Data Mining untuk Klasifikasi Penderita Kanker Payudara Menggunakan Algoritma K-Nearest Neighbor

Nadya Meilani[1], Odi Nurdiawan[2]

*Corresponding author: meilaninadya3@gmail.com

[1,2] Manajemen Informatika, STMIK IKMI Cirebon

Abstract

Cancer is a large group of diseases that can appear in almost any organ or tissue of the body when abnormal cells grow uncontrollably. Types of cancer are usually named according to the organ or tissue where the cancer is formed, one of which is breast cancer. Breast cancer is a type of cancer that forms in the breast tissue. Breast cancer is the most common type of cancer in Indonesian society. This cancer is the most common type of cancer in women. This type of cancer is one of the leading causes of death in Indonesia. Therefore, the problem in this study is the inaccuracy in the classification of breast cancer, which makes many people feel worried about the disease they are suffering from. The technique used is classification using the K-Nearest Neighbor algorithm. K-Nearest Neighbor is an algorithm for classifying objects with learning data that is closest to the object. The tool used to find the accuracy value is RapidMiner version 10. The method used in this study is the Knowledge Discovery in Databases (KDD) method. This research is used to make decisions based on the results obtained to determine policies to be taken in the management of breast cancer patients. The results show a fairly good percentage, namely producing an accuracy of 72.62% with a value of k = 5. In the cancer patient dataset, the prediction of no-recurrence-events (no recurrence) dominates

***Keywords**: K-Neares Neighbor Algorithm (KNN), Breast Cancer Analysis*

Abstrak

Kanker adalah sekelompok besar penyakit yang dapat muncul di hampir semua organ atau jaringan tubuh ketika sel-sel yang tidak normal tumbuh tak terkendali. Jenis kanker biasanya diberi nama sesuai organ atau jaringan tempat kanker itu terbentuk, salah satunya adalah kanker payudara. Kanker payudara adalah kanker yang terbentuk pada bagian jaringan payudara. Kanker payudara merupakan jenis kanker yang paling banyak diidap masyarakat Indonesia, kanker ini ialah jenis kanker yang paling umum diderita oleh kaum wanita. Kanker jenis ini menjadi salah satu penyebab kematian tertinggi di tanah air. Oleh karena itu permasalahan dalam penelitian ini adalah belum akuratnya ketepatan dalam klasifikasi kanker payudara sehingga banyak orang merasa khawatir akan penyakit yang dideritanya. Teknik yang digunakan adalah klasifikasi dengan menggunakan algoritma K-Nearest Neighbor. K-Nearest Neighbor adalah algoritma untuk melakukan klasifikasi terhadap objek dengan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Alat yang digunakan untuk mencari nilai akurasi adalah rapidminer versi 10. Metode yang digunakan dalam penelitian ini ini adalah metode Knowledge Discovery In

Database (KDD). Penelitian ini digunakan utuk mengambil keputusan berdasarkan hasil yang didapat untuk menentukan kebijakan yang akan diambil dalam penanganan pasien kanker payudara. Hasilnya menunjukan persentase yang cukup baik yaitu menghasilkan accurracy sebesar 72,62% dengan nilai k = 5, pada dataset pasien kanker prediksi no-recurrence-events (tidak kambuh) lebih mendominasi.

