

Integrasi XGBoost dan Visualisasi Gradio untuk Memprediksi Pendapatan Pembayar Asuransi: Studi Kasus Rumah Sakit Swasta di Manado

Integration of XGBoost and Gradio Visualization for Predicting Insurance Payer Revenue: A Case Study of a Private Hospital in Manado

Cherry Lumingkewas¹, Wilsen Grivin Mokodaser^{*2}

Fakultas Ekonomi dan Bisnis, Universitas Klabat

Fakultas Ilmu Komputer, Universitas Klabat

*E-mail : cherry@unklab.ac.id¹, wilsenm@unklab.ac.id^{*2}*

**Corresponding author*

Received 21 April 2025; Revised 25 April 2025; Accepted 30 April 2025

Abstrak - Prediksi pendapatan yang akurat sangat penting untuk menjaga keberlanjutan finansial, mengoptimalkan alokasi sumber daya, dan membantu perencanaan strategis rumah sakit. Untuk memprediksi pendapatan bulan Agustus, data historis pendapatan digunakan dari Januari hingga Juli 2022. Pra-pemrosesan data termasuk menangani nilai kosong atau nilai yang tidak ada, memilih fitur, dan membagi data menjadi set pelatihan dan pengujian. Metode utama untuk pelatihan model adalah algoritma XGBoost. Hasil evaluasi model diperoleh Mean Squared Error (MSE) dengan hasil 3.239, dan R-squared (R^2) 0.99884, Metrik-metrik ini menunjukkan hasil prediksi yang sangat baik. Ditemukan bahwa data bulan Juni dan Juli memberikan kontribusi terbesar. Sistem dibangun dalam bentuk dashboard interaktif berbasis Gradio untuk meningkatkan aksesibilitas dan pemanfaatan hasil model. Oleh karena itu, manajemen rumah sakit dapat menggunakan solusi ini untuk membuat keputusan yang lebih tepat dan berbasis data. Secara keseluruhan, penelitian ini menunjukkan bahwa penggabungan pembelajaran mesin dan visualisasi interaktif dapat sangat bermanfaat dalam manajemen keuangan rumah sakit. Dengan kata lain, integrasi machine learning ke dalam sistem operasional membuka jalan bagi pengambilan keputusan yang lebih berbasis data (*data-driven decision making*) dan adaptif terhadap dinamika pasar.

Kata kunci – Xgboost, Pendapatan, Prediksi, Gradio

Abstract - Accurate revenue prediction is crucial for maintaining financial sustainability, optimizing resource allocation, and supporting the hospital's strategic planning. To predict revenue for August, historical revenue data from January to July 2022 was used. Data preprocessing included handling missing or null values, feature selection, and splitting the data into training and testing sets. The primary method for training the model was the XGBoost algorithm. The model evaluation results showed a Mean Squared Error (MSE) of 3.239 and an R-squared (R^2) value of 0.99884, indicating excellent prediction performance. It was found that data from June and July contributed the most to the prediction. The system was built as an interactive dashboard using Gradio to enhance accessibility and usability of the model results. Therefore, hospital management can use this solution to make more accurate and data-driven decisions. Overall, this study demonstrates that integrating machine learning and interactive visualization can be highly beneficial for hospital financial management. In other words, the integration of machine learning into operational systems paves the way for more data-driven and adaptive decision-making in response to market dynamics.

Keywords - Xgboost, Pendapatan, Prediksi, Gradio

1. PENDAHULUAN

Masyarakat Indonesia sudah akrab dengan produk keuangan yang disebut asuransi kesehatan. Dari 262 juta penduduk Indonesia, 208 juta, atau sekitar 79,4%, memiliki asuransi kesehatan, menurut data dari Badan Pusat Statistik (BPS) pada tahun 2018. Ini berarti 794 orang dari setiap 1.000 penduduk memiliki asuransi kesehatan, peningkatan dari 716 orang pada tahun 2017.[1] Sangat penting untuk memahami asuransi kesehatan swasta bagi mereka yang mempertimbangkannya. Asuransi kesehatan adalah jenis asuransi yang melindungi pemegang polis atau tertanggung dari biaya medis, bedah, obat-obatan, dan lainnya. Asuransi ini dapat mengganti biaya medis karena sakit atau luka serta membayar biaya perawatan medis secara langsung.[2] Salah satu tujuan utama asuransi pada dasarnya adalah untuk memfasilitasi pengalihan risiko, yang memungkinkan Tertanggung untuk mentransfer risiko kesehatannya kepada Penanggung.[3] Fungsi utama asuransi adalah memberikan pertanggungan dengan premi yang dapat disesuaikan sesuai kebutuhan perlindungan, memungkinkan pemegang polis mendapatkan perlindungan yang sesuai dengan anggaran mereka tanpa harus membayar premi yang tidak proporsional. Selain itu, fungsi utama asuransi dalam pengelolaan risiko adalah mengumpulkan uang dari pemegang polis untuk digunakan untuk membayar klaim saat diperlukan.[4] Jenis asuransi bervariasi menurut kebutuhan, dan memahami jenis-jenisnya membantu seseorang memilih yang paling sesuai dengan keuangan mereka. Apakah itu asuransi kesehatan, jiwa, atau properti, memilih jenis asuransi yang tepat memberikan perlindungan yang lebih khusus dan sesuai dengan kebutuhan individu. Dengan memahami fungsi utama dan sekunder tersebut, kita dapat melihat bahwa asuransi bukan hanya melindungi dari risiko, tetapi juga memainkan peran penting dalam pembangunan ekonomi.[5]

