

Analisis Klasifikasi Stadium Kanker Payudara Menggunakan Algoritma Naïve Bayes Berdasarkan Data Rekam Medis

Mudafiq Riyan Pratama¹, Sefia Ayu Maharani², Mochammad Choirur Roziqin³, Dony Setiawan Hendyca Putra⁴

¹Jurusan Kesehatan, Politeknik Negeri Jember, mudafiq.riyan@polije.ac.id

²Jurusan Kesehatan, Politeknik Negeri Jember, sefiaayu160903@gmail.com

³Jurusan Kesehatan, Politeknik Negeri Jember, irul@polije.ac.id

⁴Jurusan Kesehatan, Politeknik Negeri Jember, dony_shp@polije.ac.id

Keywords:

Breast Cancer,
Classification,
Naïve Bayes,
Medical Records,
Data Mining

ABSTRACT

Breast cancer is one of the leading causes of cancer-related morbidity and mortality among women, highlighting the need for an approach that supports rapid and accurate cancer staging. The utilization of medical records through data mining techniques provides an opportunity to develop classification models that can assist clinical decision-making. This study aimed to analyze breast cancer stage classification using the Naïve Bayes algorithm based on patients' medical records from Baladhika Husada Level III Hospital, Jember. A quantitative approach was employed using secondary data consisting of 476 medical records obtained through data selection, data cleaning, and data transformation with categorical encoding from an initial dataset of 1,082 medical records. The classification model was developed using the Naïve Bayes algorithm and evaluated using a Confusion Matrix based on accuracy, precision, and recall across nine training-testing split ratios. The results showed that the 90:10 split ratio achieved the best performance, with an accuracy of 87.50%, precision of 86.36%, and recall of 86.36%. Clinical characteristic analysis based on likelihood values indicated that the presence of axillary lumps, changes in breast shape and size, breast skin abnormalities, and tumor size were more predominant in advanced-stage breast cancer, thereby contributing to the classification process. Overall, the findings demonstrate that the Naïve Bayes algorithm is capable of classifying breast cancer stages with good performance and has the potential to serve as the foundation for developing a medical record-based clinical decision support system to assist in breast cancer stage identification.

Kata Kunci

Kanker Payudara,
Klasifikasi,
Naïve Bayes,
Rekam Medis,
Data Mining

ABSTRAK

Kanker payudara merupakan salah satu penyebab utama morbiditas dan mortalitas akibat kanker pada perempuan sehingga diperlukan pendekatan yang mampu mendukung proses penentuan stadium secara cepat dan akurat. Pemanfaatan data rekam medis melalui teknik *data mining* memberikan peluang untuk membangun model klasifikasi yang dapat membantu pengambilan keputusan klinis. Penelitian ini bertujuan menganalisis klasifikasi stadium kanker payudara menggunakan algoritma *Naïve Bayes* berdasarkan data rekam medis pasien di Rumah Sakit Tingkat III Baladhika Husada Jember. Penelitian menggunakan pendekatan kuantitatif dengan memanfaatkan data sekunder sebanyak 476 rekam medis yang diperoleh melalui proses *data selection*, *data cleaning*, dan *data transformation* dengan *categorical encoding* dari total 1.082 data rekam medis. Model klasifikasi dibangun menggunakan algoritma *Naïve Bayes* dan dievaluasi menggunakan *Confusion Matrix* berdasarkan nilai *accuracy*, *precision*, dan *recall* pada sembilan skenario pembagian data *training-testing split*. Hasil penelitian menunjukkan bahwa skenario pembagian data 90:10 memberikan performa terbaik dengan nilai *accuracy* sebesar 87,50%, *precision* sebesar 86,36%, dan *recall* sebesar 86,36%. Analisis karakteristik klinis berdasarkan nilai *likelihood* menunjukkan bahwa variabel benjolan di ketiak, perubahan bentuk dan ukuran payudara, kelainan pada kulit payudara, serta ukuran tumor memiliki kecenderungan yang lebih dominan pada stadium lanjut sehingga memberikan kontribusi terhadap proses klasifikasi. Hasil penelitian menunjukkan bahwa algoritma *Naïve Bayes* mampu mengklasifikasikan stadium kanker payudara dengan performa yang baik serta berpotensi dimanfaatkan sebagai dasar pengembangan sistem pendukung keputusan berbasis rekam medis untuk membantu identifikasi stadium kanker payudara.

Korespondensi Penulis:

Mudafiq Riyan Pratama,
Politeknik Negeri Jember, Jl. Mastrip 164 Jember
Telepon : +6282229414949
Email: mudafiq.riyan@polije.ac.id

Submitted: 01-07-2026; Accepted: 28-07-2026;
Published: 29-07-2026

Copyright (c) 2026 The Author (s) This article is distributed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0)

1. PENDAHULUAN

Kanker payudara menjadi salah satu jenis kanker dengan angka kesakitan (morbiditas) dan kematian (mortalitas) tertinggi pada perempuan di seluruh dunia. Berdasarkan laporan *Global Cancer Observatory* (GLOBOCAN) tahun 2022, kanker payudara merupakan jenis kanker dengan insidensi tertinggi kedua di dunia, dengan total 2.296.840 kasus baru. Di Indonesia, kanker payudara menempati peringkat pertama sebagai kanker dengan insidensi tertinggi, yaitu sebanyak 66.271 kasus baru, serta menyebabkan 22.598 kematian pada tahun yang sama [1]. Besarnya jumlah kasus dan angka kematian akibat kanker payudara menunjukkan bahwa penyakit ini masih menjadi tantangan utama dalam bidang kesehatan. Oleh karena itu, diperlukan upaya deteksi dini dan penanganan yang tepat untuk meningkatkan keberhasilan terapi sekaligus menekan angka mortalitas.

Kanker payudara merupakan penyakit yang ditandai oleh pertumbuhan sel-sel abnormal pada jaringan payudara secara tidak terkendali, sehingga berpotensi menginvasi jaringan di sekitarnya maupun menyebar ke organ tubuh lain melalui proses metastasis. Risiko terjadinya kanker payudara dipengaruhi oleh berbagai faktor, antara lain usia, riwayat kanker payudara dalam keluarga, faktor genetik, riwayat reproduksi, paparan hormon, obesitas, konsumsi alkohol, serta gaya hidup. Selain itu, beberapa manifestasi klinis seperti munculnya benjolan pada payudara, perubahan bentuk payudara, perubahan puting susu, keluarnya cairan dari puting, hingga perubahan pada kulit payudara juga menjadi indikator penting dalam proses diagnosis. Keberagaman faktor risiko dan karakteristik klinis tersebut menyebabkan proses identifikasi stadium kanker memerlukan analisis yang komprehensif agar penanganan yang diberikan dapat disesuaikan dengan kondisi pasien [2] [3].

