

MODEL KLASIFIKASI PENYAKIT KANKER PAYUDARA MENGGUNAKAN ALGORITMA NAIVE BAYES

BREAST CANCER DISEASE CLASSIFICATION MODEL USING NAIVE BAYES ALGORITHM

Lili Tanti^{a,1}, Bob Subhan Riza^{b,2}, Yan Yang Thanri^{c,3}, Nurmi Panjaitan^{d,4}

^{a,b,c,d} Universitas Potensi Utama, Jl. KL. Yos Sudarso Km. 6,5 no. 3-A Tanjung Mulia, Medan, 20241, Indonesia
lilitanti82@gmail.com¹, bobsubhan779@gmail.com², ythanri@gmail.com³

ABSTRAK

Kanker payudara merupakan salah satu jenis kanker yang paling umum menyerang wanita di seluruh dunia dan menjadi penyebab utama kematian akibat kanker. Diagnosis dini dan akurasi dalam klasifikasi kanker jinak dan ganas sangat penting untuk menentukan langkah pengobatan yang tepat dan meningkatkan peluang kesembuhan. Penelitian ini bertujuan untuk mengeksplorasi penerapan algoritma Naive Bayes dalam klasifikasi kanker payudara menggunakan data klinis, dengan tujuan untuk membedakan kanker jinak dan ganas secara efisien. Pendekatan yang digunakan dalam penelitian ini adalah CRISP-DM, yang memfokuskan pada pemrosesan data dan pembangunan model klasifikasi. Algoritma Naive Bayes diterapkan untuk menganalisis atribut klinis seperti *radius_mean*, *texture_mean*, dan *perimeter_mean* untuk membedakan jenis kanker payudara. Hasil penelitian menunjukkan bahwa model yang dibangun dengan algoritma ini mencapai akurasi 91,15% dengan tingkat kesalahan sebesar 8,85%, menunjukkan efisiensi dan kecepatan yang baik dalam menangani dataset besar. Tabel yang disertakan dalam penelitian ini berisi data pasien dengan kanker payudara ganas, yang memberikan gambaran mendalam mengenai karakteristik tumor yang berperan penting dalam analisis diagnosis dan pengobatan. Pembahasan hasil penelitian menyoroti peran algoritma ini dalam meningkatkan deteksi dini dan mendukung pengambilan keputusan medis yang lebih cepat serta lebih akurat. Kesimpulan dari penelitian ini menegaskan bahwa penerapan Naive Bayes dalam klasifikasi kanker payudara dapat meningkatkan akurasi diagnosis, mempercepat proses pengobatan, serta berpotensi menurunkan angka kematian akibat kanker payudara, serta memberikan kontribusi besar dalam pengembangan teknologi kesehatan berbasis data.

Kata Kunci : Naive Bayes, Klasifikasi, Kanker Payudara, CRISP-DM, Model.

ABSTRACT

Breast cancer is one of the most common types of cancer affecting women worldwide and remains a leading cause of cancer-related deaths. Early diagnosis and accurate classification of benign and malignant tumors are crucial for determining appropriate treatment strategies and improving survival rates.

This study aims to explore the application of the Naive Bayes algorithm in breast cancer classification using clinical data, with the goal of efficiently distinguishing between benign and malignant cases. The approach adopted in this research is CRISP-DM, focusing on data processing and classification model development. The Naive Bayes algorithm was applied to analyze clinical attributes such as *radius_mean*, *texture_mean*, and *perimeter_mean* to differentiate between types of breast cancer. The results indicate that the model built using this algorithm achieved an accuracy of 91.15% with an error rate of 8.85%, demonstrating its efficiency and speed in handling large datasets. Tables included in the study present data on patients with malignant breast cancer, offering deep insights into tumor characteristics critical for diagnostic and treatment analyses. The discussion highlights the role of this algorithm in enhancing early detection and supporting faster and more accurate medical decision-making. The study concludes that implementing Naive Bayes in breast cancer classification can improve diagnostic accuracy, expedite treatment processes, and potentially reduce mortality rates from breast cancer, making a significant contribution to the development of data-driven healthcare technologies.

Keywords : Naive Bayes, Classification, Breast Cancer, CRISP-DM, Model.

*Info Artikel :**Disubmit:**Direview:**Diterima :*

Copyright © 2024 - Journal UPU. All rights reserved.

