FARABI
Jurnal Matematika dan Pendidikan Matematika
p-ISSN 2623-2332
e-ISSN 2798-5474

Prediksi Faktor Pengaruh Klaim Asuransi Menggunakan Regresi Logistik

Adelia Sutany[1], Ni Made Mila Mahadewi[2], Tania Oktavionabila Kurniawan[3], Edwin Setiawan Nugraha[4]
Program Studi Aktuaria, Fakultas Bisnis, Universitas Presiden, Cikarang-Indonesia 17530
Email: [1]adelia.sutany@student.president.ac.id, [2]ni.mahadewi@student.president.ac.id, [3]tania.kurniawan@student.president.ac.id

ABSTRAK

Bagi perusahaan asuransi, memahami faktor-faktor yang mempengaruhi pengajuan klaim sangat penting. Hal ini harus dilakukan oleh perusahaan asuransi agar mereka dapat menjaga keseimbangan dari jumlah premi yang didapatkan dan jumlah klaim yang dilakukan oleh nasabah. Ketidakpastian terkait faktor risiko yang signifikan dapat menimbulkan kesulitan dalam menentukan harga premi yang tepat dan pengelolaan risiko yang akan dilakukan oleh perusahaan. Penelitian ini bertujuan untuk menganalisis faktor apa saja yang dapat memengaruhi kemungkinan nasabah mengajukan klaim asuransi dengan menggunakan metode regresi logistik. Hasil yang didapatkan dari analisis ini menunjukkan bahwa dari 7 variabel bebas ada 4 variabel yang mempengaruhi jumlah klaim, yakni *age, children, bmi,* dan *smoker* mempengaruhi klaim. Kinerja dari model ini menunjukkan bahwa hasil dari *accuracy* sebesar 89%, *precision* sebesar 90%, *sensitivity* sebesar 91% dan *F1-score* sebesar 90%. Penelitian ini berharap bahwa model yang telah dibuat dapat menjadi acuan bagi masyarakat dan industri asuransi untuk mengetahui faktor penyebab klaim secara akurat dan efisien.

Kata kunci: Regresi Logistik, Asuransi, Klaim, GLM

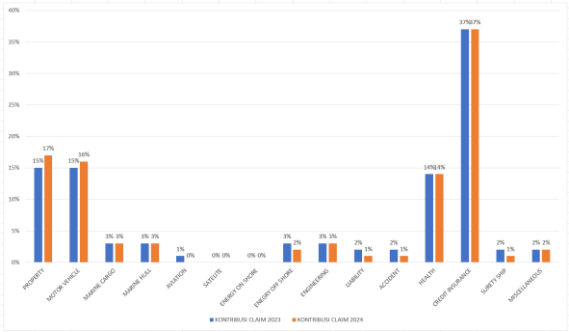
ABSTRACT

For insurance companies, understanding the factors that influence claims submissions is very important. This must be done by insurance companies so that they can mantain a balance between the amount of premiums received and the number of claim made by customers. Uncertainly regarding significant risk factors can cause difficulties in determining the appropriate premium price and risk management to be carried out by the company. This study aims to analyze the factors that can influence the likelihood of customers filing insurance claims using the logistic regression method. The results of this analysis show that out of 7 independent variables, 4 variables influence the number of claims, such as age, children, BMI, and smoker influence claims. The perfomance of this model shows an accuracy of 89%, precision of 90%, sensitivity of 91%, and an F1-Score of 90%. This study hopes that the model developed can serve as a reference for the public and the insurance industry to accurately and efficiently indentify the factors causing claims.

Keywords: *Logistic Regression, Insurance, Claim, GLM*

A. Pendahuluan

Asuransi Jiwa merupakan salah satu dari banyaknya produk asuransi. Asuransi Jiwa merupakan perjanjian hukum antara perusahaan asuransi dengan pihak pengguna asuransi, hal ini disebut kontrak asuransi jiwa (Risalah & Rahmani, 2022). Di era digital saat ini, industri asuransi dituntut untuk terus berinovasi dalam pengambilan keputusan. Salah satu tantangan yang dihadapi adalah bagaimana memanfaatkan data nasabah secara optimal untuk seawal mungkin mengidentifikasi potensi risiko.



