TIN: Terapan Informatika Nusantara
Vol 5, No 2, July 2024, page 144-154
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i2.5324

Optimasi Metode Support Vector Machine Menggunakan Seleksi Fitur Recursive Feature Elimination dan Forward Selection untuk Klasifikasi Kanker Payudara

Eva Senia Septiany[*], Hanny Hikmayanti Handayani, Tohirin Al Mudzakir, Anis Fitri Nur Masruriyah

Fakultas Ilmu Komputer, Teknik Informatika, Universitas Buana Perjuangan Karawang, Karawang, Indonesia
Email: [1,*]if20.evaseptiany@mhs.ubpkarawang.ac.id, [2]hanny.hikmayanti@ubpkarawang.ac.id, [3]tohirin@ubpkarawang.ac.id, [4]anis.masruriyah@ubpkarawang.ac.id
Email Penulis Korespondensi: if20.evaseptiany@mhs.ubpkarawang.ac.id

Abstrak−Kanker, penyebab utama kematian global yang diakibatkan oleh *proliferasi* sel abnormal yang menyebar di luar batas jaringan normal. Kanker payudara merupakan salah satu jenis kanker yang paling banyak ditemui, dengan sekitar 2,26 juta kasus dilaporkan pada tahun 2020. Penelitian ini bertujuan untuk mengembangkan algoritma *Support Vector Machine* (SVM) yang lebih efektif untuk klasifikasi kanker payudara melalui teknik seleksi fitur yang efisien. Penelitian sebelumnya telah menggunakan berbagai algoritma seperti *K-Nearest Neighbor* dan *Logistic Regression* untuk identifikasi kanker payudara. Penelitian ini berfokus pada peningkatan akurasi dengan menggunakan metode pemilihan fitur alternatif seperti *Recursive Feature Elimination* (RFE) dan *Forward Selection*. Dataset yang digunakan terdiri dari 569 *instance* dengan 32 fitur yang bersumber dari UCI *Machine Learning Repository*, dan diklasifikasikan ke dalam kategori jinak dan ganas. Metode *pre-processing* data, termasuk pembersihan data, pengkodean, dan pemilihan fitur, diterapkan pada dataset. Teknik RFE dan Forward Selection digunakan untuk mengidentifikasi fitur paling penting untuk pelatihan model. Evaluasi model SVM yang ditingkatkan menunjukkan akurasi pelatihan hampir 100% dan akurasi *Cross Validation* sebesar 97%, menunjukkan efektivitas pendekatan yang diusulkan dalam konteks kanker payudara. Selain itu, *Learning Curve* dan pengujian menunjukkan kestabilan model SVM tanpa tanda-tanda *overfitting* atau *underfitting*. Dengan demikian, penelitian ini mengembangkan algoritma SVM dengan metode pemilihan fitur yang menghasilkan hasil akurasi yang lebih baik dalam klasifikasi kanker payudara.

Kata Kunci: *Breast Cancer*; RFE; *Forward Selection*; SVM

Abstract−Cancer, the leading cause of global death, results from abnormal cell proliferation that spreads beyond the boundaries of normal tissue. Breast cancer is one of the most common types of cancer, with approximately 2.26 million cases reported in 2020. This research aims to develop a more effective Support Vector Machine (SVM) algorithm for breast cancer classification through efficient feature selection techniques. Previous research has used various algorithms such as K-Nearest Neighbor and Logistic Regression for breast cancer identification. This research focuses on improving accuracy by using alternative feature selection methods such as Recursive Feature Elimination (RFE) and Forward Selection. The dataset used consists of 569 instances with 32 features sourced from the UCI Machine Learning Repository, and classified into benign and malignant categories. Data pre-processing methods, including data cleaning, coding, and feature selection, were applied to the dataset. RFE and Forward Selection techniques were used to identify the most important features for model training. Evaluation of the improved SVM model shows a training accuracy of nearly 100% and a Cross Validation accuracy of 97%, demonstrating the effectiveness of the proposed approach in the context of breast cancer. In addition, the Learning Curve and testing showed the stability of the SVM model with no signs of overfitting or underfitting. Thus, this study developed an SVM algorithm with a feature selection method that produces better accuracy results in breast cancer classification.

Keywords: Breast Cancer; RFE; Forward Selection; SVM

1. PENDAHULUAN

Berdasarkan *World Health Organization*, kanker termasuk penyebab utama kematian di seluruh dunia. Kanker dapat muncul di hampir semua organ atau berkembang biak di jaringan tubuh ketika terdapat sel-sel abnormal yang tidak terkendali dan melampaui batas normal tubuh kemudian menyebar ke organ-organ lainnya (World Health Organization: WHO, 2022). Kanker yang paling umum yaitu *breast, lung, colon, rectum* dan *prostate cancers*. Terhitung pada tahun 2020 mencapai hampir 10 juta kematian atau hampir satu dari enam kematian, dimana *breast cancers* memiliki kasus terbanyak dengan mencapai 2.26 juta kasus. Kanker payudara adalah jenis kanker yang berasal dari jaringan kelenjar payudara juga terdapat pada jaringan lemak atau jaringan ikat dalam payudara (Setiawan, 2023).

Jumlah kanker payudara di Indonesia menempati peringkat pertama pada tahun 2020. Data tersebut didapatkan dari Global Cancer Observatory (Globocan) dengan jumlah kasus baru sebanyak 65.858 dan jumlah kematian karena kanker payudara mencapai 22.430 (*Cancer Today*, 2021). Oleh karena itu, pemilihan fitur yang tepat diperlukan untuk mengoptimalkan algoritma SVM dalam mengklasifikasikan penyakit kanker payudara.

