Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, Juni 2024 Page: 499-508
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5421

Klasifikasi Penyakit Jantung Tipe Kardiovaskular Menggunakan Adaptive Synthetic Sampling dan Algoritma Extreme Gradient Boosting

Acep Handika Permana, Fajri Rakhmat Umbara*, Fatan Kasyidi

Fakultas Sains dan Informatika, Program Studi Teknik Informatika, Universitas Jenderal Achmad Yani, Cimahi, Indonesia
Email: [1]acephandikap20@if.unjani.ac.id, [2*]fajri.rakhmat@lecture.unjani.ac.id, [3]fatan.kasyidi@lecture.unjani.ac.id
Email Penulis Korespondensi: fajri.rakhmat@lecture.unjani.ac.id
Submitted: **26/06/2024**; Accepted: **30/06/2024**; Published**: 30/06/2024**

Abstrak−Penyakit kardiovaskular adalah penyakit yang menyerang sistem kardiovaskular, seperti penyakit jantung dan stroke. Menurut data World Health Organization (WHO), 17,9 juta kematian di seluruh dunia pada tahun 2019 disebabkan oleh penyakit kardiovaskular. Deteksi dini sangat penting dilakukan, namun diagnosa penyakit jantung menjadi kompleks di negara berkembang karena keterbatasan alat diagnostik dan tenaga medis. Penelitian ini menggunakan Heart Disease Dataset dari Kaggle, terdiri dari 15 atribut dan 4238 record, untuk mengembangkan model klasifikasi penyakit jantung menggunakan XGBoost. Tahapan penelitian meliputi imputasi data, transformasi data dengan LabelEncoder, penyeimbangan data menggunakan ADASYN, pembagian data kedalam 80% pelatihan, 20% pengujian, serta tuning hyperparameter dengan Bayesian Optimization. Hasil menunjukkan bahwa model XGBoost dengan ADASYN memiliki performa lebih baik dengan ROC-AUC sebesar 0.971 dan akurasi sebesar 0.916, dibandingkan dengan model tanpa ADASYN yang hanya memiliki ROC-AUC sebesar 0.698 dan akurasi sebesar 0.841. Berdasarkan hasil penelitian, ADASYN terbukti efektif meningkatkan performa model pada dataset yang tidak seimbang. Selain itu, Bayesian Optimization juga berperan penting dalam menemukan kombinasi parameter optimal, yang dapat meningkatkan kinerja model lebih lanjut. Dengan adanya penelitian ini, dampak yang diberikan cukup signifikan dalam pengembangan metode deteksi dini penyakit jantung kardiovaskular, khususnya melalui penerapan algoritma klasifikasi XGBoost.

Kata Kunci : Penyakit Jantung; Kardiovaskular; Klasifikasi; ADASYN; XGBoost

Abstract−Cardiovascular diseases are conditions that commonly affect the cardiovascular system, such as heart disease and stroke. According to data from the World Health Organization (WHO), 17.9 million deaths worldwide in 2019 were attributable to cardiovascular disease. Early detection is crucial, but diagnosing heart disease is complex in developing countries due to the limited availability of diagnostic tools and medical personnel. This study uses the Heart Disease Dataset from Kaggle, consisting of 15 attributes and 4238 records, to develop a heart disease classification model using XGBoost. The research stages include data imputation, data transformation using LabelEncoder, data balancing using ADASYN, data splitting (80% training data, 20% testing data), and hyperparameter tuning with Bayesian Optimization. The results show that the XGBoost model with ADASYN performs better, with a ROC-AUC of 0.971 and an accuracy of 0.916, compared to the model without ADASYN, which has a ROC-AUC of 0.698 and an accuracy of 0.841. Based on the research results, ADASYN has proven effective in improving model performance on imbalanced datasets. Additionally, Bayesian Optimization plays an important role in finding the optimal parameter combination, which can further enhance model performance. With this research, the impact is quite significant in the development of early detection methods for cardiovascular heart disease, particularly through the application of the XGBoost classification algorithm.

Keywords : Heart Disease; Cardiovascular; Classification; ADASYN; XGBoost

1. PENDAHULUAN

Kardiovaskular adalah istilah yang mengacu pada sistem pembuluh darah dan jantung. Penyakit kardiovaskular adalah penyakit yang menyerang sistem kardiovaskular, seperti penyakit jantung dan stroke[1]. Menurut data World Health Organization (WHO), 17,9 juta kematian di seluruh dunia pada tahun 2019 disebabkan oleh penyakit kardiovaskular. Angka ini menunjukkan bahwa penyakit jantung dan stroke menyumbang 85% dari seluruh kematian secara global. Faktor-faktor risiko penyakit kardiovaskular meliputi usia, kebiasaan merokok, pola makan, konsumsi alkohol, kelebihan berat badan, tingkat aktivitas fisik, dan tekanan darah tinggi[2]. Secara umum, setengah dari populasi pasien yang telah diidentifikasi menderita Penyakit Jantung mengalami kematian dalam rentang waktu satu hingga dua tahun[3]. Oleh karena itu, pendeteksian dini sangat penting. Namun, penanganan dan diagnosis penyakit jantung menjadi proses yang kompleks, terutama di negara-negara berkembang, karena keterbatasan alat diagnostik dan kekurangan tenaga medis serta sumber daya lainnya. Kekurangan ini berdampak pada kemampuan untuk melakukan prediksi dan pengobatan yang efektif terhadap pasien jantung[4]. Untuk mengatasi masalah ini, teknologi Machine Learning dapat digunakan untuk mengenali dan memprediksi penyakit jantung. Pendekatan ini menganalisis data kesehatan dan memberikan prediksi akurat terkait risiko penyakit jantung[5].