Kata Kunci : *Algoritma K-Neares Neighbor (KNN), Analisa Kanker Payudara*

Pendahuluan

Kanker payudara menempati urutan pertama dengan jumlah kasus kanker terbanyak di Indonesia , dan juga menjadi salah satu penyumbang kematian pertama akibat kanker (kemkes.go.id, 2022). Kanker payudara merupakan salah satu jenis kanker yang paling banyak diderita oleh masyrakat Indonesia terutama kaum wanita (Nugraha et al., 2019). Kanker merupakan suatu penyakit pertumbuhan sel yang tidak terbatas tidak terkendali pada organ tempat asal tumbuhnya, tetapi dapat menyebar ke organ-organ lain dalam tubuh. Penyakit ini tergolong pada tigkat berbahaya karena dapat menyebabkan kematian pada penderitanya (*What Is Breast Cancer? - National Breast Cancer Foundation*, n.d.). Menurut Global Burden of Cancer Study (Globocan) dan World Healt Organization (WHO) tahun 2020 mencapat bahwa kanker payudara memiliki jumlah kasus baru tertinggi di Indonesia sebesar 65.858 kasus atau 16,6% dari total 396.914 kasus kanker. Adapun jumlah kasus kematiannya mencapai lebih dari 22 ribu jiwa kasus (World Health Organization (WHO), n.d.).

Terdapat beberapa penelitian yang telah dilakukan dengan menggunakan dataset Kanker Payudara diantaranya menggunakan metode seperti *Naïve Bayes* (Wiratama Putra & Triayudi, 2022) *Neural Network* (Nugraha et al., 2019), Random Forest (Angkasa & Junifer Pangaribuan, 2022) dan Metode *Support Vector Machine* dengan *Backward Elimination* (Resmiati & Arifin, n.d.). Dikutip dari lama resmi CNN tahun 2018, jumlah kasus penderita kanker di seluruh dunia terus melambung secara signifikan. Laporan terbaru yang dirilis oleh *International Agency for Research on Cancer*, WHO (Organisasi Kesehatan Dunia), bahwa terdapat 18,1 juta kasus kanker baru dan 9,6 juta kematian yang terjadi pada tahun itu. Serangan kanker yang masif ini diperkirakan oleh WHO akan menjadi penyebab utama kematian nomor satu di dunia pada akhir abad ini. Hasil yang diperoleh dari analisa yang didapat setelah melakukan penelitian mereka menganalisis data dari 185 negara di dunia dengan fokus pada 36 jenis kanker (WHO: Kanker Membunuh Hampir 10 Juta Orang Di Dunia Tahun Ini, n.d.). Ada beberapa jenis kanker yang pernah disebutkan seperti kanker paru, kolorektal, lambung, hati dan juga payudara. Semua informasi yang disebutkan di atas dapat dimanfaatkan untuk mengetahui pola baru dari sebuah data dengan cara melakukan pengolahan atau penggalian data. Pola terbaru ini dapat digunakan untuk tujuan yang bermanfaat seperti pengklasifikasian data penderita penyakit kanker berdasarkan kambuh atau tidak kambuhnya penyakit tersebut. Berdasarkan hasil yang dipaparkan oleh WHO serta data yang diperoleh dari University Medical Center berupa dataset pasien kanker payudara dengan jumlah record data sebanyak 186 data, penulis melakukan penelitian secara ilmiah dengan metode yang

digunakan adalah klasifikasi algoritma *Naive Bayes* agar diketahui nilai perbandingan antara kambuh atau tidak kambuhnya penyakit kanker tersebut. Nantinya hasil dari klasifikasi ini nantinya dapat dimanfaatkan untuk menangani pasien guna mengurangi jumlah pasien yang sering mengalami kambuh penyakit kanker. Kesimpulan yang dapat ditarik dari hasil penelitian Ibnu Ramadhan, Kurniawati yaitu pasien pengidap penyakit kanker yang mengalami kambuh penyakit masih lebih besar dibandingkan dengn pasien yang tidak mengalami kambuh. Analisa yang dilakukan dengan menggunakan algoritma *Naive Bayes* menunjukan tingkat keakuratan dari perhitungannya yaitu cukup besar dapat dilihat dari persentase yang mencapai lebih dari 70%. Dengan nilai sebesar 71,43% bisa juga disebabkan oleh kurangnya kekompleksan data yang menyebabkan model dapat memprediksi secara akurat. Berdasarkan jurnal terdahulu yang dikutip dapat diketahui tingkat keakuratan algoritma Naïve Bayes menunjukan persentase yang cukup bersar maka dibuatlah penelitian mengenai ketepatan algoritma *K-Nearest Neighbor* pada dataset kanker payudara untuk mengetahui apakah algoritma K-NN dapat menunjukan tingkat keakuratan lebih tinggi daripada algoritma *Naïve Bayes* (Ramadhan & Kurniawati, 2020).