1. PENDAHULUAN

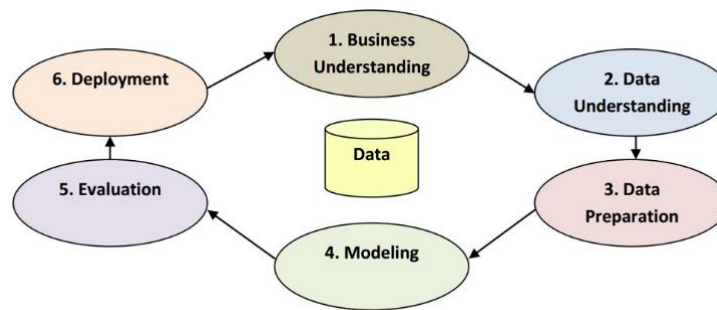
Kanker merupakan salah satu kondisi medis yang sering kali sulit dideteksi pada tahap awal. Hal ini disebabkan oleh rendahnya tingkat sensitivitas alat diagnostik serta kurangnya pemahaman masyarakat terhadap tanda-tanda awal kanker. Deteksi dini memiliki peran penting dalam meningkatkan peluang kesembuhan pasien, namun banyak kasus kanker baru teridentifikasi pada stadium lanjut, sehingga menimbulkan tantangan besar dalam pengobatan dan berdampak signifikan bagi pasien [1]. Kanker payudara adalah salah satu jenis kanker yang paling sering didiagnosis pada wanita di Australia dan menjadi penyebab kematian tertinggi kedua setelah kanker paru-paru [2]. Secara umum, kanker dapat dibagi menjadi dua jenis, yaitu kanker ganas (malignant) dan kanker jinak (benign). Kanker ganas menyerang secara agresif dan merusak sel-sel di sekitarnya, sedangkan kanker jinak cenderung tidak menyebar atau menyerang jaringan di sekitarnya. Salah satu tantangan utama dalam mendeteksi kanker payudara adalah ketiadaan gejala pasti yang dapat diidentifikasi secara akurat pada tahap awal [3].

Deteksi dini menjadi langkah strategis untuk meningkatkan peluang kesembuhan pasien kanker payudara. Dalam beberapa tahun terakhir, kecerdasan buatan (Artificial Intelligence/AI) telah membawa terobosan besar dalam proses diagnostik, termasuk melalui penerapan algoritma Naive Bayes. Algoritma ini berbasis pada teori probabilitas Bayes dan mampu menganalisis dataset dengan atribut seperti *radius_mean*, *texture_mean*, *perimeter_mean*, *area_mean*, dan *concave points_mean* untuk memprediksi diagnosis pasien dalam kategori "M" (malignant) atau "B" (benign). Keunggulan utama dari algoritma Naive Bayes adalah kesederhanaan, efisiensi, dan kemampuannya untuk menangani dataset dalam jumlah besar secara cepat dan akurat [4]. Meskipun algoritma ini berasumsi bahwa setiap atribut bersifat independen, penelitian menunjukkan bahwa asumsi ini cukup untuk menghasilkan model dengan tingkat akurasi tinggi. Naive Bayes juga dapat memberikan hasil secara real-time, yang sangat bermanfaat dalam konteks medis [5]. Selain itu, algoritma ini telah diterapkan secara luas pada berbagai kasus klasifikasi, termasuk diagnosis kanker payudara, dan terbukti memberikan hasil yang memuaskan [6]. Dalam bidang kesehatan, klasifikasi teks medis menjadi salah satu komponen penting dalam pengembangan kecerdasan buatan. Algoritma seperti Naive Bayes dapat digunakan untuk menganalisis teks medis yang kompleks, termasuk laporan diagnosis dan rekam medis pasien, guna menemukan pola yang relevan dalam data kanker payudara secara lebih efektif.

Penelitian ini bertujuan untuk memanfaatkan algoritma Naive Bayes dalam mendukung proses deteksi dini kanker payudara. Hasil penelitian ini diharapkan dapat membantu para dokter dalam membuat diagnosis yang lebih cepat dan akurat. Selain itu, penerapan teknologi berbasis kecerdasan buatan ini diharapkan mampu meningkatkan efisiensi layanan kesehatan secara keseluruhan serta menekan angka kematian akibat kanker payudara. Dengan memanfaatkan kemajuan teknologi, penelitian ini berupaya memberikan kontribusi nyata bagi komunitas medis dan masyarakat luas dalam menghadapi tantangan kesehatan modern.

2. METODE

Penelitian ini menggunakan pendekatan berbasis metodologi CRISP-DM (Cross Industry Standard Process for Data Mining), yang merupakan standar dalam pengembangan proses data mining untuk berbagai aplikasi yang dapat dilihat pada gambar 1.



Gambar 1. Metode Penelitian menggunakan CRIPS-DM

Penelitian ini sebagaimana terlihat pada gambar 1 menggunakan metodologi CRISP-DM (Cross Industry Standard Process for Data Mining), yang terdiri dari enam tahapan utama untuk memastikan pengembangan model yang sistematis dan terarah. Tahap pertama adalah Business Understanding, di mana penelitian fokus pada memahami masalah utama, yaitu pentingnya deteksi dini kanker payudara untuk meningkatkan tingkat kesembuhan pasien. Pada tahap ini, tujuan penelitian dirumuskan dengan mengidentifikasi kebutuhan diagnosis yang cepat dan akurat serta indikator keberhasilan model yang dikembangkan. Tahap berikutnya adalah Data Understanding, yang melibatkan eksplorasi data diagnosis kanker payudara, termasuk atribut seperti *radius_mean*, *texture_mean*, *perimeter_mean*, *area_mean*, dan *concave points_mean*, untuk memastikan bahwa data relevan dengan tujuan penelitian.

