Jurnal JPILKOM, 3 (1) (2025), ISSN: 3025-6887 (Online)
Jurnal JPILKOM (Jurnal Penelitian Ilmu Komputer)
Journal homepage: https://jpilkom.org

Optimasi Hyperparameter Naïve Bayes dalam Memprediksi Kanker Payudara Menggunakan GridSearch

Syahrin Daffa Raffalin Sitompul
Program Studi Sistem Informasi, STIKOM Tunas Bangsa Pematangsiantar, Indonesia
E-Mail: daffasitompull500@gmail.com

Article Info

Article history:

Received Jan 28, 2025
Revised Feb 27, 2025
Accepted Mar 19, 2025

Kata Kunci:

Naïve Bayes,
Prediksi Kanker Payudara,
Optimasi Hyperparameter,
GridSearchCV,
SMOTE,
Diagnosis Medis

Keywords:

Naïve Bayes,
Breast Cancer Prediction,
Hyperparameter Optimization,
GridSearchCV,
SMOTE,
Medical Diagnosis

ABSTRAK

Kanker payudara merupakan salah satu penyakit paling mematikan bagi perempuan, dan deteksi dini berperan penting dalam meningkatkan peluang kesembuhan. Penelitian ini mengoptimalkan algoritma *Naïve Bayes* untuk prediksi kanker payudara dengan menggunakan *GridSearchCV* dalam *tuning hyperparameter* serta *SMOTE* untuk mengatasi ketidakseimbangan kelas. *Dataset* yang digunakan adalah *Breast Cancer Wisconsin (Diagnostic)* dengan 569 data dan 30 fitur numerik. Proses *preprocessing* mencakup *encoding* label, seleksi fitur berdasarkan korelasi, *normalisasi*, dan *oversampling*. *GridSearchCV* menemukan nilai *var_smoothing* optimal sebesar 3.20×10^{-8}. Model terbaik mencapai akurasi 92,31% dengan *precision* dan *recall* seimbang untuk kedua kelas. Hasil ini menunjukkan bahwa optimasi *Naïve Bayes* efektif dalam klasifikasi data medis dan berpotensi mendukung diagnosis kanker payudara secara dini dan akurat.

ABSTRACT

Breast cancer is one of the deadliest diseases affecting women, and early detection plays a vital role in increasing survival rates. This study aims to optimize the performance of the Naïve Bayes algorithm for breast cancer prediction using GridSearchCV for hyperparameter tuning and SMOTE to address class imbalance. The dataset used is the Breast Cancer Wisconsin (Diagnostic), consisting of 569 records and 30 numerical features. The preprocessing steps include label encoding, feature selection based on correlation, normalization, and oversampling. GridSearchCV identified the optimal var_smoothing value of 3.20×10^{-8}. The optimized model achieved 92.31% accuracy with balanced precision and recall across both classes. These results indicate that the optimized Naïve Bayes is effective for medical data classification and has strong potential to support early and accurate breast cancer diagnosis

Corresponding Author:
Syahrin Daffa Raffalin Sitompul,
Program Studi Sistem Informasi, STIKOM Tunas Bangsa Pematangsiantar,
Jalan Jendral Sudirman Blok A, No. 1/2/3, Siantar Barat, Kota Pematangsiantar, Sumatera Utara, 21127, Indonesia
Email: daffasitompull500@gmail.com

1. PENDAHULUAN

Kanker payudara merupakan salah satu penyakit yang paling mematikan bagi perempuan di seluruh dunia, termasuk di Indonesia. Deteksi dini terhadap penyakit ini memiliki peran penting dalam meningkatkan peluang kesembuhan pasien. Seiring dengan berkembangnya teknologi dan ilmu data, pemanfaatan *machine learning* dalam bidang medis, khususnya untuk prediksi penyakit, menjadi topik yang sangat menarik dan sedang banyak dikaji (Andryan et al., 2022). Dalam tren ini, penggunaan algoritma *Naïve Bayes* semakin sering diangkat karena kemudahannya dalam implementasi dan efisiensinya dalam klasifikasi data berlabel.

Berbagai penelitian telah membuktikan bahwa *Naïve Bayes* memiliki performa yang cukup baik dalam klasifikasi teks dan data medis. Misalnya, salah satu penelitian menunjukkan bahwa optimasi *Naïve Bayes* dalam klasifikasi teks klinis kanker menghasilkan akurasi yang cukup tinggi (Taslim et al., 2023). Sementara itu, penelitian lain menerapkan pendekatan *GridSearch* pada berbagai varian *Naïve Bayes* untuk mendeteksi phishing email dan mendapatkan peningkatan performa yang signifikan (Rahman & Fauzi Abdulloh, 2023). Namun, meskipun metode ini populer, penerapan optimasi *hyperparameter* pada *Naïve Bayes* masih belum banyak dilakukan secara sistematis dalam konteks prediksi penyakit kanker, khususnya kanker payudara.

Sejumlah studi telah membahas pentingnya optimasi *hyperparameter* dalam meningkatkan kinerja model pembelajaran mesin. Salah satu penelitian mengenai klasifikasi sentimen menggunakan *Multinomial Naïve Bayes* dan TF-IDF, menunjukkan bahwa hasil klasifikasi dapat meningkat signifikan ketika fitur dan parameter model dioptimalkan secara tepat (Gerliandeva et al., 2024). Pendekatan serupa juga digunakan untuk kombinasi *Support Vector Machine (SVM), GridSearch,* dan *N-Gram* dalam klasifikasi sentimen komentar pengguna game, dengan hasil peningkatan akurasi mencapai lebih dari 85% (Iriananda et al., 2024). Penelitian lain juga menekankan bahwa tuning *hyperparameter* pada algoritma *supervised learning* mampu mengoptimalkan performa klasifikasi keluarga penerima bantuan pangan secara signifikan (Joshua Agung Nurcahyo & Theopilus Bayu Sasongko, 2023).

