JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 2, April 2024, Page 889-899
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7457

Komparasi Algoritma Machine Learning dalam Klasifikasi Kanker Payudara

Nurfadlan Afiatuddin[1], M.Teguh Wicaksono[1], Vitto Rezky Akbar[1], Rahmaddeni[1,*], Denok Wulandari[2]

[1]Teknik Informatika, STMIK Amik Riau, Pekanbaru, Indonesia
[2]Teknik Komputer, Institut Az Zuhra, Pekanbaru, Indonesia
Email: [1]nurfadlan.afiatuddin096@gmail.com, [2]muhammadteguuh01@gmail.com, [3]vitorezkiakbar@gmail.com,
[4,*]rahmaddeni@sar.ac.id. [5]denokwulandari18@gmail.com
Email Penulis Korespondensi: rahmaddeni@sar.ac.id

Abstrak−Setiap tahun, jutaan wanita dihadapkan pada masalah kesehatan global yang serius, yaitu kanker payudara. Penelitian ini bertujuan untuk meningkatkan efisiensi dalam mengklasifikasikan kanker payudara dengan menggunakan pembelajaran mesin. Salah satu masalah utama yang dihadapi adalah ketidakseimbangan antara jumlah kasus malignant dan benign dalam dataset. Oleh karena itu, penelitian ini bertujuan untuk membandingkan performa beberapa algoritma pembelajaran mesin dalam mengklasifikasikan kanker payudara, seperti Decision Tree, Naive Bayes, K-Nearest Neighbors, Logistic Regression, dan Random Forest. Proses preprocessing data, pembagian data dengan variasi rasio, dan pengujian berbagai algoritma klasifikasi adalah teknik yang digunakan dalam penelitian ini. Dataset yang digunakan berasal dari dataset diagnosis kanker payudara Wisconsin dari platform Kaggle. Teknik Synthetic Minority Over-Sampling Technique (SMOTE) digunakan untuk mencapai keseimbangan proporsi antara kelas yang tidak seimbang. Setelah melakukan penyetelan hyperparameter, Logistic Regression menunjukkan kinerja terbaik dengan akurasi mencapai 100% dalam beberapa situasi. Penelitian ini menyimpulkan bahwa penggunaan pembelajaran mesin, terutama dengan teknik penanganan ketidakseimbangan kelas, dapat meningkatkan kemampuan dalam mendeteksi kanker payudara lebih awal. Selain itu, penelitian ini juga membantu memahami algoritma terbaik untuk meningkatkan akurasi dalam mengklasifikasikan kanker payudara, memberikan dukungan bagi profesional kesehatan dalam diagnosis dini, serta meningkatkan kualitas perawatan bagi pasien.

Kata Kunci: Kanker Payudara; Pembelajaran Mesin; Ketidakseimbangan Kelas; SMOTE; Akurasi Klasifikasi

Abstract−Every year, millions of women are faced with a serious global health issue: breast cancer. This research aims to improve the efficiency of breast cancer classification using machine learning. One of the main challenges encountered is the imbalance between the number of malignant and benign cases in the dataset. Therefore, this study aims to compare the performance of several machine learning algorithms in classifying breast cancer, such as Decision Tree, Naive Bayes, K-Nearest Neighbors, Logistic Regression, and Random Forest. Preprocessing data, dividing data with various ratios, and testing various classification algorithms are the techniques used in this research. The dataset used originates from the Wisconsin Breast Cancer Diagnosis dataset from the Kaggle platform. The Synthetic Minority Over-Sampling Technique (SMOTE) is used to achieve balance in the proportions of imbalanced classes. After hyperparameter tuning, Logistic Regression showed the best performance with accuracy reaching 100% in several situations. This study concludes that the use of machine learning, especially with techniques for handling class imbalance, can improve the ability to detect breast cancer early. Additionally, this research also helps understand the best algorithms to improve accuracy in classifying breast cancer, providing support for healthcare professionals in early diagnosis, and enhancing the quality of patient care.

Keywords: Breast Cancer; Machine Learning; Class Imbalance; SMOTE; Classification Accuracy

1. PENDAHULUAN

Pada tahun 2020, sebanyak 2,3 juta wanita didiagnonis menderita kanker payudara dan 685.000 orang meninggal di seluruh dunia. Pada akhir tahun itu, sekitar 7,8 juta wanita hidup dengan didiagnosis kanker payudara, menjadikannya jenis kanker paling umun di dunia. Di setiap negara di seluruh dunia, Meskipun kanker payudara terjadi pada wanita usia berapa pun setelah masa pubertas, prevalensi meningkat seiring bertambahnya usia[1]. Menurut studi terbaru IARC mengenai tantangan kanker payudara di seluruh dunia, diperkirakan akan terjadi peningkatan 40% kasus baru (lebih dari 3 juta setiap tahun) dan 50% kematian (lebih dari 1 juta setiap tahun) hingga 2040. Kanker payudara merupakan satu dari delapan kasus kanker di seluruh dunia pada tahun 2020, menunjukan urgensi penanganan global, terutama di negara-negara yang mengalami transisi ekonomi[2].

Berdasarkan laporan kanker di Indonesia pada tahun 2020, kanker payudara menjadi yang paling umum. Data Globocan tahun 2020, mencatat jumlah kasus baru kanker payudara di indonesia mencapai 68.858 kasus, atau 16,6% dari total 396.914 kasus baru kanker[3]. Pemerintah Indonesia, meiskipun memprioritaskan pencegahan peningkatan kasus kanker payudara, juga melakukan upaya melawan berbagai jenis kanker sesuai dengan Rencana Aksi Nasional Kanker 2022-2022.[4]. Kanker payudara dapat terdeteksi pada tahap awal, dan terbagi menjadi dua jenis utama, yaitu tumor ganas dan jinak. Identifikasi dini jenis tumor sangat penting untuk pengobatan yang tepat bagi pasien penderita kanker payudara[5]. Namun, deteksi dini kanker dapat bermanfaat bagi pasien karena memungkinkan mereka untuk berkonsultasi dengan dokter pada waktu yang tepat[6]. Dua cara terpenting untuk mencegah kematian akibat kanker payudara adalah menemukan kanker sejak dini dan mendapatkan pengobatan kanker terbaru. Obat kanker payudara lebih efektif jika ditemukan pada usia dini, ketika kanker belum menyebar. Metode yang paling andal untuk mendeteksi kanker payudara sejak dini adalah melakukan tes skrining secara teratur. American Cancer Society memiliki pedoman skrining untuk wanita dengan risiko kanker payudara rata-rata dan tinggi[7]. Tujuan dari melakukan pemeriksaan payudara sendiri adalah untuk

Submitted: **02/02/2024**; Accepted: **30/04/2024**; Published: **30/04/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 2, April 2024, Page 889-899
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7457

mengetahui apakah bentuk dan rasanya normal. Memahami bagaimana kondisi payudara biasanya membantu kita mendeteksi perubahan atau kelainan, seperti benjolan baru atau perubahan pada kulit, jika kita menemukan perubahan seperti itu saat memeriksa payudara. Sangat penting untuk segera menghubungi dokter untuk pemeriksaan tambahan[8]. Dalam sektor layanan kesehatan, penggunaan pembelajaran mesin memiliki dampak besar dalam meningkatkan kecepatan dan akurasi pekerjaan dokter. Ini memungkinkan pengolahan data layanan kesehatan yang optimal, mendeteksi indikator awal epidemi atau pandemi, serta membuka banyak kemungkinan baru di bidang perawatan kesehatan[9]. Ketidakseimbangan metode manual mendorong peneliti untuk mengembangkan suatu metode yang tidak bergantung pada manusia. Penggunaaan komputer sebagai alat bantu dalam menganalisis data. Memberikan konstribusi besar dalam membantu para profesional kesehatan membuat diagnosis penyakit pasien[10].

