan 57 data testing, dan data heart untuk penelitian sebanyak 274 data training dan 30 data testing. Berdasarkan data kanker dan data heart dapat diterapkan dalam melakukkan prediksi mana data yang memiliki akurasi yang tinggi. Perbandingan dari keduanya data yang memiliki nilai akurasi yang tinggi adalah data kanker yang akan digunakan untuk penelitian selanjunya.

Hasil evaluasi dengan menggunakan K-fold Cross Validation menghasilkan nilai CA (Clasification Accuracy) metode Euclidean distance memiliki akurasi 79.69% untuk data kanker dan 52.17% untuk data heart. Nilai Accuracy dengan metode *Manhattan distance* yaitu 85% untuk data kanker dan 53.85% untuk data heart. Nilai Accuracy dengan metode Minkowsky distance yaitu 97.92% untuk data kanker dan

52.46% untuk data heart. Nilai Precission metode Ecludean distance memiliki akurasi 100% untuk data kanker, dan 65.79% untuk data heart. Nilai Precission dengan metode *Manhattan distance* yaitu 100% untuk data kanker dan 68.57% untuk data heart. Nilai Precission dengan metode *Minkowsky distance* yaitu 80% untuk data kanker dan 40.00% untuk data heart. Nilai recall metode Ecludean distance memiliki akurasi 23.53% untuk data kanker, dan 69.44% untuk data heart. Nilai Accuracy dengan metode *Manhattan distance* yaitu 47.06% untuk data kanker dan 66.67% untuk data heart. Nilai recall dengan metode *Minkowsky distance* yaitu 94.12% untuk data kanker dan 66.67% untuk data heart. Dari hasil kasus tersebut dapat dikatakan bahwa metode Minkowsky distance memiliki kinerja akurasi yang bagus.

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 radius_mean dengan area_mean sedangkan pada penyakit Heart Disease yaitu gejala thalach dengan gejala slope.

2.3 Saran

Sesuai dengan hasil pengujian menunjukkan terdapat beberapa kekurangan hasil penelitian ini, oleh karena itu dapat disarankan beberapa hal untuk penelitian lebih lanjut.

1. Perlu ditambah data yang lebih banyak lagi supaya penelitian yang dilakukan memiliki tingkat akurasi yang lebih tinggi.
2. Hasil penelitian bisa ditindak lanjuti dengan diimplementasikan dengan metode klasifikasi seperti Decission Trees, Neural Networks, Support Vector Machines, atau Naïve Bayes untuk melihat perbandingan hasilnya.

Daftar Pustaka

AhmedMedjahed, S., Ait Saadi, T., & Benyettou, A. (2013). Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules. *International Journal of Computer Applications*, *62*(1), 1–5. https://doi.org/10.5120/10041-4635

Atmaja, D. M. U. (2019). *Penerapan Algoritma K-Nearest Neighbor Untuk*. *1*(November), 199–208.

Binabar, S. W., & Ivandari. (2017). Optimasi Parameter K pada Algoritma KNN untuk Deteksi Penyakit Kanker Payudara. *IC-Tech*, *XII*(2), 11–18.

Fayyad, U., Haussler, D., & Stolorz, P. (1996). KDD for Science Data Analysis: Issues and Examples. *Proceedings of the Second International Conference on*

Knowledge Discovery and Data Mining (KDD-96), 50–56.

Kusrini; Luthfi, E. T. (2009). *Algoritma Data Mining* (T. Ari Prabawati (ed.)). CV Andi Offset. https://books.google.co.id/books?id=-Ojclag73O8C&printsec=frontcover&hl=id#v=onepage&q&f=false

Larose, D. T. (1999). *An Introduction to Data Mining The CRISP-DM*.

Prahmana, I. G. (2020). Analisis Algoritma K-Nearets Neighbor Dengan Pemodelan Simple Linier Regression. *Tesis*, 1–76.

Rizki. (2003). Bab iii landasan teori 3.1. *Http://E-Journal.Uajy.Ac.Id/7244/4/3TF03686.Pdf*, *492*, 15–48.

Saputro, I. W., & Sari, B. W. (2020). Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa. *Creative Information Technology Journal*, *6*(1), 1. https://doi.org/10.24076/citec.2019v6i1.178

Susilowati, E., Hapsari, A. T., Efendi, M., & Edi, P. (2019). Diagnosa Penyakit Kanker Payudara Menggunakan Metode K - Means Clustering. *Jurnal Sistem Informasi, Teknologi Informatika Dan Komputer*, *10*(1), 27–32.

Ulya, S., Soeleman, M. A., & Budiman, F. (2021). Optimasi Parameter K Pada Algoritma K-NN Untuk Klasifikasi Prioritas Bantuan Pembangunan Desa. *Techno.Com*, *20*(1), 83–96. https://doi.org/10.33633/tc.v20i1.4215

Vulandari, S.Si., M.Si, R. T. (2017). *Data Mining Teori dan Aplikasi Rapidminer*. PENERBIT GAVA MEDIA.