Penelitian dalam bidang medis juga menunjukkan peningkatan performa model ketika dilakukan tuning parameter. Karthika M S et al. (2022) Salah satu penelitian membuktikan bahwa penerapan *Adam* dan *RanAdam* untuk tuning parameter dalam deteksi kanker paru-paru dari data mikroarray mampu meningkatkan presisi model secara drastic (M S et al., 2024). Demikian pula, penelitian lain menunjukkan bahwa optimasi *Gaussian Naïve Bayes* dengan teknik *Univariate Feature Selection* menghasilkan akurasi tinggi dalam prediksi cuaca (Lindawati et al., 2023). Bahkan dalam konteks keamanan siber, *Bayesian* pada model deteksi malware mampu mendapatkan hasil yang kompetitif (ALGorain & Clark, 2021).

Dalam ranah data tidak seimbang, suatu penelitian menyatakan bahwa *Naïve Bayes* dapat dioptimalkan dengan teknik penyeimbangan kelas agar performanya tetap stabil (Andriansyah & Nurhasanah, 2020). Dalam studi yang lebih spesifik, penggunaan metode SMOTE-ENN dan *XGBoost* dengan optimasi *Bayesian* untuk mendeteksi mikrokalsifikasi, salah satu indikator kanker payudara, dan mendapatkan performa tinggi dengan nilai *F1-Score* mencapai 0,94 (Panjaitan & Sutarman, 2024). Temuan ini menunjukkan bahwa optimasi parameter pada algoritma sangat penting untuk meningkatkan performa prediksi dalam domain medis.

Namun demikian, belum banyak penelitian yang secara spesifik menggabungkan optimasi *hyperparameter* berbasis *GridSearch* pada algoritma *Naïve Bayes* dalam konteks prediksi kanker payudara. Beberapa studi yang membandingkan algoritma seperti *Naïve Bayes*, *K-Nearest Neighbor*, dan *Random Forest* (Sejati et al., 2022) atau *XGBoost* dan *SVM* (Andryan et al., 2022)) untuk diagnosis kanker payudara, masih belum menitikberatkan pada proses optimasi parameter model secara mendalam. Selain itu, penggunaan *Naïve Bayes* dan *SVM* untuk analisis sentimen terhadap aplikasi finansial juga menegaskan perlunya proses tuning parameter untuk mendapatkan hasil terbaik (Maulana et al., 2024).

Berdasarkan celah yang ada, penelitian ini bertujuan untuk mengoptimalkan algoritma *Naïve Bayes* dalam memprediksi kanker payudara menggunakan teknik *GridSearch* untuk pemilihan *hyperparameter* terbaik. Dengan menggunakan dataset diagnosis kanker, penelitian ini akan menguji sejauh mana optimasi parameter dapat meningkatkan akurasi, presisi, dan recall dari model *Naïve Bayes*. Selain itu, penelitian ini juga memberikan kontribusi berupa analisis kinerja model sesudah tuning, serta memberikan insight lebih lanjut mengenai pengaruh setiap parameter terhadap hasil prediksi.

Secara keseluruhan, penelitian ini tidak hanya menjawab kebutuhan akan model klasifikasi yang akurat dalam diagnosis kanker payudara, namun juga memberikan pendekatan sistematis dalam proses tuning parameter menggunakan *GridSearch*, yang dapat diadaptasi untuk domain medis lainnya. Diharapkan, hasil dari penelitian ini dapat memberikan kontribusi dalam pengembangan sistem pendukung keputusan berbasis *machine learning* untuk dunia medis di masa depan.

2. METODE PENELITIAN

2.1 Data Penelitian

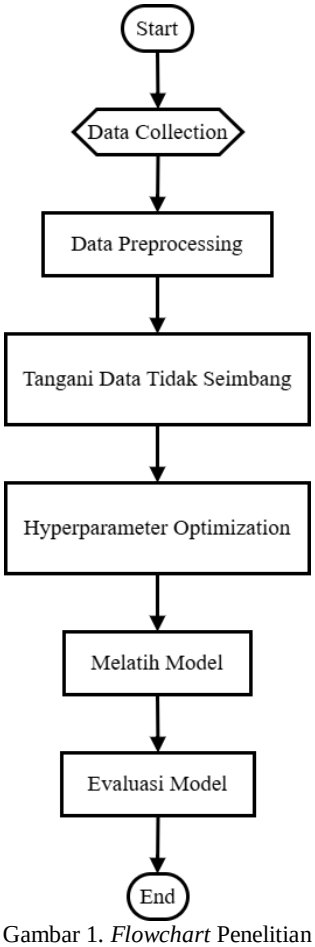
Dataset yang digunakan dalam penelitian ini adalah *Breast Cancer Wisconsin* (*Diagnostic*) Dataset yang tersedia secara publik melalui berbagai *platform* riset data, salah satunya Kaggle (Ovsen, 2016). Dataset ini terdiri dari 569 entri data yang masing-masing merepresentasikan hasil diagnosis terhadap pasien yang diuji. Terdapat 30 fitur numerik yang dihasilkan dari pencitraan digital inti sel kanker, seperti: *radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean,* dan seterusnya, hingga *fractal_dimension_worst*. Fitur-fitur ini mencerminkan berbagai karakteristik statistik sel kanker seperti

ukuran, bentuk, dan tekstur. Sementara itu, kolom diagnosis merupakan variabel target dengan dua kategori: "M" (*Malignant*) yang menunjukkan tumor ganas dan "B" (*Benign*) untuk tumor jinak.

