Jurnal Manajemen Teknologi dan Sistem Informasi (JMS)
Volume 5, Nomor 2, September 2025
ISSN 2808-5450 (media cetak), ISSN 2808-5019 (media online)
Available Online at https://ejournal.unama.ac.id/index.php/jms
DOI https://doi.org/10.33998/jms.v5i2

Peningkatan Performa *Naïve Bayes* dengan *Information Gain* Menggunakan *Machine Learning* untuk Klasifikasi Kanker Payudara

Mardiana[1], Jasmir[2], Sharipuddin[3]
Fakultas Ilmu Komputer, Magister Sistem Informasi, Universitas Dinamika Bangsa, Jambi, Indonesia
Email: [1]mardianacumel@gmail.com, [2]ijay_jasmir@yahoo.com, [3]sharifbuhaira@gmail.com
Email Penulis Korespondensi: mardianacumel@gmail.com

Submitted :
19 April 2025

Revision :
24 Juni 2025

Accepted:
25 September 2025

Published:
30 September 2025

Abstrak−Penelitian ini bertujuan untuk meningkatkan performa algoritma *Naïve Bayes* dalam klasifikasi diagnosis kanker payudara dengan mengintegrasikan metode seleksi fitur *Information Gain*. *Dataset* yang digunakan adalah Kanker Payudara Wisconsin (Diagnostik) yang terdiri dari 569 sampel. Penelitian ini menguji efektivitas metode seleksi fitur dalam meningkatkan akurasi, sensitivitas, dan spesifisitas model klasifikasi. Implementasi seleksi fitur *Information Gain* berhasil meningkatkan akurasi model *Naïve Bayes* dari 94.15% menjadi 96.49%, dengan peningkatan sebesar 2.34%. Penambahan metode seleksi fitur ini terbukti signifikan dalam meningkatkan kemampuan prediktif model. Hasil penelitian ini dapat mendukung pengambilan keputusan medis yang lebih tepat, berpotensi mempengaruhi keputusan pengobatan dan hasil pasien dalam praktik klinis. Penelitian ini memberikan wawasan baru mengenai penerapan *machine learning* dalam diagnostik medis dan menyarankan langkah-langkah lanjutan untuk penelitian lebih dalam di masa depan.

Kata Kunci: *Naïve Bayes*; *Information Gain*; Klasifikasi; Kanker Payudara

Abstract−*This study aims to enhance the performance of the Naïve Bayes algorithm in breast cancer diagnosis classification by integrating the Information Gain feature selection method. The dataset used is the Wisconsin Breast Cancer (Diagnostic) dataset, consisting of 569 samples. This study evaluates the effectiveness of feature selection in improving the accuracy, sensitivity, and specificity of the classification model. The implementation of the Information Gain feature selection method successfully increased the Naïve Bayes model's accuracy from 94.15% to 96.49%, a 2.34% improvement. The addition of feature selection significantly enhanced the predictive capability of the model. The findings of this study can support more accurate medical decision-making, potentially influencing treatment decisions and patient outcomes in clinical practice. This research provides new insights into the application of machine learning in medical diagnostics and suggests directions for future research.*

Keywords*: Naïve Bayes; Information Gain; Classification; Breast Cancer*

1. PENDAHULUAN

Data mining merupakan teknik penting dalam analisis data besar, yang membantu mengekstrak informasi dan pengetahuan bermanfaat, termasuk dalam diagnosis kanker payudara, salah satu penyebab kematian terbanyak pada wanita [1]. Metode seperti algoritma *Naïve Bayes*, didukung oleh seleksi fitur *Information Gain*, dapat meningkatkan akurasi dan efisiensi diagnosa kanker payudara. Seiring perkembangan teknologi digital, *machine learning*, sebagai cabang kecerdasan buatan, memfasilitasi analisis data besar dan kompleks, serta memungkinkan pembuatan model prediksi yang lebih akurat [2]. *Machine learning* atau pembelajaran mesin merupakan penerapan kecerdasan buatan yang memberikan sistem kemampuan belajar secara otomatis dari sekumpulan data untuk melakukan tugas tertentu tanpa diprogram secara eksplisit [3].

Secara umum, metode pada *machine learning* dibagi menjadi empat tipe berdasarkan cara pembelajarannya pada penelitian ini, menggunakan *supervised learning*, yaitu metode yang memetakan data *input* dan *output* berdasarkan contoh yang tersedia atau *dataset* [4]. *Dataset* masukan terdiri dari *training set* dan *testing set*, di mana *training set* memiliki variabel *output* yang perlu diprediksi atau diklasifikasi. Metode ini mempelajari pola dari *training set* dan menerapkannya ke setiap *testing set* untuk melakukan prediksi atau klasifikasi. Klasifikasi merupakan tugas yang mengategorikan objek data ke dalam kelas-kelas yang telah ditentukan. Terdapat dua fungsi utama dalam klasifikasi, pertama mengembangkan model yang bertindak sebagai prototipe dan disimpan dalam memori, kedua menggunakan model tersebut untuk memprediksi kategori dari objek data baru berdasarkan kelas-kelas yang ada [5]. Dalam dunia medis, penerapan *machine learning* dapat mempercepat diagnosis kanker payudara dan mendukung keputusan medis yang lebih baik.

