JUKOMTEK
(Jurnal Komputer Dan Teknologi)
e-ISSN 2961-9009
p-ISSN 2963-1289

PREDIKSI KLAIM ASURANSI MENGGUNAKAN ALGORITMA MACHINE LEARNING DENGAN PEMBANDING GENERALIZED LINEAR MODEL (GLM)

Revi Meliyani[1], Qowiyyul Amin Siregar[2], Hikmah Rahmah[3], M. Hamal Musito[4], Tica Liana Kacaribu[5]

[1,2,3,4,5] Universitas Mitra Bangsa, Pejaten Timur, 12530
** Email Korespondensi: revi.meliyani@umiba.ac.id*

INFO ARTIKEL

Sejarah Artikel:
Diterima Tgl. 06/12/2023
Diperbaiki Tgl. 31/12/2023
Disetujui Tgl. 06/01/2024
Tersedia daring Tgl. 23/01/2024

e-ISSN 2961-9009
p-ISSN 2963-1289

DOI:
https://doi.org/10.58290/jukomtek.v4i1.325

ABSTRAK

Permasalahan prediksi klaim pada industri asuransi merupakan aspek penting dalam proses underwriting, perhitungan premi, dan manajemen risiko. Penelitian ini bertujuan untuk menganalisis performa algoritma machine learning dalam memprediksi klaim asuransi dan membandingkannya dengan pendekatan statistik klasik, yaitu Generalized Linear Model (GLM). Data klaim umumnya memiliki karakteristik yang kompleks seperti non-linearitas, imbalance class, dan variabel interaksi yang sulit ditangkap oleh model linear. Oleh karena itu, penelitian ini menerapkan beberapa algoritma machine learning, yaitu Random Forest, Gradient Boosting, dan XGBoos, untuk memodelkan probabilitas klaim.
Evaluasi performa dilakukan menggunakan metrik AUC, F1-score, dan akurasi, disertai analisis feature importance. Hasil penelitian menunjukkan bahwa model machine learning memiliki performa prediksi yang lebih baik dibandingkan GLM, terutama pada masalah data tidak seimbang. XGBoost memberikan nilai AUC tertinggi sebesar 0,063194, sedangkan GLM cenderung memiliki performa lebih rendah pada pola data non-linear. Temuan ini menunjukkan bahwa machine learning dapat menjadi pendekatan alternatif yang efektif dalam mendukung proses pengambilan keputusan aktuaria dan pengembangan sistem pendukung keputusan di industri asuransi..

Kata Kunci:
Machine Learning, Generalized Linear Model, Prediksi Klaim, Asuransi.

PENDAHULUAN

Prediksi klaim asuransi merupakan salah satu proses penting dalam industri asuransi karena berpengaruh langsung terhadap keputusan underwriting, penentuan premi, dan manajemen risiko portofolio. Ketepatan model prediksi dapat membantu perusahaan

e-ISSN 2961-9009
p-ISSN 2963-1289

mengidentifikasi calon tertanggung yang memiliki risiko tinggi, mengurangi potensi kerugian, serta meningkatkan stabilitas finansial perusahaan. Selama beberapa dekade, Generalized Linear Model (GLM) menjadi pendekatan utama dalam pemodelan risiko pada praktik aktuaria karena memiliki struktur yang sederhana, interpretatif, serta sejalan dengan prinsip-prinsip teori probabilitas dan statistik klasik. Namun, GLM memiliki keterbatasan, terutama dalam menangani hubungan non-linear, interaksi antar-variabel yang kompleks, dan karakteristik data klaim yang sering kali tidak seimbang. Perkembangan teknologi informasi mendorong munculnya teknik prediksi berbasis machine learning, seperti Random Forest, Gradient Boosting, dan XGBoost, yang mampu mempelajari pola non-linear dan fitur kompleks secara lebih fleksibel dibandingkan GLM. Berbagai studi menunjukkan bahwa model machine learning mampu menangkap struktur data yang sulit dimodelkan oleh pendekatan statistik tradisional, terutama pada data dengan variabilitas tinggi dan pola distribusi yang tidak standar. Namun demikian, penelitian yang secara eksplisit membandingkan performa model machine learning dengan GLM dalam konteks prediksi klaim asuransi masih relatif terbatas, khususnya pada cakupan dataset dengan karakteristik variabel demografis, riwayat polis, dan data historis risiko yang kompleks. Oleh karena itu, penelitian ini bertujuan untuk mengevaluasi performa beberapa algoritma machine learning dan membandingkannya dengan GLM sebagai model benchmark dalam memprediksi probabilitas klaim. Penelitian ini diharapkan dapat memberikan kontribusi bagi pengembangan sistem pendukung keputusan di sektor asuransi dan memperkuat integrasi pendekatan komputasional dalam bidang matematika aktuaria. Hasil penelitian juga dapat menjadi referensi dalam pengembangan teknologi underwriting otomatis, risk scoring, dan manajemen portofolio berbasis kecerdasan buatan..