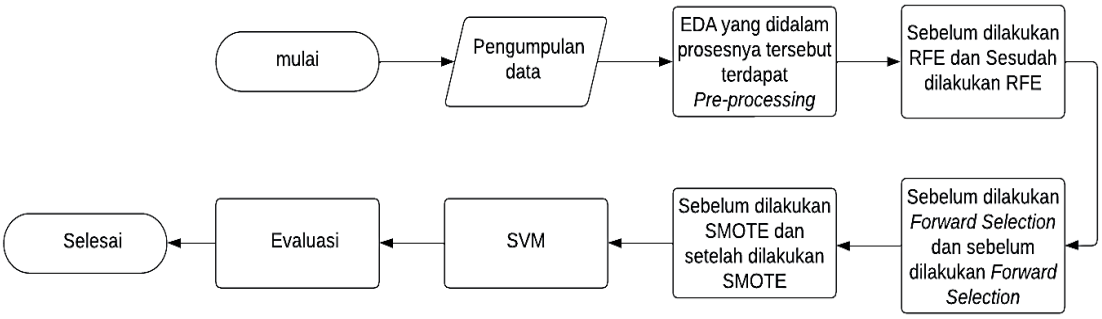
Adapun penelitian terkait sebelumnya meneliti tentang kanker payudara dan diselesaikan menggunakan algoritma *K-Nearest Neighbor* untuk menentukan jenis kanker payudara tersebut termasuk jinak atau ganas (Chazar & Erawan, 2020). Kemudian data kanker payudara yang digunakan tersebut dibagi menjadi data *test* dan data *training* untuk menghitung jarak terdekat. Sehingga hasil model algoritma K-NN dapat memprediksi jenis kanker jinak atau ganas. Penelitian berikutnya meneliti kanker payudara menggunakan algoritma Regresi Logistik dan SVM. Pada penelitian tersebut data kanker payudara yang memiliki 30 fitur diseleksi menggunakan *Backward Elimination* (Farahdiba et al., 2023). Kemudian penelitian tersebut membandingkan dengan membuat 4 model pengujian yaitu dua algoritma klasifikasi Regresi Logistik dan SVM, serta dua model dengan menambahkan seleksi fitur *Backward*

TIN: Terapan Informatika Nusantara
Vol 5, No 2, July 2024, page 144-154
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i2.5324

Elimination, sehingga hasil akurasi dari model Logistik Regression yaitu 94.04% dan hasil dengan *Backward Elimination* mendapatkan 95.43%, terlihat mengalami peningkatan sebesar 1.39%. Sedangkan hasil dari SVM yaitu 96.14% dan hasil dari penambahan *Backward Elimination* yaitu 97.02%, ada peningkatan sebesar 0.88%. Penelitian lainnya menggunakan algoritma *Support Vector Machine* menggunakan *Particle Swarm Optimization* untuk mendiagnosa penyakit kanker payudara (Maulana et al., 2022). Sehingga memperoleh hasil bahwa *Particle Swarm Optimization* dapat membantu mendiagnosa kanker payudara lebih akurat. Kemudian, didapatkan akurasi sebesar 97,28% (kenaikan akurasi 0,57%), adapun akurasi sebelum menggunakan PSO adalah 95,71%, presisi sebesar 95,55%, terdapat peningkatan 0,93% dari sebelumnya yaitu 99,21%. Selanjutnya penelitian tentang kanker payudara diselesaikan menggunakan algoritma SVM dan Logistik Regresion untuk mendeteksi kanker tersebut ganas atau jinak (Nurjanah et al., 2023). Pada penelitian tersebut data kanker payudara menunjukan adanya ketidak seimbangan, kemudian untuk menyeimbangkan data yaitu menggunakan metode *oversampling* SMOTE, adapun hasilnya yaitu *accuracy* sebesar 1.0, *precision* 1.0 dan *recall* 1.0. Setelah implementasi dilakukan langkah selanjutnya yaitu deployment menggunakan RAD. Sehingga hasil akurasi model adalah 90%. Penelitian selanjutnya tentang klsifikasi curah hujan diselesaikan mengggunakan algoritma SVM (Pratama et al., 2022). pada penelitian tersebut teknik data mining CRISP-DM digunakan dalam penelitian ini mengkategorikan data curah hujan ke dalam kategori sedang dan lebat. Kemudian, SVM dengan teknik optimasi pemilihan fitur menggunakan RFE (*Recursive Feature Elimination*) adalah algoritma yang digunakan untuk mengklasifikasikan curah hujan tersebut. sehingga akurasi model SVM menjadi 77% sebelum menerapkan RFE dan meningkat 2% menjadi 79% setelah menerapkan RFE. Adapun penelitian selanjutnya yang melakukan penelitian mengenai penyakit diabetes menggunakan algoritma *Decision Tree* dan *Random Forest* (Ismafillah et al., 2023). pada penelitian tersebut menggunakan teknik SMOTE untuk menangani ketidakseimbangan data. Kemudian, dilakukan pengujian menggunakan model *Random Forest* dan SMOTE dengan *Decision Tree* dan SMOTE. Adapun untuk memprediksi seseorang termasuk menderita diabetes dievaluasi menggunakan *K-Fold Cross Validation*. Sehingga didapatkan hasil bahwa model *Random Forest* dan SMOTE lebih unggul dalam distribusi data diabetes. Penelitian selanjutnya menggunakan semua algoritma *machine learning* untuk klasifikasi kanker payudara (koirunnisa et al., 2023). Penelitian ini menggunakan teknik PCA untuk pemilihan fitur dan SMOTE untuk menangani ketidakseimbangan data. Pengujian juga dilakukan menggunakan model SVM dengan kernel RBF dan SMOTE, serta berbagai model *machine leaning*. Penelitian ini juga menggunakan K-Fold Cross Validation untuk melakukan evaluasi prediksi kanker payudara. Sehingga hasilnya menunjukkan bahwa model SVM dengan kernel RBF dan SMOTE lebih baik dalam klasifikasi kanker payudara.

Penelitian ini dilakukan untuk mengembangkan penelitian yang dilakukan oleh Farahdiba, et al (Farahdiba et al., 2023). Adapun pada penelitian tersebut tidak membahas seleksi fitur lain selain *backward elimination* sehingga dalam penelitian tersebut menyarankan untuk menguji pemilihan fitur selain *backward elimination.* Maka dari itu pada penelitian ini bertujuan menggunakan seleksi fitur lain yaitu *forward selection* dan *recursive feature elimination* (RFE) untuk pengembangan penelitian sebelumnya.

2. METODOLOGI PENELITIAN



Gambar 1. Alur Penelitian

Pada Gambar 1 tahapan dimulai dengan pengumpulan data, dilanjutkan dengan *Exploratory Data Analysis* (EDA) yang didalam prosesnya terdapat *pre-processing* data, penggunaan fitur yang dipilih menggunakan RFE dan *forward selection*, kemudian ke tahap pemodelan menggunakan algoritma *Support Vector Machine* dan evaluasi menggunakan *Confusion Matrix, Cross Validation dan Learning Curve*. Penelitian ini bertujuan untuk pengembangan model klasifikasi penyakit kanker payudara menggunakan algoritma SVM dan pemilihan fitur yaitu *forward selection* dan RFE.