Penentuan stadium kanker payudara merupakan salah satu tahapan penting dalam proses diagnosis karena menjadi dasar dalam pemilihan terapi, penentuan prognosis, serta evaluasi keberhasilan pengobatan. Semakin dini stadium kanker dapat diidentifikasi, semakin besar peluang pasien memperoleh terapi yang efektif dan memiliki angka harapan hidup yang lebih baik. Dalam praktiknya, penentuan stadium dilakukan berdasarkan berbagai informasi klinis pasien yang terdokumentasi dalam rekam medis. Seiring dengan implementasi rekam medis elektronik di fasilitas pelayanan kesehatan, data klinis pasien tersimpan dalam jumlah besar dan memiliki potensi untuk dimanfaatkan sebagai sumber informasi dalam mendukung pengambilan keputusan klinis. Namun demikian, pemanfaatan data rekam medis masih lebih banyak difokuskan sebagai dokumentasi pelayanan kesehatan dibandingkan sebagai sumber pengetahuan melalui analisis data [4].

Perkembangan *data mining* memungkinkan proses ekstraksi informasi dan penemuan pola dari data rekam medis sehingga dapat dimanfaatkan untuk mendukung proses klasifikasi penyakit. Salah satu algoritma klasifikasi yang banyak digunakan adalah Naïve Bayes, yaitu algoritma berbasis probabilitas yang memiliki tingkat komputasi relatif sederhana, mampu bekerja dengan baik pada data berdimensi banyak, serta menghasilkan performa klasifikasi yang kompetitif pada berbagai kasus di bidang kesehatan [5]. Beberapa penelitian terdahulu menunjukkan bahwa algoritma Naïve Bayes mampu menghasilkan performa klasifikasi yang baik pada kasus kanker payudara. Ramadhan dan Kurniawati (2020) memperoleh tingkat akurasi sebesar 71,43%, sedangkan Fahrani et al. (2024) melaporkan akurasi sebesar 97,78% menggunakan data rekam medis rumah sakit. Penelitian lain oleh Munir et al. (2024) juga menunjukkan bahwa algoritma Naïve Bayes memiliki performa yang kompetitif dibandingkan algoritma klasifikasi lainnya [6].

Penelitian ini memiliki beberapa perbedaan dibandingkan penelitian sebelumnya. Pertama, penelitian menggunakan 476 data rekam medis pasien kanker payudara yang berasal dari Rumah Sakit Tingkat III Baladhika Husada Jember. Model klasifikasi dibangun menggunakan 13 variabel independen yang terdiri atas usia pasien, riwayat kanker payudara yang dialami oleh keluarga, riwayat penyakit payudara yang pernah dialami oleh pasien sebelumnya, status IMT (Indeks Massa Tubuh), mengalaminya nyeri pada payudara, terdapat benjolan pada payudara, terdapat benjolan pada ketiak, perubahan bentuk dan ukuran pada payudara, keluarnya cairan dari puting payudara, perubahan puting susu, kelainan kulit payudara, status lokalis kanker, dan ukuran tumor payudara. Sementara itu, stadium kanker digunakan sebagai variabel dependen yang menjadi target klasifikasi. Ketiga, penelitian tidak hanya mengevaluasi performa model menggunakan *confusion matrix*, tetapi juga menganalisis probabilitas *likelihood* setiap variabel sehingga karakteristik klinis yang berkontribusi terhadap stadium awal maupun stadium lanjut dapat diidentifikasi secara lebih komprehensif.

Berdasarkan uraian tersebut, pemanfaatan algoritma Naïve Bayes untuk klasifikasi kanker payudara telah banyak dilakukan. Namun, sebagian besar penelitian masih berfokus pada klasifikasi kanker jinak dan ganas maupun perbandingan performa algoritma klasifikasi. Penelitian mengenai klasifikasi stadium kanker payudara berdasarkan karakteristik klinis pasien menggunakan data rekam medis rumah sakit masih relatif terbatas. Oleh karena itu, penelitian ini bertujuan membangun model klasifikasi stadium kanker payudara menggunakan algoritma Naïve Bayes berdasarkan data rekam medis pasien di Rumah Sakit Tingkat III Baladhika Husada Jember serta mengevaluasi kinerja model dalam mengklasifikasikan stadium penyakit.

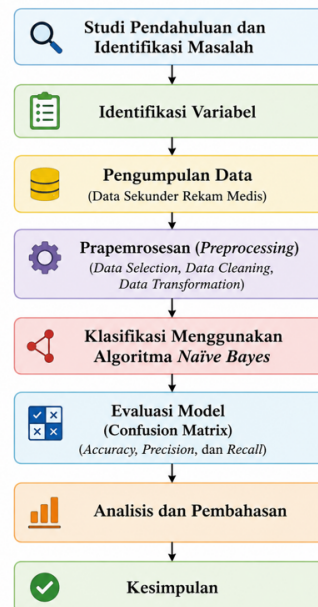
2. METODE PENELITIAN

2.1 Jenis Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan menerapkan teknik *data mining* menggunakan algoritma *Naïve Bayes* untuk mengklasifikasikan stadium kanker payudara berdasarkan data rekam medis pasien. Model klasifikasi diimplementasikan menggunakan bahasa pemrograman Python pada platform Google Colab.

Evaluasi terhadap hasil klasifikasi dilakukan menggunakan *Confusion Matrix* dengan menghitung nilai *accuracy*, *precision*, dan *recall* sebagai indikator kinerja model.

2.2 Tahapan Penelitian



Gambar 1. Tahapan Penelitian

2.3 Variabel Penelitian

Pada penelitian ini, variabel target (dependen) adalah stadium kanker payudara yang dikelompokkan menjadi dua kategori, yaitu stadium awal yang mencakup Stadium I, Stadium IIA, dan Stadium IIB; serta stadium lanjut yang terdiri atas Stadium IIIA, Stadium IIIB, Stadium IIIC, dan Stadium IV. Sementara itu, model klasifikasi dibangun menggunakan 13 variabel independen yang merepresentasikan karakteristik klinis pasien, meliputi usia pasien, riwayat kanker payudara yang dialami oleh keluarga, riwayat penyakit payudara yang pernah dialami oleh pasien sebelumnya, status IMT (Indeks Massa Tubuh), mengalaminya nyeri pada payudara, terdapat benjolan pada payudara, terdapat benjolan pada ketiak, perubahan bentuk dan ukuran pada payudara, keluarnya cairan dari puting payudara, perubahan puting susu, kelainan kulit payudara, status lokalis kanker, dan ukuran tumor payudara.

2.4 Pengumpulan Data & Preprocessing Data

Penelitian dilaksanakan di Rumah Sakit Tingkat III Baladhika Husada Jember. Data penelitian menggunakan data sekunder berupa rekam medis elektronik pasien rawat inap yang terdiagnosis kanker payudara berdasarkan ICD-10 kode C50 selama periode 2024–2025. Populasi penelitian berjumlah 1.082 rekam medis pasien. Setelah melalui proses seleksi berdasarkan kriteria kelengkapan data, diperoleh 476 data rekam medis yang digunakan sebagai dataset penelitian. Dari data yang sudah dikumpulkan, selanjutnya dilakukan tahap preprocessing yang bertujuan mempersiapkan data agar siap digunakan dalam proses klasifikasi. Tahapan ini meliputi *data selection*, yaitu pemilihan data yang sesuai dengan tujuan penelitian; *data cleaning*, yaitu menghapus data yang tidak lengkap, duplikat, atau tidak sesuai; serta *data transformation*, yaitu mengubah format data ke dalam bentuk kategorikal yang dapat diproses oleh algoritma Naïve Bayes.