Kanker payudara merupakan masalah kesehatan global dengan tingkat morbiditas dan mortalitas tinggi [6], dan kemajuan teknologi informasi serta data mining membuka peluang untuk metode diagnostik yang lebih efisien. Algoritma *Naïve Bayes*, yang menggunakan teorema *Bayes* untuk prediksi berbasis probabilitas, efektif dalam klasifikasi cepat dan akurat, terutama untuk membedakan antara tumor jinak dan ganas pada kanker payudara. *Naïve Bayes Classifier* (NBC) merupakan salah satu metode pembelajaran mesin yang memanfaatkan perhitungan probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi probabilitas

Volume 5, Nomor 2, September 2025
ISSN 2808-5450 (media cetak), ISSN 2808-5019 (media online)
Available Online at https://ejournal.unama.ac.id/index.php/jms
DOI https://doi.org/10.33998/jms.v5i2

di masa depan berdasarkan pengalaman di masa sebelumnya [7]. *Naïve Bayes* merupakan teknik prediksi berbasis *probabilistic* sederhana yang berdasar pada penerapan teorema *Bayes* (atau aturan *Bayes*) dengan asumsi independensi (ketidaktergantungan) yang kuat [8]. Penelitian ini menggunakan *dataset* Kanker Payudara Wisconsin (Diagnostik) yang mencakup 569 sampel dengan 31 fitur dari gambar digital aspirasi jarum halus (FNA). Dengan optimasi algoritma *Naïve Bayes* melalui seleksi fitur *Information Gain*, penelitian ini bertujuan meningkatkan akurasi diagnosa dan mendukung penanganan kanker yang lebih efektif, mengingat meningkatnya kasus kanker payudara di Indonesia.

Dalam proses analisis data, *feature selection* memainkan peran penting dalam memilih fitur yang relevan, atau subset calon fitur, yang akan meningkatkan efisiensi model klasifikasi [9]. *Information Gain* digunakan untuk mengukur pentingnya setiap fitur dalam pengambilan keputusan. Nilai *Information Gain* dari atribut digunakan untuk menentukan bobot setiap atribut yang digunakan untuk klasifikasi, dengan tujuan mengidentifikasi fitur yang paling diperlukan dalam proses klasifikasi [10]. Dengan menghitung nilai entropi terlebih dahulu, yang merupakan ukuran ketidakpastian dalam distribusi data, *Information Gain* membantu mengurangi *noise* yang disebabkan oleh fitur yang tidak relevan [11]. Dalam pembuatan model klasifikasi, pembagian data menjadi dua bagian, yaitu data latih dan data uji, sangat penting. Data latih digunakan untuk melatih model, sementara data uji digunakan untuk menguji keakuratan model yang telah dilatih [12]. Rasio pembagian data ini sangat bergantung pada ukuran *dataset* yang dimiliki, dan pembagian ini tidak memiliki pedoman atau metrik spesifik yang mengatur [13]. *Python*, sebagai bahasa pemrograman yang sering digunakan untuk analisis data, memberikan fleksibilitas dalam membangun model, mengotomatisasi tugas, dan memproses *dataset* besar secara efisien [14].

Penelitian ini fokus pada penggunaan algoritma *Naïve Bayes* untuk klasifikasi diagnosa kanker payudara dengan peningkatan melalui seleksi fitur *Information Gain*. Evaluasi model dilakukan menggunakan *dataset* Kanker Payudara Wisconsin (Diagnostik), dengan mengukur akurasi, sensitivitas, dan spesifisitas. *Naïve Bayes* efektif untuk *dataset* besar, namun memerlukan pemilihan fitur yang cermat untuk menghindari penurunan akurasi akibat atribut berlebihan. Menurut Kumar dalam jurnal Lila Dini Utami [15] seleksi fitur, seperti *Information Gain*, yang diakui sebagai salah satu yang terbaik, dapat meningkatkan efisiensi dan efektivitas pengklasifikasi dengan mengurangi jumlah data yang dianalisis.

Sebagai tinjauan atas penelitian-penelitian terdahulu, pada penelitian yang dilakukan oleh Muhammad Ramanda Hasibuan dan Marji [16], metode *Information Gain* berhasil meningkatkan akurasi klasifikasi penyakit gagal ginjal hingga 96,8% menggunakan *Modified K-Nearest Neighbor*, signifikan lebih tinggi dari sistem tanpa *Information Gain* yang mencapai 79,9%. Penelitian lanjutan oleh Avira Budianita [17] mengaplikasikan *Naive Bayes Classifier* dengan *Information Gain*, menghasilkan peningkatan akurasi prediksi waktu kelulusan mahasiswa dari 81,99% menjadi 83,60%, membuktikan efektivitas seleksi fitur dalam mengurangi kompleksitas data. Rahmanita [18] melanjutkan penggunaan kedua metode tersebut untuk klasifikasi penyakit dan hama tanaman jagung, mencapai akurasi 98,47% dan mempercepat proses diagnosa. Amelia Isnanda [19] meneliti penggunaan dalam analisis sentimen *e-wallet* selama pandemi, meningkatkan akurasi dari 84% menjadi 92% dengan *Information Gain*. Albet Dwi Pangestu [20] juga menerapkan *Naïve Bayes* dan *Information Gain* untuk analisis sentimen terhadap kebijakan PPKM Darurat, mencatat peningkatan performa dengan akurasi mencapai 81%. Keseluruhan penelitian ini menunjukkan signifikansi *Information Gain* dalam meningkatkan efektivitas algoritma *Naïve Bayes* di berbagai aplikasi data besar dan kompleks.