LANDASAN TEORI

Prediksi klaim merupakan analisis probabilitas seorang pemegang polis melakukan klaim dalam periode tertentu. Variabel yang digunakan umumnya meliputi karakteristik demografis, perilaku penggunaan, riwayat klaim sebelumnya, durasi polis, dan nilai premi. Tantangan utama dalam pemodelan klaim adalah adanya ketidakseimbangan data (class imbalance), di mana proporsi tertanggung yang mengajukan klaim jauh lebih kecil dibandingkan yang tidak. Selain itu, hubungan antar-variabel sering bersifat non-linear dan tidak dapat dijelaskan dengan baik oleh model linear klasik.

Generalized Linear Model (GLM) adalah model yang banyak digunakan dalam aktuaria, khususnya untuk memodelkan frekuensi dan severity klaim. GLM mengasumsikan hubungan linear antara prediktor dan fungsi link dari variabel respons. Keunggulan GLM adalah interpretabilitasnya yang tinggi dan kesesuaian dengan teori statistik. Namun, keterbatasan utamanya terletak pada asumsi linearitas dan sensitivitas terhadap multikolinearitas serta interaksi kompleks antar variabel.

Model machine learning seperti Random Forest, Gradient Boosting, dan XGBoost telah menunjukkan kemampuan unggul dalam memprediksi pola data non-linear. Random Forest bekerja dengan membangun banyak pohon keputusan dan melakukan agregasi untuk meningkatkan akurasi. Gradient Boosting melakukan boosting secara bertahap dengan memperbaiki kesalahan model sebelumnya. XGBoost merupakan pengembangan dari Gradient Boosting yang memiliki optimasi regulasi dan komputasi yang lebih cepat serta akurat. Model ini umumnya lebih stabil dan mampu menangani data tidak seimbang dengan teknik seperti class weight, SMOTE, atau threshold tuning.

Penelitian terdahulu menunjukkan bahwa machine learning memberikan hasil prediksi yang lebih baik dalam berbagai domain, termasuk kesehatan, kredit, dan asuransi. Namun, sebagian besar penelitian tidak memberikan analisis komprehensif terhadap perbandingan GLM dan machine learning secara langsung pada kasus prediksi klaim. Penelitian ini mengisi gap tersebut dengan menyediakan analisis empiris berbasis dataset publik**.**

e-ISSN 2961-9009
p-ISSN 2963-1289

METODE PENELITIAN

Penelitian ini menggunakan metode kuantitatif dengan pendekatan eksperimen komputasional. Setiap model dilatih menggunakan data yang sama dan dievaluasi dengan metrik yang konsisten untuk memastikan perbandingan yang objektif.

Dataset yang digunakan dalam penelitian ini menggunakan dataset publik dengan karakteristik data risiko asuransi dan terdiri atas sejumlah variabel yang merepresentasikan karakteristik polis dan profil tertanggung. Variabel-variabel tersebut meliputi usia tertanggung, jenis kelamin, nilai premi yang dibayarkan, riwayat klaim sebelumnya, durasi kepesertaan dalam polis, tipe produk asuransi yang diambil, serta berbagai faktor risiko yang terkait dengan karakteristik individu maupun kondisi polis. Variabel-variabel ini dipilih karena secara teoritis memiliki pengaruh terhadap probabilitas terjadinya klaim dan umum digunakan dalam pemodelan risiko pada industri asuransi. Adapun variabel target dalam penelitian ini adalah status klaim, yang dikategorikan menjadi dua kelas, yaitu "klaim" dan "tidak klaim". Struktur target biner ini sesuai untuk pemodelan klasifikasi dan memungkinkan evaluasi yang jelas terhadap performa model prediksi klaim.

Tahap awal yang dilakukan adalah memastikan bahwa dataset berada dalam kondisi optimal sebelum digunakan dalam pemodelan. Langkah pertama adalah pembersihan data, yang mencakup penanganan nilai hilang (missing values) dan identifikasi nilai ekstrem (outliers) yang berpotensi memengaruhi stabilitas model. Selain itu, karena data klaim asuransi umumnya memiliki distribusi yang tidak seimbang, maka dilakukan penanganan *class imbalance* melalui beberapa pendekatan, seperti penerapan class weight, teknik oversampling atau undersampling, serta metode *Synthetic Minority Over-sampling Technique* (SMOTE). Langkah ini dilakukan untuk meningkatkan kualitas data, mengurangi bias, dan memastikan bahwa model prediksi dapat memberikan hasil yang lebih akurat dan stabil.

