Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1398−1406
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6390

Model Prediksi Penyakit Jantung dengan Penanganan Outlier Menggunakan Interquartile Range dan Extreme Gradient Boosting

Lukman Azhari[1*], Novi Wulandari[2], Feru Adiningrat[3], Allan Desi Alexander[4]

[1]Program Studi Teknik Informatika, Universitas Muhammadiyah Tangerang, Tangerang
Jl. Perintis Kemerdekaan I No.33, Babakan, Cikokol, Kec. Tangerang, Kota Tangerang, Banten, Indonesia
[2]Program Studi Manajemen Informatika, STMIK Al Muslim, Bekasi
Jl. Raya Setu, Kp. Bahagia, Kec. Tambun Selatan, Kabupaten Bekasi, Jawa Barat, Indonesia
[3]Program Studi Sistem Informasi, STMIK Pranata Indonesia, Bekasi
Jl. Cut Mutia No.28, Margahayu, Kec. Bekasi Timur, Kota Bekasi, Jawa Barat, Indonesia
[4]Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Bhayangkara Jakarta Raya, Jakarta
Jl. Harsono RM No.67, Ragunan, Pasar Minggu, Kota Jakarta Selatan, Daerah Khusus Ibukota Jakarta, Indonesia
Email: [1,*]lukman.azhari@ft-umt.ac.id, [2]novi.wulandari@almuslim.ac.id,
[3]feru.adiningrat@gmail.com,[4]allan@ubharajaya.ac.id
Email Penulis Korespondensi: lukman.azhari@ft-umt.ac.id
Submitted: **03/12/2024;** Accepted: **31/01/2025;** Published: **31/01/2025**

Abstrak−Penyakit jantung merupakan salah satu penyebab kematian tertinggi di dunia, dengan tingkat prevalensi yang terus meningkat, termasuk di Indonesia. Keterlambatan deteksi dan diagnosis menjadi tantangan utama dalam penanganan penyakit ini, karena sebagian besar kasus baru teridentifikasi setelah pasien mengalami gejala serius atau serangan jantung. Data medis yang sering kali mengandung outlier dan noise menambah kompleksitas pengembangan model prediktif yang akurat. Penelitian ini bertujuan untuk mengembangkan model prediksi penyakit jantung menggunakan kombinasi metode Interquartile Range (IQR) untuk penanganan outlier dan algoritma Extreme Gradient Boosting (XGBoost) untuk pemodelan prediktif. Metode IQR diterapkan pada tahap pre-processing untuk mengidentifikasi dan menghilangkan outlier secara robust tanpa mengurangi integritas data, sementara XGBoost digunakan untuk membangun model prediksi yang efisien melalui pendekatan ensemble learning. Hasil penelitian menunjukkan peningkatan signifikan pada performa model, dengan akurasi meningkat dari 75.41% menjadi 89.47% dan AUC-ROC dari 0.8615 menjadi 0.9450. Model menunjukkan kemampuan prediksi yang seimbang dengan precision 95.24% dan recall 80.00% untuk kasus tidak ada penyakit, serta precision 86.11% dan recall 96.88% untuk kasus ada penyakit. Model yang dikembangkan memberikan kontribusi dengan meningkatkan kualitas data melalui penanganan outlier secara robust menggunakan metode IQR, membangun model prediksi yang lebih akurat dengan memanfaatkan keunggulan algoritma XGBoost dalam pendekatan ensemble learning.

Kata Kunci: Prediksi Penyakit Jantung; Interquartile Range; Extreme Gradient Boosting; XGBoost; Penanganan Outlier

Abstract−Heart disease remains one of the leading causes of death worldwide, with increasing prevalence rates, including in Indonesia. Delayed detection and diagnosis are the main challenges in treating this disease, as most cases are only identified after patients experience serious symptoms or heart attacks. Medical data often containing outliers and noise adds to the complexity of developing accurate predictive models. This study aims to develop a heart disease prediction model using a combination of the Interquartile Range (IQR) method for outlier handling and the Extreme Gradient Boosting (XGBoost) algorithm for predictive modeling. The IQR method is applied at the pre-processing stage to identify and eliminate outliers robustly without reducing data integrity, while XGBoost is used to build an efficient prediction model through an ensemble learning approach. The results showed significant improvements in model performance, with accuracy increasing from 75.41% to 89.47% and AUC-ROC from 0.8615 to 0.9450. The model demonstrates balanced predictive capabilities with precision of 95.24% and recall of 80.00% for cases without disease, and precision of 86.11% and recall of 96.88% for cases with disease. The developed model makes significant contributions by improving data quality through robust outlier handling using the IQR method, building a more accurate prediction model by leveraging the advantages of the XGBoost algorithm in the ensemble learning approach.

Keywords: Heart Disease Prediction; Interquartile Range; Extreme Gradient Boosting; XGBoost; Outlier Handling

1. PENDAHULUAN

Penyakit jantung tetap menjadi salah satu penyebab utama kematian di dunia. Menurut World Health Organization (WHO), pada tahun 2021, sekitar 20,5 juta orang meninggal akibat penyakit jantung, mencakup 32% dari seluruh kematian global [1]. Di Indonesia, tren ini juga menunjukkan peningkatan. Data Riset Kesehatan Dasar (Riskesdas) 2022 mencatat prevalensi penyakit jantung meningkat dari 1,5% pada 2018 menjadi 2,2% pada 2022, menegaskan pentingnya deteksi dini yang lebih efektif [2]. Mayoritas kasus baru teridentifikasi setelah pasien mengalami gejala serius atau serangan jantung, sehingga sistem deteksi dini yang cepat dan akurat sangat diperlukan [3]. Pendekatan konvensional dalam diagnosis penyakit jantung masih bergantung pada pemeriksaan klinis dan evaluasi manual, yang sering kali terlambat dalam mendeteksi risiko penyakit. Studi dari American Heart Association menunjukkan bahwa 45% kematian akibat serangan jantung terjadi sebelum pasien mencapai rumah sakit, menyoroti perlunya sistem deteksi dini berbasis teknologi [4].