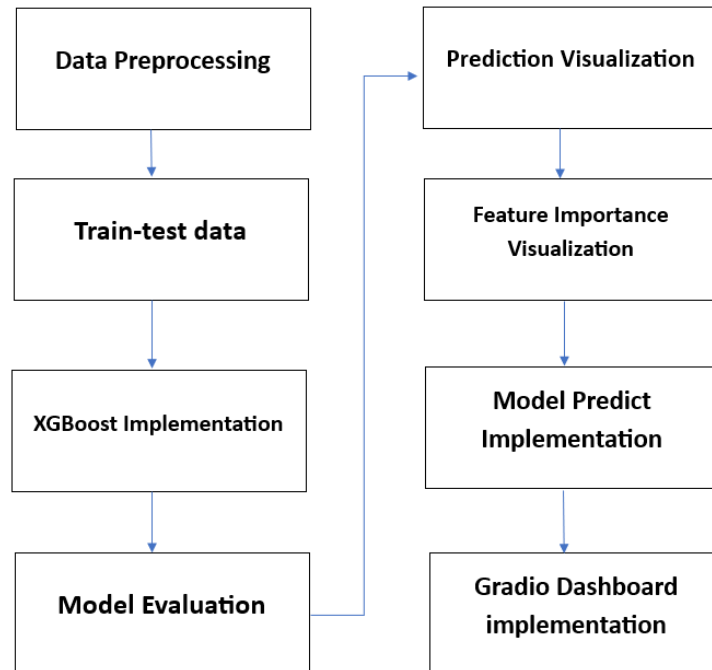
Dengan melihat seberapa pentingnya asuransi dalam membantu untuk memenuhi kebutuhan dari masyarakat Indonesia, maka dari sudut pandang kerjasama dengan pihak rumah sakit juga membuka peluang dalam menjalin kerjasama yang lebih baik. Meskipun sejumlah besar penelitian telah mempelajari prediksi pendapatan dan manajemen keuangan dalam sektor kesehatan, tidak banyak studi yang memberikan perhatian khusus pada komponen pendapatan asuransi kesehatan sebagai salah satu sumber pendanaan utama rumah sakit. Dalam praktiknya, pendapatan asuransi kesehatan sangat dinamis dan berubah-ubah, dan sangat dipengaruhi oleh klaim, peraturan, dan praktik layanan pasien, sehingga prediksinya menjadi kompleks namun penting. Beberapa penelitian yang dilakukan untuk memprediksi pendapatan diantaranya dilakukan untuk prediksi pendapatan penjualan retail dengan kesimpulan algoritma xgboost dapat mengoptimalkan manajemen persediaan, meningkatkan keakuratan perencanaan produksi, dan mengalokasikan sumber daya secara optimal.[6] penelitian terkait lainnya dilakukan untuk memprediksi minat konsumen asuransi dengan metode interpolasi data mining yang mendapatkan hasil yaitu informasi dari basisdata nasabah dan transaksi pembayaran premi asuransi sehingga dapat memprediksi seberapa besar angsuran premi asuransi yang terbaik sesuai dengan kondisi calon nasabah[7] selanjutnya algoritma xgboost juga pernah digunakan dalam penelitian untuk analisis perilaku pembayaran listrik konsumen disandingkan dengan algoritma K-Means dan memperoleh hasil bahwa kombinasi kluster k-means dengan klasifikasi XGBoost serta dengan melakukan penyetelan hyperparameter teknik bayesian memiliki tingkat akurasi model jauh lebih baik dengan nilai 89,27% dan nilai AUC sebesar 0,92 jika dibandingkan dengan metode gradient boosting yang tingkat akurasinya hanya 74,76% dan nilai AUC sebesar 0,75[8] selain itu xgboost juga pernah dilakukan penelitian dalam melakukan klasifikasi keputusan kredit dibandingkan dengan algoritma random forest dan Hasil percobaan menunjukkan bahwa kedua algoritma menghasilkan kinerja model yang cukup kompetitif; XGBoost memiliki akurasi 1.0 untuk semua metrik evaluasi pada ukuran data 10.000 dan 100.000, sedangkan Random Forest memiliki akurasi 0.998 pada ukuran data 10.000 dan 0.999 pada ukuran data 100.000.[9]

Dari penelitian yang telah dibahas sebelumnya maka pada penelitian ini akan membangun sebuah integrasi model XGBoost dan Visualisasi Gradio dalam analisis pendapatan payer asuransi, dengan adanya penelitian ini diharapkan dapat memberikan pandangan lebih baik lagi

tentang pentingnya melakukan peramalan pendapatan asuransi kesehatan guna membantu stakeholder terkait dalam menghadapi kebijakan kerja sama dengan pihak asuransi

2. METODE PENELITIAN

Explaining Pada penelitian ini peneliti membagi kedalam beberapa tahap yang dapat dilihat pada gambar dibawah ini:



Gambar 1. Tahapan penelitian

2.1 Data Preprocessing

Pada tahap preprocessing data, fungsi `dropna()` digunakan untuk menghapus baris-baris dengan nilai kosong (NaN) pada fitur input dan target output. Penghapusan ini sangat penting karena sebagian besar algoritma pembelajaran mesin, termasuk XGBoost yang digunakan dalam penelitian ini, membutuhkan data yang bersih dan lengkap untuk menghasilkan model prediktif yang akurat. Meskipun XGBoost memiliki kemampuan dasar untuk menangani nilai yang tidak ada, proses pelatihan model masih dapat dipengaruhi oleh NaN, terutama saat menggunakan fungsi-fungsi seperti `train_test_split()` atau `fit()` dari pustaka SciKit-learn yang tidak toleran terhadap nilai kosong. Selain itu, karena model harus membuat prediksi berdasarkan data yang tidak lengkap, kehilangan nilai-nilai fitur input dapat menyebabkan bias atau ketidakakuratan. Oleh karena itu, untuk memastikan bahwa model belajar dari data yang valid dan utuh, penghapusan baris dengan nilai kosong dipilih sebagai metode sederhana dan aman.[10]

2.2 Train-test Data

Pemilahan data menjadi data pelatihan (train) dan data pengujian (test) merupakan langkah krusial dalam proses pembangunan model machine learning. Tujuan utamanya adalah untuk mengukur seberapa baik model dapat melakukan generalisasi, yaitu kemampuan model dalam membuat prediksi yang akurat terhadap data baru yang belum pernah dilihat sebelumnya. Data train digunakan untuk “mengajarkan” model pola dan hubungan antara fitur-fitur input dan target output. Sementara itu, data test berfungsi sebagai simulasi kondisi dunia nyata, di mana

model diuji dengan data yang tidak pernah dilihat selama pelatihan. Dengan membandingkan hasil prediksi model pada data test dengan nilai aktualnya, kita dapat mengevaluasi performa model menggunakan metrik seperti Mean Squared Error (MSE) atau R^2 Score.[11] Tanpa pemisahan ini, model berisiko mengalami overfitting, yaitu kondisi di mana model terlalu “hapal” terhadap data latih tetapi gagal memprediksi dengan baik pada data baru. Oleh karena itu, pembagian train-test bukan hanya praktik standar, tetapi juga fondasi penting untuk membangun model prediktif yang andal dan tidak bias.