2.5 Klasifikasi Menggunakan Algoritma Naïve Bayes

Dataset yang telah melalui tahap *preprocessing* selanjutnya diproses menggunakan algoritma *Naïve Bayes*. Model dibangun dengan beberapa skenario pembagian *data training* dan *data testing*, yaitu 10:90, 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, 80:20, dan 90:10. Setiap skenario digunakan untuk memperoleh model dengan performa klasifikasi terbaik berdasarkan *probabilitas posterior* yang dihitung menggunakan algoritma *Naïve Bayes*.

2.6 Evaluasi Model Menggunakan Confusion Matrix

Model klasifikasi yang dihasilkan dievaluasi menggunakan metode *Confusion Matrix*. Evaluasi dilakukan berdasarkan nilai *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Parameter evaluasi yang digunakan meliputi *accuracy*, *precision*, dan *recall* untuk mengetahui kemampuan model dalam mengklasifikasikan stadium kanker payudara.

3. HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil implementasi algoritma *Naïve Bayes* dalam mengklasifikasikan stadium kanker payudara berdasarkan data rekam medis pasien. Pembahasan diawali dengan penyajian hasil *preprocessing* data yang mencakup *data selection*, *data cleaning*, dan *data transformation* dengan *categorical encoding*. Selanjutnya, dilakukan proses pembentukan model klasifikasi menggunakan algoritma *Naïve Bayes* serta evaluasi performa model melalui *Confusion Matrix* menggunakan parameter *accuracy*, *precision*, dan *recall* pada berbagai skenario pembagian

data training dan *testing*. Hasil yang diperoleh kemudian diinterpretasikan untuk menjelaskan kemampuan model dalam mengklasifikasikan stadium kanker payudara serta dibandingkan dengan temuan penelitian terdahulu guna memperkuat pembahasan.

3.1 Hasil Preprocessing Data

Pada penelitian ini, preprocessing dilakukan sebagai tahapan awal untuk menghasilkan dataset yang siap digunakan pada proses klasifikasi. Tahapan ini meliputi *data selection*, *data cleaning*, dan *data transformation* guna memastikan data yang digunakan memiliki kualitas yang baik, konsisten, serta sesuai dengan kebutuhan algoritma *Naïve Bayes*. Pada penelitian ini, *preprocessing* terdiri atas tiga tahapan, yaitu *data selection*, *data cleaning*, dan *data transformation*.

Tabel 1. Hasil *Preprocessing Data*

Tahap <i>Preprocessing</i>	Hasil
<i>Data Selection</i>	Data rekam medis pasien kanker payudara ICD-10 C50 periode 2024–2025 dengan jumlah 1.082 rekam medis pasien. Setelah melalui proses seleksi berdasarkan kriteria kelengkapan data, diperoleh 476 data rekam medis yang digunakan sebagai dataset penelitian.
<i>Data Cleaning</i>	Menghapus data yang tidak lengkap, duplikat, dan inkonsisten sehingga seluruh data siap dianalisis.
<i>Data Transformation</i>	Mengubah atribut ke dalam bentuk kategorikal sesuai kebutuhan algoritma <i>Naïve Bayes</i> . Setelah menjadi kategorikal, data tersebut dilakukan <i>encoding</i> dalam bentuk numerik.

Dari hasil data transformation, maka didapatkan isian tabel sebagai berikut:

Tabel 2. Tabel Hasil Transformasi Dalam Bentuk Kategorikal

	A	G	H	I	J	N	O	P	Q	R	S	T	U	V	W	Y
	No RM	Usia	Klasifikasi Usia	Riwayat Kanker Payudara dari Keluarga	Riwayat Penyakit Payudara Sebelumnya	Status IMT	Nyeri Pada Payudara	Benjolan di Payudara	Benjolan di Ketiak	Perubahan Bentuk dan Ukuran Payudara	Keluar Cairan dari Payudara	Perubahan Puting Susu	Kelainan Pada Kulit Payudara	Status Lokalis Kanker Payudara	Ukuran Tumor Payudara	Stadium
1																
2	142237	46	Lansia awal	Tidak Ada	Ada	Normal (18,5–22,9)	Ada	Ada	Tidak Ada	Tidak Ada	Tidak Ada	Tidak Ada	Ada	Unilateral	12,34	Stadium II B
3	146984	65	Lansia akhir	Tidak Ada	Ada	Normal (18,5–22,9)	Ada	Ada	Tidak Ada	Tidak Ada	Tidak Ada	Tidak Ada	Tidak Ada	Unilateral	2	Stadium I
4	132930	52	Lansia awal	Tidak Ada	Tidak Ada	Berat badan berlebih (23–24,9)	Ada	Ada	Ada	Ada	Tidak Ada	Tidak Ada	Ada	Unilateral	2,9	Stadium II B
5	135226	50	Lansia awal	Tidak Ada	Ada	Obesitas I (25–29,9)	Ada	Ada	Ada	Tidak Ada	Tidak Ada	Tidak Ada	Tidak Ada	Unilateral	3,5	Stadium II B
6	136949	59	Lansia akhir	Tidak Ada	Ada	Normal (18,5–22,9)	Ada	Ada	Tidak Ada	Tidak Ada	Tidak Ada	Tidak Ada	Tidak Ada	Unilateral	11,5	Stadium II B
7	143263	42	Dewasa akhir	Tidak Ada	Ada	Underweight (<18,5)	Ada	Ada	Ada	Tidak Ada	Tidak Ada	Tidak Ada	Tidak Ada	Unilateral	2	Stadium II B
8	143046	46	Lansia awal	Tidak Ada	Ada	Obesitas I (25–29,9)	Ada	Ada	Ada	Tidak Ada	Tidak Ada	Tidak Ada	Tidak Ada	Unilateral	4,5	Stadium II B
9	147079	69	Manula	Tidak Ada	Ada	Normal (18,5–22,9)	Ada	Ada	Tidak Ada	Tidak Ada	Tidak Ada	Tidak Ada	Tidak Ada	Unilateral	3,5	Stadium II A
10	128677	44	Dewasa akhir	Tidak Ada	Ada	Normal (18,5–22,9)	Tidak Ada	Ada	Tidak Ada	Tidak Ada	Tidak Ada	Tidak Ada	Tidak Ada	Unilateral	1,5	Stadium I
11	146700	53	Lansia awal	Tidak Ada	Ada	Obesitas I (25–29,9)	Ada	Ada	Ada	Tidak Ada	Tidak Ada	Tidak Ada	Tidak Ada	Unilateral	1,3	Stadium II A
12	145749	70	Manula	Tidak Ada	Ada	Berat badan berlebih (23–24,9)	Ada	Ada	Tidak Ada	Tidak Ada	Tidak Ada	Tidak Ada	Tidak Ada	Unilateral	3,8	Stadium II A
13	144108	49	Lansia awal	Tidak Ada	Ada	Obesitas I (25–29,9)	Ada	Ada	Tidak Ada	Tidak Ada	Tidak Ada	Tidak Ada	Tidak Ada	Unilateral	2,5	Stadium II A
14	133747	57	Lansia akhir	Tidak Ada	Ada	Normal (18,5–22,9)	Ada	Ada	Tidak Ada	Ada	Tidak Ada	Tidak Ada	Tidak Ada	Unilateral	6,5	Stadium II B
15	133863	45	Dewasa akhir	Tidak Ada	Ada	Berat badan berlebih (23–24,9)	Ada	Ada	Tidak Ada	Ada	Tidak Ada	Ada	Tidak Ada	Unilateral	4	Stadium II A
16	144844	42	Dewasa akhir	Tidak Ada	Ada	Berat badan berlebih (23–24,9)	Ada	Ada	Ada	Tidak Ada	Tidak Ada	Tidak Ada	Tidak Ada	Unilateral	3,5	Stadium II B
17	137885	58	Lansia akhir	Tidak Ada	Ada	Normal (18,5–22,9)	Ada	Ada	Tidak Ada	Ada	Tidak Ada	Tidak Ada	Tidak Ada	Bilateral	20	Stadium II B

