TIN: Terapan Informatika Nusantara
Vol 5, No 2, July 2024, page 162-171
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i2.5408

Optimasi Algoritma Machine Learning Menggunakan Seleksi Fitur Xgboost Untuk Klasifikasi Kanker Payudara

Naufal Cahya Ramadhan*, Hanny Hikmayanti H, Tatang Rohana, Amril Mutoi Siregar

Fakultas Ilmu Komputer, Teknik Informatika, Universitas Buana Perjuangan, Karawang, Indonesia
Email: 1,*if20.naufalramadhan@mhs.ubpkarawang.ac.id, 2hanny.hikmayanti@ubpkarawang.ac.id,
3tatang.rohana@ubpkarawang.ac.id, 4amrilmutoi@ubpkarawang.ac.id
Email Penulis Korespondensi: if20.naufalramadhan@mhs.ubpkarawang.ac.id

Abstrak–Penelitian ini menganalisis kinerja algoritma K-Nearest Neighbors (KNN), Naïve Bayes, dan Random Forest dalam klasifikasi diagnosis kanker payudara menggunakan dataset Wisconsin Breast Cancer. Masalah yang dibahas adalah bagaimana meningkatkan akurasi klasifikasi diagnosis kanker payudara melalui teknik preprocessing yang tepat. Tujuan penelitian adalah mengevaluasi dan membandingkan kinerja ketiga algoritma tersebut setelah penerapan preprocessing yang meliputi pembersihan data, penanganan missing value, duplikasi data, dan outlier, serta seleksi fitur menggunakan XGBoost dan oversampling SMOTE. penerapan seleksi fitur untuk mengidentifikasi fitur yang paling relevan dan SMOTE untuk menyeimbangkan distribusi kelas dalam dataset. Hasil evaluasi kinerja menggunakan confusion matriks menunjukkan bahwa Random Forest memiliki kinerja terbaik dengan akurasi, presisi, recall, dan F1-score yang tinggi, mencapai AUC sebesar 98% setelah penerapan SMOTE. Kombinasi seleksi fitur dan SMOTE terbukti meningkatkan performa model secara signifikan, meskipun KNN menunjukkan penurunan performa dengan SMOTE, sementara Naïve Bayes mengalami peningkatan yang cukup baik. penelitian ini menunjukkan tentang pentingnya teknik preprocessing dalam pengembangan model machine learning untuk aplikasi medis, menekankan bahwa teknik yang tepat dapat secara signifikan meningkatkan kinerja klasifikasi dan menghasilkan diagnosis yang lebih akurat.

Kata Kunci: Seleksi Fitur; Extreme Gradient Boosting (XGBoost); K-Nearest Neighbor (KNN); Naïve Bayes; Random Forest

Abstract–This research analyzes the performance of the K-Nearest Neighbors (KNN), Naïve Bayes, and Random Forest algorithms in the classification of breast cancer diagnosis using the Wisconsin Breast Cancer dataset. The problem discussed is how to improve the accuracy of breast cancer diagnosis classification through appropriate preprocessing techniques. The research objective is to evaluate and compare the performance of the three algorithms after the application of preprocessing which includes data cleaning, handling missing values, data duplication, and outliers, as well as feature selection using XGBoost and SMOTE oversampling. application of feature selection to identify the most relevant features and SMOTE to balance the class distribution in the dataset. Performance evaluation results using a confusion matrix show that Random Forest has the best performance with high accuracy, precision, recall, and F1-score, reaching an AUC of 98% after the application of SMOTE. The combination of feature selection and SMOTE was shown to significantly improve model performance, although KNN showed a decrease in performance with SMOTE, while Naïve Bayes experienced a considerable improvement. This study demonstrates the importance of preprocessing techniques in the development of machine learning models for medical applications, emphasizing that appropriate techniques can significantly improve classification performance and result in more accurate diagnoses.

Keywords: Feature Selection; Extreme Gradient Boosting (XGBoost); K-Nearest Neighbor (KNN); Naïve Bayes; Random Forest

1. PENDAHULUAN

Secara teori, semakin banyak fitur yang diberikan kepada pengklasifikasi, semakin banyak informasi yang dimilikinya dan semakin baik klasifikasinya, namun dalam praktiknya proses pelatihan dan hasil klasifikasi menurun seiring dengan bertambahnya dimensi fitur. Jumlah fitur yang sangat banyak meningkatkan waktu pembelajaran dan mengurangi akurasi klasifikasi karena adanya redundansi fitur (Zhang et al., 2018) Seleksi fitur adalah proses memilih kumpulan fitur terbaik dari semua fitur dalam *dataset*. Pemilihan fitur menjadi masalah karena ruang pencarian sangat besar dan rumit, membuat pencarian yang mendalam menjadi tidak berguna, sehingga membutuhkan penggunaan algoritma pencarian global yang efisien (Nguyen et al., 2020).

Seleksi fitur adalah prosedur pra pemrosesan yang memiliki dampak langsung pada hasil klasifikasi. Analisis data sangat bergantung pada seleksi fitur dalam pengenalan pola. Proses ini bertujuan untuk mengekstrak fitur terbaik dari fitur asli, mengurangi dimensionalitas data dalam jumlah besar dan menghindari persoalan dimensionalitas, yang pada akhirnya meningkatkan kinerja metode klasifikasi (Pratama & Adhitya, 2019). Tujuan dari seleksi fitur adalah untuk memilih variabel-variabel penting untuk klasifikasi menggunakan *K-Nearest Neighbor (KNN), Naïve Bayes, dan Random Forest*. Karena karakteristik yang tidak relevan dan duplikasi akan merusak hasil, seleksi fitur harus dapat membedakan data. Menggunakan strategi seleksi fitur untuk menghasilkan fitur yang relevan adalah cara terbaik untuk meningkatkan algoritma klasifikasi. Teknik KNN memiliki masalah karena dapat sangat terganggu oleh adanya *noise* atau data yang tidak relevan jika ukuran fitur tidak sesuai dengan relevansinya. Untuk mengatasi masalah ini, KNN membutuhkan metode seleksi fitur untuk mengecualikan data yang berisik dan fitur yang tidak perlu. Teknik seleksi fitur digunakan untuk mengurangi efek noise data dan fitur yang tidak relevan dalam algoritma KNN, Naive Bayes, dan Random Forest. Pada Naive Bayes, asumsi independensi antar fitur dapat mengganggu estimasi probabilitas jika fitur yang tidak relevan masih dipertimbangkan. Sementara itu, pada Random Forest, meskipun memiliki kelebihan dalam mengurangi overfitting, fitur yang tidak relevan masih dapat mempengaruhi pengambilan keputusan pada pohon keputusan. Oleh karena itu, dengan menggunakan teknik seleksi fitur, kedua algoritma ini dapat mengatasi masalah tersebut dan meningkatkan akurasi dan konsistensi model yang dihasilkan. Teknik seleksi fitur membantu dalam

TIN: Terapan Informatika Nusantara
Vol 5, No 2, July 2024, page 162-171
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i2.5408

memperbaiki prosedur dan membuatnya lebih akurat. Sistem penambah pohon yang dapat diskalakan *Extreme Gradient Boosting (XGBoost)* menggabungkan pohon secara berurutan untuk membentuk model akhir dengan kesalahan yang minimal. Dengan *XGboost*, model klasifikasi *ensemble* dan pendekatan pohon dapat meningkatkan kinerja klasifikasi dan mengurangi fitur, sehingga mengurangi waktu komputasi (Manju et al., 2019). *XGBoost* memilih fitur terbaik berdasarkan peringkat, meningkatkan akurasi *K-Nearest Neighbor, Naïve Bayes*, dan *Random Forest*, yang dapat dipengaruhi oleh fitur dan *noise* (Rifatama et al., 2023).