Berdasarkan kajian penelitian sebelumnya dan analisis permasalahan yang telah dijelaskan, penerapan metode data mining menggunakan algoritma *Naïve Bayes* dengan teknik seleksi fitur *Information Gain* dinilai tepat untuk klasifikasi diagnosa kanker payudara. *Naïve Bayes* dipilih karena efisiensinya dalam mengolah data besar dan kemudahan implementasinya, namun tetap memerlukan optimisasi fitur. *Information Gain* efektif dalam meningkatkan akurasi dengan memilih atribut yang paling relevan, sehingga mengurangi kompleksitas dan mempercepat proses. Penerapan *Information Gain* telah terbukti efektif di berbagai bidang klasifikasi, seperti pada penyakit gagal ginjal, analisis waktu kelulusan mahasiswa, klasifikasi penyakit dan hama tanaman, serta analisis sentimen *e-wallet* dan kebijakan publik. Dengan teknik ini, penelitian diharapkan menghasilkan model klasifikasi yang lebih akurat dan efisien, serta mendukung pengambilan keputusan dalam diagnosis kanker payudara secara lebih tepat.

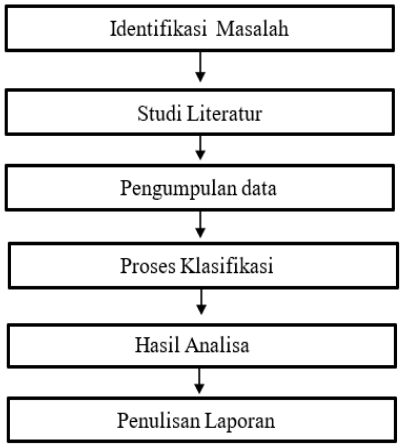
Penelitian ini bertujuan untuk menghasilkan model klasifikasi yang efektif dalam mendiagnosa kanker payudara menggunakan algoritma *Naïve Bayes* yang dikombinasikan dengan metode seleksi fitur *Information Gain*, mengetahui tingkat akurasi model klasifikasi yang dikembangkan dalam mengidentifikasi kanker payudara berdasarkan *dataset* yang tersedia, serta mengevaluasi kinerja algoritma *Naïve Bayes* dengan seleksi fitur *Information Gain* dibandingkan dengan metode klasifikasi lainnya. Peneliti mengharapkan hasil penelitian ini dapat meningkatkan keefektifan diagnostik kanker payudara di fasilitas kesehatan dengan menyediakan model klasifikasi yang lebih akurat, membantu pengambilan keputusan medis yang lebih tepat, memberikan wawasan mendalam mengenai tingkat akurasi dan keandalan model klasifikasi yang dikembangkan, serta memberikan perbandingan yang signifikan antara algoritma *Naïve Bayes* yang disempurnakan dengan *Information Gain* dan metode klasifikasi lainnya, yang berkontribusi pada penelitian lebih lanjut dalam peningkatan metode diagnostik untuk kanker payudara.

Jurnal Manajemen Teknologi dan Sistem Informasi (JMS)
Volume 5, Nomor 2, September 2025
ISSN 2808-5450 (media cetak), ISSN 2808-5019 (media online)
Available Online at https://ejournal.unama.ac.id/index.php/jms
DOI https://doi.org/10.33998/jms.v5i2

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Kerangka kerja penelitian yang digunakan dalam proses penelitian pada dasarnya merupakan urutan langkah-langkah yang harus dilakukan sehingga tujuan akhir dari penelitian dapat tercapai dan siap untuk diimplementasikan. Adapun kerangka kerja penelitian yang peneliti gunakan dapat dilihat pada gambar berikut :

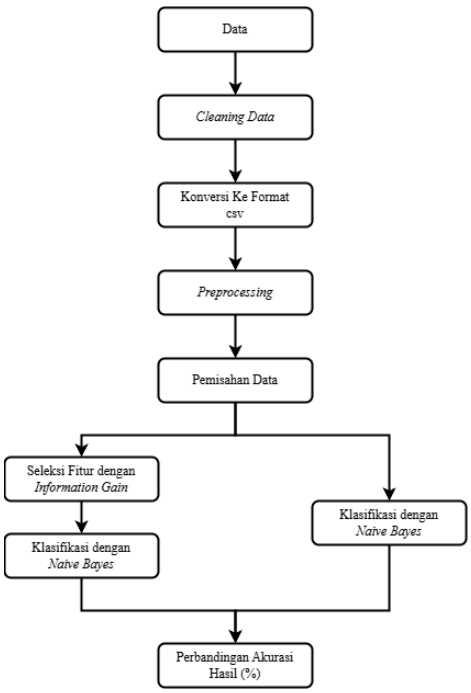


Gambar 1. Tahapan Penelitian

Berdasarkan tahapan penelitian yang telah digambarkan di atas, maka dapat diuraikan pembahasan dari masing-masing tahapan penelitian yaitu dimulai dengan identifikasi masalah, yaitu mengidentifikasi tantangan dalam diagnosis kanker payudara menggunakan teknik data mining, khususnya dalam hal akurasi dan reliabilitas algoritma klasifikasi dalam data medis kompleks. Selanjutnya, dilakukan studi literatur untuk memperoleh landasan teori terkait teknik data mining, proses klasifikasi menggunakan algoritma *Naïve Bayes*, dan seleksi fitur *Information Gain*. Pengumpulan data dilakukan dengan mengambil *dataset* Kanker Payudara Wisconsin (Diagnostik) dari Kaggle untuk analisis lebih lanjut. Proses klasifikasi dilakukan dengan mengikuti beberapa tahapan yang telah ditentukan, dan hasil analisis dievaluasi berdasarkan keakuratan model yang dihasilkan. Hasil terbaik akan dipilih untuk evaluasi lebih lanjut. Terakhir, laporan penelitian disusun sebagai dokumentasi yang dapat digunakan oleh peneliti di masa depan.