Setelah seluruh data dinyatakan lengkap dan konsisten, dilakukan tahap transformasi data dengan mengubah atribut kategorikal menjadi representasi numerik (*categorical encoding*). Proses ini bertujuan agar setiap variabel dapat diproses secara komputasional pada tahap pembentukan model klasifikasi menggunakan algoritma *Naïve Bayes*. Pengkodean numerik dilakukan terhadap seluruh variabel independen dan variabel dependen dengan memberikan nilai numerik pada setiap kategori tanpa mengubah makna maupun karakteristik data asli.

Tabel 3. Transformasi Data Penelitian ke Bentuk Numerik

No.	Variabel	Kategori Asal	Hasil Encoding
1.	Usia	< 40 tahun	0
		≥ 40 tahun	1
2.	Riwayat kanker payudara dalam keluarga	Tidak	0
		Ya	1
3.	Riwayat penyakit payudara sebelumnya	Tidak	0
		Ya	1
4.	Indeks Massa Tubuh (IMT)	Normal	0
		Gemuk	1
5.	Nyeri payudara	Tidak	0
		Ya	1
6.	Benjolan pada ketiak	Tidak	0

		Ya	1
7.	Benjolan pada payudara	Tidak	0
		Ya	1
8.	Perubahan bentuk dan ukuran payudara	Tidak	0
		Ya	1
9.	Keluar cairan dari puting payudara	Tidak	0
		Ya	1
10.	Perubahan puting susu	Tidak	0
		Ya	1
11.	Kelainan kulit payudara	Tidak	0
		Ya	1
12.	Status lokalis kanker	Tidak	0
		Ya	1
13.	Ukuran tumor	≤ 5 cm	0
		> 5 cm	1
14.	Stadium kanker (Kelas)	Stadium Awal	0
		Stadium Lanjut	1

Hasil encoding menunjukkan bahwa sebanyak 13 variabel independen dan 1 variabel dependen telah berhasil dikonversi ke dalam format numerik. Dataset hasil transformasi selanjutnya digunakan sebagai masukan (input) pada proses pembentukan model klasifikasi menggunakan algoritma Naïve Bayes. Contoh hasil transformasi data setelah proses encoding ditampilkan pada tabel di bawah ini.

Tabel 4. Contoh Hasil Data Setelah *Categorical Encoding*

A	B	C	D	E	F	G	H	I	J	K	L	M	N
Usia	Riwayat Kanker Payudara dari Keluarga	Riwayat Penyakit Payudara Sebelumnya	Status IMT	Nyeri Pada Payudara	Benjolan di Payudara	Benjolan di Ketiak	Perubahan Bentuk dan Ukuran Payudara	Keluar Cairan dari Payudara	Perubahan Puting Susu	Kelainan Pada Kulit Payudara	Status Lokalis Kanker Payudara	Ukuran Tumor Payudara	Stadium
5	0	1	2	1	1	0	0	0	0	1	1	3	0
6	0	1	2	1	1	0	0	0	0	0	1	1	0
5	0	0	3	1	1	1	1	0	0	1	1	2	0
5	0	1	4	1	1	1	0	0	0	0	1	2	0
6	0	1	2	1	1	0	0	0	0	0	1	3	0
4	0	1	1	1	1	1	0	0	0	0	1	1	0
5	0	1	4	1	1	1	0	0	0	0	1	2	0
7	0	1	2	1	1	0	0	0	0	0	1	2	0
4	0	1	2	0	1	0	0	0	0	0	1	1	0
5	0	1	4	1	1	1	0	0	0	0	1	1	0
7	0	1	3	1	1	0	0	0	0	0	1	2	0
5	0	1	4	1	1	0	0	0	0	0	1	2	0
6	0	1	2	1	1	0	1	0	0	0	1	3	0
4	0	1	3	1	1	0	1	0	1	0	1	2	0
4	0	1	3	1	1	1	0	0	0	0	1	2	0
6	0	1	2	1	1	0	1	0	0	0	1	3	0
5	0	1	3	1	1	0	1	0	1	1	1	2	0
4	0	1	2	1	1	0	1	0	0	0	1	3	0
6	0	0	3	1	1	0	0	0	0	0	1	2	0

Berdasarkan hasil transformasi tersebut, seluruh data telah memiliki format yang seragam sehingga siap digunakan pada proses pelatihan (*training*) dan pengujian (*testing*) model klasifikasi menggunakan algoritma Naïve Bayes.

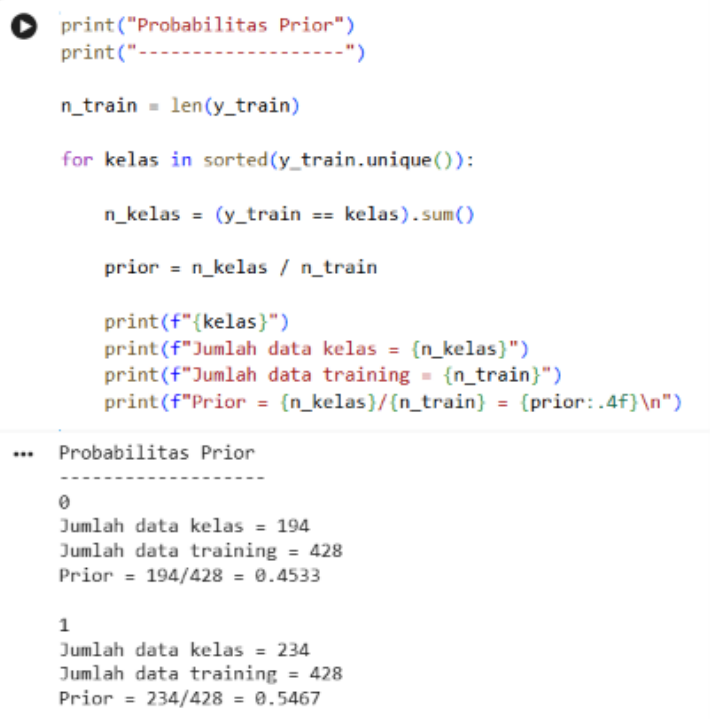
3.2 Hasil Klasifikasi Menggunakan Algoritma Naïve Bayes

Model klasifikasi dibangun menggunakan dataset hasil *preprocessing* yang terdiri atas 476 data rekam medis pasien. Implementasi dilakukan menggunakan bahasa pemrograman Python pada platform Google Colab dengan memanfaatkan algoritma Naïve Bayes. Algoritma ini mengklasifikasikan stadium kanker payudara berdasarkan probabilitas setiap variabel independen terhadap kelas target menggunakan Teorema Bayes.