2.1 Pengumpulan Data

Penelitian ini menggunakan dataset *Breast Cancer Winconsin Diagnostic* yang didapatkan dari website UCI *Machine Learning Respository*. Dataset ini memiliki 32 fitur dan 569 data dan terdapat dua kelas yaitu *Benig* (B) dengan total 357 dan *Malignant* (M) dengan total 212. Dalam dataset ini semua variabel independen merupakan kategori data numerik, sedangkan variabel dependen diubah ke biner. Selain itu, dataset ini terdapat tiga fitur yaitu *Mean* (rata-rata),

TIN: Terapan Informatika Nusantara
Vol 5, No 2, July 2024, page 144-154
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i2.5324

se (standar deviasi), dan *worst* (terburuk). Fitur tersebut terbagi ke dalam setiap kolom dengan berbagai ukuran yang mencangkupp *radius* (rata-rata jarak dari pusat ke titik-tik di sekeliling), *texture* (standar deviasi dari *gray-scale values*), *perimeter*, *area*, *smoothness* (variasi lokal dalam panjang radius), *compactness*, *concavity*, *concave points* (jumlah bagian cekung dari kontur), *symmetry*, dan *fractal dimension*. Fitur tersebut diperoleh dari gambar digital *Fine Needle Aspirate* (FNA) menggunakan massa payudara.

2.2 *Exploratory Data Analysis*

Analisis data eksplorasi yang komprehensif dilakukan untuk mendapatkan pemahaman yang lebih dalam tentang korelasi antara berbagai parameter pada kanker payudara (Nurjanah et al., 2023). Analisis Data Eksplorasi merupakan peran penting sebelum terlibat dalam proses pemodelan, karena berfungsi sebagai langkah awal dalam memahami data dan struktur yang mendasarinya (Qadrini et al., 2023). Berikut merupakan analisis data eksplorasi dari peneltian ini:

2.2.1 Membaca Data

Pada bagian ini menampilkan jumlah data, kolom, dan menghapus kolom *unamed* dan *id* menggunakan *df.drop* karena tidak diperlukan. Pada bagian ini juga menampilkan korelasi data secara keseluruhan dan distribusi data tersebut.

2.2.2 Menampilkan Info Data

Pada tahapan ini menampilkan tipe data, menujukan target data terpilih yaitu diagnosis yang terdapat dua kelas yaitu *Benig* dengan jumlah 357 data dan *Malignant* 212 data, sehingga dengan hasil data tersebut menunjukan bahwa data tersebut tidak seimbang.

2.2.3 Pre-Processing

Terdapat beberapa teknik yang digunakan pada tahap ini yaitu sebagai berikut:

a. *Cleaning Data*
   Tujuan dari tahap pembersihan data dalam penelitian ini adalah untuk memastikan bahwa data tidak mengandung masalah yang dapat mengganggu analisis. Prosedur yang diikuti termasuk menghilangkan *Noisy*, *Missing Value*, *Duplicate Data*, dan menemukan serta mengelola data yang tidak relevan. Data yang dihasilkan setelah melalui proses ini akan lebih akurat dan siap untuk digunakan pada tahap analisis selanjutnya.
b. *Encoding*
   Encoding dalam penelitian ini digunakan untuk mempermudah pembacaan data. *Encoding* dalam hal ini melibatkan pengubahan satu tipe data dari "objek" menjadi "kategori". Selain itu, transformasi data dilakukan untuk mengubah data numerik dari data kategorikal, yang membuat pemodelan data menjadi lebih mudah.
c. *Normalization*
   Tahapan normalisasi data dilakukan untuk mengatasi perbedaan skala antar atribut dalam dataset, sehingga analisis data menjadi lebih konsisten dan akurat. Metode normalisasi yang digunakan dalam penelitian ini adalah *Min-Max Scaler*. Dengan menggunakan teknik pra-pemrosesan *Min-Max Scaler*, nilai fitur dalam dataset disesuaikan agar berada di antara 0 dan 1 (Aji & Suprianto, 2023).

2.3 Seleksi Fitur

Seleksi fitur merupakan salah satu metode dalam tahap *pre-processing* data mining. Seleksi fitur bertujuan untuk mengurangi kompleksitas atribut yang diolah selama proses analisis. Proses ini bertujuan untuk mengidentifikasi fitur yang paling penting dalam kolom yang terdapat pada dataset penyakit kanker payudara. Seleksi fitur juga sering digunakan untuk mengurangi dimensi model, membantu dalam mengurangi jumlah fitur domain, serta menghilangkan fitur yang tidak relevan (Adnyana, 2019). Adapun seleksi fitur yang digunakan yaitu sebagai berikut:

2.3.1 *Recursive Feature Elimination* (RFE)

Recursive feature elimination (RFE) adalah teknik seleksi fitur yang bekerja dengan cara menghapus fitur-fitur yang memiliki hubungan yang lemah atau tidak memiliki hubungan sama sekali. Sehingga ditemukan subset fitur yang optimal yang akan digunakan dalam proses pembangunan model. Model kemudian dibangun dengan menggunakan subset fitur yang tersisa dan akurasinya dapat ditentukan. RFE digunakan untuk mengevaluasi fitur-fitur dalam hubungannya satu sama lain. Proses pemilihan fitur ini bertujuan untuk membuat model yang cocok dengan semua atribut yang relevan (Pratama et al., 2022).

Tabel 1. *Pseudocode* Seleksi Fitur RFE

Input :
x: Data fitur (tanpa kolom 'diagnosis')
y: Label ('diagnosis')
Output:
Selected_features_rfe: Nama fitur yang telah terpilih setelah RFE
X_rfe: Data fitur yang hanya berisi fitur terpilih
Method:

TIN: Terapan Informatika Nusantara
Vol 5, No 2, July 2024, page 144-154
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i2.5324

1. Menginisialisasi model SVM dengan kernel linear dan parameter C=1
2. Membuat objek RFE dengan estimator (model) dan jumlah fitur yang ingin di pilih (dalam penelitian ini memilih 16 fitur).
3. Lakukan RFE pada data fitur dan label
4. Dapatkan indeks fitur yang terpilih
5. Nama fitur yang dipilih akan didapatkan
6. Membuat dataset baru dengan fitur yang terpilih
7. Menampilkan informasi hasil RFE
8. Selesai

Proses seleksi fitur dengan metode RFE dan SVM sebagai estimator ditunjukkan oleh *pseudocode* di atas. Untuk memprediksi label 'diagnosis', RFE akan memilih fitur yang paling relevan. Sebuah dataset baru dengan fitur-fitur yang dipilih akan dihasilkan setelah proses ini selesai.