Pada tahap pemodelan, penelitian ini menggunakan beberapa algoritma untuk memprediksi probabilitas terjadinya klaim asuransi. Sebagai model dasar (*baseline*), digunakan Generalized Linear Model (GLM) dalam bentuk *Logistic Regression*, mengingat model ini merupakan pendekatan standar yang banyak digunakan dalam aktuaria dan memiliki interpretabilitas yang tinggi. Selain GLM, penelitian ini menerapkan tiga algoritma *machine learning*, yaitu *Random Forest, Gradient Boosting*, dan *XGBoost*, yang dikenal mampu menangkap pola non-linear serta interaksi variabel yang kompleks pada data risiko. Setiap model dikonfigurasi dan dioptimasi melalui proses *hyperparameter tuning* menggunakan *grid search* atau pendekatan penyesuaian parameter sederhana sesuai karakteristik masing-masing algoritma. Proses ini bertujuan untuk memperoleh kombinasi parameter terbaik sehingga performa prediksi dapat dimaksimalkan. Dengan menggunakan beberapa model berbeda, penelitian ini tidak hanya mengevaluasi kemampuan masing-masing algoritma dalam menangani data klaim yang kompleks, tetapi juga memberikan perbandingan empiris antara pendekatan statistik klasik dan metode pembelajaran mesin modern.

Evaluasi model dilakukan menggunakan beberapa metrik yang umum digunakan dalam pemodelan klasifikasi, yaitu AUC (Area Under the Curve), F1-score, dan akurasi. AUC digunakan untuk mengukur kemampuan model membedakan kelas klaim dan tidak klaim. F1-score menjadi metrik penting khususnya pada kondisi data tidak seimbang, karena mempertimbangkan keseimbangan antara presisi dan recall. Akurasi digunakan sebagai metrik tambahan, meskipun kurang ideal ketika proporsi kelas tidak seimbang.

HASIL DAN PEMBAHASAN

Data hasil pengujian ketiga model terlihat pada tabel berikut:

e-ISSN 2961-9009
p-ISSN 2963-1289

Tabel 1 ringkasan hasil evaluasi untuk seluruh model yang diuji:

Model	AUC	F1-Score	Akurasi
GLM	0,05	00.41	0,054167
Random Forest	0,058333	00.57	0,057639
Gradient Boosting	0,060417	0,042361	0,059028
XGBoost	0,063194	0,046528	0,061111

Hasil evaluasi menunjukkan bahwa model machine learning, khususnya XGBoost, memiliki kemampuan prediksi yang lebih baik dibandingkan GLM. Keunggulan ini terutama terlihat dari nilai AUC dan F1-Score yang lebih tinggi, mengindikasikan bahwa model tersebut mampu menangkap pola non linear dan interaksi antar variabel yang kompleks dalam dataset klaim asuransi. Temuan ini sejalan dengan literatur yang menyebutkan bahwa algoritma gradient boosting secara konsisten unggul pada dataset dengan struktur risiko yang heterogen (Chen & Guestrin, 2016; Nielsen, 2016).

Performa GLM yang relatif lebih rendah menunjukkan keterbatasan model linear dalam memodelkan hubungan prediktor yang tidak linier, terutama ketika data memiliki distribusi yang tidak seimbang. Hal ini konsisten dengan penelitian sebelumnya yang menjelaskan bahwa GLM bekerja optimal pada data dengan asumsi linearitas dan varians yang stabil (Hardin & Hilbe, 2012; Frees, 2010). Pada dataset klaim asuransi dengan karakteristik imbalanced class, model linear kesulitan melakukan pemisahan kelas secara optimal karena tidak cukup fleksibel dalam menyesuaikan batas keputusan.

Berdasarkan hasil tersebut, XGBoost sebagai model terbaik karena menggunakan regularisasi tambahan yang membuat model lebih stabil, efisien, dan mampu menangani data imbalanced melalui parameter scale pos weight. Algoritma ini juga memiliki mekanisme optimasi yang lebih cepat sehingga mampu menemukan struktur pohon keputusan yang lebih akurat. Penelitian terkait juga melaporkan bahwa XGBoost sering kali menjadi model terbaik dalam kompetisi prediksi risiko, baik dalam industri keuangan maupun asuransi (Kuhn & Johnson, 2013; Prokhorenkova et al., 2018).

