JURNAL ILMIAH BETRIK
Besemah Teknologi Informasi dan Komputer
Editor Office : LPPM Sekolah Tinggi Teknologi Pagar Alam, Jln. Masik Siagim No. 75 Simpang Mbacang, Pagar Alam, SUM-SEL, Indonesia
Phone : +62 852-7901-1390.
Email : betrik@sttpagaralam.ac.id | admin.jurnal@sttpagaralam.ac.id
Website : https://ejournal.sttpagaralam.ac.id/index.php/betrik/index
ISSN : 2339 - 1871

PERBANDINGAN ALGORITMA *RANDOM FOREST* DAN *XGBOOST* UNTUK KLASIFIKASI PENYAKIT PARU-PARU BERDASARKAN DATA DEMOGRAFI PASIEN

Risky Harahap[1], M. Irpan[2], M. Azzuhri Dinata[3], Lusiana Efrizoni [4], Rahmaddeni[5]
Program Studi Teknik Informatika Universitas Sains Dan Teknologi Indonesia[12345]
Jalan Purwodadi No. KM, 10, Sidomulyo Barat, Tampan, Pekanbaru-Riau 28294, Indonesia
Surel : 2110031802097@sar.ac.id[1], 2110031802100@sar.ac.id[2], 2110031802095@sar.ac.id[3], lusiana@stmik-amik-riau.ac.id[4], rahmaddeni@sar.ac.id[5]

Abstrak: Dalam penelitian ini, algoritma *Random Forest* dan *XGBoost* dibandingkan dalam klasifikasi penyakit paru-paru menggunakan data demografi pasien. Dataset yang digunakan terdiri dari 30.000 data pasien dengan 9 atribut dan 1 label yang diambil dari Kaggle. Tahapan penelitian termasuk pengumpulan data, *Preprocessing*, pembagian data, dan klasifikasi data menggunakan kedua algoritma. Hasil menunjukkan bahwa algoritma XGBoost memiliki akurasi 94% dan *AUC* 0.98, sedangkan algoritma *Random Forest* memiliki akurasi 91% dan *AUC* 0.97. Meskipun *Random Forest* lebih cepat dan lebih mudah diinterpretasikan, XGBoost bekerja lebih baik dengan data yang kompleks dengan hasil yang lebih konsisten. Melalui penggunaan teknik regularisasi dan penanganan outliers yang lebih baik, *XGBoost* juga dapat mengatasi masalah overfitting dengan lebih baik. Studi ini memberikan panduan untuk peneliti dan praktisi dalam memilih algoritma terbaik untuk tugas klasifikasi medis, terutama yang berkaitan dengan penyakit paru-paru.

Kunci Utama: Klasifikasi, Penyakit Paru-Paru, *Random Forest, XGBoost*, Data Demografi.

***Abstract**: In this study, the Random Forest and XGBoost algorithms were compared in lung disease classification using patient demographic data. The dataset used consists of 30,000 patient data with 9 attributes and 1 label taken from Kaggle. Research stages include data collection, preprocessing, data sharing, and data classification using both algorithms. The results show that the XGBoost algorithm has an accuracy of 94% and an AUC of 0.98, while the Random Forest algorithm has an accuracy of 91% and an AUC of 0.97. While Random Forest is faster and easier to interpret, XGBoost performs better with complex data with more consistent results. Through the use of regularization techniques and better handling of outliers, XGBoost can also better address the overfitting problem. This study provides guidance for researchers and practitioners in selecting the best algorithm for medical classification tasks, especially those related to lung diseases.*

***Keywords** : Classification, Lung Disease, Random Forest, XGBoost, Demographic Data*

1. PENDAHULUAN

Paru-paru adalah salah satu bagian tubuh yang melakukan fungsi pernafasan. Paru-paru juga turut terlibat dalam proses respirasi. Respirasi adalah proses pelepasan energi yang tersimpan pada zat sumber energi melalui proses kimia yang menggunakan oksigen. Dalam proses ini, senyawa organik dipecah menjadi *CO2, H2O*, dan energy [1]. Selain itu, paru-paru mengubah karbon dioksida dalam darah dan oksigen dari udara. Beberapa penyakit mematikan dapat terjadi pada paru-paru, salah satunya Pneumonia. Penyakit paru-paru adalah penyakit yang paling umum pada manusia dan biasanya disebabkan karena menghirup udara yang tercemar oleh debu, asap, virus, atau bakteri, yang dapat menyebabkan infeksi saluran pernapasan [2]. Penyakit paru-paru dapat menyerang siapa pun, mulai dari bayi hingga orang dewasa. Penyakit ini jelas tidak mudah disembuhkan. Selain Pneumonia adapula penyakit berbahaya yang dapat menganggu fungsi paru-paru yakni, kanker paru-paru yang umumnya disebabkan oleh kebiasaan merokok. Perokok di Indonesia mengalami peningkatan. Di Indonesia, konsumsi rokok terus meningkat di kalangan masyarakat dari segala usia, termasuk orang-orang yang kurang mampu, sebagai akibat dari interaksi sosial yang berkelanjutan yang mengarah pada kecanduan nikotin. Perokok mengabaikan bahaya merokok yang tidak terlihat bagi kesehatan mereka [3].

Selain pneumonia dan kanker paru-paru, gangguan atau penyakit pada paru-paru manusia adalah sebagai berikut:

1. Tuberkulosis (TBC) adalah penyakit paru-paru yang disebabkan oleh bakteri *Mycobacterium Tuberculosis*. Bakteri ini tidak hanya menyerang paru-paru tetapi juga dapat menyerang kelenjar getah bening, sistem saraf pusat, tulang, dan ginjal [4].
2. Asma adalah kondisi yang terjadi dan menyebabkan peradangan kronik pada saluran pernafasan. Suara mengi dan batuk adalah gejalanya, serta rasa berat pada dada, terutama pada malam hari, dan sesak pada saluran pernafasan. Asma biasanya didiagnosis dengan sifat variabilitas [5].
3. Infeksi saluran pernapasan yang menyerang bronkus dikenal sebagai bronkitis, yang paling sering menyerang anak-anak karena lingkungan mereka yang kotor. Contohnya adalah asap yang berasal dari pembakaran kayu saat memasak, asap dari kendaraan bermotor, dan asap dari orang tua yang merokok di rumah [6].