Data yang digunakan pada tugas akhir ini yaitu *Breast Cancer Dataset* yang merupakan data publik atau data sekunder yang bukan berasal langsung dari penelitian tetapi dari sumber lain, yaitu diambil dari situs web *University of California Irvine Machine Learning Repository* atau lebih dikenal dengan UCI Repository yang merupakan sekumpulan database, teori domain dan juga data generator yang biasanya digunakan oleh komunitas machine learning untuk analisis empiris dari algoritma machine learning. Adapun dataset *Breast Cancer* dapat diakses di alamat https://archive.ics.uci.edu/ml/datasets/Breast+Cancer.

Adapun permasalahan dalam penelitian ini adalah belum diketahuinya tingkat keakuratan dalam pengklasifikasian pasien kanker payudara yang sudah dinyatakan sembuh apakah pasien tersebut dapat terjangkit Kembali oleh kanker payudara atau tidak, hal ini dapat diketahui jika pasien kanker berada pada stadium 1, 2 dan bahkan stadium 3 sampai dengan 4 jika sudah dinyatakan sembuh dan bebas dari sel kanker akan dilakukan pengklasifikasian dan juga akan diketahui berapa tingkat keakuratan dari pengklasifikasian tersebut. Sehingga memudahkan dalam hal mendiagnosa pasien kanker payudara.

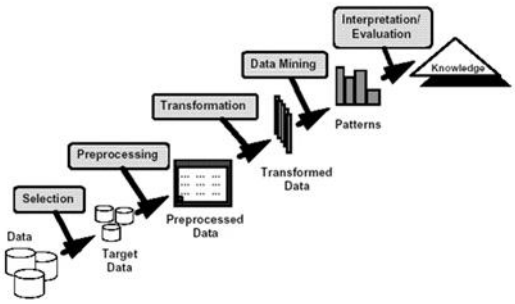
Tujuan utama dari penelitian ini adalah untuk mengetahui bagaimana pengklasifikasian kanker payudara dengan algoritma *K-Nearest Neighbor (K-NN)* dan berapakah tingkat akurasi dengan menggunakan algoritma *K-Nearest Neighbor (K-NN).* Tujuan pembuatan jurnal ini diharapkan dapat membantu pihak medis terutama dokter spesialis Onkologi dalam melakukan diagnosa kanker dengan tepat dan benar sehingga dapat dilakukan tindakan sesuai dengan diagnosa penyakit kanker payudara.

Metode Penelitian

A. Data Mining

Data mining merupakan suatu kegiatan untuk mengidentifikasi berbagai jenis data. Mulai dari banyaknya data yang tersimpan di database, data-data tersebut diidentifikasi dari kemungkinan adanya pola ataupun lainnya yang dianggap berpotensi untuk menghasilkan sesuatu yang dapat dipakai oleh perusahaan atau organisasi pemilik database tersebut. Data mining sendiri meiliki berbagai jenis metode yang dapat digunakan yaitu KDD, CRISP-DM, SEMMA, dll. Pada setiap prosesnya memiliki metode yang berbeda-beda dalam mencari informasi penting yang ada pada database orgnisasi. Pada penelitian ini akan digunakan metode *Knowledge Discovery in Database Process (KDD).*