2.2 Tahapan Klasifikasi

Pengklasifikasian data dilakukan dengan beberapa tahapan. Adapun alur tahapan pengklasifikasian dapat dilihat pada gambar berikut :



Gambar 2. Tahapan Klasifikasi

Jurnal Manajemen Teknologi dan Sistem Informasi (JMS)
Volume 5, Nomor 2, September 2025
ISSN 2808-5450 (media cetak), ISSN 2808-5019 (media online)
Available Online at https://ejournal.unama.ac.id/index.php/jms
DOI https://doi.org/10.33998/jms.v5i2

Berdasarkan tahapan klasifikasi yang telah digambarkan di atas, maka dapat diuraikan pembahasan dari masing-masing tahapan klasifikasi yang dimulai dengan pengumpulan dan persiapan *dataset* yang mencakup berbagai atribut penting. Selanjutnya, dilakukan *cleaning* data untuk menghilangkan inkonsistensi dan kesalahan dalam *dataset*, kemudian data dikonversi ke format CSV untuk memudahkan pengolahan. Setelah itu, *preprocessing* data dilakukan dengan normalisasi, pengkodean variabel kategorikal, dan pengolahan data teks. Data yang telah diproses kemudian dibagi menjadi set pelatihan dan pengujian untuk menghindari *overfitting*. Seleksi fitur dilakukan menggunakan *Information Gain* untuk memilih fitur yang relevan, diikuti dengan klasifikasi menggunakan algoritma *Naïve Bayes*. Akhirnya, dilakukan perbandingan akurasi model dengan membandingkan prediksi terhadap data uji untuk mengevaluasi kinerja model.

2.3 Bahan Penelitian

Pada penelitian ini, digunakan Kumpulan Data Kanker Payudara Wisconsin (Diagnostik) yang berisi 569 sampel dengan 31 fitur yang diekstrak dari gambar *Fine Needle Aspiration* (FNA) massa payudara. *Dataset* publik ini, yang telah diberi label untuk mengklasifikasikan kanker payudara sebagai jinak (B = *Benign*) dan ganas (M = *Malignant*), diperoleh dari Kaggle, sebuah platform online yang menyediakan berbagai kumpulan data, kode, dan kompetisi untuk para penggiat *data science* dan *machine learning*. Berikut adalah informasi dari *dataset* yang dapat dilihat pada tabel berikut :

Tabel 1. Informasi *Dataset*

Nama	:	Breast Cancer Wisconsin (Diagnostic) Data Set
Ukuran	:	125.2 kB
Pemilik	:	Ovsen dan UCI Machine Learning
Bahasa	:	Inggris
Lisensi	:	https://creativecommons.org/licenses/by-nc-sa/4.0/
Tautan	:	https://www.kaggle.com/*datasets*/uciml/breast-cancer-wisconsin-data

3. HASIL DAN PEMBAHASAN

3.1 Representasi Data

Kumpulan data yang digunakan dalam penelitian ini adalah Kumpulan Data Kanker Payudara Wisconsin (Diagnostik), yang berisi 569 sampel dengan 31 fitur yang diekstrak dari gambar digital hasil aspirasi jarum halus *Fine Needle Aspiration* (FNA) pada massa payudara. Data ini digunakan untuk mengklasifikasikan kanker payudara sebagai ganas atau jinak, yang merupakan diagnosis penting dalam penelitian medis dan klinis. Proses seleksi fitur dengan metode *Information Gain* diterapkan untuk mengidentifikasi fitur-fitur yang paling relevan dan informatif bagi model, dengan tujuan meningkatkan performa klasifikasi menggunakan model *Naïve Bayes* dan memastikan bahwa fitur yang digunakan dalam prediksi memberikan kontribusi signifikan terhadap akurasi model. Dengan analisis data dan seleksi fitur yang tepat, penelitian ini bertujuan untuk meningkatkan akurasi klasifikasi diagnosis kanker payudara serta mengoptimalkan hasil penelitian sebelumnya, memastikan efisiensi dalam proses klasifikasi dan mengidentifikasi fitur yang dapat memberikan informasi lebih akurat tentang kanker payudara. *Dataset* ini, yang berasal dari sumber terpercaya seperti Kaggle, memiliki distribusi kelas dengan 357 sampel jinak (B = *Benign*) dan 212 sampel ganas (M = *Malignant*). Proses seleksi fitur dan penggunaan metode *Naïve Bayes*, serta peningkatan performa menggunakan *Information Gain*, diharapkan dapat menghasilkan model yang lebih akurat dalam mengklasifikasikan jenis kanker, yang merupakan langkah penting dalam mendukung keputusan medis yang lebih baik.