Terdapat penelitian terdahulu yang mengkaji penggunaan pendekatan Machine Learning untuk memprediksi penyakit kardiovaskular. Dataset Cleveland, yang dikumpulkan dari repositori data mining dan pembelajaran mesin online University of California, Irvine (UCI), menyediakan data untuk penelitian ini. Penelitian tersebut menggunakan Metode XGBoost yang ditingkatkan dengan Bayesian Optimization. Akurasi yang tercapai dalam penelitian ini mencapai 91.8%. Selain itu, metode ini menunjukkan kinerja yang baik dalam memprediksi penyakit jantung dengan Area di Bawah Kurva (AUC) dari grafik ROC sebesar 0.9134[6]. Terdapat penelitian terdahulu lain yang menggunakan data yang sama dengan penelitian ini, yaitu dataset bernama "Heart Disease Dataset" yang dipublikasikan oleh Mirza HASNINE dalam website Kaggle.com. Dataset tersebut memiliki 4,238 record dan 16 atribut yang berbeda. Untuk mengidentifikasi data yang terkait dengan penyakit jantung, penelitian ini menggunakan berbagai metode pembelajaran mesin, seperti Logistic Regression (LR), Artificial Neural Network, KNN, Decision

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, Juni 2024 Page: 499-508
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5421

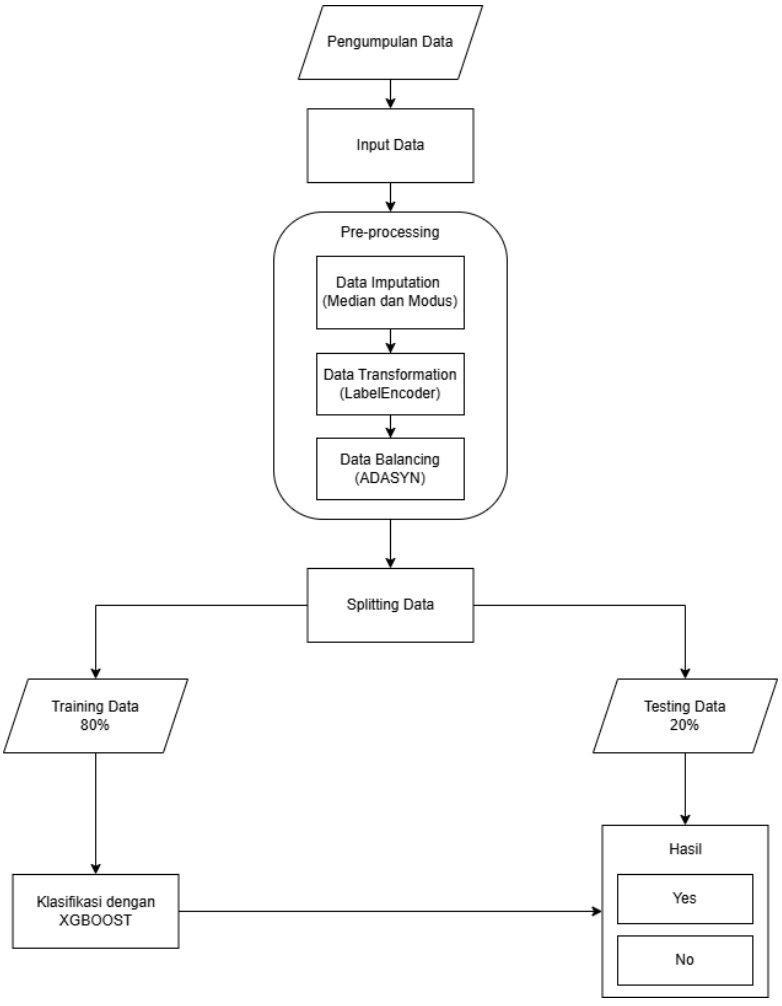
Trees DT, Naive Bayes, dan SVM. Dengan akurasi sebesar 86,1%, hasil penelitian tersebut menunjukkan keunggulan logistic regression%[7]. Extreme Gradient Boosting (XGBoost) pertama kali diperkenalkan pada tahun 2014 oleh Tianqi Chen dari University of Washington[8]. XGBoost merupakan algoritma pembelajaran mesin yang menggunakan gabungan pohon keputusan dan teknik gradient boosting untuk membuat model prediksi yang kuat. Algoritma ini secara berurutan membangun sebuah ensemble dari pohon keputusan yang lemah, di mana setiap pohon keputusan berikutnya memperbaiki kesalahan yang dibuat oleh pohon keputusan sebelumnya[9]. XGBoost dirancang untuk memiliki efisiensi, fleksibilitas, dan portabilitas tinggi untuk menyelesaikan masalah klasifikasi dan regresi[10]. XGBoost dianggap mampu mengatasi volume data yang besar, bahkan mencapai miliaran entri, dengan menggunakan sumber daya yang lebih efisien. Kecepatannya dianggap melebihi solusi populer lainnya lebih dari sepuluh kali lipat, menjadikannya pilihan yang sangat efektif untuk menangani tugas analisis data yang kompleks[11].

Dalam penelitian ini, dataset yang digunakan menghadapi tantangan ketidakseimbangan data, dimana kelas mayoritas memiliki lebih banyak data daripada kelas minoritas. Ketidakseimbangan ini dapat berpengaruh pada hasil klasifikasi karena nilai sensitivitas dan Area Under Curve (AUC) dapat menurun sebagai akibat dari ketidakseimbangan ini[12]. Salah satu cara untuk mengatasi ketidakseimbangan data adalah dengan teknik oversampling, tetapi ini sering menyebabkan overfitting karena hanya memperbanyak data yang sudah ada[13]. Sebagai alternatif, dikembangkan Metode Adaptive Synthetic Sampling (ADASYN). Berbeda dengan oversampling biasa yang hanya menyalin data yang sudah ada, ADASYN menciptakan data sintetis baru untuk kelas minoritas. Pendekatan ini dianggap dapat mengurangi risiko overfitting[14].

Berdasarkan uraian di atas, penelitian ini telah melakukan klasifikasi terhadap Penyakit Jantung Kardiovaskular dengan menggunakan Metode Extreme Gradient Boosting (XGBoost). Mengingat adanya ketidakseimbangan kelas dalam dataset yang digunakan, langkah-langkah untuk menangani masalah ini akan diambil dengan menerapkan teknik ADASYN (Adaptive Synthetic Sampling).

2. METODOLOGI PENELITIAN

Metodologi penelitian merupakan serangkaian langkah atau pendekatan sistematis yang digunakan oleh peneliti dalam suatu penelitian. Gambar 1 mengilustrasikan tahapan penelitian secara lebih rinci.



Gambar 1. Metodologi Penelitian

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, Juni 2024 Page: 499-508
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5421

2.1 Pengumpulan Data

Pengumpulan data merupakan langkah awal dalam penelitian karena dataset yang dipilih akan menjadi dasar dari analisis dan temuan yang dihasilkan[9]. Penelitian ini menggunakan Heart Disease Dataset dari platform Kaggle oleh Mirza Hasnine (https://www.kaggle.com/datasets/mirzahasnine/heart-disease-dataset). Dataset ini memiliki 4238 record dan16 atribut.