Setelah memahami data, tahap Data Preparation dilakukan untuk membersihkan dan mempersiapkan data. Proses ini mencakup penghapusan nilai yang hilang, outlier, serta pemilihan atribut yang relevan untuk meningkatkan kualitas dan akurasi model. Kemudian, tahap Modeling dilakukan dengan menggunakan algoritma Naive Bayes untuk membangun model klasifikasi. Model ini dirancang untuk mengkategorikan data diagnosis ke dalam dua kelas, yaitu "M" (malignant) dan "B" (benign), berdasarkan atribut yang telah dipilih.

Tahap selanjutnya adalah Evaluation, di mana kinerja model yang dikembangkan dievaluasi menggunakan metrik seperti akurasi, presisi, recall, dan F1-score untuk memastikan bahwa model mampu memberikan prediksi yang sesuai dengan kebutuhan klinis. Akhirnya, hasil penelitian diimplementasikan pada tahap Deployment dalam bentuk model berbasis kecerdasan buatan yang siap digunakan oleh tenaga medis untuk mendukung diagnosis kanker payudara. Metodologi ini dirancang untuk memastikan setiap tahap pengembangan dilakukan secara terstruktur dan memberikan hasil yang dapat diterapkan secara praktis untuk meningkatkan efisiensi dan akurasi diagnosis medis.

Algoritma Naïve Bayes

Berbasis teori probabilitas Bayes, algoritma Naive Bayes adalah teknik klasifikasi yang didasarkan pada asumsi "naif", atau sederhana bahwa setiap fitur bersifat independen satu sama lain. Namun, asumsi ini jarang benar dalam dunia data nyata, tetapi algoritma ini masih efektif dalam banyak situasi, terutama untuk masalah klasifikasi teks seperti analisis sentimen, deteksi spam, dan pengenalan dokumen. Naive Bayes menghitung probabilitas suatu data termasuk dalam kelas tertentu berdasarkan fitur-fiturnya. Persamaan utamanya berasal dari Teorema Bayes:

$$P(X|C) = \frac{P(X|C).P(C)}{P(X)} \quad (1)$$

$P(C|X)$: Probabilitas bahwa data X termasuk kelas C (probabilitas posterior).

$P(X|C)$: Probabilitas data X diberikan kelas C (likelihood).

$P(C)$: Probabilitas awal dari kelas C (prior).

$P(X)$: Probabilitas total dari data X (normalisasi).

3. HASIL DAN PEMBAHASAN

A. Business Understanding

Salah satu masalah utama adalah membedakan kanker ganas (malignant) dan jinak (jinak) secara cepat dan akurat berdasarkan ciri-ciri visual. Sementara tujuannya ialah menciptakan model prediktif berbasis data yang menggunakan algoritma naïve bayes untuk mengklasifikasikan jenis kanker berdasarkan karakteristik visual, sehingga model ini membantu dokter membuat diagnosis yang lebih cepat, akurat, dan berbasis data, yang dapat meningkatkan efisiensi pengobatan dan meningkatkan hasil kesehatan pasien.

B. Data Understanding

Dataset ini berisi data tentang pasien yang didiagnosis dengan kanker payudara dengan Alamat dataset public yaitu <https://www.kaggle.com/datasets/erdemtaha/cancer-data>. Tiap entri dalam dataset mewakili satu pasien dengan berbagai informasi yang relevan untuk membantu diagnosis dan klasifikasi kanker.

- 1) Fitur-Fitur Utama:
 - a) ID : Menandakan ID unik untuk setiap pasien. Ini hanya digunakan sebagai pengidentifikasi dan tidak berperan langsung dalam analisis atau prediksi.
 - b) Diagnosis : Label yang menunjukkan apakah kanker tersebut jinak (B - Benign) atau ganas (M - Malignant). Ini adalah variabel target yang ingin diprediksi oleh model. Diagnosis ini penting karena membantu dalam pembelajaran model untuk mengklasifikasikan kanker.
- 2) Fitur-Fitur Karakteristik Visual:
 - a) radius_mean : Rata-rata dari radius tumor atau massa kanker. Radius dapat memberikan gambaran tentang ukuran tumor.
 - b) texture_mean : Rata-rata tekstur permukaan tumor, yang bisa mengindikasikan ketidakteraturan permukaan kanker.
 - c) perimeter_mean : Rata-rata panjang keliling tumor. Ini dapat mengindikasikan kelengkungan dan bentuk tumor.
 - d) area_mean: Rata-rata area tumor. Area yang lebih besar bisa mengindikasikan kanker yang lebih invasif.
 - e) smoothness_mean: Rata-rata kehalusan tumor, yang menunjukkan tingkat kehalusan permukaan tumor. Tumor ganas biasanya memiliki permukaan yang lebih kasar.
 - f) compactness_mean: Rata-rata kepadatan atau kekompakan tumor. Tumor yang lebih padat mungkin memiliki sifat ganas yang lebih tinggi.
 - g) concavity_mean: Rata-rata kedalaman cekungan pada permukaan tumor. Tumor ganas sering kali memiliki cekungan yang lebih dalam pada permukaannya.
 - h) concave points_mean: Rata-rata jumlah titik cekung pada permukaan tumor. Semakin banyak cekungan, kemungkinan semakin ganas tumor tersebut.
- 3) Fitur-Fitur Rentang Nilai Rata-Rata:
 - a) Fitur-fitur ini menunjukkan distribusi nilai-nilai dalam rentang tertentu (misalnya, berapa banyak tumor yang memiliki nilai radius antara 10 dan 15, atau yang memiliki tekstur di antara rentang tertentu). Fitur ini bisa memberikan wawasan lebih lanjut mengenai distribusi karakteristik tumor dalam populasi yang lebih luas.