Data mining dan machine learning telah digunakan secara luas dalam pengembangan sistem prediksi penyakit jantung, dengan berbagai algoritma menunjukkan tingkat akurasi yang beragam. Studi yang dilakukan dengan menggunakan algoritma K-Nearest Neighbors (KNN), khususnya dengan parameter K=5, mencapai

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1398−1406
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6390

performa dengan akurasi 64,03%, presisi 64,03%, dan recall 64,58% [5]. Penelitian lain yang mengimplementasikan algoritma C4.5 menunjukkan peningkatan signifikan dengan akurasi mencapai 79% [6]. Pendekatan Naïve Bayes Classifier dalam studi berikutnya, dengan tiga variasi pembagian data, berhasil mencapai akurasi tertinggi sebesar 83,1% [7]. Eksperimen dengan algoritma Random Forest awalnya menghasilkan akurasi 82,6087%, yang kemudian meningkat menjadi 85,058% setelah dioptimasi menggunakan teknik K-Fold Cross Validation dan GridSearchCV [8]. Pengembangan model menggunakan Support Vector Machine (SVM) juga menunjukkan hasil yang menjanjikan dengan akurasi mencapai 85% [9].

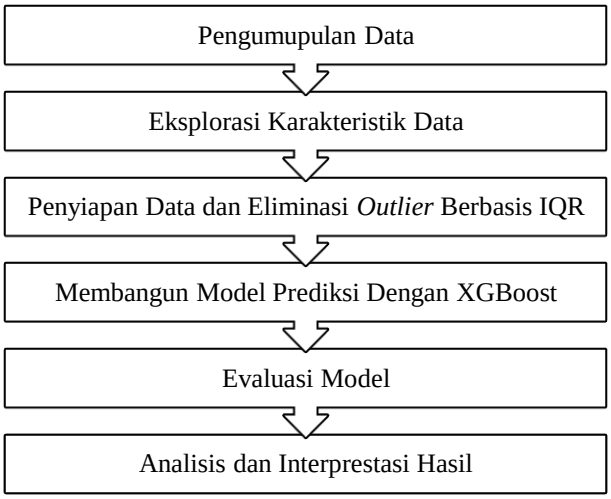
Namun, penelitian sebelumnya masih memiliki beberapa keterbatasan, seperti kurangnya penanganan terhadap outlier dalam data medis serta belum mengintegrasikan teknik boosting untuk meningkatkan akurasi prediksi. Outlier dan noise dalam data medis menjadi tantangan signifikan bagi akurasi prediksi [10]. Untuk mengatasi tantangan ini, metode Interquartile Range (IQR) digunakan sebagai teknik pra-pemrosesan untuk mendeteksi dan menangani outlier, sementara Extreme Gradient Boosting (XGBoost) diterapkan untuk meningkatkan akurasi model prediksi. Interquartile Range (IQR) hadir sebagai metode statistik yang robust untuk mendeteksi dan menangani outlier dalam dataset [11]. Metode ini bekerja dengan membagi data menjadi kuartil dan mengidentifikasi nilai-nilai di luar rentang yang dapat diterima [12]. Pendekatan IQR sangat relevan untuk mengatasi data outlier karena kemampuannya dalam mempertahankan integritas data sambil menghilangkan nilai-nilai ekstrem yang dapat menggangu proses analisis [13]. Di sisi lain, Extreme Gradient Boosting (XGBoost) telah menunjukkan performa yang unggul dalam berbagai tugas klasifikasi dan prediksi [14]. Algoritma ini menggunakan pendekatan ensemble learning yang mengkombinasikan multiple decision tree dengan teknik gradient boosting untuk menghasilkan model yang powerful dan efisien [15]. XGBoost menawarkan sejumlah keunggulan, termasuk kemampuannya untuk menangani nilai yang hilang (missing values) dengan efisien, sehingga meminimalkan dampak data yang tidak lengkap [16]. Selain itu, algoritma ini dilengkapi dengan teknik regularisasi yang efektif untuk mencegah overfitting, memastikan model tetap dapat menggeneralisasi dengan baik pada data yang tidak terlihat sebelumnya [17]. XGBoost juga mengoptimalkan penggunaan sumber daya komputasi, dengan cara yang efisien dalam hal memori dan waktu proses, memungkinkan penerapan model dalam skala besar dengan hasil yang cepat dan akurat [18].

Penelitian ini bertujuan untuk mengembangkan model prediksi penyakit jantung yang mengintegrasikan metode IQR untuk penanganan outlier dengan algoritma XGBoost untuk pemodelan prediktif. Model ini berkontribusi dalam meningkatkan kualitas data, menghasilkan prediksi yang lebih akurat, serta membantu tenaga medis dalam pengambilan keputusan klinis secara lebih efektif. Dengan demikian, model ini dapat diterapkan dalam sistem kesehatan guna mendukung upaya deteksi dini dan pencegahan penyakit jantung.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Tahapan penelitian ini dirancang secara sistematis dan terstruktur untuk memastikan pengembangan model prediksi penyakit jantung yang optimal [19]. Metode penelitian terdiri dari beberapa tahap terintegrasi dengan tujuan spesifik untuk mendukung pengembangan model. Pendekatan bertahap ini memungkinkan pemantauan dan validasi yang ketat pada setiap langkah pengembangan, sehingga menghasilkan model prediksi yang tidak hanya akurat tetapi juga dapat diandalkan dalam aplikasi praktis. Alur tahapan penelitian disajikan pada Gambar 1.



Gambar 1. Tahapan Penelitian

Gambar 1 menunjukkan alur penelitian secara keseluruhan, sementara rincian setiap langkah dalam prosedur penelitian dijelaskan lebih detail pada penjabaran berikut:

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1398−1406
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6390

