Analisis Performa *K-Nearest Neighbor*, *Naive Bayes*, *Decision Tree*, dan *Logistic Regression* Untuk Klasifikasi Kanker Payudara Menggunakan *Orange Data Mining*

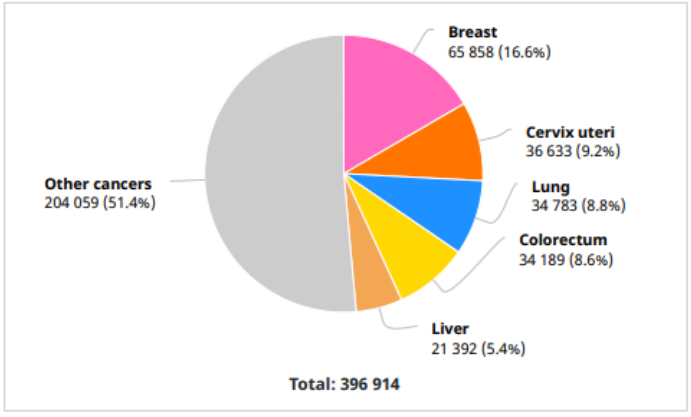
Angga Firmansyah
Teknik Informatika, Program Pascasarjana, Universitas Pamulang
e-mail: anggafirmansyah.it@gmail.com

***Abstrak*— Kanker payudara merupakan jenis kanker yang sering muncul di dalam sel-sel payudara, dan pertumbuhan sel kanker tersebut seringkali sulit untuk dikendalikan. Di Indonesia, kanker payudara menjadi kasus kanker yang paling umum, dan juga salah satu penyebab utama kematian. Oleh karena itu, sangat penting untuk mengklasifikasikan pasien yang menderita kanker payudara dan orang yang sehat sebagai langkah awal dalam mendeteksi penyakit ini lebih awal, yang dapat meningkatkan peluang kesembuhan pasien. Dalam penelitian ini, dilakukan evaluasi terhadap empat algoritma klasifikasi yang umum digunakan, yaitu *K-Nearest Neighbor (K-NN)*, *Naive Bayes*, *Decision Tree*, dan *Logistic Regression*, untuk mengklasifikasikan kanker payudara. Menggunakan perangkat lunak Orange Data Mining sebagai alat utama dalam implementasi dan percobaan. Data yang digunakan dalam penelitian ini berasal dari *Breast Cancer Coimbra* dari *UCI Machine Learning Repository*. Tujuan dari penelitian ini adalah untuk membandingkan kinerja masing-masing algoritma klasifikasi dalam pengenalan kanker payudara dan mengevaluasi tingkat akurasi mereka. Hasil eksperimen menunjukkan bahwa tiap algoritma memiliki keunggulan dan kelemahan tersendiri dalam hal akurasi, sensitivitas, dan spesifisitas. Analisis kinerja ini akan memberikan panduan yang berharga dalam pemilihan algoritma klasifikasi yang paling sesuai untuk tugas mengklasifikasikan kanker payudara.**

***Kata Kunci*— *K-Nearest Neighbor (K-NN)*; *Naive Bayes*; *Decision Tree*; *Logistic Regression*; Kanker Payudara.**

I. PENDAHULUAN

Kanker merupakan jenis penyakit tidak menular yang ditandai oleh pertumbuhan sel atau jaringan yang tidak normal, yang bersifat agresif dan tidak dapat dikendalikan, serta mampu mengganggu fungsi jaringan tertentu. Sel-sel kanker dapat berasal dari berbagai komponen dalam proses pembentukan organ, yang pada gilirannya dapat menyebabkan perkembangan massa tumor sebagai hasil dari proliferasi sel, dan mereka dapat menyebar melalui sistem peredaran darah atau sistem getah bening [1]. Kanker payudara termasuk dalam kategori penyakit kanker yang sangat serius. Menurut data yang disediakan oleh *Global Cancer Observatory* (GCO) yang dikeluarkan oleh *World Health Organization* (WHO) pada tahun 2020, ditemukan sebanyak 396.914 kasus penyakit kanker di Indonesia. Di antara jenis kanker tersebut, kanker payudara merupakan yang paling umum terjadi di negara ini, dengan angka sebanyak 65.858 kasus. Persentase ini setara dengan sekitar 16,6% dari total kasus penyakit kanker yang terjadi di Indonesia [2]. Kanker payudara merupakan suatu jenis tumor ganas yang muncul di dalam sel-sel payudara. Kejadian kanker payudara terjadi ketika sel-sel di dalam jaringan payudara mengalami pertumbuhan yang tidak terkendali, dan hal ini dapat mengganggu jaringan sehat yang ada di sekitarnya [3][4].



Gambar 1.
Data WHO *Global Cancer Observatory* 2020 Indonesia [2]

Berdasarkan informasi yang diterbitkan dalam makalah CA yang berjudul "*A Cancer Journal for Clinicians*" pada tahun 2020, terungkap bahwa kanker payudara pada wanita merupakan masalah kesehatan serius dengan jumlah kasus sebanyak 2,3 juta atau 11,7% dari total kasus kanker baru yang terdiagnosis. Lebih mengejutkan lagi, kanker payudara memiliki tingkat diagnosis yang lebih tinggi daripada kanker paru-paru, yang hanya mencapai 11,5%. Hyuna Sung, seorang ilmuwan terkemuka dan ahli epidemiologi di *American Cancer Society*, mengungkapkan bahwa kanker payudara menduduki peringkat teratas dalam hal jumlah kasus penyakit kanker, mencapai 2,3 juta kasus pada tahun 2020, naik dari 2.088.849 kasus pada tahun 2018 [5]. Kanker payudara merupakan salah satu jenis kanker yang sangat berbahaya, seperti yang terdokumentasikan dalam data *Global Cancer Observatory* pada tahun 2018. Data tersebut mencantumkan bahwa di Indonesia, kanker payudara menduduki peringkat kedua dalam hal jumlah kematian pasien kanker. Kanker payudara menyebabkan 22.692 kasus kematian, yang merupakan sekitar 11% dari total 207.210 kasus kematian yang tercatat pada tahun tersebut [6].