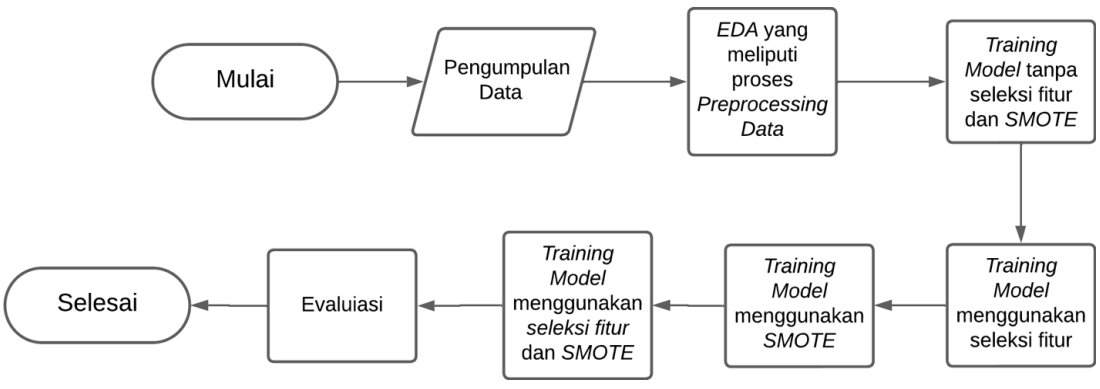
Machine Learning adalah bidang studi yang mempelajari penggunaan algoritma komputasi untuk mengubah data empiris menjadi model yang dapat digunakan. Algoritma *machine learning* dapat digunakan untuk mengumpulkan pemahaman tentang fenomena dunia maya yang menghasilkan data yang diteliti, mengabstraksikan pemahaman tentang fenomena yang mendasari dalam bentuk model, memprediksi nilai masa depan dari suatu fenomena dengan menggunakan model yang dihasilkan diatas mendeteksi perilaku anomali yang ditunjukkan oleh suatu fenomena yang diamati (Edgar & Manz, 2017).

Pada penelitian ini akan menggunakan XGBoost untuk mengklasifikasikan fitur dengan menggunakan metode *K-Nearest Neighbors*, *Naïve Bayes*, dan *Random Forest*. Pada penelitian Yohanes Setiawan pada tahun 2023, yang menggunakan metode pemilihan fitur *information gain* pada data mining berbasis *nearest neighbor*, hasil penelitiannya menunjukkan bahwa *Multi Local Mean k-Harmonic Nearest Neighbor* (MLM-KHNN) pada k=3 memiliki nilai *recall* yang paling baik yaitu sebesar 94%, begitu juga dengan metrik-metrik yang lain di atas 70% (Setiawan & Yohanes, 2023). Pada penelitian lain Muhammad Fauzan dkk melakukan penelitian menggunakan metode *information gain* pada klasifikasi penerima bantuan sosial pangkalan sesai dengan KNN, temuan menunjukkan bahwa klasifikasi terbaik diperoleh dengan rasio 7:3 dan 8:2 dengan nilai k=5, yang memperoleh nilai akurasi terbaik sebesar 98,21%, sedangkan nilai terendah diperoleh dengan rasio 9:1 dengan nilai k=5, tanpa menggunakan *information gain* sebesar 89,82%. (Fauzan et al., 2023). Alex Wira Wilantapoera dkk. menggunakan pendekatan *mutual information feature selection* untuk menganalisa sentimen kategori aspek pada evaluasi produk dengan menggunakan KNN. Hasil penelitian mengungkapkan bahwa hasil pengujian dengan penilaian *k-fold cross validation* dengan k=5 menghasilkan akurasi terbesar untuk aspek harga pada k=1 dengan akurasi 92,89%. Hasil terbaik pada aspek *packaging*, dengan nilai k=1, menghasilkan akurasi sebesar 81.39%. Hasil terbaik diperoleh pada aspek aroma dengan nilai k = 1 yang menghasilkan akurasi sebesar 80,74% (Wilantapoera et al., 2023).

Penelitian yang dilakukan oleh Rahmat Hidayat et al. dilakukan untuk klasifikasi kanker payudara dengan menggunakan metode optimalisasi fitur *partikel swarm biner* pada KNN. Hasil penelitian menunjukkan bahwa dengan 30 fitur, model KNN menghasilkan akurasi terbaik sebesar 94,15% pada k=13, dan model KNN+BPSO menghasilkan akurasi terbaik sebesar 95,32% pada k=9 dengan 5 fitur, yaitu compactness se, simetri se, texture paling buruk, dan simetri paling buruk (Hidayata et al., 2023). Dalam penelitian tambahan yang dilakukan oleh Deni Kurnia dkk., mereka menggunakan metode optimasi *particle swarm* seleksi fitur dengan XGBoost untuk mengklasifikasikan penyakit parkinson. Nilai AUC percobaan klasifikasi pada model dengan SMOTE dan pengaturan *hyperparameter* hanya 0,9250, sedangkan pada model tanpa SMOTE dan pengaturan hyperparameter hanya 0,9325. Ketika kedua teknik SMOTE dan pengaturan hyperparameter digunakan secara bersamaan, nilai AUC percobaan klasifikasi adalah 0,9325 (Kurnia et al., 2023). Berdasarkan penelitian sebelumnya, penelitian ini akan melakukan optimasi algoritma KNN, *Naïve Bayes*, dan *Random Forest* dengan seleksi fitur menggunakan metode *XGBoost*.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian



Gambar 1. Alur Penelitian

Pada gambar 1 menggambarkan alur proses pengembangan dan evaluasi model machine learning yang dimulai dengan tahap pengumpulan data. Setelah data terkumpul, dilakukan *Exploratory Data Analysis* (EDA) yang mencakup *preprocessing* data untuk memastikan data siap digunakan dalam pemodelan. Selanjutnya, model dilatih tanpa menggunakan seleksi fitur dan teknik SMOTE untuk mendapatkan *baseline* performa. Kemudian, model dilatih menggunakan seleksi fitur, diikuti dengan pelatihan model menggunakan teknik SMOTE untuk mengatasi masalah

TIN: Terapan Informatika Nusantara
Vol 5, No 2, July 2024, page 162-171
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i2.5408

ketidakseimbangan kelas dalam dataset. Tahap berikutnya adalah melatih model dengan menggabungkan kedua teknik, yaitu XGBoost dan SMOTE, untuk melihat apakah kombinasi ini memberikan hasil yang lebih baik. Setelah semua model dilatih, dilakukan evaluasi menggunakan *confusion matrix* dan ROC AUC untuk menentukan model dengan performa terbaik.