1) Pengumupulan Data
Pada tahap pengumpulan data, penelitian ini menggunakan dataset penyakit jantung yang diperoleh dari platform Kaggle dengan nama "Heart Disease Dataset" [20]. Dataset ini terdiri dari 303 data pasien dengan 13 variabel input dan 1 variabel output. Variabel input meliputi usia (age), jenis kelamin (sex), tipe nyeri dada (cp), tekanan darah (trtbps), kolesterol (chol), gula darah puasa (fbs), hasil ECG istirahat (restecg), detak jantung maksimum (thalachh), angina akibat olahraga (exng), depresi ST (oldpeak), slope segmen ST (slp), jumlah pembuluh darah utama (caa), dan hasil thalassemia (thall). Sedangkan variabel output berupa status penyakit jantung yang dikategorikan dalam dua kelas: 0 untuk pasien tanpa penyakit jantung dan 1 untuk pasien dengan penyakit jantung. Dataset memiliki distribusi kelas yang cukup seimbang dengan 165 kasus positif (54,5%) dan 138 kasus negatif (45,5%).
2) Eksplorasi Karakteristik Data
Tahap eksplorasi karakteristik data melibatkan analisis mendalam terhadap properti dan pola dalam dataset [21]. Dilakukan analisis statistik deskriptif untuk memahami distribusi setiap variabel, identifikasi nilai-nilai unik, pemeriksaan korelasi antar variabel, dan visualisasi data untuk memperoleh wawasan awal. Tahap ini juga mencakup identifikasi potensi outlier, missing values, dan ketidakseimbangan kelas yang mungkin mempengaruhi performa model. Hasil dari eksplorasi ini akan menentukan strategi preprocessing yang tepat untuk tahap selanjutnya.
3) Penyiapan Data dan Eliminasi Outlier Berbasis IQR
Tahap ini berfokus pada penyiapan data agar siap digunakan untuk pemodelan, dengan penekanan khusus pada penanganan outlier menggunakan metode Interquartile Range (IQR). Proses dimulai dengan pembersihan data dasar seperti penanganan missing values dan standardisasi format data. Selanjutnya, metode IQR diterapkan untuk mengidentifikasi dan menangani outlier pada setiap fitur numerik. Nilai-nilai yang berada di luar rentang Q1 - 1.5 IQR hingga Q3 + 1.5 IQR diidentifikasi sebagai outlier dan ditangani sesuai dengan karakteristik data. Tahap ini juga mencakup normalisasi atau standardisasi fitur numerik dan encoding fitur kategorikal agar data siap untuk proses pemodelan.
4) Membangun Model Prediksi Dengan XGBoost
Pada tahap ini, dilakukan pengembangan model prediksi menggunakan algoritma XGBoost. Proses dimulai dengan pembagian dataset menjadi data training dan testing menggunakan stratified sampling untuk menjaga distribusi kelas. Model XGBoost dikonfigurasi dengan parameter awal dan dilakukan proses training menggunakan data latih. Tahap ini juga mencakup eksperimen dengan berbagai konfigurasi parameter untuk menemukan model yang optimal.
5) Evaluasi Model
Tahap evaluasi model melibatkan pengujian komprehensif terhadap performa model yang telah dikembangkan. Evaluasi dilakukan menggunakan berbagai metrik seperti accuracy, precision, recall, F1-score, dan ROC-AUC untuk mendapatkan gambaran lengkap tentang kemampuan model. Confusion matrix dianalisis untuk memahami pola kesalahan prediksi [22]. Validasi silang diterapkan untuk memastikan konsistensi performa model. Hasil evaluasi ini memberikan pemahaman mendalam tentang kekuatan dan keterbatasan model yang dikembangkan.
6) Analisis dan Interprestasi Hasil
Tahap akhir penelitian fokus pada analisis mendalam terhadap hasil yang diperoleh dan interpretasinya dalam konteks prediksi penyakit jantung. Dilakukan analisis terhadap feature importance untuk memahami kontribusi setiap variabel dalam proses prediksi. Interpretasi hasil mencakup pemahaman tentang kasus-kasus di mana model berhasil atau gagal dalam melakukan prediksi. Tahap ini juga melibatkan perbandingan performa model dengan penelitian-penelitian sebelumnya dan pembuatan rekomendasi untuk pengembangan lebih lanjut. Hasil analisis dan interpretasi ini menjadi dasar untuk penarikan kesimpulan dan penentuan kontribusi penelitian dalam bidang prediksi penyakit jantung.

2.2 Metode Interquartile Range (IQR)

Interquartile Range (IQR) adalah ukuran variabilitas statistik yang mengukur sebaran data antara kuartil pertama (Q1) dan kuartil ketiga (Q3) [11]. IQR memberikan gambaran tentang seberapa tersebar data tersebut, khususnya dengan fokus pada data yang berada di tengah distribusi, mengabaikan data ekstrem atau outlier [13]. Keunggulan utama IQR adalah kemampuannya mendeteksi outlier secara robust tanpa terpengaruh oleh nilai ekstrem yang dapat mendistorsi analisis statistik tradisional [23].

Pendekatan IQR berguna untuk mengidentifikasi dan menangani titik data yang secara statistik signifikan berbeda dari mayoritas dataset. IQR dihitung menggunakan persamaan (1).

$$\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1} \quad (1)$$

di mana Q1 adalah kuartil pertama (persentil ke-25), sedangkan Q3 adalah kuartil ketiga (persentil ke-75).

Untuk mengidentifikasi outlier, digunakan bawah (Lower Fence) dan batas atas (Upper Fence) dengan persamaan (2) dan persamaan (3).

$$\text{Lower Fence} = \mathrm{Q1} - (1{,}5 \times \mathrm{IQR}) \quad (2)$$

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1398−1406
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6390

$$\text{Upper Fence} = Q3 + (1{,}5 \times \text{IQR}) \quad (3)$$

Dalam identifikasi outlier, data dianggap menyimpang apabila nilainya berada di luar rentang Lower Fence dan Upper Fence, dengan menggunakan faktor pengali 1.5 yang secara empiris terbukti efektif dalam mendeteksi data yang menyimpang pada distribusi mendekati normal.

2.3 Pendekatan Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) adalah algoritma machine learning yang menggunakan prinsip gradient boosting dengan optimasi lanjutan [24]. Secara prinsip, boosting adalah teknik yang membangun model secara berurutan, di mana setiap model yang baru berusaha untuk memperbaiki kesalahan yang dibuat oleh model sebelumnya [25]. XGBoost mengoptimalkan proses boosting dengan beberapa teknik tambahan yang meningkatkan efisiensi dan akurasi model. XGBoost membangun model secara sekuensial melalui ensemble dari weak learners (umumnya decision trees) untuk membentuk strong learner.

XGBoost memiliki beberapa keunggulan yang membuatnya menjadi algoritma yang handal dalam tugas klasifikasi. Algoritma ini dilengkapi dengan kemampuan menangani missing values secara otomatis tanpa memerlukan pra-pemrosesan khusus, serta memiliki mekanisme pencegahan overfitting melalui teknik regularisasi yang terintegrasi [17]. Dalam hal kinerja komputasi, XGBoost mengimplementasikan optimasi paralel yang memungkinkan pemrosesan data lebih cepat dan efisien [18]. Selain itu, algoritma ini menerapkan teknik tree pruning yang memungkinkan pemangkasan cabang pohon keputusan yang tidak efektif, sehingga menghasilkan model yang lebih efisien dan optimal dalam penggunaan sumber daya komputasi.

