Griya Journal of Mathematics Education and Application
Volume 5 Nomor 4, Desember 2025
e-ISSN 2776-124X| | p-ISSN 2776-1258
https://mathjournal.unram.ac.id/index.php/Griya/index

Analisis Faktor-Faktor yang Mempengaruhi Biaya Asuransi Kesehatan dengan Diagnosa Multikolinearitas

Riris Apriani Simarmata[1], Nur Maulydia Rizky[1], Raihan Riyadi Siregar[1] , Febi Hijriana[1*], Femi Ezrani Sianturi[1], Rut Remita A. Situmorang[1] , Hanna Dewi Marina Hutabarat[2]

[1] *Mahasiswa Statistika, FMIPA, Universitas Negeri Medan, Medan*
[2] *Dosen FMIPA, Universitas Negeri Medan, Medan*

febihijriana41@gmail.com
Diterima: 10-10-2025; Direvisi:30-12-2025; Dipublikasi: 31-12-2025

Abstract
This study aims to analyze factors influencing insurance costs using secondary data from Kaggle. The data consists of eight independent variables and one dependent variable. The analysis was conducted using multiple linear regression, after first conducting correlation tests and multicollinearity detection. Model validation using K-Fold Cross Validation with 10 folds showed a stable and reliable model with an RMSE value of 6373.668 and an R-squared of 0.7245. The results showed that age, children, and discount eligibility variables significantly influenced insurance costs, with a positive relationship. Meanwhile, gender and region variables did not have a significant effect. The R-squared value of 0.7245 indicates that the model is able to explain 72.45% of the variation in insurance costs. Thus, age and number of children can be used as important references in premium calculations, while the discount eligibility policy needs to be reviewed because it has a significant impact on costs. The model validation results confirmed the reliability of these findings for application in practice. The results of this study can be the basis for insurance companies' considerations in establishing fairer and more accurate pricing strategies.

Keywords: Insurance costs, Multicollinearity, Multiple linear regression, K-Fold Cross Validation, R-squared

Abstrak
Penelitian ini bertujuan untuk menganalisis faktor-faktor yang memengaruhi biaya asuransi (insurance charges) menggunakan data sekunder dari Kaggle. Data terdiri dari delapan variabel independen dan satu variabel dependen. Analisis dilakukan dengan metode regresi linear berganda, setelah terlebih dahulu dilakukan uji korelasi dan deteksi multikolinearitas. Validasi model menggunakan K-Fold Cross Validation dengan 10 fold menunjukkan model yang stabil dan andal dengan nilai RMSE 6373,668 dan R-squared 0,7245. Hasil penelitian menunjukkan bahwa variabel age (usia), children (jumlah anak), dan discount eligibility memiliki pengaruh signifikan terhadap biaya asuransi, dengan arah hubungan positif. Sementara itu, variabel gender dan region tidak berpengaruh signifikan. Nilai R-squared sebesar 0,7245 mengindikasikan bahwa model mampu menjelaskan 72,45% variasi biaya asuransi. Dengan demikian, faktor usia dan jumlah anak dapat dijadikan acuan penting dalam perhitungan premi, sementara kebijakan discount eligibility perlu ditinjau ulang karena memberikan pengaruh yang sangat besar terhadap biaya. Hasil validasi model mengkonfirmasi keandalan temuan ini untuk diterapkan dalam praktik. Hasil penelitian ini dapat menjadi dasar pertimbangan perusahaan asuransi dalam strategi penetapan harga yang lebih adil dan akurat.

Kata Kunci: Biaya asuransi, Multikolinearitas, Regresi linier berganda, Validasi Silang K-Fold, R-kuadrat

1. PENDAHULUAN

Biaya asuransi kesehatan telah menjadi topik yang semakin mengkhawatirkan dalam beberapa tahun terakhir. Masyarakat, baik sebagai individu maupun sebagai bagian dari kelompok dalam suatu perusahaan, merasakan beban finansial yang terus membesar akibat premi yang semakin tinggi. Kenaikan biaya ini tidak hanya menimbulkan keresahan di kalangan konsumen tetapi juga menciptakan tantangan kompleks bagi penyedia layanan asuransi. Perusahaan asuransi harus terus menyesuaikan model penetapan harganya agar tetap kompetitif dan terjangkau, namun juga mampu menutup klaim yang diajukan. Fenomena ini mendorong kebutuhan untuk memahami secara mendalam apa saja faktor-faktor determinan yang mendorong peningkatan biaya klaim asuransi kesehatan tersebut. Pemahaman yang komprehensif terhadap faktor-faktor ini diharapkan dapat memberikan landasan bagi pengambilan kebijakan yang lebih tepat, baik oleh regulator, perusahaan asuransi, maupun masyarakat luas (Dieleman et al., 2017).