- b) Fitur Numerik: Semua fitur karakteristik visual (seperti radius_mean, texture_mean, area_mean, dll.) adalah fitur numerik. Mereka menggambarkan ukuran dan sifat visual tumor.
- c) Fitur Kategorikal: diagnosis adalah fitur kategorikal dengan dua kelas: "M" untuk Malignant (ganas) dan "B" untuk Benign (jinak). Ini adalah target variabel yang digunakan untuk klasifikasi.

Row No.	diagnosis	id	radius_mean	texture_mean	perimeter_...	area_mean	smoothness...	compactne
1	M	842302	17.990	10.380	122.800	1001	0.118	0.278
2	M	842517	20.570	17.770	132.900	1326	0.085	0.079
3	M	84300903	19.690	21.250	130	1203	0.110	0.160
4	M	84348301	11.420	20.380	77.580	386.100	0.142	0.284
5	M	84358402	20.290	14.340	135.100	1297	0.100	0.133
6	M	843786	12.450	15.700	82.570	477.100	0.128	0.170
7	M	844359	18.250	19.980	119.600	1040	0.095	0.109
8	M	84458202	13.710	20.830	90.200	577.900	0.119	0.165
9	M	844981	13	21.820	87.500	519.800	0.127	0.193
10	M	84501001	12.460	24.040	83.970	475.900	0.119	0.240

Gambar 2. Data Set Kanker Payudara (M – Malignant)

Pada gambar 2 menunjukkan tabel berisi data pasien dengan kanker payudara ganas (diagnosis "M"). Setiap baris menunjukkan informasi tentang karakteristik tumor, seperti rata-rata jari-jari (radius_mean), tekstur (texture_mean), keliling (perimeter), luas (area_mean), kehalusan (smoothness), dan kepadatan (compactness) tumor. Data ini membantu dokter dalam menganalisis stadium kanker dan menentukan metode pengobatan yang tepat. Semua pasien dalam tabel ini didiagnosis dengan kanker payudara ganas, dan informasi tersebut sangat penting untuk merencanakan pengobatan yang sesuai. Sedangkan untuk kanker payudara jinak (diagnosis "B") ditunjukkan pada gambar 3.

Row No.	id	diagnosis ↑	radius_mean	texture_mean	perimeter_...	area_mean	smoothness...	compactnes...	cor
20	8510426	B	13.540	14.360	87.460	566.300	0.098	0.081	0.0
21	8510653	B	13.080	15.710	85.630	520	0.107	0.127	0.0
22	8510824	B	9504	12.440	60.340	273.900	0.102	0.065	0.0
38	854941	B	13.030	18.420	82.610	523.800	0.090	0.038	0.0
47	85713702	B	8196	16.840	51.710	201.900	0.086	0.059	0.0
49	857155	B	12.050	14.630	78.040	449.300	0.103	0.091	0.0
50	857156	B	13.490	22.300	86.910	561	0.088	0.077	0.0
51	857343	B	11.760	21.600	74.720	427.900	0.086	0.050	0.0
52	857373	B	13.640	16.340	87.210	571.800	0.077	0.061	0.0
53	857374	B	11.940	18.240	75.710	437.600	0.083	0.048	0.0
56	85759902	B	11.520	18.750	73.340	409	0.095	0.055	0.0
59	857810	B	13.050	19.310	82.610	527.200	0.081	0.038	0.0
60	858477	B	8618	11.790	54.340	224.500	0.098	0.053	0.0
61	858970	B	10.170	14.880	64.550	311.900	0.113	0.081	0.0

Gambar 3. Data Set Kanker Payudara (jinak (B - Benign)

C. Data Preparation

Tahap Data Preparation merupakan langkah penting dalam metodologi CRISP-DM yang bertujuan untuk mempersiapkan data agar siap digunakan dalam proses pemodelan. Pada tahap ini, data yang telah dikumpulkan dari dataset diagnosis kanker payudara dianalisis secara mendalam untuk memastikan

kualitas dan kelengkapannya. Proses ini meliputi beberapa langkah utama, seperti menangani data yang hilang (missing values) dengan metode imputasi atau penghapusan data yang tidak lengkap, serta mengidentifikasi dan mengatasi outlier yang dapat memengaruhi kinerja model. Selain itu, data juga dinormalisasi untuk memastikan bahwa setiap atribut memiliki skala yang sama, sehingga algoritma Naive Bayes dapat bekerja secara optimal. Tahap ini juga mencakup pemilihan atribut (feature selection) untuk memilih variabel yang paling relevan, seperti `radius_mean`, `texture_mean`, `perimeter_mean`, `area_mean`, dan `concave points_mean`, yang diketahui memiliki pengaruh signifikan terhadap diagnosis kanker payudara. Selanjutnya, data diubah menjadi format yang sesuai untuk pemodelan, termasuk encoding data kategorikal seperti kelas diagnosis "M" (malignant) dan "B" (benign) menjadi nilai numerik. Dengan memastikan data bersih, lengkap, dan relevan, tahap Data Preparation ini menjadi dasar yang kuat untuk membangun model klasifikasi yang akurat dan andal.

