Idealis: Indonesia Journal Information System
Volume 8, Nomor 2, Juli 2025
ISSN 2684-7280 (*online*)
Halaman 267-276
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

MODEL RANDOM FOREST DATA HISTORIS MULTIVARIAT UNTUK PREDIKSI PENDAPATAN ASURANSI

Wilsen Grivin Mokodaser[1*], Hartiny Pop Koapaha[2], Stenly Ibrahim Adam[3]
[1,3]Informatika, Fakultas Ilmu Komputer, Universitas Klabat, Airmadidi, Indonesia
[2]Akuntansi, Fakultas Ekonomi dan Bisnis, Universitas Klabat, Airmadidi, Indonesia
Email: [1*]wilsenm@unklab.ac.id, [2]hartinikoapaha@unklab.ac.id, [3]stenly.adam@unklab.ac.id

Abstrak-Perusahaan asuransi adalah perusahaan keuangan non-bank yang melindungi nasabah dari risiko dan mengumpulkan uang dari premi nasabah selama periode tertentu, sesuai dengan ketentuan polis. Karena perusahaan asuransi telah lama terlibat dalam perekonomian negara, masyarakat tidak begitu ragu akan layanan yang mereka tawarkan. Disebabkan oleh ketidakpastian yang terkait dengan hal-hal seperti kesehatan, pendidikan, harta-benda, dan kematian, kesadaran masyarakat tentang pentingnya asuransi terus meningkat. Asuransi menjadi alat penting bagi masyarakat untuk mengantisipasi risiko atau kerugian di masa depan. model *Random Forest* diterapkan untuk memprediksi pendapatan asuransi bulan berikutnya berdasarkan data historis multivariat dari bulan Januari hingga Juli/Agustus. Hasil evaluasi menunjukkan bahwa model memiliki performa yang cukup baik dalam menangkap pola pendapatan, dengan skor evaluasi *Mean Absolute Error* (MAE) sebesar ±25.139.426 menunjukkan bahwa rata-rata kesalahan prediksi hanya sekitar 25 juta rupiah, angka yang masih tergolong wajar jika dibandingkan dengan skala pendapatan keseluruhan. *Mean Squared Error* (MSE) sebesar 2.9815×10^{15} mencerminkan adanya beberapa *error* besar, meskipun hal ini wajar mengingat skala data dan keberadaan outlier yang sulit dihindari. R^2 *Score* sebesar 0.85 menandakan bahwa 85% variabilitas pendapatan dapat dijelaskan oleh model dari data historis, yang menunjukkan performa prediksi yang sangat baik. Kontribusi ilmiah dari penelitian ini adalah penerapan pendekatan regresi non-linear berbasis *Random Forest* untuk melakukan peramalan pendapatan asuransi menggunakan data multivariat historis bulanan, yang jarang dibahas secara mendalam dalam konteks industri asuransi. Pendekatan ini tidak hanya menyoroti efektivitas *Random Forest* dalam menangkap pola musiman dan hubungan non-linier antar variabel waktu, tetapi memberikan landasan eksplorasi metode *machine learning* lanjutan dalam analisis data asuransi.

Kata Kunci: asuransi, *random forest*, multivariat, prediksi, variabilitas.

Abstract-*Insurance companies, as non-bank financial institutions, provide risk protection by collecting customer premiums over a specified policy period. Their long-standing role in the economy has fostered public trust, while growing uncertainties in health, education, property, and mortality have heightened awareness of insurance's importance as a risk mitigation tool. This study employs a Random Forest model to predict monthly insurance revenue using multivariate historical data from January to July/August. The model demonstrates strong performance in capturing revenue patterns, with a Mean Absolute Error (MAE) of ±25,139,426, indicating an average prediction deviation of approximately 25 million rupiah—a reasonable margin given the revenue scale. The Mean Squared Error (MSE) of 2.9815×10^{15} suggests some significant errors, which is expected due to data scale and inherent outliers. Additionally, the model achieves an R^2 score of 0.85, explaining 85% of revenue variability based on historical data, reflecting robust predictive accuracy. The study's key contribution lies in applying a Random Forest-based nonlinear regression approach to insurance revenue forecasting, an underexplored area in the industry. This method effectively captures seasonal trends and nonlinear relationships between time-dependent variables, demonstrating the algorithm's suitability for complex financial data. Furthermore, the findings establish a foundation for advancing machine learning applications in insurance analytics, offering potential for enhanced predictive modeling and strategic decision-making. The results underscore the viability of Random Forest in revenue prediction while highlighting opportunities for further refinement using advanced techniques.*

Keywords*: insurance, random forest, multivariate, prediktion, variabily.*

1. PENDAHULUAN

Perusahaan asuransi adalah perusahaan keuangan non-bank yang melindungi nasabah dari risiko dan mengumpulkan uang dari premi nasabah selama periode tertentu, sesuai dengan ketentuan polis. Risiko adalah ketidakpastian yang dapat menyebabkan sesuatu yang buruk. Perusahaan asuransi memerlukan dana yang sangat besar untuk menutupi seluruh risiko tersebut. Oleh karena itu, agar perusahaan asuransi terus memperoleh keuntungan yang diharapkan, mereka harus dikelola dengan baik dan profesional. Bisnis asuransi dan bisnis lainnya bertahan karena laba, yang menarik investor [1] Penghasilan dihitung berdasarkan kemampuan suatu organisasi untuk mengelola bisnis secara keseluruhan, terutama dalam bisnis asuransi kesehatan, karena dapat menunjukkan bagaimana organisasi mengendalikan risiko [2]. Laba yang dihasilkan menjadi salah satu cara untuk menilai keberhasilan kinerja sebuah perusahaan. Jika laba terus meningkat selama beberapa waktu, itu berarti manajemen telah mengelola keuangan dan kerja sama dengan baik, yang meningkatkan nilainya. Seiring dengan itu, karena perusahaan asuransi telah lama terlibat dalam perekonomian negara, masyarakat tidak begitu ragu akan layanan yang di tawarkan. Ketidakpastian yang terkait dengan hal-hal seperti kesehatan, pendidikan, harta benda, dan kematian, turut mendorong meningkatnya kesadaran masyarakat tentang pentingnya asuransi. Asuransi pun menjadi alat penting bagi masyarakat untuk mengantisipasi risiko atau kerugian di masa depan [3] Sebagai cerminan dari peningkatan tersebut, data Otoritas Jasa Keuangan tahun 2016 menunjukkan bahwa di Indonesia terdapat 24 perusahaan asuransi jiwa syariah, 28 asuransi

Idealis: Indonesia Journal Information System
Volume 8, Nomor 2, Juli 2025
ISSN 2684-7280 (*online*)
Halaman 267-276
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

umum syariah, dan 3 reasuransi. Asuransi jiwa syariah terdiri dari 19 perusahaan asuransi jiwa unit usaha syariah dan 5 asuransi jiwa full syariah. Sementara itu, asuransi umum syariah terdiri dari 25 perusahaan asuransi umum unit usaha syariah dan 3 perusahaan asuransi umum *full* syariah [4]. Jumlah asuransi ini belum termasuk dengan beberapa jaminan kesehatan yang dikelola secara khusus oleh perusahaan terhadap karyawan mereka.