Kanker payudara merupakan salah satu isu serius dalam bidang kesehatan yang mempengaruhi populasi wanita di seluruh dunia. Pengidentifikasian awal dan klasifikasi yang akurat terhadap kanker payudara memiliki peran penting dalam meningkatkan tingkat keselamatan serta kesejahteraan pasien. Penerapan teknologi dalam pengolahan dan analisis data telah memainkan peran yang semakin vital dalam hal ini. Salah satu pendekatan efektif yang digunakan adalah penggunaan algoritma klasifikasi dalam menganalisis data medis untuk mengklasifikasikan kasus kanker payudara. Sebuah pembandingan yang komprehensif mengenai kinerja algoritma seperti *K-Nearest Neighbor (K-NN)*, *Naive Bayes*, *Decision Tree*, dan *Logistic Regression* dalam konteks klasifikasi kanker payudara belum selalu tersedia. Oleh karena itu, penelitian ini bertujuan untuk memberikan pemahaman yang lebih mendalam mengenai kinerja algoritma-algoritma ini dalam tugas klasifikasi kanker payudara.

Dalam konteks perkembangan teknologi, perangkat lunak *Orange Data Mining* telah menjadi alat yang populer dan bermanfaat dalam proses pengolahan data serta analisis. Penggunaannya dalam penelitian medis telah menjadi topik yang menarik, dan penelitian ini ingin mengeksplorasi sejauh mana efektivitasnya dalam konteks klasifikasi kanker payudara. Harapannya, penelitian ini dapat memberikan sumbangan yang berarti dalam pemilihan algoritma yang paling cocok untuk mengklasifikasikan kanker payudara berdasarkan data medis. Hasil penelitian ini diharapkan dapat menjadi panduan berharga bagi praktisi medis dan peneliti di bidang ini untuk meningkatkan efektivitas dalam deteksi dini dan manajemen kasus kanker payudara. Oleh karena itu, penelitian ini memiliki relevansi yang signifikan dalam upaya pencegahan dan penanganan penyakit yang berdampak besar pada banyak individu, khususnya wanita di seluruh dunia.

A. Dasar Teori

Dasar teori dalam penelitian ini merupakan bagian yang menjelaskan konsep-konsep, literatur terkait, dan teori-teori yang relevan terhadap topik penelitian. Ini memberikan landasan teoritis bagi penelitian dan membantu dalam pemahaman konteks serta relevansi dari penelitian.

1) Data Mining

Data mining merupakan tahap penggunaan teknologi, metode statistik, dan matematika guna menyaring data dengan tujuan menemukan relasi, pola, dan tren yang inovatif dan penting [7]. *Data mining* bisa juga diartikan sebagai langkah eksplorasi dan analisis data yang dilakukan dengan metode otomatis atau semi-otomatis pada *dataset* yang besar, dengan tujuan mengidentifikasi pola dan peraturan yang memiliki signifikansi [8]. Dalam konteks ini, data mining dapat digambarkan sebagai upaya untuk menemukan pola yang ada dalam *dataset* dengan memanfaatkan metode statistik dan matematika.

2) K-Nearest Neighbor (K-NN)

K-Nearest Neighbor (K-NN) adalah suatu algoritma pembelajaran semi-*supervised* yang memerlukan data latihan dan nilai k yang telah ditetapkan sebelumnya [9]. Dalam KNN, k merupakan jumlah tetangga yang diambil dalam membuat sebuah keputusan [10]. Algoritma KNN memiliki prinsip kerja yaitu mencari jarak tetangga terdekat antara data yang akan dievaluasi dengan data latih [11]. Berikut merupakan persamaan dalam menentukan jarak Euclidean pada *K-Nearest Neighbors*.

Formula *Euclidean* [12]:

$$Euclidean = \sqrt{\sum_{i=1}^{p} (X_{2i} - X_{1i})^2} \qquad (1)$$

Keterangan :

p = dimensi data

X_1 = Data train

X_2 = Data test

3) Naive Bayes

Naive Bayes adalah algoritma yang berbasis pada konsep probabilitas untuk mengembangkan model prediksi dalam tugas klasifikasi [13]. *Naive Bayes* merupakan salah satu algoritma yang metode pembelajarannya bersifat *supervised*, dimana pada saat proses pembelajaran dibutuhkan data latih untuk dapat mengambil keputusan. Nilai probabilitas dari setiap kelas target yang ada akan dihitung terhadap input yang diberikan pada tahap klasifikasi. Kelas target yang memiliki probabilitas paling besarlah yang menjadi kelas pada data inputan tersebut [14]. Keunggulan dari *Naive Bayes* adalah cepat dan efektif dalam mengolah data dalam jumlah besar [13]. Di bawah ini merupakan formula persamaan *Naive Bayes*:

Formula *Naive Bayes* [13] :

$$\text{P (J|K)} = \frac{P(K|J)P(J)}{P(K)} = \frac{P(J \cap 1)}{P(K)} \quad (2)$$

Keterangan :
P(J|K) = Probabilitas J terjadi bila K terjadi
P(K|J) = Probabilitas K terjadi bila J terjadi
P(J) = Probabilitas J terjadi
P(K) = Probabilitas K terjadi
P(J∩K) = Probabilitas P(J) dan P(K) terjadi secara bersamaan

4) Decision Tree

Decision Tree merupakan algoritma yang digunakan untuk membentuk model keputusan dengan memanfaatkan struktur berbentuk pohon atau hierarki [15]. Pohon pada decision tree memiliki root node dan node, root node merupakan puncak dari pohon sedangkan node merupakan percabangan dari root node itu sendiri. Pada setiap node decision tree terdapat proses pembuatan keputusan yang menghasilkan dua cabang yaitu "ya" atau "tidak", pembuatan keputusan sendiri dilakukan dengan menguji suatu variabel, proses pengujian ini terus berlanjut hingga node paling bawah atau disebut dengan leaf node [13]. Dalam proses *decision tree* hal yang paling pertama dilakukan adalah memilih root node, salah satu cara dalam pemilihan root node ini dapat dilakukan dengan menghitung nilai gain pada setiap atribut, sebelum melakukan perhitungan pada nilai gain perlu dilakukan perhitungan pada nilai entropy terlebih dahulu. Berikut formula *entropy* pada algoritma *decision tree*:

Formula *Entropy* [13]:

$$\text{Entropy(S)} = \sum_{i=1}^{n} P_i * \log_2(P_i) \quad (3)$$

Keterangan :
n = Jumlah kelas S
P = Proporsi nilai-nilai masuk ke dalam kelas di tingkat i

5) Logistic Regression

Logistic regression merupakan metode analisis yang sesuai ketika variabel terikat memiliki dua kemungkinan nilainya (biner). Regresi Logistik digunakan untuk memodelkan data dan menguraikan hubungan antara satu variabel terikat biner dengan satu atau lebih variabel bebas yang dapat berupa nominal, ordinal, interval, atau rasio [16]. Model persamaan aljabar layaknya OLS yang biasa digunakan adalah sebagai berikut: Y = B0 + B1X + e. Dimana e adalah error varians atau residual. Dengan model regresi ini, tidak menggunakan interpretasi yang sama seperti halnya persamaan regresi OLS. Model Persamaan yang terbentuk berbeda dengan persamaan OLS [17].

Formula *Logistic Regression* [17]:

$$In\left(\frac{\hat{p}}{1-\hat{p}}\right) = B_0 + B_1X \quad (4)$$

Keterangan :
In = logaritma natural
B_0 = konstanta
B_1 = koefisien masing-masing variable
X = variable independen
$\hat{p}$= probabilitas logistik yang dirumuskan sebagai berikut :

$$\hat{p} = \frac{exp(B_0 + B_1X)}{1 + exp(B_0 + B_1X)} = \frac{e^{B_0+B_1X}}{1 + e^{B_0+B_1X}} \quad (5)$$

Keterangan :
exp atau e : fungsi exponen

6) Elemen Pengukuran

Confusion Matrix merupakan suatu tabel yang berisi empat kombinasi berbeda antara prediksi dan nilai sebenarnya. Berdasarkan tabel 1 dapat dijelaskan bahwa terdapat 4 representasi hasil proses klasifikasi pada *confusion matrix* ialah *True Positif* (TP), *True Negatif* (TN), *False Positif* (FP), *serta False Negatif* (FN) [18], [19]. *Confusion matrix* pula kerap disebut *error* matrix.

Tabel 1.
Pengujian *Confusion Matrix*

Classification	***Predicted Class***	
	True	***False***
Actual: True	*True Positif (TP)*	*False Negatif (FN)*
Actual: False	*False Positif (FP)*	*True Negatif (TN)*

Selanjutnya, *accuracy* adalah sejauh mana hubungan antara nilai yang diperkirakan dan nilai sebenarnya. *Precision* adalah sejauh mana tingkat kesesuaian antara informasi yang diinginkan oleh pengguna dan respons yang diberikan oleh sistem. Sementara itu, recall adalah sejauh mana tingkat kemampuan sistem untuk mengingat kembali data tertentu [20], rumus yang digunakan sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

II. METODE PENELITIAN

Metodologi penelitian ini akan memberikan panduan tentang bagaimana penelitian dilakukan, dari pemilihan atribut hingga pengujian dan evaluasi model klasifikasi. Ini adalah bagian yang kunci untuk memahami langkah-langkah yang diperlukan dalam penelitian ini dan hasil yang diharapkan.

A. Pengumpulan Data

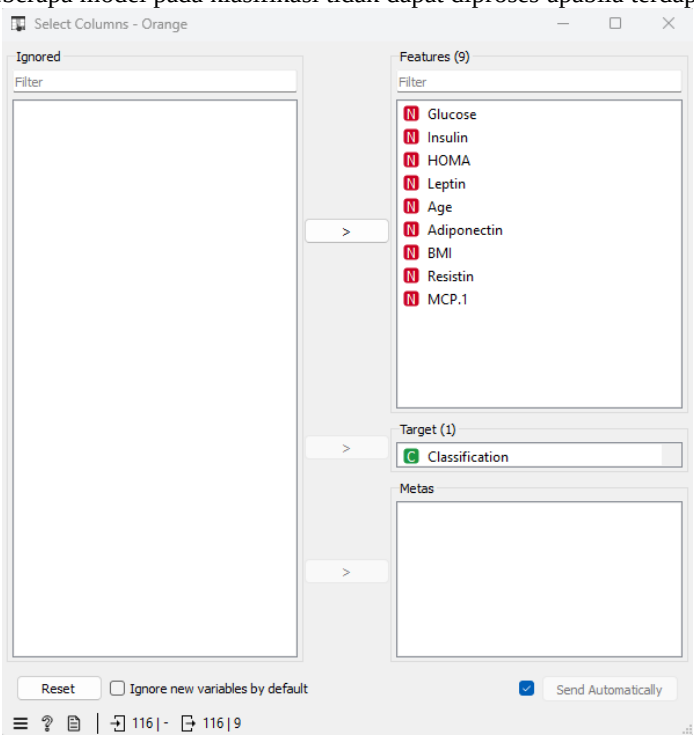
Pada tahap ini, menjelaskan atribut-atribut yang digunakan dalam penelitian ini. Atribut-atribut ini akan mencakup data medis yang berkaitan dengan kanker payudara. Atribut yang relevan yang digunakan untuk mengidentifikasi dan mengklasifikasikan kanker payudara. Data acuan merupakan *Breast Cancer Coimbra* dari *UCI Machine Learning Repository*. *Dataset* tersebut terdiri dari total seratus enam belas data dengan sembian variabel prediktor dan satu variabel kelas seperti tabel 2.