Menghindari asap rokok, baik pasif maupun aktif, adalah salah satu pencegahan dan pengobatan yang dapat dilakukan oleh orang yang menderita penyakit paru-paru. Ini karena asap rokok mengandung karsinogen yang paling aktif. Hidup di area yang aman dari polusi udara. memupuk kebiasaan makan makanan yang bergizi dan berserat. Beberapa metode pengobatan yang dapat digunakan termasuk bedah, radioterapi, kemotrepi, dan bedah laser [7].

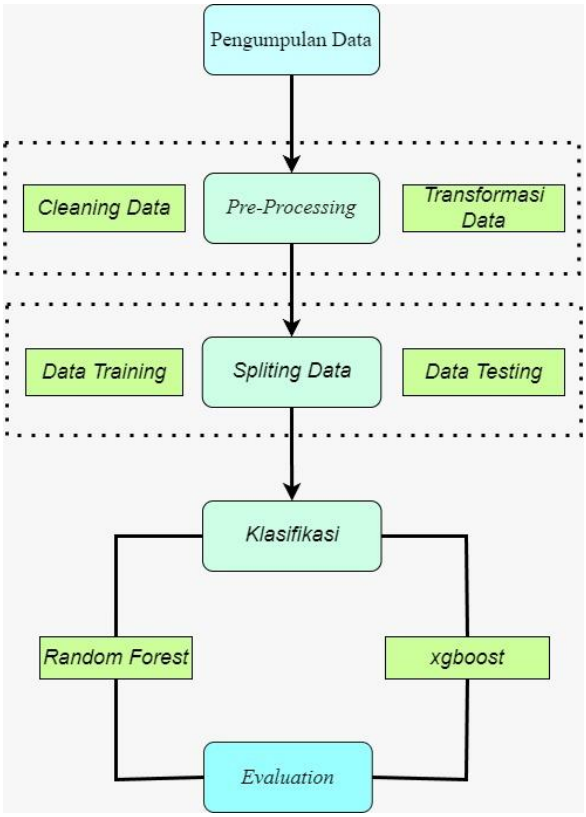
Oleh karena mengkhawatirkannya penyakit-penyakit yang dapat bersarang pada paru-paru, maka sekiranya perlu dibuat suatu model untuk mengklasifikasi penyakit paru-paru dalam perbandingan algoritma *Random Forest* dan *XGBoost. Random Forest* adalah salah satu metode klasifikasi yang melakukan pemilihan secara acak dan membangkitkan simpul anak untuk setiap node (simpul di atasnya) [8]. Namun, *XGBoost*, singkatan dari "*Extreme Gradient Boosting*", adalah varian dari peningkatan pohon *Gradient* [9]. *Gradient Tree Boosting* adalah metode peningkatan pohon yang menggabungkan sekumpulan pengklasifikasian yang lemah untuk membuat pengklasifikasian yang kuat. *XGBoost* meningkatkan kinerja dengan mengatur kompleksitas pohon dengan menggunakan berbagai metode regularisasi dan memiliki kinerja pemilihan yang lebih akurat [10].

Data mining adalah proses menemukan pola atau aturan dalam kumpulan data yang sangat mungkin. menemukan data baru [11]. Data mining biasanya digunakan untuk menemukan data dari basis data berskala besar, sehingga sering disebut sebagai *Knowledge Discovery Databases (KDD)* [12]. Mencari pola yang dapat menggambarkan atau membedakan setiap kelas data adalah proses yang dikenal sebagai

klasifikasi, yang merupakan salah satu teknik penting dalam data mining. Tujuan dari klasifikasi adalah untuk menemukan kelas objek yang peruntukannya tidak diketahui [13].

2. METODOLOGI PENELITIAN

Adapun beberapa tahapan-tahapan dalam metode penelitian dengan menjelaskan pada gambar 1.



Gambar 1. Metodologi Penelitian

2.1 Pengumpulan Data

Data yang dikumpulkan melalui website Kaggle diupload dengan Andot03 Bsrc, yang berjumlah 30.000 data dengan 9 atribut dan 1 label. Tujuan pengumpulan data ini adalah untuk mendapatkan informasi tentang suata data yang lebih mendalam.

2.2 *Preprocessing Data*

Preprocessing adalah proses penggunaan data dalam informasi yang mentah sebelum melakukan analisis, yang mencakup beberapa tahapan seperti pembersihan dan pengubahan data. Tujuan dari *Preprocessing* adalah untuk membuat data dalam format yang mudah digunakan supaya prosesnya lebih mudah [14].

2.2.1 *Cleaning Data*

Pembersihan Data adalah proses membersihkan data dari suara atau data yang tidak relevan seperti nilai yang hilang (yang tidak ada), outliers, duplikat, dan data yang tidak konsisten, dan kemudian mempersiapkan data untuk diolah dan dianalisis dengan baik [15].

2.2.2 *Transformasi Data*

Pengubahan Data yang bertujuan untuk mengubah data asli menjadi bentuk lain sehingga data tersebut dapat memenuhi kriteria dalam sebuah penelitian [16].

2.3 *Spliting Data*

Dalam penelitian, pembagian data digunakan. Untuk menguji model atau algoritma, penelitian biasanya membagi set data menjadi dua atau lebih bagian. Data latih digunakan untuk melatih algoritma, sementara data uji digunakan untuk menguji kinerja algoritma [17].

Oleh karena itu, untuk dua algoritma yang digunakan dalam penelitian ini, *Random Forest* dan *XGBoost*, digunakan pembagian data 80:20.

2.4 Data Mining

Data mining adalah proses yang menggunakan matematika, statistik, kecerdasan buatan, dan pembelajaran mesin untuk mengekstraksi dan mengidentifikasi informasi terkait dan bermanfaat dari berbagai database yang sangat besar. Tujuan dari data mining adalah untuk mendapatkan nilai tambahan dari kumpulan data, yang terdiri dari pengetahuan yang sebelumnya tidak diketahui secara manual [18].

Berdasarkan definisi yang telah diberikan, beberapa hal penting tentang Data Mining adalah sebagai berikut [19]:

1. Data Mining adalah suatu proses otomatis yang menggabungkan data yang sudah ada.
2. Data yang akan diproses adalah data yang sangat besar.