Penelitian ini bertujuan untuk memprediksi jenis tumor (ganas atau jinak) berdasarkan fitur-fitur yang tersedia dengan menggunakan algoritma *Naïve Bayes*. Untuk meningkatkan performa prediksi model, dilakukan optimasi hyperparameter menggunakan metode *GridSearch*. Selain itu, guna mengatasi ketidakseimbangan kelas dalam data diagnosis, digunakan pula teknik *oversampling* dengan metode SMOTE (*Synthetic Minority Over-sampling Technique*). Kombinasi dari *preprocessing* yang tepat dan tuning parameter ini diharapkan dapat meningkatkan akurasi dan ketepatan model dalam mendeteksi kanker payudara secara dini dan otomatis.

2.2 Flowchart Penelitian

Penelitian ini mengikuti alur kerja yang terstruktur untuk memastikan setiap langkah ditangani secara sistematis dan guna meningkatkan kinerja model. Gambar 1. menggambarkan proses yang dilakukan secara bertahap dalam penelitian ini.



Gambar 1. *Flowchart* Penelitian

Penelitian ini dimulai dengan tahap awal (*Start*) yang menandai dimulainya proses pemodelan prediktif untuk mendeteksi jenis tumor pada pasien kanker payudara. Langkah pertama adalah pengumpulan data (*Data Collection*), di mana data diambil dari *Breast Cancer Wisconsin (Diagnostic)* Dataset yang terdiri dari 569 entri dengan 30 fitur numerik serta satu label target (diagnosis) yang menunjukkan jenis tumor, yaitu "M" (*Malignant*) dan "B" (*Benign*). Setelah data dikumpulkan, dilakukan tahap pra-pemrosesan data (*Data Preprocessing*). Pada tahap ini, data dibersihkan dari nilai yang tidak konsisten, serta dilakukan transformasi yang diperlukan, seperti pengkodean label diagnosis dan normalisasi fitur numerik. Semua fitur digunakan karena dataset ini telah memiliki fitur yang saling independen secara statistik dan siap digunakan dalam proses pembelajaran mesin.

Selanjutnya, dilakukan penanganan ketidakseimbangan kelas (Tangani Data Tidak Seimbang), mengingat proporsi data antara tumor jinak dan ganas tidak seimbang. Untuk mengatasi hal ini, digunakan teknik oversampling SMOTE (*Synthetic Minority Over-sampling Technique*) guna memperbanyak data dari

kelas minoritas, sehingga distribusi kelas menjadi seimbang dalam data latih. Kemudian, model mengalami tahap optimasi hyperparameter (*Hyperparameter Optimization*) dengan menggunakan metode *GridSearch* untuk menemukan kombinasi parameter terbaik pada algoritma *Naïve Bayes*, seperti pemilihan varian *smoothing* (nilai alpha) yang optimal agar model dapat menangani fitur dengan distribusi yang berbeda secara lebih akurat. Setelah parameter optimal diperoleh, model dilatih pada data yang telah diproses dalam tahap pelatihan model (Melatih Model). Model belajar mengenali pola-pola dalam fitur statistik sel kanker yang berkaitan dengan klasifikasi tumor jinak atau ganas.

Langkah berikutnya adalah evaluasi model (Evaluasi Model), di mana performa model diuji menggunakan metrik evaluasi seperti akurasi, *precision, recall*, dan *f1-score* guna mengukur efektivitas model dalam mengklasifikasi jenis tumor. Akhir dari proses ini ditandai dengan tahap selesai (*End*), di mana model telah selesai dibangun dan dievaluasi, sehingga siap untuk digunakan sebagai alat bantu dalam diagnosis kanker payudara secara otomatis dan akurat.

2.3 *Naïve Bayes*

Penggunaan *Naïve Bayes* dalam klasifikasi tetap kompetitif ketika dikombinasikan dengan pemilihan fitur yang tepat, terutama pada data berdimensi tinggi. Pendekatan ini mampu meningkatkan efisiensi model tanpa mengorbankan akurasi, selama hanya fitur-fitur relevan yang digunakan.

Selain itu, model ini menunjukkan performa yang stabil meskipun dihadapkan pada data yang mengandung noise atau ketidakseimbangan kelas. Temuan ini memperkuat posisi *Naïve Bayes* sebagai metode klasifikasi yang layak digunakan, baik dalam tahap awal pemodelan maupun dalam perbandingan dengan algoritma lain (Blanquero et al., 2021).

2.4 Oversampling dengan SMOTE

SMOTE digunakan untuk menyeimbangkan jumlah data pada setiap kelas variabel target. Proses dilakukan setelah *encoding* dan sebelum pembagian data menjadi data latih dan data uji, agar model tidak belajar dari data uji. Pemilihan SMOTE didasarkan pada hasil penelitian sebelumnya yang menunjukkan bahwa metode ini efektif meningkatkan kinerja klasifikasi pada data tidak seimbang, baik pada model *Naïve Bayes*, CNN, maupun kombinasi dengan metode lain (Joloudari et al., 2023; Nguyen et al., 2023; Nurdian et al., 2022; Salman et al., 2024; Wongvorachan et al., 2023).