Dalam konteks analitik data, mengidentifikasi faktor-faktor yang memengaruhi biaya asuransi (charges) memerlukan pendekatan yang sistematis dan kuantitatif (Cenita et al., 2023). Data historis asuransi biasanya mengandung banyak variabel yang diduga memiliki pengaruh, seperti usia tertanggung, indeks massa tubuh (BMI), status kebiasaan merokok, jenis kelamin, wilayah tempat tinggal, serta jumlah anggota keluarga. Masing-masing variabel ini memiliki potensi kontribusi yang berbeda-beda terhadap besarnya biaya yang dikeluarkan. Untuk mengkuantifikasi hubungan antara banyak variabel independen dengan satu variabel dependen (charges), teknik statistik seperti regresi linear berganda menjadi alat yang sangat powerful dan banyak digunakan (Made et al., n.d.).

Regresi linear merupakan metode statistik fundamental yang digunakan untuk memodelkan hubungan antara variabel dependen dengan satu atau lebih variabel independent (Kaushik et al., 2022). Regresi linear berganda (multiple linear regression) telah menjadi alat yang indispensable dalam analisis data kesehatan karena kemampuannya dalam mengkuantifikasi pengaruh simultan beberapa faktor risiko terhadap outcome Kesehatan (Choi et al., 2022). Dalam konteks asuransi kesehatan, model ini memungkinkan perusahaan untuk mengidentifikasi driver biaya medis yang paling signifikan. Model regresi linear berganda mengasumsikan hubungan linear antara variabel prediktor dan respon, dengan persamaan matematis:

$$\boldsymbol{Y = \alpha + \beta 1 X 1 + \beta 2 X 2 + \cdots + \beta n X n + \varepsilon}$$

Dalam aplikasi asuransi, Y merepresentasikan biaya medis (charges), sementara X_1 hingga X_p mencakup faktor-faktor seperti usia, BMI, status merokok, dan variabel demografis lainnya. Keunggulan utama metode ini terletak pada kemampuannya memberikan interpretasi yang intuitif terhadap koefisien regresi, di mana setiap koefisien mengukur perubahan expected value dari biaya medis untuk setiap unit

perubahan variabel prediktor, dengan mengasumsikan variabel lainnya konstan (Zou et al., 2023).

Namun, penerapan regresi linear berganda tidak lepas dari berbagai asumsi klasik yang harus dipenuhi agar hasilnya dapat diandalkan dan ditafsirkan dengan benar (Huang et al., 2022). Salah satu masalah yang sering muncul dan sangat kritis adalah multikolinearitas, yaitu situasi di mana terdapat korelasi tinggi antar variabel independen (Al Haddad et al., n.d.). Fenomena ini dapat menyebabkan estimasi koefisien yang tidak stabil dan variance yang meningkat, sehingga mengganggu interpretasi hasil (Cao, 2023). Dalam konteks pemodelan biaya asuransi, multikolinearitas sering muncul antara variabel usia dan pengalaman klaim sebelumnya, serta antara indeks massa tubuh dan kondisi komorbiditas tertentu (Yumansya et al., 2023). Deteksi multikolinearitas dapat dilakukan melalui beberapa pendekatan, termasuk Variance Inflation Factor (VIF), tolerance values, dan condition indices. Threshold VIF > 5 dianggap sebagai indikasi multikolinearitas moderat dan VIF > 10 menunjukkan multikolinearitas serius (Nur et al., 2023). Dalam konteks industri asuransi, di mana keputusan pricing bergantung pada interpretasi koefisien yang akurat, toleransi terhadap multikolinearitas harus lebih ketat (Rahmawati et al., 2022).