2.2 *Pre-processing*

Setelah data awal berhasil diperoleh dari Heart Disease Dataset yang ditemukan di website Kaggle, langkah berikutnya adalah tahap pra-proses. Tahap ini terdiri dari beberapa langkah penting yang meliputi data imputation, data transformation, dan data balancing.

a. Data Imputation

Tahap awal dalam pra-proses data adalah imputasi data. Pada langkah ini, perhatian utama adalah menangani masalah nilai yang hilang (missing values) dengan mengimputasinya dari dataset. Metode imputasi yang digunakan adalah modus untuk data kategorik dan median untuk data numerik[15]. Penggunaan metode ini dilakukan karena terdapat banyak nilai yang hilang pada beberapa atribut dalam dataset yang digunakan.

b. Data Transformation

Setelah membersihkan data, langkah berikutnya adalah transformasi data. Langkah ini krusial untuk mengubah format data generik seperti teks atau kategori menjadi format numerik yang dapat diproses oleh algoritma pemodelan. Teknik transformasi data yang digunakan dalam penelitian ini disebut LabelEncoder. LabelEncoder berfungsi mengubah nilai-nilai dalam kolom kategori menjadi nilai numerik, dimulai dari 0 hingga jumlah kategori yang ada[16].

c. Data Balancing

Data balancing mengacu pada upaya menyeimbangkan distribusi kelas atau target variabel dalam dataset. Salah satu cara umum untuk mengatasi ketidakseimbangan kelas adalah dengan oversampling, di mana sampel dari kelas minoritas diperbanyak hingga jumlahnya sebanding dengan kelas mayoritas. Metode ADASYN (Adaptive Synthetic Sampling) adalah salah satu teknik resampling yang digunakan dalam pemrosesan data tidak seimbang (imbalanced data) untuk menangani masalah kelas minoritas[14]. ADASYN bekerja dengan membuat sampel sintetis baru dari kelas minoritas.

2.3 Proses Klasifikasi

Setelah menyelesaikan fase pra-pemrosesan, data dibagi menjadi set pelatihan dan pengujian. Berdasarkan penelitian sebelumnya, yang mengusulkan perbandingan 80% untuk data latih dan 20% untuk data uji[17]. Dalam penelitian ini, selain menggunakan rasio 80:20, juga digunakan rasio 70:30 dan 60:40 untuk menguji rasio mana yang paling baik dalam menghasilkan model yang optimal.Setelah pemisahan data dilakukan, langkah selanjutnya adalah melakukan hyperparameter tunning menggunakan Bayesian Optimization. Bayesian Optimization dipilih karena memiliki beberapa keunggulan dibandingkan dengan metode tuning hyperparameter lainnya, seperti Grid Search dan Random Search. Bayesian Optimization lebih efisien karena menggunakan model probabilistik untuk memprediksi performa model dengan berbagai kombinasi hyperparameter. Dengan pendekatan ini, proses pencarian hyperparameter optimal dapat dilakukan lebih cepat dan akurat[17]. Tabel 1 menampilkan daftar parameter yang digunakan untuk hyperparameter tunning.

Tabel 1. Parameter XGBoost

Parameter	Keterangan	Nilai Bayesian Optimization
n_estimators	Jumlah pohon keputusan yang digunakan dalam model.	Integer(10, 5000)
max_depth	Kedalaman maksimum pohon keputusan.	Integer(70, 100)
learning_rate	Kecepatan pembelajaran; menentukan seberapa cepat model belajar.	Real(0.01, 1.0)
gamma	Penghalang untuk split node; nilai lebih tinggi membuat model lebih konservatif.	Real(0.05, 0.1)
subsample	Proporsi data pelatihan yang digunakan untuk setiap pohon.	Real(0.5, 1.0)
colsample_bytree	Proporsi fitur yang digunakan untuk setiap pohon.	Real(0.7, 1.0)
min_child_weight	Bobot minimum sum dari anak-anak di setiap node.	Integer(10, 15)
reg_lambda	Parameter regulasi L2 (Ridge); mengurangi kompleksitas model.	Real(0.01, 100)
reg_alpha	Parameter regulasi L1 (Lasso); meningkatkan kejarangan bobot fitur.	Real(0.01, 100)

Setelah proses tuning hyperparameter selesai, langkah berikutnya adalah melakukan modeling klasifikasi menggunakan XGBoost. XGBoost (Extreme Gradient Boosting) pertama kali oleh Tianqi Chen dari University of Washington[18]. XGBoost adalah teknik ensemble yang berbasis pada gradient boosting, yang memungkinkan pembangunan model yang sangat akurat dan efisien. XGBoost memiliki kemampuan untuk menangani berbagai jenis data, termasuk data yang sangat besar dan kompleks. Algoritma ini juga memiliki kemampuan untuk melakukan

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, Juni 2024 Page: 499-508
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5421

penyetelan parameter yang sangat baik, sehingga memungkinkan pengguna untuk menyesuaikan model secara optimal[8].

2.5 Pengujian dan Evalusi

Tahap terakhir adalah pengujian dan evaluasi, di mana pada tahap ini dilakukan evaluasi kinerja model klasifikasi menggunakan beberapa metode evaluasi. Berikut adalah metode evaluasi yang digunakan :

a. ROC – AUC

Kinerja model klasifikasi dapat diukur dengan menggunakan kurva ROC (Characteristics of Receiver Operating) dan AUC (Area Under the Curve)[19]. Evaluasi menggunakan kurva ROC dan AUC dapat memberikan informasi yang berguna tentang kinerja model, terutama dalam kasus di mana kelas-kelas yang diamati tidak seimbang. Sebagai contoh, ketika jumlah sampel positif jauh lebih sedikit daripada jumlah sampel negatif[20].

b. Confusion Matrix

Confusion matrix merupakan sebuah metode evaluasi atau pengujian yang memberikan wawasan berharga tentang kinerja model klasifikasi. Confusion Matrix ini digunakan untuk mengevaluasi kemampuan model untuk mengkategorikan data ke dalam kategori yang tepat. Dengan memanfaatkan confusion matrix, kita dapat memahami secara rinci di mana model berperforma baik dan di mana model mungkin membuat kesalahan[21].

3. HASIL DAN PEMBAHASAN

Pada bagian ini akan diberikan penjelasan rinci mengenai temuan dan analisis dari setiap fase penelitian "Klasifikasi Penyakit Jantung Kardiovaskular Menggunakan Algoritma XGBoost". Penjelasan meliputi proses pengumpulan data, pre-processing, proses klasifikasi, serta pengujian dan evaluasi kinerja algoritma klasifikasi.