2.5 Hyperparameter Tuning

Tuning hyperparameter dilakukan menggunakan metode *GridSearch* untuk memperoleh kombinasi parameter terbaik pada model *Naïve Bayes*. Proses ini dilakukan dengan mengatur beberapa nilai kandidat untuk parameter seperti *var_smoothing*. Pemilihan *GridSearch* didasarkan pada keunggulannya dalam menjelajahi seluruh ruang kombinasi parameter secara sistematis, sehingga mampu meningkatkan akurasi model secara signifikan (Nurdian et al., 2022). Selain itu, *GridSearch* juga efektif meningkatkan performa klasifikasi sentimen menggunakan *Random Forest*, dengan hasil akurasi yang lebih stabil dibandingkan metode tuning lainnya (Siji George & Sumathi, 2020). Dalam penelitian ini, *GridSearch* dikombinasikan dengan validasi silang untuk memastikan model yang diperoleh memiliki performa yang konsisten dan tidak *overfitting*.

2.6 Evaluasi Model

Evaluasi model dilakukan untuk mengukur kinerja *Naïve Bayes* dalam mendeteksi kanker payudara berdasarkan fitur-fitur dari data citra histopatologi. Beberapa metrik yang digunakan meliputi akurasi, *precision, recall*, dan *F1-Score*. Seluruh metrik dievaluasi pada data uji untuk memastikan kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya. Akurasi digunakan sebagai metrik utama karena mencerminkan proporsi prediksi yang benar terhadap keseluruhan data. Namun, untuk memastikan evaluasi yang seimbang terhadap masing-masing kelas, metrik *precision, recall*, dan *F1-Score* juga dianalisis secara per kelas. Selain itu, digunakan *confusion matrix* untuk melihat distribusi prediksi model pada masing-masing kelas. Evaluasi dilakukan setelah proses tuning hyperparameter agar hasil mencerminkan performa terbaik dari model yang telah dioptimasi.

3. HASIL DAN PEMBAHASAN

3.1 Data Collection

Data yang digunakan dalam penelitian ini berasal dari kumpulan data diagnostik kanker payudara yang berisi berbagai fitur pengukuran morfologi sel. Dataset mencakup 569 sampel dengan 31 atribut numerik yang menggambarkan karakteristik tumor, seperti radius, tekstur, perimeter, area, hingga dimensi fraktal. Data ini

telah diorganisir dalam format terstruktur dan diimpor menggunakan perangkat lunak analisis data untuk proses selanjutnya. Gambar 2 memperlihatkan beberapa baris dan kolom dari dataset yang digunakan dalam penelitian ini.

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	texture
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	

5 rows × 33 columns

Gambar 2. Beberapa Baris Dataset

Dan Gambar 3 menunjukkan keseluruhan 32 fitur serta 1 target yaitu (diagnosis) yang digunakan dalam penelitian ini.

0	id	569 non-null	int64
1	diagnosis	569 non-null	object
2	radius_mean	569 non-null	float64
3	texture_mean	569 non-null	float64
4	perimeter_mean	569 non-null	float64
5	area_mean	569 non-null	float64
6	smoothness_mean	569 non-null	float64
7	compactness_mean	569 non-null	float64
8	concavity_mean	569 non-null	float64
9	concave points_mean	569 non-null	float64
10	symmetry_mean	569 non-null	float64
11	fractal_dimension_mean	569 non-null	float64
12	radius_se	569 non-null	float64
13	texture_se	569 non-null	float64
14	perimeter_se	569 non-null	float64
15	area_se	569 non-null	float64
16	smoothness_se	569 non-null	float64
17	compactness_se	569 non-null	float64
18	concavity_se	569 non-null	float64
19	concave points_se	569 non-null	float64
20	symmetry_se	569 non-null	float64
21	fractal_dimension_se	569 non-null	float64
22	radius_worst	569 non-null	float64
23	texture_worst	569 non-null	float64
24	perimeter_worst	569 non-null	float64
25	area_worst	569 non-null	float64
26	smoothness_worst	569 non-null	float64
27	compactness_worst	569 non-null	float64
28	concavity_worst	569 non-null	float64
29	concave points_worst	569 non-null	float64
30	symmetry_worst	569 non-null	float64
31	fractal_dimension_worst	569 non-null	float64
32	Unnamed: 32	0 non-null	float64

Gambar 3. Keseluruhan Fitur dan Target

Dari dataset yang telah dikumpulkan tidak terdapat baris ataupun kolom yang berisi *missing values*, duplikasi data dan *outliers*. Sehingga dapat dilanjutkan ke proses berikutnya.