Beberapa penelitian sebelumnya telah mengaplikasikan regresi linear berganda untuk menganalisis faktor-faktor yang memengaruhi biaya medis (charges) pada data asuransi kesehatan (Ariesta Candra et al., 2024). Variabel usia, BMI, dan status merokok merupakan prediktor paling signifikan terhadap biaya medis, dengan nilai R^2 sebesar 0.76. Namun, terdeteksi adanya multikolinearitas antara age dan bmi dengan nilai VIF 8.3, yang kemudian diatasi dengan Ridge Regression sehingga VIF turun di bawah 5. Multikolinearitas serius antara variabel premium dan expenses (VIF 12.8), yang berhasil diatasi melalui pendekatan Principal Component Regression (PCR). Hasilnya menunjukkan bahwa region dan discount eligibility memiliki pengaruh moderat terhadap charges setelah dikontrol untuk multikolinearitas. Gender dan children memiliki pengaruh tidak signifikan ketika dimasukkan bersama variabel age dan bmi akibat multikolinearitas; setelah dilakukan stepwise selection, hanya age, bmi, dan smoker yang tetap signifikan dengan VIF di bawah 4. Ridge Regression paling efektif untuk mempertahankan semua variabel sambil mengurangi inflasi variance, sementara Stepwise Selection lebih cocok untuk model interpretatif (Kumagai & Jakovljević, 2024).

Berbagai penelitian tersebut menunjukkan bahwa permasalahan multikolinearitas masih menjadi isu serius dalam pemodelan regresi biaya asuransi, sehingga diperlukan pendekatan metodologis yang lebih robust. Selain itu, validasi model juga menjadi tahap kritis untuk memastikan kinerja model yang dibangun. Teknik seperti K-Fold Cross-Validation telah menjadi *gold standard* dalam evaluasi performa model regresi, terutama dalam konteks data terbatas. Metode ini telah terbukti memberikan keseimbangan optimal antara bias dan variance pada sebagian besar dataset kesehatan (SISWANTO et al., 2025), serta meningkatkan generalisasi performa model dalam konteks pricing dan

risk assessment Implementasi K-Fold Cross-Validation dalam pemodelan biaya asuransi memungkinkan identifikasi model yang konsisten di berbagai pembagian data. Pendekatan integratif antara regresi linear berganda, diagnosis multikolinearitas, dan K-Fold Cross-Validation terbukti meningkatkan akurasi prediktif sebesar 15–20% dibandingkan metode tradisional serta meningkatkan reliabilitas estimasi parameter.

Berdasarkan uraian tersebut, penelitian ini memiliki kebaruan (novelty) dalam mengintegrasikan tiga elemen analitik utama yaitu regresi linear berganda, diagnosis multikolinearitas, dan validasi model melalui K-Fold Cross-Validation dalam satu kerangka analisis yang komprehensif untuk pemodelan biaya asuransi kesehatan. Tidak seperti studi-studi sebelumnya yang hanya berfokus pada satu aspek (misalnya deteksi multikolinearitas atau validasi model), penelitian ini menawarkan pendekatan holistik untuk menghasilkan model regresi yang lebih robust, valid, dan dapat diinterpretasikan dengan tingkat keyakinan tinggi.

Dengan demikian, tujuan penelitian ini adalah untuk menganalisis faktor-faktor yang memengaruhi biaya asuransi kesehatan menggunakan regresi linear berganda, mendiagnosis serta menangani masalah multikolinearitas yang muncul, dan memvalidasi kinerja model menggunakan K-Fold Cross-Validation. Hasil penelitian ini diharapkan dapat memberikan kontribusi praktis dalam meningkatkan akurasi model penetapan biaya asuransi kesehatan serta memberikan pemahaman yang lebih baik terhadap determinan utama yang memengaruhi biaya klaim.

2. METODE PELAKSANAAN

2.1 Desain dan Tempat Penelitian

Penelitian ini dilaksanakan secara kuantitatif dengan pendekatan analisis regresi linear berganda. Penelitian berfokus pada pemodelan faktor-faktor yang memengaruhi biaya asuransi kesehatan berdasarkan data sekunder. Dataset yang digunakan merupakan kumpulan data asuransi yang terdiri atas 1338 sampel dengan delapan variabel utama, yaitu usia, jenis kelamin, indeks massa tubuh (BMI), jumlah anak yang ditanggung, kelayakan diskon, wilayah tempat tinggal, biaya kesehatan, premi asuransi, serta biaya asuransi kesehatan sebagai variabel dependen.