Tabel 2.
Atribut Data Penelitian

No	Atribut	Nilai	Keterangan
1	*Age*	Angka numerik	Umur pasien (tahun)
2	*BMI*	Angka numerik	Body Mass Index (berat badan) pasien (kg/m^2)
3	*Glucose*	Angka numerik	Kadar gula dalam tubuh pasien (mg/dL)
4	*Insulin*	Angka numerik	Kadar insulin (hormon polipeptida) dalam tubuh pasien (µU/mL)
5	*HOMA*	Angka numerik	Pengukuran resistensi insulin dan fungsi sel beta dalam tubuh pasien
6	*Leptin*	Angka numerik	Kadar leptin (hormon yang dibuat sel lemak) dalam tubuh pasien (ng/mL)
7	*Adiponectin*	Angka numerik	Kadar adiponectin (hormon protein dan adipokin) dalam tubuh pasien(µg/mL)
8	*Resistin*	Angka numerik	Kadar resistin (protein kaya asam amino) pada tubuh pasien (ng/mL)
9	*MCP.1*	Angka numerik	Kadar MCP.1 pada tubuh (pg/dL)
10	*Classification*	Angka numerik	Label menunjukkan seseorang sehat atau sakit kanker yaitu 1: sehat 2 : sakit kanker

B. Preprocessing

Bagian ini akan membahas langkah-langkah yang diambil dalam pemilihan dan pra-pemrosesan data. Akan mencakup pengumpulan data medis yang relevan dan memastikan kualitas dan kebersihan data. Nilai yang hilang dalam data *instance* akan menggangu proses klasifikasi. Beberapa model pada klasifikasi tidak dapat diproses apabila terdapat data dan nilai yang hilang.

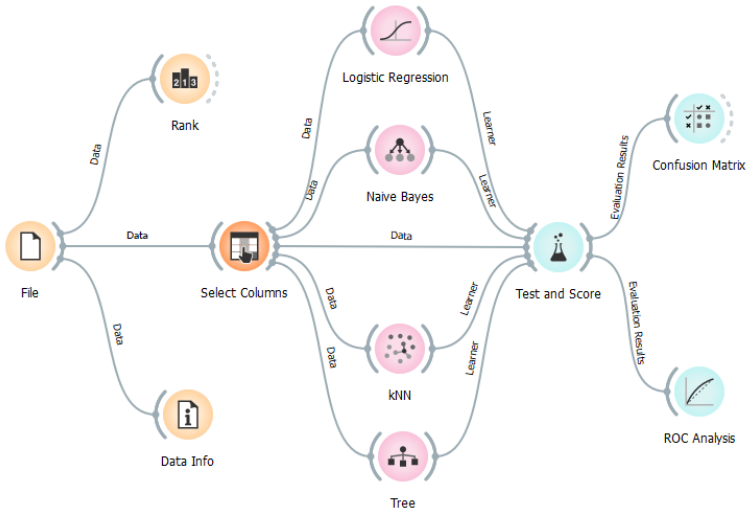


Gambar 2.
Proses Pemilihan Variabel

Berdasarkan gambar 2 merupakan proses pemilihan data menggunakan *widget*, pertama pilih kolom atribut yang akan digunakan ke *features* terdiri dari sembian variabel prediktor yaitu *Age*, *BMI*, *Glucose*, *Insulin*, *HOMA*, *Leptin*, *Adiponectin*, *Resistin*, *MCP.1* dan satu variabel kelas yaitu *Classification* ke *target*.

C. Modeling

Proses penggunaan algoritma *K-Nearest Neighbor (K-NN)*, *Naive Bayes*, *Decision Tree*, dan *Logistic Regression* untuk melakukan klasifikasi kanker payudara. Akan mencakup implementasi algoritma tersebut menggunakan perangkat lunak *Orange Data Mining*, dilakukan perbandingan beberapa metode *data mining* untuk memilih metode yang terbaik dengan akurasi yang tinggi, dalam mengklasifikasi *dataset* status kanker payudara seperti terlihat pada gambar 3.



Gambar 3.
Proses *Modeling Orange Data Mining*

Pada gambar 3 merupakan desain widget yang sudah terdapat algoritma *K-Nearest Neighbor (K-NN)*, *Naive Bayes*, *Decision Tree*, dan *Logistic Regression*. Lengkap dengan widget *Test and Score* untuk menghasilkan nilai akurasi, presisi, *recall* dan AUC, dari setiap metode. Selanjutnya evaluasi hasil dengan widget *Confusion Matrix* dan *ROC Analysis* untuk membandingkan performa ketiga algoritma.

D. Evaluasi & Analisis Model

Analisis hasil dilakukan dengan cara membandingkan nilai akurasi, presisi, recall dan AUC, dari setiap metode. Nilai akurasi, presisi, recall didapat dari *Confusion Matrix*. Nilai AUC (*Area Under the Curve*) diperoleh dari hasil pengukuran kinerja model klasifikasi, khususnya ketika menggunakan metrik ROC (*Receiver Operating Characteristic*). AUC mengukur sejauh mana model mampu membedakan antara dua kelas yang berbeda (misalnya, kelas positif dan negatif), klasifikasi kategori kelas AUC dapat dilihat pada Tabel 3.

Tabel 3.
Klasifikasi Kategori AUC

Rentang Kelas	Kategori
0.90 – 1.00	*Excellent classification*
0.80 – 0.90	*Good classification*
0.70 – 0.80	*Fair classification*
0.60 – 0.70	*Poor classification*
0.50 – 0.60	*Failure*

III. HASIL DAN PEMBAHASAN

A. Analisis Distribusi Data & Missing Value

Analisis distribusi data bertujuan untuk mengetahui keseimbangan jumlah kelas yang terindikasi sakit kanker dan sehat. Distribusi kelas data *Breast Cancer Coimbra*, dapat dilihat pada tabel 4.

Tabel 4.
Analisis Distribusi Data

No.	Kelas	Jumlah
1	1 (Sehat)	52
2	2 (Sakit Kanker)	64

Analisis distribusi data adalah proses untuk memahami bagaimana data yang dimiliki tersebar atau terdistribusi. Tujuan utama dari analisis distribusi data adalah untuk mendapatkan wawasan tentang karakteristik data yang akan digunakan dalam penelitian. Selanjutnya, analisis atribut *missing value* adalah proses untuk mengidentifikasi dan memahami atribut yang hilang atau kosong dalam *dataset*. Jumlah *missing value* dalam setiap atribut data *Breast Cancer Coimbra* dapat dilihat pada tabel 5.