Proses klasifikasi diawali dengan menghitung probabilitas awal (*prior probability*) masing-masing kelas, yaitu stadium awal dan stadium lanjut. Selanjutnya, dihitung probabilitas bersyarat (*likelihood*) setiap atribut terhadap masing-masing kelas berdasarkan *data training*. Nilai *prior* dan *likelihood* kemudian digunakan untuk memperoleh probabilitas *posterior* yang menjadi dasar dalam menentukan hasil klasifikasi.

3.3.1 Probabilitas Prior

Probabilitas prior dihitung berdasarkan proporsi jumlah data pada setiap kelas target. Nilai ini menggambarkan peluang awal suatu data termasuk ke dalam kelas stadium awal maupun stadium lanjut sebelum mempertimbangkan karakteristik klinis pasien. Hasil perhitungan probabilitas *prior* dituangkan dalam perhitungan manual sekaligus penerapan code pemrograman Python pada Google Colab berikut.

<i>Prior</i> Stadium Awal $= \frac{X}{A}$ $= \frac{194}{428}$ $= 0,4533$ <i>Prior</i> Stadium Lanjut $= \frac{X}{A}$ $= \frac{234}{428}$ $= 0,5467$	 <pre> print("Probabilitas Prior") print("-----") n_train = len(y_train) for kelas in sorted(y_train.unique()): n_kelas = (y_train == kelas).sum() prior = n_kelas / n_train print(f"{kelas}") print(f"Jumlah data kelas = {n_kelas}") print(f"Jumlah data training = {n_train}") print(f"Prior = {n_kelas}/{n_train} = {prior:.4f}\n") ... Probabilitas Prior ----- 0 Jumlah data kelas = 194 Jumlah data training = 428 Prior = 194/428 = 0.4533 1 Jumlah data kelas = 234 Jumlah data training = 428 Prior = 234/428 = 0.5467 </pre>
---	--

Gambar 2. Perhitungan dan Code Python untuk Penerapan Probabilitas *Prior*

Berdasarkan hasil perhitungan probabilitas *Prior* tersebut, diketahui bahwa kelas stadium lanjut memiliki nilai probabilitas *prior* yang lebih tinggi yaitu sebesar 0,5467, sedangkan kelas stadium awal memiliki nilai probabilitas *prior* sebesar 0,4533. Nilai tersebut menunjukkan bahwa distribusi kelas pada data *training* tidak memiliki jumlah yang sama.

3.3.2 Probabilitas *Likelihood*

Tahap berikutnya adalah menghitung probabilitas bersyarat (*Likelihood*) setiap kategori atribut terhadap kelas stadium awal maupun stadium lanjut. Perhitungan dilakukan terhadap seluruh variabel independen yang telah melalui proses transformasi data. Nilai *likelihood* digunakan sebagai dasar dalam menghitung probabilitas *Posterior* pada setiap data uji. Untuk menghitung nilai *Likelihood* menggunakan formula:

$$P(x_j|C_i) = \frac{nx_{j,C_i+1}}{n_{C_i+k}} \quad (1)$$

Berikut sebagian perhitungan *Likelihood* untuk kategori usia:

$$\begin{aligned}
 P(\text{Usia=Dewasa Awal}|\text{Stadium Awal}) &= \frac{nx_{Ci+1}}{n_{Ci+k}} = \frac{15+1}{220+7} = \frac{16}{227} = 0,0705 \\
 P(\text{Usia=Dewasa Awal}|\text{Stadium Lanjut}) &= \frac{nx_{Ci+1}}{n_{Ci+k}} = \frac{9+1}{256+7} = \frac{10}{263} = 0,0380 \\
 P(\text{Usia=Dewasa Akhir}|\text{Stadium Awal}) &= \frac{nx_{Ci+1}}{n_{Ci+k}} = \frac{44+1}{220+7} = \frac{45}{227} = 0,1982 \\
 P(\text{Usia=Dewasa Akhir}|\text{Stadium Lanjut}) &= \frac{nx_{Ci+1}}{n_{Ci+k}} = \frac{56+1}{256+7} = \frac{57}{263} = 0,2167 \\
 P(\text{Usia=Lansia Awal}|\text{Stadium Awal}) &= \frac{nx_{Ci+1}}{n_{Ci+k}} = \frac{94+1}{220+7} = \frac{95}{227} = 0,4185 \\
 P(\text{Usia=Lansia Awal}|\text{Stadium Lanjut}) &= \frac{nx_{Ci+1}}{n_{Ci+k}} = \frac{93+1}{256+7} = \frac{94}{263} = 0,3574 \\
 P(\text{Usia=Lansia Akhir}|\text{Stadium Awal}) &= \frac{nx_{Ci+1}}{n_{Ci+k}} = \frac{51+1}{220+7} = \frac{52}{227} = 0,2291 \\
 P(\text{Usia=Lansia Akhir}|\text{Stadium Lanjut}) &= \frac{nx_{Ci+1}}{n_{Ci+k}} = \frac{81+1}{256+7} = \frac{82}{263} = 0,3118 \\
 P(\text{Usia=Manula}|\text{Stadium Awal}) &= \frac{nx_{Ci+1}}{n_{Ci+k}} = \frac{16+1}{220+7} = \frac{17}{227} = 0,0749 \\
 P(\text{Usia=Manula}|\text{Stadium Lanjut}) &= \frac{nx_{Ci+1}}{n_{Ci+k}} = \frac{17+1}{256+7} = \frac{18}{263} = 0,0684
 \end{aligned}$$

Tahap berikutnya adalah menghitung probabilitas bersyarat (*likelihood*) untuk setiap kategori atribut terhadap masing-masing kelas stadium kanker payudara. Perhitungan dilakukan terhadap seluruh variabel independen yang digunakan dalam penelitian sehingga diperoleh peluang kemunculan setiap kategori pada kelas stadium awal maupun stadium lanjut. Dari hasil probabilitas *Likelihood* pada semua kategori atribut, hasilnya dituangkan dalam bentuk tabel di bawah ini:

Tabel 5. Hasil Probabilitas *Likelihood* Untuk Semua Kategori Atribut

1.	Usia	Stadium Awal	Stadium Lanjut
	• Dewasa Awal	0,0705	0,0380
	• Dewasa Akhir	0,1982	0,2167
	• Lansia Awal	0,4185	0,3574
	• Lansia Akhir	0,2291	0,3118
	• Manula	0,0749	0,0684
2.	Riwayat dari Keluarga	Stadium Awal	Stadium Lanjut
	• Ada	0,0180	0,0271
	• Tidak ada	0,9820	0,9729
3.	Riwayat Penyakit Payudara Sebelumnya	Stadium Awal	Stadium Lanjut
	• Ada	0,7883	0,8101
	• Tidak ada	0,2117	0,1899
4.	Status IMT	Stadium Awal	Stadium Lanjut
	• Underweight (<18,5)	0,1022	0,1609
	• Normal (18,5–22,9)	0,4089	0,4253
	• Berat badan berlebih (23–24,9)	0,1778	0,1571
	• Obesitas I (25–29,9)	0,2400	0,1839
	• Obesitas II (≥ 30)	0,0711	0,0728
5.	Nyeri pada payudara	Stadium Awal	Stadium Lanjut
	• Ada	0,9324	0,9689
	• Tidak Ada	0,0676	0,0310
6.	Benjolan dipayudara	Stadium Awal	Stadium Lanjut
	• Ada	0,9955	0,9961
	• Tidak Ada	0,0045	0,0039
7.	Benjolan diketiak	Stadium Awal	Stadium Lanjut
	• Ada	0,2703	0,7054
	• Tidak Ada	0,7297	0,2946
8.	Perubahan bentuk dan ukuran payudara	Stadium Awal	Stadium Lanjut
	• Ada	0,4279	0,6395
	• Tidak Ada	0,5721	0,3605
9.	Keluar cairan dari payudara	Stadium Awal	Stadium Lanjut
	• Ada	0,0315	0,0310
	• Tidak Ada	0,9685	0,9690
10.	Perubahan puting susu	Stadium Awal	Stadium Lanjut
	• Ada	0,1126	0,2674
	• Tidak Ada	0,8874	0,7326
11.	Kelaian pada kulit payudara	Stadium Awal	Stadium Lanjut
	• Ada	0,1081	0,6202
	• Tidak Ada	0,8919	0,3798
12.	Status Lokalis Kanker Payudara	Stadium Awal	Stadium Lanjut
	• Unilateral	0,9730	0,9612
	• Bilateral	0,0270	0,0388
13.	Ukuran Tumor Payudara	Stadium Awal	Stadium Lanjut
	• ≤ 2 cm	0,2332	0,0965
	• 2–5 cm	0,5202	0,3614
	• > 5 cm	0,2466	0,5521

Nilai *likelihood* menunjukkan kontribusi setiap karakteristik klinis pasien terhadap kemungkinan suatu data termasuk ke dalam kelas stadium awal atau stadium lanjut. Informasi tersebut menjadi dasar dalam proses perhitungan probabilitas *posterior* pada tahap klasifikasi.

3.3.3 Menghitung Probabilitas *Posterior* dengan Normalisasi (*Confidence*)

Setelah nilai *prior* dan *likelihood* diperoleh, langkah berikutnya adalah menghitung probabilitas *posterior* menggunakan Teorema Bayes. Probabilitas *posterior* dihitung dengan menggabungkan nilai probabilitas *prior* dan

likelihood dari setiap atribut pada masing-masing kelas berdasarkan data uji (*testing data*). Pada tahap ini, seluruh nilai *likelihood* yang diperoleh dari distribusi *Gaussian* dikalikan dengan nilai *prior* untuk menghasilkan nilai probabilitas awal setiap kelas. Setelah nilai *posterior* untuk setiap kelas diperoleh, dilakukan proses normalisasi untuk mengubahnya menjadi nilai *confidence*. Proses normalisasi dilakukan dengan cara membandingkan nilai probabilitas *posterior* dari suatu kelas terhadap jumlah keseluruhan nilai *posterior* pada semua kelas yang tersedia. Proses ini penting karena nilai hasil perkalian sering kali sangat kecil sehingga sulit untuk diinterpretasikan secara langsung. Normalisasi dilakukan agar total probabilitas dari seluruh kelas bernilai satu, sehingga hasil dapat dibandingkan secara proporsional. Dengan demikian, nilai *confidence* yang dihasilkan dapat menunjukkan peluang relatif masing-masing kelas secara lebih jelas dan mudah dipahami. Temuan ini sejalan dengan penelitian Indey & Supangat (2024) yang menyebutkan bahwa proses normalisasi pada nilai probabilitas bertujuan untuk memudahkan interpretasi hasil serta meningkatkan ketepatan dalam membandingkan peluang antar kelas [7]. Berikut contoh perhitungan *posterior* sebelum dinormalisasi:

$$\begin{aligned}
 \text{Posterior (Stadium Awal)} &= 0,4533 \times (0,0705 \times 0,0180 \times 0,7883 \times 0,1022 \times 0,9955 \times 0,2703 \times 0,4279 \times \\
 &\quad 0,0315 \times 0,1126 \times 0,1081 \times 0,9730 \times 0,0182 \times 0,0705) \\
 &= 0,000000231 \\
 \text{Posterior (Stadium Lanjut)} &= 0,5467 \times (0,0380 \times 0,0271 \times 0,8101 \times 0,1609 \times 0,9961 \times 0,7054 \times 0,6395 \times \\
 &\quad 0,0310 \times 0,2674 \times 0,6202 \times 0,9612 \times 0,0388 \times 0,0684) \\
 &= 0,000000584 \\
 \text{Total posterior} &= 0,000000231 + 0,000000584 \\
 &= 0,000000815
 \end{aligned}$$

Selanjutnya dilakukan normalisasi *posterior* yang kemudian menghasilkan *confidence*. Berikut contoh perhitungan *posterior* normalisasi:

$$\text{Posterior (Stadium Awal)} = \frac{0,000000231}{0,000000815} = 0,283 / 28,3\%$$

$$\text{Posterior (Stadium Lanjut)} = \frac{0,000000584}{0,000000815} = 0,717 / 71,7\%$$

Hasil $0,283 < 0,717$ maka diklasifikasikan sebagai Stadium Lanjut.

Untuk menghitung keseluruhan nilai *posterior* dengan normalisasi, maka implementasi perhitungan dilakukan menggunakan Python sebagaimana ditunjukkan pada gambar kode program berikut:

```

confidence_df = pd.DataFrame({
    'No': range(1, len(y_test)+1),
    'Stadium': y_test.values,
    'Confidence Stadium Awal': y_prob[:,0],
    'Confidence Stadium Lanjut': y_prob[:,1]
})

confidence_df = confidence_df.round(3)

confidence_df

```

Gambar 3. Kode Program Perhitungan Nilai Posterior Pada Data *Testing*

Berdasarkan hasil eksekusi kode program tersebut, didapatkan nilai *posterior* pada seluruh data testing yang dituangkan pada tabel 6. Kelas dengan nilai probabilitas *posterior* tertinggi dipilih sebagai hasil klasifikasi atau prediksi stadium kanker payudara.