2.1.1 Pengumpulan Data

Data dari *Breast Cancer Wisconsin* (Diagnosis) digunakan dalam penelitian ini, dengan 32 karakteristik dan 569 catatan. Ada dua jenis kategori: 212 ganas dan 357 jinak, yang dikumpulkan dari situs web kaggle.com. Ciri-ciri dalam kumpulan data Kanker Payudara Wisconsin (Diagnostik) dihitung dari gambar digital yang diperoleh melalui Biopsi Aspirasi Jarum Halus (FNAB) atau aspirasi jarum halus pada tumor payudara. Ciri-ciri ini menentukan sifat-sifat inti sel pasien. Kumpulan data diagnosis kanker payudara Wisconsin mencakup 12 fitur untuk setiap inti sel 10 atribut bernilai nyata dihitung, seperti radius (jarak rata-rata antara pusat dan titik-titik di sekitarnya), tekstur (deviasi standar nilai skala abu-abu), keliling, *area*, *smoothness* (variasi lokal dalam panjang jari-jari), *compactness* (perimeter ^ 2 / area - 1.0), cekung (tingkat keparahan bagian cekung dari kontur), titik cekung (jumlah bagian cekung dari kontur), simetri, dimensi fraktal (perkiraan garis pantai - 1). Setiap gambar menerima nilai rata-rata, kesalahan standar, dan nilai paling negatif dari fitur-fitur ini, yang mewakili 32 fitur.

2.1.2 Preprocessing Data

Pada tahap preprocessing terdapat proses analisis data yang melibatkan banyak langkah, termasuk pembersihan data. Pembersihan data memerlukan tindakan penting seperti menemukan dan menghapus nilai yang hilang, serta mencari dan menghapus data duplikat, untuk memastikan keakuratan dan kebersihan data analisis. Preprocessing diperlukan sebelum menggunakan data untuk memastikan bahwa nilainya seimbang. Data yang tidak seimbang dapat mempengaruhi kinerja model. *MinMax Scaling* adalah metode untuk mengubah data menjadi rentang 0 hingga 1 dengan mengurangi nilai saat ini dari nilai minimum dan kemudian membaginya dengan nilai maksimum dikurangi nilai minimum.

2.1.3 Seleksi Fitur Xboost

Setelah memisah data pelatihan dan pengujian, pilih fitur dengan XGBoost. Pendekatan XGBoost memilih karakteristik tergantung pada relevansi atau pengaruhnya, seperti yang ditentukan oleh algoritma. Tujuan dari tahap feature importance adalah untuk menentukan variabel fitur yang paling berpengaruh terhadap variabel target atau kelas model dalam machine learning. Dengan menggunakan metode peringkat signifikansi fitur, maka dapat mengevaluasi dan meningkatkan kinerja model dengan fokus pada fitur yang paling penting. Penggunaan threshold memungkinkan kita untuk menetapkan batas nilai relevansi yang signifikan, sehingga fitur-fitur dengan nilai di atas batas tersebut dianggap penting, sementara yang di bawahnya diabaikan (Syafei & Efrilianda, 2023).

2.1.4 SMOTE

Menurut Nithes V. Chawla, *Synthetic Minority Oversampling* (SMOTE) telah diusulkan sebagai metode untuk menangani data yang tidak seimbang. Saat menggunakan metode SMOTE, jumlah distribusi data di kelas minoritas yang disampling diseimbangkan dengan melakukan oversampling pada data hingga jumlah sampel data setara dengan jumlah sampel data di kelas mayoritas. Data yang paling dekat disusun berdasarkan jarak Euclidean dari data lainnya menggunakan teknik slicing (Nurdian et al., 2022).

2.1.5 K-Nearest Neighbor

K-Nearest Neighbor (K-NN) adalah metrik yang menggunakan data pelatihan untuk mengkategorikan objek berdasarkan jarak terdekatnya. Teknik KNN memberikan hasil prediksi dengan mengklasifikasikan jarak terpendek antar tetangga (Adhikary & Banerjee, 2023). Teknik ini menghitung jarak terkecil antara informasi yang akan dinilai dengan tetangganya dalam data pelatihan (Ali et al., 2023).
ini merupakan teorema KNN yang universal untuk menghitung jarak :

$$d_i = \sqrt{\sum_{i=1}^{n}(x_{ij} - p_j)^2} \quad (1)$$

Keterangan:
d_i = jarak sampel
x_{ij} = data sampel pengetahuan
p_j = data input var j
n = jumlah sampel
Langkah-langkah untuk mengimplementasikan metode K-NN sebagai berikut (Ali et al., 2023):
1. Menentukan parameter k (jumlah tetangga terdekat)
2. Menghitung kuadrat dari jarak Euclidean objek terhadap data latih yang diberikan
3. Mengurutkan hasil kuadrat pada langkah 2 dari yang tertinggi ke yang rendah
4. Mengumpulkan kategori klasifikasi tetangga terdekat berdasarkan nilai k
5. Memprediksi kategori objek menggunakan kategori tetangga terdekat dengan nilai tertinggi

TIN: Terapan Informatika Nusantara
Vol 5, No 2, July 2024, page 162-171
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i2.5408

2.1.6 Naïve Bayes

Thomas Bayes seorang ilmuwan Inggris, memelopori *Naïve Bayes*, sebuah pendekatan *machine learning* dan data mining yang sangat efisien dan efektif berdasarkan probabilitas dan statistik. Pada tahap ini, pendekatan tersebut diterapkan pada dataset penelitian dengan menggunakan *Naïve Bayes* (Carli et al., 2023). *Naïve Bayes Classifier* sangat bergantung pada asumsi bahwa setiap kondisi/kejadian adalah independen (Shanshool et al., 2023). Membuat model *Naive Bayes* itu sederhana dan sangat bermanfaat untuk data dalam jumlah besar. Selain kesederhanaannya, *Naive Bayes* terkenal karena mengungguli sistem klasifikasi yang paling canggih (Siregar et al., 2020). Keuntungan menggunakan *Naive Bayes* adalah membutuhkan lebih sedikit data pelatihan untuk mengevaluasi parameter mean dan varians dari variabel yang digunakan dalam klasifikasi. Di sini, kami menyajikan bentuk universal dari teorema Bayes (V & S, 2022).