3.1 Pengumpulan Data

Dalam penelitian ini menggunakan Heart Disease Dataset yang didapatkan dari platform Kaggle oleh Mirza Hasnine. Dataset ini memiliki 16 atribut dan 4238 record. Setiap atribut dalam dataset ini memiliki peran penting dalam menentukan kondisi kardiovaskular seseorang. Tabel 2 berikut adalah deskripsi dari masing-masing atribut yang terdapat dalam dataset:

Tabel 2. Deskripsi Atribut Data

Atribut	Deskripsi
Gender	Jenis kelamin subjek
Age	Usia subjek
Education	Tingkat pendidikan subjek
CurrentSmoker	Status perokok saat ini
CigsPerDay	Jumlah konsumsi rokok per hari
BPMeds	Penggunaan obat tekanan darah
PrevalentStroke	Riwayat stroke sebelumnya
PrevalentHyp	Riwayat hipertensi (tekanan darah tinggi)
Diabetes	Diabetes
TotChol	Jumlah total kolestrol
SysBP	Tekanan darah sistolik
DiaBP	Tekanan darah diastolik
BMI	Index masa tubuh
HeartRate	Denyut jantung per menit
Glucose	Kadar glukosa dalam tubuh
Heart_stroke	Penyakit ini disebabkan oleh penghentian suplai darah ke otak, yang dapat menyebabkan gejala seperti kelemahan otot, kesulitan berbicara, dan pingsan.

Penjelasan di atas memberikan gambaran tentang variabel yang ada dalam dataset dan bagaimana masing-masing variabel berkontribusi terhadap klasifikasi penyakit jantung kardiovaskular. Dataset ini memfasilitasi analisis mendalam terhadap faktor-faktor risiko yang mempengaruhi kesehatan jantung, memungkinkan peneliti untuk membangun model prediksi yang lebih akurat. Berikut tabel 3 merupakan dataset yang digunakan dalam penelitian ini:

Tabel 3. Data Penyakit Jantung

No	Gender	age	education	currentSmoker	…	Heart_stroke
1	Male	39	postgraduate	0	…	No
2	Female	46	primaryschool	0	…	No
3	Male	48	uneducated	1	…	No

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, Juni 2024 Page: 499-508
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5421

4	Female	61	graduate	1	…	yes
5	Female	46	graduate	1	…	No
…	…	…	…	…	…	…
4234	Male	50	uneducated	1	…	yes
4235	Male	51	graduate	1	…	No
4236	Female	48	primaryschool	1	…	No
4237	Female	44	uneducated	1	…	No
4238	Female	52	primaryschool	0	…	No

3.2 Pre-Processing

Tahap pra-proses bertujuan untuk mempersiapkan data sebelum dilakukan proses klasifikasi. Pra-proses melibatkan tiga tahapan utama, yaitu pembersihan data, transformasi data, dan penyeimbangan data.

3.2.1 Data Imputation

Tahap awal dalam pra-proses data adalah imputasi data. Pada langkah ini, perhatian utama adalah menangani masalah nilai yang hilang (missing values) dengan mengimputasinya dari dataset. Metode imputasi yang digunakan adalah modus untuk data kategorik dan median untuk data numerik. Seperti terlihat pada Gambar 2, pendekatan ini dipilih karena banyaknya nilai yang hilang untuk sejumlah atribut dalam dataset.

```
Missing Value :
Gender              0
age                 0
education         105
currentSmoker       0
cigsPerDay         29
BPMeds             53
prevalentStroke     0
prevalentHyp        0
diabetes            0
totChol            50
sysBP               0
diaBP               0
BMI                19
heartRate           1
glucose           388
Heart_stroke        0
```

Gambar 2. Jumlah Missing Value

Dapat dilihat pada gambar 2 diatas terdapat missing value pada atribut education sebanyak 105 buah, cigsPerDay 29 buah, BPMeds 53 buah, totChol 50 buah, BMI 19 buah, heartRate 1 buah dan glucose 388 buah. Untuk menangani masalah tersebut, digunakan metode imputasi dengan median dan modus.

3.2.2 Data Transformation

Langkah selanjutnya dalam tahap pra-proses adalah melakukan transformasi data. Transformasi ini bertujuan untuk mengubah data kategorikal menjadi data numerik agar dapat diproses oleh algoritma machine learning dengan lebih efisien. Hal ini diperlukan karena terdapat beberapa atribut yang memiliki data kategorikal.

Tabel 4. Dataset sebelum *Data Transformation*

No	Gender	…	education	…	prevalentStroke	…	Heart_stroke
1	Male	…	postgraduate	…	no	…	No
2	Female	…	primaryschool	…	no	…	No
3	Male	…	uneducated	…	no	…	No
4	Female	…	graduate	…	no	…	yes
5	Female	…	graduate	…	no	…	No
…	…	…	…	…	…	…	…
4234	Male	…	uneducated	…	no	…	yes
4235	Male	…	graduate	…	no	…	No
4236	Female	…	primaryschool	…	no	…	No
4237	Female	…	uneducated	…	no	…	No
4238	Female	…	primaryschool	…	no	…	No

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, Juni 2024 Page: 499-508
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5421

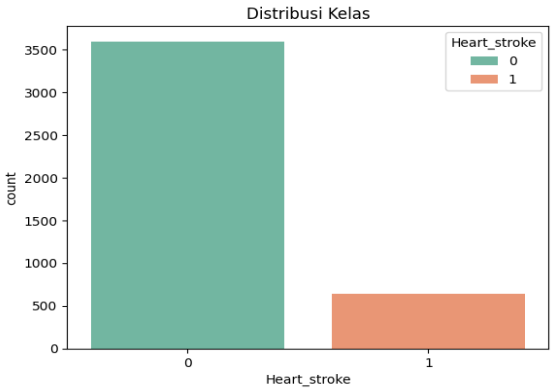
Dari Tabel 4 di atas, terlihat bahwa dalam dataset masih terdapat 4 atribut yang memiliki tipe data kategorikal, yaitu Gender, education, prevalentStroke, dan Heart_Stroke. Atribut-atribut ini perlu diubah menjadi data numerik agar dapat diproses secara efisien oleh algoritma machine learning. Tahap transformasi data ini akan menggunakan metode LabelEncoder dari paket scikit-learn. Tabel 5 menampilkan hasil transformasi data yang dilakukan LabelEncoder.