Knowledge Discovery in Database Process (KDD) merupakan salah satu metode yang dapat digunakan dalam melakukan pemprosesan data atau data mining. Knowledge Discovery in Database Process sebagai proses dari menggunakan metode data mining untuk mencari informasi-informasi yang berguna dan berharga, pola yang ada pada data, yang melibatkan algoritma untuk mengidentifikasi suatu pola yang ada pada data. Proses *Knowledge Discovery in Database Process (KDD)* terdiri dari beberapa step, yaitu: seleksi data, pra-proses data, transformasi data, data mining, dan yang terakhir interpretasi dan evaluasi. Berikut adalah gambaran serta penjelasan mengenai proses data mining dengan metode *Knowledge Discovery in Database Process (KDD)* secara detail:



Gambar 1. Tahapan Proses KDD

1. Data Cleansing, Proses ini merupakan pengolahan data yang nantinya dipilihlah data yang dianggap bisa digunakan.
2. Data Integration, Proses menggabungkan data menjadi satu pada data yang dianggap berulang.
3. Selection, Proses seleksi atau pemilihan data yang dianggap relevan terhadap analisis.
4. Data Trasnformation, Proses transformasi data terpilih ke dalam bentuk mining procedure.
5. Data Mining, Proses dimana dilakukannya berbagai macam teknik untuk mengekstrak pola-pola potensial untuk menghasilkan data yang berguna.
6. Pattern Evolution, Proses dimana berbagai pola yang telah diidentifikasi berdasarkan measure yang telah diberikan.

7. Knowledge Presentation, Proses terakhir dari proses KDD, data-data yang sudah diproses divisualisasikan agar supaya mudah untuk dipahami oleh pengguna dan diharapkan dapat diambil keputusan berdasarkan (Andi Fahlevi, 2021) (Rinaldi Dikananda et al., n.d.).

B. Klasifikasi

Klasifikasi ialah proses menemukan model atau fungsi yang mendeskripsikan dan membedakan data ke dalam kelas-kelas. Klasifikasi melibatkan proses pemeriksaan karakteristik dari objek dan memasukkan objek ke dalam salah satu kelas yang sudah didefinisikan sebelumnya Algoritma *machine learning* terbagi menjadi dua, yaitu *supervised* dan *unsupervised learning*. Klasifikasi termasuk dalam algoritma supervised learning, artinya model yang bisa memprediksi keluaran, tujuannya adalah melatih model sehingga dapat memprediksi keluaran ketika diberikan data baru. Klasifikasi dalam data science adalah suatu proses memprediksi kelas atau kategori data dengan memanfaatkan nilai yang ada pada data. Proses klasifikasi pada dasarnya dilakukan agar analisis data menjadi lebih mudah dan juga memberikan hasil yang akurat.

C. Algoritma K-Nearest Neighbor (K-NN)

Algoritma K-Nearest Neighbor (K-NN) merupakan salah satu metode yang digunakan untuk mengelompokkan data. Prinsip kerja *K-Neares Neighbor (K-NN)* adalah mencari jarak terpendek antara data yang akan dievaluasi dengan knn (tetangga) terdekat dalam data latih. Pada fase pelatihan, algoritma hanya menyimpan vektor fitur untuk mengklasifikasikan data pelatihan. Klasifikasi vektor yang sama menghitung pada data pelatihan. Dalam vektor yang sama, kalsifikasi data uji dihitung. Jarak dari vektor terbaru ke semua vektor data pelatihan dihitung, dan sejumlah k terdekat diambil (Cahya, 2018) (Yuliarina & Hendry, 2022).

Hasil dan Pembahasan

A. Implementasi Klasifikasi Kanker Payudara Dengan Algoritma *K-Nearest Neighbor (K-NN)*

1) Data Selection

Pada tahap select data ini dilakukan pemilihan data yang akan digunakan dan kemudian akan dianalisis yang nantinya diharapkan akan memberikan keterangan penilaian pada data kanker payudara dengan kategori *recurrence-events* (kambuh) atau *no-recurrence-events* (tidak kambuh).