$$P(X) = \frac{P(H) \times P(H)}{P(X)} \quad (2)$$

Keterangan :
X = data dengan kelas yang tidak diketahui
H = data hipotesis adalah kelas tertentu
P(H|X) = probabilitas hipotesis H berdasarkan kondisi X (probabilitas posteriors)
P(H) = probabilitas hipotesis (probabilitas sebelumnya)
P(X|H) = probabilitas X berdasarkan kondisi hipotesis H
P(X) = probabilitas dari X

2.1.7 Random Forest

*Random Fores*t adalah kumpulan prediktor pohon di mana setiap pohon didasarkan pada nilai vektor acak yang disampel secara otonom dan didistribusikan secara seragam di antara semua tanaman di hutan.
Estimasi algoritma random forests yang dilakukan atau *class* dapat dijelaskan sebagai berikut:
1. Bootstrap pohon ditentukan dengan mengambil sampel data yang berbeda
2. Untuk setiap sampel bootstrap memberikan pohon kategori yang belum dipangkas, dengan cara sebagai berikut

Perubahan: pada setiap simpul, di antara semua prediktor pemilihan pemisahan kelas pertama tidak lebih disukai, secara sewenang-wenang pola m percobaan dari prediktor dan dari variabel seseorang, pilih pemecahan yang dapat diterima dan data baru diharapkan dengan cara menjumlahkan prediksi pohon n-pohon penggunaan suara mayoritas untuk jenis (R et al., 2019). *Random Forest* juga dikenal sebagai serangkaian pohon keputusan. Ini dapat digunakan untuk memprediksi kategori dengan berbagai nilai potensial dan probabilitas yang disesuaikan. Mengambil tindakan yang tepat untuk menghindari overfitting pada model yang dikembangkan sangat penting. Pembelajaran mesin *random forest* melibatkan pembuatan beberapa pohon keputusan dan kemudian menggabungkannya untuk mendapatkan prediksi yang lebih akurat dan konsisten (Ismafillah et al., 2023).

2.2 Evaluasi

Confusion matrix adalah tampilan data dalam bentuk matriks untuk mengidentifikasi data yang terdeteksi secara akurat dan seberapa sering data dilabeli sebagai data lain oleh sistem (Suhliyyah et al., 2023). *Confusion Matrix* digunakan untuk menilai performa dari setiap teknik kategorisasi yang diujicobakan. Perhitungan presisi, recall, dan akurasi akan didasarkan pada data dalam *confusion matrix* (Meuthia Zulma et al., 2021).

Tabel 1. Rumus Evaluasi Confusion Matrix

		Data Aktual	
Data Prediksi	Positif	*TP (True Positif)*	*FP (False Positive)*
	Negatif	*FN (False Negative)*	*TN (True Negative)*

Berikut ini adalah rumus mengukur performa berdasarkan confusion matrix dalam akurasi, presisi, dan recall:

$$Akurasi = \frac{True\ Positive + True\ Negative}{False\ Positive + False\ Negative\ True\ Positive + True\ Negative} \quad (3)$$

Akurasi digunakan untuk menilai kecukupan prediksi dalam kaitannya dengan data sampel.

$$Presisi = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (4)$$

Presisi digunakan untuk menentukan nilai prediksi benar bernilai positif terhadap semua prediksi positif.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (5)$$

Recall digunakan untuk menentukan nilai prediksi benar bernilai positif terhadap semua data bernilai positif.

TIN: Terapan Informatika Nusantara
Vol 5, No 2, July 2024, page 162-171
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i2.5408

Setelah itu, *Receiver Operating Characteristics* (ROC) akan digunakan untuk menggambarkan kinerja dari setiap klasifikasi yang diuji dalam dua dimensi. Di sini, sumbu horizontal mewakili nilai false positive dan sumbu vertikal menunjukkan true positive. *Area Under Curve* (AUC) adalah nilai yang mengukur area di bawah kurva ROC (Meuthia Zulma et al., 2021). Untuk mengkategorikan hasil AUC, kualitas klasifikasi dapat dilihat berdasarkan nilai AUC dalam tabel 2 yang disediakan.

Table 2. Kriteria AUC

Nilai AUC	Penjelasan
90% - 100%	*Excellent*
80% - 90%	*Good*
70% - 80%	*Fair*
60% - 70%	*Poor*
<60%	*Failure*

Pada tabel 2 di atas menjelaskan kriteria penilaian Area Under Curve (AUC) yang digunakan untuk mengevaluasi model prediksi. AUC adalah ukuran untuk mengukur seberapa baik model klasifikasi dapat membedakan antara kelas positif dan negatif. Model dengan AUC antara 90% - 100% dikategorikan sebagai sangat baik, karena mampu membedakan kelas dengan tingkat akurasi yang tinggi. Model dengan AUC antara 80% - 90% dianggap baik, cukup dapat diandalkan dalam membedakan kelas, meskipun tidak sebaik kategori sangat baik. Model dengan AUC antara 70% - 80% dianggap cukup, mampu membedakan kelas tetapi dengan akurasi sedang. Model dengan AUC antara 60% - 70% dianggap buruk, karena kurang efektif dalam membedakan kelas. Dan model dengan AUC di bawah 60% dianggap gagal, karena tidak dapat membedakan kelas dengan baik, kinerjanya mendekati tebakan acak.

3. HASIL DAN PEMBAHASAN

Dimulai dengan pembagian dataset yang sudah di eksplorasi data, yang terdiri dari 569 data yaitu *texture* (standar deviasi dari *grayscale values*), *perimeter*, *area*, *smoothness* (variasi lokal dalam panjang radius), *compactness*, *concavity*, *concave points* (jumlah bagian cekung dari kontur), *symmetry*, dan *fractal dimension.* Yang terdiri 2 kategori: benign and malignant, menjadi tiga dataset: latihan, validasi, dan uji. Model dilatih menggunakan *Split Data.*

3.1 Pengujian Preprocessing

Bagian ini merupakan pengujian tentang pengaruh adanya *Preprocessing* pada data penyakit kanker payudara yang digunakan ialah Cleaning data yang mana terdiri dari : *Missing Value, Duplicate Data, Outlier Check, Handling Outlier.*

Missing Value ialah tahap pra-pemrosesan data dimana nilai-nilai yang hilang atau absen dalam dataset diidentifikasi dan ditangani dengan cara tertentu sebelum data tersebut dimasukkan ke dalam model atau algoritma *Machine Learning*. Dalam *pre-proccesing* ini dilakukan : Penghapusan Data yang tidak memiliki nilai apapun.

```
diagnosis                0
radius_mean              0
texture_mean             0
perimeter_mean           0
area_mean                0
smoothness_mean          0
compactness_mean         0
concavity_mean           0
concave points_mean      0
symmetry_mean            0
fractal_dimension_mean   0
radius_se                0
texture_se               0
perimeter_se             0
area_se                  0
smoothness_se            0
```