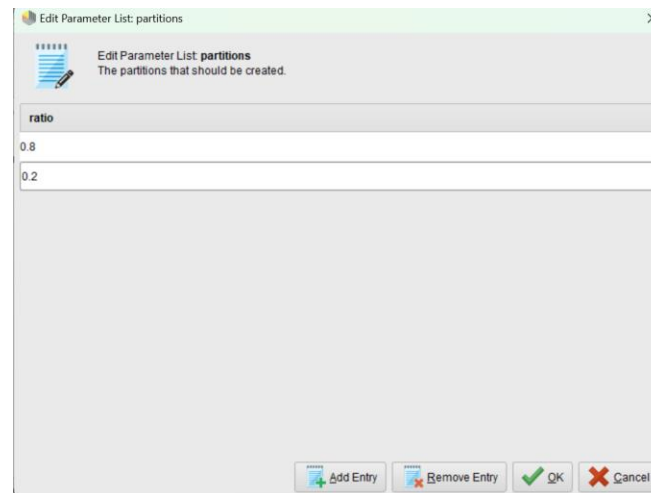
Label	Polynomial	0	Least	Most	Values
diagnosis			M (212)	B (357)	B (357)
id	Integer	0	Min 8670	Max 911320502	Average 30371
radius_mean	Real	0	Min 6.981	Max 28.110	Average 14.12
texture_mean	Real	0	Min 9.710	Max 39.280	Average 19.29
perimeter_mean	Real	0	Min 43.790	Max 188.500	Average 91.96
area_mean	Real	0	Min 143.500	Max 2501	Average 654.8
smoothness_mean	Real	0	Min 0.053	Max 0.163	Average 0.096

Gambar 4. Missing Value dan Label pada Data

Pada gambar 4 menunjukkan bahwa label berwarna hijau menunjukkan bahwa kolom dalam data tersebut tidak mengandung missing value atau nilai yang hilang. Ini berarti setiap baris data memiliki nilai untuk setiap variabel, memastikan data lengkap dan valid. Tidak adanya missing value memberikan keuntungan dalam analisis, karena seluruh dataset dapat digunakan tanpa perlu melakukan imputasi atau penghapusan baris yang memiliki nilai hilang. Hal ini menyederhanakan proses analisis, baik untuk analisis statistik maupun pembelajaran mesin, karena tidak ada gangguan akibat data yang hilang. Selain itu, analisis menjadi lebih akurat dan hasilnya lebih dapat diandalkan, karena semua data yang digunakan adalah data yang valid dan konsisten. Dengan demikian, tidak adanya missing value memudahkan interpretasi hasil dan membuat proses pengolahan data lebih efisien, mengurangi kebutuhan akan teknik penanganan data yang hilang seperti imputasi atau penghapusan.

Pembagian data menjadi set pelatihan dan pengujian adalah langkah penting dalam proses pembangunan model machine learning. Langkah ini bertujuan untuk memastikan bahwa model yang dikembangkan memiliki kemampuan generalisasi yang baik, sehingga dapat memberikan prediksi yang akurat terhadap data baru yang belum pernah dilihat sebelumnya. Dengan membagi data ke dalam dua bagian, yaitu 80% untuk pelatihan dan 20% untuk pengujian, proses pelatihan model dapat difokuskan pada pengoptimalan parameter berdasarkan data pelatihan, sementara evaluasi dilakukan menggunakan data pengujian untuk menilai performa model secara objektif. Strategi ini sangat penting untuk mencegah masalah overfitting, di mana model hanya "menghafal" data pelatihan tanpa memahami pola yang mendasarinya. Pembagian data yang proporsional juga memastikan bahwa distribusi data di kedua subset

tetap konsisten, sehingga hasil evaluasi dapat merepresentasikan performa model secara keseluruhan dapat dilihat pada gambar 4.



Gambar 4. SplitData

Pada gambar 5 menunjukkan sebuah jendela dialog berjudul "Edit Daftar Parameter : partisi" muncul di bawah gambar. Data dibagi menjadi beberapa bagian (partisi) dengan rasio tertentu dengan bantuan jendela ini. Dalam mesin pengajaran, partisi ini biasanya digunakan untuk memisahkan data menjadi data latihan (data pelatihan), data uji (data pengujian), atau bahkan validasi data.