3. Tujuan dari proses ini adalah untuk menemukan pola atau hubungan yang dapat memberikan indikasi yang bermanfaat.

2.5 Klasifikasi

Klasifikasi adalah proses menempatkan sesuatu ke dalam kategori atau kelas yang telah ditentukan sebelumnya [20]. Klasifikasi juga mengacu pada proses pembuatan model yang mengelompokkan objek berdasarkan karakteristiknya [21]. Proses klasifikasi data atau dokumen dapat dimulai dengan membuat aturan klasifikasi dengan menggunakan data latih dan data uji [22]. Teknik klasifikasi menjadi dua algoritma yaitu *Random Forest* dan *XGBoost*.

2.5.1 R*andom Forest*

Random Forest (RF) adalah algoritma yang bergantung pada pohon regresi dan klasifikasi untuk mencapai node akhir dalam struktur pohon melalui metode pemisahan biner rekursif. Algoritma *Random Forest* memiliki beberapa kelebihan, termasuk kemampuan untuk menghasilkan error yang relatif rendah, kinerja yang baik dalam klasifikasi, kemampuan untuk mengatasi data pelatihan dalam jumlah besar, dan metode yang efektif untuk mengestimasi data yang tidak ada. Banyak pohon independen dengan subset yang dipilih secara acak menggunakan bootstrap dari sampel pelatihan dan variable input di setiap node dibuat [23]. Metode ini terdiri dari root node, internal node, dan leaf node. Root node adalah simpul tertinggi, atau biasa disebut sebagai akar dari pohon keputusan. Leaf node, juga dikenal sebagai terminal node, adalah simpul terakhir yang tidak memiliki output atau input, sedangkan internal node adalah simpul percabangan dengan minimal dua output dan satu input. Nilai entropy digunakan sebagai penentu tingkat ketidakmurnian atribut dan nilai gain informasi. Nilai entropy dihitung dengan rumus berikut dalam persamaan 1 dan 2 [24]:

$$Entropy(Y) = -\sum_i p(c|Y)\log 2p(c|Y) \text{ ……..(1)}$$

Di mana Y adalah himpunan kasus, dan $p(c|Y)$adalah persentase nilai Y terhadap kelas c.

$$Information\ Gain\ (Y,a) = Entropy(Y) - \sum_v e\ Values(a)\frac{|Yv|}{|Ya|}Entropy(Yv) \text{ …..(2)}$$

Nilai (a) adalah semua nilai yang mungkin dalam himpunan kasus a. Yv adalah subkelas dari Y dengan kelas v yang terkait dengan kelas a. Sedangkan Ya adalah semua nilai yang sama dengan a.

2.5.2 *XGBoost*

XGBoost merupakan sebuah kelompok *Decision Tree* yang didasarkan pada *Gradient Boosting* dan dirancang untuk menjadi sangat *scalable*. Seperti gradient boosting, *XGBoost* membuat ekspansi tujuan secara additif dengan minimizing loss function oleh itu *XGBoost* hanya menangani *Decision Tree* sebagai *Base Classifers*, variasi loss function digunakan untuk mengontrol kompleksitas tree [25]. Adapun rumus dalam menggunakan algoritma *XGBoost* adalah sebagai berikut[26]:

$$L_{xgb} = \sum_{i=0}^{n} L(y_{i,}F(x_i)) + \sum_{m=1}^{M}\Omega(h_m) \text{……..(3)}$$

Keterangan:

L_{xgb} = Ini adalah fungsi kerugian yang ingin diminimalkan oleh model *XGBoost* secara keseluruhan.

$\sum_{i=0}^{n} L(y_{i,}F(x_i))$ = Menunjukkan penjumlahan dari 0 hingga n, di mana n adalah jumlah data secara keseluruhan. Ini adalah fungsi kerugian antara nilai sebenarnya $y_{i,}$ dan prediksi F(x_i) pada data ke-iii.

$\sum_{m=1}^{M} \Omega(h_m)$ = Menunjukkan penjumlahan dari 1 hingga M, di mana M adalah jumlah total pohon dalam model. Ini adalah fungsi regularisasi untuk pohon ke-m yang membantu mengontrol kompleksitas model.

$$\Omega(h) = yT + \frac{1}{2}\lambda\|w\|^2 \text{.....(4)}$$

Keterangan:
$\Omega(h)$ = Setiap pohon memiliki kompleksitas yang dikontrol oleh fungsi regularisasi.

yT = Parameter regularisasi yang mengatur jumlah daun yang ada pada pohon. Pohon yang lebih sederhana akan memiliki nilai yang lebih tinggi.

$\frac{1}{2}\lambda\|w\|^2$ = Konstanta untuk penyesuaian skala. Parameter regularisasi yang mengontrol ukuran nilai parameter w. Norma kuadrat dari vektor parameter w yang menunjukkan ukuran dari semua skor output dari daun-daun pohon.

2.6 Evaluasi

Evaluasi adalah prosedur yang digunakan dalam pembelajaran mesin untuk mengevaluasi kinerja klasifikasi. Sangat penting untuk melakukan evaluasi untuk menilai kemampuan mereka untuk membuat klasifikasi yang akurat. Dalam penelitian ini, metrix evaluasi diklasifikasikan menjadi laporannya dalam lima bagian, yaitu akurasi, kecepatan, *Recall*, skor F1, dukungan, dan konsistensi matriks. Tabel *Confusion Matrix* memungkinkan penulis melihat kinerja algoritma pembelajaran yang diawasi. Setiap baris memiliki contoh class sebenarnya, tetapi setiap kolom matiks mewakili instance dalam kelas yang diharapkan[27].

2.6.1 *Confusion Matriks*

kinerja dalam situasi matriks confusion. Perfomance adalah alat yang biasanya digunakan oleh penelitian untuk menentukan kualitas algrotima. Ini memiliki tiga metrik, yaitu akurasi, presisi, dan *Recall*, yang digunakan untuk melakukan klasifikasi disajikan tabel 1 [28].