2.3 XGBoost Implementation

Awal yang sangat sederhana, biasanya terdiri dari pohon keputusan kecil. Selanjutnya, model diuji dengan data pelatihan dan kesalahan prediksi dihitung. XGBoost akan terus menerus menambahkan model-model kecil berikutnya, yang disebut weak learners, untuk meminimalkan kesalahan prediksi dari model sebelumnya. Setiap pohon baru berusaha memperbaiki kesalahan pohon sebelumnya dalam proses yang disebut peningkatan. XGBoost menghitung gradien dari fungsi loss (fungsi kerugian) pada setiap iterasi untuk mengidentifikasi arah perbaikan yang paling efektif. Selain itu, XGBoost menggunakan berbagai teknik regularisasi (seperti L1 dan L2) untuk menghindari overfitting dan melakukan pruning dan optimasi paralel (tidak ada perpanjangan cabang) untuk mempercepat pelatihan dan meningkatkan efisiensi. Setelah beberapa iterasi, semua model kecil tersebut digabungkan menjadi satu model prediktif yang kuat. Model ini memiliki kemampuan untuk membuat estimasi yang sangat akurat terhadap data yang belum pernah dilihat sebelumnya.[12]

2.4 Model Evaluation

Salah satu tahap penting dalam proses pembangunan sistem prediktif adalah evaluasi model, yang bertujuan untuk mengevaluasi seberapa baik model berfungsi terhadap data yang belum pernah dilihat sebelumnya. Setelah model dilatih menggunakan data pelatihan, atau training data, langkah pertama dalam evaluasi adalah mengujinya pada data pengujian, atau test data, untuk mengetahui seberapa baik model dapat digeneralisasi. Dalam regresi, seperti prediksi pendapatan, evaluasi dilakukan dengan membandingkan nilai prediksi model yang dibuat dengan nilai sebenarnya dalam data uji. Average Absolute Error (MAE), Mean Squared Error (MSE), dan R-squared score adalah beberapa metrik regresi yang umum digunakan. R^2 mengukur seberapa baik variabel independen mampu menjelaskan variasi dari variabel dependen, MAE menghitung kesalahan absolut rata-rata dari prediksi, dan MSE menghitung kesalahan yang lebih besar karena menggunakan kuadrat dari selisih.[13] Selain itu, visualisasi sering digunakan untuk memberikan gambaran yang mudah dipahami tentang akurasi model, seperti grafik yang menunjukkan hubungan antara nilai aktual dan nilai prediksi. Selain membandingkan angka, evaluasi model juga berfungsi sebagai dasar untuk menentukan apakah model sudah dapat digunakan secara efektif atau apakah perlu dilakukan optimasi tambahan melalui penyesuaian parameter, pemilihan fitur, atau metode validasi silang.[14]

2.5 Prediction Visualization

Memahami kinerja dan interpretabilitas model prediktif secara intuitif memerlukan tahapan visualisasi prediksi. Setelah model dilatih dan menghasilkan prediksi untuk data uji, tahap pertama visualisasi biasanya dilakukan dengan menggunakan grafik untuk membandingkan nilai prediksi dengan nilai aktual. Grafik sering digunakan dalam regresi untuk menunjukkan tren data historis, hasil prediksi, dan, jika tersedia, nilai aktual dari periode yang diprediksi. Misalnya, jika setiap payer asuransi memiliki prediksi pendapatan bulan Agustus yang didasarkan pada data dari Januari hingga Juli, sebuah grafik dibuat untuk setiap payer. Tujuan dari proses ini adalah untuk membantu pengguna atau pengambil keputusan mengetahui apakah model menangkap pola data dengan baik dan menemukan anomali atau outlier. Visualisasi juga sangat penting untuk membuat

dashboard interaktif yang memungkinkan pengguna memilih entitas tertentu, seperti nama payer, dan melihat hasil prediksi langsung dalam grafik yang informatif dan mudah dipahami. Akibatnya, visualisasi prediksi meningkatkan kejelasan dan nilai praktis dari hasil pemodelan machine learning selain mendukung aspek analitis.[15]

2.6 Feature Importance Visualization

Salah satu langkah penting dalam interpretasi model pembelajaran mesin adalah visualisasi pentingnya fitur. Ini terutama berlaku untuk algoritma seperti XGBoost, yang rumit tetapi memiliki kemampuan bawaan untuk menilai kontribusi masing-masing fitur terhadap hasil prediksi. Setelah model dilatih, XGBoost menghitung skor pentingnya setiap fitur berdasarkan seberapa sering dan besar kontribusinya dalam proses pembentukan pohon keputusan. Skor ini menunjukkan seberapa besar pengaruh variabel input, seperti pendapatan bulan Januari hingga Juli, terhadap nilai prediksi, seperti pendapatan bulan Agustus. Selanjutnya, hasil perhitungan disajikan dalam bentuk grafik batang. Nilai kepentingan relatif (skor kepentingan) ditunjukkan pada sumbu horizontal, dan nama-nama fitur ditunjukkan pada sumbu vertikal. Visualisasi ini membantu peneliti dan praktisi memahami aspek mana yang paling penting untuk hasil prediksi. Dengan demikian, mereka dapat menggunakannya dalam proses pengambilan keputusan strategis atau optimalisasi model. Misalnya, jika pendapatan bulan Juli mencapai nilai signifikan tertinggi, maka dapat disimpulkan bahwa tren terbaru memengaruhi prediksi. Oleh karena itu, visualisasi fitur penting tidak hanya memperjelas struktur model tetapi juga meningkatkan interpretabilitas dan kejelasan tentang penggunaan machine learning dalam dunia nyata.[16]

2.7 Model Predict Implementation

Dalam penelitian ini, tahap implementasi prediksi model adalah proses penerapan model pembelajaran mesin yang telah dilatih untuk menghasilkan perkiraan atau output berdasarkan data input baru. Dalam tahap ini, setelah model XGBoost dilatih menggunakan data historis pendapatan, model siap digunakan untuk memprediksi pendapatan pada bulan selanjutnya. Tahap ini dimulai dengan mengumpulkan atau memilih baris data yang memiliki informasi yang diperlukan untuk memprediksi pendapatan.[17] Selanjutnya, fungsi `predict()` objek model XGBoost dijalankan untuk menghasilkan nilai prediksi. Untuk mendapatkan interpretasi yang lebih baik, hasil prediksi dapat dianalisis atau divisualisasikan segera. Ketika pengguna memilih entitas tertentu, seperti nama payer asuransi, dashboard interaktif menampilkan hasil prediksi secara real-time. Ini membuat proses prediksi lebih dinamis dan lebih mudah digunakan. Tahapan ini menunjukkan bagaimana model pembelajaran mesin dapat diintegrasikan ke dalam aplikasi nyata untuk membantu prediktif dan pengambilan keputusan berbasis data.