Untuk mengklasifikasikan kanker payudara, penelitian ini menyelidiki berbagai algoritma pembelajaran mesin, seperti Decision Tree, Naive Bayes, K-NN, Logistic Regression, dan Random Forest. Data dibagi ke dalam kategori rasio 60:40, 70:30, 80:20, dan 90:10 antara data latihan dan data uji. Perbandingan dilakukan untuk mengevaluasi kinerja masing-masing algoritma dalam berbagai skenario pembagian data. Hal ini penting karena dapat memberikan wawasan tentang algoritma mana yang lebih efektif dalam mengklasifikasikan kanker payudara, tergantung pada proporsi data latih dan uji yang digunakan. Dengan membandingkan berbagai skenario, peneliti dapat menentukan rasio pembagian data yang optimal dan algoritma yang paling sesuai untuk mencapai kinerja klasifikasi yang optimal. Jika perbandingan tidak dilakukan, peneliti tidak akan dapat mengetahui algoritma mana yang memberikan hasil terbaik dalam mengklasifikasikan kanker payudara. Hal ini dapat mengakibatkan penggunaan algoritma yang kurang efektif atau tidak optimal, yang dapat menghasilkan hasil klasifikasi yang tidak akurat atau tidak dapat diandalkan. Oleh karena itu, untuk memastikan metode yang optimal digunakan dalam penelitian ini, perbandingan yang cermat antara skenario pembagian data dan berbagai algoritma sangat penting.

Dalam beberapa tahun terakhir, banyak peneliti mengembangkan berbagai teknik pembelajaran mesin untuk memprediksi kanker payudara lebih akurat[11]. Teknologi pembelajaran mesin dapat membantu dalam diagnosis kanker payudara melalui komputer, memungkinkan proses diagnosis yang lebih cepat dan akurat, tanpa menggantikan peran dokter atau ahli dengan komputer, tetapi membantu mereka memahami kasus dengan baik dan menemukan kanker lebih awal[12]. Saat ini, alat pembelajaran mesin dapat digunakan untuk mendeteksi dan memprediksi kanker[13]. Metode klasifikasi pembelajaran mesin sangat membantu dalam deteksi kanker payudara[14]. Untuk menghasilkan model klasifikasi terbaik, diperlukan algoritma yang efektif[15]. Banyak model pembelajaran mesin telah dikembangkan untuk kanker payudara dengan berbagai jenis data. Validasi eksternal model yang berhasil menunjukan kemampuan generalisasinya[16]. Dengan menggunakan alat pembelajaran mesin yang mempetimbangkan hyperparametik dengan tepat, tumor dapat diindentifikasi lebih efektif[17]. Mekipun demikian, ketidakseimbangan dalam kumpulan data dapat menjadi hambatan. [18]. Upaya penelitian dilakukan untuk mengatasi masalah ini, salah satunya dengan menerapkan teknik yang Synthetic Minority Over-Sampling Technique (SMOTE), yang membantu mengatasi ketidakseimbangan dalam dataset[19].

Dalam mengawali telaah literatur ini, kami memulai perjalanannya dengan merinci fondasi penelitian sebelumnya yang membentuk dasar pengembangan model pembelajaran mesin untuk kanker payudara. Penelitian ini diawali dengan menerapkan metode Support Vector Machine dan Backward Elimination pada Dataset Kanker payudara Coimbra, dengan tujuan meningkatkan klasifikasi kanker payudara. Hasil awal menunjukan tingkat akurasi sebesar 65,22% dan AUC 0,700 (klasifikasi yang wajar). Namun, dengan penerapan Backward Elimination, terjadi peningkatan yang signifikan, membawa akurasi mencapai 95,65% dan AUC 1,000 (klasifikasi yang luar biasa). Dengan pencapaian ini, pendekatan ini mencatat langkah yang sangat berarti dalam meningkatkan efektivitas model klasifikasi kanker payudara[20]. Penelitian ini ingin mengetahui seberapa baik metode klasifikasi Support Vector Machine dan Multilayer Percepton dalam mengidentifikasi kanker payudara. Hasil menunjukan bahwa Multilayer Percepton dengan fungsi aktivasi Logistic dan optimasi Adam (Adaptive Moment Estimation) memiliki akurasi, presisi dan recall terbaik dengan nilai 97,7%[21]. Penelitian ini fokus pada klasifkasi dataset kanker payudara Coimbra dengan menggunakan metode K-NN dan forward selection. Tujuan nya adalah membandingkan kinerja K-NN dengan forward selection terhadap algoritma pembelajaran mesin lainnnya dan juga meningkatkan kemampuan K-NN dalam mengklasifikasikan dataset tersebut. Hasilnya menunjukan bahwa penggunaan metode seleksi fitur dapat membantu meningkatkan kinerja K-NN. Kombinasi KNN+FS dan K-NN+OS memiliki akurasi tertinggi (91,43%) dengan proporsi data 70/30% dan nilai K=1. Sebaliknya, K-NN+BE mengalami overfitting[22]. Penelitian ini bertujuan meningkatkan akurasi metode pembelajaran mesin dengan cara memproses data sebelumnya. Proses ini melibatkan penggantian nilai yang hilang, pengubahan bentuk data, penghalusan data yang berisik, pemilihan fitur, validasi data, dan penanganan ketidakseimbangan kelas untuk memprediksi kanker payudara. Hasil penelitian menunjukkan bahwa menggunakan metode Jaringan Saraf Tiruan—Z-Score—Seleksi Maju pada Dataset Kanker Payudara Wisconsin memberikan tingkat akurasi tertinggi sebesar 98,24%.[23]. Dalam penelitian ini, lima algoritma pembelajaran mesin digunakan untuk memprediksi kanker payudara pada dataset Diagnostik Kanker Payudara Wisconsin, di mana Support Vector Machine (SVM) memiliki akurasi tertinggi sebesar 97,2%.[24].