Tabel 5. Dataset sesudah *Data Transformation*

No	Gender	…	education	…	prevalentStroke	…	Heart_stroke
1	1	…	1	…	0	…	0
2	0	…	2	…	0	…	0
3	1	…	3	…	0	…	0
4	0	…	0	…	0	…	1
5	0	…	0	…	0	…	0
…	…	…	…	…	…	…	…
4234	1	…	3	…	0	…	1
4235	1	…	0	…	0	…	0
4236	0	…	2	…	0	…	0
4237	0	…	3	…	0	…	0
4238	0	…	2	…	0	…	0

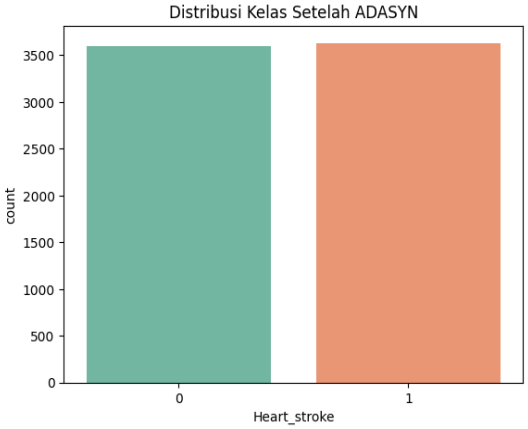
3.2.3 Data Balancing

Dalam Dataset yang digunakan dalam penelitian ini terdapat masalah ketidakseimbangan kelas, yang dapat dilihat pada Gambar 3 di bawah ini:



Gambar 3. Distribusi Kelas sebelum *Balancing Data*

Pada gambar 3, terlihat adanya ketidakseimbangan antara kelas “0” dan kelas "1". Jumlah kelas 0 dalam dataset di atas adalah 3594, sedangkan jumlah kelas 1 adalah 644, sehingga dataset tersebut memiliki rasio ketidakseimbangan sebesar 5.58 : 1. Dalam penelitian ini, untuk mengatasi masalah tersebut, digunakan Metode Adaptive Synthetic Sampling (ADASYN). Teknik ini bekerja dengan menghasilkan sampel sintetis dari kelas minoritas untuk menyeimbangkan distribusi kelas. Dengan menerapkan ADASYN, data yang telah diseimbangkan dapat diperoleh, sebagaimana ditunjukkan pada Gambar 4.



Gambar 4. Distribusi Kelas setelah *Balancing Data*

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, Juni 2024 Page: 499-508
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5421

Setelah penerapan ADASYN, jumlah data yang semula 4238 menjadi 7227 catatan, dengan kelas "0" berjumlah 3594 dan kelas "1" berjumlah 3633. Untuk lebih jelasnya bisa dilihat dalam table 5.

Tabel 6. Dataset sesudah *Data Balancing*

No	Gender	age	education	currentSmoker	…	Heart_stroke
1	1	39	1	0	…	0
2	0	46	2	0	…	0
3	1	48	3	1	…	0
4	0	61	0	1	…	1
5	0	46	0	1	…	0
…	…	…	…	…	…	…
7223	1	52	3	0	…	1
7224	0	56	2	0	…	1
7225	0	50	3	0	…	1
7226	0	61	3	0	…	1
7227	0	50	3	0	…	1

3.3 Proses Klasifikasi

Data tersebut harus dipecah (split) menurut rasio tertentu sebelum proses kategorisasi dapat dimulai. Beberapa rasio pembagian data, antara lain 60%:40%, 70%:30%, dan 80%:20%, digunakan dalam penelitian ini. Pembagian dengan tiga rasio ini bertujuan untuk menguji mana rasio yang paling baik dalam menghasilkan model klasifikasi yang optimal. Setelah data dibagi, langkah berikutnya adalah melakukan tuning hyperparameter menggunakan Bayesian Optimization. Dalam penggunaan XGBoost, penggabungan dengan Bayesian Optimization bisa sangat berguna untuk membangun model yang optimal. Dalam penelitian terdahulu disebutkan bahwa efektivitas metode XGBoost sangat tergantung pada jumlah parameter yang dimodifikasi oleh pengguna[6]. Dengan Bayesian Optimization, pengguna bisa menyesuaikan parameter kunci secara efisien untuk meningkatkan kinerja model XGBoost. Berikut tabel 7 hasil tuning parameter dari Bayesian Optimization untuk model XGBoost :

Tabel 7. Hasil Hyperparameter Tunning dengan Bayesian Optimization

Parameter	Nilai Bayesian Optimization	Nilai Parameter Terbaik
n_estimators	Integer(10, 5000)	4505
max_depth	Integer(70, 100)	96
learning_rate	Real(0.01, 1.0)	0.01
gamma	Real(0.05, 0.1)	0.0585
subsample	Real(0.5, 1.0)	1.0
colsample_bytree	Real(0.7, 1.0)	0.7308
min_child_weight	Integer(10, 15)	14
reg_lambda	Real(0.01, 100)	0.0292
reg_alpha	Real(0.01, 100)	0.2036

Setelah dilakukan hyperparameter tuning menggunakan Bayesian Optimization, langkah selanjutnya adalah melakukan klasifikasi dengan menggunakan XGBoost. Model XGBoost yang telah dioptimalkan dengan parameter terbaik dari proses tuning akan digunakan untuk melatih data latih. Proses pelatihan ini memungkinkan model untuk belajar dari pola yang ada dalam data latih. Setelah model XGBoost dilatih dengan data latih, langkah selanjutnya adalah mengevaluasi performa model menggunakan data uji. Untuk melakukan evaluasi, hasil prediksi model dibandingkan dengan nilai target dalam data uji. Akurasi adalah metrik evaluasi yang paling umum digunakan, yang menunjukkan seberapa baik model dapat mengklasifikasikan data dengan tepat.

Tabel 8. Hasil Akurasi Klasifikasi

Metode	Akurasi		
	60:40	70:30	80:20
XGBoost	85.9%	84.9%	84.1%
XGBoost + ADASYN	89.4%	90.3%	91.6%

Tabel 8 menunjukkan hasil akurasi dari klasifikasi menggunakan model XGBoost dan XGBoost dengan ADASYN setelah dievaluasi dengan berbagai rasio pembagian data.