Tabel 1. Dataset Kanker Payudara

Class	***Age***	***Menopause***	***Tumor-size***	***Inv-nodes***	***Node-caps***	***Deg-malig***	***Breast***	***Breast-quad***	***Irradiant***
no-recurrence-events	30-39	premeno	30-34	0-2	no	3	left	left_low	no

no-recurrence-events	40-49	premeno	20-24	0-2	no	2	right	right_up	no
no-recurrence-events	40-49	premeno	20-24	0-2	no	2	left	left_low	no
no-recurrence-events	60-69	ge40	15-19	0-2	no	2	right	left_up	no
no-recurrence-events	40-49	premeno	0-4	0-2	no	2	right	right_low	no
no-recurrence-events	60-69	Ge40	15-19	0-2	no	2	left	left_low	no
no-recurrence-events	50-59	premeno	25-29	0-2	no	2	left	left_low	no

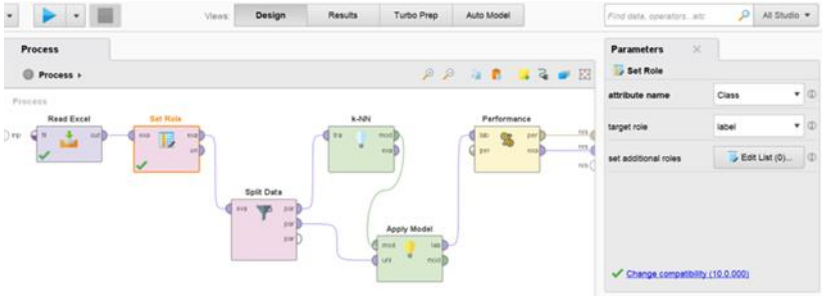
2) Data Preprocessing

Tabel 2. Pembersihan Data

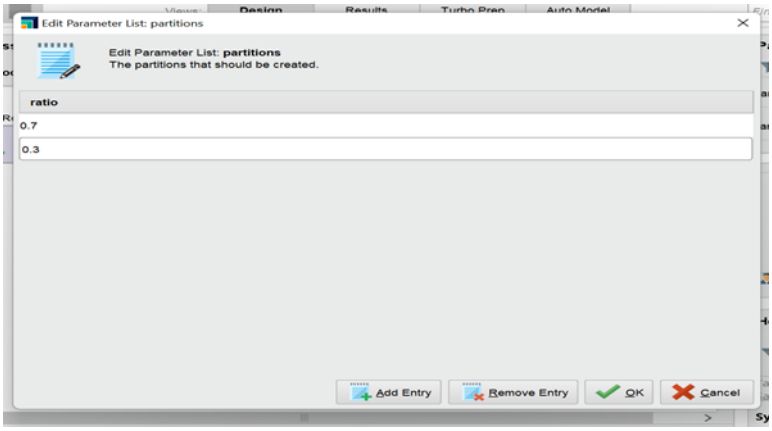
Inv-nodes	***Node-caps***	***Deg-malig***	***Breast***	***Breast-quad***	***Irradiant***
0-2	no	2	left	left_up	yes
3-5	?	1	left	left_up	yes
3-5	?	1	left	left_low	yes
3-5	2	2	left	left_up	yes

Pada tahap ini, dataset yang digunakan terdapat *missing value* sehingga pada tahapan ini dilakukan pembersihan data pada bagian atribut Node-caps tetapi bukan menggunakan operator Replace Missing Values melainkan secara manual dengan cara menghapus row cells yang terdapat tanda tanya pada excel.