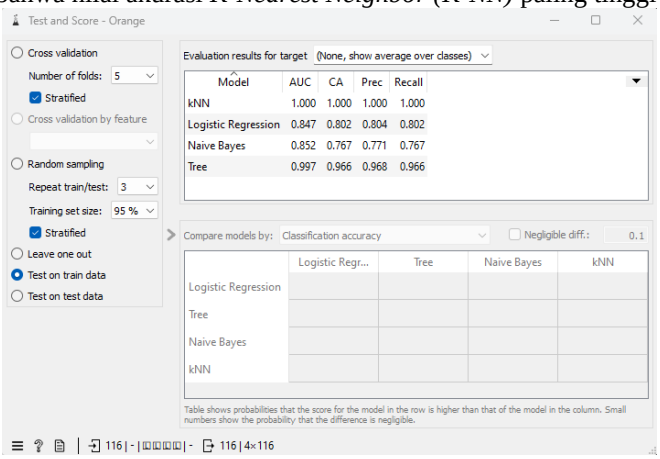
Tabel 5.
Analisis Atribut *Missing Value*

No.	Atribut	Jumlah
1	*Age*	0
2	*BMI*	0
3	*Glucose*	0
4	*Insulin*	0
5	*HOMA*	0
6	*Leptin*	0
7	*Adiponectin*	0
8	*Resistin*	0
9	*MCP.1*	0
10	*Classification*	0

Analisis atribut *missing value* penting karena data yang hilang dapat memengaruhi keakuratan dan validitas penelitian. Oleh karena itu, peneliti perlu menentukan cara yang tepat untuk mengatasi data yang hilang agar hasil penelitian tetap dapat diandalkan.

B. Hasil Simulasi 4 Model Klasifikasi

Hasil dari simulasi menggunakan empat model klasifikasi *K-Nearest Neighbor (K-NN)*, *Naive Bayes*, *Decision Tree*, dan *Logistic Regression* akan dibahas. Akan mencakup hasil prediksi untuk data pengujian dan tingkat akurasi dari masing-masing model. Berdasarkan data yang sudah dilakukan pengujian menggunakan sembian variabel prediktor dan satu variabel kelas. Perhitungan meliputi 4 skenario pengujian untuk menentukan nilai akurasi, presisi, recall dan AUC tertinggi dari masing-masing model seperti terlihat pada gambar 4. Hasil klasifikasi model *K-Nearest Neighbor (K-NN)*, *Naive Bayes*, *Decision Tree*, dan *Logistic Regression* menunjukkan bahwa nilai akurasi *K-Nearest Neighbor (K-NN)* paling tinggi yaitu 100%.

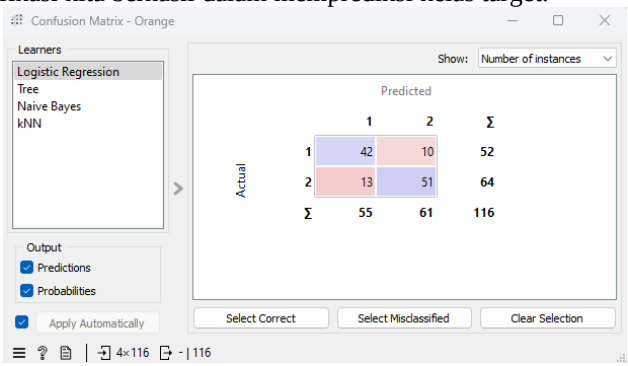


Gambar 4.
Hasil Pengujian Model

Berdasarkan gambar 4 juga memperlihatkan perbandingan 4 model AUC, diketahui bahwa nilai AUC yang paling tinggi adalah metode *K-Nearest Neighbor (K-NN)* yaitu 1.000. AUC merupakan salah satu metrik evaluasi yang umum digunakan dalam *data mining*, terutama dalam konteks evaluasi model klasifikasi. AUC mengukur kualitas dan performa suatu model klasifikasi dalam membedakan antara kelas positif dan negatif. Semakin besar AUC, semakin baik hasil klasifikasi yang digunakan.

C. Hasil Evaluasi Confusion Matrix

Confusion matrix adalah alat penting dalam evaluasi kinerja model klasifikasi dalam machine learning dan data mining. *Confusion matrix* adalah tabel matriks yang digunakan untuk menggambarkan hasil dari proses klasifikasi, yang mencakup empat elemen utama, yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Ini membantu kita memahami sejauh mana model klasifikasi kita berhasil dalam memprediksi kelas target.



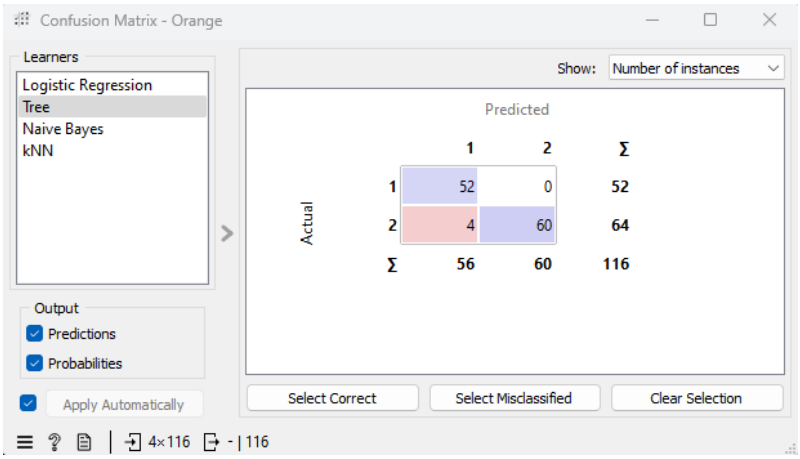
Gambar 5.
Nilai *Confusion Matrix Logistic Regression*