Ini adalah bagian penting dari penentuan rasio pembagian data. Dalam gambar, terdapat dua baris dengan nilai 0,8 dan 0,2, yang menunjukkan bahwa data akan dibagi menjadi dua partisi dengan rasio 80% dan 20%, masing-masing. Nilai 0,8 biasanya mewakili rasio data latihan (data pelatihan), yang berarti 80% dari total data akan digunakan untuk melatih model pembelajaran mesin. Nilai 0,2 biasanya mewakili rasio data uji (data pengujian), yang berarti 20% dari total data yang akan digunakan. Berikut adalah rumus untuk membagi data menjadi data latih dan data uji berdasarkan rasio tertentu:

$$\text{Data Latih} = \text{Total Data} \times \text{Rasio Latih} \quad (2)$$

$$\text{Data Uji} = \text{Total Data} \times \text{Rasio Uji} \quad (3)$$

Maka Penyelesaiannya dengan menggunakan persamaan (2) dan persamaan (3)

$$\text{Data Latih} = 570 \times 0.8 = 456 \text{ Data}$$

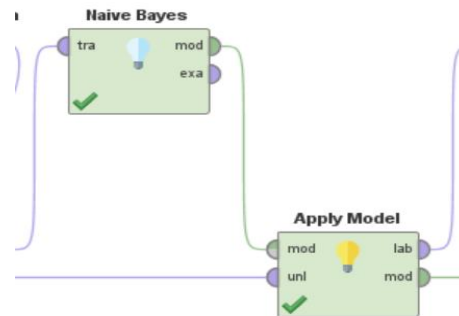
$$\text{Data Uji} = 570 \times 0.2 = 114 \text{ Data}$$

Berdasarkan hasil perhitungan manual dari 570 data sampel, dataset tersebut dibagi menjadi 456 data untuk Data Training dan 114 data untuk Data Testing, yang berasal dari data publik di platform Kaggle. Pembagian ini dilakukan untuk memastikan bahwa model Naïve Bayes dapat dilatih secara optimal menggunakan Data Training dan kemudian diuji pada Data Testing untuk mengevaluasi performanya dalam klasifikasi kanker payudara.

D. Modelling

Modeling adalah tahap penting dalam proses analisis data yang bertujuan untuk membangun model prediktif atau klasifikasi berdasarkan dataset yang telah dipersiapkan sebelumnya. Pada tahap ini, algoritma pembelajaran mesin diterapkan untuk mengenali pola-pola dalam data dan menghasilkan model yang mampu membuat prediksi atau keputusan berdasarkan data baru. Dalam penelitian ini, algoritma Naive Bayes digunakan karena keunggulannya dalam klasifikasi yang cepat, sederhana, dan efektif meskipun dengan asumsi independensi antar atribut. Proses modeling melibatkan pelatihan model

menggunakan data pelatihan yang telah dibagi sebelumnya dan pengujian model menggunakan data uji untuk mengevaluasi kinerjanya. Hasil dari tahap ini akan memberikan gambaran mengenai seberapa baik model dapat memprediksi diagnosis kanker payudara, baik untuk kategori "M" (malignant) maupun "B" (benign), berdasarkan atribut-atribut seperti radius_mean, texture_mean, perimeter_mean, area_mean, dan concave points_mean terlihat pada gambar 5.



Gambar 5. Modeling pada Data Set

Pada gambar 5 merupakan alur kerja pemodelan yang digunakan oleh algoritma Naive Bayes dalam perangkat lunak pembelajaran mesin dijelaskan di sini. Pelatihan dan penerapan model adalah dua langkah dalam proses. Pada tahap pelatihan, data pelatihan dimasukkan ke blok Naive Bayes. Algoritma ini menggunakan teorema Bayes sebagai dasar untuk menghitung kemungkinan fitur untuk setiap kelas. Model yang dihasilkan dari pelatihan ini kemudian dikirim melalui output "mod".

Pada tahap penerapan, model yang telah dibuat dimasukkan ke blok Apply Model bersama dengan data baru tanpa label (input "unl"). Model memprediksi kelas data baru berdasarkan probabilitas yang telah dihitung sebelumnya, dan data dengan label prediksi dikeluarkan melalui output "lab". Karena cepat, sederhana, dan efektif, Naive Bayes sering digunakan untuk klasifikasi, terutama untuk data besar. Meskipun mengasumsikan fitur independensi, yang jarang terjadi di dunia nyata, algoritma ini tetap bekerja dengan baik.

Radius-mean

1. Mean (Rata – Rata)

Mean adalah rata-rata dari sekumpulan data, yang dihitung dengan menjumlahkan semua nilai data lalu membaginya dengan jumlah data.

Berikut ini adalah rumusnya :

$$Mean = \frac{\sum x_i}{n} \quad (4)$$

xi: Nilai data ke-i

n: Banyaknya data

2. Standar Deviasi, adalah ukuran seberapa tersebar data dari nilai rata-ratanya.

a) Rumus Standar Deviasi (Populasi):

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}} \quad (5)$$

b) Rumus Standar Deviasi (Sampel):

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \quad (6)$$

xi: Nilai data ke-i

μ : Mean (rata-rata populasi) atau $\bar{x}$ (mean sampel)

n: Banyaknya data

Misalkan kita menggunakan fitur radius_mean dengan nilai $x=15.0$ untuk kelas "M" dan "B". Kita hitung likelihood menggunakan mean dan standard deviation dari tabel. Berikutnya menghitung probabilitas untuk fitur radius_mean dengan $x=15.0$.