2.3.2 *Forward Selection*

Forward selection merupakan salah satu metode *wrapper* yang digunakan untuk memilih fitur yang dilakukan sebelum proses klasifikasi. Tujuan utama dari pendekatan ini adalah untuk memilih karakteristik yang relevan dengan data dan memiliki dampak besar pada hasil klasifikasi. *Forward selection* menghilangkan elemen-elemen yang kurang relevan dari pertimbangan dan hanya mempertahankan elemen-elemen yang benar-benar menambah nilai pada model. Pratama K. I. & Kusnawi menyatakan bahwa hal ini berusaha untuk meningkatkan kinerja metode klasifikasi secara keseluruhan (Kusnawi & Khrisna Irham Fadhil Pratama, 2023). Pendekatan ini dapat mengurangi waktu komputasi dan kompleksitas selain meningkatkan akurasi model, yang akan mempercepat proses klasifikasi. Adapun pada penelitian ini *forward selection* yang digunakan yaitu *Sequential Feature Selection* (SFS).

Tabel 2. *Pseudocode Forward Selection*

Input :
x: Data fitur (tanpa kolom ‘diagnosis’)
y: Label (‘diagnosis’)
Output:
Selected_features_sfs: Nama fitur yang telah terpilih setelah SFS
X_sfs: Data fitur yang hanya berisi fitur terpilih
Method:
1. Menginisialisasi model SVM dengan kernel linear
2. Membuat objek SFS dengan estimator (model), jumlah fitur yang ingin dipilih, dan arah seleksi (forward)
3. Melakukan SFS pada data fitur dan label
4. Dapatkan indeks fitur yang terpilih
5. Nama fitur yang dipilih akan didapatkan
6. Membuat dataset baru dengan fitur yang terpilih
7. Menampilkan informasi hasil RFE
8. Selesai

Pseudocode di atas menunjukan proses menggunakan *Support Vector Machine* (SVM) sebagai estimator dan teknik *Sequential Feature Selection* (SFS) untuk pemilihan fitur. Untuk mengantisipasi label 'diagnosis', SFS akan memilih fitur yang paling relevan.

2.4 SMOTE

SMOTE merupakan teknik oversampling yang digunakan untuk menghasilkan sampel sintesis dari kelas minoritas sehingga menciptakan keseimabangan antara kelas mayoritas dan minoritas. Dengan menerapkan SMOTE, representasi kelas minoritas dalam dataset dapat ditingkatkan, yang meningkatkan kinerja model dan hasil evaluasinya (Fitriyani et al., 2023).

Tabel 3. *Pseudocode* SMOTE

Input:
T: Jumlah sampel kelas minoritas
N: Persentase SMOTE (misalnya, 100%)
k: Jumlah tetangga terdekat
Output:
Sampel minoritas sintetis sebanyak (N/100) * T
Method:
1. Inisialisasi dataset (X) dan label (y)
   X: adalah matriks fitur dengan m baris (jumlah data) dan n kolom (jumlah fitur)
   y: adalah vektor label dengan m elemen (kelas/target)

TIN: Terapan Informatika Nusantara
Vol 5, No 2, July 2024, page 144-154
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i2.5324

```
2. Untuk setiap sampel kelas minoritas:
   for each minority sample in X:
       Pilih k tetangga terdekat dari sampel minoritas
       neighbors = k_nearest_neighbors(minority sample, X)
3. Untuk setiap tetangga:
   for each neighbor in neighbors:
       Hitung perbedaan antara fitur sampel minoritas dan tetangga
       difference = neighbor - minority sample
4. Generate sampel sintetis pada garis segmen antara dua fitur
   synthetic_sample = minority sample + random_number(0, 1) * difference
5. Tambahkan sampel sintetis ke dataset
   X_synth.append(synthetic_sample)
   y_synth.append(minority label)
6. Selesai
```

Pseudocode yang ditampilkan menunjukkan teknik SMOTE (*Synthetic Minority Over-sampling Technique*), yang mencoba membangun sampel sintetis dari sampel kelas minoritas. Dengan memanfaatkan tetangga terdekat, metode ini memperluas zona keputusan kelas minoritas dan menghasilkan sampel sintetis.

2.5 *Support Vector Machine* (SVM)

Seperti yang ditunjukkan oleh penelitian sebelumnya, metode SVM secara efisien menemukan hyperplane dengan margin terbesar di antara dua kelas (Ramadhan, 2021). Jarak antara hyperplane dan vektor pendukung titik data terdekat untuk setiap kelas mendefinisikan margin. Dengan menggunakan kernel, manfaat SVM dapat ditingkatkan. Sebagai sebuah fungsi, kernel meningkatkan dimensi data asli, sehingga memudahkan pemisahan data. Berbagai macam fungsi kernel yang tersedia ditampilkan pada Tabel 3. SVM dapat menangani masalah klasifikasi non-linear dalam dimensi asli dengan metode ini. Setelah hyperplane diidentifikasi, SVM menggunakan posisinya di hyperplane untuk meramalkan kelas data uji (Masruriyah et al., 2024).

Tabel 4. Jenis Kernel

SVM	Jenis Kernel	Rumus
Linear	*Linear*	K (x,y) = x.y
	Polynomial	K (x,y) = $(x.y + 1)^p$
Non-Linear	RBF	K (x,y) = $e^{-\|x.y\|2/2\sigma 2}$
	Sigmoid	(x,y) = tanh(Kx.y - δ)

Adapun berikut adalah Pseudocode dari pemodelan SVM.

Tabel 5. *Pseudocode* SVM

```
Input :
x: Data fitur (tanpa kolom 'diagnosis')
y: Label ('diagnosis')
Output:
w: Vektor bobot (koefisien) dari model SVM
b: Konstanta bias dari model SVM
Prediksi: Array prediksi untuk setiap data kelas positif atau negatif
Method:
1. Inisialisasi dataset (x) dan label (y)
2. Inisialisasi model SVM dengan kernel linear
3. Latih model pada data fitur (x) dan label (y)
4. Dapatkan vektor bobot (w) dan konstanta bias (b) dari model
5. Hitung nilai prediksi untuk setiap data
   jika nilai prediksi > 0, maka data termasuk ke dalam kelas  positif
   jika nilai prediksi < 0, maka data termasuk ke dalam kelas negatif
```

Pseudocode di atas menggambarkan proses pelatihan model SVM dengan kernel linear dan penggunaannya untuk memprediksi kelas positif atau negatif berdasarkan fitur-fitur yang diberikan.