Tabel 1. *Confusion Matriks*

	Prediksi	
True	*TP*	*FN*
False	*FP*	*PN*

Apabila:
TP : True Positive
FP : False Positive
FN : False Negative
TN : True Negative

A. Nilai yang dikenal sebagai akurasi adalah ukuran seberapa dekat nilai prediksi sistem dengan nilai prediksi yang benar.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \text{ ……..(5)}$$

B. *Precision* ini mencakup nilai sensitivitas atau ketepatan sistem terhadap informasi yang diberikan oleh sistem untuk menunjukkan data positif atau negatif dengan benar.

$$Precision = \frac{TP}{TP+FP} \text{ ……(6)}$$

C. Nilai pengembalian adalah nilai yang menunjukkan tingkat keberhasilan atau spesifisitas untuk mengetahui kembali informasi tentang data atau teks yang positif atau negatif.

$$Recall = \frac{TP}{TP+FN} \text{ ……..(7)}$$

2.6.2 *ROC*

Pada *Confusion Matrix*, performa informasi hanya disajikan dalam bentuk angka. Untuk menampilkan informasi kinerja algoritma klasifikasi dalam bentuk grafik dapat digunakan *Receiver Operating Characteristic (ROC)* atau *Precision-Recall Curve*. Kurva *ROC* dibuat berdasarkan nilai yang telah didapatkan dari perhitungan dengan confusion matrix, yaitu antara False Positive Ratedengan True Positive

Rate. Untuk membandingkan nilai kinerja masing-masing algoritmadapat dilakukan dengan membandingkan luas di bawah kurva atau *AUC (Area Under Curve)*[29].

Kurva ROC (Receiver Operating Characteristic) adalah grafik yang digunakan untuk mengevaluasi kinerja model klasifikasi, khususnya dalam konteks klasifikasi biner. Kurva *ROC* menggambarkan trade-off antara tingkat positif salah *(False Positive Rate, FPR)* dan tingkat positif benar *(True Positive Rate, TPR)* pada berbagai ambang batas keputusan. Komponen utama kurva *ROC* antara lain :

A. *True Positive Rate (TPR)* : Juga dikenal sebagai sensitivitas atau recall, *TPR* adalah rasio jumlah prediksi positif benar terhadap total jumlah kasus positif sebenarnya.
Rumusnya:

$$TPR = \frac{TP}{TP+FN} \text{.....(8)}$$

B. *False Positive Rate (FPR)*: *FPR* adalah rasio jumlah prediksi positif salah terhadap total jumlah kasus negatif sebenarnya.
Rumusnya:

$$FPR = \frac{FP}{FP+PN} \text{.....(9)}$$

C. *Area Under The Curve*: Luas di bawah kurva ROC dan memiliki nilai AUC 0–1. Nilai AUC yang lebih tinggi menunjukkan kinerja model yang lebih baik.

3. HASIL DAN PEMBAHASAN

Adapun juga hasil dan pembahasan yang berdasarkan tahapan-tahapan metode penelitian.

3.1 Pengumpulan Data

Pada tahap pengumpulan data yang digunakan dalam dataset predict terkena penyakit paru-paru yang di upload oleh Andot03 Bsrc melalui website kaggle https://www.kaggle.com/datasets/andot03bsrc/dataset-predic-terkena-penyakit-paruparu yang berjumlah 30000 data masing-masing memiliki 9 atribut dengan 1 label. Data ini disajikan pada gambar 2.

	No	Usia	Jenis_Kelamin	Merokok	Bekerja	Rumah_Tangga	Aktivitas_Begadang	Aktivitas_Olahraga	Asuransi	Penyakit_Bawaan	Hasil
0	1	Tua	Pria	Pasif	Tidak	Ya	Ya	Sering	Ada	Tidak	Ya
1	2	Tua	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Ada	Tidak
2	3	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak	Tidak
3	4	Tua	Pria	Aktif	Ya	Tidak	Tidak	Jarang	Ada	Ada	Tidak
4	5	Muda	Wanita	Pasif	Ya	Tidak	Tidak	Sering	Tidak	Ada	Ya
...	...	...	...	...	...	...	...	...	...	...	...
29995	29996	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak	Tidak
29996	29997	Tua	Wanita	Aktif	Ya	Tidak	Ya	Jarang	Ada	Ada	Tidak
29997	29998	Tua	Wanita	Aktif	Ya	Tidak	Ya	Jarang	Ada	Ada	Tidak
29998	29999	Muda	Wanita	Pasif	Ya	Tidak	Tidak	Sering	Tidak	Ada	Tidak
29999	30000	Tua	Wanita	Pasif	Tidak	Ya	Tidak	Sering	Tidak	Tidak	Ya

30000 rows × 11 columns

Gambar 2. Dataset Penyakit Paru-Paru

Berdasarkan Gambar 2 merupakan pada dataset penyakit paru-paru ini terdiri beberapa fitur seperti Usia, Jenis_Kelamin, Merokok, Bekerja, Rumah_Tangga, Aktivitas_Begadang, Aktivitas_Olahraga, Asuransi, Penyakit_Bawaan Dan Hasil. Berikut ini menjelaskan masing-masing fitur:

1. Usia: Usia subjek dalam dataset.
2. Jenis Kelamin: Jenis kelamin subjek, misalnya pria atau wanita.
3. Merokok: Status merokok subjek (Ya/Tidak).
4. Bekerja: Status pekerjaan subjek (Ya/Tidak).

5. Rumah Tangga: Informasi mengenai kondisi rumah tangga subjek.
6. Aktivitas Bergadang: Informasi tentang kebiasaan bergadang subjek.
7. Aktivitas Olahraga: Informasi tentang kebiasaan berolahraga subjek.
8. Asuransi: Status kepemilikan asuransi subjek.
9. Penyakit Bawaan: Kondisi kesehatan bawaan subjek.
10. Hasil: Hasil dari penelitian atau tes terkait penyakit kanker paru-paru.

```
Usia                0
Jenis_Kelamin       0
Merokok             0
Bekerja             0
Rumah_Tangga        0
Aktivitas_Begadang  0
Aktivitas_Olahraga  0
Asuransi            0
Penyakit_Bawaan     0
Hasil               0
dtype: int64
```

Gambar 3. Hasil Pembersihan Data

3.2 *Preprocessing*

Sesudah dalam pengumpulan data ada beberapa tahapan preprocessing. Tujuannya untuk melakukan data yang mentah menjadi data relevan untuk digunakan dalam modeling.