(a)

```
compactness_se           0
concavity_se             0
concave points_se        0
symmetry_se              0
fractal_dimension_se     0
radius_worst             0
texture_worst            0
perimeter_worst          0
area_worst               0
smoothness_worst         0
compactness_worst        0
concavity_worst          0
concave points_worst     0
symmetry_worst           0
fractal_dimension_worst  0
dtype: int64
```

(b)

Gambar 2. (a) dan (b): Hasil Missing Value

Duplicate Data ialah tahap pra-pemrosesan data dimana sebuah data dilakukan pengecekan di dalam data tersebut memiliki satu atau beberapa data yang sama persis.

```
Jumlah duplikat data: 0
```

Gambar 3. Hasil *Duplicate Data*

TIN: Terapan Informatika Nusantara
Vol 5, No 2, July 2024, page 162-171
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i2.5408

Outlier Check ialah tahap pra-pemrosesan data dimana proses untuk mengidentifikasi dan menangani nilai-nilai yang dianggap sebagai outlier dalam dataset sebelum data tersebut dimasukkan ke dalam model atau algoritma *Machine Learning.* Dilakukan : Visualisasi Data dan Teknik Deteksi Outlier yang menggunakan metode Z-Score.

```
(569, 31)
```
(a)

```
(535, 31)
```
(b)

Gambar 4. (a) Data sebelum dihapus *Outlier***.** dan (b) Data sesudah dihapus *Outlier*

Handling Outlier ialah tahap pra-pemrosesan data dimana proses untuk menghapus sebuah data outlier yang tidak termasuk kategori manapun lalu melakukan metode handling menggunakan Z-Score yang memiliki data awal 569 menjadi 535 Data.

Pada preprocessing ditambahkan sebuah seleksi fitur XGBoost. Seleksi fitur tersebut ialah salah satu teknik yang digunakan dalam pre-processing dan pemodelan data untuk memilih fitur yang paling penting atau relevan dalam dataset data yang terpilih berjumlah 19 Data.

Tabel 3. Data XGBoost

Hasil Seleksi Fitur *XGBoost*	
Perimeter worst	Texture mean
Area worst	Radius se
Radius worst	Compactness se
Concave points worst	Concavity se
Concavity mean	Area se
Concave points mean	Smoothness worst
Concave point se	Symmetry worst
Perimeter se	Compactness worst
Texture worst	Fractal dimension worst
Concavity worst	

Encoder adalah tahap berikutnya dalam rantai pra-pemrosesan. *Encoder* adalah transformasi tipe data yang mengubah tipe data dari 'objek' menjadi 'kategori' untuk memungkinkan pembacaan data. Transformasi data yang dilakukan pada penelitian ini adalah mengubah data kategorikal menjadi data numerik, dengan 0 = jinak dan 1 = ganas, sehingga data dapat dengan mudah dimodelkan. Pendekatan yang digunakan adalah label encoding, seperti yang terlihat pada Gambar 4 di bawah ini.

```
diagnosis
0    345
1    190
```

Gambar 6. Hasil Label Encoder

Setelah memberi label pada Encoder, Standarisasi dan Normalisasi dilakukan lagi. Standarisasi dan normalisasi adalah langkah normalisasi data yang membantu mengurangi perbedaan skala di seluruh karakteristik. Teknik Penskalaan Min-Max digunakan untuk menormalkan data, seperti yang diilustrasikan pada Gambar 5.

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean
0	1.0	0.603597	0.027801	0.618667	0.494379	0.754503	0.884255	0.851831	0.797290	0.794416
1	1.0	0.745052	0.334440	0.697753	0.681753	0.368361	0.202931	0.246665	0.380325	0.409264
2	1.0	0.696804	0.478838	0.675045	0.610839	0.653551	0.481200	0.560318	0.693225	0.572335
3	1.0	0.729700	0.192116	0.714979	0.665033	0.546862	0.388398	0.562021	0.565312	0.407360
4	1.0	0.299852	0.248548	0.303657	0.192332	0.862338	0.515787	0.447914	0.438428	0.583756

Gambar 7. *Standarization and Normalization*

Tahap selanjutnya sebelum menguji model adalah SMOTE, yang juga dikenal sebagai pendekatan oversampling, yang melibatkan peningkatan jumlah data di kelas minoritas dengan menduplikasi jumlah data kelas minoritas secara acak hingga sama atau mendekati data kelas mayoritas. Temuannya adalah sebagai berikut.

```
diagnosis
0    345
1    190
```
(a)

```
diagnosis
1.0    345
0.0    345
```
(b)

Gambar 8. (a) Sebelum Melakukan SMOTE dan (b) Sesudah Melakukan SMOTE

TIN: Terapan Informatika Nusantara
Vol 5, No 2, July 2024, page 162-171
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i2.5408

Gambar 8 menunjukkan distribusi kelas dari sebuah dataset sebelum dan sesudah menerapkan *Synthetic Minority Over-sampling Technique* (SMOTE). Sebelum SMOTE, dataset tidak seimbang dengan 345 contoh kelas "0" (kelas mayoritas) dan 190 contoh kelas "1" (kelas minoritas). Setelah menerapkan SMOTE, dataset menjadi seimbang, dengan kedua kelas masing-masing memiliki 345 data. SMOTE menghasilkan sampel sintetis untuk kelas minoritas, memastikan bahwa model pembelajaran mesin yang dilatih pada data ini tidak bias terhadap kelas mayoritas.

3.2 Evaluasi Model

Bagian ini merupakan pengujian terhadap sebuah data penyakit kanker payudara yang dilakukan oleh peneliti dimulai dari preprocessing sampai dengan evaluasi model. Pada bagian ini akan diuraikan mengenai pembahasan dari penelitian maupun pengujian evaluasi model yang telah dilakukan. Prediksi model dilakukan menggunakan *K-Nearest Neighbor*, *Naive Bayes*, dan *Random Forest* dengan hasil akurasi pada tabel berikut.