2.6 Evaluasi

Pada penelitian ini evaluasi yang digunakan yaitu *Confusion Matrix*. Salah satu cara untuk menentukan apakah konsep dalam data mining sudah akurat adalah dengan menggunakan confusion matrix. Nilai akurasi, recall, dan presisi dapat diperoleh melalui evaluasi dengan menggunakan *Confusion Matrix*. Akurasi dalam klasifikasi data mining mengacu pada proporsi data yang diproses dengan akurat, setelah itu hasil uji klasifikasi dieksekusi. Presisi atau bisa juga disebut

TIN: Terapan Informatika Nusantara
Vol 5, No 2, July 2024, page 144-154
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i2.5324

confidence ialah nilai relatif pada kasus yang diprediksi positif. Recall atau sensitivity yaitu kasus positif yang diprediksi secara benar (Ginting et al., 2020) (Robbani, 2021).

Tabel 6. Rumus Evaluasi *Confusion Matrix*

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

Keterangan:
TP (*True Positive*) = jumlah dokumen dari kelas 1 yang tepat diklasifikasikan sebagai kelas 1
TN (*True Negative*) = jumlah dokumen dari kelas 0 yang tepat diklasifikasikan sebagai kelas 0
FP (*False Positive*) = jumlah dokumen dari kelas 0 yang tepat diklasifikasikan sebagai kelas 1
FN (*False Negative*) = jumlah dokumen dari kelas 1 yang tepat diklasifikasikan sebagai kelas 0

Sedangkan data yang dilatih dan diuji di evaluasi menggunakan teknik *Cross Validation* dan *Learning Curve*. Teknik *Cross Validation* atau dapat dikenal juga sebagai estimasi rotasi ialah teknik validasi model untuk mengevaluasi dan memastikan seberapa baik hasil analisis statistik digeneralisir ke dataset yang berbeda (Tuntun et al., 2022). *Learning Curve* merupakan grafik garis yang menggambarkan korelasi antara jumlah data dalam set pelatihan dan akurasi model klasifikasi (Abidin et al., 2021). Selain itu *Learning Curve* dapat digunakan untuk menganalisis data yang mengalami *underfitting* atau *overfitting* (Barinov et al., 2023).

3. HASIL DAN PEMBAHASAN

Tahap awal dimulai dengan dataset yang dieksplorasi, yang terdiri dari 569 data dengan 31 fitur yang digunakan untuk analisis. Dataset tersebut kemudian dikategorikan menjadi dua kelas, B (Benign) dan M (Malignant), yang masing-masing mewakili kondisi jinak dan ganas. Proses eksplorasi data ini bertujuan untuk memahami distribusi dan karakteristik dari masing-masing fitur, sehingga analisis lebih lanjut dapat dilakukan secara akurat dan tepat sasaran.

3.1 Hasil

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 tidak mencukupi dalam dataset di identifikasi dan ditangani dengan cara tertentu sebelum data tersebut dimasukkan ke dalam model atau algoritma (Nikfalazar et al., 2019). Oleh karena itu, mengelola nilai yang hilang menjadi sangat penting dalam data mining dan 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 *pre-processing* data dimana sebuah data dilakukan pengecekan didalam 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 144-154
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i2.5324