XGBoost membangun model prediksi melalui kombinasi bertahap dari weak learners, yang direpresentasikan dalam persamaan (4).

$$\hat{y_i} = \sum_t f_t(x_i) \quad (4)$$

di mana y_i menunjuukan prediksi untuk instance i, f_t merupakan weak learner ke- t, x_i merupakan fitur input, dan t menunjukkan jumlah total trees

Pada XGBoost, fungsi objektif terdiri dari dua komponen utama. Komponen pertama adalah fungsi loss, yang digunakan untuk mengevaluasi seberapa baik model memprediksi data latih. Komponen kedua adalah fungsi regularisasi, yang bertugas mengontrol kompleksitas model guna menghindari overfitting. Formulasi fungsi objektif XGBoost dirumuskan dalam persamaan (5).

$$L = \sum_i l(y_i, \hat{y}_i) + \sum_t \Omega(f_t) \quad (5)$$

di mana l merupakan fungsi loss, Ω menunjukkan fungsi regularisasi, y_i adalah nilai actual, dan $\hat{y}_i$ adalah nilai prediksi.

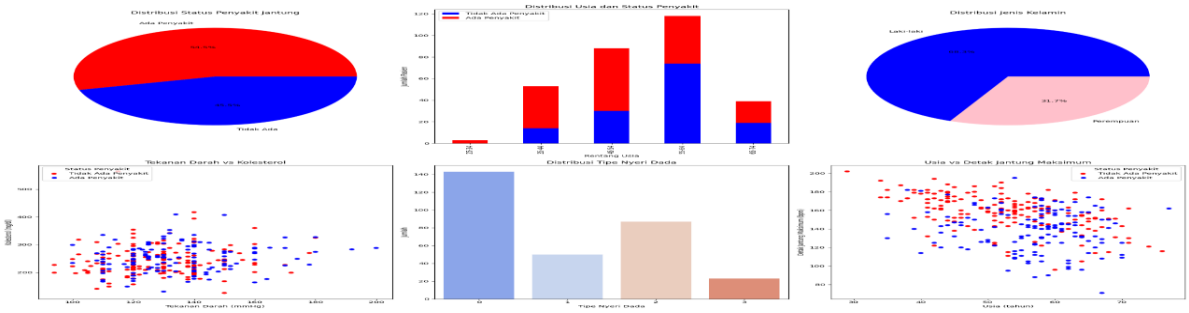
Fungsi regularisasi pada XGBoost bertujuan untuk mengendalikan kompleksitas model agar terhindar dari overfitting. Fungsi ini dirumuskan dalam persamaan (6).

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \quad (6)$$

di mana Ω(f) merupakan fungsi regularisasi untuk pohon keputusan f, γ menunjukkan kompleksitas tree, T menunjukkan jumlah leaf, λ merupakan parameter regularisasi, dan w merupakan leaf weights.

3. HASIL DAN PEMBAHASAN

Untuk mengembangkan model prediksi penyakit jantung, langkah pertama adalah mempersiapkan dataset yang digunakan untuk proses pelatihan dan pengujian. Penelitian ini menggunakan Heart Disease Dataset dari platform Kaggle yang berisi data klinis 303 pasien dengan 14 variabel untuk memprediksi kemungkinan terjadinya penyakit jantung [20]. Dataset ini mencakup berbagai parameter medis yang relevan untuk analisis risiko penyakit jantung. Tahap berikutnya adalah eksplorasi karakteristik data untuk memperoleh pemahaman mendalam tentang pola dan hubungan antar variabel. Hasil visualisasi dari distribusi dan hubungan berbagai variabel tersaji pada Gambar 2.

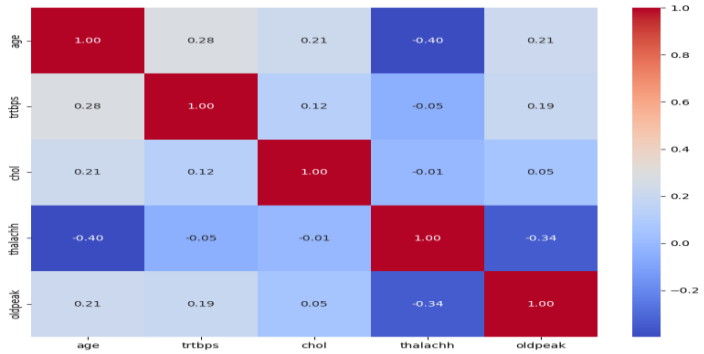


Gambar 2. Visualisasi Distribusi dan Hubungan Variabel Dengan Status Penyakit Jantung

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1398−1406
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6390

Visualisasi hasil eksplorasi data ditunjukkan pada Gambar 2, yang mengungkapkan beberapa karakteristik penting dataset. Distribusi status penyakit menunjukkan 54.5% pasien memiliki penyakit jantung dan 45.5% tidak memiliki penyakit jantung, hal ini mengindikasikan dataset yang cukup seimbang sehingga mengurangi risiko bias dalam pengembangan model. Analisis distribusi usia mengidentifikasi kelompok usia 55-64 tahun sebagai kelompok dengan frekuensi tertinggi dan dominasi kasus positif penyakit jantung, menunjukkan rentang usia kritis untuk pemantauan kesehatan jantung. Dari aspek gender, terdapat ketidakseimbangan yang signifikan dengan dominasi pasien laki-laki (68.3%) dibanding perempuan (31.7%), mengindikasikan potensi faktor risiko yang berbeda antar gender. Visualisasi hubungan tekanan darah dengan kolesterol mengungkapkan adanya outlier signifikan, terutama pada nilai kolesterol di atas 400 mg/dl, yang memerlukan penanganan khusus dalam tahap preprocessing data. Distribusi tipe nyeri dada menunjukkan variasi gejala dalam empat kategori dengan tipe 0 sebagai manifestasi yang paling umum. Scatter plot usia terhadap detak jantung maksimum memperlihatkan tren penurunan kapasitas jantung seiring pertambahan usia. Pola-pola data ini dapat menjadi dasar untuk strategi preprocessing dan pemilihan fitur yang tepat dalam pengembangan model prediksi penyakit jantung.