Tabel 6. Nilai Posterior dengan Normalisasi

No	Aktual	Confidence		Prediksi
		Stadium Awal	Stadium Lanjut	
1	Stadium Lanjut	0.228	0.772	Stadium Lanjut
2	Stadium Awal	0.540	0.460	Stadium Awal
3	Stadium Awal	0.743	0.257	Stadium Awal
4	Stadium Awal	0.680	0.320	Stadium Awal
5	Stadium Awal	0.709	0.291	Stadium Awal
6	Stadium Awal	0.673	0.327	Stadium Awal
7	Stadium Lanjut	0.747	0.253	Stadium Awal
8	Stadium Awal	0.694	0.306	Stadium Awal
9	Stadium Lanjut	0.314	0.686	Stadium Lanjut
10	Stadium Awal	0.602	0.398	Stadium Awal
11	Stadium Lanjut	0.205	0.795	Stadium Lanjut
12	Stadium Lanjut	0.332	0.668	Stadium Lanjut
13	Stadium Lanjut	0.123	0.877	Stadium Lanjut
14	Stadium Lanjut	0.376	0.624	Stadium Lanjut
15	Stadium Awal	0.636	0.364	Stadium Awal
16	Stadium Lanjut	0.477	0.523	Stadium Lanjut

No	Aktual	Confidence		Prediksi
		Stadium Awal	Stadium Lanjut	
17	Stadium Awal	0.730	0.270	Stadium Awal
18	Stadium Awal	0.622	0.378	Stadium Awal
19	Stadium Lanjut	0.142	0.858	Stadium Lanjut
20	Stadium Awal	0.454	0.546	Stadium Lanjut
21	Stadium Lanjut	0.699	0.301	Stadium Awal
22	Stadium Lanjut	0.068	0.932	Stadium Lanjut
23	Stadium Awal	0.709	0.291	Stadium Awal
24	Stadium Lanjut	0.144	0.856	Stadium Lanjut
25	Stadium Awal	0.313	0.687	Stadium Lanjut
26	Stadium Lanjut	0.116	0.884	Stadium Lanjut
27	Stadium Awal	0.762	0.238	Stadium Awal
28	Stadium Awal	0.746	0.254	Stadium Awal
29	Stadium Lanjut	0.286	0.714	Stadium Lanjut
30	Stadium Awal	0.600	0.400	Stadium Awal
31	Stadium Lanjut	0.244	0.756	Stadium Lanjut
32	Stadium Awal	0.816	0.184	Stadium Awal
33	Stadium Lanjut	0.243	0.757	Stadium Lanjut
34	Stadium Awal	0.659	0.341	Stadium Awal
35	Stadium Awal	0.695	0.305	Stadium Awal
36	Stadium Lanjut	0.522	0.478	Stadium Awal
37	Stadium Awal	0.730	0.270	Stadium Awal
38	Stadium Lanjut	0.085	0.915	Stadium Lanjut
39	Stadium Awal	0.693	0.307	Stadium Awal
40	Stadium Awal	0.695	0.305	Stadium Awal
41	Stadium Awal	0.563	0.437	Stadium Awal
42	Stadium Lanjut	0.068	0.932	Stadium Lanjut
43	Stadium Awal	0.693	0.307	Stadium Awal
44	Stadium Lanjut	0.107	0.893	Stadium Lanjut
45	Stadium Awal	0.675	0.325	Stadium Awal
46	Stadium Lanjut	0.329	0.671	Stadium Lanjut
47	Stadium Lanjut	0.155	0.845	Stadium Lanjut
48	Stadium Awal	0.394	0.606	Stadium Lanjut

3.3 Evaluasi Model dan Pembahasan

3.3.1 Hasil Evaluasi Model Berdasarkan Skenario *Training-Testing Split*

Untuk mengetahui pengaruh proporsi data *training* terhadap performa model, dilakukan pengujian menggunakan sembilan skenario pembagian data *training* dan data *testing* dengan rasio persentase berikut: 10:90, 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, 80:20, dan 90:10. Setiap skenario dievaluasi menggunakan *Confusion Matrix* dengan parameter *accuracy*, *precision*, dan *recall*. Hasil evaluasi seluruh skenario disajikan pada Tabel berikut.

Tabel 7. Perbandingan Performa *Confusion Matrix* Berdasarkan *Training-Testing Split*

Rasio data <i>training</i> : <i>testing</i>	Accuracy	Precision	Recall
10:90 (47 : 429)	72,26%	72,11%	72,84%
20:80 (95 : 381)	72,70%	74,27%	75,00%
30:70 (142 : 334)	74,55%	77,33%	74,30%
40:60 (190 : 286)	74,48%	76,13%	76,62%
50:50 (238 : 238)	76,89%	75,76%	81,30%
60:40 (185 : 191)	76,44%	75,93%	81,19%
70:30 (333 : 143)	81,82%	80,00%	84,51%
80:20 (380 : 96)	80,21%	82,00%	80,39%
90:10 (428 : 48)	87,50%	86,36%	86,36%

Berdasarkan hasil pengujian, performa model mengalami variasi pada setiap skenario pembagian data. Hasil terbaik diperoleh pada rasio persentase **90:10** dengan nilai **accuracy sebesar 87,50%**, **precision sebesar 86,36%**, dan **recall sebesar 86,36%**. Hasil tersebut menunjukkan bahwa proporsi data *training* paling besar yang mampu meningkatkan kemampuan model dalam mengenali pola karakteristik pasien sehingga menghasilkan performa klasifikasi yang lebih baik.

3.3.2 Pembahasan Performa Model

Hasil pengujian menunjukkan bahwa pembagian data dengan rasio *split validation* 90:10 menghasilkan performa paling optimal, dengan nilai *accuracy* 87,50%, *precision* 86,36%, serta *recall* 86,36%. Nilai *accuracy* menunjukkan bahwa model mampu mengklasifikasikan 87,50% data pengujian dengan benar. Nilai *precision* mencapai 86,36% menunjukkan bahwa model memiliki kemampuan yang baik dalam menghasilkan prediksi yang

tepat serta mengenali data pada kelas yang sebenarnya. Nilai *recall* yang sama-sama mencapai 86,36% menunjukkan bahwa model memiliki kemampuan yang baik dalam menangkap sebagian besar data yang sebenarnya termasuk dalam kelas positif. Hal ini sejalan dengan penelitian Laksono et al (2025) yang menyatakan bahwa penerapan *Naïve Bayes Classifier* dengan skema pembagian data 90% data latih dan 10% data uji mampu menghasilkan performa yang baik, dengan tingkat akurasi mencapai 85% [8].

3.3.3 Analisis Karakteristik Klinis Pasien Berdasarkan Hasil Likelihood

Selain menunjukkan performa klasifikasi yang baik, hasil penelitian juga memperlihatkan adanya perbedaan karakteristik klinis antara pasien stadium awal dan stadium lanjut. Berdasarkan distribusi karakteristik pasien yang selanjutnya direpresentasikan dalam nilai *likelihood* pada Tabel 5, seluruh variabel penelitian berkontribusi terhadap proses klasifikasi menggunakan algoritma *Naïve Bayes*. Namun, beberapa variabel menunjukkan perbedaan probabilitas yang lebih besar antara stadium awal dan stadium lanjut sehingga memberikan kontribusi yang lebih dominan dalam proses klasifikasi.

Salah satu variabel yang menunjukkan perbedaan cukup jelas adalah **benjolan di ketiak**. Hasil penelitian menunjukkan bahwa pasien yang tidak mengalami benjolan di ketiak lebih banyak berada pada kelompok stadium awal, sedangkan keberadaan benjolan di ketiak lebih dominan ditemukan pada pasien stadium lanjut. Temuan ini mengindikasikan bahwa keterlibatan kelenjar getah bening aksila merupakan salah satu karakteristik klinis yang berkaitan dengan progresivitas kanker payudara.