3.2.1 *Cleaning Data* (Pembersihan Data)

Pada tahap melakukan proses pembersihan data dalam jumlah 30.000 data. Bahwasanya pada step tidak terdapat *missing value*, semua data telah dikonfirmasikan dan siap digunakan dalam tahap selanjutnya pada disajikan gambar 3.

3.2.2 *Tranformasi Data* (Pengubahan Data)

Setelah melakukan pembersihan data, selanjutnya dalam tahap pengubahan data sehingga data yang berubah menjadi numerik untuk meningkatkan data yang lebih akurat pada disajikan gambar 4.

	No	Usia	Jenis_Kelamin	Merokok	Bekerja	Rumah_Tangga	Aktivitas_Begadang	Aktivitas_Olahraga	Asuransi	Penyakit_Bawaan	Hasil
0	1	1	1	0	0	1	1	1	1	0	1
1	2	1	1	1	0	1	1	0	1	1	0
2	3	0	1	1	0	1	1	0	1	0	0
3	4	1	1	1	1	0	0	0	1	1	0
4	5	0	0	0	1	0	0	1	0	1	1
...	...	...	...	...	...	...	...	...	...	...	...
29995	29996	0	1	1	0	1	1	0	1	0	0
29996	29997	1	0	1	1	0	1	0	1	1	0
29997	29998	1	0	1	1	0	1	0	1	1	0
29998	29999	0	0	0	1	0	0	1	0	1	0
29999	30000	1	0	0	0	1	0	1	0	0	1

30000 rows × 11 columns

Gambar 4. Hasil Pengubahan Data

3.3 *Spliting Data*

Setelah melakukan *Preprocessing* dan analisis klasifikasi, data dibagi menjadi dua bagian yaitu pelatihan data dan penilaian data. Pelatihan data membantu algortima membuat model, dan penilaian data mengukur tingkat keakuratan dan performa yang diperoleh dari pelatihan data.

Tabel 2. Pembagian Data Training dan Data Testing

Keterangan	Data Training	Data Testing	Total
Proporsi	80%	20%	100%
Jumlah	24.000	6.000	30.000

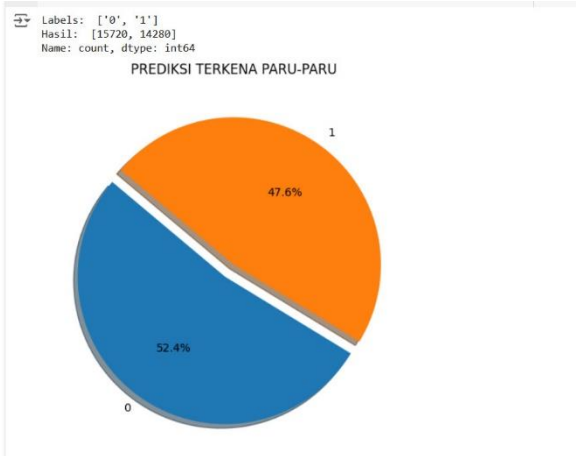
Berdasarkan Tabel 2, dari 30.000 dataset, 24.000 data dibagi untuk instruksi dan 6.000 data untuk pengujian. *Software phyton* digunakan untuk membagi data instruksi dan pengujian secara acak.

3.4 Klasifikasi

Setelah tahapan *Preprocessing* menyelesaikan pembagian data, langkah berikutnya adalah mengklasifikasikan dalam perbandingan algoritma *Random Forest* dan *XGBoost*.

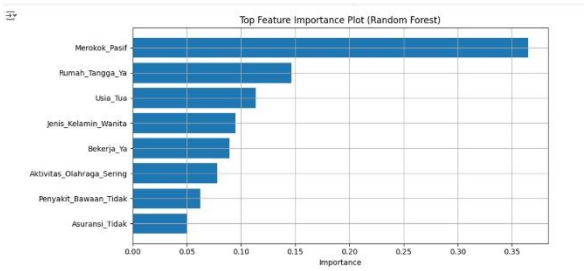
3.4.1 *Random Forest*

Pada tahap ini, kami menganalisis algoritma *Random Forest* untuk menunjukkan klasifikasi penyakit paru-paru. Diagram *Random Forest* menunjukkan bahwa klasifikasi pasien tidak terkena penyakit paru-paru sebesar 52,4%, sedangkan klasifikasi pasien terkena penyakit paru-paru sebesar 47,6%. Kesimpulannya, jumlah pasien tidak terkena penyakit paru-paru adalah 15.721, dan jumlah pasien terkena penyakit paru-paru adalah 14.280, seperti yang ditunjukkan pada gambar 5.



Gambar 5. Diagram *Random Forest*

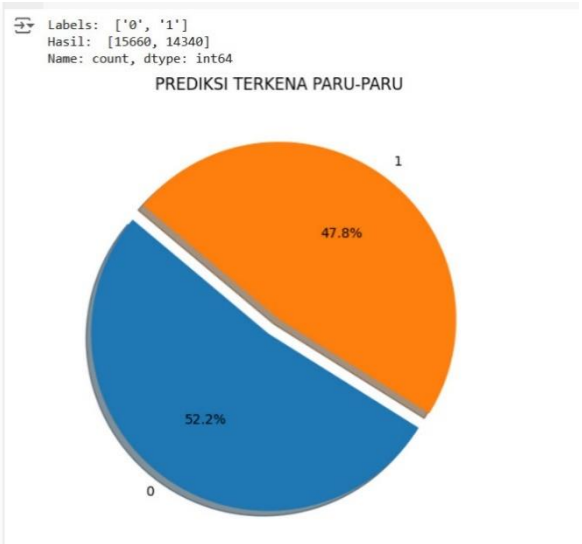
Gambar 6 menunjukkan kesimpulan dari hasil pemilihan ranking dari data demografi pasien ini bahwa Merokok, Rumah_Tangga, dan Usia adalah variabel yang paling mempengaruhi penyakit paru-paru.