Pada gambar 5 menunjukkan bahwa nilai dari *True Positif* (TP) adalah 42, *True Negatif* (TN) adalah 51, *False Positif* (FP) adalah 13, dan *False Negatif* (FN) adalah 10. Maka nilai *Accuracy*, *Precision* dan *Recall* dari metode *Logistic Regression* adalah sebagai berikut:

$$Accuracy = \frac{(42\ +\ 51)}{(42\ +\ 51\ +\ 13\ +\ 10)} x100\% = 80\%$$

$$Precision = \frac{(42)}{(42\ +\ 13)} x100\% = 76\%$$

$$Recall = \frac{(42)}{(42\ +\ 10)} x100\% = 81\%$$



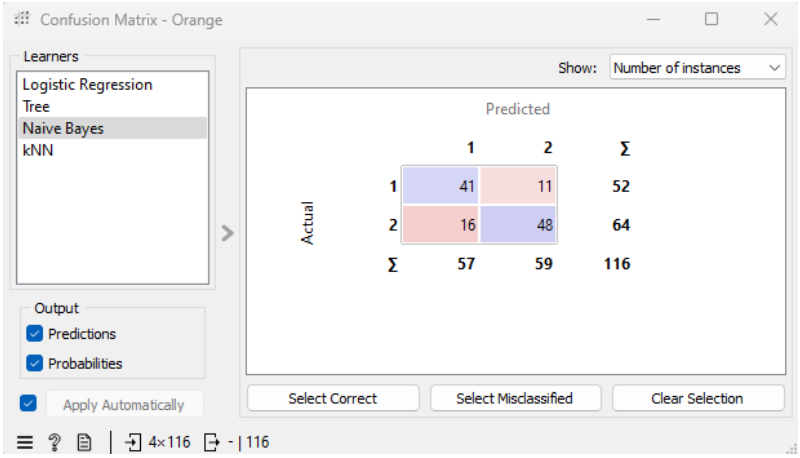
Gambar 6.
Nilai *Confusion Matrix Decision Tree*

Pada gambar 6 menunjukkan bahwa nilai dari *True Positif* (TP) adalah 52, *True Negatif* (TN) adalah 60, *False Positif* (FP) adalah 4, dan *False Negatif* (FN) adalah 0. Maka nilai *Accuracy*, *Precision* dan *Recall* dari metode *Decision Tree* adalah sebagai berikut:

$$Accuracy = \frac{(52\ +\ 60)}{(52\ +\ 60\ +\ 4\ +\ 0)} x100\% = 97\%$$

$$Precision = \frac{(52)}{(52\ +\ 4)} x100\% = 93\%$$

$$Recall = \frac{(52)}{(52\ +\ 0)} x100\% = 100\%$$



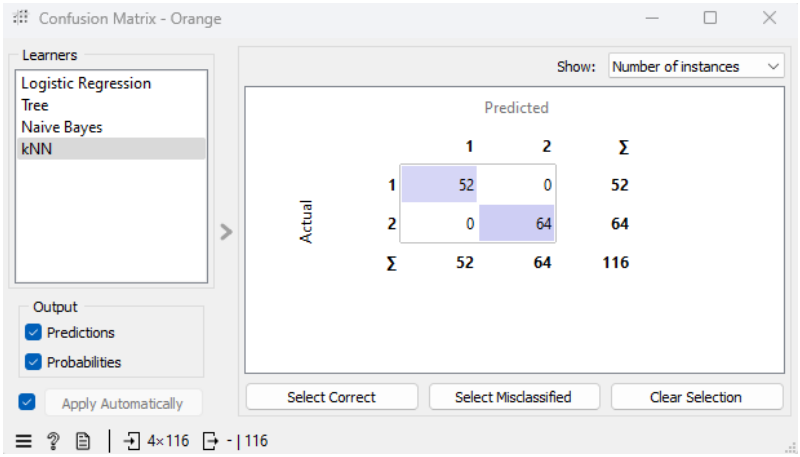
Gambar 7.
Nilai *Confusion Matrix Naive Bayes*

Pada gambar 7 menunjukkan bahwa nilai dari *True Positif* (TP) adalah 41, *True Negatif* (TN) adalah 48, *False Positif* (FP) adalah 16, dan *False Negatif* (FN) adalah 11. Maka nilai *Accuracy*, *Precision* dan *Recall* dari metode *Naive Bayes* adalah sebagai berikut:

$$Accuracy = \frac{(41 + 48)}{(41 + 48 + 16 + 11)} x100\% = 77\%$$

$$Precision = \frac{(41)}{(41 + 16)} x100\% = 72\%$$

$$Recall = \frac{(41)}{(41 + 11)} x100\% = 79\%$$



Gambar 8.
Nilai *Confusion Matrix K-Nearest Neighbor (K-NN)*

Pada gambar 8 menunjukkan bahwa nilai dari *True Positif* (TP) adalah 52, *True Negatif* (TN) adalah 64, *False Positif* (FP) adalah 0, dan *False Negatif* (FN) adalah 0. Maka nilai *Accuracy*, *Precision* dan *Recall* dari metode *K-Nearest Neighbor (K-NN)* adalah sebagai berikut:

$$Accuracy = \frac{(52 + 64)}{(52 + 64 + 0 + 0)} x100\% = 100\%$$

$$Precision = \frac{(52)}{(52 + 0)} x100\% = 100\%$$

$$Recall = \frac{(52)}{(52 + 0)} x100\% = 100\%$$

Berdasarkan hasil evaluasi dan validasi dengan menggunakan *Confusion Matrix* diperoleh nilai perbandingan *Accuracy*, *Precision* dan *Recall* dari 4 metode *K-Nearest Neighbor (K-NN)*, *Naive Bayes*, *Decision Tree*, dan *Logistic Regression* seperti terlihat pada tabel 6.