2.7 Gradio Dashboard Implementation

Tahapan implementasi Gradio adalah proses mengintegrasikan model pembelajaran mesin ke dalam antarmuka pengguna berbasis web yang interaktif dan ramah pengguna. Dengan menggunakan Gradio untuk menghubungkan model prediktif dengan pengguna akhir, proses prediksi dapat dilakukan secara real-time tanpa memerlukan pengetahuan teknis tentang kode program.[18] Tahapan awal dimulai dengan mendefinisikan fungsi Python yang akan digunakan ketika pengguna memberikan input. Fungsi ini memungkinkan untuk memuat data dari payer tertentu, membuat prediksi pendapatan, dan menampilkan visualisasi dari data tersebut. Selanjutnya, komponen `gr.Dropdown` digunakan untuk mengkonfigurasi elemen input seperti dropdown menu, sehingga pengguna dapat memilih nama payer yang ingin dianalisis. Parameter output `gr.Interface` digunakan untuk menyusun elemen output yang terdiri dari grafik dan teks deskriptif. Setelah semua komponen terintegrasi, antarmuka dibuat menggunakan `gr.Interface()` dan dijalankan melalui `launch()`. Ini memungkinkan dashboard dibuka secara lokal dan dibagikan

melalui link publik. Metode ini memungkinkan hasil prediksi yang sebelumnya hanya dapat dilihat melalui pemrograman untuk ditampilkan secara visual dan interaktif. Metode ini juga meningkatkan kemudahan penggunaan dan mendukung pemanfaatan model oleh pihak non-teknis yang terlibat dalam pengambilan keputusan berbasis data

3. HASIL DAN PEMBAHASAN

Tahap implementasi dari penelitian ini dimulai dengan struktur data yang akan dilakukan pengelolaan dengan algoritma XGBoost, data dapat dilihat pada gambar berikut:

Payer	Januari	Februari	Maret	April	Mei	Juni	Juli	Agustus
ASURANSI EQUITY LIFE INDONESIA - ADMEDICA	308,083	88,743,701	107,228,222	7,797,961	2,087,224	2,072,229	2,072,221	2,087,220
EQUITY LIFE INDONESIA	91,798,000	48,448,100	41,000,000	240,000	7,744,800	7,744,800	7,744,800	7,744,800
AA INTERNATIONAL HUB SUN BHU	7,400,000	2,000,000	204,640	375,000	383,420	350,000	354,800	384,000
ADMEDICA, PT - ASURANSI JMA SVARTAIL, PT	1,000,000	3,300,000	4,000,000	440,000	2,220,000	2,220,000	2,220,000	2,220,000
ADMINISTRASI NEDOKA, PT	1,000,000	2,000,000	1,000,000	1,000,000	2,000,000	2,000,000	2,000,000	2,000,000
ASURANSI ANIMA BUANA, PT (SARANA MEDIKA)	5,000,000	72,000,000	212,000,000	800,000	100,000	100,000	100,000	100,000
ASURANSI AXA INDONESIA - ADMEDICA	100,000,000	48,000,000	30,000,000	1,000,000	100,000	100,000	100,000	100,000
ASURANSI COMA, PT	200,000	1,000,000	600,000	2,000,000	1,000,000	1,000,000	1,000,000	1,000,000
ASURANSI JIWA ASTRA, PT (RAPID TEST)	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000
AXA MALAYSIA (AXA ASSISTANCE)	80,000	80,000	80,000	80,000	80,000	80,000	80,000	80,000
BINA MUSAHAHA - ADMEDICA	3,000,000	3,000,000	3,000,000	3,000,000	3,000,000	3,000,000	3,000,000	3,000,000
FULLERTON HEALTH INDONESIA - DONGGI SENORO	3,000,000	3,000,000	3,000,000	3,000,000	3,000,000	3,000,000	3,000,000	3,000,000
HSD SELATAN HOSPITALS LEPPO VILLAGE	10,000,000	10,000,000	10,000,000	10,000,000	10,000,000	10,000,000	10,000,000	10,000,000
HSD SELATAN HOSPITALS S-KARIRI	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000
INTL SOS (PT ASIH EKA ABADI) - FREEDPORT IND	40,000,000	40,000,000	40,000,000	40,000,000	40,000,000	40,000,000	40,000,000	40,000,000
INTERBANK LAYANAN GLOBAL RSK - ADMEDICA	300,000	300,000	300,000	300,000	300,000	300,000	300,000	300,000
LEPO GENERAL INSURANCE TRAC, PT - TELICONSULTATION	25,000,000	25,000,000	25,000,000	25,000,000	25,000,000	25,000,000	25,000,000	25,000,000
MANDORI AXA GENERAL INSURANCE, PT	10,000,000	10,000,000	10,000,000	10,000,000	10,000,000	10,000,000	10,000,000	10,000,000
MANIKULA (DIA KIRKANA, PT)	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000
PENGHUBUNG KAWILAYAT SULAWESI UTARA	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000
PLN (PERUSERO), PT - ASURANSI PERISAI LISTRIK NASIONAL (APLN), PT	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000
PLN (PERUSERO), PT - ASURANSI PERISAI LISTRIK NASIONAL (PILN), PT	40,000,000	40,000,000	40,000,000	40,000,000	40,000,000	40,000,000	40,000,000	40,000,000
PRUDENTIAL LIFE ASSURANCE, PT	10,000,000	10,000,000	10,000,000	10,000,000	10,000,000	10,000,000	10,000,000	10,000,000
SH AMSON (HSD) - JHEDETA	400,000,000	277,000,000	414,000,000	2,000,000	600,000	600,000	600,000	600,000
SRI MAKASSAR (FAMILY)	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000
TDOT.COM - PERICO	100,000	100,000	100,000	100,000	100,000	100,000	100,000	100,000
WELLES TOWERS WATSON, PT - VERA CHERY INDONESIA, PT	100,000	100,000	100,000	100,000	100,000	100,000	100,000	100,000

Gambar 2. Struktur Data

Data yang akan diimplementasikan seperti pada gambar 2 merupakan data pendapatan dalam bentuk urutan waktu yang terdiri dari nama payer yang merupakan asuransi swasta kemudian total pendapatan masing-masing payer mulai dari bulan Januari sampai dengan Agustus.