Tabel 2. Data Diagnosis Kanker Payudara

No	id	diagnosis	radius_mean	texture_mean	...	fractal_dimension_worst
1	842302	M	0,777083	10.38	...	0,825694
2	842517	M	20.57	0,761806	...	0.08902
3	84300903	M	0,839583	21.25	...	0.08758
4	84348301	M	11.42	20.38	...	0,120139
5	84358402	M	20.29	14.34	...	0.07678
6	843786	M	12.45	15.07	...	0,863889
7	844359	M	18.25	0,859722	...	0.08368
8	84458202	M	0,590972	0,890972	...	0,799306
9	844981	M	13	0,931944	...	0,744444
10	84501001	M	12.46	24.04.00	...	1,440972
11	845636	M	16.02	23.24	...	0,727778
12	84610002	M	0,679167	0,770139	...	0,710417
13	846226	M	19.17	24.08.00	...	0.06287

Jurnal Manajemen Teknologi dan Sistem Informasi (JMS)
Volume 5, Nomor 2, September 2025
ISSN 2808-5450 (media cetak), ISSN 2808-5019 (media online)
Available Online at https://ejournal.unama.ac.id/index.php/jms
DOI https://doi.org/10.33998/jms.v5i2

14	846381	M	0,684028	1,024306	...	0,99375
15	84667401	M	0,592361	0,959028	...	0,93125
16	84799002	M	14.54	27.54.00	...	0.08216
17	848406	M	0,630556	20.13	...	0,793056
18	84862001	M	16.13	0,880556	...	0.07615
19	849014	M	0,847917	22.15	...	0.07259
20	8510426	B	13.54	14.36	...	0.08183
...	...	...	...	...	...	...
569	92751	B	0,344444	24.54.00	...	0.07039

3.1.1 Pemilihan Atribut

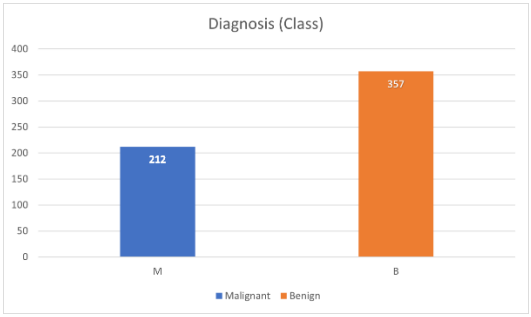
Proses seleksi fitur menggunakan *Information Gain* menghasilkan perangkingan fitur berdasarkan bobot, dan dalam penelitian ini, dipilih 8 fitur teratas dari 30 fitur yang ada. Fitur-fitur terpilih ini kemudian digunakan dalam proses klasifikasi. Hasil seleksi fitur dapat dilihat pada tabel berikut :

Tabel 3. Atribut Terpilih

No.	Atribut	Bobot
1	perimeter_worst	0.453885
2	concave points_mean	0.447369
3	concave points_worst	0.444939
4	area_worst	0.438863
5	radius_worst	0.437821
6	perimeter_mean	0.371635
7	concavity_mean	0.370734
8	area_mean	0.329785

3.1.2 Distribusi Kelas

Analisis distribusi kelas dalam *dataset* kanker payudara menunjukkan ketidakseimbangan antara kedua kelas, dengan 357 sampel *Benign* (B) dan 212 sampel *Malignant* (M). Proporsi kelas *Benign* lebih dominan dibandingkan *Malignant*. Adapun distribusi kelas dapat dilihat pada gambar berikut :



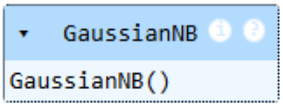
Gambar 3. Visualisasi Kelas Dalam *Dataset*

3.2 Implementasi Algoritma *Naïve Bayes*

Pada tahap implementasi, digunakan model Gaussian *Naïve Bayes* (GaussianNB) untuk melatih model menggunakan data pelatihan yang tersedia. Model ini dilatih dengan menggunakan perintah model.fit(X_train, y_train), yang memungkinkan model untuk mempelajari pola dari data dan siap untuk melakukan prediksi.

```
model = GaussianNB()
model.fit(X_train, y_train)
```

Berikut adalah output yang dihasilkan dari proses pembuatan dan pelatihan model dapat dilihat pada gambar berikut :



Gambar 4. *Output* Pembuatan dan Pelatihan Model

Jurnal Manajemen Teknologi dan Sistem Informasi (JMS)
Volume 5, Nomor 2, September 2025
ISSN 2808-5450 (media cetak), ISSN 2808-5019 (media online)
Available Online at https://ejournal.unama.ac.id/index.php/jms
DOI https://doi.org/10.33998/jms.v5i2

Setelah model dilatih, evaluasi kinerjanya dilakukan dengan menghitung akurasi, confusion matrix, dan classification report menggunakan data uji. Nilai akurasi dihitung menggunakan fungsi accuracy_score(y_test, y_pred), sementara confusion matrix dan classification report memberikan gambaran tentang kinerja model pada setiap kelas, termasuk precision, *recall*, dan f1-score.

```
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)
class_report = classification_report(y_test, y_pred)
```

Hasil evaluasi model ditampilkan dengan mencetak confusion matrix dan laporan klasifikasi, serta akurasi model yang dihitung. Output ini memberikan informasi penting mengenai performa model *Naïve Bayes* pada *dataset* yang digunakan, dengan akurasi yang dihasilkan tercatat pada Gambar 5.

```
print('Confusion Matriks :')
print(conf_matrix)
print('\nLaporan Klasifikasi :')
print(class_report)
print('Akurasi =', accuracy*100, '%')
```