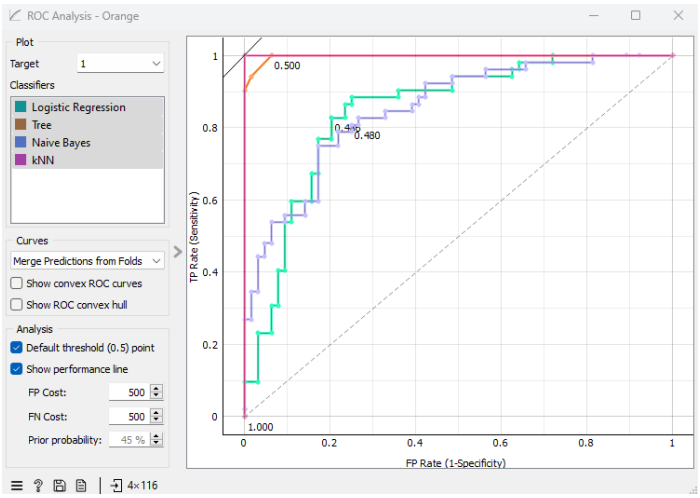
Tabel 6.
Perbandingan Kinerja

Metode	*Accuracy*	*Precision*	*Recall*
Logistic Regression	80%	76%	81%
Decision Tree	97%	93%	100%
Naive Bayes	77%	72%	79%
K-Nearest Neighbor (K-NN)	100%	100%	100%

Berdasarkan tabel 6 dapat diketahui bahwa kinerja dari model *K-Nearest Neighbor (K-NN)* lebih baik dari model *Naive Bayes*, *Decision Tree*, dan *Logistic Regression*. Akurasi klasifikasi tidak bisa mencapai hasil yang sempurna karena pasti ada nilai *error*. Hal tersebut dipengaruhi oleh banyaknya data uji dan data latih yang digunakan dalam proses simulasi.

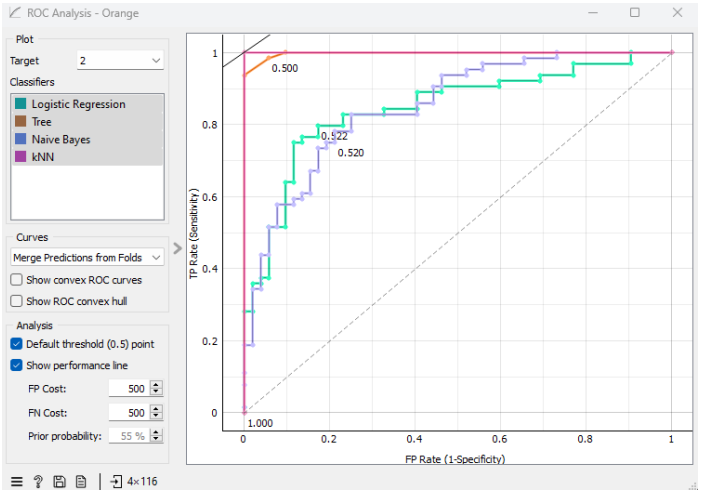
D. Hasil Evaluasi *ROC Curve*

Evaluasi model-model tersebut akan menggunakan *Receiver Operating Characteristic Curve (ROC Curve)* untuk mengukur kinerja model dalam mengklasifikasikan kanker payudara. Masing-masing model dianalisis menggunakan kurva ROC, yang menggambarkan hubungan antara sensitivitas dan spesifisitas model pada berbagai ambang batas keputusan. Melihat kurva ROC merupakan cara yang mudah untuk membandingkan akurasi dari setiap model klasifikasi karena disediakan dalam bentuk grafis.



Gambar 9.
Analisa ROC Kanker Payudara Klasifikasi Sehat

Pada gambar 9 menunjukkan bahwa hasil analisis ROC Kanker Payudara Klasifikasi Sehat pada masing-masing model sebagai berikut: *Logistic Regression* adalah 0.486, *Decision Tree* adalah 0.500, *Naive Bayes* adalah 0.480 dan *K-Nearest Neighbor (K-NN)* adalah 1.000. Oleh karena itu, riset klasifikasi untuk studi kasus ini model yang memiliki nilai akurasi paling baik adalah *K-Nearest Neighbor (K-NN)* dan *Decision Tree* karena kurvanya mendekati titik 0.1.



Gambar 10.
Analisa ROC Kanker Payudara Klasifikasi Sakit

Pada gambar 10 menunjukkan bahwa hasil analisis ROC Kanker Payudara Klasifikasi Sakit pada masing-masing model sebagai berikut: *Logistic Regression* adalah 0.522, *Decision Tree* adalah 0.500, *Naive Bayes* adalah 0.520 dan *K-Nearest Neighbor (K-NN)* adalah 1.000. Oleh karena itu, riset klasifikasi untuk studi kasus ini model yang memiliki nilai akurasi paling baik adalah *K-Nearest Neighbor (K-NN)* dan *Decision Tree* karena kurvanya mendekati titik 0.1.

IV. KESIMPULAN

Berdasarkan hasil pengujian dan analisis yang telah dilakukan, menunjukkan bahwa setelah menggunakan model *K-Nearest Neighbor (K-NN)*, *Naive Bayes*, *Decision Tree*, dan *Logistic Regression* untuk mengklasifikasi status kanker payudara diperoleh hasil bahwa kinerja *K-Nearest Neighbor (K-NN)* lebih unggul dari *Naive Bayes*, *Decision Tree*, dan *Logistic Regression*. Terbukti bahwa dari data uji yang digunakan *K-Nearest Neighbor (K-NN)* memiliki nilai *accuracy* 100%, *precision* 100%, *recall* 100% dan AUC 1.000 yang termasuk kategori *excellent classification*, sedangkan *Decision Tree* nilai *accuracy* 97%, *precision* 93%, *recall* 100% dan AUC 0.997 yang termasuk kategori *excellent classification*, lalu *Logistic Regression* memiliki nilai *accuracy* 80%, *precision* 76%, *recall* 81% dan AUC 0.847 yang termasuk kategori *good classification* dan *Naive Bayes* memiliki nilai *accuracy* 77%, *precision* 72%, *recall* 79% dan AUC 0.852 yang termasuk kategori *good classification*. Riset ini memiliki potensi untuk meningkatkan diagnosis dini dan perawatan kanker payudara, yang pada gilirannya dapat menyelamatkan nyawa dan meningkatkan kualitas hidup pasien. Selain itu, kontribusi riset ini dapat mendorong pengembangan teknologi medis yang lebih canggih dan efektif dalam melawan penyakit ini.