Tujuan penelitian ini adalah untuk membandingkan berbagai algoritma pembelajaran mesin yang menggunakan metode pengajaran kelompok untuk mengklasifikasikan kanker payudara. Ini penting karena masih

Submitted: **02/02/2024**; Accepted: **30/04/2024**; Published: **30/04/2024**

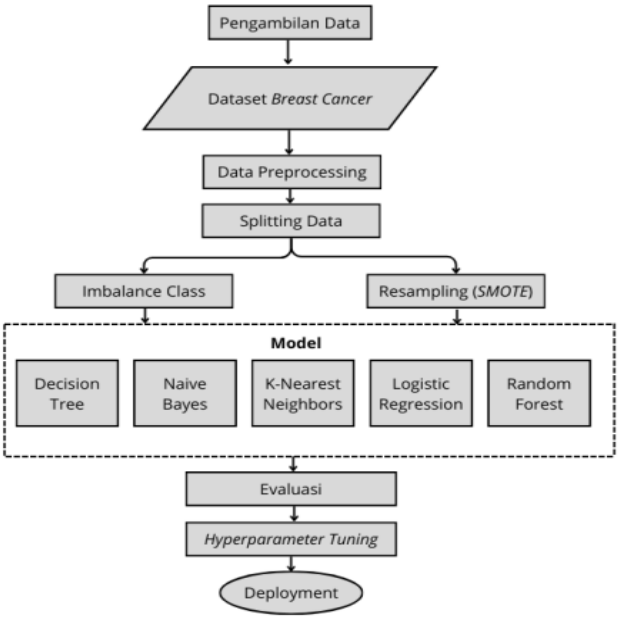
JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 2, April 2024, Page 889-899
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7457

sedikit penelitian yang secara khusus membandingkan algoritma ini dalam konteks pengajaran kelompok. Selain itu, pembelajaran kelompok belum banyak menyelidiki penanganan ketidakseimbangan data pada kumpulan data kanker payudara. Untuk menangani ketidakseimbangan data, kami menggunakan metode SMOTE. Kami ingin melihat bagaimana ukuran set data latihan dan uji mempengaruhi kinerja klasifikasi dengan membagi data menjadi beberapa rasio berbeda: 60%–40%, 70%–30%, 80%–20%, dan 90%–10%. Selain itu, karena algoritma Logistic Regression sangat konsisten dalam menghasilkan akurasi yang tinggi, kami melakukan penyesuaian hyperparameter khusus padanya. Kami menilai hasilnya menggunakan skor F1, akurasi, presisi, dan recall. Kami berharap bahwa penelitian

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Melalui penelitian tentang perbandingan berbagai algoritma pembelajaran mesin untuk mengklasifikasikan kanker payudara, kita akan dibawa melalui serangkaian langkah yang penting dan menarik, seperti yang tergambar dalam Gambar 1. Langkah-langkah ini akan membantu kita menemukan solusi yang lebih baik dalam diagnosis dan penanganan penyakit serius ini.



Gambar 1. Tahapan Penelitian

Pada Gambar 1, penelitian ini dimulai dengan meninjau literatur untuk memperoleh pemahaman tentang klasifikasi kanker payudara. Selanjutnya, data untuk penelitian diambil dari platform Kaggle, sebuah informasi dan dataset terkemuka. Setelah mengambil dataset kanker payudara dari Kaggle, langkah berikutnya melibatkan preprocessing data untuk mengatasi duplikasi, missing value, outlier dan standardisasi data. Setelah mendapatkan data yang berkualitas dan siap untuk model, maka dilakukan splitting data. Dalam penelitian ini, dilakukan spilitting data ke dalam beberapa kategori rasio yang berbeda. Proses spiliting data melibatkan data latih dan data uji, dengan komposisi masing-masing sebesar 60:40, 70:30, 80:20 dan 90:10.

Langkah berikutnya melibatkan pembuatan model dengan algoritma pembelajaran mesin menggunakan Decision Tree, Naive Bayes, K-NN, Logistic Regression, dan Random Forest dalam mengklasifikasikan kanker payudara dengan mempertimbangkan ketidakseimbangan kelas pada target atau label. Kelas M (Malignant) memiliki jumlah yang lebih sedikit dibandingkan dengan kelas B (Benign). Untuk mengatasi ketidakseimbangan kelas selanjutnya menggunakan teknik SMOTE (Synthetic Minority Over-sampling Technique) agar mendapatkan hasil yang maksimal. Kemudian penelitian dilanjutkan dengan tahap optimisasi Hyperparameter Tuning pada Logistic Regression untuk meningkatkan model.

Langkah terakhir dalam siklus pengembangan perangkat lunak adalah deployment. Deployment adalah proses penting yang melibatkan persiapan, konfigurasi, pengujian, dan peluncuran aplikasi atau sistem agar siap digunakan oleh pengguna.

2.2 Pengumpulan Data

Penelitian ini menggunakan Dataset “Breast Cancer Wisconsin Diagnostic” yang berasal dari Kaggle. Data set ini berisi 569 baris 32 kolom. Setiap kolom menunjukan fitur yang digunakan untuk menjelaskan sampel tersebut.

Submitted: **02/02/2024**; Accepted: **30/04/2024**; Published: **30/04/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 2, April 2024, Page 889-899
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7457

Unnamed: 0	x.radius_mean	x.texture_mean	x.perimeter_mean	x.area_mean	x.smoothness_mean	x.compactness_mean	x.concavity_mean	x.concave_pts_mean
1	13.540	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.047810
2	13.080	15.71	85.63	520.0	0.10750	0.12700	0.04568	0.031100
3	9.504	12.44	60.34	273.9	0.10240	0.06492	0.02956	0.020760
4	13.030	18.42	82.61	523.8	0.08983	0.03766	0.02562	0.029230
5	8.196	16.84	51.71	201.9	0.08600	0.05943	0.01588	0.005917

Gambar 2. Dataset Breast Cancer Wisconsin Diagnostic

Gambar ini menunjukan lima baris pertama dari Dataset “Breast Cancer Wisconsin Diagnostic” Dataset ini bertipe Float dan tidak ada missing value.

a) Atribut yang mungkin berfungsi sebagai indeks baris dan tidak memiliki label disebut "Unnamed".
b) Radius rata-rata sel yang ditemukan dalam gambar disebut x.radius_mean.
c) Nilai rata-rata tekstur sel-sel yang terdeteksi dalam gambar disebut x.texture_mean.
d) X.perimeter_mean adalah keliling rata-rata dari sel-sel yang terdeteksi dalam gambar.
e) X.area_mean adalah rata-rata luas sel yang ditemukan dalam gambar.
f) X.smoothness_mean menunjukkan tingkat ketidakteraturan (kehalusan) sel yang terlihat dalam gambar.
g) X.compactness_mean adalah nilai rata-rata kepadatan sel yang terlihat dalam gambar.
h) X.concavity_mean adalah nilai rata-rata konkavitasi sel yang terlihat dalam gambar.
i) X.concave_pts_mean adalah nilai rata-rata dari jumlah titik konkaf sel-sel yang terdeteksi dalam gambar.