3.2 Data Preprocessing

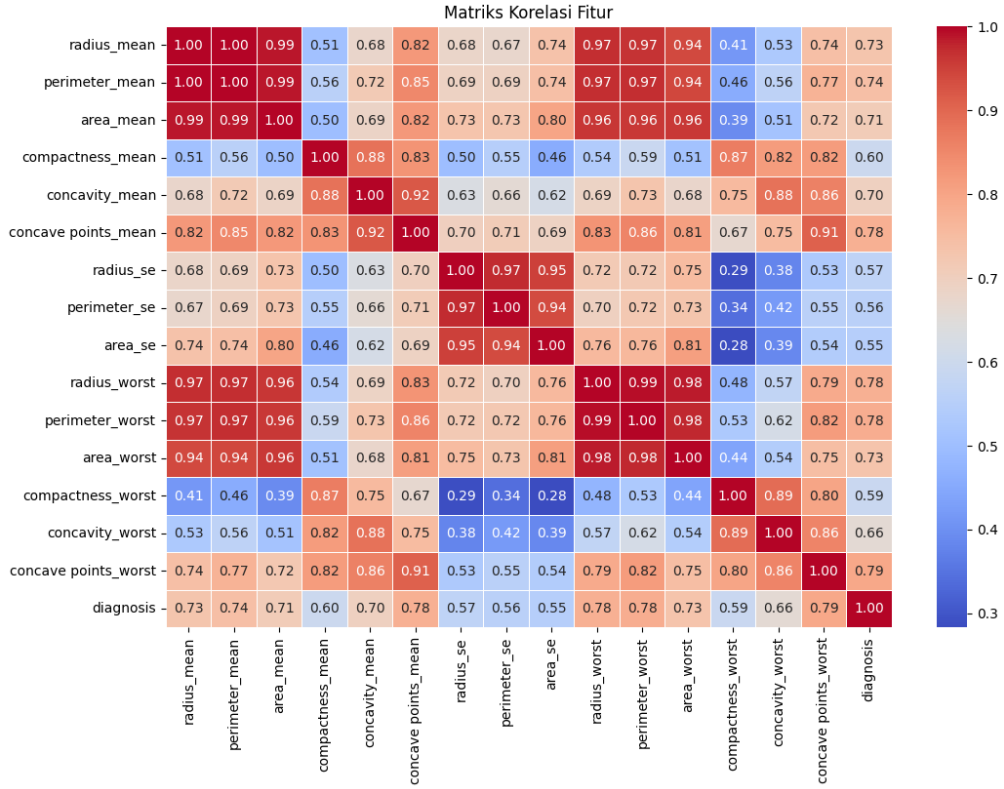
Proses *preprocessing* dimulai dengan menghapus kolom (*feature selection*) yang tidak memiliki kontribusi terhadap prediksi, seperti kolom ID dan Unanamed. Selanjutnya, fitur kategorikal dikonversi menjadi numerik menggunakan teknik *Label Encoding*, pada dataset tersebut hanya kolom diganosis yang merupakan fitur kategorikal, dan hanya memiliki 2 variasi data yaitu *Malignant* dan *Benign*, sehingga ketika diencode akan menghasilkan 0 dan 1, yang mana 0 untuk *Malignant* dan 1 untuk *Benign*. Gambar 4. Menunjukkan beberapa baris dan kolom dataset yang sudah bersih dari *variable* yang tidak dibutuhkan dan yang telah dilakukan *encoding* dengan teknik *Label Encoding*.

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean
0	1	17.99	10.38	122.80	1001.0	0.11840	0.27760
1	1	20.57	17.77	132.90	1326.0	0.08474	0.07864
2	1	19.69	21.25	130.00	1203.0	0.10960	0.15990
3	1	11.42	20.38	77.58	386.1	0.14250	0.28390
4	1	20.29	14.34	135.10	1297.0	0.10030	0.13280

5 rows × 31 columns

Gambar 4. Data *Feature Selection* dan Hasil *Encoding*

Selanjutnya dilakukan analisis korelasi antar variabel untuk mengidentifikasi fitur-fitur yang memiliki hubungan kuat dengan variabel target, yaitu “diagnosis”. Dalam penelitian ini, hanya fitur-fitur dengan nilai korelasi minimal 0.5 terhadap variabel target yang dipilih untuk digunakan pada tahap pembangunan model. Gambar 5. menunjukkan fitur-fitur terpilih hasil dari proses seleksi berdasarkan nilai korelasi.



Gambar 5. Korelasi Antar Fitur

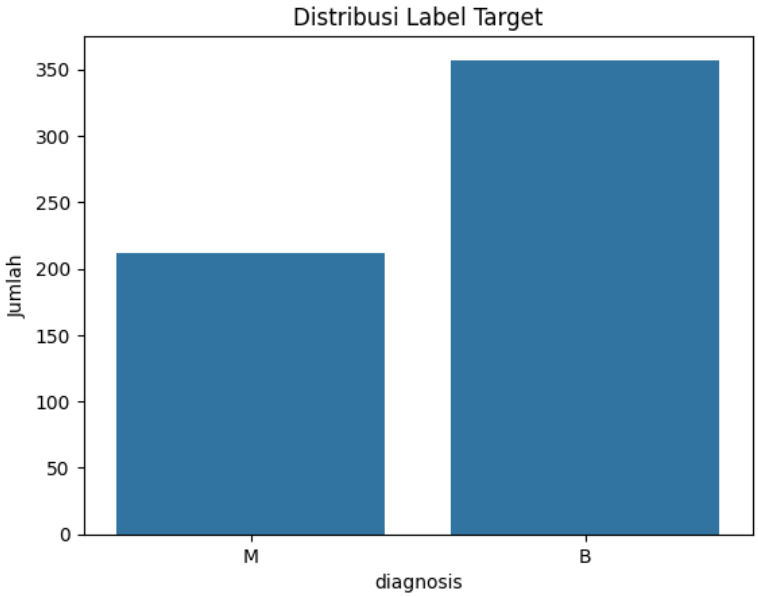
Normalisasi juga diterapkan pada fitur numerik untuk menyamakan skala antar variabel. Proses ini penting untuk memastikan setiap fitur memiliki kontribusi yang seimbang terhadap model. Gambar 6. Menunjukkan beberapa baris dan kolom data bersih setelah dilakukan normalisasi data.