Dari berbagai jenis asuransi baik di bidang kesehatan maupun bidang lainnya, setiap tahun dihasilkan data yang sangat bervariasi. Data-data ini, jika dikelola dengan baik, dapat memberikan masukan yang berharga bagi setiap elemen yang terkait, baik itu pemilik asuransi maupun pihak yang menjalin kerja sama. Salah satu metode yang digunakan untuk menghasilkan informasi penting dari setiap data adalah dengan melakukan regresi. Analisis regresi digunakan untuk mengetahui hubungan linier antara dua variabel atau lebih. Satu variabel yang berperan sebagai variabel terikat (dependen) biasanya dilambangkan dengan notasi "Y", sedangkan yang lainnya berperan sebagai variabel bebas (independen) dan dilambangkan dengan notasi "X". Umumnya, analisis regresi digunakan untuk melakukan prediksi atau ramalan, sedangkan hubungan variabel tersebut bersifat fungsional yang diwujudkan dalam suatu model matematis. Selain itu, analisis regresi juga dipakai untuk memahami variabel yang berhubungan dengan variabel terkait agar lebih mengetahui bentuk-bentuk hubungan tersebut [5] Data yang telah dikelola dengan teknik regresi dapat memberikan informasi penting, baik untuk pengembangan kerja sama maupun dalam merumuskan kebijakan bagi para pemangku kepentingan. Dalam hal ini, peramalan berbasis *Multivariate Time Series* menjadi salah satu pendekatan yang relevan. Jenis peramalan ini melibatkan lebih dari satu kriteria yang berubah dari waktu ke waktu sehingga memungkinkan prediksi berdasarkan pola riwayat urutan data [6]. Peramalan *(forecasting)* harga saham, misalnya, merupakan salah satu topik yang sangat populer dalam penelitian karena bertujuan memperoleh prediksi pendapatan dengan memanfaatkan teknik-teknik statistik dan komputasi [7].

Beberapa penelitian telah dilakukan untuk proses peramalan data multivariat, salah satunya menggunakan metode *Random Forest* untuk memprediksi harga Bitcoin dengan memilih fitur dari dataset Bitcoin. Dengan menggunakan pemodelan regresi hutan acak, diperoleh nilai MAPE sebesar 1,50% dan akurasi 98,50%. Hasil ini menunjukkan bahwa algoritma hutan acak ini adalah salah satu pemodelan yang dapat melakukan prediksi yang baik untuk data acak [8]. Dalam penelitian selanjutnya, klasifikasi dilakukan dengan menggunakan citra hasil augmentasi, kemudian dengan metode ekstraksi fitur Histogram Warna, dan selanjutnya dengan algoritma *Random Forest*. Untuk mendapatkan hasil terbaik, penelitian ini juga melakukan beberapa perbandingan, termasuk perbandingan ekstraksi fitur dan algoritma, yang menghasilkan akurasi sebesar 99,65% dari metode yang disarankan [9]. Kemudian, *Random Forest* digunakan untuk regresi dan klasifikasi pada data kualitas air sumur di Provinsi DKI Jakarta. Sebanyak 267 data digunakan, terdiri dari 214 data pelatihan dan 53 data pengujian. Hasilnya menunjukkan bahwa algoritma tersebut memiliki presisi 0,823 dan sensitivitas 0,83, yang berarti mampu memprediksi 82% dari 83% data yang dapat diklasifikasikan sebagai air yang layak atau tidak layak konsumsi. Hasil prediksi menunjukkan bahwa 40 data berhasil memprediksi nilai nol sesuai dengan targetnya, dan 4 data berhasil memprediksi nilai satu sesuai dengan targetnya pada data pengujian. Selain itu, metode klasifikasi dengan algoritma hutan acak juga dievaluasi berdasarkan variabel-variabel kreditur, yaitu V1 hingga V20. Penelitian menggunakan tahapan metode CRISP-DM, dengan pembagian 80% data untuk pelatihan dan 20% untuk pengujian dari total 1000 data. Hasilnya menunjukkan bahwa algoritma hutan acak memiliki tingkat akurasi sebesar 0,83 [10].

Lebih jauh lagi, *Random Forest* dikenal sebagai salah satu metode pembelajaran mesin yang efektif untuk menangani masalah klasifikasi dan regresi. Metode ini bekerja dengan membuat beberapa pohon keputusan yang dibangun pada subset data pelatihan secara acak, dan prediksi akhir didasarkan pada hasil kolektif dari pohon-pohon tersebut [11] Dengan melihat beberapa penelitian terkait diatas maka didapati bahwa salah satu metode yang dapat gunakan untuk mengelola data multivariat yang dihasilkan dari pendapatan asuransi kesehatan yang bisa digunakan adalah dengan menggunakan model random forest untuk melakukan pengeolaan data historis multivariat untuk melakukan prediksi pendapatan asuransi kesehatan yang dihasilkan.

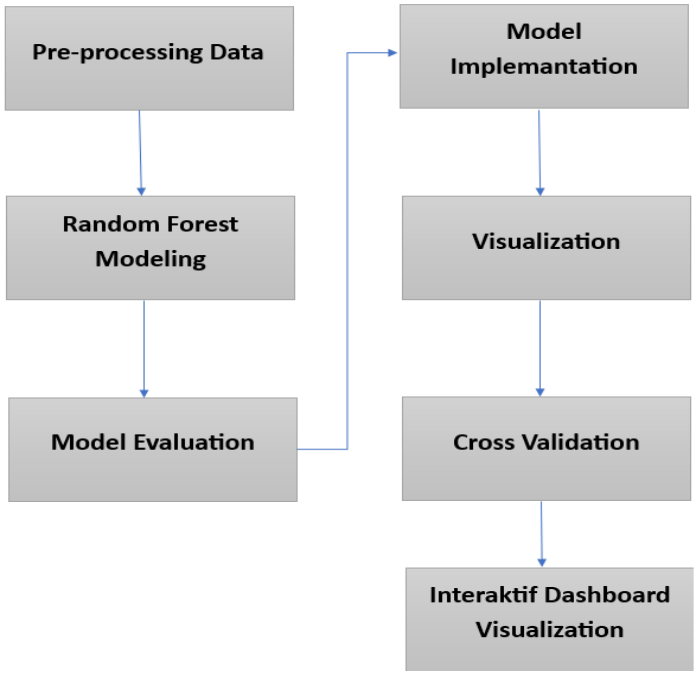
2. METODE PENELITIAN

Dalam penelitian ini akan dilakukan beberapa tahapan penting untuk dapat menghasilkan sebuah model peramalan data multivariat, tahapan penelitian terdapat pada Gambar 1.