3.1 Data Preprocessing

```

features = ['Januari', 'Februari', 'Maret', 'April', 'Mei', 'Juni', 'Juli']
target = 'Agustus'

df_clean = df[features + [target]].dropna()
X = df_clean[features]
y = df_clean[target]

```

Gambar 3. Data Cleansing

Pada tahap ini seperti yang dilihat pada gambar 3, dilakukan proses seleksi fitur dan target dari dataset yang telah dibersihkan. Baris kode `features = ['Januari', 'Februari', 'Maret', 'April', 'Mei', 'Juni', 'Juli']` menunjukkan bahwa variabel `features` berisi daftar kolom yang merepresentasikan data pendapatan dari bulan Januari hingga Juli, yang akan dijadikan sebagai input atau variabel prediktor (independen) untuk model pembelajaran mesin. Sementara itu, variabel `target = 'Agustus'` mendefinisikan kolom 'Agustus' sebagai variabel target (dependen) yang ingin diprediksi. Selanjutnya, baris `df_clean = df[features + [target]].dropna()` digunakan untuk membuat subset baru dari dataframe asli `df` dengan hanya mengambil kolom-kolom yang dibutuhkan (Januari sampai Agustus). Kemudian diterapkan fungsi `dropna()` untuk menghapus baris-baris yang memiliki nilai kosong (NaN) pada salah satu kolom yang disebutkan, guna menjaga kualitas dan integritas data yang akan digunakan dalam pelatihan model. Setelah data

bersih diperoleh, dilakukan pemisahan antara fitur dan target. Baris `X = df_clean[features]` menyimpan data input (fitur bulan Januari–Juli) ke dalam variabel `X`, sedangkan `y = df_clean[target]` menyimpan data target (bulan Agustus) ke dalam variabel `y`. Proses ini sangat penting untuk membentuk struktur data yang siap digunakan dalam pelatihan model XGBoost, di mana `X` akan digunakan untuk membangun pola prediktif dan `y` sebagai acuan keluaran yang ingin dipelajari oleh model.

3.2 Train-test Data

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Gambar 4. implementasi train-test data

Pada gambar 4 dilakukan pembagian dataset menjadi data pelatihan dan data pengujian. Fungsi `train_test_split` dari library `scikit-learn` digunakan untuk secara acak membagi data fitur (`X`) dan target (`y`) ke dalam dua bagian: 80% untuk data pelatihan (`X_train`, `y_train`) dan 20% untuk data pengujian (`X_test`, `y_test`), sebagaimana ditentukan oleh parameter `test_size=0.2`. Tujuan utama dari pembagian ini adalah untuk melatih model pada satu subset data (training) dan kemudian mengukur performa prediksi model pada data yang tidak pernah dilihat sebelumnya (testing). Hal ini bertujuan untuk menghindari *overfitting*, yaitu kondisi di mana model terlalu baik pada data pelatihan namun gagal melakukan generalisasi terhadap data baru. Selain itu, parameter `random_state=42` digunakan untuk mengatur seed dari pengacakan, agar proses pembagian data menjadi konsisten setiap kali kode dijalankan. Ini memastikan bahwa hasil eksperimen dapat direproduksi dengan membagi data yang sama pada setiap eksekusi. Dengan cara ini, validitas dan reliabilitas evaluasi model dapat terjamin dalam proses pengujian performa prediksi terhadap data nyata.

3.3 XGBoost Implementation

```
XGBRegressor(base_score=None, booster=None, callbacks=None,
               colsample_bylevel=None, colsample_bynode=None,
               colsample_bytree=None, device=None, early_stopping_rounds=None,
               enable_categorical=False, eval_metric=None, feature_types=None,
               gamma=None, grow_policy=None, importance_type=None,
               interaction_constraints=None, learning_rate=None, max_bin=None,
               max_cat_threshold=None, max_cat_to_onehot=None,
               max_delta_step=None, max_depth=3, max_leaves=None,
               min_child_weight=None, missing=nan, monotone_constraints=None,
               multi_strategy=None, n_estimators=50, n_jobs=None,
               num_parallel_tree=None, random_state=42, ...)
```

Gambar 5. implementasi algoritma

Tahapan ini merupakan inti dari proses pembelajaran mesin, yaitu pembangunan dan pelatihan model prediktif menggunakan algoritma XGBoost Regressor. Pada baris pertama model = `XGBRegressor(...)` dengan menggunakan parameter `objective='reg:squarederror'` menunjukkan bahwa model ini dikonfigurasi untuk menyelesaikan masalah regresi, dengan fungsi objektif berupa mean squared error (MSE), yang menghitung selisih kuadrat antara nilai prediksi dan nilai aktual. Parameter `n_estimators=50` menentukan jumlah total pohon keputusan (trees) yang akan

dibangun dalam proses boosting; setiap pohon berkontribusi secara bertahap untuk memperbaiki kesalahan dari pohon sebelumnya. Sementara itu, `max_depth=3` membatasi kedalaman masing-masing pohon hingga tiga level untuk mengontrol kompleksitas model dan mencegah overfitting. Parameter `random_state=42` digunakan untuk memastikan bahwa hasil pelatihan bersifat deterministik dan dapat direproduksi. Kemudian dilanjutkan dengan `model.fit(X_train, y_train)` yang merupakan proses pelatihan model di mana XGBoost mempelajari pola hubungan antara fitur-fitur (Januari–Juli) dengan target variabel (Agustus) menggunakan data pelatihan. Proses ini menghasilkan model yang mampu melakukan prediksi pendapatan payer berdasarkan data historis sebelumnya.

3.4 Model Evaluation

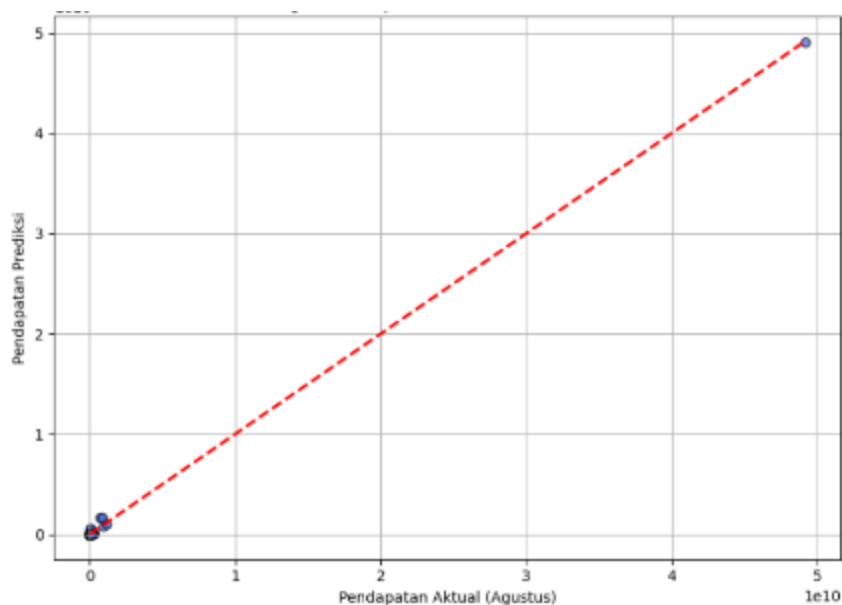
```
print(f"📊 Mean Squared Error (MSE): {mse}")
print(f"📊 R2 Score: {r2}")
```