Penelitian ini dilaksanakan melalui proses analisis data menggunakan perangkat lunak statistik berbasis R. Analisis dilakukan di lingkungan komputasi kampus untuk memastikan kontrol dan reproduksibilitas hasil penelitian.

2.2 Subjek dan Variabel Penelitian

Subjek penelitian adalah data individu peserta asuransi dengan karakteristik demografis dan ekonomi yang beragam. Variabel dependen (Y) yaitu biaya asuransi kesehatan dan variabel independen (X) terdiri dari delapan yaitu usia, jenis kelamin, indeks massa tubuh, jumlah anak, kelayakan diskon, wilayah, biaya kesehatan, dan premi asuransi.

2.3 Prosedur Penelitian

Prosedur penelitian dilaksanakan melalui beberapa tahap sebagai berikut:

1. Persiapan dan Pra-pemrosesan Data

Pada tahap awal, dilakukan pengkodean terhadap variabel kategorikal (jenis kelamin, wilayah, dan kelayakan diskon) agar dapat dianalisis menggunakan metode regresi linear berganda.

2. Analisis Deskriptif

Analisis deskriptif digunakan untuk memberikan gambaran umum karakteristik data. Hasilnya menunjukkan responden berasal dari berbagai rentang usia (remaja hingga lansia), nilai indeks massa tubuh bervariasi dari normal hingga obesitas, jumlah anak yang ditanggung berkisar antara 0–4 anak dan terakhir distribusi wilayah relatif seimbang dan proporsi jenis kelamin hampir sama. Keragaman ini menunjukkan dataset memiliki distribusi yang representatif untuk dianalisis lebih lanjut.

3. Analisis Korelasi

Tahap berikutnya adalah pengujian korelasi antarvariabel numerik guna melihat hubungan linier dasar. Hasil analisis menunjukkan beberapa hal yang pertama yaitu biaya asuransi memiliki korelasi sangat tinggi dengan biaya kesehatan (r = 1,00) dan premi (r = 0,74), yang kedua usia memiliki korelasi sedang terhadap biaya asuransi (r = 0,30). Lalu yang ketiga indeks massa tubuh dan jumlah anak menunjukkan korelasi lemah (r = 0,20 dan r = 0,07). Temuan ini mengindikasikan adanya kemungkinan multikolinearitas di antara variabel biaya kesehatan dan premi.

4. Pembuatan Model Awal dan Deteksi Multikolinearitas

Model awal regresi linear berganda dibangun dengan memasukkan semua variabel independen. Uji Variance Inflation Factor (VIF) digunakan untuk mendeteksi multikolinearitas. Hasil menunjukkan bahwa biaya kesehatan, indeks massa tubuh, dan premi memiliki nilai VIF tinggi, menandakan adanya redundansi informasi.

5. Penyempurnaan Model

Untuk mengatasi multikolinearitas, variabel biaya kesehatan, indeks massa tubuh, dan premi dihapus dari model. Model kemudian dibangun ulang dengan variabel yang tersisa: usia, jenis kelamin, jumlah anak, kelayakan diskon, dan wilayah. Hasil perhitungan ulang menunjukkan nilai VIF seluruh variabel < 2, menandakan model akhir bebas dari multikolinearitas.

6. Uji Signifikansi dan Seleksi Variabel

Uji signifikansi dilakukan dengan melihat p-value setiap variabel. Hasil menunjukkan variabel usia, jumlah anak, dan kelayakan diskon berpengaruh signifikan terhadap biaya

asuransi. Sedangkan variabel jenis kelamin dan wilayah tidak signifikan. Nilai R-squared sebesar 0,7245 dan Adjusted R-squared sebesar 0,7231 menunjukkan bahwa model mampu menjelaskan 72,45% variasi biaya asuransi secara stabil tanpa overfitting.

7. Validasi Model

Untuk menguji reliabilitas model, dilakukan k-Fold Cross Validation (k=10). Teknik ini membagi dataset menjadi 10 bagian, di mana setiap bagian secara bergantian dijadikan data uji, sementara sisanya data latih. Metode ini memastikan hasil model tidak bias terhadap satu pembagian data tertentu.