	radius_mean	perimeter_mean	area_mean	compactness_mean	concavity_mean	concave points_mean	radius_se
0	0.521037	0.545989	0.363733	0.792037	0.703140	0.731113	0.356147
1	0.643144	0.615783	0.501591	0.181768	0.203608	0.348757	0.156437
2	0.601496	0.595743	0.449417	0.431017	0.462512	0.635686	0.229622
3	0.210090	0.233501	0.102906	0.811361	0.565604	0.522863	0.139091
4	0.629893	0.630986	0.489290	0.347893	0.463918	0.518390	0.233822

Gambar 6. Data Hasil Normalisasi Data

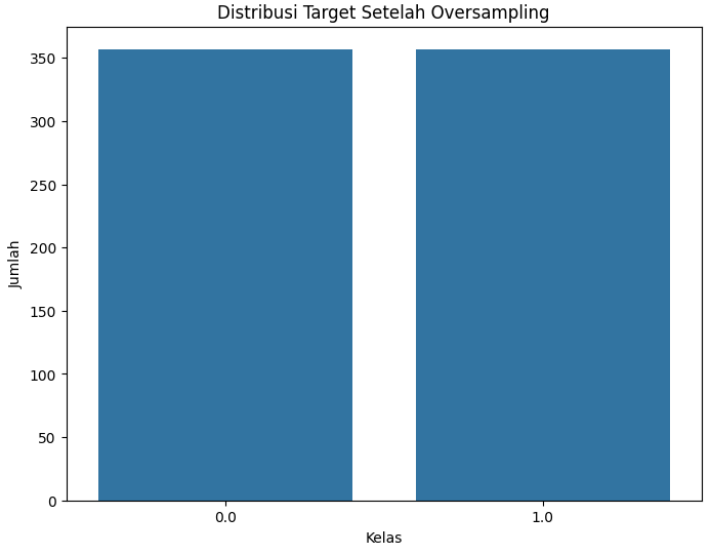
3.3 *Tangani Data Tidak Seimbang*

Setelah proses normalisasi dilakukan, data kemudian dipisahkan antara fitur dan target. Dalam dataset ini, fitur yang digunakan terdiri dari 30 atribut numerik seperti *radius_mean*, *texture_mean*, *area_mean*, hingga *fractal_dimension_worst*. Sedangkan variabel target adalah kolom diagnosis, yang terdiri dari dua kelas, yaitu *M* atau kelas 1 (*Malignant*) dan *B* atau kelas 0 (*Benign*). Distribusi awal data pada variabel target menunjukkan ketidakseimbangan, di mana jumlah data untuk kelas 0 yaitu 357 data yang lebih banyak dibandingkan kelas 1 yaitu 212 data, yang dapat menimbulkan bias dalam proses prediksi model. Gambar 7. Menampilkan visualisasi distribusi kelas sebelum dilakukan penanganan ketidakseimbangan data



Gambar 7. Distribusi Kelas Sebelum *Oversampling*

Oleh karena itu, digunakan metode SMOTE untuk melakukan *oversampling* terhadap kelas minoritas. SMOTE menghasilkan data sintetis berdasarkan hubungan antar sampel di kelas minoritas, sehingga model dapat belajar secara adil terhadap semua kelas. Setelah dilakukan *oversampling* maka distribusi kelas sekarang sudah seimbang yaitu 357 data untuk kedua kelas. Gambar 7. Menunjukkan distribusi kelas setelah dilakukan *oversampling*.



Gambar 8. Distribusi Kelas Setelah *Oversampling*

3.4 Hyperparameter Tuning

Setelah data sudah dilakukan penyeimbangan terhadap distribusi data, maka selanjutnya data dibagi menjadi data train dan data test, dengan proporsi 80% untuk data training dan 20% untuk data testing. Untuk memperoleh kinerja optimal dari model *Naive Bayes*, dilakukan tuning hyperparameter menggunakan metode *GridSearchCV* dengan validasi silang sebanyak 5-fold terhadap 100 kombinasi nilai *var_smoothing* dalam rentang 10^{-8} hingga 10^{-6}, sehingga dilakukan total 500 proses pelatihan. Parameter terbaik diperoleh pada nilai *var_smoothing* sebesar 3.20×10^{-8}, dengan akurasi validasi silang sebesar 92.12%. Model terbaik ini kemudian digunakan untuk pengujian, menghasilkan akurasi 92.31% pada data uji, dengan performa klasifikasi yang seimbang antara *precision* dan *recall* untuk kedua kelas.

3.5 Melatih Model

Model *Naïve Bayes* dilatih menggunakan data latih yang telah melalui proses *oversampling* dengan SMOTE. Pelatihan dilakukan menggunakan kombinasi hyperparameter terbaik yang diperoleh dari *GridSearch*. Selama proses pelatihan, kinerja model dimonitor menggunakan metrik evaluasi terhadap data validasi untuk memastikan kestabilan dan menghindari *overfitting*.