3) Data Transformation



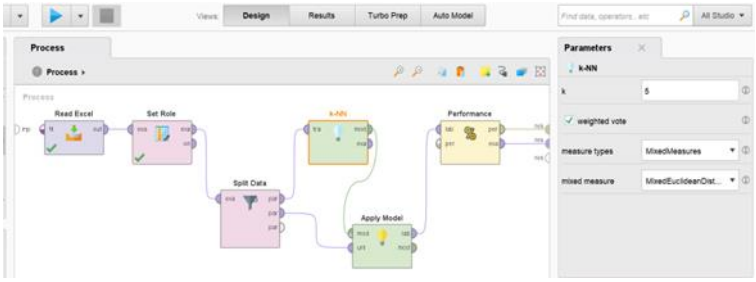
Gambar 2. Mengatur Target Role pada Operator Set Role



Gambar 3. Mengatur Parameter pada Operator Set Data

Pada tahap ini data akan diubah ke bentuk yang sesuai dengan data mining, dan nantinya akan diberikan atribut agar skala pengukuran data asli biasa diubah kedalam bentuk lain. Atribut yang digunakan adalah polynominal dan integer . Kemudian diatur role label pada atribut Class. Pada operator Split Data juga diatur agar dapat membagi dataset menjadi data trainning sebesar 70% dan data testing sebesar 30%.

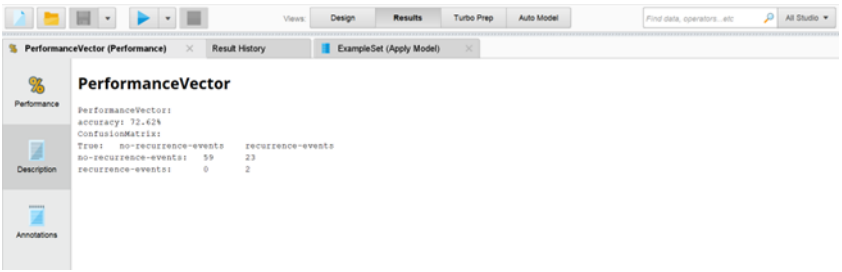
4) Data Mining



Gambar 4. Proses Data Mining

Pada tahap ini akan dijalankan proses rancang desain dengan menggunakan aplikasi Rapidminer dan diterapkan model klasifikasi *K-Nearest Neighbor* untuk menghasilkan pengklasifikasian dan tingkat akurasi performancenya.

5) Interpretation/Evaluation



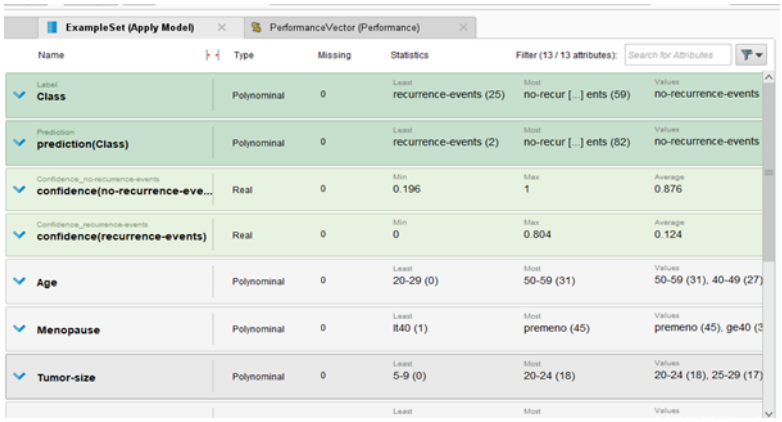
Gambar 5. Hasil Performance Vector

Hasil dari klasifikasi algoritma *K-Nearest Neighbor (K-NN)* dengan menggunakan dataset kanker payudara terdapat 279 sample data dengan 2 klasifikasinya dikategorikan kedalam *recurrence-events* (kambuh) atau *no-recurrence-events* (tidak kambuh).

B. Akurasi Algoritma K-NN

Hasilnya disini berisikan tentang hasil dari penelitian yang disesuaikan dengan tujuan penulisan jurnal dan mengacu pada tahapan perancangan.