3.4 Pengujian dan Evaluasi

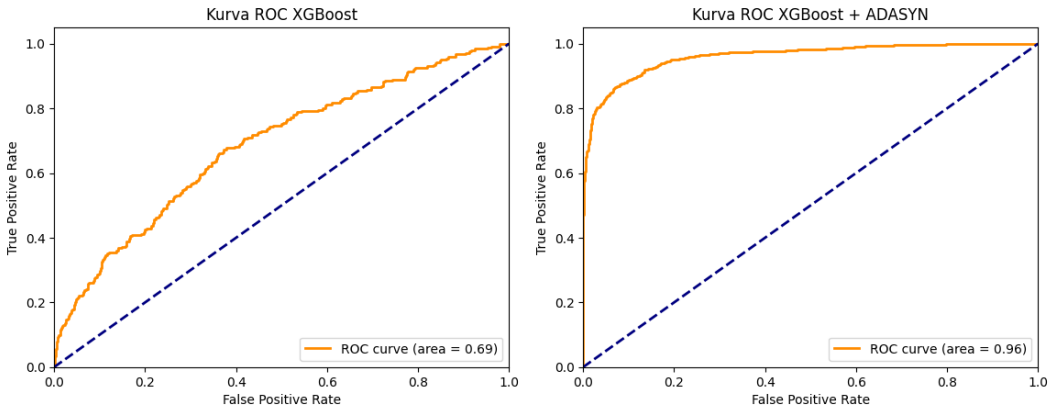
Pada tahap pengujian model algoritma menggunakan metode XGBoost, dilakukan dua pendekatan yaitu XGBoost dengan ADASYN dan tanpa ADASYN untuk mengatasi ketidakseimbangan kelas dalam dataset. Dengan menggunakan Confusion Matrix, kinerja model dinilai dengan metrik-metrik tertentu yang diperoleh dari true labels

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, Juni 2024 Page: 499-508
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5421

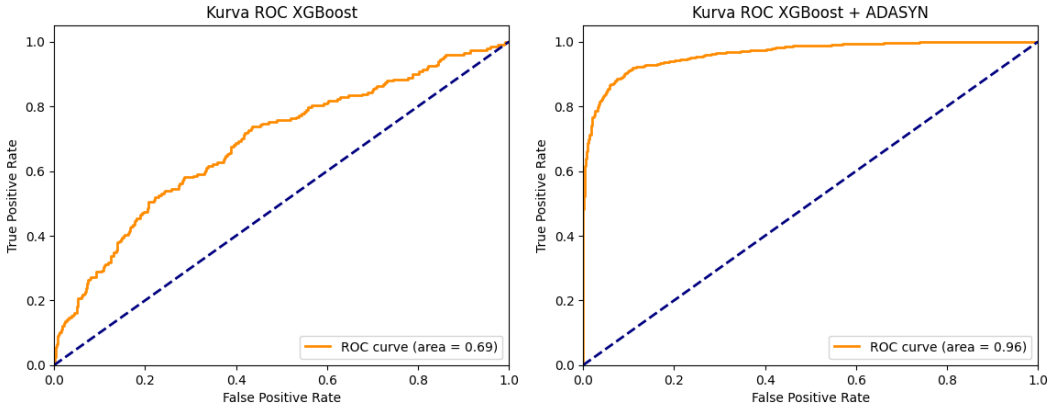
dan predicted labels. Selain itu, kemampuan model untuk membedakan antara kelas positif dan negatif diukur dengan ROC-AUC.

3.4.1 Pengujian ROC - AUC

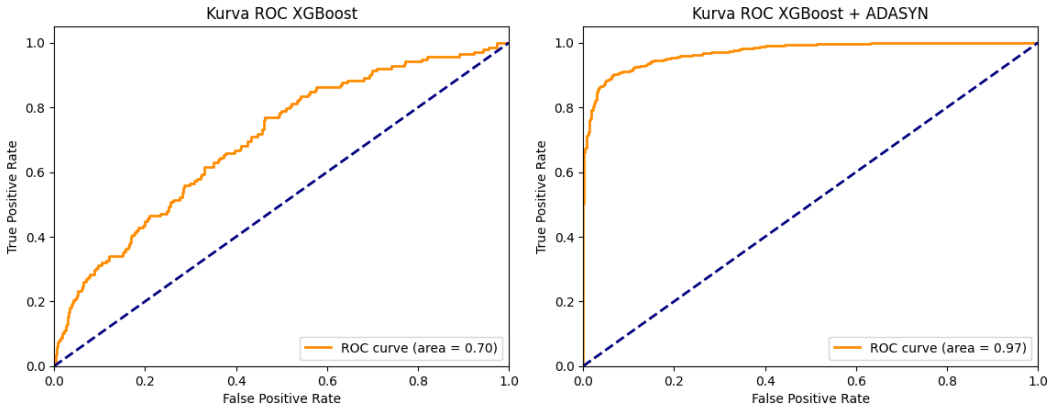
Pada tahap ini, pengujian dilakukan menggunakan ROC-AUC untuk mengevaluasi kemampuan model XGBoost untuk membedakan antara kelas negatif dan positif. Dalam klasifikasi biner, metrik evaluasi penting adalah ROC-AUC (Receiver Operating Characteristic—Area Under Curve). Gambar 5 hingga gambar 7 merupakan kurva ROC menunjukkan perbandingan True Positive Rate (TPR) dan False Positive Rate (FPR) di berbagai treshold prediksi model.