Output ini memberikan informasi penting mengenai performa model *Naïve Bayes* pada *dataset* yang digunakan, dengan akurasi yang dihasilkan dapat dilihat pada gambar berikut :

```
Confusion Matriks :
[[104   4]
 [  6  57]]

Laporan Klasifikasi :
              precision    recall  f1-score   support

           B       0.95      0.96      0.95       108
           M       0.93      0.90      0.92        63

    accuracy                           0.94       171
   macro avg       0.94      0.93      0.94       171
weighted avg       0.94      0.94      0.94       171

Akurasi = 94.15204678362574 %
```

Gambar 5. Hasil Evaluasi Model *Naïve Bayes*

3.3 Implementasi Algoritma *Naïve Bayes* dengan Seleksi Fitur *Information Gain*

Pada tahap ini, dilakukan perhitungan mutual information menggunakan mutual_info_classif untuk mengevaluasi pentingnya fitur dalam memprediksi kelas target pada data pelatihan.

```
mutual_info = mutual_info_classif(X_train, y_train)
```

Selanjutnya, teknik seleksi fitur SelectKBest digunakan untuk memilih 8 fitur teratas berdasarkan nilai mutual information yang dihitung. Data pelatihan dan pengujian kemudian diproses untuk menggunakan fitur-fitur terpilih, yaitu X_train_selected dan X_test_selected.

```
selector = SelectKBest(mutual_info_classif, k=8)
X_train_selected = selector.fit_transform(X_train, y_train)
X_test_selected = selector.transform(X_test)
```

Setelah seleksi fitur dilakukan, model Gaussian *Naïve Bayes* (GaussianNB) dilatih menggunakan data yang telah diseleksi fiturnya dengan perintah clf.fit(X_train_selected, y_train).

```
clf = GaussianNB()
clf.fit(X_train_selected, y_train)
```

Output dari pembuatan dan pelatihan model ini dapat dilihat pada gambar berikut :

```
▾ GaussianNB
GaussianNB()
```

Gambar 6. *Output* Pembuatan dan Pelatihan Model

Jurnal Manajemen Teknologi dan Sistem Informasi (JMS)
Volume 5, Nomor 2, September 2025
ISSN 2808-5450 (media cetak), ISSN 2808-5019 (media online)
Available Online at https://ejournal.unama.ac.id/index.php/jms
DOI https://doi.org/10.33998/jms.v5i2

Setelah model dilatih, nilai mutual information yang dihitung sebelumnya diubah menjadi format yang mudah dibaca dan diurutkan berdasarkan kontribusinya terhadap klasifikasi.

```
mutual_info = pd.Series(mutual_info)
mutual_info.index = X.columns
mutual_info.sort_values(ascending=False)
```

Hasil pengurutan nilai *Information Gain* dapat dilihat pada gambar berikut :

	0
perimeter_worst	0.451349
concave points_mean	0.446502
concave points_worst	0.445520
area_worst	0.439701
radius_worst	0.435074
perimeter_mean	0.372261
concavity_mean	0.370585
area_mean	0.329743
radius_mean	0.326418
area_se	0.315506
concavity_worst	0.314809
perimeter_se	0.251590
radius_se	0.251439
compactness_worst	0.212275
compactness_mean	0.205864
texture_worst	0.168651
concavity_se	0.123108
concave points_se	0.108744
texture_mean	0.103488

Gambar 7. *Output* Pengurutan Nilai *Information Gain*

Kemudian, dilakukan evaluasi model dengan menghitung akurasi, confusion matrix, dan classification report pada data uji yang telah dipilih fiturnya. Evaluasi ini memberikan gambaran lengkap mengenai kinerja model *Naïve Bayes* yang sudah dioptimalkan dengan seleksi fitur *Information Gain*.

```
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)
class_report = classification_report(y_test, y_pred)

print('Confusion Matriks :')
print(conf_matrix)
print('\nLaporan Klasifikasi :')
print(class_report)
print('Akurasi =', accuracy*100, '%')
```

Hasil evaluasi model yang menunjukkan akurasi dan kinerja model secara keseluruhan dapat dilihat pada gambar berikut :

```
Confusion Matriks :
[[107   1]
 [  5  58]]

Laporan Klasifikasi :
              precision    recall  f1-score   support

           B       0.96      0.99      0.97       108
           M       0.98      0.92      0.95        63

    accuracy                           0.96       171
   macro avg       0.97      0.96      0.96       171
weighted avg       0.97      0.96      0.96       171

Akurasi = 96.49122807017544 %
```

Gambar 8. Hasil Evaluasi Model *Naïve Bayes* Dengan Seleksi Fitur *Information Gain*

3.4 Hasil Evaluasi Akurasi

Setelah dilakukan analisis klasifikasi menggunakan *Naïve Bayes* dan seleksi fitur menggunakan *Information Gain*, dilakukan perbandingan antara model tanpa seleksi fitur dan model yang menggunakan seleksi fitur. Hasil dari analisis ini dapat dilihat pada tabel berikut :

Volume 5, Nomor 2, September 2025
ISSN 2808-5450 (media cetak), ISSN 2808-5019 (media online)
Available Online at https://ejournal.unama.ac.id/index.php/jms
DOI https://doi.org/10.33998/jms.v5i2