Gambar 6. *Rank Selection Random Forest*

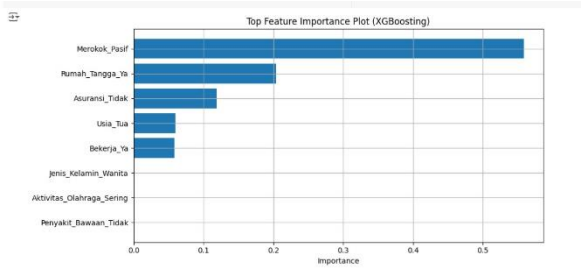
3.4.2 *XGBoost*

Pada tahap ini, kami menggunakan algoritma *XGBoost* untuk menunjukkan klasifikasi penyakit paru-paru. Diagram *XGBoost* menunjukkan bahwa klasifikasi pasien tidak terkena penyakit paru-paru sebesar 52,2%, dan klasifikasi pasien terkena penyakit paru-paru sebesar 47,8%. Dengan demikian, total jumlah pasien tidak terkena penyakit paru-paru adalah 15.660, dan jumlah pasien terkena penyakit paru-paru adalah 14.340, seperti yang ditunjukkan pada gambar 7.



Gambar 7. Diagram *XGBoost*

Gambar 8 menunjukkan kesimpulan dari hasil pemilihan ranking dari data demografi pasien ini bahwa Merokok, Rumah_Tangga, dan Asuransi adalah variabel yang paling mempengaruhi penyakit paru paru.



Gambar 8. *Rank Selection XGBoost*

Berdasarkan Tabel 3, hasil perbandingan menunjukkan bahwa *Random Forest* mengklasifikasikan 52.4% (15,720 pasien) sebagai tidak terkena penyakit paru-paru, sedangkan *XGBoost* mengklasifikasikan 52.2% (15,660 pasien). *Random Forest* juga mengklasifikasikan 47.6% (14,280 pasien) dan *XGBoost* mengklasifikasikan 47.8% (14,340 pasien). Kedua metode menggunakan variabel yang sebagian besar serupa, seperti Merokok, Rumah_Tangga, dan faktor Usia. Namun, *XGBoost* menambahkan variabel Asuransi sebagai komponen tambahan yang memengaruhi hasil klasifikasi.

Tabel 3. Hasil Klasifikasi Algoritma *Random Forest* Dan *XGBoost*

Aspek	*Random Forest*	*XGBoost*
Prediksi tidak terkena penyakit paru-paru	52.4% (15.720 pasien)	52.2% (15.660 Pasien)
Prediksi terkena penyakit paru-paru	47.6% (14.280 pasien)	47.8% (14.340 Pasien)
Variable yang mempengaruhi	Merokok, Rumah_Tangga, Usia	Merokok, Rumah_Tangga, Asuransi

3.5 Evaluasi

Sesudah melakukan hasil klasifikasi maka akan menggunakan pegujian dari *Confusion Matriks* maupun *ROC* dalam algoritma *Random Forest* dan *XGBoost*.

3.5.1 Confusion Matriks

Pada melakukan analisis *Confusion Matriks* dalam suatu model algoritma *Random Forest* dan *XGBoost*. Tujuan *Confusión Matriks* untuk menyajikan alat yang berguna dalam mengevaluasikan dengan perfoma model klasifikasi.

A. *Random Forest*

Berdasarkan Tabel 4 dan Gambar 9, algoritma *Random Forest* menghasilkan hasil klasifikasi 91% pada penyakit paru-paru.

Tabel 4. *Accuracy Random Forest*

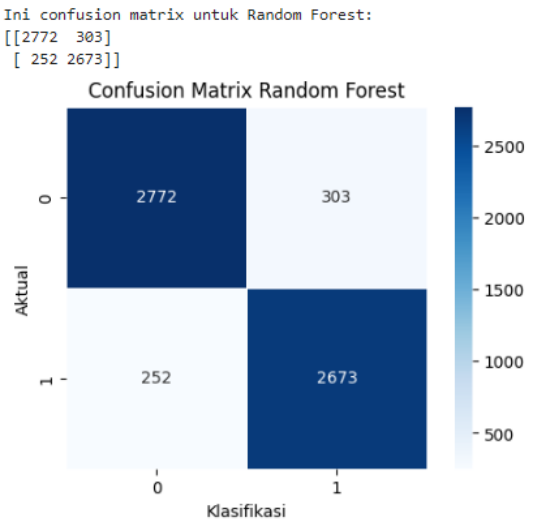
	Random Forest			
	Precision	Recall	F1-Score	Support
0	0.92	0.90	0.91	3075
1	0.90	0.91	0.91	2925
Accuracy			0.91	6000
Macro Avg	0.91	0.91	0.91	6000
Weighted Avg	0.91	0.91	0.91	6000

Berdasarkan hasil menggunakan algoritma *Random Forest* dengan *Splitting Data 80:20* mendapatkan hasil akurasi 91%. Hasil ini dapat dihitung akurasi sebagai berikut:

TP = 2673 FP = 303

TN = 2772 FN = 252

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{2673+2772}{2673+2772+303+252} = \frac{5445}{6000} = 91\%$$



Gambar 9. Hasil *Confusion Matriks Random Forest*

B. *XGBoost*

Berdasarkan Tabel 5 dan Gambar 10, algoritma *XGBoost* menghasilkan hasil klasifikasi 94% pada penyakit paru-paru.

Tabel 5. *Accuracy XGBoost*

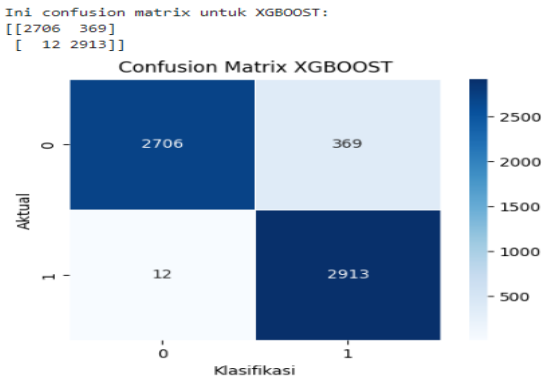
	XGBoost			
	Precision	Recall	F1-Score	Support
0	1.00	0.88	0.93	3075
1	0.89	1.00	0.94	2925
Accuracy			0.94	6000
Macro Avg	0.94	0.94	0.94	6000
Weighted Avg	0.94	0.94	0.94	6000

Berdasarkan hasil menggunakan algoritma *XGBoost* dengan *Splitting Data 80:20* mendapatkan hasil akurasi 94%. Hasil ini dapat dihitung akurasi sebagai berikut:

TP = 2913 FP = 369

TN = 2706 FN = 12

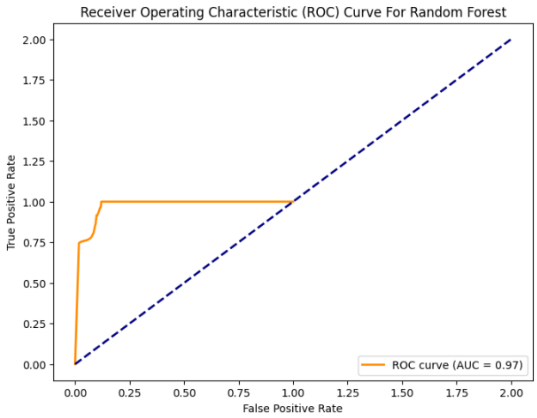
$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{2913+2706}{2913+2706+369+12} = \frac{5619}{6000} = 94\%$$



Gambar 10. Hasil *Confusion Matriks XGBoost*

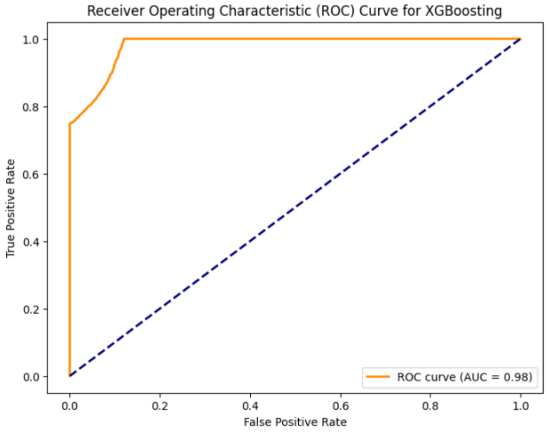
3.5.2 *ROC*

Pada gambar ini, kita dapat melihat gambar yang dihasilkan dari klasifikasi algoritma Random Forest dan XGBoost.



Gambar 11. *ROC Random Forest*

Gambar 11 menunjukkan bahwa model *Random Forest* memiliki performa yang sangat baik dengan *AUC* sebesar 0.97 pada kurva *ROC*, yang menunjukkan kemampuan prediksi yang sangat baik untuk membedakan kelas positif dan negatif dengan tingkat kesalahan yang rendah. Oleh karena itu, *Random Forest* dianggap sebagai pilihan yang tepat untuk klasifikasi dalam situasi ini.



Gambar 12. *ROC XGBoost*

Gambar 12 menunjukkan bahwa model *XGBoost* memiliki performa yang sangat baik dengan *AUC* sebesar 0.98 pada kurva *ROC*, yang menunjukkan kemampuan prediksi yang hampir sempurna untuk membedakan kelas positif dan negatif dengan tingkat kesalahan yang sangat rendah. Oleh karena itu, *XGBoost* adalah pilihan yang tepat untuk klasifikasi dalam situasi ini.

3.5.3 Analisis Hasil Perbandingan *Accuracy* Dan *AUC*

Setelah melakukan evaluasi menunjukkan hasil perbandingan Akurasi dan *AUC* dalam algoritma *Random Forest* dan *XGBoost* dalam menggunakan *Splitting Data 80:20*. Hasil Akurasi dan *AUC* di sajikan tabel 6.

Tabel 6. *Hasil Perbandingan Accuracy Dan AUC*

	Random Forest	*XGBoost*
Accuracy	*91%*	*94%*
AUC	*97%*	*98%*

Hasil di atas menunjukkan bahwa *XGBoost* lebih akurat dan memiliki *AUC* yang lebih besar daripada *Random Forest*.

4. KESIMPULAN

Dalam penelitian ini, algoritma *Random Forest* dan *XGBoost* dibandingkan dalam mengklasifikasikan penyakit paru-paru berdasarkan data demografi pasien. Penelitian ini menemukan bahwa *Random Forest* lebih cepat dan lebih mudah untuk ditafsirkan, tetapi *XGBoost* lebih akurat dan memiliki *AUC* yang lebih besar. *Random Forest* juga lebih baik dalam menangani data yang memberik dan kompleks, tetapi *XGBoost* lebih baik menangani data yang rumit dan memberik. Kedua algoritma ini berguna untuk mengklasifikasikan penyakit paru-paru, dan pemilihan algoritma dapat disesuaikan dengan kebutuhan aplikasi yang digunakan, apakah lebih mengutamakan interpretabilitas *Random Forest* dan kecepatan atau kemampuan menangani data yang lebih kompleks.

DAFTAR RUJUKAN

[1] M. Y. Haffandi, E. Haerani, F. Syafria, and L. Oktavia, "Klasifikasi Penyakit Paru-Paru Dengan Menggunakan Metode Naïve Bayes Classifier," *J. Tek. Inf. dan Komput.*, vol. 5, no. 2, p. 176, 2022, doi: 10.37600/tekinkom.v5i2.649.

[2] A. S. Pratama, S. Safrizal, and J. Iriani, "Sistem Pakar Mendiagnosa Penyakit Gangguan Pernafasan Oleh Asap Rokok Menggunakan Metode Dempster Shafer," *It (Informatic Tech. J.*, vol. 9, no. 1, p. 79, 2021, doi: 10.22303/it.9.1.2021.79-88.

[3] E. W. Tipa, P. A. Kawatu, and A. F. C. Kalesaran, "Hubungan Kebiasaan Merokok Dengan Kapasitas Vital Paru Pada Penambang Emas Di Desa Tatelu Kabupaten Minahasa Uatara," *J. KESMAS*, vol. 10, no. 3, pp. 140–146, 2021.

[4] RSUD, "Bangsal Kemuning (Penyakit Paru)," *RSUD*, 2024. https://rsud.banjarkota.go.id/rawat-inap/bangsal-kemuning/

[5] N. Ratama, "Analisa Dan Perbandingan Sistem Aplikasi Diagnosa Penyakit Asma Dengan Algoritma Certainty Factor Dan Algoritma Decision Tree Berbasis Android," *J. Inform. J. Pengemb. IT*, vol. 3, no. 2, pp. 177–183, 2018, doi: 10.30591/jpit.v3i2.848.