Probabilitas untuk fitur radius_mean dengan nilai $x=15.0$ adalah sebagai berikut:

- Untuk kelas "M": $P(x|M) = 17,346$
- Untuk kelas "B": $P(x|B) = 12,101$

Ini menunjukkan bahwa nilai $x=15.0$ lebih mungkin berasal dari kelas "M" dibandingkan dengan kelas "B" berdasarkan distribusi Gaussian

Attribute	Parameter	M	B
id	mean	30325575.647	16323283.094
	standard deviation	120812130.618	77592992.382
radius_mean	mean	17.346	12.101
	standard deviation	3.149	1.784
texture_mean	mean	21.558	17.924
	standard deviation	3.893	4.006
perimeter_mean	mean	114.585	77.766
	standard deviation	21.536	11.813
area_mean	mean	964.883	459.472
	standard deviation	357.851	134.667
smoothness_mean	mean	0.103	0.092
	standard deviation	0.013	0.013
compactness_mean	mean	0.145	0.079

Gambar 6. Hasil Model Klasifikasi Penyakit Kanker Payudara

Algoritma Naive Bayes digunakan untuk memprediksi kelas data, seperti menentukan apakah data termasuk kelompok M atau B, berdasarkan nilai atribut. Naive Bayes mengasumsikan independensi antar atribut dan menghitung probabilitas berdasarkan statistik deskriptif, seperti mean dan standar deviasi. Dalam tabel yang disajikan, atribut seperti radius_mean, texture_mean, dan area_mean memberikan informasi fisik suatu objek. Perbedaan nilai rata-rata antara kelompok M dan B menunjukkan bahwa kedua kelompok memiliki karakteristik berbeda, yang membantu membedakan kelas. Misalnya, jika radius_mean lebih tinggi pada kelompok M, atribut ini dapat menjadi pembeda yang signifikan. Standar deviasi juga mencerminkan variasi data di dalam kelompok; standar deviasi besar menunjukkan data lebih tersebar. Dengan memahami distribusi dan kontribusi setiap atribut, Naive Bayes dapat memanfaatkan informasi ini untuk mengklasifikasi data secara efektif. Kombinasi mean dan standar deviasi memainkan peran penting dalam analisis dan prediksi.

E. Evaluation

Dalam analisis performa model klasifikasi, metrik seperti accuracy, precision, dan recall sangat penting untuk mengevaluasi sejauh mana model dapat membedakan antara kelas Malignant (M - ganas) dan Benign (B - jinak) dalam diagnosis kanker payudara. Berdasarkan gambar 7 yang diberikan, hasil evaluasi menunjukkan akurasi sebesar 91.15%, yang menunjukkan bahwa model memiliki performa yang cukup baik.

	true M	true B	class precision
pred. M	39	7	84.78%
pred. B	3	64	95.52%
class recall	92.86%	90.14%	

Gambar 7. Evaluasi Model dalam bentuk Confusion Matrix

Berdasarkan confusion matrix pada gambar 7 yang diberikan, model klasifikasi menghasilkan 39 True Positive (TP), yang berarti model berhasil mengidentifikasi 39 kasus kanker ganas (M) dengan benar. Namun, terdapat 7 False Positive (FP), di mana model salah memprediksi sebagai kanker ganas (M) padahal sebenarnya jinak (B). Selain itu, ada 3 False Negative (FN), di mana model salah memprediksi sebagai jinak (B) padahal sebenarnya ganas (M), yang dapat berisiko karena kasus kanker ganas tidak terdeteksi dengan benar. Sementara itu, model juga berhasil mengklasifikasikan 64 True Negative (TN), yang berarti model dengan tepat mengidentifikasi 64 kasus kanker jinak (B). Dari hasil ini, kita dapat menghitung metrik evaluasi seperti akurasi, presisi, dan recall untuk menilai performa model dalam mendeteksi kanker payudara.

Akurasi (Accuracy)

Akurasi mengukur seberapa sering model memberikan prediksi yang benar secara keseluruhan dengan menggunakan persamaan (7)

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (7)$$

$$Accuracy = \frac{39 + 64}{39 + 7 + 3 + 64} = \frac{103}{113} = 0,9115 \approx 91,15\%$$

Presisi (Precision)

Precision mengukur seberapa akurat model saat memprediksi kelas positif (Malignant - M) menggunakan persamaan (8)

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

$$Precision = \frac{39}{39 + 7} = \frac{39}{46} = 0,8478 \approx 84,78\%$$

Recall (Sensitivity)

Recall mengukur seberapa baik model dalam mendeteksi kasus positif sebenarnya dengan menggunakan persamaan (9)

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

$$Recall = \frac{39}{39 + 3} = \frac{39}{42} = 0,9286 \approx 92,86\%$$

Model memiliki akurasi 91.15%, yang berarti 91.15% dari semua prediksi benar. Precision 84.78% menunjukkan bahwa ketika model memprediksi kanker ganas, 84.78% prediksinya benar. Recall 92.86% menunjukkan bahwa model berhasil mengidentifikasi 92.86% dari semua kasus kanker ganas. Dengan hasil ini, model menunjukkan performa yang cukup baik dalam mendeteksi kanker payudara.