Tabel 4. Confusion Matrix

Algoritma	*Dataset*	*Precision*	*recall*	*F1-score*	*Accuracy*
KNN without XGBoost & SMOTE	*Benig*	98%	98%	98%	98%
	Malignant	98%	98%	98%	
Naïve Bayes without XGBoost & SMOTE	*Benig*	95%	94%	94%	93%
	Malignant	91%	93%	92%	
Random Forest without XGBoost & SMOTE	*Benig*	97%	97%	97%	96%
	Malignant	95%	95%	95%	
KKN using XBoost	*Benig*	97%	97%	97%	96%
	Malignant	95%	95%	95%	
Naïve Bayes Using XGBoost	*Benig*	95%	94%	94%	93%
	Malignant	91%	93%	92%	
Randon Forest Using XGBoost	*Benig*	97%	97%	97%	96%
	Malignant	95%	95%	95%	
KNN Using SMOTE	*Benig*	99%	96%	97%	97%
	Malignant	96%	99%	97%	
Naïve Bayes Using SMOTE	*Benig*	90%	93%	92%	91%
	Malignant	92%	90%	91%	
Random Forest Using SMOTE	*Benig*	100%	96%	98%	98%
	Malignant	96%	100%	98%	
KNN using XGBoost & SMOTE	*Benig*	99%	97%	98%	98%
	Malignant	97%	99%	98%	
Naïve Bayes using XGBoost & SMOTE	*Benig*	93%	90%	91%	91%
	Malignant	90%	93%	91%	
Random Forest using XGBoost & SMOTE	*Benig*	100%	96%	98%	98%
	Malignant	96%	100%	98%	

Tabel 4 memberikan rincian hasil kinerja dari berbagai algoritma *machine learning* dalam empat skenario yang berbeda dalam berbagai kondisi pra pemrosesan, seperti penggunaan *XGBoost*, *SMOTE*, dan kombinasinya. Tabel 4 ini memberikan informasi tentang presisi, *recall*, *F1-score*, dan akurasi untuk set data jinak dan ganas dalam setiap kondisi.

Dalam kondisi tanpa *XGBoost* dan *SMOTE*, *KNN* menunjukkan performa yang tinggi dan seimbang dengan presisi, *recall*, *F1-score*, dan akurasi sebesar 98% untuk kedua kategori, menunjukkan performa yang kuat tanpa preprocessing. *Naïve Bayes* memiliki kinerja yang sedikit lebih rendah dengan presisi, *recall*, dan *F1-score* masing-masing sebesar 95%, 94%, dan 94% untuk jinak, serta 91%, 93%, dan 92% untuk ganas. Akurasi keseluruhannya adalah 93%, menunjukkan kinerja yang masuk akal tetapi tidak sekuat *KNN*. Sementara itu, *Random Forest* menunjukkan kinerja yang sangat baik dengan presisi, *recall*, dan *F1-score* sebesar 97% untuk jinak dan 95% untuk ganas, serta akurasi keseluruhan 96%. Hal ini menunjukkan keefektifannya dalam menangani dataset tanpa *preprocessing* seperti seleksi fitur dan *SMOTE*.

Ketika *XGBoost* digunakan dalam kombinasi dengan algoritma, tidak ada perubahan yang signifikan dalam kinerja *KNN* dan *Naïve Bayes*. *XGBoost* hanya sedikit menurunkan performa *KNN* dengan presisi, *recall*, dan *F1-score* sebesar 97% untuk kedua kategori, dan akurasi sebesar 96%. Hal ini menunjukkan bahwa *XGBoost* tidak secara signifikan meningkatkan kinerja *KNN*. Demikian juga, kinerja *Naïve Bayes* sedikit menurun dengan *XGBoost*, menghasilkan *F1-score* 94% dan 93% untuk jinak dan ganas, masing-masing, dan akurasi keseluruhan pada 93%. *Random Forest* tetap memiliki kinerja yang stabil dengan *XGBoost*, menunjukkan presisi, *recall*, dan *F1-score* sebesar 97% untuk jinak dan 95% untuk ganas, serta akurasi keseluruhan sebesar 96%.

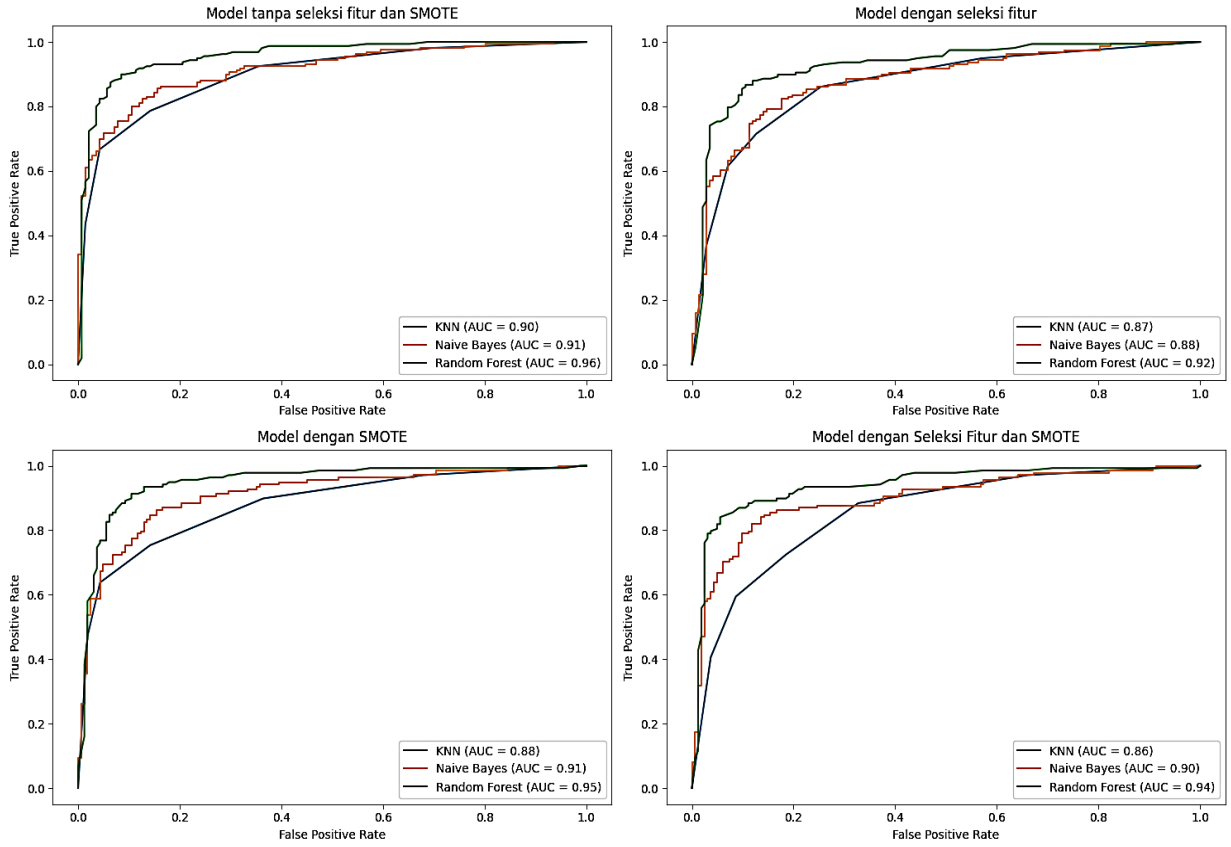
Penerapan *SMOTE* secara signifikan mempengaruhi kinerja *KNN* dan *Random Forest*. *KNN* mengalami peningkatan presisi untuk kasus jinak menjadi 99%, tetapi recall dan *F1-score* sedikit menurun menjadi 96% untuk kasus ganas. Akurasi keseluruhan tetap tinggi pada 97%, menunjukkan bahwa *SMOTE* menyeimbangkan kelas-kelas secara efektif. Sedangkan, *Random Forest* menunjukkan kinerja yang kuat dengan *SMOTE*, mencapai presisi 100%

TIN: Terapan Informatika Nusantara
Vol 5, No 2, July 2024, page 162-171
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i2.5408

untuk jinak dan 96% untuk ganas, dengan *recall* dan *F1-score* yang tinggi, serta akurasi keseluruhan sebesar 98%, menunjukkan dampak positif dari *SMOTE*.