Gambar 1. Klaim Dibayar Indutri Asuransi 2024

Dikutip dari Asosiasi Asuransi Umum Indonesia, mencatat bahwa angka klaim dibayar

industri asuransi umum triwulan 4 tahun 2024 tercatat sebesar 49,9 Triliun Rupiah meningkat sebesar 8,5 % dibandingkan dengan periode yang sama di tahun sebelumnya. Peningkatan persentase klaim dapat dilihat pada per Lini Usaha Asuransi Umum dan peningkatan terjadi disebagian besar lini usaha di Triwulan 4 tahun 2024. Beberapa diantaranya disebutkan dalam laporan yang disampaikan oleh Asosiasi Asuransi Umum Indonesia yaitu, asuransi Harta Benda (24,7%) diikuti oleh asuransi Marine Hull (19,5%) (Purwaningsih, n.d.). Kenaikan angka ini menunjukan pentingnya menganalisa faktor – faktor penyebab terjadinya klaim asuransi. Proses evaluasi klaim yang dilakukan secara manual sering kali memakan waktu cukup lama dan rentan terhadap kesalahan manusia, terutama ketika data nasabah bersifat kompleks dan memiliki skala yang besar. Oleh karena itu, diperlukan pendekatan analitis yang cepat, tepat, akurat, dan dapat diandalkan untuk membantu Perusahaan asuransi dalam mengidentifikasi faktor – faktor klaim atau mengidentifikasi potensi klaim sejak dini.

Seiring dengan berkembangnya teknologi, *machine learning* menjadi solusi yang semakin banyak digunakan untuk mendukung pengambilan keputusan berbasis data. Pendekatan ini memungkinkan sistem untuk melakukan identifikasi berbasis data historis dalam memprediksi pola-pola tertentu, termasuk dalam hal pengajuan klaim asuransi. Salah satu yang diunggulkan dari kecerdasan buatan adalah kemampuan untuk mempelajari data dengan sendirinya. Kemampuan itu dikenal dengan *Machine learning*. *Machine Learning* merupakan model statistik yang digunakan untuk memberikan prediksi data menggunakan komputer. (Santoso et al., 2020)

Salah satu algoritma yang dapat digunakan adalah Regresi Logistik dimana metode ini dapat diimplementasikan di dalam *machine learning* untuk melakukan klasifikasi kepada data. Regresi logistik merupakan analisis yang menjelaskan korelasi antara satu atau lebih variabel bebas terhadap satu variabel terikat yang merupakan variabel dikotomis (Situngkir & Sembiring, 2023). Pada Regresi Logistik, respon biner digunakan dan prediktor yang dipakai terdiri dari data kontinu, kategori, atau kombinasi keduanya. Pada analisis ini asumsi distribusi multivariat normal atau kesamaan matriks varians kovarians tidak dibutuhkan, dan dapat diterapkan dalam berbagai jenis skala data. (Suhendra et al., 2020)

Pada studi regresi logistik yang ditulis oleh Mohamad et al dalam (Mohamad et al., 2024) dalam mengetahui faktor yang berpengaruh terhadap ranting pembeli Kopi Kenangan, menghasilkan *accuracy* sebesar 89,7%. Penelitian lain yang dilakukan oleh Pratiwi dalam (Dewi & Pratiwi, 2021) menunjukan bahwa tingkat kepuasan pelanggan jasa layanan ojek online (GRAB) di Kabupaten Lamongan bergantung pada kualitas pelayanan dan promosi, hasil studi ini menunjukan kedua hal tersebut memiliki pengaruh hingga 92%. Promosi dan harga rendah yang ditawarkan oleh Grab akan berdampak pada kepuasan pelanggan hingga 1,475 kali lebih tinggi dibanding dengan harga yang tidak mendapatkan promo.

Mengacu pada berbagai penelitian terdahulu, pada penelitian ini akan menerapkan regresi logistik untuk memprediksi kemungkinan seseorang akan mengajukan klaim asuransi atau tidak. Sistem ini dapat dimanfaatkan oleh perusahaan asuransi sebagai alat bantu dalam mengevaluasi risiko nasabah serta menyusun strategi manajemen risiko dan pemasaran yang lebih efektif. Sistem ini juga berperan dalam meningkatkan kesadaran masyarakat terhadap faktor-faktor yang mempengaruhi pengajuan klaim sebagai bentuk edukasi terhadap masyarakat.