2.1 *Preprocessing* data

Pada tahapan *data preprocessing* adalah proses awal yang sangat penting dalam analisis data maupun pengembangan model *machine learning*. Tujuan utamanya adalah menyiapkan data agar bersih, konsisten, dan layak untuk dianalisis atau digunakan dalam pemodelan. Berikut adalah tahapan umum dalam *preprocessing* data: Pemeriksaan dan Penanganan *Missing Value* (Data Hilang) Data sering kali memiliki nilai kosong (missing values) akibat kesalahan pencatatan atau pengumpulan data. Langkah ini mencakup Mengidentifikasi kolom atau baris dengan nilai kosong. Menghapus data yang hilang (jika jumlahnya kecil) atau Mengisi nilai kosong menggunakan metode seperti mean, median, atau interpolasi. Pemeriksaan dan Penanganan *Outlier* (Nilai Pencilan) Outlier adalah data yang

Idealis: Indonesia Journal Information System
Volume 8, Nomor 2, Juli 2025
ISSN 2684-7280 (*online*)
Halaman 267-276
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index



Gambar 1. Alur proses metode penelitian

sangat berbeda dari mayoritas, dan dapat mengganggu akurasi model. Penanganannya bisa melalui: Deteksi menggunakan z-score atau IQR *(interquartile range*). Menghapus *outlier* atau mengubahnya (*winsorization*).[12]

2.2 *Random Forest Modeling*

Pada Tahapan awal dalam pemodelan menggunakan *Random Forest* dimulai dengan menetapkan tujuan pemodelan dan menyiapkan data yang relevan. Penting untuk terlebih dahulu memahami apakah tujuan model adalah untuk melakukan regresi (memprediksi nilai numerik). Setelah itu, data harus diproses dan dibersihkan melalui tahapan *preprocessing* agar siap digunakan, termasuk pemisahan antara fitur (X) dan target (y). Langkah berikutnya adalah membagi data ke dalam dua bagian, yaitu data latih (*training set*) dan data uji (test set), yang bertujuan untuk melatih model dan menguji performanya secara objektif terhadap data yang belum pernah dilihat sebelumnya. Setelah data terbagi, proses membangun model *Random Forest* dilakukan dengan menentukan parameter awal seperti jumlah pohon (n_estimators), kedalaman maksimum pohon (*max_depth*), dan parameter lain yang relevan. Model kemudian dilatih menggunakan data latih dengan metode *ensemble tree*, di mana setiap pohon dilatih pada subset data yang berbeda untuk meningkatkan akurasi keseluruhan.[13] Setelah pelatihan selesai, model digunakan untuk melakukan prediksi terhadap data uji guna mengukur kemampuan generalisasinya terhadap data baru.

2.3 *Model Evaluation*

Tahapan evaluasi model adalah proses penting untuk menilai seberapa baik model *machine learning* bekerja dalam memprediksi data baru secara akurat dan konsisten. Evaluasi dilakukan setelah model dilatih, dan bertujuan untuk mengukur kinerjanya menggunakan data uji yang tidak digunakan selama proses pelatihan. Pada model regresi seperti *Random Forest Regression*, evaluasi biasanya menggunakan beberapa metrik utama, yaitu *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), dan R^2 *Score* (koefisien determinasi). *Mean Absolute Error* (MAE) mengukur rata-rata selisih absolut antara nilai prediksi dan nilai aktual, yang memberikan gambaran seberapa jauh prediksi menyimpang dari nilai sebenarnya. Nilai MAE yang lebih rendah menunjukkan prediksi yang lebih akurat. Sementara itu, *Mean Squared Error* (MSE) menghitung rata-rata kuadrat dari selisih tersebut. Karena selisihnya dikuadratkan, MSE lebih sensitif terhadap outlier atau kesalahan besar. R^2 *Score* atau koefisien determinasi menunjukkan seberapa besar proporsi variansi pada data target yang dapat dijelaskan oleh model. Nilai R^2 mendekati 1 menandakan model yang sangat baik, sedangkan nilai mendekati 0 berarti model tidak mampu menjelaskan variabilitas data.[14]

2.4 Model *Implementation* dan *Visualization*

Tahapan implementasi dan visualisasi model merupakan langkah lanjutan setelah model selesai dievaluasi dan dinyatakan layak untuk digunakan. Pada tahap ini, model digunakan untuk melakukan prediksi terhadap data baru

Idealis: Indonesia Journal Information System
Volume 8, Nomor 2, Juli 2025
ISSN 2684-7280 (*online*)
Halaman 267-276
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

atau data yang belum pernah dilihat selama pelatihan. Implementasi model dimulai dengan menyiapkan data input baru (misalnya data bulan September) dan menjalankan model untuk menghasilkan prediksi berdasarkan pola yang telah dipelajari sebelumnya. Proses ini sangat penting untuk menguji kemampuan model dalam konteks dunia nyata atau skenario simulasi. Setelah model memberikan hasil prediksi, langkah berikutnya adalah visualisasi, yang berfungsi untuk menyampaikan hasil model dengan cara yang lebih mudah dipahami, terutama oleh pengguna non-teknis.

2.5 *Cross Validation*

Pada tahapan ini akan dilakukan validasi silang terhadap model yang ada. *Cross-validation* digunakan karena hal ini merupakan teknik evaluasi dalam *machine learning* yang digunakan untuk mengukur kemampuan generalisasi suatu model terhadap data baru yang belum pernah dilihat sebelumnya. Teknik ini membagi dataset menjadi beberapa bagian atau fold, lalu secara bergantian menggunakan sebagian data untuk melatih model dan sisanya untuk mengujinya. Salah satu metode yang paling umum adalah *k-fold cross-validation*, di mana data dibagi menjadi k bagian dan proses pelatihan-pengujian dilakukan sebanyak k kali, masing-masing dengan kombinasi data latih dan data uji yang berbeda.[15] Dengan cara ini, hasil evaluasi model menjadi lebih akurat, stabil, dan tidak bergantung pada satu pembagian data tertentu. *Cross-validation* juga berfungsi untuk mendeteksi kemungkinan *overfitting*, yaitu kondisi saat model terlalu cocok dengan data latih namun gagal menggeneralisasi pada data baru. Selain itu, teknik ini sangat berguna dalam pemilihan model dan tuning parameter, karena memberikan gambaran performa model yang lebih objektif dan menyeluruh

2.6 Interaktif *Dashboard*

Pada tahapan ini dashboard interaktif menggunakan *Gradio* memiliki berbagai manfaat yang signifikan, terutama dalam konteks aplikasi *data science* dan *machine learning*. Salah satu manfaat utamanya adalah mempermudah interaksi antara pengguna (*user*) dengan model atau data tanpa perlu pengetahuan pemrograman. Dengan antarmuka visual yang intuitif, pengguna dapat mengeksplorasi hasil prediksi, memilih input secara dinamis (seperti memilih *payer* asuransi), dan langsung melihat hasilnya dalam bentuk grafik atau tabel. Hal ini sangat berguna untuk komunikasi hasil analisis secara lebih transparan dan informatif, terutama kepada pihak non-teknis seperti manajer, pemangku kepentingan, atau pengguna akhir. Selain itu, *dashboard* interaktif juga memungkinkan pengujian cepat dan *real-time* terhadap berbagai skenario input tanpa harus mengubah kode. *Gradio* sendiri mudah diintegrasikan di Google Colab atau disebarluaskan melalui web, sehingga sangat cocok untuk keperluan presentasi, demonstrasi proyek, atau prototipe aplikasi machine learning yang ringan dan efisien.