📊 Mean Squared Error (MSE): 3.2398596979532576e+16
 📊 R² Score: 0.9988455728401666

Gambar 6. hasil evaluasi model

Gambar 6 merupakan hasil evaluasi model XGBoost dimana menunjukkan performa yang sangat baik dalam melakukan prediksi terhadap pendapatan payer asuransi untuk bulan Agustus berdasarkan data historis bulan Januari hingga Juli. Nilai Mean Squared Error (MSE) yang diperoleh sebesar 3.2398596979532576 mencerminkan rata-rata kuadrat dari selisih antara nilai prediksi dan nilai aktual. Meskipun angka ini terlihat besar secara absolut, hal tersebut masih dapat diterima apabila skala pendapatan yang diprediksi memang besar, seperti dalam satuan jutaan atau miliaran rupiah. Lebih penting lagi, nilai R-squared (R^2) yang mencapai 0.9988 menunjukkan bahwa model mampu menjelaskan sebesar 99.88% variasi dari data target, yang berarti hampir seluruh pola dan tren dalam data berhasil dipelajari oleh model. Nilai R^2 yang sangat tinggi ini menjadi indikator bahwa model memiliki tingkat akurasi dan kemampuan generalisasi yang sangat baik dalam konteks dataset yang digunakan. Oleh karena itu, dapat disimpulkan bahwa model XGBoost yang telah dibangun sangat layak digunakan untuk keperluan prediktif dalam pengambilan keputusan, khususnya dalam konteks estimasi pendapatan bulanan berdasarkan tren historis.

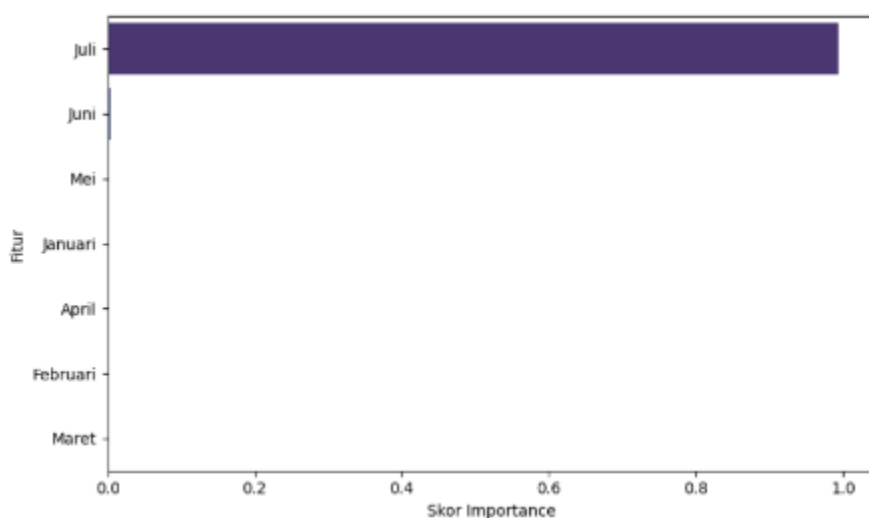
3.5 Prediction Visualization



Gambar 7. Visualisasi hasil prediksi

Visualisasi yang ditampilkan pada gambar 7 menunjukkan hubungan antara pendapatan aktual bulan Agustus dan pendapatan yang diprediksi oleh model XGBoost. Pada grafik tersebut, titik-titik biru merepresentasikan pasangan nilai aktual dan prediksi untuk masing-masing data uji, sedangkan garis merah putus-putus merupakan garis referensi ideal yang menunjukkan posisi di mana nilai prediksi sama persis dengan nilai aktual. Sebagian besar titik tampak sangat dekat atau bahkan berada tepat di atas garis ini, yang menandakan bahwa prediksi model memiliki tingkat akurasi yang sangat tinggi. Tidak terlihat adanya penyimpangan besar atau outlier yang menonjol, sehingga dapat disimpulkan bahwa model tidak menunjukkan kesalahan sistematis dan memiliki kemampuan generalisasi yang baik. Visualisasi ini selaras dengan hasil evaluasi sebelumnya, yaitu nilai R^2 sebesar 0.9988, yang menunjukkan bahwa hampir seluruh variasi dalam data dapat dijelaskan oleh model. Dengan demikian, grafik ini menjadi bukti visual yang kuat bahwa model XGBoost yang diterapkan sangat andal dalam memprediksi pendapatan bulanan berdasarkan data historis.

3.6 Feature Importance Visualization



Gambar 8. Visualisasi feature importance

Visualisasi yang ditampilkan pada gambar 8 merupakan grafik Feature Importance dari model XGBoost yang digunakan untuk memprediksi pendapatan bulan Agustus. Grafik ini menunjukkan seberapa besar kontribusi masing-masing fitur (dalam hal ini pendapatan dari bulan Januari hingga Juli) terhadap proses pembentukan prediksi.

Dari grafik tersebut terlihat bahwa fitur 'Juli' memiliki skor importance yang sangat dominan dibandingkan fitur-fitur lainnya. Skor ini hampir mencapai nilai maksimum (1.0), sedangkan fitur lainnya seperti Juni, Mei, April, dan seterusnya memiliki nilai mendekati nol. Artinya, model XGBoost sangat mengandalkan nilai pendapatan bulan Juli untuk memprediksi pendapatan bulan Agustus, sementara kontribusi bulan-bulan sebelumnya dianggap sangat kecil atau bahkan tidak signifikan dalam pengambilan keputusan oleh model.

Secara praktis, hal ini masuk akal apabila pola pendapatan dari satu bulan ke bulan berikutnya bersifat konsisten atau mengalami tren berurutan, di mana data bulan Juli bisa mencerminkan kondisi keuangan terkini dari masing-masing payer. Namun, dominasi satu fitur ini juga bisa menjadi indikasi bahwa model terlalu fokus pada data terakhir, yang dalam beberapa kasus bisa meningkatkan akurasi, tapi juga dapat menyebabkan overfitting jika data mendadak berubah. Oleh karena itu, interpretasi feature importance ini penting untuk mempertimbangkan keseimbangan data historis dalam pelatihan model prediktif.