B. Metode Penelitian

1. Regresi Logistik

Regresi Logistik adalah metode statistika yang dapat digunakan untuk memodelkan keterkaitan antara beberapa variabel bebas dengan variabel terikat. Regresi Logistik memiliki 3 tipe, regresi logistik Biner, Multinomial dan Ordinal. Regresi Logistik Biner adalah tipe logistik dimana variabel terikatnya memiliki hanya 2 kemungkinan hasil. Contohnya 0 atau 1, iya atau tidak, dan benar atau salah. Regresi logistik multinomial adalah tipe yang variabel terikatnya memiliki lebih dari 2 hasil tapi tidak tersusun dalam urutan tertentu. Regresi Logistik Ordinal adalah jenis model yang juga mempunyai lebih dari 2 hasil dimana hasilnya memiliki urutan yang ditentukan, seperti 1,2, dan 3 ataupun A sampai F. (Tutz, 2022)

Perubahan pada variabel terikat (Y) adalah biner, dimana Y=1 menunjukkan

bahwa adanya hubungan antara variabel bebas dan variabel terikat (X), dan Y = 0 menunjukkan tidak adanya relasi antara variabel bebas dan variabel terikat. Dengan perubahan hasil dari Y mengikuti distribusi *Bernoulli* dengan parameter $\pi(x_i)$ dengan fungsi probalibitas(McCullagh & Nelder, 1989):

$$f(y_i|\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} = (1-\pi_i)exp\left[y_i ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] \quad (1)$$

Dalam kasus ini, hasil dari y_i adalah 0 atau 1, dan π_i merupakan probabilitas dari y = 1 saat percobaan ke i. Hal ini mengikuti distribusi eksponensial. $Ln\left(\frac{\pi_i}{1-\pi_i}\right)$ merupakan *log odds* pada Y=1 dimana ini juga dikenal sebagai logit dari π_i.

Dari persamaan (1), respon model biner untuk regresi logistic dengan X sebagai variable terikat bisa di modelkan sebagai berikut (Pindyck & Rubinfeld, 1989):

$$\frac{1}{1+e^{-(\alpha+\sum_{j=1}^{n}\beta_i X_{ji}+\sum_{k=1}^{m}\gamma_k D_{ki})}} \quad (2)$$

Asumsikan

$$\left(\alpha+\sum_{j=1}^{n}\beta_i X_{ji}+\sum_{k=1}^{m}\gamma_k D_{ki}\right)$$

Maka, model

logit (2) bisa ditulis sebagai:

$$\pi_i = \frac{1}{1+e^{-z}} \quad (3)$$

Setelah itu, dari persamaan (3) bisa ditulis sebagai berikut:

$$1-\pi_i = 1-\frac{1}{1+e^{-z}} \quad (4)$$

Maka

$$\frac{\pi_i}{1-\pi_i} = \frac{1}{[1+e^{-z}]e^{-z}} = e^{-z} \quad (5)$$

Secara sederhana, persamaan (5) bisa diubah menjadi persamaan logaritma sebegai berikut:

$$Ln\frac{\pi_i}{(1-\pi_i)} = \alpha+\sum_{j=1}^{n}\beta_i X_{ji}+\sum_{k=1}^{m}\gamma_k D_{ki}+e \quad (6)$$

Di studi ini, π_i = Probabilitas dari klaim ($\pi_i = 1$ jika terjadinya klaim, $\pi_i = 0$ jika tidak ada klaim).

$1-\pi_i$ = Probabilitasa dari terjadinya klaim

$\frac{\pi_i}{1-\pi_i}$ = *Odds Ratio* (Risiko)

X_j = vektor dari *free changes* (j = 1, 2, … , n)

D_K = vektor dari *dummy changes* (k = 1, 2, … , m)

α, β_i, dan $\gamma_k = e$ = *random error* parameter dari regresi logistik

Model dari logistik regresi dapat dibentuk seperti persamaan di bawah:

$$\log\left(\frac{p}{1-p}\right) = \beta_0+\beta_1 X_1 + \cdots + \beta_n X_n$$

Di mana:

- p adalah probabilitas seseorang mengajukan klaim.
- $X_1, X_2, \ldots, X_n$ adalah variabel prediktor (seperti usia, BMI, dll).
- β_0 adalah *intercept*, dan $\beta_1, \beta_2, \ldots$ adalah koefisien regresi.

Koefisien dari setiap variabel menggambarkan *log odds* dari probabilitas klaim terhadap nilai dari variabel tersebut. Jika nilai dari koefisien memiliki *p-value* < 0.05, maka variabel koefisien tersebut secara signifikan berpengaruh terhadap variabel terikat.