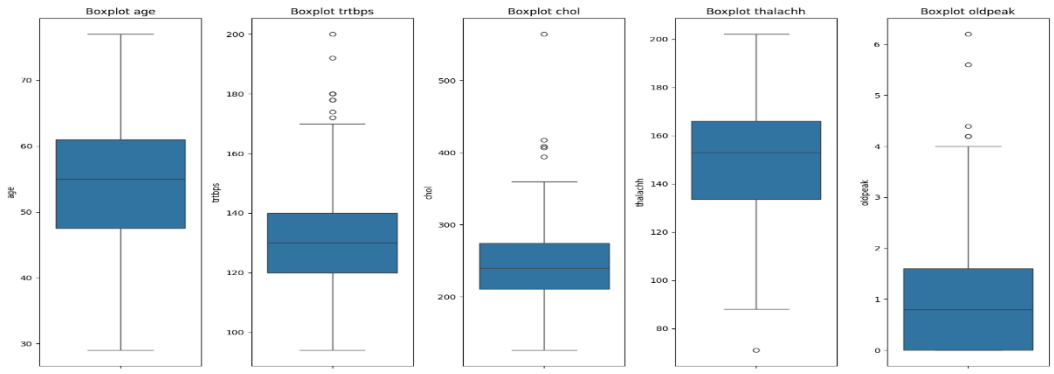
Selanjutnya dilakukan eksplorasi data untuk menganalisis korelasi antar variabel numerik, hal ini penting untuk memahami keterkaitan antara berbagai parameter medis dan mengidentifikasi potensi multikolinearitas dalam dataset yang dapat mempengaruhi performa model prediksi. Hasil visualisasi dalam bentuk heatmap mengenai korelasi antar variabel pada dataset ini ditampilkan pada Gambar 3.



Gambar 3. Visualisasi Korelasi Antar Variabel Numerik

Heatmap pada Gambar 3 mengungkapkan korelasi terkuat terdapat antara usia dan detak jantung maksimum dengan nilai -0.40, mengindikasikan hubungan negatif yang signifikan. Detak jantung maksimum juga menunjukkan korelasi negatif dengan depresi ST sebesar -0.34. Tekanan darah memiliki korelasi positif moderat dengan usia (0.28), sementara kolesterol menunjukkan korelasi yang relatif lemah dengan variabel lainnya. Pola korelasi ini memberikan wawasan penting untuk pemilihan fitur dan pengembangan model prediksi penyakit jantung yang lebih akurat.

Berikutnya dilakukan eksplorasi data untuk mengidentifikasi dan menganalisis outlier pada dataset, hal ini penting untuk memahami kualitas data dan menentukan strategi preprocessing yang tepat karena outlier dapat mempengaruhi performa model prediksi. Hasil visualisasi dalam bentuk boxplot yang menunjukkan distribusi data dan keberadaan outlier pada lima variabel numerik dapat dilihat pada Gambar 4.

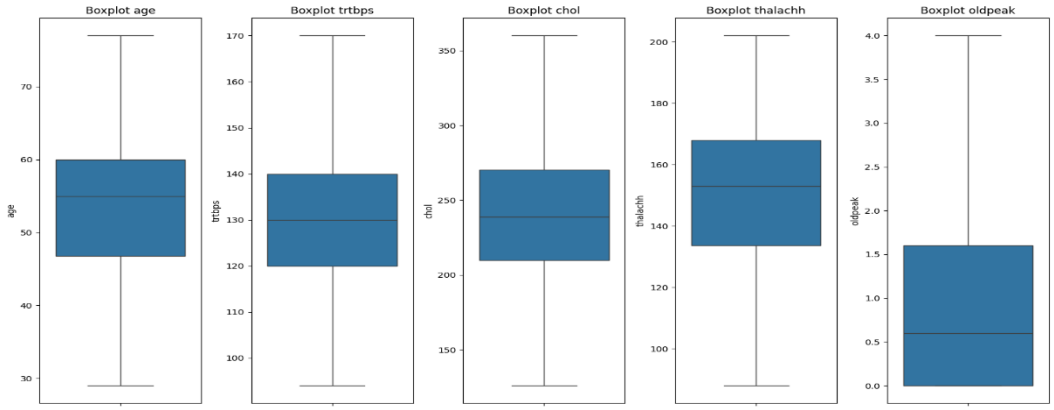


Gambar 4. Boxplot Distribusi Data dan Keberadaan Outlier Pada Varibel Numerik

Visualisasi boxplot pada Gambar 4 menunjukkan distribusi dan outlier pada lima variabel numerik dalam dataset. Variabel kolesterol (chol) memiliki outlier paling signifikan dengan beberapa titik di atas 400 mg/dl, jauh di atas batas atas normal. Pada variabel tekanan darah (trtbps), terlihat beberapa outlier di atas 180 mmHg. Variabel thalachh (detak jantung maksimum) menunjukkan satu outlier di bawah 80 bpm. Variabel depresi ST (oldpeak) memiliki beberapa outlier di atas 4.0. Sementara variabel usia (age) menunjukkan distribusi yang relatif normal dengan rentang 29-77 tahun tanpa outlier yang signifikan. Visualisasi ini mengindikasikan perlunya penanganan

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1398−1406
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6390

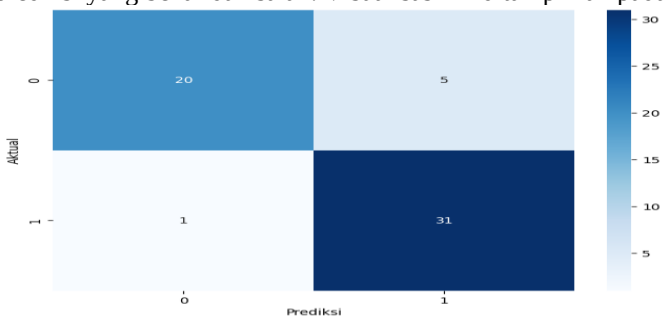
outlier, terutama pada variabel kolesterol dan tekanan darah, sebelum melakukan pemodelan. Untuk menghilangkan outlier digunakan teknik Interquartile Range (IQR). IQR bekerja dengan menghitung rentang antara kuartil pertama (Q1) dan kuartil ketiga (Q3), kemudian menentukan batas atas dan bawah menggunakan formula: batas bawah = Q1 - 1.5 × IQR dan batas atas = Q3 + 1.5 × IQR. Data yang berada di luar rentang tersebut dianggap sebagai outlier dan dihapus dari dataset. Hasil boxplot setelah penerapan IQR tersaji pada Gambar 5.