2.3 Data Preprocessing

Data preprocessing merupakan teknik untuk menggali informasi dari data sebelum pembuatan model. Adapaun tahapan-tahapan yang dilakukan pada saat preprocessing adalah:

a) Data Cleaning: Merupakan proses menemukan, mengoreksi, dan menghapus kesalahan, ketidakakuratan, atau anomali dalam kumpulan data. Tujuan dari data cleaning adalah memastikan bahwa data yang digunakan untuk analisis atau pemrosesan lebih lanjut adalah akurat, lengkap, konsisten dan relevan.
b) Features Selection: Proses modifikasi fitur sebelum pembuatan model. Teknik yang digunakan pada penelitian ini menggunakan korelasi spearman.
c) Standardisasi Data: Proses mentrasformasi nilai-nilai dalam dataset sehingga memiliki rata-rata 0 dan deviasi standar 1. Tujuannya adalah membuat distribusi data lebih mudah diinterprestasikan dan diproses oleh algoritma pembelajaran mesin.

2.4 Spiltting Data

Data yang digunakan dalam membuat model dibagi menjadi dua, yaitu data training dan data testing. Pembagian dilakukan dalam rasio:

a) Data Training 60% - Data Testing 40%.
b) Data Training 70% - Data Testing 30%.
c) Data Training 80% - Data Testing 20%.
d) Data Training 90% - Data Testing 10%.

2.5 Algoritma Klasifikasi

Algoritma klasifikasi dalam pembelajaran mesin merupakan metode yang mempelajari pola pada data pelatihan dan menggunakannya untuk memprediksi kelas atau label dari data yang terlihat. Tujuannya adalah untuk memisahkan atau mengelompokan data ke dalam kelas yang berbeda berdasarkan fitur atau atribut yang dimilikinya.Contoh algortima klasfikasi meliputi:

a) Decision Tree: Algoritma klasifikasi yang menggunakan keputusan berdasarkan struktur yang dimodelkan seperti pohon, dikenal sebagai keputusan pohon.

$$Gain(S,A) = Error(S) \sum_{i=1}^{n} \frac{|Si|}{|S|} Error(Si) \quad (1)$$

Keterangan:
$Gain(S,A)$ adalah gain informasi dari atribut A pada himpunan data S, $Error(S)$ adalah nilai kesalahan (misclassification error) dari himpunan data S, $|S|$ adalah jumlah total instansi dalam himpunan data S, $|Si|$ adalah jumlah instansi dalam subset Si yang dihasilkan setelah membagi S berdasarkan atribut, A. $Error(Si)$adalah nilai kesalahan dari subset Si.

b) Naive Bayes: Algoritma klasifikasi yang sering digunakan untuk masalah klasifikasi dan merupakan metode klasfikasi pada teorema Bayes.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \quad (2)$$

Keterangan:

Submitted: **02/02/2024**; Accepted: **30/04/2024**; Published: **30/04/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 2, April 2024, Page 889-899
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7457

$P(A|B)$ adalah probabilitas kondisional dari A bila diketahui B, $P(B|A)$ adalah probabilitas kondisional dari B bila diketahui A dan $P(A)$ dan $P(B)$ adalah probabilitas dari A dan B masing-masing.

c) K-Nearest Neighbors: Algoritma klasifikasi K-Nearest Neighbors (KNN) berbasis jarak, Di mana klasifikasi data baru didasarkan pada kelas mayoritas berdasarkan data latih yang memiliki jarak paling dekat dengan data tersebut. Dalam KNN, jarak euclidean adalah salah satu teknik pengukuran jarak yang paling umum digunakan.

$$Euclidean\ Dist = \sqrt{\sum_{i=1}^{n}(X_1 - X_2)^2} \quad (3)$$

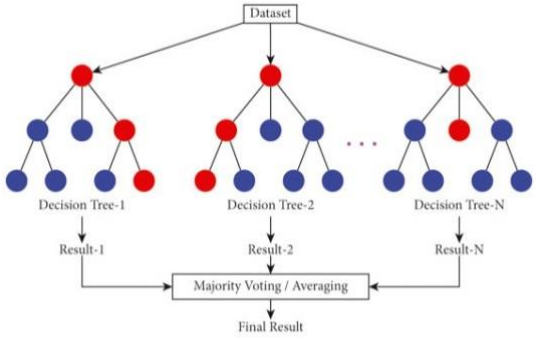
Keterangan:
n adalah jumlah atribut pada setiap titik data B, X_1 adalah nilai atribut ke- i dari data sampel (data latih) dan X_2 adalah nilai atribut ke- i dari data uji (objek baru yang akan diklasifikasikan).

d) Logistic Regression: Logistic regression adalah algoritma klasifikasi yang digunakan untuk memprediksi probabilitas hubungan antara variable dependen dan independen.

$$\boldsymbol{in}\left(\frac{P}{1-p}\right) = \boldsymbol{W^T X}, \boldsymbol{p} = \frac{1}{1+e^{-wTx}} \quad (4)$$

Keterangan:
P adalah probabilitas kejadian yang bernilai 1, X adalah vektor kolom dari variabel independen, W adalah vektor kolom dari koefisien regresi dan T menunjukkan operasi transpose.

e) Random Forest: Random Forest, diperkenalkan oleh Leo Breiman, menggunakan pohon keputusan secara acak. Dalam Random Forest, sampel pelatihan diambil sebanyak pohon yang diinginkan dengan metode random sampling with replacement (SRS WR), disebut bagging. Setiap sampel menghasilkan satu pohon keputusan. Pada pengujian, satu objek diuji oleh semua pohon, dan keputusan akhir didasarkan pada mayoritas pilihan pohon-pohon. Perhitungan error dilakukan pada objek yang tidak digunakan dalam pembuatan pohon, disebut Out-of-bag (OOB) error.



Gambar 3. Cara Kerja Random Forest

Gambar 3 menunjukkan Random Forest memilih sejumlah sampel secara acak dari dataset yang disediakan, dan kemudian membuat serangkaian pohon keputusan untuk setiap sampel tersebut. Prediksi dibuat melalui kerja keras pohon-pohon tersebut. Setelah itu, melalui proses voting yang teliti, hasil prediksi dari masing-masing pohon digabungkan. Modus prediksi digunakan untuk klasifikasi, dan rata-rata prediksi digunakan untuk regresi. Akhirnya, algoritma dengan bijak memilih prediksi yang paling konsisten dari hasil voting.

2.6 Imbalance Class

Ketika jumlah sampel atau observasi yang termasuk dalam setiap kelas atau label data tidak seimbang secara signifikan, itu disebut ketidak seimbangan kelas. Ini terjadi ketika jumlah sampel dari satu kelas jauh lebih besar atau jauh lebih sedikit daripada jumlah sampel dari kelas lainnya dalam konteks klasifikasi. Jumlah label kelas M (Malignant) lebih sedikit dari pada label kelas B (Benign) dalam penelitian ini.