2. *Package* di R Studio

Library adalah paket tambahan dalam R yang berisi fungsi-fungsi khusus untuk memudahkan analisis data, visualisasi, manipulasi data, *machine learning*, dan lain lain. Beberapa *library* yang digunakan untuk mengolah data pada studi ini antara lain adalah *library(dplyr)* digunakan untuk manipulasi data, seperti menyaring baris dengan *filter()*, memilih kolom dengan *select()*, membuat kolom baru dengan *mutate()*, serta menyambung perintah menggunakan %>%. *Library(ggplot2)* dimanfaatkan untuk membuat visualisasi data yang fleksibel dan rapi. *Library(summarytools)* digunakan untuk menampilkan ringkasan statistik dan eksplorasi data.

3. Model GLM

Model Linear Tergeneralisasi atau *Generalized Linear Model* (GLM) adalah kelas model regresi yang dapat dipakai untuk memodelkan berbagai keterikatan antara variabel respons dan satu atau lebih variabel prediktor. Dalam GLM, biasanya Variabel respons mengikuti anggota distribusi keluarga eksponensial seperti distribusi Normal, Poisson, Binomial, Gamma, and Inverse Gaussian. Dalam GLM, mean dari variable berhubungan dengan kombinasi linear kovariat melalui link function. GLM memungkinkan hubungan non-linear yang lebih fleksibel dengan menggunakan distribusi statistik dasar yang berbeda. GLM memiliki beberapa fitur diantaranya; fleksibilitas, interpretabilitas, ketahanan, dll. Adapun kelemahan dari GLM, salah satunya adalah *Overfitting* yang mana ini rentan terjadi jika modelnya terlalu rumit atau memiliki terlalu banyak variabel prediktor. Berikut adalah rumus dari GLM (Wilandari et al., 2020):

$$y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon_0 \quad (6)$$

Dimana y adalah link function dari fungsi tersebut. Jika data hanya menghasilkan dua kemungkinan, maka *family* yang digunakan adalah *Binomial*. Jika data bersifat numerik dan menghasilkan lebih dari dua kemungkinan, maka kita dapat menggunakan *Poisson, Gamma,* dll sebagai *family* menyesuaikan dengan dataset yang digunakan.

4. Metrik Evaluasi

Confusion Matrix merupakan tabel sederhana yang berfungsi untuk menjelaskan evaluasi dari klasifikasi algoritma. Matriks ini membandingkan prediksi yang dibuat oleh model dengan hasil aktual dan menunjukkan dimana model tersebut benar atau salah. Hal ini membantu untuk memahami dimana model membuat kesalahan, sehingga dapat diperbaiki. Algoritma ini membagi prediksi ke dalam 4 kategori (Sathyanarayanan, 2024):

- *True Positive (TP)*: model dengan tepat memprediksi hasil yang positif, yang berarti bahwa hasil yang aktualnya adalah positif.
- *True Negative (TN)*: model dengan tepat memprediksi hasil yang negatif, yang berarti bahwa hasil yang aktualnya adalah negatif.
- *False Positive (FP)*: model salah memprediksi hasil positif, yaitu hasil aktualnya negatif. Kesalahan ini dikenal sebagai kesalahan Tipe I.
- *False Negative (FN)*: model salah memprediksi hasil negatif, hasil aktualnya adalah positif. Kesalahan ini dikenal sebagai kesalahan Tipe II.

Tabel 1. Confusion matrix

	Prediksi Positif	Prediksi Negatif
Aktual Positif	*True Positive (TP)*	*False Negative (FN)*
Aktual Negatif	*False Positive (FP)*	*True Negative (TN)*

Accuracy menunjukkan banyak prediksi yang benar dari semua prediksi yang dihasilkan model. *Accuracy* memberikan gambaran kinerja keseluruhan, tetapi dapat menyesatkan ketika 1 variabel lebih dominan daripada yang lain. Rumus (3) dapat digunakan untuk mendapakan nilai dari *Accuracy*. (Arisandi & Dewi, 2024)

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

Recall atau *sensitivity* mengukur seberapa baik model dalam memperkirakan hal positif. Hal Ini menunjukkan proporsi positif yang terdeteksi dengan benar dari semua kasus positif yang sebenarnya terjadi. *Recall* yang tinggi sangat penting terutama jika melewatkan kasus positif dapat menyebabkan konsekuensi yang signifikan seperti dalam tes medis. Berikut adalah rumus dari *Recall*:

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

Precision berfokus pada kualitas prediksi positif model. Hal ini memberi tahu kita berapa banyak prediksi "positif" yang benar-benar aktual. Hal ini penting dalam situasi di mana *False Positive* perlu diminimalkan. Berikut adalah rumus dari *Precision*:

$$Precision = \frac{TP}{TP + FP} \qquad (9)$$

F1-Score adalah gabungan dari *recall* dan *Precision* ke dalam satu metrik untuk menyeimbangkan keduanya. Metode ini memberikan Gambaran yang lebih baik tentang kinerja keseluruhan model terutama untuk set data yang tidak seimbang. Hal ini sangat membantu ketika *False Positive* dan *False Negative* penting, meskipun asumsi dari *Precision* dan *Recall* juga penting. Berikut adalah rumus dari *F1-Score*:

$$F1 - Score = \frac{2 \times Precision \times Recal}{Precision + Recall} \qquad (10)$$

5. *Odds Ratio*

Odds ratio adalah suatu ukuran statistik yang mengukur faktor risiko dengan menghitung perbandingan dari jumlah yang terpapar faktor risiko dengan tidak terpapar faktor risiko. *Odds Ratio* memiliki rumus seperti berikut (Aidid, n.d.):

$$OR = \frac{ad}{bc} \qquad (11)$$

Tabel 2. Tabel *odds ratio*

Faktor	Risiko	
	Tidak	Ya
Tidak	a	b
Ya	c	d

Tetapi pada *odds ratio*, di perlukan juga *P value*, yang menunjukkan apakah nilai *odds ratio* yang didapatkan dari sampel dapat dipakai untuk keseluruhan populasi atau tidak. Maka dari itu, diperlukan juga taraf signifikansi dengan batas yang telah ditentukan.

Ada beberapa kriteria hasil dari Odds Ratio diantaranya:

- Jika OR = 1, itu menunjukan bahwa *exposure* tidak mempengaruhi kemungkinan hasil.
- Jika OR > 1, itu menunjukan bahwa *exposure* berhubungan dengan kemungkinan hasil yang lebih tinggi.
- Jika OR < 1, itu menunjukan bahwa *exposure* berhubungan dengan kemungkinan hasil yang lebih rendah.

Jadi dari beberapa faktor risiko tersebut, hasil ini dapat membantu dalam menentukan interpretasi dari hasil Odds Ratio. (*Odds Ratio*, n.d.)

6. Data *Acquisition*

Data *acquisition* adalah tahap di mana data yang diperlukan dalam analisa akan dikumpulkan. Penelitian ini menggunakan data sekunder, yaitu data yang telah dikumpulkan dan dipublikasikan oleh pihak lain. Dataset yang dianalisis adalah dataset klaim asuransi yang diperoleh dari situs Kaggle dan data diolah menggunakan R Studio. Jumlah dataset yang digunakan adalah 1338 data.

Tabel 3. Karakteristik Dataset Klaim Asuransi

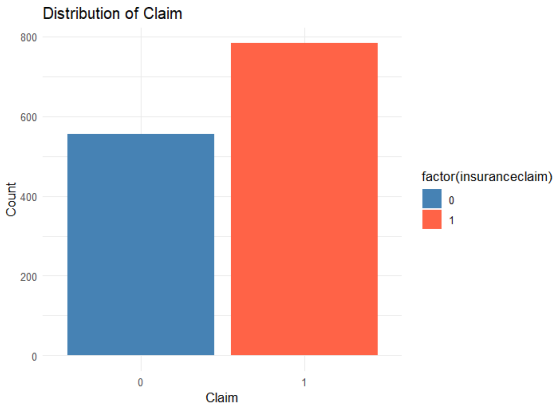
Variabel	Jenis dan Pengukuran Variabel
Age	Numerik
Sex	Kategorik (0,1)
BMI	Numerik
Children	Numerik
Smoker	Kategorik (0,1)
Region	Kategorik (0,1,2,3)
Charges	Numerik
Insurance Claim	Kategorik (0,1)

Adapun variabel yang dianalisis meliputi usia (*age*), jenis kelamin (*sex)* dengan pengkodean yaitu 0 untuk perempuan dan 1 untuk laki-laki, indeks masa tubuh (*bmi*), jumlah anak (*Children*), perokok dengan pengkodean yaitu 0 untuk tidak merokok dan 1 untuk perokok, wilayah (*Region*) dengan 0 untuk timur laut, 1 untuk barat laut, 2 untuk tenggara, dan 3 untuk darat daya, biaya mendis individu (*Charges*) dan faktor terikat yaitu klaim asuransi dengan pengkodean 1 untuk iya klaim dan 0 untuk tidak klaim.