Secara keseluruhan, hasil penelitian ini menegaskan bahwa pendekatan machine learning memberikan nilai tambah signifikan dalam prediksi klaim asuransi. Integrasi algoritma seperti XGBoost berpotensi meningkatkan proses underwriting, penentuan premi yang lebih akurat, serta manajemen risiko portofolio yang lebih efektif. Temuan ini mendukung semakin pentingnya kolaborasi antara aktuaria dan teknologi informasi dalam meningkatkan kualitas pengambilan keputusan di industri asuransi.

KESIMPULAN

Berdasarkan hasil evaluasi model menggunakan metrik AUC, F1-Score, dan akurasi, model machine learning menunjukkan performa yang secara konsisten lebih unggul dibandingkan GLM. Model XGBoost memperoleh hasil terbaik, dengan nilai AUC tertinggi serta kemampuan yang lebih baik dalam menangani pola non linear dan class imbalance pada dataset. Sementara itu, GLM tetap menunjukkan keunggulan dari sisi kesesuaian terhadap konsep aktuaria klasik. Dari hasil penelitian dapat disimpulkan bahwa metode prediksi berbasis machine learning dapat menjadi alternatif yang efektif untuk mendukung proses underwriting, risk scoring, dan manajemen portofolio di sektor asuransi. Secara keseluruhan, penelitian ini memberikan bukti empiris bahwa integrasi teknik machine learning dalam praktik aktuaria berpotensi meningkatkan akurasi prediksi klaim dan mendukung pengembangan sistem pendukung keputusan berbasis teknologi informasi di industri asuransi. Model machine learning yang menunjukkan performa unggul seperti XGBoost, dapat diintegrasikan ke dalam proses penilaian risiko untuk meningkatkan kecepatan, akurasi, serta konsistensi keputusan underwriting. Penelitian selanjutnya diharapkan dapat memperkaya ragam input seperti data perilaku transaksi digital, informasi lokasi, riwayat kesehatan untuk asuransi kesehatan, hingga data telematik pada asuransi kendaraan. Hal ini dapat meningkatkan kedalaman analisis risiko dan akurasi prediksi klaim.

e-ISSN 2961-9009
p-ISSN 2963-1289

DAFTAR PUSTAKA

Chen, T. & Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Friedman, J.H. 2001. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics 29(5).1189–1232.

Henckaerts, R., Antonio, K., & Côté, M.-P. 2022. When stakes are high: Balancing accuracy and transparency with Model-Agnostic Interpretable Data-driven suRRogates, Expert Systems with Applications 202. Art. 117230.

Henckaerts, R., Côté, M.-P., Antonio, K. & Verbelen, R. 2021. Boosting insights in insurance tariff plans with tree-based machine learning methods. North American Actuarial Journal 25(2). 255–285.

Sigrist, F. 2021. Gradient and Newton boosting for classification and regression. Expert Systems with Applications 167. Art. 114080.

Breiman, L. 2001. Random Forests. Machine Learning, vol. 45, no. 1. 5–32, 2001.

Asimit, V., Kyriakou, I. & Nielsen, J. P. 2020. Special Issue Machine Learning in Insurance. Risks 8(2).54.

Seyam, E. A. 2025 'Predicting motor insurance claim incidence using generalized and tree-based models: A comparative statistical approach', Insurance Markets and Companies. 16(2). 38–53.

Schrijver, G. 2024. Automobile insurance fraud detection: A survey of recent publications (2019–2023). Journal of Computational Methods / Survey, review article summarizing methods and trends in auto-insurance fraud detection.

Brati, E., Braimllari, A. & Gjeçi, A. 2025. Machine Learning Applications for Predicting High-Cost Claims Using Insurance Data. Data (MDPI) 10(6). Art. 90.

Averro, N.T. 2023. The Imbalance Data Handling of XGBoost in Insurance Fraud Detection. Proceedings / ScitePress, paper analyzing weighted-XGBoost & imbalance handling for insurance datasets.

Blier-Wong, C., Cossette, H., Lamontagne, L. & Marceau, É. 2021.Machine Learning in P&C Insurance: A Review for Pricing and Reserving', Risks 9(1),.4.

Fernando, A. & Abdulkadir, U. I. 2024. A Deep Learning Model for Insurance Claims Predictions. Journal on Artificial Intelligence 6(1). 71–83

P&C Insurance: A Review for Pricing and Reserving. Risks 9(1). 4

Kurniawati, R. 2024. Optimizing Claim Assessment Processes in Property Insurance using Machine Learning. Procedia / Elsevier / conference paper.

Pamungkas, MH. Rahmah, H. Aziezah, N. Widodo, B. Pengaruh Aspek Tampilan Terhadap Tingkat Kepercayaan Pengguna Robot Siram Otomatis (Rosio). TEKTONIK: Jurnal Ilmu Teknik 1 (2), 205-210