Outlier Check ialah tahap *pre-processing* 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)          (535, 31)
```

(a) (b)

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

Handling Outlier ialah tahap *pre-processing* 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 yang digunakan.

Pada *pre-processing* ditambahkan sebuah Seleksi Fitur *Recursive Feature Elimination* (RFE) dan *Forward selection*. *Recursive Feature Elimination* (RFE) ialah salah satu teknik yang digunakan dalam *pre-processing* dan pemodelan data untuk memilih fitur yang paling penting atau relevan dalam dataset data yang dipilih berjumlah 16 Data. Setelah melakukan *Recursive Feature Elimination* (RFE) dilakukan kembali *Forward Selection* yang memilih fitur yang relevan dengan data untuk mempengaruhi hasil klasifikasi dan meningkatkan kinerja metode klasifikasi. Prosedur pemilihan fitur ini dilakukan sebelum proses klasifikasi. dengan total fitur yang sama yaitu 16.

Tabel 7. Hasil Seleksi Fitur

No.	RFE	*Forward Selection*
1	*radius mean*	*radius mean*
2	*texture mean*	*texture mean*
3	*area mean*	*perimeter mean*
4	*concavity mean*	*compactness mean*
5	*concave points mean*	*compactness se*
6	*radius se*	*concavity se*
7	*area se*	*concave points se*
8	*compactness se*	*fractal dimension se*
9	*radius worst*	*radius worst*
10	*texture worst*	*texture worst*
11	*perimeter worst*	*perimeter worst*
12	*area worst*	*area worst*
13	*smoothness worst*	*smoothness worst*
14	*concavity worst*	*compactness worst*
15	*concave points worst*	*concavity worst*
16	*symmetry worst*	*concave points worst*

Setelah melakukan kedua seleksi fitur tersebut didapatkan 10 fitur yang sama dan 6 fitur yang berbeda dari masing-masing teknik seleksi fitur tersebut karena perbedaan dari kedua teknik teresbut. Kemudian setelah digabungkan mendapatkan total 22 fitur dari gabungan fitur yang terpilih.

Tabel 8. *Combined feature* hasil RFE dan *Forward Selection*

Hasil seleksi fitur *RFE* dan *Forward Selection*	
concavity se	*area mean*
concave points mean	*compactness se*
smoothness worst	*symmetry worst*
concavity worst	*radius mean*
area se	*area worst*
concavity mean	*radius se*
perimeter worst	*perimeter mean*
concave points se	*texture mean*
concave points worst	*compactness mean*
compactness worst	*fractal dimension se*
texture worst	*radius worst*

Tahap selanjutnya di *pre-processing* yaitu *Encoding*, *encoding* adalah perubahan tipe data. Untuk mempermudah pembacaan data, hal ini dilakukan dengan mengubah tipe data dari “objek” menjadi “kategori”. Untuk memudahkan pemodelan data, penelitian ini juga melakukan transformasi data, seperti mengubah data kategorikal menjadi data numerik. Adapun teknik yang digunakan yaitu label *encoding* seperti gambar 5 dibawah ini.

TIN: Terapan Informatika Nusantara
Vol 5, No 2, July 2024, page 144-154
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i2.5324

Malignant = 1

Benig = 0

Gambar 5. Hasil *encoding*

Setelah melakukan pelabelan *Encoding* dilakukan kembali *Standarization* dan *Normalization*. *Standarization* dan *Normalization* ialah proses normalisasi data yang membantu mengurangi perbedaan antara atribut. Seperti yang ditunjukkan pada gambar 6 di bawah ini, metode normalisasi data yang digunakan adalah metode skala Min-Max..

	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	...

compactness_worst	concavity_se	smoothness_worst	perimeter_worst	perimeter_mean	area_se	concave points_se	radius_worst	area_mean	texture_worst
0.757130	0.374425	0.572692	0.831980	0.618667	0.845442	0.455119	0.756066	0.494379	0.141525
0.182638	0.129617	0.301026	0.672019	0.697753	0.387998	0.384284	0.739168	0.681753	0.303571
0.467965	0.267038	0.446763	0.632959	0.675045	0.503051	0.590192	0.677643	0.610839	0.360075
0.204706	0.396376	0.397241	0.631099	0.714979	0.505415	0.540579	0.633016	0.665033	0.123934
0.588381	0.255889	0.692253	0.328539	0.303657	0.117579	0.326068	0.326690	0.192332	0.312633

Gambar 6. *Standarization and Normalization*

Selanjutnya dilakukan metode *oversampling* menggunakan SMOTE dikarenakan data kanker payudara yang diperoleh dari sumber UCI *Machine Learning* ada ketidak seimbangan. Setelah diterapkannya metode SMOTE, data menjadi seimbang. Adapun hasilnya dapat ditunjukan pada gambar dibawah ini:

diagnosis	
0	345
1	190

(a)

diagnosis	
1.0	345
0.0	345

(b)

Gambar 7. (a) Sebelum SMOTE dan (b) Setelah SMOTE

3.2 Evaluasi Model

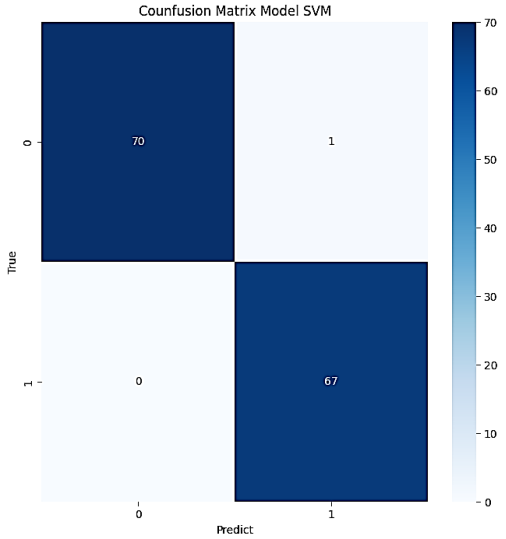
Pada bagian ini, penulis melakukan pengujian terhadap penyakit kanker payudara mulai dari preprocessing hingga evaluasi model. Pembahasan mengenai penelitian dan pengujian evaluasi model dapat dijelaskan pada bagian ini. Klasifikasi model dengan SV/SVC dengan kombinasi hyperparameter terbaik dengan akurasi 97%.

Tabel 9. *Confusion Matrix*

Algoritma	Dataset	Precision	recall	F1-score	Accuracy
Sebelum menggunakan Seleksi Fitur dan SMOTE	Benig Malignant	0.97 1.00	1.00 0.95	0.98 0.98	0.98
Setelah menggunakan RFE	Benig Malignant	0.97 1.00	1.00 0.95	0.98 0.98	0.98
Setelah menggunakan *Forward Selection*	Benig Malignant	1.00 0.98	0.98 1.00	0.99 0.99	0.99
Setelah menggunakan SMOTE	Benig Malignat	1.00 0.99	0.99 1.00	0.99 0.99	0.99
Setelah menggunakan RFE + SMOTE	Benig Malignant	1.00 0.97	0.97 1.00	0.99 0.99	0.99
Setelah menggunakan *Forward Selection* + SMOTE	Benig Malignant	1.00 0.99	0.99 1.00	0.99 0.99	0.99

Tabel 9 menunjukkan bahwa prediksi positif yang benar (Jinak dan Ganas) ditemukan untuk setiap prediksi positif yang dibuat oleh model, dengan akurasi 1,00 untuk "Jinak" dan 0,99 untuk "Ganas". Berdasarkan kategori yang ditetapkan pada Gambar 8 di bawah ini, model ini sangat akurat dalam memprediksi penyakit kanker payudara.

TIN: Terapan Informatika Nusantara
Vol 5, No 2, July 2024, page 144-154
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i2.5324



Gambar 8. Hasil *Confusion Matrix* Prediksi Model

Recall mengukur persentase positif nyata (*Benig* dan *Malignant*) yang dideteksi dengan benar oleh model. Akurasi *recall* pada Tabel 9 sebesar 1.00 untuk "*benig*" dan 0.99 untuk "*malignant*" menunjukkan bahwa model dapat secara efektif mengingat kembali kategori optimasi.

F1-Score adalah rata-rata seimbang dari *recall* dan presisi ketika set data yang dianggap positif jarang ada dalam data yang diidentifikasi dengan benar oleh model. Dengan akurasi F1-Score "*benig*" dan "*malignant*" masing-masing sebesar 0,99 dan 0,99, model ini menunjukkan tingkat kemahiran memori yang tinggi untuk kategori penyakit kanker payudara.

Hasil *clasification report* untuk pengujian prediksi model menunjukkan akurasi pelatihan sebesar 97%. Untuk dataset dengan kategori "jinak", akurasi pelatihan mencapai 100%, sedangkan untuk dataset dengan kategori "ganas", akurasi pelatihan adalah 99%.

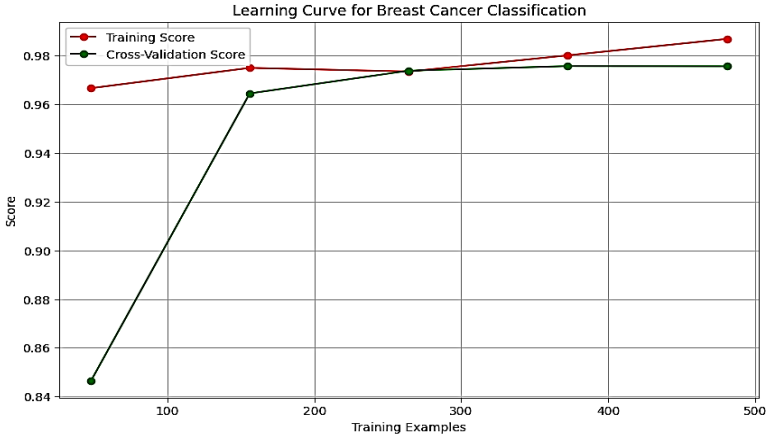
Hasil pengujian *pre-processing* terhadap performa optimasi algoritma *Support Vector Machine* menggunakan Seleksi Fitur RFE dan *Forward Selection* pada klasifikasi kanker payudara dapat dilihat pada Gambar 10. Gambar tersebut menunjukkan bahwa metode algoritma *machine learning* dapat mengingat kategori sebuah dataset yang sudah ditetapkan.

Adapun untuk hasil evaluasi dari model SVM menggunakan *Cross Validation Score* yaitu 97% seperti yang ditunjukan pada gambar 9 di bawah:

```
Cross Validation Scores:  0.9756464011180992
```

Gambar 9. Hasil Evaluasi *Cross Validation*

Selanjutnya model dilakukan proses *Training* menggunakan metode *learning curve* untuk melihat *training score* dan *test score.*



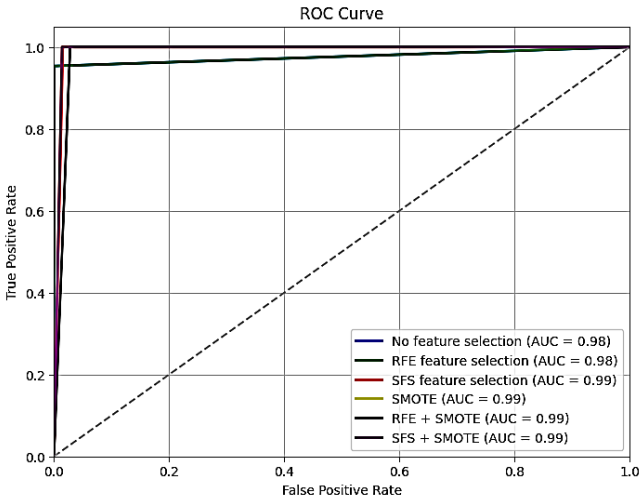
Gambar 10. Hasil *Learning Curve*

Berdasarkan gambar 10, dapat dilihat jika hasil yang ditunjukan untuk model SVM sangat baik tidak terjadi *underfitting* ataupun *overfitting*, Analisis *Learning Curve* menunjukkan bagaimana kinerja model berubah seiring

TIN: Terapan Informatika Nusantara
Vol 5, No 2, July 2024, page 144-154
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i2.5324

dengan meningkatnya data latih. Grafik ini menggambarkan bagaimana kinerja model yang lebih tinggi dicapai dengan lebih banyak penggunaan data pelatihan.

Kurva *Area Under the Curve* (AUC) dan kurva *Receiver Operating Characteristic* (ROC) digunakan dalam penelitian ini sebagai metode evaluasi tambahan untuk menilai performa model secara menyeluruh, meskipun angka akurasi dan presisi hanya memberikan penilaian awal. Temuan evaluasi ditunjukkan pada grafik di bawah.



Gambar 11. Hasil ROC dan AUC

Meskipun nilai AUC dari model tanpa menggunakan teknik oversampling adalah 0.98, yang merupakan angka yang cukup tinggi, nilai AUC 0.99 lebih besar ketika teknik oversampling seperti SMOTE diterapkan. Hal ini menunjukkan bahwa model yang dilatih dengan SMOTE, dengan atau tanpa seleksi fitur, lebih baik dalam membedakan antar kelas.

Perlu dicatat bahwa setiap model memiliki nilai AUC yang sangat tinggi, mulai dari 0,98 hingga 0,99. Hal ini menunjukkan bahwa model-model tersebut dapat membedakan antara kelas positif dan negatif dengan akurasi yang masuk akal. Meskipun demikian, model yang menggunakan pendekatan SMOTE menunjukkan peningkatan dalam nilai AUC ketika dipasangkan dengan RFE atau SFS.

Hasil ini memberikan gambaran yang lebih mendalam dan jelas tentang keandalan model, terutama ketika menangani ketidakseimbangan kelas. Kapasitas model untuk membedakan antar kelas dapat disimpulkan lebih lanjut melalui penggunaan ROC dan AUC. Meskipun demikian, angka akurasi dan presisi dapat memberikan gambaran yang luas tentang kinerja model. Hasil ini dapat mengarahkan pengembangan dan peningkatan model klasifikasi yang lebih efektif untuk menangani masalah yang berhubungan dengan ketidakseimbangan kelas.

4. KESIMPULAN

Algoritma *Support Vector Machine* (SVM) dan teknik pemilihan fitur yang efektif, seperti *Recursive Feature Elimination* (RFE) dan *Forward Selection*, digunakan dalam penelitian ini untuk mengklasifikasikan kasus kanker payudara. Dataset UCI *Machine Learning Repository Breast Cancer Wisconsin Diagnostic*, yang memiliki 569 kasus dengan 32 atribut yang dibagi ke dalam kelompok jinak dan ganas. Setelah langkah *preprocessing* yang melibatkan pemilihan karakteristik dan pembersihan data, penelitian ini menemukan 22 fitur penting. Dengan menggunakan *Training Score* sebesar 0,99 dan *Cross-Validation* Score sebesar 0,97, optimasi algoritma SVM dengan menggunakan seleksi fitur RFE dan Forward Selection memberikan hasil yang baik sesuai dengan hasilnya. Model SVM yang efektif untuk klasifikasi kanker payudara tanpa adanya tanda-tanda *overfitting* atau *underfitting*, teknik RFE dan *Forward Selection* berhasil mengidentifikasi kualitas penting untuk pelatihan model. Hal ini menunjukkan bahwa algoritma SVM yang dikombinasikan dengan teknik seleksi fitur dan SMOTE ini menghasilkan hasil yang sangat baik dalam klasifikasi kanker payudara.

REFERENCES

Abidin, M. I., Notodiputro, K. A., & Sartono, B. (2021). Improving Classification Model Performances using an Active Learning Method to Detect Hate Speech in Twitter. Indonesian Journal of Statistics and Its Applications, 5(1), 26–38. https://doi.org/10.29244/ijsa.v5i1p26-38

Adnyana, I. M. B. (2019). Penerapan Feature Selection untuk Prediksi Lama Studi Mahasiswa. *Jurnal Sistem Dan Informatika (JSI)*, *13*(2), 72–76.

Aji, P. W. S., & Suprianto, S. (2023). Stroke disease prediction using random forest method. Universitas Muhammadiyah Sidoarjo. http://dx.doi.org/10.21070/ups.2643

Barinov, R., Gai, V., Kuznetsov, G., & Golubenko, V. (2023). Automatic evaluation of neural network training results.

TIN: Terapan Informatika Nusantara
Vol 5, No 2, July 2024, page 144-154
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v5i2.5324

Computers, 12(2), 26. https://doi.org/10.3390/computers12020026