2.7 SMOTE (Synthetic Minority Over-sampling Technique)

Ada banyak cara untuk mengatasi ketidakseimbangan kelas (Imbalace Class). Salah satu metode yang digunakan adalah SMOTE, metode yang menyeimbangkan distribusi data sampel pada kelas minoritas dengan memilih sampel hingga sebanding dengan sampel kelas mayoritas. Overfitting dapat terjadi ketika metode SMOTE digunakan karena data kelas minoritas diduplikasi sehingga data latih yang sama ada. Proses SMOTE dimulai dengan menghitung jarak antara data minoritas, menemukan nilai presentase SMOTE, menemukan jumlah k terdekat, dan terakhir, menggunakan persamaan berikut untuk membuat data sintesis:

$$X_{syn} = X_i + (X_{knn} - X_i) \times \delta \quad (5)$$

Di mana $\boldsymbol{X_{syn}}$ adalah hasil dari data sintesis, $\boldsymbol{X_i}$ adalah data yang direplikasi, dan $\boldsymbol{X_{knn}}$ adalah data yang paling dekat dengan data yang direplikasi, dan $\boldsymbol{\delta}$ adalah nilai acak antara 0 dan 1.

Submitted: **02/02/2024**; Accepted: **30/04/2024**; Published: **30/04/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 2, April 2024, Page 889-899
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7457

2.8 Evaluasi Hasil dan Klasifikasi

Pada penelitian ini dalam mengevaluasi model klasifikasi yaitu menggunakan confusion matrix. Confusion matrix adalah matriks yang menggambarkan kinerja dari suatu model klasifikasi pada rangkaian data uji yang menyatakan data uji yang dinyatakan terklasifikasi dengan benar dan data uji yang dinyatakan terklasifikasi salah ditampilkan dari confusion matrix untuk dua kelas.

Tabel 1. Confusion Matriks

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

Dalam konteks evaluasi model klasifikasi, confusion matriks, memberikan gambaran yang jelas tentang bagaimana model berfungsi dalam memprediksi berbagai kelas. Hasil prediksi model digambarkan dengan empat istilah:

a) True Positive (TP): Ini menggambarkan jumlah data yang sebenarnya positif (pasien menderita kanker) dan berhasil diprediksi sebagai positif oleh model.
b) False Negative (FN): Ini menggambarkan jumlah data yang sebenarnya positif (pasien menderita kanker) tetapi salah diprediksi sebagai negatif oleh model.
c) False Positive (FP): Ini menggambarkan jumlah data yang sebenarnya negatif (pasien tidak menderita kanker) tetapi salah diprediksi sebagai positif oleh model.
d) True Negative (TN): Ini menggambarkan jumlah data yang sebenarnya negatif (pasien tidak menderita kanker) dan berhasil diprediksi sebagai negatif oleh model.

Kita dapat menghitung akurasi, presisi, recall, dan nilai F1 dengan menggunakan confusion matrix untuk mengetahui seberapa baik model bekerja. Akurasi menunjukkan seberapa tepat model dalam mengklasifikasikan dengan benar.

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \quad (6)$$

Perbandingan antara jumlah dokumen yang dianggap relevan dan total dokumen yang ditemukan dalam suatu sistem klasifikasi dikenal sebagai presesi.

$$Presisi = \frac{TP}{TP+FP} \quad (7)$$

Recall adalah perbandingan antara jumlah dokumen yang ditemukan kembali oleh sistem klasifikasi dan jumlah dokumen yang relevan secara keseluruhan.

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

Selanjutnya diberikan F1-score, yang merupakan rata-rata dari presisi dan recall.

$$F1 - Score = \frac{Recall\ x\ Presisi}{Recall+Presisi} x2 \quad (9)$$

2.9 Hyper-parameter Tuning

Tahap ini dimaksudkan untuk memaksimalkan parameter model pembelajaran mesin. Untuk meningkatkan performa model, penelitian ini akan melakukan penyesuaian hyperparameter setelah menemukan model yang memiliki kinerja terbaik di antara model yang menangani ketidakseimbangan kelas dan model yang menangani kelas seimbang (SMOTE).

3. HASIL DAN PEMBAHASAN

Hasil penelitian menunjukkan bahwa menggunakan pembelajaran mesin, terutama dengan teknik penanganan ketidakseimbangan kelas seperti SMOTE, dapat secara signifikan meningkatkan efisiensi dalam mengklasifikasikan kanker payudara. Dari berbagai algoritma yang diujikan, Logistic Regression menunjukkan kinerja terbaik dengan akurasi hingga 100% dalam beberapa situasi setelah disesuaikan hyperparameter. Temuan ini mengindikasikan bahwa metode tersebut mungkin menjadi solusi efektif dalam mengatasi ketidakseimbangan antara jumlah kasus kanker payudara ganas dan jinak dalam dataset. Hasil penelitian ini memberikan kontribusi penting dalam pemahaman algoritma terbaik untuk meningkatkan akurasi dalam diagnosis kanker payudara. Ini dapat memberikan dukungan yang berarti bagi para profesional kesehatan dalam mendeteksi kanker lebih awal dan memberikan perawatan yang lebih baik bagi pasien. Dengan menggunakan teknik pembelajaran mesin dan strategi penanganan ketidakseimbangan kelas, penelitian ini menunjukkan potensi besar untuk meningkatkan kemampuan sistem dalam mendeteksi kanker payudara dengan lebih efisien dan efektif. Ini menyoroti pentingnya

Submitted: **02/02/2024**; Accepted: **30/04/2024**; Published: **30/04/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 2, April 2024, Page 889-899
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7457

integrasi teknologi dalam bidang kesehatan untuk meningkatkan diagnosis dini dan perawatan menyeluruh bagi pasien.

Tabel 2. Dataset Breast Cancer Wisconsin Diagnostic menggunakan SMOTE

Label	Imbalance	Resampling (SMOTE)
M (Malignant)	212	357
B (Benign)	357	357

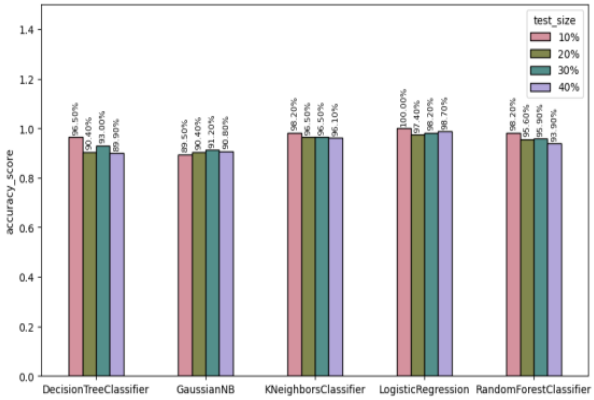
Tabel 2 menyoroti upaya untuk menyeimbangkan jumlah sampel antara kelas Malignant dan Benign menggunakan teknik SMOTE (Synthetic Minority Over-sampling Technique). awalnya, ada ketidakseimbangan pada label data, dengan 212 sampel untuk kelas Malignant dan 357 sampel untuk kelas Benign. Ketidakseimbangan ini dapat menyebabkan model klasifikasi cenderung memihak kelas mayoritas dan mengabaikan kelas minoritas, yang dapat menyebabkan hasil yang buruk.