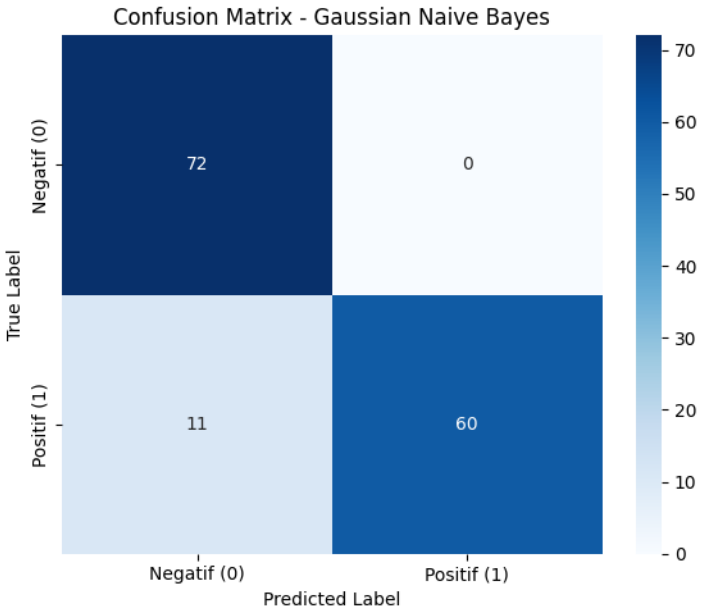
3.6 Evaluasi Model

Evaluasi dilakukan pada data uji menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score* untuk mengukur performa model dalam memprediksi penyakit kanker payudara. Hasil evaluasi menunjukkan bahwa model *Naïve Bayes* mampu memberikan kinerja yang sangat baik dengan akurasi mencapai **92,31%.** Nilai *precision*, *recall*, dan *F1-score* pada masing-masing kelas tergolong tinggi dan seimbang, mengindikasikan bahwa model tidak berat sebelah terhadap salah satu kelas. Gambar 9. Menunjukkan *Classification Report* berdasarkan model yang dibangun menggunakan parameter terbaik dari hyperparameter tuning yang sudah dilakukan.

```
Fitting 5 folds for each of 100 candidates, totalling 500 fits
Best Parameters: {'var_smoothing': np.float64(3.1992671377973846e-08)}
Best Cross-Validation Accuracy: 92.12%

Test Accuracy: 92.3076923076923

Classification Report:
              precision    recall  f1-score   support

           0       0.87      1.00      0.93        72
           1       1.00      0.85      0.92        71

    accuracy                           0.92       143
   macro avg       0.93      0.92      0.92       143
weighted avg       0.93      0.92      0.92       143
```

Gambar 9. *Classification Report*

Berdasarkan *Classification Report*, kelas 0 memiliki *precision* sebesar 0,87 dan *recall* 1,00, sedangkan kelas 1 memiliki *precision* 1.00 dan *recall* 0,85. Hal ini menunjukkan bahwa model mampu mengidentifikasi kedua kelas dengan akurasi yang konsisten. Gambar 10. Menunjukkan Hasil *Confusion Matrix* dari model yang sudah dilakukan.



Gambar 10. *Confusion Matrix*

Confusion matrix menunjukkan bahwa dari 72 sampel kelas 0, seluruhnya (72 sampel) berhasil diklasifikasikan dengan benar. Sementara itu, dari 71 sampel kelas 1, sebanyak 60 diklasifikasikan dengan benar dan 11 lainnya mengalami salah klasifikasi sebagai kelas 0.

4. KESIMPULAN

Penelitian ini berhasil menunjukkan bahwa optimasi *hyperparameter* pada algoritma *Naïve Bayes* menggunakan teknik *GridSearch* dapat meningkatkan performa prediksi kanker payudara secara signifikan. Hasil evaluasi model menunjukkan nilai akurasi, *precision*, *recall*, dan *F1-score* yang tinggi dan seimbang, membuktikan efektivitas pendekatan ini dalam klasifikasi data medis. Permasalahan terkait kurangnya penerapan optimasi parameter pada *Naïve Bayes* dalam konteks medis, khususnya kanker payudara, berhasil dijawab melalui pendekatan yang sistematis. Keunggulan penelitian ini terletak pada proses tuning yang terstruktur dan hasil yang stabil, sementara kekurangannya adalah keterbatasan jenis algoritma dan dataset yang digunakan. Untuk penelitian selanjutnya, disarankan untuk melakukan perbandingan dengan algoritma lain, menggunakan dataset yang lebih besar dan representatif, serta mengintegrasikan teknik penyeimbangan data dan seleksi fitur guna meningkatkan generalisasi model di dunia nyata.

REFERENCES

ALGorain, F. T., & Clark, J. A. (2021). Bayesian Hyper-Parameter optimisation for Malware Detection. *CEUR Workshop Proceedings*, *3125*, 69–84.

Andriansyah, S., & Nurhasanah. (2020). Seminar Nasional Industri dan Teknologi (SNIT), Politeknik Negeri Bengkalis. *Konsep Desain Menentukan Hull Type, Material, Dan Propulsi Unmanned Surface Vehicle (Usv) Untuk Patroli Di Wilayah Rokan Hiir Dengan Metode Desicion Tree*, *Lcm*, 478–486.