Karakteristik serupa juga terlihat pada variabel **perubahan bentuk dan ukuran payudara**. Pasien yang tidak mengalami perubahan bentuk dan ukuran payudara lebih banyak ditemukan pada stadium awal, sedangkan pasien yang mengalami perubahan tersebut cenderung berada pada stadium lanjut. Hal ini menunjukkan bahwa perubahan morfologi payudara semakin nyata seiring dengan perkembangan penyakit, sehingga dapat menjadi salah satu indikator klinis dalam proses klasifikasi stadium.

Variabel **kelainan pada kulit payudara** juga memperlihatkan pola yang serupa. Pasien yang tidak mengalami kelainan pada kulit payudara lebih banyak termasuk dalam kelompok stadium awal, sedangkan kelainan pada kulit payudara lebih sering ditemukan pada pasien stadium lanjut. Temuan ini menunjukkan bahwa perubahan pada jaringan kulit merupakan manifestasi klinis yang umumnya muncul pada stadium penyakit yang lebih lanjut akibat penyebaran tumor ke jaringan di sekitarnya.

Selain itu, **ukuran tumor payudara** menjadi salah satu variabel yang memiliki perbedaan distribusi paling jelas. Pasien dengan ukuran tumor 2–5 cm lebih banyak ditemukan pada kelompok stadium awal, sedangkan ukuran tumor lebih dari 5 cm didominasi oleh pasien stadium lanjut. Hasil tersebut menunjukkan bahwa peningkatan ukuran tumor berkaitan dengan perkembangan stadium kanker payudara. Temuan ini juga sesuai dengan konsep klasifikasi TNM yang menempatkan ukuran tumor (Tumor) sebagai salah satu komponen utama dalam menentukan stadium penyakit.

Secara keseluruhan, karakteristik klinis yang menunjukkan perbedaan distribusi antara stadium awal dan stadium lanjut memiliki kontribusi terhadap proses perhitungan probabilitas pada algoritma *Naïve Bayes*. Semakin besar perbedaan distribusi suatu variabel terhadap masing-masing kelas, semakin besar pula pengaruhnya dalam membedakan stadium kanker payudara. Hal ini menunjukkan bahwa data rekam medis yang memuat karakteristik klinis pasien dapat dimanfaatkan tidak hanya sebagai dokumentasi pelayanan kesehatan, tetapi juga sebagai sumber data yang mendukung pengembangan model klasifikasi berbasis *machine learning*.

4. KESIMPULAN

Penelitian ini berhasil membangun model klasifikasi stadium kanker payudara menggunakan algoritma *Naïve Bayes* berdasarkan data rekam medis pasien di Rumah Sakit Tingkat III Baladhika Husada Jember. Melalui tahapan preprocessing yang meliputi *data selection*, *data cleaning*, dan *data transformation* dengan *categorical encoding*, diperoleh dataset sebanyak 476 data rekam medis yang digunakan dalam proses pembentukan model klasifikasi. Hasil pengujian pada sembilan skenario pembagian *data training-testing split* menunjukkan bahwa rasio 90:10 menghasilkan performa terbaik dengan nilai *accuracy* sebesar 87,50%, *precision* sebesar 86,36%, dan *recall* sebesar 86,36%.

Analisis karakteristik klinis menunjukkan bahwa variabel benjolan di ketiak, perubahan bentuk dan ukuran payudara, kelainan pada kulit payudara, serta ukuran tumor memiliki kecenderungan distribusi yang lebih dominan pada kelompok stadium lanjut sehingga memberikan kontribusi terhadap proses klasifikasi menggunakan algoritma *Naïve Bayes*. Temuan tersebut menunjukkan bahwa karakteristik klinis yang terdokumentasi dalam rekam medis memiliki potensi untuk dimanfaatkan sebagai dasar pembentukan model klasifikasi stadium kanker payudara.

Secara keseluruhan, penelitian ini menunjukkan bahwa algoritma *Naïve Bayes* memiliki performa yang baik dalam mengklasifikasikan stadium kanker payudara berdasarkan data rekam medis. Model yang dihasilkan berpotensi mendukung pengembangan sistem pendukung keputusan klinis untuk membantu proses identifikasi stadium secara lebih cepat dan objektif, sehingga dapat menjadi pelengkap dalam pengambilan keputusan oleh tenaga kesehatan.

REFERENSI

- [1] M. Cherlady Anastasia and C. Catherine, “Gambaran Tingkat Pengetahuan Mahasiswa Fakultas Kedokteran Tentang Penggunaan Ultrasonografi Payudara Dalam Deteksi Dini Tumor Payudara,” *Journal of Public Health Science (JoPHS)*, vol. 2, no. 4, pp. 655–661, Dec. 2025, doi: 10.70248/jophs.v2i4.3512.
- [2] Menteri Kesehatan Republik Indonesia, *Keputusan Menteri Kesehatan Republik Indonesia Nomor Hk.01.07/Menkes/414/2018 Tentang Pedoman Nasional Pelayanan Kedokteran Tata Laksana Kanker Payudara*. Indonesia, 2018.
- [3] A. Herawati *et al.*, “Karakteristik Kanker Panyudara,” *Fakumi Medical Journal: Jurnal Mahasiswa Kedokteran*, vol. 1, no. 1, pp. 44–53, Oct. 2021, doi: 10.33096/FMJ.V1I1.76.
- [4] K. Suparna and L. M. K. K. S. Sari, “Kanker Payudara: Diagnostik, Faktor Risiko, dan Stadium,” *Ganesha Medicina*, vol. 2, no. 1, pp. 42–48, Jun. 2022, doi: 10.23887/GM.V2I1.47032.
- [5] Muhammad Arhami and Muhammad Nasir, *Data Mining, Algoritma Dan Implementasi*, no. 9. Andi Offset, 2020. Accessed: Jul. 29, 2026. [Online]. Available: <https://repositori.telkomuniversity.ac.id/home/catalog/id/191497/slug/data-mining-algoritma-dan-implementasi.html>
- [6] A. Sirojul Munir *et al.*, “Perbandingan Akurasi Algoritma Naive Bayes dan Algoritma Decision Tree dalam Pengklasifikasian Penyakit Kanker Payudara,” *Jurnal Ilmiah Informatika Global*, vol. 15, no. 1, pp. 23–29, Apr. 2024, doi: 10.36982/JIIG.V15I1.3578.
- [7] J. F. Indey and S. Supangat, “Implementasi Algoritma Naïve Bayes dalam Sistem Pengarsipan Surat Berbasis AI di GPI Papua Klasik Mimika Papua Tengah,” *JURNAL ILMIAH INFORMATIKA*, vol. 12, no. 02, pp. 102–113, Sep. 2024, doi: 10.33884/JIF.V12I02.9087.
- [8] N. P. Laksono, A. A. Syaaiyullah, and A. Y. P. Yusuf, “Klasifikasi Penentuan Siswa Berprestasi Menggunakan Algoritma Naïve Bayes Classifier Di PT.Yes Study Education Group Indonesia,” *Jurnal Riset Informatika dan Teknologi Informasi*, vol. 2, no. 3, pp. 244–252, Jul. 2025, doi: 10.58776/JRITI.V2I3.158.