1. Statistik Data

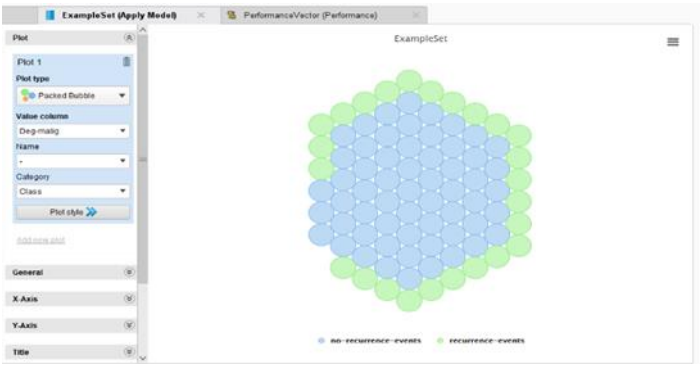


Gambar 6. Statistik Data Kanker

Berdasarkan Gambar 4. 5 dapat diketahui bahwa:

a. Data label klasifikasi dengan tipe polynominal terdapat 84 data, dengan rinciannya *recurrence-events* (kambuh) sebanyak 25 dan *no-recurrence-events* (tidak kambuh) sebanyak 59.
b. Data predikasi dengan tipe polynominal terdapar 84 data, dengan rinciannya *recurrence-events* (kambuh) sebanyak 2 dan *no-recurrence-events* (tidak kambuh)sebanyak 82.
c. Confidence *recurrence-events* (kambuh) dengan nilai terkecil 0, terbesar 0,804 dan rata-rata yang didapat adalah sebesar 0,124.
d. Confidence *no-recurrence-events* (tidak kambuh) dengan nilai tekecil 0,196 terbesar 1 dan rata-rata yang didapat adalah sebesar 0,876.

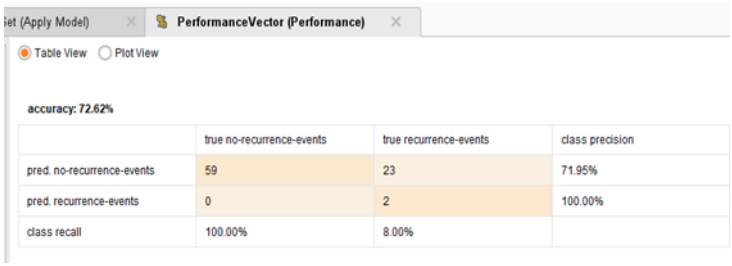
2. Visualisasi Data



Gambar 7. Visualisai Data dalam Bentuk Packed Bubble

Hasil dari visualisai data dengan plot type packed bubble menghasilkan example set apply model data bahwa *no-recurrence-events* (tidak kambuh) lebih mendominasi daripada *recurrence-events* (kambuh).

3. Nilai Akurasi Performance



	true no-recurrence-events	true recurrence-events	class precision
pred. no-recurrence-events	59	23	71.95%
pred. recurrence-events	0	2	100.00%
class recall	100.00%	8.00%	

Gambar 8. Akurasi Performace Vector

Hasil dari akurasi performance vector yaitu sebesar 72,62% dengan rincian class precission adalah pred *no-recurrence-events* (tidak kambuh) 71,95%, *pred recurrence-events* (kambuh) 100,00% dan rincian class *recal true no-recurrence-events* (tidak kambuh) sebesar 100,00% dan *recurrence-events* (kambuh) sebesar 8,00%.

Bagian ini tempat menuliskan hasil penelitian yang dijabarkan secara detail, jelas dan terurut. Hasil penelitian disajikan dalam bentuk tabel, grafik atau ilustrasi lain dan disertai dengan pembahasan yang disajikan secara terstruktur dan sistematis. Uraian performansi, kelemahan, dan kelebihan dari hasil penelitian harus dijelaskan.