Untuk mengatasi ketidakseimbangan tersebut, teknik SMOTE—teknik sampel lebih banyak minoritas sintetis—membuat sampel sintetis baru untuk kelas minoritas berdasarkan pola yang ada di sekitar titik data minoritas yang sudah ada. Dengan melakukan ini, jumlah sampel kedua kelas menjadi seimbang, dengan 357 sampel untuk kelas Malignant dan 357 sampel untuk kelas Benign.

Tabel 3. Hasil pengujian dengan Imbalance Class

Algoritma	Splitting Data		Class Label	Accuracy	Precision	Recall	F1 Score
	Training	Testing					
Decision Tree	60%	40%	Imbalance	89.9%	86.6%	85.5%	86.1%
	70%	30%	Imbalance	93.0%	89.2%	92.1%	90.6%
	80%	20%	Imbalance	90.4%	88.6%	81.6%	84.9%
	90%	10%	Imbalance	96.5%	89.5%	100%	94.4%
Naïve Bayes	60%	40%	Imbalance	90.8%	86.0%	89.2%	87.6%
	70%	30%	Imbalance	91.2%	86.4%	90.5%	88.4%
	80%	20%	Imbalance	90.4%	82.9%	89.5%	86.1%
	90%	10%	Imbalance	89.5%	76.2%	94.1%	84.2%
K-Nearest Neighbors	60%	40%	Imbalance	96.1%	95.1%	94.0%	94.5%
	70%	30%	Imbalance	96.5%	96.7%	93.7%	95.2%
	80%	20%	Imbalance	96.5%	97.2%	92.1%	94.6%
	90%	10%	Imbalance	98.2%	94.4%	100%	97.1%
Logistic Regression	60%	40%	Imbalance	98.7%	100%	96.4%	98.2%
	70%	30%	Imbalance	98.2%	100%	95.2%	97.6%
	80%	20%	Imbalance	97.4%	97.3%	94.7%	96.0%
	90%	10%	Imbalance	100%	100%	100%	100%
Random Forest	60%	40%	Imbalance	93.9%	91.6%	91.6%	91.6%
	70%	30%	Imbalance	95.9%	93.8%	95.2%	94.5%
	80%	20%	Imbalance	95.6%	94.6%	92.1%	93.3%
	90%	10%	Imbalance	98.2%%	94.4%	100%	97.1%

Hasil pengujian dengan Klasifikasi Imbalance Class ditampilkan dalam Tabel 3. Pengujian ini membandingkan kinerja model algoritma Decision Tree, Naive Bayes, K-Nearest Neighbors, Logistic Regression, dan Random Forest. Pengujian dilakukan dengan komposisi data pelatihan dan pengujian yang berbeda, masing-masing 60:40, 70:30, 80:20, dan 90:10. Tabel ini juga menampilkan metrik evaluasi kinerja model untuk setiap algoritma klasifikasi yang diuji, termasuk akurasi, presisi, recall, dan skor F1.



Gambar 4. Akurasi dengan Imbalance Class

Submitted: **02/02/2024**; Accepted: **30/04/2024**; Published: **30/04/2024**

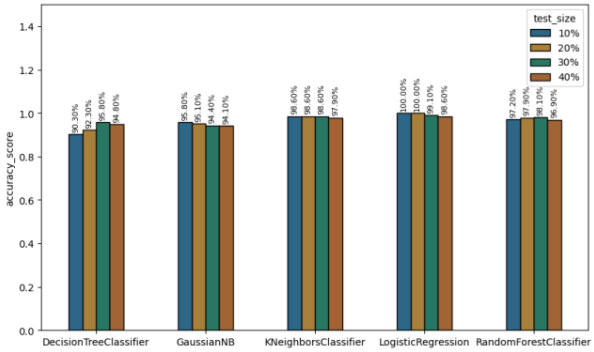
JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 2, April 2024, Page 889-899
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7457

Gambar 4 menunjukkan akurasi dengan Imbalance Class. Hasil pengujian menunjukkan bahwa algoritma Logistic Regression menunjukkan kinerja yang baik dengan komposisi data pelatihan dan pengujian 90:10, dengan skor akurasi, presisi, recall, dan skor F1 masing-masing 100%. Hasil pengujian dengan Imbalance Class dapat dilihat pada tabel 3.

Tabel 4. Hasil pengujian dengan SMOTE

Algoritma	Splitting Data		Class Label	Accuracy	Precision	Recall	F1 Score
	Training	Testing					
Decision Tree	60%	40%	SMOTE	94.8%	91.9%	97.9%	94.8%
	70%	30%	SMOTE	95.8%	92.8%	99.0%	95.8%
	80%	20%	SMOTE	92.3%	90.3%	94.2%	92.2%
	90%	10%	SMOTE	90.3%	85.4%	97.2%	90.9%
Naïve Bayes	60%	40%	SMOTE	94.1%	93.6%	94.3%	94.0%
	70%	30%	SMOTE	94.4%	94.2%	94.2%	94.2%
	80%	20%	SMOTE	95.1%	97.0%	92.8%	94.8%
	90%	10%	SMOTE	95.8%	97.1%	94.4%	95.8%
K-Nearest Neighbors	60%	40%	SMOTE	97.9%	97.2%	98.6%	97.9%
	70%	30%	SMOTE	98.6%	98.1%	99.0%	98.6%
	80%	20%	SMOTE	98.6%	98.6%	98.6%	98.6%
	90%	10%	SMOTE	98.6%	97.3%	100%	98.6%
Logistic Regression	60%	40%	SMOTE	98.6%	98.6%	98.6%	98.6%
	70%	30%	SMOTE	99.1%	98.1%	100%	99.0%
	80%	20%	SMOTE	100%	100%	100%	100%
	90%	10%	SMOTE	100%	100%	100%	100%
Random Forest	60%	40%	SMOTE	96.9%	95.8%	97.9%	96.8%
	70%	30%	SMOTE	98.1%	97.2%	99.0%	98.1%
	80%	20%	SMOTE	97.9%	97.1%	98.6%	97.8%
	90%	10%	SMOTE	97.2%	97.2%	97.2%	97.2%

Tabel 4 menunjukkan hasil pengujian yang dilakukan setelah menggunakan teknik SMOTE. Pengujian ini menunjukkan perbandingan kinerja model antara algoritma Decision Tree, Naive Bayes, K-Nearest Neighbors, Logistic Regression, dan Random Forest. Komposisi yang digunakan untuk data pelatihan dan pengujian adalah 60:40, 70:30, 80:20, dan 90:10. Tabel ini tidak hanya menunjukkan skor F1, akurasi, presisi, dan recall untuk masing-masing algoritma, tetapi juga menunjukkan bagaimana teknik SMOTE mempengaruhi kinerja model dalam menangani masalah ketidakseimbangan kelas.