Tabel 1. Struktur data

Payer	Januari	Februari	Maret	April	Mei	Juni	Juli	Agustus
Asuransi Quiti Life Indonesia-Admedika	985.851	68.740.701	197.238.290	7.997.961	2.073.328	2.073.328	2.070.581	2.060.280
Equity Life Indonesia	69.496.848	46.784.516	70.596.700	260.000	1.123.600	1.123.600	1.123.600	1.060.060
AA International Hub SDN BHD	7.430.300	2.565.080	264.940	273.000	383.950	350.000	354.900	384.650
Admedika PT, Asuransi Jma Syariah	1.903.000	3.319.008	7.608.627	445.398	1.328.549	1.273.654	3.263.412	1.229.000
Administrasi Medika, PT	1.903.000	3.319.008	7.608.627	445.398	1.328.549	1.273.654	3.263.412	1.229.000
Asuransi Astra Buana, PT	5.383.174	72.907.960	212.479.715	863.148	100.900	100.900	100.900	100.900
Asuransi AXA Indonesia - Admedika	108.483.749	36.579.715	36.579.715	1.255.395	100.900	100.000	100.900	103.400
Asuransi Cigna, PT								
Asuransi Jiwa Astra, PT	2.291.342	1.964.409	2.211.886	1.390.000	1.538.848	1.538.848	1.674.267	1.592.708
Axa Malaysia (Asia Assistance)	964,197	964,197	1.463.670	1.428.920	88.648.326	166.861.000	85.309.406	85.000

Idealis: Indonesia Journal Information System
Volume 8, Nomor 2, Juli 2025
ISSN 2684-7280 (*online*)
Halaman 267-276
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

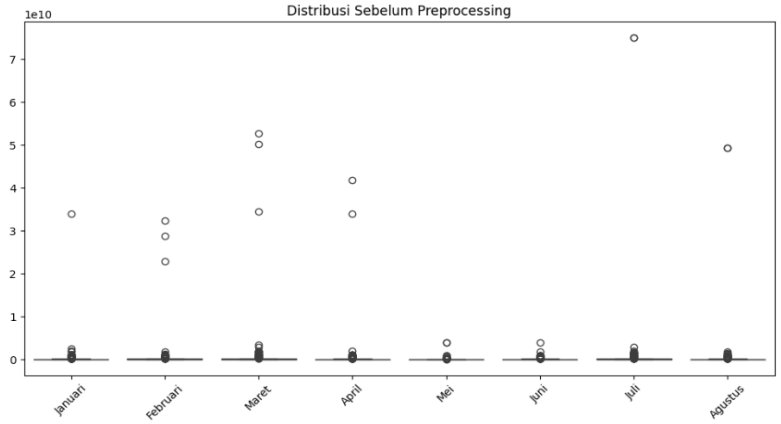
2.7 Dataset

Berdasarkan kumpulan data pada dataset penelitian yang ditampilkan pada Tabel 1, proses pemodelan diawali dengan pemisahan fitur dan target dari dataset. Variabel X merepresentasikan fitur (variabel independen) yang digunakan dalam model, yaitu semua kolom mulai dari kolom kedua hingga kolom sebelum terakhir. Sementara itu, variabel y digunakan sebagai target (variabel dependen), yang diambil dari kolom terakhir dalam dataset. Dengan asumsi bahwa dataset memuat data bulanan dari Januari hingga Agustus, maka fitur yang digunakan dalam model adalah nilai-nilai dari bulan Januari hingga Juli, sedangkan nilai bulan Agustus dijadikan sebagai target yang akan diprediksi. Setelah pemisahan fitur dan target, data kemudian dibagi menjadi data latih dan data uji menggunakan metode *train_test_split*. Selanjutnya, dilakukan proses standardisasi terhadap fitur menggunakan *StandardScaler* agar model dapat dilatih dengan data yang telah dinormalisasi. Model yang digunakan adalah *Random Forest Regressor*, yang kemudian dilatih dengan data latih yang telah distandardisasi. Pendekatan ini memungkinkan model untuk belajar dari pola data historis (Januari–Juli) dalam rangka memprediksi nilai pada bulan Agustus.

3. HASIL DAN PEMBAHASAN

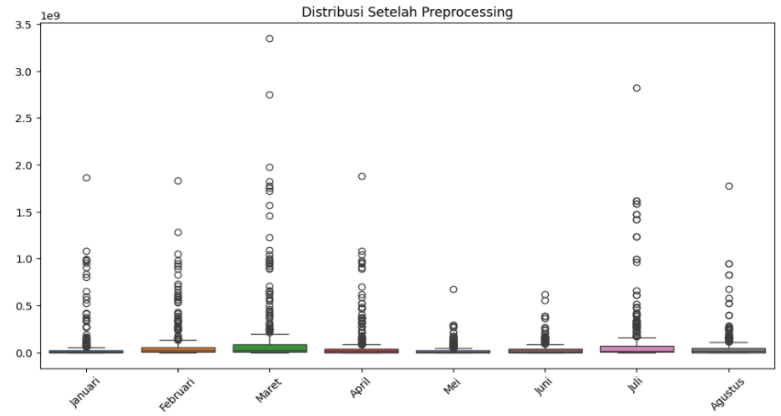
3.1 Preprocessing data

Tabel 1 merupakan *sample* dari data yang akan diproses dengan algoritma *random forest*, data terdiri dari nama dari masing-masing asuransi beserta dengan hasil pendapatan yang diperoleh dari bulan januari hingga agustus dan terdiri dari 423 *record.* Gambar 2 menampilkan dua grafik *boxplot* yang menggambarkan distribusi data pendapatan per bulan dari Januari hingga Agustus, sebelum dan sesudah dilakukan *preprocessing*. Tujuan utama dari visualisasi ini adalah untuk menunjukkan bagaimana proses *preprocessing*, seperti *outlier removal*, memengaruhi skala dan penyebaran data.