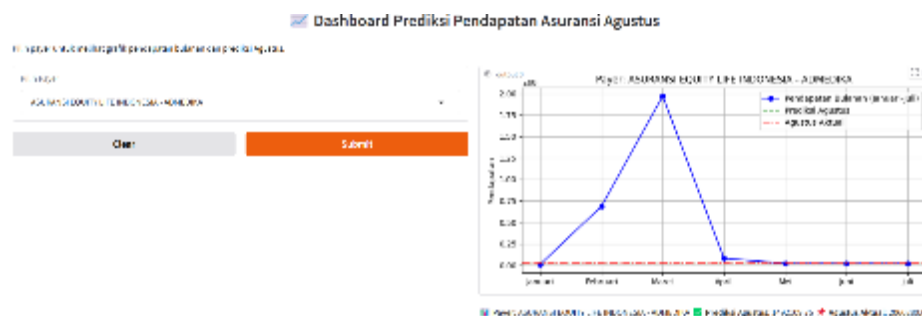
3.7 Model Predict Implementation

	Payer	Januari	Februari	Maret	April	Mei	Juni	Juli	Agustus_M aktual	Prediksi_Agustus
41	KHUSNADAH KHEITIF (KHUSNADAH KHEITIF)	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
1	ELITE (ELITE)	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
2	AK. MERSING (AK. MERSING)	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
3	ALBIR (ALBIR)	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
4	ALBIR (ALBIR)	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
418	YASOGAN (YASOGAN)	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
419	YOGI YASOGAN (YOGI YASOGAN)	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
420	YOGI YASOGAN (YOGI YASOGAN)	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
421	YOGI YASOGAN (YOGI YASOGAN)	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
422	YOGI YASOGAN (YOGI YASOGAN)	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00

Gambar 9. Implementasi model

Tahapan hingga menghasilkan nilai prediksi seperti yang ditunjukkan pada gambar 9 dimulai dengan melakukan penyusunan hasil akhir dari prediksi model XGBoost. Langkah pertama yang dilakukan adalah memastikan bahwa kolom identitas, yaitu Payer, tersedia dalam dataset asli (df). Jika kolom tersebut ada, maka nilainya akan disalin ke dalam dataframe hasil (df_all) dengan mencocokkan indeks, sehingga setiap baris hasil prediksi tetap terkait dengan nama payer asuransi yang sesuai. Selanjutnya, dilakukan hal serupa untuk kolom target (Agustus). Jika kolom ini tersedia dalam dataset awal, maka nilai aktual pendapatan bulan Agustus juga ditambahkan ke df_all berdasarkan urutan indeks, sehingga kita bisa melakukan perbandingan langsung antara nilai aktual dan prediksi. Setelah proses penggabungan ini selesai, sistem menampilkan hasil akhir dalam bentuk tabel yang mencakup nama payer, data historis pendapatan dari Januari hingga Juli, nilai aktual Agustus, serta hasil prediksi Agustus. Langkah ini memberikan konteks penuh terhadap performa model, memungkinkan pengguna untuk mengevaluasi hasil prediksi secara lebih mudah dan informatif.

3.8 Gradio Dashboard Implementation



Gambar 10. Hasil pembuatan dashboard

Pada tahapan ini model yang telah dihasilkan akan diimplementasikan kedalam bentuk visualisasi dashboard interaktif yang akan mempermudah user dalam melakukan interpretasi hasil prediksi dari masing-masing payer asuransi. Dengan memanfaatkan library Gradio dibuatlah sebuah pilihan dalam bentuk menu dropdown yang akan menampilkan payer asuransi, kemudian di sediakan tombol untuk clear dan submit, saat tombol submit dipilih maka visualisasi hasil prediksi akan ditampilkan pada grafik yang terdiri dari prediksi jumlah total pendapatan setiap bulan dan pendapatan prediksi bulan selanjutnya, sementara garis merah menunjukkan total pendapatan actual dari bulan yang sebenarnya.

4. KESIMPULAN

Penerapan algoritma XGBoost dalam memprediksi pendapatan bulanan per payer asuransi menunjukkan hasil yang sangat menjanjikan. Melalui proses pelatihan menggunakan data historis pendapatan dari bulan Januari hingga Juli, model mampu melakukan prediksi pendapatan bulan Agustus dengan tingkat akurasi yang baik, sebagaimana dibuktikan melalui evaluasi menggunakan metrik regresi seperti Mean Absolute Error (MAE), Mean Squared Error (MSE), dan R-squared (R^2). Visualisasi prediksi menunjukkan bahwa model dapat menangkap pola tren pendapatan secara konsisten dan memberikan estimasi yang logis terhadap nilai-nilai baru. Selain itu, fitur-fitur seperti pendapatan bulan Juni dan Juli teridentifikasi sebagai variabel paling berpengaruh berdasarkan analisis *feature importance*, menunjukkan bahwa tren terbaru memberikan kontribusi signifikan dalam pembentukan keputusan model. Integrasi model ke dalam dashboard interaktif berbasis Gradio juga membuktikan bahwa solusi prediktif ini dapat diakses dengan mudah oleh pengguna non-teknis, serta siap diterapkan dalam konteks pengambilan keputusan bisnis secara real-time. Secara keseluruhan, penggunaan XGBoost terbukti efektif, efisien, dan dapat diandalkan dalam tugas prediktif berbasis data historis finansial, serta memberikan potensi besar untuk pengembangan sistem prediksi otomatis dalam skala yang lebih luas. Dari sisi bisnis, implementasi sistem prediksi ini memberikan nilai tambah yang signifikan. Manajemen Rumah Sakit yang bekerja sama dengan pihak asuransi kini dapat mengantisipasi fluktuasi pendapatan lebih awal, sehingga dapat merencanakan strategi keuangan, alokasi sumber daya, dan kebijakan pemasaran dengan lebih presisi. Dashboard interaktif yang dibangun dengan Gradio juga memungkinkan pihak manajemen untuk mengevaluasi kinerja tiap payer secara cepat dan responsif tanpa harus memahami kompleksitas teknis dari model. Dengan kata lain, integrasi machine learning ke dalam sistem operasional membuka jalan bagi pengambilan keputusan yang lebih berbasis data (*data-driven decision making*) dan adaptif terhadap dinamika pasar.