C. Hasil dan Pembahasan

1. Persiapan Data

Pada tahap ini dibagi menjadi beberapa tahap, dimulai dari data *cleaning, transformation* dan *spliting.* Pada tahapan *cleaning* bertujuan untuk mencari jika ada data yang hilang atau *outlier* di dalam dataset. *Outlier* adalah analisa yang dilakukan pada sebuah data dimana data memiliki variabel dengan nilai yang jauh berbeda (ekstrim) dibandingkan dengan data pengamatan lainnya. *Outlier* juga dapat diartikan data yang tidak normal, tidak selaras dengan pola umum model (Smiti, 2020). Kemudian analisa data, jika diperlukan dapat melakukan *transformation* untuk mengubah variabel dalam bentuk angka bisa 0,1 dan lainnya. Selanjutnya periksa apakah variabel terikat memiliki perbedaan yang signifikan. Jika ternyata data memiliki perbedaan yang sangat signifikan, dapat melakukan *balancing* data. Pada dataset klaim asuransi, data yang didapatkan tidak menunjukann perbedaan yang sangat signifikan, dapat dilihat pada Gambar 1.



Gambar 2. Distribusi klaim

2. Generalized Linear Model

Pemodelan dilakukan dengan metode *Generalized Linear Model* (GLM) dengan pendekatan regresi logistik dengan total 7 variabel bebas. Hasil analisa disajikan pada Tabel 4.

Tabel 4. Hasil Pemodelan GLM

term	*estimate*	*std.error*	*statistic*	*p.value*
(Intercept)	-7,387	0,579	-12,755	2,94E-37
age	0,027	0,007	3,690	0,000225
sex	-0,016	0,158	-0,099	0,920836
bmi	0,259	0,018	14,204	8,63E-46
children	-1,424	0,094	-15,106	1,48E-51
smoker	4,048	0,420	9,644	5,19E-22
region	-0,094	0,072	-1,300	0,193642
charges	5,8E-06	1,55E-05	0,374	0,708774

Berdasarkan hasil pemodelan regresi logistik, variabel *age*, *bmi*, *children*, dan *smoker* memiliki nilai *p-value* di bawah 0,05. Hal ini menyatakan bahwa keempat variabel pada data diatas secara statistik memiliki pengaruh signifikan terhadap seseorang melakukan klaim asuransi. Berikut rentang nilai setiap variabel pada Tabel 4 dijelaskan sebagai berikut. Variabel *age* memiliki nilai *p-value* sebesar 0.000225 dan koefisien positif (0.02677), berarti jika semakin bertambah usia seseorang maka kemungkinan untuk mengajukan kalim juga meningkat secara signifikan. Demikian pula bmi (*Body Mass Index*) memiliki koefisien sebesar 0.2586 dengan *p-value* yang sangat kecil (<2e-16), merepresentasikan semakin tinggi nilai bmi, semakin besar kemungkinan seseorang mengajukan klaim. Variabel jumlah anak (*children*) menunjukan koefisien negatif sebesar -1.424 dengan *p-value* <2e-16, yang mengindikasikan bahwa semakin banyak jumlah anak, maka kecenderungan untuk melakukan klaim justru menurun secara signifikan. Variabel merokok tidak merokok juga sangat signifikan dengan *p-value* <2e-16 dan koefisien sebesar 4.048 menunjukan bahwa individu yang merokok memiliki kemungkinan jauh lebih tinggi untuk mengajukan klaim dibandingkan dengan yang tidak merokok.

3. Metrik Evaluasi

Setelah dilakukan pemodelan menggunakan metode *Generalized Linear Model* (GLM), dengan pendekatan regresi logistik, dilakukan evaluasi kinerja model menggunakan *confusion matrix*. Hasil analisa disajikan pada Tabel 5.

Tabel 5. Hasil dari *confusion matrix*

Prediksi	Aktual	
	Tidak klaim	Klaim
Tidak laim	474	71
Klaim	81	712

Tabel 5 menunjukkan hasil *confusion matrix* dari model regresi logistik yang dibangun untuk memprediksi apakah seseorang akan mengajukan klaim asuransi atau tidak. Dalam Tabel angka 474 menunjukan jumlah kasus yang diprediksi tidak klaim (0) dan aktualnya juga tidak klaim (0) disebut *True Negative* (TN). Untuk angka 712 menunjukan jumlah kasus yang diprediksi klaim (1) dan aktualnya juga klaim (1), disebut *True Positive* (TP). Angka 81 adalah jumlah kasus yang diprediksi klaim (1), namun aktualnya tidak klaim (0), disebut sebagai *False Positive* (FP). Terakhir, angka 71 adalah jumlah kasus yang diprediksi tidak klaim (0) padahal aktualnya klaim (1), disebut juga sebagai *False Negative* (FN)