Gambar 5. Boxplot Distribusi Data Setelah Menggunakan Teknik Interquartile Range (IQR)

Visualisasi pada Gambar 5 merupakan boxplot penerapan IQR yang menunjukkan distribusi data yang lebih seimbang dan terpusat tanpa adanya outlier. Variabel kolesterol (chol) yang sebelumnya memiliki outlier ekstrem di atas 400 mg/dl kini memiliki rentang nilai yang lebih normal antara 150-350 mg/dl. Tekanan darah (trtbps) juga menunjukkan distribusi yang lebih baik dengan rentang 95-170 mmHg. Detak jantung maksimum (thalachh) dan depresi ST (oldpeak) menunjukkan rentang nilai yang lebih teratur tanpa adanya nilai ekstrem. Variabel usia (age) tetap menunjukkan distribusi yang relatif sama karena sebelumnya memang tidak memiliki outlier yang signifikan. Penanganan outlier ini mengurangi jumlah data dari 303 menjadi 235 instances, namun menghasilkan dataset yang lebih representatif dan sesuai untuk pemodelan.

Langkah selanjutnya adalah membangun model prediksi menggunakan XGBoost, dimulai dengan membagi dataset menjadi data pelatihan dan pengujian dengan rasio 80:20. Pembagian ini dilakukan menggunakan fungsi ‘train_test_split’, di mana parameter stratify digunakan untuk menjaga keseimbangan distribusi kelas. Model XGBoost diinisialisasi dengan parameter ‘use_label_encoder=False’ untuk menghindari warning terkait label encoding, eval_metric='logloss' untuk metrik evaluasi, dan random_state=42 untuk reproducibility. Setelah inisialisasi, model dilatih menggunakan metode fit() dengan data training. Model kemudian melakukan prediksi pada data testing menggunakan predict() untuk kelas output dan predict_proba() untuk probabilitas prediksi. Selanjutnya model yang telah dibangun diuji menggunakan data testing untuk mengukur kemampuannya dalam melakukan prediksi penyakit jantung. Performa model divisualisasikan menggunakan confusion matrix, yang menggambarkan jumlah prediksi yang benar dan salah. Visualisasi ini ditampilkan pada Gambar 6.



Gambar 6. Confusion Matrix Model Prediksi Yang Dikembangkan

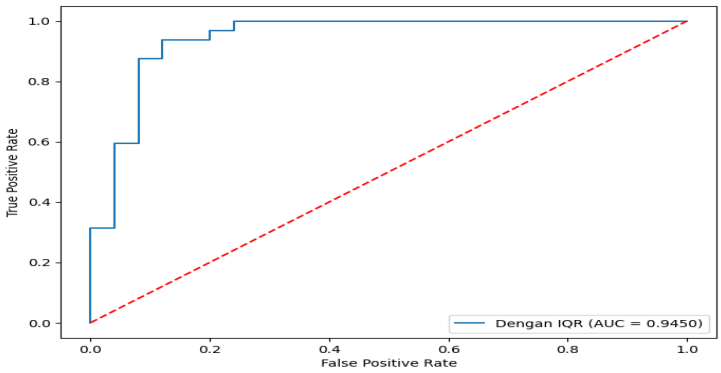
Gambar 6 menampilkan confusion matrix yang menunjukkan performa model yang dikembangkan dalam memprediksi penyakit jantung. Dari 57 total kasus, model berhasil memprediksi dengan benar 51 kasus (20 kasus tidak ada penyakit dan 31 kasus ada penyakit), sementara 6 kasus diprediksi salah (5 false positive dan 1 false negative). Berdasarkan hasil confusion matrix ini, dilakukan perhitungan metrik evaluasi model yang meliputi precision, recall, F1-score, dan accuracy. Hasil perhitungan metrik evaluasi tersebut ditampilkan pada Tabel 1.

Tabel 1. Hasil Evaluasi Model

Kelas	Precision	Recall	F1-Score	Accuracy
Tidak Ada Penyakit	0.9524	0.8000	0.8696	0.8947
Ada Penyakit	0.8611	0.9688	0.9118	

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1398−1406
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6390

Tabel 1 menunjukkan hasil evaluasi performa model yang baik dengan akurasi keseluruhan 89.47%. Model lebih akurat dalam mendeteksi pasien dengan penyakit jantung (precision 86.11%, recall 96.88%) dibanding tanpa penyakit jantung (precision 95.24%, recall 80.00%). F1-score yang tinggi untuk kedua kelas (86.96% dan 91.18%) mengindikasikan keseimbangan yang baik antara precision dan recall. Tingginya recall untuk kelas penyakit jantung (96.88%) menunjukkan model sangat efektif dalam mengidentifikasi kasus positif, yang kritis untuk diagnosis medis. Selanjutnya model juga dievaluasi menggunakan ROC Curve, agar dapat mengukur kemampuan model dalam membedakan antara kelas dengan menunjukkan hubungan antara True Positive Rate (TPR) dan False Positive Rate (FPR) pada berbagai threshold. Hasil ROC Curve disajikan pada Gambar 7.



Gambar 6. ROC Curve Model Prediksi Yang Dikembangkan

Gambar 6 menunjukkan performa model yang sangat baik dengan nilai AUC 0.9450. Kurva biru yang jauh di atas garis merah putus-putus (baseline) mengindikasikan kemampuan model yang kuat dalam membedakan kelas positif dan negatif. Kenaikan tajam di awal kurva menunjukkan model dapat mencapai true positive rate yang tinggi dengan false positive rate yang rendah. Hal ini menegaskan bahwa model dengan penanganan outlier menggunakan IQR dan XGBoost memiliki kemampuan diskriminasi yang sangat baik dalam memprediksi penyakit jantung.

Untuk menganalisis pengaruh penerapan metode IQR terhadap performa model XGBoost, dilakukan perbandingan akurasi dan nilai AUC-ROC antara model tanpa IQR dan model dengan IQR. Hasil perbandingan ini disajikan dalam Tabel 2.