Gambar 5. Hasil Kurva ROC rasio 60:40



Gambar 6. Hasil Kurva ROC rasio 70:30



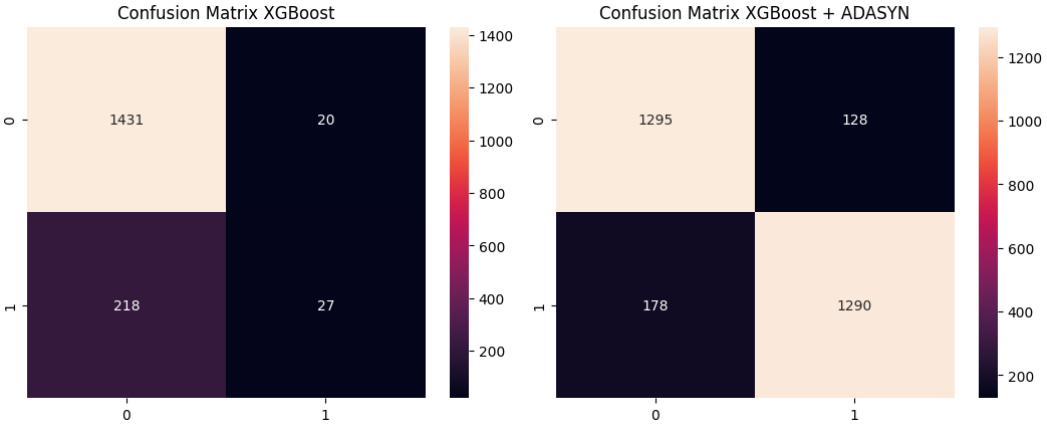
Gambar 7. Hasil Kurva ROC rasio 80:20

3.4.2 Pengujian Confusion Matrix

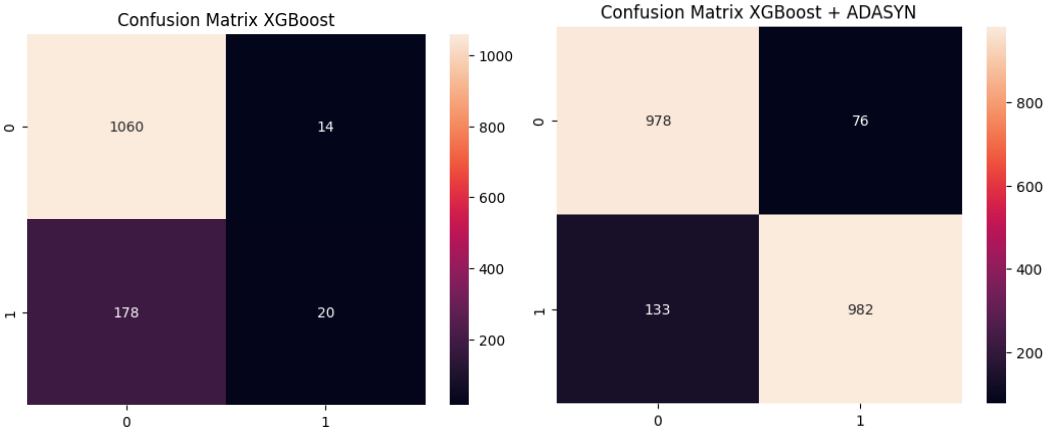
Pada tahap ini, analisis dilakukan menggunakan Confusion Matrix untuk mengevaluasi kinerja model XGBoost. Confusion Matrix menunjukkan berapa banyak prediksi model yang akurat (true positives dan true negatives) dan

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, Juni 2024 Page: 499-508
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5421

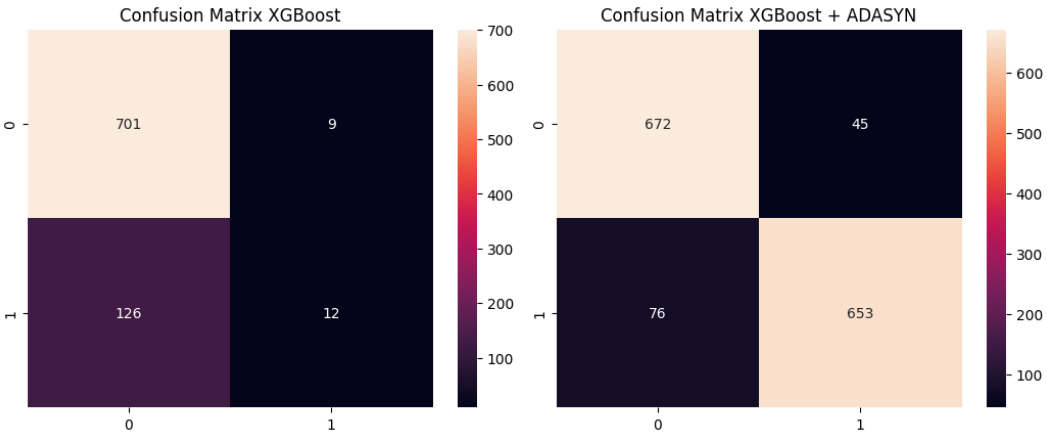
tidak akurat (false positives dan false negatives). Metrik evaluasi, termasuk akurasi, recall, presisi, dan F1-score, dapat dihitung menggunakan Confusion Matrix yang terlihat dari gambar 8 hingga gambar 10.



Gambar 8. Hasil Confusion Matrix rasio 60:40



Gambar 9. Hasil Confusion Matrix rasio 70:30



Gambar 10. Hasil Confusion Matrix rasio 80:20

4. KESIMPULAN

Penelitian ini menggunakan Heart Disease Dataset yang terdiri dari 15 atribut dan 4238 record, diambil dari platform Kaggle, untuk mengembangkan model Klasifikasi Penyakit Jantung Kardiovaskular dengan XGBoost. Beberapa tahapan penting dilakukan dalam penelitian ini untuk memaksimalkan performa model. Tahapan tersebut meliputi imputation data untuk menangani data yang hilang, transformation data menggunakan LabelEncoder, balancing data menggunakan ADASYN, splitting data dengan rasio 80% pelatihan dan 20% pengujian, serta hyperparameter tuning menggunakan Bayesian Optimization. Hasilnya penelitian menunjukkan bahwa, dibandingkan dengan model tanpa ADASYN, model XGBoost yang dikombinasikan dengan ADASYN memiliki kinerja lebih baik. Model dengan

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, Juni 2024 Page: 499-508
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5421

ADASYN mencapai ROC-AUC sebesar 0.971 dan akurasi sebesar 0.916, sementara model tanpa ADASYN hanya mencapai ROC-AUC sebesar 0.698 dan akurasi sebesar 0.841. Perbedaan ini menunjukkan bahwa penggunaan ADASYN untuk penyeimbangan data sangat efektif dalam meningkatkan performa model, terutama dalam situasi dataset yang tidak seimbang. Selain itu, penerapan Bayesian Optimization untuk tuning hyperparameter memainkan peran krusial dalam menemukan kombinasi parameter yang optimal, yang membantu meningkatkan kinerja model XGBoost lebih lanjut. Dengan menggabungkan teknik ADASYN untuk menangani ketidakseimbangan kelas dan Bayesian Optimization untuk meningkatkan tuning model, model dapat memanfaatkan data latih secara lebih efektif, sehingga meningkatkan akurasi prediksi dan kemampuan model untuk mengidentifikasi kasus penyakit jantung. Secara keseluruhan, penelitian ini menegaskan bahwa kombinasi ADASYN untuk menyeimbangkan data dan Bayesian Optimization untuk mengoptimalkan parameter model adalah kunci untuk membangun model klasifikasi penyakit jantung yang dapat diandalkan dan akurat.

REFERENCES

[1] N. L. K. A. Arsani, N. P. D. S. Wahyuni, N. N. M. Agustin, and M. Budiawan, "Deteksi Dini dan Pencegahan Penyakit Kardiovaskular," *Proceeding Senadimas Undiksha*, vol. 1, no. 1, pp. 663–668, 2022.