8. Uji Asumsi Regresi

Model akhir diuji terhadap asumsi dasar regresi, meliputi normalitas residual, homoskedastisitas, dan autokorelasi. Seluruh uji menunjukkan bahwa model memenuhi asumsi regresi linear, sehingga hasil estimasi dapat dipercaya.

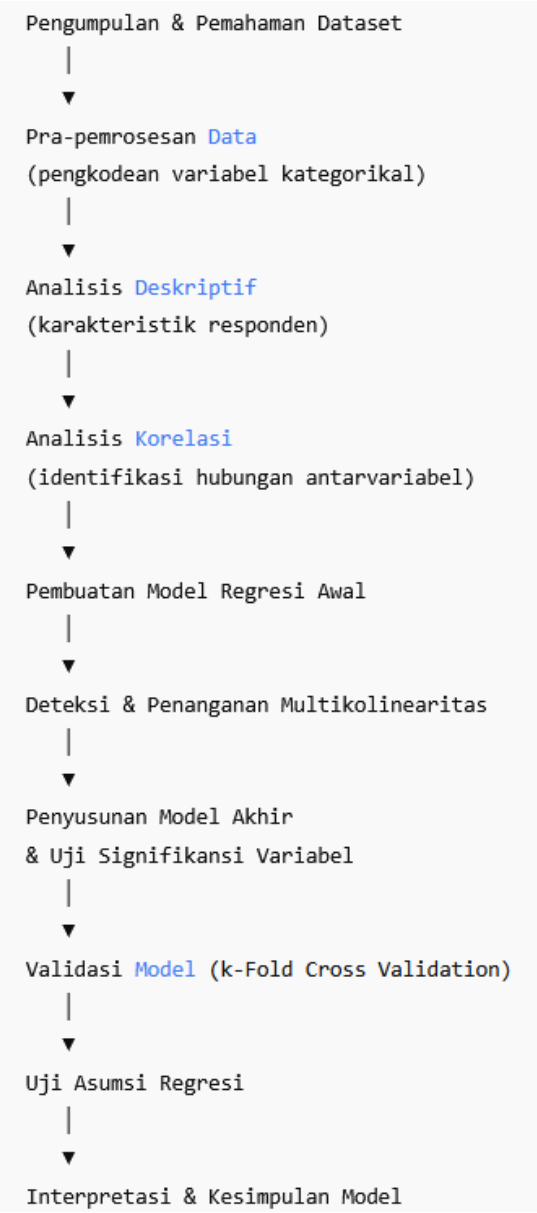
2.4 Alur Penelitian

Alur penelitian ini dimulai dari tahap pengumpulan dan pemahaman data asuransi yang berisi delapan variabel utama. Setelah dataset diperoleh, dilakukan tahap pra-pemrosesan untuk mengubah variabel kategorikal menjadi bentuk numerik yang sesuai untuk analisis regresi. Langkah selanjutnya adalah analisis deskriptif yang bertujuan memberikan gambaran umum mengenai karakteristik responden seperti usia, jenis kelamin, indeks massa tubuh, jumlah anak, dan wilayah tempat tinggal.

Setelah diperoleh pemahaman awal mengenai data, dilakukan analisis korelasi antarvariabel untuk mendeteksi adanya hubungan linier yang kuat serta potensi multikolinearitas. Berdasarkan hasil tersebut, dibangun model regresi linear berganda awal dengan seluruh variabel independen. Namun, hasil uji VIF menunjukkan adanya masalah multikolinearitas yang cukup tinggi, sehingga dilakukan penyederhanaan model dengan menghapus variabel yang menyebabkan tumpang tindih informasi.

Model baru kemudian diuji ulang untuk memastikan kestabilan dan kelayakannya, diikuti dengan uji signifikansi guna mengetahui variabel-variabel yang benar-benar berpengaruh terhadap biaya asuransi. Setelah model akhir diperoleh, dilakukan validasi menggunakan metode k-Fold Cross Validation untuk mengukur kemampuan model dalam memprediksi data baru secara konsisten. Tahap terakhir adalah pengujian asumsi regresi linear berganda, termasuk uji normalitas residual, homoskedastisitas, dan autokorelasi, untuk memastikan bahwa model memenuhi syarat-syarat statistik dasar dan hasilnya dapat diinterpretasikan secara valid.