4. KESIMPULAN

Model Naïve Bayes telah terbukti sebagai metode yang efektif dalam klasifikasi kanker payudara, dengan akurasi 91,15% dalam membedakan antara tumor ganas (Malignant - M) dan jinak (Benign - B). Dengan analisis fitur yang relevan, model ini dapat membantu tenaga medis dalam mengambil keputusan berbasis data untuk meningkatkan efisiensi diagnosis. Proses deployment juga memungkinkan model ini untuk diintegrasikan dalam sistem rekam medis elektronik (EHR) atau aplikasi berbasis cloud, sehingga dapat diakses dengan mudah oleh profesional medis. Namun, meskipun model ini memiliki performa yang baik, masih terdapat beberapa keterbatasan seperti asumsi independensi antar fitur yang mungkin tidak selalu realistis dalam data medis yang kompleks.

Untuk penelitian selanjutnya, model ini dapat dikembangkan lebih lanjut dengan menggunakan pendekatan ensemble learning, seperti kombinasi Naïve Bayes dengan Random Forest atau Gradient Boosting, guna meningkatkan akurasi prediksi. Selain itu, eksplorasi teknik deep learning, seperti Convolutional Neural Networks (CNNs) untuk analisis gambar mammogram, dapat menjadi langkah inovatif dalam mendukung diagnosis berbasis citra medis. Penelitian mendatang juga dapat berfokus pada implementasi sistem berbasis edge computing, yang memungkinkan prediksi dilakukan secara langsung di perangkat medis tanpa harus mengandalkan server cloud. Dengan pengembangan lebih lanjut, diharapkan sistem klasifikasi kanker payudara berbasis kecerdasan buatan ini dapat semakin akurat, efisien, dan mudah diakses oleh tenaga kesehatan di seluruh dunia.

UCAPAN TERIMA KASIH

Kami mengucapkan terima kasih yang sebesar-besarnya kepada semua pihak yang telah berkontribusi dalam penelitian dan pengembangan model klasifikasi kanker payudara ini. Terima kasih kepada para tenaga medis, peneliti, serta tim pengembang yang telah memberikan wawasan dan data yang berharga untuk meningkatkan akurasi model. Kami juga menghargai dukungan dari institusi akademik dan teknologi yang telah menyediakan sumber daya guna mewujudkan penelitian ini. Semoga hasil penelitian ini dapat memberikan manfaat bagi dunia medis dan menjadi langkah awal dalam pengembangan teknologi kecerdasan buatan untuk deteksi dini kanker payudara.

REFERENSI

- [1] Anggraini, R. A. (2023). Algoritma Naïve Bayes Dengan Backward Elimination Pada Dataset Breast Cancer. *Jurnal Kajian Ilmiah*, 23(1), 87-94.
- [2] Jabbar, M. A., Hasmin, E., Susanto, C., Musu, W., & Artikel, I. (2022). Komparasi Algoritma Decision Tree, Naive Bayes, dan K-Nearest Neighbors dalam Klasifikasi Kanker Payudara Comparison of Decision Tree Algorithms, Naive Bayes, and K-Nearest Neighbors in Breast Cancer Classification.
- [3] Derisma, D., & Febrian, F. (2020). Perbandingan Teknik Klasifikasi Neural Network, Support Vector Machine, dan Naive Bayes dalam Mendeteksi Kanker Payudara. *Bina Insani ICT Journal*, 7(1), 53-62.
- [4] Wulandari, E. (2022). Klasifikasi Kanker Paru-Paru Menggunakan Metode Naive Bayes. *International Research on Big-Data and Computer Technology: I-Robot*, 6(2), 20-24.

- [5] Sudewo, R. T. P., Pratama, Y., & Yanti, E. (2023). Analisis Data Mining Untuk Prediksi Kanker Payudara Menggunakan Algoritma Klasifikasi. *Jurnal Pustaka Data (Pusat Akses Kajian Database, Analisa Teknologi, dan Arsitektur Komputer)*, 3(2), 62-69.
- [6] Taslim, T., Handayani, S., & Fajrizal, F. (2023). Kinerja Komparatif Optimasi Algoritma Naive Bayes dalam Klasifikasi Teks untuk Uji Klinis Kanker. *Jurnal Eksplora Informatika*, 13(1), 113-123.
- [7] Juliani, D. (2024). Implementasi Machine Learning untuk Klasifikasi Penyakit Kanker Paru Menggunakan Metode Naïve Bayes dengan Tambahan Fitur Chatbot. *JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI (IPTEK)*, 8(2), 12-17.