Gambar 2. Distribusi data Sebelum *Preprocessing*

Distribusi Sebelum *Preprocessing* data memiliki banyak *outlier* ekstrem, yaitu titik-titik data yang berada jauh di atas mayoritas nilai lainnya. Misalnya, pada bulan Januari dan Agustus terdapat nilai yang mencapai di atas 7×10^{10} atau 70 miliar rupiah, yang sangat jauh dari konsentrasi data lainnya. *Outlier* seperti ini dapat mendistorsi analisis dan membuat model *machine learning* menjadi bias atau tidak akurat, karena model akan mencoba menyesuaikan prediksi terhadap nilai ekstrem tersebut.



Gambar 3. Distribusi data setelah *preprocessing*

Idealis: Indonesia Journal Information System
Volume 8, Nomor 2, Juli 2025
ISSN 2684-7280 (*online*)
Halaman 267-276
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

Distribusi Setelah *Preprocessing* grafik kedua pada Gambar 3 menunjukkan distribusi data setelah dilakukan tahapan *preprocessing*, termasuk pembersihan *outlier*. Tampak bahwa skala nilai sudah lebih terkendali, dengan sebagian besar data berada dalam rentang yang lebih sempit (sekitar 0 hingga 3×10^9 atau 3 miliar rupiah). Meskipun masih terdapat *outlier*, namun jumlah dan skala ekstremnya sudah jauh berkurang. Selain itu, persebaran data menjadi lebih merata antar bulan, yang akan meningkatkan kemampuan model dalam mempelajari pola secara konsisten. Visualisasi ini memperlihatkan bahwa tahapan *preprocessing* sangat penting dalam menangani data keuangan yang rawan *outlier* ekstrem. Dengan mengurangi pengaruh nilai-nilai yang menyimpang, model dapat bekerja lebih akurat dan stabil. Distribusi yang lebih normal dan seimbang setelah *preprocessing* meningkatkan kualitas pelatihan model serta keandalannya dalam membuat prediksi selanjutnya.

3.2 Random Forest *Modeling.*

```
model = RandomForestRegressor(random_state=42)
model.fit(X_train_scaled, y_train)
```

```
RandomForestRegressor
RandomForestRegressor(random_state=42)
```

Gambar 4. Penerapan Model Random Forest

```
print(f"MAE: {mae:.2f}")
print(f"MSE: {mse:.2f}")
print(f"R² Score: {r2:.2f}")
```

```
MAE: 25139425.90
MSE: 2981545048649158.00
R² Score: 0.85
```

Gambar 5. Evaluasi Model

Pada Gambar 4 dan Gambar 5 menunjukkan proses penerapan model dan berdasarkan hasil evaluasi model menggunakan tiga metrik utama MAE, MSE, dan R^2 *Score* dapat disimpulkan bahwa model memiliki performa yang cukup baik dalam memprediksi pendapatan bulan Agustus. Pertama, nilai MAE (*Mean Absolute Error*) sebesar ±25.139.426 menunjukkan bahwa rata-rata kesalahan absolut antara nilai prediksi dan nilai aktual berada di kisaran 25 juta rupiah. Ini berarti, secara umum, model meleset sekitar 25 juta dalam setiap prediksinya. MAE merupakan metrik yang mudah dipahami dan cukup stabil karena tidak terlalu terpengaruh oleh nilai pencilan (*outlier*), sehingga cocok digunakan untuk menilai keakuratan model dalam skala besar.

Kedua, MSE (*Mean Squared Error*) memiliki nilai sebesar 2.9815×10^{15}. Metrik ini menghitung rata-rata dari kuadrat selisih antara nilai aktual dan prediksi, yang membuatnya sangat sensitif terhadap kesalahan besar. Tingginya nilai MSE dalam kasus ini menunjukkan bahwa ada beberapa prediksi dengan *error* yang cukup besar. Hal ini wajar mengingat skala pendapatan yang besar dan kemungkinan adanya *outlier* yang memengaruhi hasil kuadrat error secara signifikan.

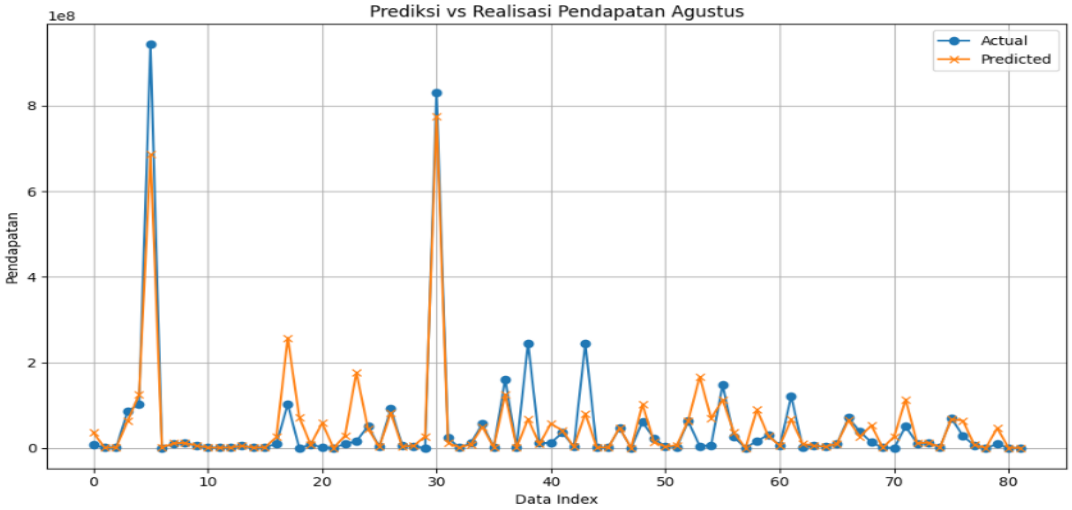
Ketiga, R^2 Score yang mencapai 0.85 menunjukkan bahwa model mampu menjelaskan 85% variasi pendapatan bulan Agustus berdasarkan data dari bulan Januari hingga Juli. Nilai ini mengindikasikan bahwa model memiliki kualitas prediksi yang tinggi dan cukup dapat diandalkan, meskipun masih ada ruang untuk perbaikan. Secara keseluruhan, dengan R^2 di atas 0.8 dan MAE yang masih dalam batas wajar, model dinilai cukup akurat untuk digunakan dalam konteks data keuangan nyata. Meskipun MSE tinggi, hal ini lebih disebabkan oleh skala data dan bukan berarti model secara umum buruk.

3.3 Model *Implementation* dan *Visualization*

Grafik pada Gambar 6 menunjukkan hasil perbandingan antara nilai aktual dan nilai prediksi pendapatan bulan Agustus dari model *Random Forest*. Sumbu horizontal (X) merepresentasikan indeks data (setiap baris data per payer asuransi), sedangkan sumbu vertikal (Y) menunjukkan nilai pendapatan.

Dari visualisasi ini, dapat disimpulkan beberapa hal penting Polanya sudah cukup mirip Secara umum, garis oranye (prediksi) mengikuti tren garis biru (aktual), terutama pada mayoritas data yang memiliki skala pendapatan rendah hingga sedang. Ini menunjukkan bahwa model sudah cukup baik dalam mengenali pola dasar dari data historis.