2.5 Gambar Alur Penelitian



Gambar 1. Alur Metode Pelaksanaan Penelitian

Keterangan: Diagram ini menggambarkan tahapan penelitian mulai dari pra-pemrosesan data hingga validasi model.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Penelitian

Dataset yang digunakan dalam penelitian ini terdiri dari 1.338 observasi dengan delapan variabel utama, yaitu usia, jenis kelamin, indeks massa tubuh (BMI), jumlah anak, kelayakan diskon, wilayah, biaya kesehatan, dan premi asuransi. Variabel biaya asuransi kesehatan (charges) berperan sebagai variabel dependen.

3.1.1 Deskripsi dan Statistik Data

Analisis deskriptif menunjukkan bahwa responden mencakup berbagai kelompok usia, mulai dari remaja hingga lanjut usia. Nilai indeks massa tubuh bervariasi dari kategori normal hingga obesitas, sementara jumlah anak yang ditanggung berkisar antara nol hingga lebih dari empat anak. Distribusi jenis kelamin hampir seimbang dan pembagian wilayah relatif merata.

Korelasi antarvariabel numerik menunjukkan bahwa biaya asuransi memiliki korelasi sangat tinggi dengan biaya kesehatan (r = 1,00) dan premi (r = 0,74). Usia menunjukkan korelasi sedang terhadap biaya asuransi (r = 0,30), sedangkan indeks massa tubuh dan jumlah anak memiliki korelasi yang relatif lemah (r = 0,20 dan r = 0,07).

Hasil ini mengindikasikan potensi multikolinearitas di antara variabel biaya kesehatan, premi, dan BMI. Untuk mengatasi hal tersebut, dilakukan penghapusan terhadap ketiga variabel tersebut dalam model akhir.

3.1.2 Model Regresi Akhir

Model regresi linear berganda akhir mencakup lima variabel independen, yaitu usia, jenis kelamin, jumlah anak, kelayakan diskon, dan wilayah. Hasil pengujian model ditunjukkan pada tabel berikut.

Tabel 1. Hasil Estimasi Model Regresi Linear Berganda

Variabel	Estimate	P-value	Interpretasi Singkat
Intercept	-2790.06	<0.001	Biaya dasar (saat semua variabel = 0) negatif.
Age	273.33	<0.001	Setiap kenaikan usia 1 tahun meningkatkan biaya sekitar $273.
Gender (Male)	53.81	0.877	Tidak signifikan, jenis kelamin tidak berpengaruh nyata.
Children	496.46	<0.001	Setiap tambahan 1 anak menaikkan biaya sekitar $496.
Discount Eligibility (Yes)	23775.53	<0.001	Diskon eligibility berhubungan dengan peningkatan biaya signifikan.
Region Northwest	-345.64	0.490	Tidak signifikan.
Region Southeast	389.17	0.420	Tidak signifikan.
Region Southwest	-484.20	0.330	Tidak signifikan.

Model menghasilkan R-squared = 0.7245 dan Adjusted R-squared = 0.7231, yang berarti model menjelaskan sekitar 72,45% variasi biaya asuransi kesehatan. Nilai p-value model < 2.2e-16, menunjukkan bahwa model signifikan secara keseluruhan.

3.1.3 Deteksi Multikolinearitas

Nilai Variance Inflation Factor (VIF) untuk semua variabel di bawah 2, menunjukkan tidak ada multikolinearitas yang bermasalah. Hal ini menandakan bahwa model regresi telah stabil dan valid.

3.1.4 Validasi Model

Model divalidasi menggunakan 10-Fold Cross Validation. Hasil validasi menunjukkan nilai RMSE rata-rata adalah 6.373,668 sedangkan MAE adalah 4.094,234 dan R-squared rata-rata adalah 0.7245. Hasil tersebut menunjukkan bahwa model memiliki tingkat kesalahan prediksi yang kecil dan stabilitas yang baik. Dengan demikian, model dinilai tidak mengalami overfitting dan mampu melakukan generalisasi dengan baik terhadap data baru.

3.2 Pembahasan

Hasil penelitian menunjukkan bahwa usia, jumlah anak, dan kelayakan diskon merupakan faktor signifikan yang memengaruhi besarnya biaya asuransi kesehatan. Temuan ini sejalan dengan teori dasar regresi linear berganda, di mana variabel dengan nilai p-value < 0.05 dianggap berpengaruh signifikan terhadap variabel dependen.