Tabel 4. Perbandingan Hasil Selisih

Metrik Evaluasi	***Naïve Bayes* tanpa IG**	***Naïve Bayes* dengan IG**	**Selisih**
Jumlah Fitur	30	8	-22
Presisi *Benign* (B)	0.95	0.96	+0.01
Presisi *Malignant* (M)	0.93	0.98	+0.05
Recall Benign (B)	0.96	0.99	+0.03
Recall Malignant (M)	0.90	0.92	+0.02
F1-Score *Benign* (B)	0.95	0.97	+0.02
F1-Score *Malignant* (M)	0.92	0.95	+0.03
Akurasi	94.15%	96.49%	+2.34%

Penggunaan seleksi fitur *Information Gain* dalam model *Naïve Bayes* telah secara signifikan meningkatkan kinerja klasifikasi kanker payudara, memberikan peningkatan akurasi yang substansial dari 94.15% ke 96.49%, dengan selisih peningkatan sebesar +2.34%. Ini terutama terlihat pada kelas *Benign*, dimana terdapat peningkatan presisi, *recall*, dan F1-Score yang menandakan kemampuan model yang lebih baik dalam mengidentifikasi kasus yang benar-benar *Benign.* Seleksi fitur ini efektif dalam mengidentifikasi dan memprioritaskan fitur-fitur yang paling berkontribusi terhadap informasi tentang target klasifikasi, mengurangi kompleksitas model dan meningkatkan keakuratannya secara signifikan. Kehadiran peningkatan ini menunjukkan betapa krusialnya teknik seleksi fitur dalam meningkatkan kinerja model klasifikasi, terutama dalam aplikasi diagnostik medis di mana akurasi adalah sangat penting.

Penambahan seleksi fitur *Information Gain* pada model *Naïve Bayes* secara signifikan meningkatkan akurasi dalam klasifikasi kanker payudara karena *Information Gain* secara efektif mengidentifikasi dan memprioritaskan fitur yang paling berkontribusi terhadap peningkatan informasi tentang kategori target. Teknik ini mengukur penurunan entropi yang menghasilkan *dataset* yang lebih murni setelah fitur tertentu digunakan untuk pemisahan, membantu mengurangi kompleksitas model dengan mengeliminasi fitur yang kurang informatif atau redundan yang mungkin menyebabkan kebisingan dan *overfitting*. Oleh karena itu, *Naïve Bayes* yang dimodifikasi dengan seleksi fitur ini dapat fokus pada variabel yang paling relevan, yang memungkinkan model untuk membuat prediksi yang lebih akurat dan efisien. Hasilnya, seleksi fitur tidak hanya meningkatkan kinerja model secara keseluruhan tetapi juga memperkuat kemampuan model untuk menggeneralisasi ke data baru, yang sangat penting dalam konteks medis di mana keakuratan diagnosa memainkan peran krusial.

4. KESIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa penelitian berhasil mengembangkan model klasifikasi efektif untuk mendiagnosa kanker payudara menggunakan algoritma *Naïve Bayes* yang dipadukan dengan seleksi fitur *Information Gain*. *Dataset* yang digunakan, yang berisi 569 sampel dengan 30 fitur dari gambar aspirasi jarum halus massa payudara, diambil dari Kaggle dan mencakup dua klasifikasi kanker payudara, *Benign* (B) dan *Malignant* (M). Melalui seleksi fitur, berhasil dipilih 8 fitur terpenting, yang meliputi perimeter_worst, concave points_mean, dan radius_worst, yang mengurangi kompleksitas model dan mempercepat proses pelatihan tanpa mengurangi akurasi. Model yang dikembangkan menunjukkan peningkatan akurasi yang signifikan, mencapai 96.49%, lebih tinggi dari akurasi sebelumnya yang hanya 94.15%. Evaluasi kinerja model menunjukkan bahwa seleksi fitur *Information Gain* meningkatkan semua metrik evaluasi, seperti presisi, *recall*, dan skor F1, terutama dalam mengidentifikasi kanker payudara ganas. Hasil ini membuktikan bahwa kombinasi algoritma *Naïve Bayes* dengan seleksi fitur *Information Gain* memberikan kinerja yang lebih baik dibandingkan metode klasifikasi lainnya yang diuji. Selain itu, penerapan seleksi fitur meningkatkan efisiensi dan kecepatan model, yang sangat relevan untuk aplikasi medis dalam diagnosis kanker payudara.

REFERENCES

[1] A. J. W. Amna, Wahyuddin S, I Gede Iwan Sudipa, Tri Andi E. Putra, *Buku Data Mining*. 2023. [Online]. Available: https://www.cambridge.org/core/product/identifier/CBO9781139058452A007/type/book_part

[2] B. Raharjo, *Buku Pembelajaran Mesin (Machine Learning)*. 2021. [Online]. Available: https://www.codepolitan.com/mengenal-teknologi-machine-learning-pembelajaran-mesin

[3] R. K. Dinata, "Buku Machine Learning." pp. 1–156, 2020.

[4] J. Narabel and S. Budi, "Deteksi Dini Status Keanggotaan Industri Kebugaran Menggunakan Pendekatan Supervised Learning," *J. Tek. Inform. dan Sist. Inf.*, vol. 6, no. 2, pp. 266–277, 2020, doi: 10.28932/jutisi.v6i2.2675.