Berikut perhitungan nilai *confusion matrix* ke persamaan (3) (4) (5) (6), untuk mengetahui nilai *accuracy, precision, recall,* dan *F1-score.*

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

$$= \frac{712+474}{474+712+71+81} = 0{,}8863$$

$$Precision = \frac{TP}{TP+FP} = \frac{712}{712+81} = 0{,}8978$$

$$Recall = \frac{TP}{TP+FN} = \frac{712}{712+71} = 0{,}9093$$

$$F1 - Score = \frac{2\times(Precision\times Recall)}{(Precision+Recall)}$$

$$= \frac{712}{712+81} = 0{,}8978$$

Dari hasil yang didapatkan, bisa dilihat hasil dari *accuracy* adalah 88,63%. Ini menunjukkan bahwa model yang telah dibuat, 88,63% dari hasilnya berhasil memprediksi dengan benar. Selain itu, hasil dari *precision* adalah 89,78% yang menunjukkan bahwa ada sekitar 89,78% dari prediksi "klaim terjadi" yang benar-benar merupakan klaim yang sebenarnya. Hasil ini penting untuk memastikan tidak terlalu banyak jumlah *false positive*. Hasil dari *recall* adalah 90,93% yang menunjukkan adanya 90,93% dari total klaim yang dapat diprediksi dengan benar oleh model yang telah dibuat. Hasil dari *F1-score* adalah 89,78% yang berarti bahwa model ini memiliki keseimbangan yang baik antara *recall* dan *precision*.

4. *Odds Ratio*

Setelah melakukan metode *confusion matrix,* dilanjutkan dengan menghitung nilai *odds ratio* untuk mengukur seberapa besar pengaruh variabel tersebut terhadap kemungkinan atau peluang seseorang klaim asuransi. Hasil perhitungan tersebut disajikan pada Tabel 7.

Tabel 6. Hasil Odds Ratio

Variabel	Nilai yang diperoleh
(Intercept)	6,192002e-04
Usia	1,027136e+00
Jenis Kelamin	9,843833e-01
BMI	1,295146e+00
Jumlah anak	2,406667e-01
Perokok	5,728467e+01
Wilayah	9,105341e-01
Biaya mendis individu	1,000006e+00

Berdasarkan Tabel 6, dapat dilihat bahwa nilai dari variabel usia memiliki nilai *Odds ratio* sebesar 1,027 yang berarti setiap penambahan satu tahun usia akan meningkatkan peluang terjadinya kejadian sebesar 2,7%. Sementara variabel jenis kelamin memiliki nilai *odds ratio* 0,984 yang berarti bahwa jenis kelamin laki-laki memiliki peluang sedikit lebih rendah dibandingkan perempuan, namun perbedaannya tidak signifikan.

Variabel BMI memiliki nilai *odds ratio* sebesar 1,295, yang artinya bahwa setiap kenaikan satu unit *BMI* dapat meningkatkan peluang sebesar 29,5%. Sementara itu, variable jumlah anak memiliki nilai *odds ratio* sebesar 0,241 yang artinya semakin banyak jumlah anak semakin rendah peluang terjadinya kejadian.

Variabel yang paling mencolok adalah variabel perokok, dimana nilai *odds ratio* variabel tersebut sebesar 57,284 yang artinya orang yang merokok memiliki peluang 57 kali lebih besar mengalami kejadian dibandingkan yang tidak merokok. Variabel wilayah dan biaya mendis individu memiliki nilai *odds ratio* mendekati 1 yang artinya pengaruh terhadap kejadian relatif kecil atau netral.

D. Kesimpulan dan Saran

1. Kesimpulan:

Penilitian ini bertujuan untuk menganalisa apakah nasabah akan mengajukan klaim atau tidak dan juga untuk mengetahui seberapa efektif model regresi logistik pada dataset asuransi klaim. Analisis dataset tersebut dilakukan berdasarkan beberapa variabel seperti usia, jenis kelamin, indeks massa tubuh (BMI), jumlah anak, status merokok, dan wilayah tempat tinggal. Berdasarkan hasil penelitian, diperoleh nilai *accuracy* sebesar 89%, *precision* sebesar 90%, *recall*/*sensitivity* sebesar 91%, dan *F1-Score* sebesar 90%. Nilai-nilai yang diperoleh tersebut menunjukan bahwa model regresi logistik memiliki performa yang baik, kerena mampu menghasilkan prediksi yang cukup akurat dan mendekati hasil aktual. Dengan demikian, model ini tidak hanya berguna di dunia akademik, tetapi juga memiliki potensi penerapan langsung dalam dunia asuransi.