Kesimpulan

Dapat disimpulkan bahwa dari data awal sebanyak 279 yang dibagi menjadi 30% data training dan 70% data testing dengan nilai K = 5 pengujian dengan model algoritma klasifikasi *K-Nearest Neighbor (KNN)* pada aplikasi Rapidminer versi 10 menghasilkan nilai accuracy sebesar 72,62% yang menunjukan bahwa hasil klasifikasinya baik, dimana prediksi pasien kanker *recurrence-events* (kambuh)

lebih sedikit dibandingkan dengan prediksi pasien kankker *no-recurrence-events* (tidak kambuh).

Daftar Pustaka

Andi Fahlevi. (2021). *Proses Data Mining KDD – School of Information Systems*. 30 Sep 2021. https://sis.binus.ac.id/2021/09/30/proses-data-mining-kdd/

Angkasa, V., & Junifer Pangaribuan, J. (2022). *KOMPARASI TINGKAT AKURASI RANDOM FOREST DAN KNN UNTUK MENDIAGNOSIS PENYAKIT KANKER PAYUDARA*.

Cahya. (2018). *Algoritma kNN*. https://extra.cahyadsn.com/knn

kemkes.go.id. (2022). *Kementerian Kesehatan Republik Indonesia*. Jakarta, 2 Februari 2022. https://www.kemkes.go.id/article/view/22020400002/kanker-payudara-paling-banyak-di-indonesia-kemenkes-targetkan-pemerataan-layanan-kesehatan.html

Nugraha, F. S., Shidiq, M. J., & Rahayu, S. (2019). ANALISIS ALGORITMA KLASIFIKASI NEURAL NETWORK UNTUK DIAGNOSIS PENYAKIT KANKER PAYUDARA. *Jurnal Pilar Nusa Mandiri*, *15*(2), 149–156. https://doi.org/10.33480/pilar.v15i2.601

Ramadhan, I., & Kurniawati, K. (2020). Data Mining untuk Klasifikasi Penderita Kanker Payudara Berdasarkan Data dari University Medical Center Menggunakan Algoritma Naïve Bayes. *JURIKOM (Jurnal Riset Komputer)*, *7*(1), 21. https://doi.org/10.30865/jurikom.v7i1.1755

Resmiati, R., & Arifin, T. (n.d.). *SISTEMASI: Jurnal Sistem Informasi Klasifikasi Pasien Kanker Payudara Menggunakan Metode Support Vector Machine dengan Backward Elimination*. http://sistemasi.ftik.unisi.ac.id

Rinaldi Dikananda, A., Wijaya, N. A., & Faqih, A. (n.d.). *IMPLEMENTASI DATA MINING PADA KETEPATAN PENGIRIMAN BARANG DENGAN MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBORS*. https://ejournal.stmikgici.ac.id/

What is Breast Cancer? - National Breast Cancer Foundation. (n.d.). Retrieved February 15, 2023, from https://www.nationalbreastcancer.org/what-is-breast-cancer/

WHO: Kanker Membunuh Hampir 10 Juta Orang di Dunia Tahun Ini. (n.d.). Retrieved February 25, 2023, from https://www.cnnindonesia.com/gaya-hidup/20180913133914-255-329910/who-kanker-membunuh-hampir-10-juta-orang-di-dunia-tahun-ini

Wiratama Putra, T., & Triayudi, A. (2022). Analisis Sentimen Pembelajaran Daring menggunakan Metode Naïve Bayes, KNN, dan Decision Tree. *Jurnal Teknologi Informasi Dan Komunikasi)*, *6*(1), 2022. https://doi.org/10.35870/jti

World Health Organization (WHO), (2020). (n.d.). *Ini Jenis Kanker yang Paling Banyak Diderita Penduduk Indonesia*.

Yuliarina, A. N., & Hendry, H. (2022). COMPARISON OF PREDICTION ANALYSIS OF GOFOOD SERVICE USERS USING THE KNN & NAIVE BAYES ALGORITHM WITH RAPIDMINER SOFTWARE. *Jurnal Teknik Informatika (Jutif)*, *3*(4), 847–856. https://doi.org/10.20884/1.jutif.2022.3.4.294