Gambar 5. Akurasi dengan SMOTE

Gambar 5 menunjukkan bahwa setelah teknik SMOTE diterapkan pada data, hasil klasifikasi kanker payudara dengan Logistic Regression meningkat sekitar 1-2%, sementara Algoritma Decision Tree memiliki hasil terendah.

Tabel 5. Perbandingan Logistic Regression dengan tuning hyper-parameter

Algoritma	Class Label	Splitting Data		Parameter	Accuracy	Precision	Recall	F1 Score
		Training	Testing					
Logistic Regression	SMOTE	60%	40%	Default	98.6%	98.6%	98.6%	98.6%
				Tuning	98.6%	98.6%	98.6%	98.6%
		70%	30%	Default	99.1%	98.1%	100%	99.0%
				Tuning	99.1%	98.1%	100%	99.0%
		80%	20%	Default	100%	100%	100%	100%

Submitted: **02/02/2024**; Accepted: **30/04/2024**; Published: **30/04/2024**

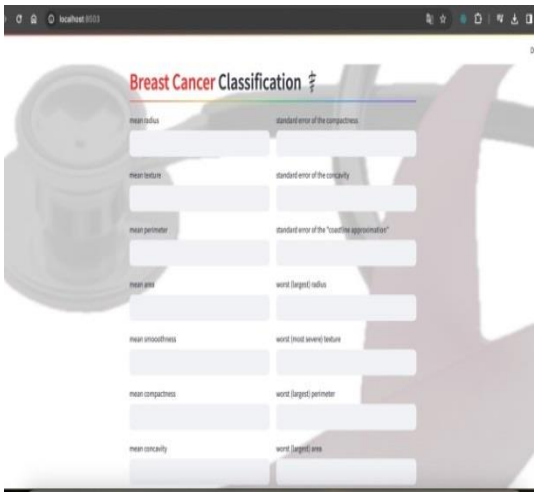
JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 2, April 2024, Page 889-899
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7457

Algoritma	Class Label	Splitting Data		Parameter	Accuracy	Precision	Recall	F1 Score
		Training	Testing					
		90%	10%	Tuning	100%	100%	100%	100%
				Default	100%	100%	100%	100%
				Tuning	100%	100%	100%	100%

Tabel 5 menunjukkan bahwa regresi logistik dapat mencapai performa klasifikasi terbaik, terutama dengan komposisi data pelatihan dan pengujian 80:20 dan 90:10. Namun, mengatur hyperparameter adalah langkah selanjutnya dalam pengoptimalan model. Pada Logistic Regression, parameter "C" diatur menjadi 0,5 dan "intercept_scaling" diatur menjadi 1.

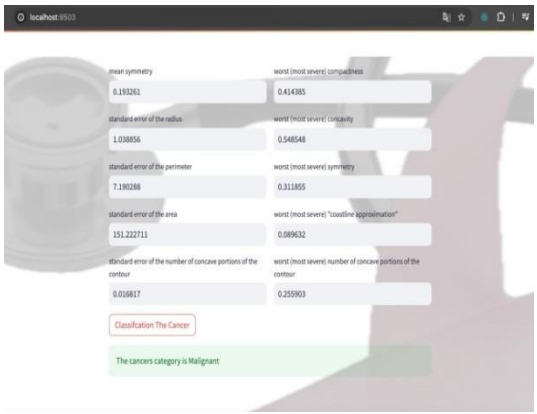
Hasil penyesuaian hyperparameter menggunakan Logistic Regression ditunjukkan dalam Tabel 5. Dengan penyesuaian ini, kita dapat melihat bagaimana perubahan parameter tersebut memengaruhi kinerja model klasifikasi. Ini adalah langkah penting dalam proses pengoptimalan model untuk mencapai kinerja terbaik dalam menangani masalah klasifikasi, terutama dalam kasus kanker payudara.

Deployment adalah langkah penting dalam siklus pengembangan perangkat lunak. Ini mencakup persiapan, konfigurasi, pengujian, dan peluncuran aplikasi atau sistem agar siap digunakan oleh pengguna. Pengguna dapat menggunakan Streamlit untuk mengisi data pengklasifikasian kanker payudara. Aplikasi ini memiliki antarmuka yang mudah digunakan, seperti yang terlihat pada Gambar 5 dan 6.



Gambar 6. Antar Muka Klasifikasi Kanker Payudara

Gambar 6 menunjukkan antarmuka (interface) dari aplikasi klasifikasi kanker payudara, yang mengandung kolom dari 32 fitur dari dataset "Breast Cancer Wisconsin Diagnostic". Fitur-fitur ini digunakan dalam proses klasifikasi untuk menentukan berdasarkan data yang diberikan apakah tumor payudara jinak (benign) atau ganas (malignant).



Gambar 7. Hasil Klasifikasi Kanker Payudara

Pada Gambar 7, Kita akan memasukkan nilai-nilai untuk setiap fitur yang mewakili data pasien dari dataset "Diagnosis Kanker Payudara Wisconsin", seperti yang ditunjukkan pada Gambar 6. Sistem akan menganalisis data setelah dimasukkan untuk menentukan apakah tumor payudara pasien benign (jinak) atau malignant (ganas).

Setelah analisis selesai, hasilnya menunjukkan bahwa tumor payudara pasien menunjukkan kanker payudara ganas (malignant). Dengan kata lain, sistem menemukan bahwa tumor payudara pasien adalah ganas dan membutuhkan perawatan medis tambahan. Aplikasi ini memudahkan pemeriksaan tumor payudara karena

Submitted: **02/02/2024**; Accepted: **30/04/2024**; Published: **30/04/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 2, April 2024, Page 889-899
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7457

antarmuka yang mudah digunakan dan proses yang cepat. Dokter dapat membuat diagnosis dan keputusan perawatan yang tepat berdasarkan informasi yang mereka dapatkan.

4. KESIMPULAN

Dataset yang diambil dari Kaggle.com untuk kanker payudara memiliki imbalance class. Sehingga penelitian ini diterapkan Teknik SMOTE untuk mengatasi imbalance class pada label target dataset. Hasil dari penelitian dengan menggunakan algoritma Decision Tree, Naïve Bayes, K-Nearest Neighbors, Logistic Regression dan Random Forest menggunakan komposisi data training dan data testing masing-masing dari 60:40, 70:30, 80:20 dan 90:10. Didapatkan algoritma terbaik menggunakan Teknik SMOTE untuk algoritma Logistic Regression pada komposisi data 80:20 dan 90:10 dengan metriks akurasi, presisi, recall, f1-score masing-masing sebanyak 100%. Kemudian setelah mencoba tuning hyper-parameter, hasil klasifikasi dari Logistic Regression tidak terdapat peningkatan.