Ketika *XGBoost* dan *SMOTE* digabungkan dengan algoritma, terjadi peningkatan kinerja. *KNN* mencapai presisi 99% dan *recall* 97% untuk kasus jinak, serta 97%, 99%, dan 98% untuk kasus ganas, dengan akurasi keseluruhan 98%. *Naïve Bayes* menunjukkan metrik yang seimbang dengan presisi, *recall*, dan F1-skor sekitar 90-93% untuk kedua kategori, dan akurasi keseluruhan sebesar 91%. Random Forest memiliki peningkatan kinerja yang signifikan dengan presisi 100% untuk jinak dan 96% untuk ganas, serta skor F1 98% dan akurasi keseluruhan pada 98%. Hasil ini menekankan pentingnya memilih teknik *preprocessing* yang sesuai dengan algoritma yang digunakan untuk mengoptimalkan kinerja.

Kurva *Receiver Operating Characteristic* (ROC) dan kurva *Area Under the Curve* (AUC) digunakan sebagai metode evaluasi tambahan untuk menilai kinerja model secara keseluruhan dalam penelitian ini. Namun, angka akurasi dan presisi hanya memberikan penilaian awal. Grafik di bawah ini menunjukkan hasil evaluasi.



Gambar 9. Kurva ROC dan Nilai AUC

Kurva ROC AUC yang disajikan dalam gambar 9 menunjukkan kinerja tiga algoritma, yaitu *K-Nearest Neighbors* (KNN), *Naïve Bayes*, dan *Random Forest*, dalam empat skenario pra pemrosesan yang berbeda. Pada skenario tanpa seleksi fitur dan *SMOTE*, *Random Forest* memiliki kinerja terbaik dengan AUC tertinggi sebesar 96%. Hal ini menunjukkan bahwa algoritma ini efektif dalam mengklasifikasikan antara kategori jinak dan ganas. *Naïve Bayes* juga memiliki kinerja yang bagus dengan AUC 91%, sedangkan KNN memiliki AUC terendah sebesar 90%, menunjukkan bahwa algoritma ini kurang efektif tanpa pra pemrosesan.

Ketika menggunakan seleksi fitur, performa semua model sedikit menurun. AUC *Random Forest* tetap tinggi di 92%, tetapi *Naïve Bayes* dan KNN mengalami penurunan menjadi 88% dan 87%. Hal ini menunjukkan bahwa seleksi fitur dapat menghilangkan informasi yang berguna bagi model-model tersebut. Penerapan *SMOTE* meningkatkan performa *Naïve Bayes* dengan mempertahankan AUC 91%. Namun, *Random Forest* mengalami sedikit penurunan menjadi 95%. KNN juga mengalami penurunan menjadi AUC 88%, menunjukkan bahwa meskipun *SMOTE* membantu menyeimbangkan kelas, tetapi juga memperkenalkan *noise* yang mempengaruhi KNN dan *Random Forest*.

Pada saat seleksi fitur digabungkan dengan *SMOTE*, hasilnya bervariasi. AUC *Random Forest* menurun menjadi 94%, sementara AUC *Naïve Bayes* sedikit meningkat menjadi 90%, dan AUC KNN semakin menurun menjadi 86%. Kombinasi ini tampaknya paling mengoptimalkan kinerja *Naïve Bayes*, sementara model-model lainnya terpengaruh secara negatif, mungkin karena penyederhanaan yang berlebihan dari pemilihan fitur dan noise data sintetis dari *SMOTE*. Seleksi fitur dan *SMOTE* dapat mempengaruhi kinerja model secara positif atau negatif tergantung pada model itu sendiri, dan oleh karena itu perlu dilakukan dengan pertimbangan yang cermat untuk menghindari penurunan kinerja yang tidak diinginkan.


TIN: Terapan Informatika Nusantara
Vol 5, No 2, July 2024, page 162-171
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i2.5408

4. KESIMPULAN

Penelitian ini mengeksplorasi berbagai teknik preprocessing dan algoritma pembelajaran mesin untuk diagnosis kanker payudara, dengan menekankan dampak metode seperti XGBoost dan SMOTE pada kinerja model. Analisis dimulai dengan pembersihan data dan penanganan nilai yang hilang, outlier, dan data duplikat untuk memastikan integritas dataset. Pemilihan fitur menggunakan XGBoost mengidentifikasi fitur yang paling signifikan, meningkatkan efisiensi model. Teknik seperti SMOTE mengatasi ketidakseimbangan kelas dengan mengambil sampel berlebih pada kelas minoritas, sementara algoritme seperti K-Nearest Neighbor (KNN), Naïve Bayes, dan Random Forest digunakan untuk mengklasifikasikan data. Evaluasi menggunakan confusion matrix dan kurva ROC-AUC untuk menilai akurasi, presisi, recall, dan F1-score. pada evaluasi confusion matrix hasil kinerja berbagai algoritma machine learning (KNN, Naïve Bayes, dan Random Forest) dalam empat skenario berbeda dengan kondisi pra-pemrosesan yang bervariasi (XGBoost, SMOTE, dan kombinasinya). Tanpa XGBoost dan SMOTE, KNN menunjukkan performa unggul dengan presisi, recall, F1-score, dan akurasi 98%, sementara Naïve Bayes dan Random Forest masing-masing memiliki akurasi 93% dan 96%. Penerapan XGBoost tidak signifikan meningkatkan kinerja, namun SMOTE secara signifikan meningkatkan performa KNN dan Random Forest, dengan akurasi keseluruhan meningkat menjadi 97% dan 98% masing-masing. Kombinasi XGBoost dan SMOTE memberikan peningkatan kinerja terbesar, dengan KNN dan Random Forest mencapai akurasi 98%, menekankan pentingnya teknik pra-pemrosesan yang tepat. Selain itu, kurva ROC dan AUC digunakan untuk evaluasi tambahan, menunjukkan pentingnya memilih teknik yang sesuai untuk mengoptimalkan kinerja model. sedangkan hasil evaluasi ROC AUC menunjukkan kinerja tiga algoritma (KNN, Naïve Bayes, dan Random Forest) dalam empat skenario pra-pemrosesan berbeda berdasarkan kurva ROC AUC. Tanpa seleksi fitur dan SMOTE, Random Forest unggul dengan AUC 96%, Naïve Bayes memiliki AUC 91%, dan KNN terendah dengan 90%. Seleksi fitur menurunkan AUC semua model: Random Forest ke 92%, Naïve Bayes ke 88%, dan KNN ke 87%, mengindikasikan hilangnya informasi penting. SMOTE meningkatkan AUC Naïve Bayes ke 91% tetapi menurunkan Random Forest ke 95% dan KNN ke 88%, menunjukkan pengenalan noise. Kombinasi seleksi fitur dan SMOTE memberikan hasil beragam: Random Forest AUC 94%, Naïve Bayes 90%, dan KNN 86%, menunjukkan bahwa kombinasi ini mengoptimalkan Naïve Bayes tetapi menurunkan performa model lainnya karena penyederhanaan berlebihan dan noise sintetis.