[6] M. S. C Imiliati, NB Nugroho, "Bronchitis Pada Anak Menggunakan Metode Certainty Factor Dan Teorema Bayes Di Upt .," *J. Cyber Tech*, 2021.

[7] Yellia Mangan, *Solusi Sehat Mencegah & Mengatasi Kanker*. AgroMedia, 2009.

[8] G. A. Sandag, "Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest," *CogITo Smart J.*, vol. 6, no. 2, pp. 167–178, 2020, doi: 10.31154/cogito.v6i2.270.167-178.

[9] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[10] Z. Salam Patrous, "Evaluating XGBoost for User Classification by using Behavioral Features Extracted from Smartphone Sensors," p. 67, 2018, [Online]. Available: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1240595&dswid=-6444

[11] F. Solikhah, M. Febianah, A. L. Kamil, W. A. Arifin, and Shelly Janu Setyaning Tyas, "Analisis Perbandingan Algoritma Naive Bayes Dan C.45 Dalam Klasifikasi Data Mining Untuk Memprediksi Kelulusan," *Tematik*, vol. 8, no. 1, pp. 96–103, 2021, doi: 10.38204/tematik.v8i1.576.

[12] Dimsyiar M Al Hafiz, Khoirul Amaly, Javen Jonathan, M Teranggono Rachmatullah, and Rosidi, "Sistem Prediksi Penyakit Jantung Menggunakan Metode Naive Bayes," *J. Rekayasa*

Elektro Sriwij., vol. 2, no. 2, pp. 151–157, 2021, doi: 10.36706/jres.v2i2.29.

[13] R. Nofitri and J. Eska, "Implementasi Data Mining Klasifikasi C4.5 Dalam Menentukan Kelayakan Pengambilan Kredit," *Semin. Nas. R.*, vol., no., p., 2018.

[14] F. Akbar, H. W. Saputra, A. K. Maulaya, M. F. Hidayat, and R. Rahmaddeni, "Implementasi Algoritma Decision Tree C4.5 dan Support Vector Regression untuk Prediksi Penyakit Stroke," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 2, no. 2, pp. 61–67, 2022, doi: 10.57152/malcom.v2i2.426.

[15] N. M. Farhan and B. Setiaji, "Indonesian Journal of Computer Science," *Indones. J. Comput. Sci.*, vol. 12, no. 2, pp. 284–301, 2023, [Online]. Available: http://ijcs.stmikindonesia.ac.id/ijcs/index.php/ijcs/article/view/3135

[16] M. Z. Al-Taie, S. Kadry, and J. P. Lucas, "Online data preprocessing: A case study approach," *Int. J. Electr. Comput. Eng.*, vol. 9, no. 4, pp. 2620–2626, 2019, doi: 10.11591/ijece.v9i4.pp2620-2626.

[17] A. Putri *et al.*, "Komparasi Algoritma K-NN, Naive Bayes dan SVM untuk Prediksi Kelulusan Mahasiswa Tingkat Akhir," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 1, pp. 20–26, 2023, doi: 10.57152/malcom.v3i1.610.

[18] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," *Edik Inform.*, vol. 2, no. 2, pp. 213–219, 2017, doi: 10.22202/ei.2016.v2i2.1465.

[19] M. Syahril, K. Erwansyah, and M. Yetri, "Penerapan Data Mining Untuk Menentukan Pola Penjualan Peralatan Sekolah Pada Brand Wigglo Dengan Menggunakan Algoritma Apriori," *J-SISKO TECH (Jurnal Teknol. Sist. Inf. dan Sist. Komput. TGD)*, vol. 3, no. 1, p. 118, 2020, doi: 10.53513/jsk.v3i1.202.

[20] N. W. S. Agustini, D. Priadi, and R. V. Atika, "Profil Kimia dan Aktivitas Antibakteri Fraksi Aktif Nannochloropsis sp. sebagai Senyawa Penghambat Bakteri Penyebab Gangguan Kesehatan Mulut," *J. Pascapanen dan Bioteknol. Kelaut. dan Perikan.*, vol. 17, no. 1, p. 19, 2022, doi: 10.15578/jpbkp.v17i1.781.

[21] E. Susilowati, M. K. Sabariah, and A. A. Gozali, "Implementasi Metode Support Vector Machine untuk Melakukan Klasifikasi Kemacetan Lalu Lintas Pada Twitter," *E-Proceeding Eng.*, vol. 2, no. 1, pp. 1478–1484, 2015.

[22] S. Raharjo and E. Winarko, "Klasterisasi, klasifikasi dan peringkasan teks berbahasa indonesia," *Kommit 2014*, vol. 8, no. Kommit, pp. 391–401, 2014.

[23] S. Devella, Y. Yohannes, and F. N. Rahmawati, "Implementasi Random Forest Untuk Klasifikasi Motif Songket Palembang Berdasarkan SIFT," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 7, no. 2, pp. 310–320, 2020, doi: 10.35957/jatisi.v7i2.289.

[24] Y. S. Nugroho and N. Emiliyawati, "Sistem Klasifikasi Variabel Tingkat Penerimaan Konsumen Terhadap Mobil Menggunakan Metode Random Forest," *J. Tek. Elektro*, vol. 9, no. 1, pp. 24–29, 2017.

[25] C. Chen, Tianqi; Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[26] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, *A comparative analysis of gradient boosting algorithms*, vol. 54, no. 3. Springer Netherlands, 2021. doi: 10.1007/s10462-020-09896-5.

[27] F. Pratama, Z. Hadryan Nst, Z. Khairi, and L. Efrizoni, "Perbandingan Algoritma Random Forest Dan K-Nearest Neighbor Dalam Klasifikasi Kesehatan Mental Mahasiswa," *J. Ilm. Betrik*, vol. 15, no. 1, pp. 31–37, 2024.

[28] M. Azhari, Z. Situmorang, and R. Rosnelly, "Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes," *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 640, 2021, doi: 10.30865/mib.v5i2.2937.

[29] K. Kristiawan and A. Widjaja, "Perbandingan Algoritma Machine Learning dalam Menilai Sebuah Lokasi Toko Ritel," *J. Tek. Inform. dan Sist. Inf.*, vol. 7, no. 1, pp. 35–46, 2021, doi: 10.28932/jutisi.v7i1.3182.