Idealis: Indonesia Journal Information System
Volume 8, Nomor 2, Juli 2025
ISSN 2684-7280 (*online*)
Halaman 267-276
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index



Gambar 6. Visualisasi data prediksi dan *data existing*

Prediksi untuk nilai ekstrem masih kurang presisi, Terlihat jelas bahwa pada beberapa titik dengan nilai aktual yang sangat tinggi (misalnya pada indeks sekitar 7 dan 30), model menghasilkan prediksi yang tidak sepenuhnya akurat baik *underprediction* (terlalu rendah) maupun *overprediction* (terlalu tinggi). Hal ini wajar, mengingat nilai ekstrem sering kali lebih sulit diprediksi dan bisa menjadi *outlier*.

Sebagian besar prediksi mendekati nilai aktual Untuk mayoritas data di mana nilai pendapatan relatif kecil dan tersebar merata, prediksi cenderung dekat dengan nilai aktual, yang ditunjukkan oleh garis prediksi yang hampir tumpang tindih dengan data aktual.

Secara keseluruhan, grafik ini menggambarkan bahwa model cukup akurat dalam memprediksi pendapatan mayoritas *payer*, namun performanya masih bisa ditingkatkan pada kasus-kasus dengan nilai pendapatan yang sangat tinggi. Ini juga sejalan dengan evaluasi metrik, di mana MAE tergolong moderat, MSE cukup tinggi (karena pengaruh error besar), dan R^2 menunjukkan kecocokan model yang kuat.

2.4 *Cross Validation*

```
print("\nCross-Validation R² Scores:", cv_scores)
print("Mean R²:", np.mean(cv_scores))
print("Std Dev:", np.std(cv_scores))
```

```
Cross-Validation R² Scores: [0.87541952 0.59884943 0.67459241 0.42793665 0.78714742]
Mean R²: 0.67278908575474
Std Dev: 0.15463502857909925
```

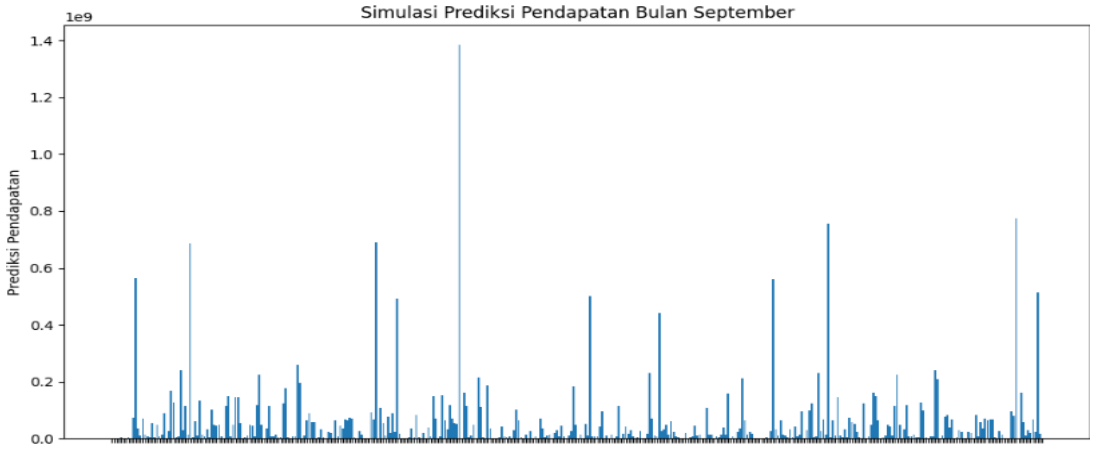
Gambar 7. Hasil *Cross Validation*

Hasil *cross-validation* 5-*fold* pada gambar 7 menunjukkan bahwa model dievaluasi sebanyak lima kali pada lima subset data yang berbeda, menghasilkan nilai R^2 berturut-turut sebesar 0.875, 0.599, 0.675, 0.428, dan 0.787. Nilai rata-rata R^2 dari seluruh lipatan adalah 0.673, yang menunjukkan bahwa secara umum model mampu menjelaskan sekitar 67,3% variabilitas data target di setiap subset, meskipun tidak setinggi saat dievaluasi satu kali dengan data uji (yang mencapai $R^2 = 0.85$).

Standar deviasi dari nilai-nilai R^2 adalah 0.155, yang berarti terdapat sedikit variasi performa model di antara kelima subset. Hal ini bisa disebabkan oleh perbedaan distribusi atau jumlah data yang lebih informatif dalam masing-masing subset. Namun, variasi ini masih dalam batas wajar dan tidak menunjukkan overfitting ekstrem atau inkonsistensi model yang besar.

Secara keseluruhan, hasil *cross-validation* ini menunjukkan bahwa model cukup stabil dan mampu mengeneralisasi dengan baik ke data baru, meskipun ada beberapa subset yang menghasilkan skor R^2 lebih rendah (misalnya 0.428). Hasil ini menguatkan bahwa model layak digunakan dalam prediksi, namun mungkin masih bisa ditingkatkan dengan tuning parameter atau fitur tambahan.

Idealis: Indonesia Journal Information System
Volume 8, Nomor 2, Juli 2025
ISSN 2684-7280 (*online*)
Halaman 267-276
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index



Gambar 8. Simulasi pendapatan bulan berikutnya

Gambar 8 menampilkan hasil simulasi prediksi pendapatan untuk bulan September berdasarkan model *Random Forest* yang telah dibangun menggunakan data dari bulan-bulan sebelumnya (Januari hingga Agustus). Sumbu horizontal menunjukkan index data, yang merepresentasikan masing-masing *payer* (misalnya asuransi atau klien), sementara sumbu vertikal menunjukkan nilai prediksi pendapatan dalam satuan rupiah.

Penjelasan grafik ini berupa bar chart vertikal, di mana setiap batang mewakili hasil prediksi pendapatan satu entitas (payer) untuk bulan September. Terdapat variasi signifikan antar payer, dengan beberapa memiliki prediksi pendapatan yang sangat tinggi, bahkan mencapai lebih dari 1,3 miliar rupiah, sementara yang lain jauh lebih rendah. Distribusi prediksi terlihat tidak seragam, yang mencerminkan bahwa model menangkap variasi historis dari masing-masing *payer* dan menerapkannya untuk memprediksi pendapatan mendatang.

Model berhasil memetakan perbedaan potensi pendapatan antar *payer*, sesuai pola yang dipelajari dari data sebelumnya. Adanya puncak-puncak tinggi menandakan bahwa model mengidentifikasi beberapa entitas dengan potensi pendapatan besar, yang bisa menjadi fokus strategi bisnis ke depan. Grafik ini membantu pengambilan keputusan untuk alokasi sumber daya atau strategi pemasaran berdasarkan potensi kontribusi finansial masing-masing *payer*.