[2] J. P. Pane, L. Simorangkir, and P. I. S. B. Saragih, "Faktor-Faktor Risiko Penyakit Kardivaskular Berbasis Masyarakat," *J. Penelit. Perawat Prof.*, vol. 4, no. 4, pp. 1183–1192, 2022.

[3] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, pp. 1–16, 2020, doi: 10.1186/s12911-020-1023-5.

[4] W. Nugraha, "Prediksi Penyakit Jantung Cardiovascular Menggunakan Model Algoritma Klasifikasi," *J. Manag. dan Inform.*, vol. 9, no. 2, pp. 3–8, 2021.

[5] A. M. A. Rahim, Inggrid Yanuar Risca Pratiwi, and Muhammad Ainul Fikri, "Klasifikasi Penyakit Jantung Menggunakan Metode Synthetic Minority Over-Sampling Technique Dan Random Forest Clasifier," *Indones. J. Comput. Sci.*, vol. 12, no. 5, pp. 2995–3011, 2023, doi: 10.33022/ijcs.v12i5.3413.

[6] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized XGBoost based diagnostic system for effective prediction of heart disease," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 7, pp. 4514–4523, 2022, doi: 10.1016/j.jksuci.2020.10.013.

[7] K. Erdem, M. B. YILDIZ, E. T. YASIN, and M. Koklu, "A Detailed Analysis of Detecting Heart Diseases Using Artificial Intelligence Methods," *Intell. Methods Eng. Sci.*, no. December, 2023, doi: 10.58190/imiens.2023.71.

[8] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A Comparative Analysis of XGBoost," no. February, 2019, doi: 10.1007/s10462-020-09896-5.

[9] R. Gupta, H. Bansal, A. K. Singh, N. Bansal, and A. Saini, "An Efficient Prediction of Cardiovascular Diseases using Machine Learning Models," *2023 Int. Conf. Network, Multimed. Inf. Technol. NMITCON 2023*, no. Ml, pp. 1–6, 2023, doi: 10.1109/NMITCON58196.2023.10276141.

[10] H. Zheng, S. W. A. Sherazi, and J. Y. Lee, "A Stacking Ensemble Prediction Model for the Occurrences of Major Adverse Cardiovascular Events in Patients with Acute Coronary Syndrome on Imbalanced Data," *IEEE Access*, vol. 9, pp. 113692–113704, 2021, doi: 10.1109/ACCESS.2021.3099795.

[11] E. S. Ompusunggu, A. Nainggolan, and ..., "Penentuan Kelayakan Promosi Pegawai Menggunakan Algoritma Random Forest Classifier Dan Xgboost Classifier," *... (Teknik Inf. dan ...*, vol. 6, pp. 773–783, 2023, doi: 10.37600/tekinkom.v6i2.949.

[12] R. D. P. S. W. R. Naomi Nessyana Debataraja, "Penerapan Synthetic Minority Oversampling Technique Dalam Mengatasi Data Tidak Seimbang Pada Metode Classification and Regression Tree," *Bimaster Bul. Ilm. Mat. Stat. dan Ter.*, vol. 9, no. 1, pp. 231–238, 2020, doi: 10.26418/bbimst.v9i1.38949.

[13] N. P. Y. T. WIJAYANTI, E. N. KENCANA, and I. W. SUMARJAYA, "Smote: Potensi Dan Kekurangannya Pada Survei," *E-Jurnal Mat.*, vol. 10, no. 4, p. 235, 2021, doi: 10.24843/mtk.2021.v10.i04.p348.

[14] F. Y. Pamuji and S. D. A. Putri, "Komparasi Metode Smote Dan Adasyn Untuk Penanganan Data Tidak Seimbang Multiclass," *J. Inform. Polinema*, vol. 9, no. 3, pp. 331–338, 2023, doi: 10.33795/jip.v9i3.1330.

[15] R. A. Maula *et al.*, "Handling Missing Value dengan Pendekatan Regresi pada Dataset Akuakultur Berukuran Kecil," *J. Rekayasa Elektr.*, vol. 18, no. 3, pp. 175–184, 2022, doi: 10.17529/jre.v18i3.25903.

[16] S. Pushpalatha and A. Stella, "Kidney Disease Diagnosis using Classification Algorithm," *Proc. 5th Int. Conf. I-SMAC (IoT Soc. Mobile, Anal. Cloud), I-SMAC 2021*, pp. 1285–1288, 2021, doi: 10.1109/I-SMAC52330.2021.9640879.

[17] S. Doki, S. Devella, S. Tallam, S. S. Reddy Gangannagari, P. Sampathkrishna Reddy, and G. P. Reddy, "Heart Disease Prediction Using XGBoost," *Proc. 2022 3rd Int. Conf. Intell. Comput. Instrum. Control Technol. Comput. Intell. Smart Syst. ICICICT 2022*, pp. 1317–1320, 2022, doi: 10.1109/ICICICT54557.2022.9917678.

[18] N. N. Pandika Pinata, I. M. Sukarsa, and N. K. Dwi Rusjayanthi, "Prediksi Kecelakaan Lalu Lintas di Bali dengan XGBoost pada Python," *J. Ilm. Merpati (Menara Penelit. Akad. Teknol. Informasi)*, vol. 8, no. 3, p. 188, 2020, doi: 10.24843/jim.2020.v08.i03.p04.

[19] L. Qadrini, A. Sepperwali, and A. Aina, "Decision Tree Dan Adaboost Pada Klasifikasi Penerima Program Bantuan Sosial," *J. Inov. Penelit.*, vol. 2, no. 7, pp. 1959–1966, 2021.

[20] M. S. Mohosheu, F. Abrar Shams, M. A. Al Noman, S. R. Abir, and Al-Amin, "ROC Based Performance Evaluation of Machine Learning Classifiers for Multiclass Imbalanced Intrusion Detection Dataset," *8th Int. Conf. Recent Adv. Innov. Eng. Empower. Comput. Anal. Eng. Through Digit. Innov. ICRAIE 2023*, vol. 2023, pp. 1–6, 2023, doi: 10.1109/ICRAIE59459.2023.10468177.

[21] R. Suprayoga, S. Zega, Muhathir, and S. Mardiana, "Classification of Mango Leaf Diseases Using XGBoost Method and HoG Feature Extraction," *Proc. ICMERALDA 2023 - Int. Conf. Model. E-Information Res. Artif. Learn. Digit. Appl.*, pp. 197–202, 2023, doi: 10.1109/ICMERALDA60125.2023.10458172.