[5] Ardi Ramdani, Christian Dwi Sofyan, Fauzi Ramdani, Muhamad Fauzi Arya Tama, and Muhammad Angga Rachmatsyah, "Algoritma Klasifikasi Data Mining Untuk Memprediksi Masyarakat Dalam Menerima Bantuan Sosial," *J. Ilm. Sist. Inf.*, vol. 1, no. 2, pp. 39–47, 2022, doi: 10.51903/juisi.v1i2.363.

[6] J. Kusumawaty, E. Noviati, I. Sukmawati, Y. Srinayanti, and Y. Rahayu, "Efektivitas Edukasi SADARI (Pemeriksaan Payudara Sendiri) Untuk Deteksi Dini Kanker Payudara," *ABDIMAS J. Pengabdi. Masy.*, vol. 4, no. 1, pp. 496–501, 2021, doi: 10.35568/abdimas.v4i1.1177.

Volume 5, Nomor 2, September 2025
ISSN 2808-5450 (media cetak), ISSN 2808-5019 (media online)
Available Online at https://ejournal.unama.ac.id/index.php/jms
DOI https://doi.org/10.33998/jms.v5i2

[7] Asiva Noor Rachmayani, *Buku Data Mining Algoritma C4.5*. 2019.
[8] F. Handayani, D. Feddy, and S. Pribadi, "Implementasi Algoritma Naive Bayes Classifier dalam Pengklasifikasian Teks Otomatis Pengaduan dan Pelaporan Masyarakat melalui Layanan Call Center 110," *J. Tek. Elektro*, vol. 7, no. 1, pp. 19–24, 2015.
[9] M. R. Maulana and M. A. Al Karomi, "*Information Gain* Untuk Mengetahui Pengaruh Atribut," *J. Litbang Kota Pekalongan*, vol. 9, pp. 113–123, 2015.
[10] F. D. Astuti, "Seleksi Atribut Menggunakan *Information Gain* Untuk Clustering Penduduk Miskin Dengan Validity Index Xie Beni," *Teknika*, vol. 6, no. 1, pp. 61–65, 2017, doi: 10.34148/teknika.v6i1.58.
[11] M. T. A. Herfandi, Zaen, Y. Yuliadi, M. Julkarnain, and F. Hamdani, "Application of *Information Gain* to Select Attributes in Improving *Naïve Bayes* Accuracy in Predicting Customer's Payment Capability," *JISA(Jurnal Inform. dan Sains)*, vol. 4, no. 2, pp. 155–163, 2021, doi: 10.31326/jisa.v4i2.1044.
[12] A. Nugroho, "Analisa Splitting Criteria Pada Decision Tree dan Random Forest untuk Klasifikasi Evaluasi Kendaraan," *JSITIK J. Sist. Inf. dan Teknol. Inf. Komput.*, vol. 1, no. 1, pp. 41–49, 2022, doi: 10.53624/jsitik.v1i1.154.
[13] R. F. Putra, I. R. Mukhlis, A. I. Datya, and S. J. Pipin, *BUKU ALGORITMA PEMBELAJARAN MESIN*. 2024.
[14] J. T. Santoso, *Buku Proyek Coding dengan Python*. 2022.
[15] L. D. Utami *et al.*, "Integrasi Metode *Information Gain* untuk Seleksi Fitur dan AdaBoost untuk Mengurangi Bias pada Analisis Sentimen Review Restoran Menggunakan Algoritma Naive Bayes," *J. Intell. Syst.*, vol. 1, no. 2, pp. 120–126, 2015.
[16] M. Ramanda Hasibuan and Marjin, "Pemilihan Fitur dengan *Information Gain* untuk Klasifikasi Penyakit Gagal Ginjal menggunakan Metode Modified K-Nearest Neighbor (MKNN)," *J. Pengemb. Teknol. Inf. dam Ilmu Komput.*, vol. 3, no. 11, pp. 3659–875, 2019, [Online]. Available: http://j-ptiik.ub.ac.id
[17] A. Budianita, "*Information Gain* Berbasis Algoritma Naive Bayes Classifier Pada Pemodelan Prediksi Kelulusan," *J. Ilm. Intech Inf. Technol. J. UMUS*, vol. 5, no. 1, pp. 1–10, 2023, doi: 10.46772/intech.v5i1.1116.
[18] E. Rahmanita, Y. D. P. Negara, Y. Kustiyahningsih, V. Sasmeka, and B. K. Khotimah, "Implementasi Metode *Naïve Bayes* dan *Information Gain* Untuk Klasifikasi Penyakit dan Hama Tanaman Jagung," *Teknika*, vol. 12, no. 3, pp. 198–204, 2023, doi: 10.34148/teknika.v12i3.684.
[19] A. Isnanda, Y. Umaidah, and J. H. Jaman, "Implementasi *Naïve Bayes* Classifier Dan *Information Gain* Pada Analisis Sentimen Penggunaan E-Wallet Saat Pandemi," *J. Teknol. Inform. dan Komput.*, vol. 7, no. 2, pp. 144–153, 2021, doi: 10.37012/jtik.v7i2.648.
[20] A. Dwi Pangestu, I. Ernawati, and N. Chamidah, "Analisis Sentimen Terhadap PPKM Darurat Pada Media Sosial Twitter Menggunakan Metode *Naïve Bayes* Dengan Seleksi Fitur *Information Gain*," *Semin. Nas. Mhs. Ilmu Komput. dan Apl.*, vol. 3, no. 2, pp. 662–671, 2022.