REFERENCES

[1] Organisasi Kesehatan Dunia (World Health Organization), "Breast cancer." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer
[2] M. Arnold et al., "Current and future burden of breast cancer: Global statistics for 2020 and 2040," Breast, vol. 66, no. September, pp. 15–23, 2022, doi: 10.1016/j.breast.2022.08.010.
[3] W. KOUWENAAR, "On cancer incidence in Indonesia.," Acta Unio Int. Contra Cancrum, vol. 7, no. 1 Spec. No., pp. 61–71, 1951, [Online]. Available: https://gco.iarc.fr/today/data/factsheets/populations/360-indonesia-fact-sheets.pdf
[4] M. Kepala Biro Komunikasi dan Pelayanan Masyarakat drg. Widyawati, "Kanker Payudara Paling Banyak di Indonesia, Kemenkes Targetkan Pemerataan Layanan Kesehatan." [Online]. Available: https://sehatnegeriku.kemkes.go.id/baca/umum/20220202/1639254/kanker-payudaya-paling-banyak-di-indonesia-kemenkes-targetkan-pemerataan-layanan-kesehatan/
[5] S. Ara, A. Das, and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," 2021 Int. Conf. Artif. Intell. ICAI 2021, pp. 97–101, 2021, doi: 10.1109/ICAI52203.2021.9445249.
[6] S. R. Gupta, "Prediction time of breast cancer tumor recurrence using Machine Learning," Cancer Treat. Res. Commun., vol. 32, no. July, p. 100602, 2022, doi: 10.1016/j.ctarc.2022.100602.
[7] "American Cancer Society Recommendations for the Early Detection of Breast Cancer," 2023, [Online]. Available: https://www.cancer.org/cancer/types/breast-cancer/screening-tests-and-early-detection/american-cancer-society-recommendations-for-the-early-detection-of-breast-cancer.html
[8] National Breast Cancer Foundation, "Breast Self-Exam," 2024, [Online]. Available: https://www.nationalbreastcancer.org/breast-self-exam/
[9] M. Javaid, A. Haleem, R. Pratap Singh, R. Suman, and S. Rab, "Significance of machine learning in healthcare: Features, pillars and applications," Int. J. Intell. Networks, vol. 3, no. May, pp. 58–73, 2022, doi: 10.1016/j.ijin.2022.05.002.
[10] M. Nurkholifah, Jasmarizal, Y. Umar, and Rahmaddeni, "Analisa Performa Algoritma Machine Learning Dalam Prediksi Penyakit Liver," J. Indones. Manaj. Inform. dan Komun., vol. 4, no. 1, pp. 164–172, 2023, doi: 10.35870/jimik.v4i1.149.
[11] K. M. M. Uddin, N. Biswas, S. T. Rikta, and S. K. Dey, "Machine learning-based diagnosis of breast cancer utilizing feature optimization technique," Comput. Methods Programs Biomed. Updat., vol. 3, no. February, p. 100098, 2023, doi: 10.1016/j.cmpbup.2023.100098.
[12] K. Kousalya, B. Krishnakumar, C. I. Shanthosh, R. Sharmila, and V. Sneha, "Diagnosis of breast cancer using machine learning algorithms," Int. J. Adv. Sci. Technol., vol. 29, no. 3 Special Issue, pp. 970–974, 2020.
[13] M. M. Hassan et al., "A comparative assessment of machine learning algorithms with the Least Absolute Shrinkage and Selection Operator for breast cancer detection and prediction," Decis. Anal. J., vol. 7, no. April, p. 100245, 2023, doi: 10.1016/j.dajour.2023.100245.
[14] V. Birchha and B. Nigam, "Performance Analysis of Averaged Perceptron Machine Learning Classifier for Breast Cancer Detection," Procedia Comput. Sci., vol. 218, no. 2022, pp. 2181–2190, 2022, doi: 10.1016/j.procs.2023.01.194.
[15] Y. Hendra Kusuma, S. Supraapto, and Y. Setiawan, "Analisis Kepuasan Penumpang pada Maskapai Penerbangan Menggunakan Algoritma C4.5 dan Naïve Bayes," SENTIMAS Semin. Nas. Penelit. dan Pengabdi. Masy., pp. 162–171, 2022, [Online]. Available: https://journal.irpi.or.id/index.php/sentimas/article/view/320/125
[16] R. Gonzalez, P. Nejat, A. Saha, C. J. V. Campbell, A. P. Norgan, and C. Lokker, "Performance of externally validated machine learning models based on histopathology images for the diagnosis, classification, prognosis, or treatment outcome prediction in female breast cancer: A systematic review," J. Pathol. Inform., vol. 15, no. August 2023, p. 100348, 2024, doi: 10.1016/j.jpi.2023.100348.
[17] P. Gupta and S. Garg, "Breast Cancer Prediction using varying Parameters of Machine Learning Models," Procedia Comput. Sci., vol. 171, pp. 593–601, 2020, doi: 10.1016/j.procs.2020.04.064.
[18] E. Gentili et al., "Machine learning from real data: A mental health registry case study," Comput. Methods Programs Biomed. Updat., vol. 5, no. August 2023, p. 100132, 2023, doi: 10.1016/j.cmpbup.2023.100132.
[19] A. N. Kasanah, M. Muladi, and U. Pujianto, "Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 3, no. 2, pp. 196–201, 2019, doi: 10.29207/resti.v3i2.945.
[20] R. Resmiati and T. Arifin, "Klasifikasi Pasien Kanker Payudara Menggunakan Metode Support Vector Machine dengan Backward Elimination," Sistemasi, vol. 10, no. 2, p. 381, 2021, doi: 10.32520/stmsi.v10i2.1238.
[21] J. KUSUMA, B. H. HAYADI, W. WANAYUMINI, and R. ROSNELLY, "Komparasi Metode Multi Layer Perceptron (MLP) dan Support Vector Machine (SVM) untuk Klasifikasi Kanker Payudara," MIND J., vol. 7, no. 1, pp. 51–60, 2022, doi: 10.26760/mindjournal.v7i1.51-60.

Submitted: **02/02/2024**; Accepted: **30/04/2024**; Published: **30/04/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 2, April 2024, Page 889-899
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7457

[22] H. Harafani and H. A. Al-Kautsar, "Meningkatkan Kinerja K-Nn Untuk Klasifikasi Kanker Payudara Dengan Forward Selection," J. Pendidik. Teknol. dan Kejuru., vol. 18, no. 1, p. 99, 2021, doi: 10.23887/jptk-undiksha.v18i1.29905.

[23] A. I. S. Azis, Irma Surya Kumala Idris, Budy Santoso, and Yasin Aril Mustofa, "Pendekatan Machine Learning yang Efisien untuk Prediksi Kanker Payudara," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 3, no. 3, pp. 458–469, 2019, doi: 10.29207/resti.v3i3.1347.

[24] M. A. Naji, S. El Filali, K. Aarika, E. H. Benlahmar, R. A. Abdelouhahid, and O. Debauche, "Machine Learning Algorithms for Breast Cancer Prediction and Diagnosis," Procedia Comput. Sci., vol. 191, pp. 487–492, 2021, doi: 10.1016/j.procs.2021.07.062.

Submitted: **02/02/2024**; Accepted: **30/04/2024**; Published: **30/04/2024**