2. Saran

Dari hasil penelitian ini telah menunjukan bahwa regresi dapat digunakan untuk memprediksi klaim asuransi dengan tingkat *accuracy* yang baik. Namun, untuk meningkatkan kualitas hasil penelitian di masa mendatang, penulis memberikan saran untuk menguji model lain selain regresi logistik seperti *random forest*, *decision tree*, atau *gradient boosting*, agar mendapat hasil atau model terbaik untuk prediksi klaim asuransi. Jika data tidak seimbang, dapat diatasi dengan menggunakan metode *SMOTE* agar hasilnya seimbang.

E. Daftar Pustaka

Aidid, M. K. (n.d.). ANALISIS REGRESI LOGISTIK BINER UNTUK MENENTUKAN MODEL PENGGUNA KB DI KELURAHAN LANGNGA KABUPATEN PINRANG.

Arisandi, R., & Dewi, A. L. (2024). ANALISIS FAKTOR RISIKO GAGAL JANTUNG DENGAN REGRESI LOGISTIK BERBASIS IoMT. Jurnal Gaussian, 12(4), 549–559. https://doi.org/10.14710/j.gauss.12.4.549-559

Dewi, A. F., & Pratiwi, R. (2021). Analisis Regresi Logistik Biner pada Pengaruh Harga, Kualitas Pelayanan dan Promosi terhadap Kepuasan Pelanggan dalam Menggunakan Jasa Layanan Grab di Kabupaten Lamongan. Inferensi, 4(2), 77. https://doi.org/10.12962/j27213862.v4i2.8637

McCullagh, P., & Nelder, J. A. (1989). Generalized linear models (2nd ed). Chapman and Hall.

Mohamad, A., Setiyanto, A., Raharjo, B. W., & Heikal, J. (2024). Analisis Faktor yang Mempengaruhi Kepuasan Pembeli di Kopi Kenangan Menggunakan Metode Regresi Logistik Biner. Jurnal Syntax Admiration, 5(8), 2964–2972. https://doi.org/10.46799/jsa.v5i8.1376

Pindyck, R. S., & Rubinfeld, D. L. (1989). Econometric models and economic forecasts (2. ed., [Nachdr.]). McGraw-Hill.

Purwaningsih, S. (n.d.). BIDANG STATISTIK, RISET & ANALISA.

Risalah, M., & Rahmani, N. A. B. (2022). Actuarial Aspects in Life Insurance. Jurnal Ekonomi, Manajemen, Akuntansi Dan Keuangan, 3(3). https://doi.org/10.53697/emak.v3i3.644

Santoso, R. R., Megasari, R., & Hambali, Y. A. (2020). Implementasi Metode Machine Learning Menggunakan Algoritma Evolving Artificial Neural Network Pada Kasus Prediksi Diagnosis Diabetes.

Sathyanarayanan, S. (2024). Confusion Matrix-Based Performance Evaluation Metrics. African Journal of Biomedical Research, 4023–4031. https://doi.org/10.53555/AJBR.v27i4S.4345

Situngkir, R. H., & Sembiring, P. (2023). Analisis Regresi Logistik untuk Menentukan Faktor-Faktor yang Mempengaruhi Kesejahteraan Masyarakat Kabupaten/Kota di Pulau Nias. FARABI: Jurnal Matematika dan Pendidikan Matematika, 6(1), 25–31. https://doi.org/10.47662/farabi.v6i1.432

Smiti, A. (2020). A critical overview of outlier detection methods. Computer Science

Review, 38, 100306. https://doi.org/10.1016/j.cosrev.2020.100306

Suhendra, M. A., Ispriyanti, D., & Sudarno, S. (2020). KETEPATAN KLASIFIKASI PEMBERIAN KARTU KELUARGA SEJAHTERA DI KOTA SEMARANG MENGGUNAKAN METODE REGRESI LOGISTIK BINER DAN METODE CHAID. Jurnal Gaussian, 9(1), 64–74. https://doi.org/10.14710/j.gauss.v9i1.27524

Tutz, G. (2022). Ordinal regression: A review and a taxonomy of models. WIREs Computational Statistics, 14(2), e1545. https://doi.org/10.1002/wics.1545

Wilandari, Y., Kartiko, S. H., & Effendie, A. R. (2020). ESTIMASI CADANGAN KLAIM MENGGUNAKAN GENERALIZED LINEAR MODEL (GLM) DAN COPULA. Jurnal Gaussian, 9(4), 411–420. https://doi.org/10.14710/j.gauss.v9i4.29260