Tabel 2. Perbandingan Performa Model XGBoost dengan dan Tanpa Penanganan Outlier

Kelas	Accuracy	AUC-ROC
XGBoost	0,7541	0,8615
IQR + XGBoost	0,8947	0,9450

Hasil perbandingan menunjukkan peningkatan performa yang signifikan setelah penerapan IQR untuk penanganan outlier. Model XGBoost dengan IQR mencapai akurasi 89.47% dan AUC-ROC 0.9450, meningkat dari model tanpa IQR yang hanya mencapai akurasi 75.41% dan AUC-ROC 0.8615. Peningkatan sekitar 14,06% pada akurasi dan 8,35% pada AUC-ROC mengindikasikan bahwa penanganan outlier menggunakan IQR efektif dalam meningkatkan kemampuan model untuk memprediksi penyakit jantung dengan lebih akurat dan reliabel.

Berdasarkan seluruh hasil evaluasi, kombinasi Interquartile Range (IQR) dan Extreme Gradient Boosting (XGBoost) menunjukkan efektivitas yang signifikan dalam pengembangan model prediksi penyakit jantung. Model menunjukkan performa yang seimbang untuk kedua kelas dengan nilai precision dan recall yang tinggi, serta nilai AUC-ROC 0.9450 yang mengindikasikan kemampuan diskriminasi yang sangat baik. IQR efektif dalam menangani outlier karena sifatnya yang robust dan tidak terpengaruh oleh nilai ekstrem, memungkinkan pembersihan data yang lebih akurat dengan mempertahankan pola data yang sebenarnya. Metode ini berhasil mengidentifikasi dan menghilangkan noise dalam data medis yang dapat menggangu proses pembelajaran model. Di sisi lain, XGBoost unggul karena kemampuannya dalam menangani data yang kompleks melalui pendekatan ensemble learning yang mengkombinasikan multiple decision tree. Algoritma ini memiliki mekanisme regularisasi built-in untuk mencegah overfitting, kemampuan menangani missing values, dan optimasi gradient boosting yang efisien. Kombinasi kedua metode ini menciptakan pipeline yang kuat dimana data yang telah dibersihkan dari outlier memungkinkan XGBoost untuk lebih efektif dalam mempelajari pola yang sebenarnya dalam data, menghasilkan model prediksi yang lebih akurat dan reliable.

Namun, terdapat beberapa aspek yang masih perlu ditingkatkan, seperti recall untuk kasus negatif (0.8000) yang masih lebih rendah dibanding kasus positif (0.9688), serta berkurangnya jumlah data training setelah penghapusan outlier. Untuk pengembangan lebih lanjut, beberapa rekomendasi yang dapat dilakukan meliputi implementasi teknik augmentasi data untuk meningkatkan jumlah sampel, eksplorasi metode penanganan outlier alternatif seperti winsorization yang tidak mengurangi jumlah data, penerapan cross-validation untuk validasi model yang lebih komprehensif, serta optimasi hyperparameter XGBoost untuk meningkatkan performa model.

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1398−1406
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6390

4. KESIMPULAN

Penelitian ini telah berhasil mengembangkan model prediksi penyakit jantung menggunakan kombinasi Interquartile Range (IQR) dan Extreme Gradient Boosting (XGBoost) dengan hasil yang sangat baik. Integrasi penggunaan IQR untuk penanganan outlier dan model prediksi dengan XGBoost terbukti efektif dalam meningkatkan kualitas data, yang ditunjukkan dengan peningkatan performa model dari akurasi 75.41% menjadi 89.47% (meningkat 14.06%) dan AUC-ROC dari 0.8615 menjadi 0.9450 (meningkat 8.35%). Model menunjukkan kemampuan prediksi yang seimbang untuk kedua kelas, dengan precision 95.24% dan recall 80.00% untuk kasus tidak ada penyakit, serta precision 86.11% dan recall 96.88% untuk kasus ada penyakit. Tingginya nilai AUC-ROC (0.9450) mengindikasikan kemampuan diskriminasi yang sangat baik dalam membedakan antara pasien dengan dan tanpa penyakit jantung. Performa ini dikarenakan IQR memiliki kemampuan dalam mendeteksi dan menangani outlier secara robust tanpa terpengaruh nilai ekstrem, sementara XGBoost dapat mengoptimalkan pembelajaran dari data yang telah dibersihkan melalui mekanisme ensemble learning dan regularisasi yang efektif. Meskipun demikian, untuk penelitian kedepan perlu mempertimbangkan beberapa aspek, diantaranya karena IQR melakukan penghapusan outlier sehingga berkurangnya jumlah data training, untuk itu dapat mempertimbangkan implementasi teknik augmentasi data untuk meningkatkan jumlah sampel, eksplorasi metode penanganan outlier alternatif seperti winsorization yang tidak mengurangi jumlah data, penerapan cross-validation untuk validasi model yang lebih komprehensif, serta optimasi hyperparameter XGBoost untuk meningkatkan performa model lebih lanjut.