Visualisasi ini menunjukkan bahwa model prediksi dapat digunakan untuk proyeksi keuangan bulan mendatang, dengan mempertimbangkan data historis. Meski tidak disertai nilai aktual (karena bulan September masih bersifat simulasi), grafik ini berguna dalam menyusun strategi bisnis yang lebih terukur.

Untuk proses pelatihan model, data historis pendapatan atau nilai dari bulan Januari hingga Juli digunakan dari model prediksi yang dibangun menggunakan algoritma *Random Forest Regressor*. Fitur-fitur ini dipilih karena data historis sering mengandung pola musiman atau tren jangka pendek yang dapat digunakan untuk memprediksi nilai bulan berikutnya, dalam hal ini bulan Agustus. Dengan menggunakan fitur-fitur tersebut, model belajar memahami hubungan temporal antar bulan dan bagaimana nilai-nilai sebelumnya berkontribusi terhadap nilai masa depan.

Secara mendalam, setiap fitur mewakili nilai pendapatan (atau variabel yang diamati) pada bulan tertentu. Misalnya, fitur dari bulan Januari hingga Maret mungkin menunjukkan tren awal tahun, sedangkan April hingga Juli bisa mencerminkan fluktuasi musiman menjelang kuartal ketiga. *Random Forest* memiliki kemampuan untuk menangkap non-linearitas serta interaksi antar fitur-fitur tersebut tanpa memerlukan transformasi khusus, yang menjadi salah satu keunggulannya dibandingkan metode linear. Sebagai pembanding, metode seperti Linear Regression sering kali digunakan dalam konteks prediksi time series atau regresi multivariat. *Linear Regression* memiliki kelebihan dalam hal interpretabilitas karena menghasilkan koefisien yang jelas untuk setiap fitur, namun memiliki keterbatasan dalam menangani hubungan non-linier atau data dengan distribusi tidak normal. Ketika digunakan pada data yang kompleks seperti prediksi pendapatan asuransi, model linear cenderung menghasilkan akurasi yang lebih rendah dibandingkan model non-linear seperti *Random Forest.*

Support Vector Regression (SVR) dapat digunakan sebagai baseline selain *Linear Regression*. SVR dapat menangani non-linearitas dengan kernel trick, tetapi cenderung lebih peka terhadap parameter dan skala data daripada *Random Forest*. Selain itu, SVR lebih lambat saat menggunakan dataset berukuran besar. Random Forest menunjukkan hasil prediksi yang lebih akurat dalam pengujian awal dibandingkan *baseline* metode linear dan SVR. Ini terutama berlaku untuk data dengan fluktuasi musiman dan *outlier* yang telah diminimalkan sebelumnya. Selain itu, *Random Forest* menawarkan keuntungan berupa pengukuran fitur penting, yang membantu dalam menilai kontribusi relatif masing-masing bulan terhadap hasil prediksi. Misalnya, bulan Juli dan Juni cenderung memiliki nilai kontribusi yang lebih besar karena dekat dengan target prediksi, bulan Agustus. Oleh karena itu, terbukti bahwa fitur yang dipilih dari bulan Januari hingga Juli tepat dan relevan untuk membuat model prediksi berbasis *Random Forest*.

Idealis: Indonesia Journal Information System
Volume 8, Nomor 2, Juli 2025
ISSN 2684-7280 (*online*)
Halaman 267-276
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

Ini terutama berlaku jika dibandingkan dengan baseline model lain, yang lebih sederhana atau kurang responsif terhadap kompleksitas data.

2.5 Interaktif *Dashboard*

Gambar 9. *Gradio* interaktif *dashboard*

Gambar 9 merupakan tampilan dari *dashboard* interaktif berbasis *Gradio* yang digunakan untuk memvisualisasikan prediksi pendapatan bulan berikutnya (misalnya bulan September) untuk setiap *payer* asuransi secara individual. *Dropdown* Pilih *Payer* Asuransi (kiri) Pengguna dapat memilih salah satu entitas asuransi dari daftar *dropdown*, dalam hal ini contohnya adalah ASURANSI CIGNA, PT. Setelah dipilih, sistem akan menampilkan grafik histori pendapatan serta hasil prediksi untuk bulan berikutnya. Tombol *Submit* dan *Clear Submit* digunakan untuk menjalankan visualisasi setelah memilih *payer*. *Clear* digunakan untuk mengosongkan pilihan atau reset tampilan. Grafik Pendapatan (kanan) Grafik garis pada sisi kanan menampilkan tren pendapatan bulanan untuk payer yang dipilih, mulai dari Januari hingga Juli atau Agustus, dan disambung dengan titik prediksi untuk bulan berikutnya (misalnya September). Dalam grafik contoh, terlihat tren peningkatan pendapatan yang signifikan dari bulan Maret ke Juli.

Tampilan Dinamis Antarmuka ini bersifat interaktif dan memungkinkan pengguna untuk eksplorasi tiap *payer* secara spesifik, memberikan insight kinerja dan proyeksi keuangan mendatang. Fungsi dan manfaat dashboard ini berguna untuk analisis individual, sehingga manajer atau analis dapat mengevaluasi performa dan mempersiapkan strategi berdasarkan proyeksi masing-masing *payer*. Visualisasi yang jelas dan interaktif membuat interpretasi data lebih mudah dan mendukung pengambilan keputusan yang berbasis data. *Dashboard Gradio* ini merupakan alat bantu visual yang intuitif untuk memprediksi dan memantau pendapatan masa depan, memungkinkan analisis terfokus terhadap tiap entitas asuransi secara efisien dan akurat.

4. KESIMPULAN

Dalam penelitian ini, model *Random Forest* diterapkan untuk memprediksi pendapatan asuransi bulan berikutnya berdasarkan data historis multivariat dari bulan Januari hingga Juli/Agustus. Hasil evaluasi menunjukkan bahwa model memiliki performa yang cukup baik dalam menangkap pola pendapatan, dengan skor evaluasi *Mean Absolute Error* (MAE) sebesar ±25.139.426 menunjukkan bahwa rata-rata kesalahan prediksi hanya sekitar 25 juta rupiah, angka yang masih tergolong wajar jika dibandingkan dengan skala pendapatan keseluruhan. *Mean Squared Error* (MSE) sebesar 2.9815×10^{15} mencerminkan adanya beberapa *error* besar, meskipun hal ini wajar mengingat skala data dan keberadaan outlier yang sulit dihindari. R^2 *Score* sebesar 0.85 menandakan bahwa 85% variabilitas pendapatan dapat dijelaskan oleh model dari data historis, yang menunjukkan performa prediksi yang sangat baik. Visualisasi hasil prediksi memperlihatkan bahwa sebagian besar nilai prediksi mengikuti pola nilai aktual, meskipun terdapat beberapa deviasi pada kasus *outlier* atau perubahan tren tajam. Selain itu, cross-validation 5-fold menghasilkan rata-rata R^2 sebesar 0.673 dengan standar deviasi 0.155, yang menunjukkan bahwa model cukup stabil dan mampu melakukan generalisasi terhadap data baru, meskipun ada variasi antar subset. Namun, terdapat beberapa kekurangan yang dapat ditingkatkan, antara lain belum adanya pemanfaatan fitur eksternal seperti inflasi, jumlah klaim, atau kondisi ekonomi makro yang mungkin berpengaruh terhadap pendapatan. Model belum sepenuhnya optimal dalam menangani *outlier* ekstrem, yang terlihat dari tingginya MSE. Pendekatan saat ini bersifat statis dan belum memanfaatkan algoritma berbasis waktu seperti LSTM yang mungkin lebih cocok untuk pola data sekuensial. Ke depannya, pengembangan model dapat diarahkan pada kombinasi model *Random Forest* dengan teknik *time series* dan integrasi data eksternal, serta penguatan proses *outlier detection* dan *feature engineering* untuk hasil prediksi yang lebih presisi dan robust.