Cancer Today. (2021, March). Global Cancer Observatory. https://gco.iarc.fr/today/data/factsheets/populations/360-indonesia-fact-sheets.pdf
Chazar, C., & Erawan, B. (2020). Machine learning diagnosis kanker payudara menggunakan algoritma support vector machine. INFORMASI (Jurnal Informatika Dan Sistem Informasi), 12(1), 67–80. https://doi.org/10.37424/informasi.v12i1.48
Farahdiba, S., Kartini, D., Nugroho, R. A., Herteno, R., & Saragih, T. H. (2023). Backward elimination for feature selection on breast cancer classification using logistic regression and support vector machine algorithms. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 17(4), 429. https://doi.org/10.22146/ijccs.88926
Fitriyani, N., Amalia, D. R., Handayani, H. H., & Masruriyah, A. F. N. (2023). Aplikasi Berbasis Web Berdasarkan Model Klasifikasi Algoritma SVM dan Logistic Regression Terhadap Data Diabetes. REMIK: Riset Dan E-Jurnal Manajemen Informatika Komputer, 7(4), 1762–1771. https://doi.org/10.33395/remik.v7i4.13001
Ginting, V. S., Kusrini, K., & Taufiq, E. (2020). Implementasi Algoritma C4.5 untuk Memprediksi Keterlambatan Pembayaran Sumbangan Pembangunan Pendidikan Sekolah Menggunakan Python. Inspiration: Jurnal Teknologi Informasi Dan Komunikasi, 10(1). https://doi.org/10.35585/inspir.v10i1.2535
Ismafillah, D., Tatang Rohana, & Yana Cahyana. (2023). Analisis algoritma pohon keputusan untuk memprediksi penyakit diabetes menggunakan oversampling smote. INFOTECH : Jurnal Informatika & Teknologi, 4(1), 27–36. https://doi.org/10.37373/infotech.v4i1.452
koirunnisa, Siregar, A. M., & Faisal, S. (2023). Optimized Machine Learning Performance with Feature Selection for Breast Cancer Disease Classification. Jurnal Ilmiah Teknik Elektro Dan Informatika (JITEKI), 9(4), 1131–1143.
Kusnawi, K., & Khrisna Irham Fadhil Pratama. (2023). Komparasi Algoritma Supervised Learning dan Feature Selection pada Klasifikasi Penyakit Gagal Jantung. Indonesian Journal of Computer Science, 12(6). https://doi.org/10.33022/ijcs.v12i6.3487
Masruriyah, A., Novita, H., Sukmawati, C., Ramadhan, A., Arif, S., & Dermawan, B. (2024). Pengukuran Kinerja Model Klasifikasi dengan Data Oversampling pada Algoritma Supervised Learning untuk Penyakit Jantung. Computer Science (CO-SCIENCE), 4(1), 62–70. https://doi.org/10.31294/coscience.v4i1.2389
Maulana, A., Nugroho, A., & Romli, I. (2022). Optimalisasi support vector machine menggunakan particle swarm optimization untuk mendiagnosa penyakit Kanker Payudara. Journal of Practical Computer Science, 1(2), 1–11. https://doi.org/10.37366/jpcs.v1i2.940
Nikfalazar, S., Yeh, C.-H., Bedingfield, S., & Khorshidi, H. A. (2019). Missing data imputation using decision trees and fuzzy clustering with iterative learning. Knowledge and Information Systems, 62(6), 2419–2437. https://doi.org/10.1007/s10115-019-01427-1
Nurjanah, N., Rani, A. N., Masruriyah, A. F. N., & Handayani, H. H. (2023). Implementasi Model Klasifikasi Jenis Kanker Payudara Menggunakan Algoritma SVM dan Logistic Regression berbasis Web. Riset Dan E-Jurnal Manajemen Informatika Komputer.
Pratama, A. R. I., Latipah, S. A., & Sari, B. N. (2022). OPTIMASI KLASIFIKASI CURAH HUJAN MENGGUNAKAN SUPPORT VECTOR MACHINE (SVM) DAN RECURSIVE FEATURE ELIMINATION (RFE). JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika), 7(2), 314–324. https://doi.org/10.29100/jipi.v7i2.2675
Qadrini, L., Hijrah, M., Hikmah, L., & Handayani, H. (2023). The application of the neighborhood cleaning rule in conjunction with random forest, k-fold cross-validation, and grid search for addressing imbalanced datasets. TIN: Terapan Informatika Nusantara, 3(8), 286–293. https://doi.org/10.47065/tin.v3i8.4124
Ramadhan, N. G. (2021). Comparative analysis of ADASYN-SVM and SMOTE-SVM methods on the detection of type 2 diabetes mellitus. Scientific Journal of Informatics, 8(2), 276–282. https://doi.org/10.15294/sji.v8i2.32484
Robbani, A. A. (2021). Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma C4.5. Universitas Buana Perjuangan Karawang.
Setiawan, Y. (2023). Data Mining berbasis Nearest Neighbor dan Seleksi Fitur untuk Deteksi Kanker Payudara. Jurnal Informatika: Jurnal Pengembangan IT, 8(2), 89–96. https://doi.org/10.30591/jpit.v8i2.4994
Tuntun, R., Kusrini, K., & Kusnawi, K. (2022). Analisis Perbandingan Kinerja Algoritma Klasifikasi dengan Menggunakan Metode K-Fold Cross Validation. JURNAL MEDIA INFORMATIKA BUDIDARMA, 6(4), 2111. https://doi.org/10.30865/mib.v6i4.4681
World Health Organization: WHO. (2022, February 3). Cancer. World Health Organization: WHO. https://www.who.int/news-room/fact-sheets/detail/cancer(N.d.).