REFERENCES

[1] W. L. N. Husain, S. Buraena, R. F. Syamsu, N. Nurmadilla, and A. F. Arsal, “Gambaran Faktor Risiko Penyakit Jantung Koroner Akut Di RSUD Aloe Saboe Gorontalo,” Indones. J. Heal., vol. 2, no. 03, pp. 162–173, 2022, doi: 10.33368/inajoh.v2i03.75.
[2] S. N. Tarmizi, “Kenali Gejala Jantung Sejak Dini,” Kemetrian Kesehatan. [Online]. Available: https://kemkes.go.id/id/rilis-kesehatan/kenali-gejala-jantung-sejak-dini
[3] M. Melyani, L. N. Tambunan, and E. P. Baringbing, “Hubungan Usia dengan Kejadian Penyakit Jantung Koroner pada Pasien Rawat Jalan di RSUD dr. Doris Sylvanus Provinsi Kalimantan Tengah,” J. Surya Med., vol. 9, no. 1, pp. 119–125, 2023, doi: 10.33084/jsm.v9i1.5158.
[4] K. Astle, “Experiencing pain after a heart attack may predict long-term survival,” American Heart Association. [Online]. Available: https://newsroom.heart.org/news/experiencing-pain-after-a-heart-attack-may-predict-long-term-survival
[5] A. Yogianto, A. Homaidi, and Z. Fatah, “Implementasi Metode K-Nearest Neighbors (KNN) untuk Klasifikasi Penyakit Jantung,” G-Tech J. Teknol. Terap., vol. 8, no. 3, pp. 1720–1728, 2024, doi: 10.33379/gtech.v8i3.4495.
[6] A. Sepharni, I. E. Hendrawan, and C. Rozikin, “Klasifikasi Penyakit Jantung dengan Menggunakan Algoritma C4.5,” STRING (Satuan Tulisan Ris. dan Inov. Teknol., vol. 7, no. 2, p. 117, 2022, doi: 10.30998/string.v7i2.12012.
[7] D. Cahya Putri Buani, “Penerapan Algoritma Naïve Bayes dengan Seleksi Fitur Algoritma Genetika Untuk Prediksi Gagal Jantung,” Evolusi J. Sains dan Manaj., vol. 9, no. 2, pp. 43–48, 2021, doi: 10.31294/evolusi.v9i2.11141.
[8] E. Edric and S. P. Tamba, “Prediksi Penyakit Gagal Jantung Dengan Menggunakan Random Forest,” J. Sist. Inf. dan Ilmu Komput. Prima(JUSIKOM PRIMA), vol. 5, no. 2, pp. 176–181, 2022, doi: 10.34012/jurnalsisteminformasidanilmukomputer.v5i2.2445.
[9] A. Putranto, N. L. Azizah, and A. I. Ratna Ika, “Sistem Prediksi Penyakit Jantung Berbasis Web Menggunakan Metode SVM dan Framework Streamlit,” J. Penerapan Sist. Inf. (Komputer Manajemen), vol. 4, no. 2, pp. 442–452, 2023, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+disease
[10] A. Razaki, Y. H. Chrisnanto, and M. Melina, “Penanganan Outlier Pada Metode Algoritma K- Nearest Neighbors (KNN) Dengan Metode Kernel Density Estimation Pada Kasus Penyakit Diabetes,” INTECOMS J. Inf. Technol. Comput. Sci., vol. 7, no. 4, pp. 1177–1188, 2024, doi: 10.31539/intecoms.v7i4.10866.
[11] R. Efendi, A. Junaidi, and A. M. Rizki, “Penentuan Pusat Klaster Secara Otomatis Pada Algoritma Density Peaks Clustering Berbasis Metode Inter Quartile Range,” J. Inform. dan Tek. Elektro Terap., vol. 12, no. 3, 2024, doi: 10.23960/jitet.v12i3.4997.
[12] M. Nijhuis and I. van Lelyveld, “Outlier Detection with Reinforcement Learning for Costly to Verify Data,” Entropy, vol. 25, no. 6, pp. 1–17, 2023, doi: 10.3390/e25060842.
[13] V. Magar, D. Ruikar, S. Bhoite, and R. Mente, “Innovative Inter Quartile Range-based Outlier Detection and Removal Technique for Teaching Staff Performance Feedback Analysis,” J. Eng. Educ. Transform., vol. 37, no. 3, pp. 176–184, 2024, doi: 10.16920/jeet/2024/v37i3/24013.
[14] M. D. Maulana, A. I. Hadiana, and F. R. Umbara, “Algoritma Xgboost Untuk Klasifikasi Kualitas Air Minum,” JATI (Jurnal Mhs. Tek. Inform., vol. 7, no. 5, pp. 3251–3256, 2024, doi: 10.36040/jati.v7i5.7308.
[15] F. A. P. Prasetya and P. H. P. Rosa, “Klasifikasi Kegagalan Pembayaran Kredit Nasabah Bank dengan Algoritma XGBoost,” in Seminar Nasional Informatika Bela Negara (SANTIKA), 2024, pp. 110–115.
[16] D. T. Murdiansyah, “Prediksi Stroke Menggunakan Extreme Gradient Boosting,” JIKO (Jurnal Inform. dan Komputer), vol. 8, no. 2, p. 419, 2024, doi: 10.26798/jiko.v8i2.1295.
[17] M. Salsabil, N. Lutvi, and A. Eviyanti, “Implementasi Data Mining dalam Melakukan Prediksi Penyakit Diabetes Menggunakan Metode Random Forest dan XGBoost,” J. Ilm. KOMPUTASI, vol. 23, no. 1, pp. 51–58, 2024.
[18] D. Kurnia, M. Itqan Mazdadi, D. Kartini, R. Adi Nugroho, and F. Abadi, “Seleksi Fitur dengan Particle Swarm Optimization pada Klasifikasi Penyakit Parkinson Menggunakan XGBoost,” J. Teknol. Inf. dan Ilmu Komput., vol. 10, no. 5, pp. 1083–1094, 2023, doi: 10.25126/jtiik.20231057252.
[19] R. I. Borman, R. Napianto, N. Nugroho, D. Pasha, Y. Rahmanto, and Y. E. P. Yudoutomo, “Implementation of PCA and

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1398−1406
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6390

KNN Algorithms in the Classification of Indonesian Medicinal Plants," in International Conference on Computer Science, Information Technology and Electrical Engineering (ICOMITEE), IEEE, 2021, pp. 46–50.

[20] M. Yasser, "Heart Disease Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/yasserh/heart-disease-dataset

[21] R. I. Borman and M. Wati, "Penerapan Data Maining Dalam Klasifikasi Data Anggota Kopdit Sejahtera Bandarlampung Dengan Algoritma Naïve Bayes," J. Ilm. Fak. Ilmu Komput., vol. 9, no. 1, pp. 25–34, 2020.

[22] R. I. Borman, F. Rossi, D. Alamsyah, R. Nuraini, and Y. Jusman, "Classification of Medicinal Wild Plants Using Radial Basis Function Neural Network with Least Mean Square," in International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS), IEEE, 2022.

[23] Fredianto and D. A. P. Putri, "Comparison of the interquartile range algorithm and local outlier factor on Australian weather data sets," AIP Conf. Proc., vol. 2727, no. 1, p. 40010, Jun. 2023, doi: 10.1063/5.0141897.

[24] M. Ridwansyah and H. Zakaria, "Implementasi Algortima Gradient Boosting Pada Aplikasi Hutang Piutang Perorangan Secara Berbasis Web Untuk Meningkatan Akurasi Prediksi Pelunasan Hutang (Studi Kasus : PT Naila Kreasi Mandiri)," JURIHUM J. Inov. dan Hum., vol. 1, no. 4, pp. 440–451, 2023.

[25] A. F. L. Ptr, M. M. Siregar, and I. Daniel, "Analysis of Gradient Boosting, XGBoost, and CatBoost on Mobile Phone Classification," J. Comput. Networks, Archit. High Perform. Comput., vol. 6, no. 2, pp. 661–670, 2024.