DAFTAR PUSTAKA

[1] S. Nurwenda, "Deteksi Dini Kanker: Mengapa dan Bagaimana?," Direktorat Jenderal Pelayanan Kesehatan Kemenkes RI, 2022. https://yankes.kemkes.go.id/view_artikel/173/deteksi-dini-kanker-mengapa-dan-bagaimana (accessed Oct. 25, 2023).
[2] C. M. Annur, "Kanker Payudara, Penyakit Kanker Paling Banyak Dialami Masyarakat Indonesia," databoks.katadata.co.id, 2022. https://databoks.katadata.co.id/datapublish/2022/10/11/kanker-payudara-penyakit-kanker-paling-banyak-dialami-masyarakat-indonesia (accessed Oct. 25, 2023).
[3] R. Fadli, "Kanker Payudara," halodoc.com, 2022. https://www.halodoc.com/kesehatan/kanker-payudara (accessed Oct. 25, 2023).
[4] Pittara, "Kanker Payudara," alodokter.com, 2023. https://www.alodokter.com/kanker-payudara (accessed Oct. 30, 2023).
[5] H. K. Nurwigati Sumartiningtyas, "Kanker Payudara Paling Banyak Didiagnosis di Dunia, Studi Jelaskan," kompas.com, 2021. https://www.kompas.com/sains/read/2021/02/05/192600023/kanker-payudara-paling-banyakdidiagnosis-di-dunia-studi-jelaskan (accessed Oct. 30, 2023).
[6] D. Andriani, "Ini Jenis Kanker yang Paling Banyak Diderita Masyarakat Indonesia," lifestyle.bisnis.com, 2020. https://lifestyle.bisnis.com/read/20200225/106/1205840/ini-jenis-kanker-yang-paling-banyak-diderita-masyarakat-indonesia (accessed Oct. 30, 2023).
[7] N. Wardani, "Penerapan Data Mining Dalam Analytic CRM," Yayasan Kita Menulis. pp. 47–48, 2020.
[8] R. Ordila, R. Wahyuni, Y. Irawan, and M. Y. Sari, "Penerapan Data Mining Untuk Pengelompokan Data Rekam Medis Pasien Berdasarkan Jenis Penyakit Dengan Algoritma Clustering …," J. Ilmu Komput., vol. 9, no. 2, pp. 148–153, 2020.
[9] Sukamto, Y. Adriyani, and R. Aulia, "Prediksi Kelompok UKT Mahasiswa Menggunakan Algoritma K-Nearest Neighbor," JUITA J. Inform., vol. 8, no. 1, p. 121, 2020.
[10] R. Sari, "Analisis Sentimen pada Review Objek Wisata Dunia Fantasi Menggunakan Algoritma K-Nearest Neighbor (K-NN)," Evolusi J. Sains dan Manaj., Vol. 8, No. 1, pp. 10–17, 2020.
[11] O. S. Bachri and R. M. Herdian Bhakti, "Penentuan Status Stunting pada Anak dengan Menggunakan Algoritma KNN," J. Ilm. Intech Inf. Technol. J. UMUS, vol. 3, no. 02, pp. 130–137, 2021.
[12] S. H. Rukmawan, F. R. Aszhari, Z. Rustam, and J. Pandelaki, "Cerebral Infarction Classification Using the K-Nearest Neighbor and Naive Bayes Classifier," J. Phys. Conf. Ser., vol. 1752, no. 1, 2021.
[13] D. Kurniawan, Pengenalan Machine Learning Python. Jakarta: PT ELEX MEDIA KOMPUTINDO, 2020.
[14] J. O. Onah, S. M. Abdulhamid, M. Abdullahi, I. H. Hassan, and A. Al-Ghusham, "Genetic Algorithm based feature selection and Naïve Bayes for anomaly detection in fog computing environment," Mach. Learn. with Appl., vol. 6, p. 100156, 2021.
[15] A. Z. Zami, O. Nurdiawan, and G. Dwilestari, "Klasifikasi Kondisi Gizi Bayi Bawah Lima Tahun Pada Posyandu Melati Dengan Menggunakan Algoritma Decision Tree," J. Sist. Komput. Dan Inform., vol. 3, pp. 305–310, 2022.
[16] S. A. T. Al Azhima, D. Darmawan, N. F. A. Hakim, I. Kustiawan, M. Al Qibtiya, and N. S. Syafei, "Hybrid Machine Learning Model untuk Memprediksi Penyakit Jantung dengan Metode Logistic Regression dan Random Forest," J. Teknol. Terpadu, vol. 8, no. 1, pp. 40–46, 2022.
[17] R. H. Situngkir and P. Sembiring, "Analisis Regresi Logistik Untuk Menentukan Faktor-Faktor Yang Mempengaruhi Kesejahteraan Masyarakat Kabupaten/Kota Di Pulau Nias," Jurnal Matematika dan Pendidikan Matematika, vol. 6, no. 1, pp. 25–31, 2023.
[18] I. B. P. Jayawiguna, "Comparison of Model Prediction for Tile Production in Tabanan Regency with Orange Data Mining Tool," in International Journal of Engineering and Emerging Technology, vol. 5, no. 2, pp. 72–76, 2020.
[19] R. Puspita and A. Widodo, "Perbandingan Metode KNN, Decision Tree, dan Naive Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS," J. Inform. Univ. Pamulang, vol. 5, no. 4, p. 646-654, 2021.
[20] H. Hozairi, A. Anwari, and S. Alim, "Implementasi Orange Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Dengan Model K-nearest Neighbor, Decision Tree Serta Naive Bayes," Jurnal Ilmiah NERO, vol. 6, no. 2, 2021.