Andryan, M. R., Fajri, M., & Sulistyowati, N. (2022). Komparasi Kinerja Algoritma Xgboost Dan Algoritma Support Vector Machine (Svm) Untuk Diagnosis Penyakit Kanker Payudara. *JIKO (Jurnal Informatika Dan Komputer)*, *6*(1), 1. https://doi.org/10.26798/jiko.v6i1.500

Blanquero, R., Carrizosa, E., Ramírez-Cobo, P., & Sillero-Denamiel, M. R. (2021). Variable selection for Naïve Bayes classification. *Computers and Operations Research*, *135*, 105456. https://doi.org/10.1016/j.cor.2021.105456

Gerliandeva, A., Chrisnanto, Y. H., & Ashaury, H. (2024). *Optimasi Klasifikasi Sentimen pada Komentar Online menggunakan Multinomial Naïve Bayes dan Ekstraksi Fitur TF-IDF serta N-grams Optimization of Sentiment Classification on Online Comments using Multinomial Naïve Bayes and TF-IDF Feature Extraction and N-g*. *X*, 259–272. https://doi.org/10.56873/jpkm.v9i2.5585

Iriananda, S. W., Budiawan, R. W., Rahman, A. Y., & Istiadi, I. (2024). Optimasi Klasifikasi Sentimen Komentar Pengguna Game Bergerak Menggunakan Svm, Grid Search Dan Kombinasi N-Gram. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *11*(4), 743–752. https://doi.org/10.25126/jtiik.1148244

Joloudari, J. H., Marefat, A., Nematollahi, M. A., Oyelere, S. S., & Hussain, S. (2023). Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks. *Applied Sciences (Switzerland)*, *13*(6). https://doi.org/10.3390/app13064006

Joshua Agung Nurcahyo, & Theopilus Bayu Sasongko. (2023). Hyperparameter Tuning Algoritma Supervised Learning untuk Klasifikasi Keluarga Penerima Bantuan Pangan Beras. *The Indonesian Journal of Computer Science*, *12*(3), 1351–1365. https://doi.org/10.33022/ijcs.v12i3.3254

Lindawati, L., Fadhli, M., & Wardana, A. S. (2023). Optimasi Gaussian Naïve Bayes dengan Hyperparameter Tuning dan Univariate Feature Selection dalam Prediksi Cuaca. *Edumatic: Jurnal Pendidikan Informatika*, *7*(2), 237–246. https://doi.org/10.29408/edumatic.v7i2.21179

M S, K., Rajaguru, H., & Nair, A. R. (2024). Enhancement of Classifier Performance with Adam and RanAdam Hyper-Parameter Tuning for Lung Cancer Detection from Microarray Data—In Pursuit of Precision. *Bioengineering*, *11*(4). https://doi.org/10.3390/bioengineering11040314

Maulana, B. A., Fahmi, M. J., Imran, A. M., & Hidayati, N. (2024). Analisis Sentimen Terhadap Aplikasi Pluang Menggunakan Algoritma Naive Bayes dan Support Vector Machine (SVM). *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(2), 375–384. https://doi.org/10.57152/malcom.v4i2.1206

Nguyen, T., Mengersen, K., Sous, D., & Liquet, B. (2023). SMOTE-CD: SMOTE for compositional data. *PLoS ONE*, *18*(6 June), 1–19. https://doi.org/10.1371/journal.pone.0287705

Nurdian, R. A., Mujib Ridwan, & Ahmad Yusuf. (2022). Komparasi Metode SMOTE dan ADASYN dalam Meningkatkan Performa Klasifikasi Herregistrasi Mahasiswa Baru. *Jurnal Teknik Informatika Dan Sistem Informasi*, *8*(1), 24–32. https://doi.org/10.28932/jutisi.v8i1.4004

Ovsen, U. M. L. ·. (2016). *Breast Cancer Wisconsin (Diagnostic) Data Set*. Kaggle.Com. https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data?resource=download

Panjaitan, C. N. Y., & Sutarman. (2024). Klasifikasi Kelas Pada Data Tidak Seimbang Dalam Deteksi Mikrokalsifikasi Menggunakan Smote-Enn Dan Xgboost Dengan Optimasi Bayesian. *Jurnal Lingkar Pembelajaran Inovatif Volume*, *5*, 105–117.

Rahman, R., & Fauzi Abdulloh, F. (2023). Performance of Various Naïve Bayes Using GridSearch Approach In Phishing

Email Dataset. *Sinkron*, *8*(4), 2336–2344. https://doi.org/10.33395/sinkron.v8i4.12958

Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, *2024*, 69–79. https://doi.org/10.58496/bjml/2024/007

Sejati, P., Munawar, Pilliang, M., & Akbar, H. (2022). Studi Komparasi Naive Bayes , K-Nearest Neighbor, dan Random Forest Untuk Prediksi Calon Mahasiswa Yang Diterima Atau Comparative Study Of Naive Bayes , K-Nearest Neighbor , And Random Forest For The Prediction Of Prospective Students. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK)*, *9*(7), 1341–1348. https://doi.org/10.25126/jtiik.202296737

Siji George, C. G., & Sumathi, B. (2020). Grid search tuning of hyperparameters in random forest classifier for customer feedback sentiment prediction. *International Journal of Advanced Computer Science and Applications*, *11*(9), 173–178. https://doi.org/10.14569/IJACSA.2020.0110920

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. https://doi.org/10.30864/eksplora.v13i1.994

Wongvorachan, T., He, S., & Bulut, O. (2023). A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information (Switzerland)*, *14*(1). https://doi.org/10.3390/info14010054