REFERENCES

Adhikary, S., & Banerjee, S. (2023). Introduction to Distributed Nearest Hash: On Further Optimizing Cloud Based Distributed kNN Variant. *Procedia Computer Science*, 1571-1580.

Ali, A., Hamraz, M., Gul, N., Khan, D. M., Aldahmani, S., & Khan, Z. (2023). A k nearest neighbour ensemble via extended neighbourhood rule and feature subsets. *Pattern Recognition*.

Carli, F., Leonelli, M., & Varando, G. (2023). A new class of generative classifiers based on staged tree models. *Knowledge-Based Systems*, 110-488.

Edgar, T. W., & Manz, D. O. (2017). *Research Methods for Cyber Security*. Elsevier Science.

Fauzan, M., Gusti, S. K., Jasril, & Pizaini. (2023). Penerapan Seleksi Fitur Untuk Klasifikasi Penerima Bantuan Sosial Pangkalan Sesai Menggunakan Metode K-Nearest Neighbor. *Jurnal Sistem Komputer dan Informatika (JSON)*, 1-10.

Hidayata, R., Kartinia, D., Mazdadia, M. I., Budiman, I., & Ramadhania, R. (2023). Implementasi Seleksi Fitur Binary Particle Swarm Optimization pada Algoritma K-NN untuk Klasifikasi Kanker Payudara. *Jurnal Sistem dan Teknologi Informasi*, 135-139.

Ismafillah, D., Rohana, T., & Cahyana, Y. (2023). Analisis algoritma pohon keputusan untuk memprediksi penyakit diabetes menggunakan oversampling smote. *NFOTECH:Jurnal Informatika Teknologi*, 27-36.

Kurnia, D., Madadia, M. I., Kartini, D., Nugroho, R. A., & Abadi, F. (2023). SELEKSI FITUR DENGAN PARTICLE SWARM OPTIMIZATION PADA KLASIFIKASI PENYAKIT PARKINSON MENGGUNAKAN XGBOOST. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, 1083-1094.

Manju, N., Harish, B. S., & Prajwal, V. (2019). Ensemble Feature Selection and Classification of Internet Traffic using XGBoost Classifier. *I. J. Computer Network and Information Security*, 37-44.

Meuthia Zulma, G. D., Angelika, & Chamidah, N. (2021). Perbandingan Metode Klasifikasi Naive Bayes, Decision Tree Dan K-Nearest Neighbor Pada Data Log Firewall. *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)*, 679-688.

Nguyen, B. H., Xue, B., & Zhang, M. (2020). A survey on swarm intelligence approaches to feature selection in data mining. *Swarm and Evolutionary Computation*.

Nurdian, R. A., Ridwan, M., & Yusuf, A. (2022). Komparasi Metode SMOTE dan ADASYN dalam Meningkatkan Performa Klasifikasi Herregistrasi Mahasiswa Baru. *Jurnal Teknik Informatika dan Sistem Informasi*, 24-32.

Pratama, & Adhitya, Y. (2019). Analisis Metode Seleksi Fitur untuk Meningkatkan Akurasi pada Variant Metode Klasifikasi K-Nearest Neighbor (kNN).

R, D., Avilala, S. V., & Subramaniyaswamy, V. (2019). Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes, KNN and Random Forest. *IEEE*, 679-684.

TIN: Terapan Informatika Nusantara
Vol 5, No 2, July 2024, page 162-171
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i2.5408

Rifatama, M. I., Faisal, M. R., Hertono, R., Budiman, I., & Mazdadi, M. I. (2023). OPTIMASI ALGORITMA K-NEAREST NEIGHBOR DENGAN SELEKSI FITUR MENGGUNAKAN XGBOOST. *JIRE (Jurnal Informatika & Rekayasa Elektronika)*.

Setiawan, & Yohanes. (2023). Data Mining berbasis Nearest Neighbor dan Seleksi Fitur untuk Deteksi Kanker Payudara. *Jurnal Informatika: Jurnal pengembangan IT (JPIT)*.

Shanshool, A. M., Hussien Saeed, E. M., & Khaleel, H. H. (2023). Comparison of various data mining methods for early diagnosis of human cardiology. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 1343-1351.

Siregar, A. M., Tukino, Faisal, S., Fauzi, A., & Kadori, I. (2020). Klasifikasi untuk Prediksi Cuaca Menggunakan Esemble Learning. *PETIR:Jurnal Pengkajian dan Penerapan Teknik Informatika*, 138-147.

Suhliyyah, Handayani, H. H., & Baihaqi, K. A. (2023). Implementasi Algoritma Logistic Regression Untuk Klasifikasi Penyakit Stroke. *Syntax: Jurnal Informatika*, 15-23.

Syafei, R. M., & Efrilianda, D. A. (2023). Machine Learning Model Using Extreme Gradient Boosting (XGBoost) Feature Importance and Light Gradient Boosting Machine (LightGBM) to Improve Accurate Prediction of Bankruptcy. *Recursive Journal of Informatics*, 64-72.

V, K., & S, S. P. (2022). Adaptive boosted random forest-support vector machine based classification scheme for speaker identification. *Applied Soft Computing*, 109-826.

Wilantapoera, A. W., Astuti, W., & Dwifebri, M. (2023). Analisis Sentimen Kategori Aspek Pada Ulasan Produk Menggunakan Metode KNN Dengan Seleksi Fitur Mutual Information. *e-Proceeding of Engineering*, 1673-1681.

Zhang, X., Shi, Z., Liu, X., & Li, X. (2018). A Hybrid Feature Selection Algorithm For Classification Unbalanced Data Processsing. *IEEE International Conference on Smart Internet of Things (SmartIoT)*, 269-275.