Idealis: Indonesia Journal Information System
Volume 8, Nomor 2, Juli 2025
ISSN 2684-7280 (*online*)
Halaman 267-276
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

Kontribusi ilmiah dari penelitian ini terletak pada penerapan pendekatan regresi non-linear berbasis *Random Forest* untuk melakukan peramalan pendapatan asuransi menggunakan data multivariat historis bulanan, yang jarang dibahas secara mendalam dalam konteks industri asuransi lokal. Pendekatan ini tidak hanya menyoroti efektivitas *Random Forest* dalam menangkap pola musiman dan hubungan non-linier antar variabel waktu, tetapi juga memberikan landasan untuk eksplorasi metode *machine learning* lanjutan dalam analisis data asuransi. Secara praktis, hasil dari penelitian ini dapat dimanfaatkan oleh perusahaan asuransi sebagai dasar dalam pengambilan keputusan strategis, seperti perencanaan keuangan, penentuan premi, serta evaluasi performa cabang berdasarkan pola pendapatan historis. Model ini juga dapat menjadi dasar pengembangan sistem prediktif yang lebih adaptif terhadap dinamika data keuangan, sehingga mendukung efisiensi dan ketepatan dalam pengelolaan risiko.

DAFTAR PUSTAKA

[1] A. R. S. Ardi, M. Batubara, and M. I. Harahap, "Pengaruh Pendapatan Premi, Hasil Investasi dan Klaim Terhadap Laba Pada PT Asuransi Multi Artha Guna Tbk (AMAG)," *J. Ekon. Syariah dan Bisnis*, vol. 5, no. 2, pp. 179–192, 2022.

[2] F. O. Denovis, S. Arsita, and N. Nurhayati, "Pengaruh Pendapatan Premi, Hasil Underwriting, Hasil Investasi Dan Risk Based Capital Terhadap Laba Perusahaan Asuransi," *JRAK J. Ris. Akunt. dan Komputerisasi Akunt.*, vol. 13, no. 1, pp. 27–35, 2022, doi: 10.33558/jrak.v12i2.3211.

[3] H. Prasetyo, J. E. Tulung, and I. D. Palandeng, "Analisis Pengaruh Pendapatan Premi, Investasi, Dan Hasil Underwriting Terhadap Laba Perusahaan Asuransi Umum Di Otoritas Jasa Keuangan Periode 2017-2021," *J. EMBA J. Ris. Ekon. Manajemen, Bisnis dan Akunt.*, vol. 11, no. 2, pp. 11–22, 2023, doi: 10.35794/emba.v11i02.47200.

[4] N. D. Fatmawati and H. S. Devy, "Pengaruh Pendapatan Premi, Klaim, Invetasi dan Biaya Operasional Terhadap Pertumbuhan Aset Perusahaan Asuransi Jiwa Syariah Di Indonesia," *Veloc. J. Sharia Financ. Bank.*, vol. 1, no. 1, pp. 35–43, 2021, doi: 10.28918/velocity.v1i1.3589.

[5] N. G. Purnomo, Wargijono Utomo, Rachmat Farich, Ratna Fajarwati, *Analisis Data Multivariat*, 1st ed. Jawa Tengah: Omera Pustaka, 2022.

[6] S. Zahara and Sugianto, "Peramalan Data Indeks Harga Konsumen Berbasis Time Series Multivariate Menggunakan Deep Learning," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 24–30, 2021, doi: 10.29207/resti.v5i1.2562.

[7] A. Marjuni, "Peramalan Harga Saham Serentak Menggunakan Model Multivariate Singular Spectrum Analysis," *J. Sist. Inf. Bisnis*, vol. 12, no. 1, pp. 17–25, 2022, doi: 10.21456/vol12iss1pp17-25.

[8] S. Saadah and H. Salsabila, "Prediksi Harga Bitcoin Menggunakan Metode Random Forest," *J. Komput. Terap.*, vol. 7, no. 1, pp. 24–32, 2021, doi: 10.35143/jkt.v7i1.4618.

[9] *et al.*, "Klasifikasi Penyakit Daun Padi menggunakan Random Forest dan Color Histogram," *J. Komputasi*, vol. 10, no. 1, 2022, doi: 10.23960/komputasi.v10i1.2961.

[10] B. Prasojo and E. Haryatmi, "Analisa Prediksi Kelayakan Pemberian Kredit Pinjaman dengan Metode Random Forest," *J. Nas. Teknol. dan Sist. Inf.*, vol. 7, no. 2, pp. 79–89, 2021, doi: 10.25077/teknosi.v7i2.2021.79-89.

[11] H. Tantyoko, D. K. Sari, and A. R. Wijaya, "Prediksi Potensial Gempa Bumi Indonesia Menggunakan Metode Random Forest Dan Feature Selection," *IDEALIS Indones. J. Inf. Syst.*, vol. 6, no. 2, pp. 83–89, 2023, doi: 10.36080/idealis.v6i2.3036.

[12] S. P. Tamba and E. -, "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.

[13] Suci Amaliah, M. Nusrang, and A. Aswi, "Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng," *VARIANSI J. Stat. Its Appl. Teach. Res.*, vol. 4, no. 3, pp. 121–127, 2022, doi: 10.35580/variansiunm31.

[14] E. Fitri, "Analisis Perbandingan Metode Regresi Linier, Random Forest Regression dan Gradient Boosted Trees Regression Method untuk Prediksi Harga Rumah," *J. Appl. Comput. Sci. Technol.*, vol. 4, no. 1, pp. 58–64, 2023, doi: 10.52158/jacost.v4i1.491.

[15] M. Fadli and R. A. Saputra, "Klasifikasi Dan Evaluasi Performa Model Random Forest Untuk Prediksi Stroke," *JT J. Tek.*, vol. 12, no. 02, pp. 72–80, 2023, [Online]. Available: http://jurnal.umt.ac.id/index.php/jt/index