DAFTAR PUSTAKA

- [1] B. P. Statistik, "Jumlah Penduduk Yang Dicakup Asuransi Kesehatan Atau Sistem Kesehatan Masyarakat," *Badan Pusat Statistik*, 2017. <https://www.bps.go.id/id/statistics-table/2/MTQzMyMy/jumlah-penduduk-yang-dicakup-asuransi-kesehatan-atau-sistem->

- kesehatan-masyarakat.html (accessed Mar. 13, 2025).
- [2] Y. F. Laturrahmi, S. Swastikawara, and N. Wardasari, "Analisis Perilaku Masyarakat Pedesaan Terhadap Asuransi Kesehatan Nasional Dalam Perspektif Komunikasi Kesehatan," *Komuniti J. Komun. dan Teknol. Inf.*, vol. 12, no. 2, pp. 87–100, 2020, doi: 10.23917/komuniti.v12i2.10040.
 - [3] D. T. Fransiska Andriyani Odos, Chandra Setya Wijaya, Miqsyah Khalid, "Prediksi Minat Dan Pembelian Asuransi Syariah Di Kalangan Aparatur Sipil Negara (Asn) Di Provinsi Aceh," *Edunomika*, vol. 08, no. 01, pp. 1–21, 2023.
 - [4] N. Rullisha, "Pengaruh Pendapatan, Resiko, Premi dan informasi Terhadap Preferensi Masyarakat Golongan Menengah Ke Atas Pada Produk Asuransi," *J. Ilm.*, vol. 3, no. 1, p. 8, 2015, [Online]. Available: <https://jimfeb.ub.ac.id/index.php/jimfeb/article/view/1711>
 - [5] S. Wahyuningsih, S. P. Ediwijoyo, P. P. Ganesha, and J. Tengah, "Jurnal E-Bis : Ekonomi Bisnis Kajian Prediksi Kebangkrutan Industri Asuransi Di Indonesia Tahun 2019-," *E-Bisnis Ekon. Bisnis*, vol. 6, no. 2, pp. 555–570, 2022.
 - [6] H. Wijaya, D. P. Hostiadi, and E. Triandini, "Meningkatkan Prediksi Penjualan Retail Xyz Dengan Teknik Optimasi Random Search Pada Model Xgboost," *SPINTER (Prosiding Semin. Has. Penelit. Inform. dan Komputer)*, vol. 1, no. 2, pp. 829–833, 2024.
 - [7] S. Kurniawan and T. Hidayat, "Penerapan data mining dengan metode interpolasi untuk memprediksi minta konsumen asuransi (Studi Kasus Asuransi Metlife)," *Media Inform.*, vol. 5, no. 2, pp. 113–128, 2007, [Online]. Available: <https://journal.uui.ac.id/media-informatika/article/view/114>
 - [8] R. H. Nugraha, D. Purwitasari, and A. B. Raharjo, "K-Means Dan Xgboost Untuk Analisis Perilaku Pembayaran Rekening Listrik Pelanggan (Studi Kasus : Pln Ulp Panakkukang)," *JUTI J. Ilm. Teknol. Inf.*, vol. 20, no. 2, pp. 84–98, 2022.
 - [9] Jan Melvin Ayu Soraya Dachi and Pardomuan Sitompul, "Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest Ensemble Learning pada Klasifikasi Keputusan Kredit," *J. Ris. Rumpun Mat. Dan Ilmu Pengetah. Alam*, vol. 2, no. 2, pp. 87–103, 2023, doi: 10.55606/jurrimipa.v2i2.1470.
 - [10] A. R. F. Riska Yulianti, Erdanisa Aghnia Ilmani, Megawati Zein Waliulu, Bagus Sartono, "PERBANDINGAN ALGORITMA RANDOM FOREST DAN XGBOOST DALAM KLASIFIKASI PENERIMA BANTUAN PANGAN NON-TUNAI," *Lebesgue J. Ilm. Pendidik. Mat. Mat. dan Stat.*, vol. 6, no. 1, pp. 21–34, 2025.
 - [11] L. N. Wakhidah, A. K. Zyen, and B. B. Wahono, "Evaluation of Telecommunication Customer Churn Classification with SMOTE Using Random Forest and XGBoost Algorithms," *J. Appl. Informatics Comput.*, vol. 9, no. 1, pp. 89–95, 2025.
 - [12] A. V. Arfian Haris Prayoga, "Pengembangan Aplikasi Bank Account Fraud Detection," *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 8, no. 3, pp. 2916–2922, 2024.
 - [13] Z. A. Tarigan, Z. Ameta Tarigan, and J. R. Sagala, "Peramalan (Forecasting) Jumlah Kunjungan Pasien Di Klinik Kasih Ibu Menggunakan Metode Weight Moving Average," *J. Media Inform. [Jumin]*, vol. 3, no. 1, pp. 38–44, 2021, [Online]. Available: <http://ejournal.sisfokomtek.org/index.php/jumin>
 - [14] Y. Aryani, "Sistem Informasi Penjualan Barang Dengan Metode Regresi Linear Berganda Dalam Prediksi Pendapatan Perusahaan," *J. Ris. Sist. Inf. dan Teknol. Inf.*, vol. 2, no. 2, pp. 39–51, 2020, doi: 10.52005/jursistekni.v2i2.47.
 - [15] M. Garma, A. Rianto, A. L. Prasasti, and A. Novianty, "Implementasi Model XGBoost untuk

- Prediksi Jumlah Transaksi dan Total Pendapatan di Jaringan Restoran CV Balibul,” *J. Nas. SAINS dan Tek.*, vol. 2, no. 2, pp. 43–46, 2024.
- [16] I. Muhammad, R. Dahlia, Muhammad Ifan Rifani Ihsan, Lisnawanty, and Rabiatus Sa’adah, “Performance Analysis of Ensemble Learning and Feature Selection Methods in Loan Approval Prediction at Banks,” *J. Artif. Intell. Eng. Appl.*, vol. 3, no. 2, pp. 557–564, 2024, doi: 10.59934/jaiea.v3i2.426.
- [17] F. Aloysia, P. Prasetya, and P. H. P. Rosa, “Klasifikasi Kegagalan Pembayaran Kredit Nasabah Bank dengan Algoritma XGBoost,” *Semin. Nas. Inform. Bela Negara*, vol. 4, pp. 110–115, 2024.
- [18] H. A. C. Utomo, Y. M. Saputra, and A. Prasetiadi, “Implementasi Sistem Konfigurasi Router Berbasis Natural Language Processing dengan Pendekatan Low Rank Adaptation Finetuning dan 8-Bit Quantization,” *J. Internet Softw. Eng.*, vol. 4, no. 2, pp. 1–7, 2023, doi: 10.22146/jise.v4i2.9093.