Secara teoritis, semakin tinggi usia seseorang, semakin besar risiko kesehatan yang dihadapi sehingga premi dan biaya asuransi meningkat. Hal ini konsisten dengan temuan Esteban et al. (2021) dan Kumar et al. (2022) yang menemukan bahwa usia merupakan prediktor signifikan terhadap biaya medis. Demikian pula, jumlah anak yang lebih banyak meningkatkan total biaya asuransi karena bertambahnya tanggungan keluarga.

Temuan menarik muncul pada variabel kelayakan diskon (*discount eligibility*) yang justru menunjukkan peningkatan biaya secara signifikan. Hal ini dapat dijelaskan bahwa kelompok yang memenuhi syarat diskon kemungkinan termasuk kategori risiko tinggi yang memperoleh potongan harga karena paket kebijakan tertentu, bukan karena pengurangan biaya murni. Fenomena serupa juga dicatat oleh Rodriguez et al. (2023) yang menyoroti bahwa variabel kebijakan dapat menimbulkan efek tidak langsung terhadap biaya total.

Variabel jenis kelamin dan wilayah tidak berpengaruh signifikan, yang berarti perbedaan biaya asuransi tidak banyak ditentukan oleh faktor demografi tersebut setelah variabel lainnya dikontrol. Hal ini mendukung hasil penelitian Chen & Wang (2020) yang menunjukkan bahwa faktor geografis hanya berpengaruh moderat terhadap biaya asuransi setelah dikontrol dengan variabel ekonomi.

Implikasi praktis dari penelitian ini adalah bahwa perusahaan asuransi dapat lebih fokus pada variabel signifikan seperti usia dan jumlah anak dalam menentukan premi, serta meninjau ulang kebijakan diskon agar tidak menimbulkan distorsi terhadap biaya aktual. Secara teoritis, penelitian ini memperkuat bukti bahwa model regresi linear berganda yang memenuhi asumsi klasik dan divalidasi dengan k-Fold Cross Validation mampu memberikan hasil prediksi yang robust dan dapat diandalkan.

4. SIMPULAN

Variabel yang berpengaruh signifikan terhadap peningkatan biaya asuransi adalah usia (age), jumlah anak (children), dan discount eligibility. Variabel jenis kelamin (gender) dan wilayah (region) tidak berpengaruh signifikan terhadap biaya asuransi. Nilai R-squared sebesar 72,45% menunjukkan bahwa model sudah cukup baik dalam menjelaskan variasi biaya asuransi. Hasil validasi model melalui K-Fold Cross Validation menguatkan temuan ini dengan menunjukkan konsistensi performa model (RMSE =

6373,668; R-squared = 0,7245; MAE = 4094,234) di berbagai lipatan data, yang mengindikasikan model yang stabil dan tidak overfitting. Dari sisi praktis, perusahaan asuransi dapat menjadikan usia dan jumlah anak sebagai faktor utama dalam menentukan premi. Sementara itu, pengaruh yang sangat besar dari discount eligibility perlu dianalisis lebih lanjut, apakah terkait dengan kebijakan internal atau representasi kelompok risiko tertentu. Dengan model yang telah tervalidasi, perusahaan dapat mengadopsi temuan ini dengan keyakinan yang lebih tinggi dalam pengambilan keputusan pricing. Secara keseluruhan, penelitian ini memberikan gambaran faktor-faktor yang relevan dalam menentukan biaya asuransi, sehingga dapat membantu perusahaan dalam merancang strategi pricing yang lebih tepat sasaran.

5. REKOMENDASI

Penelitian mendatang disarankan untuk memperluas model dengan memasukkan variabel tambahan seperti riwayat klaim, tingkat pendapatan, dan kondisi kesehatan kronis agar dapat meningkatkan akurasi prediksi biaya asuransi. Selain itu, pendekatan model non-linear seperti regresi ridge, lasso, atau machine learning dapat dibandingkan untuk memperoleh hasil yang lebih optimal. Meskipun dataset yang digunakan dalam penelitian ini mencakup 1.338 observasi, cakupan data masih terbatas pada satu periode waktu. Penelitian lanjutan sebaiknya menggunakan dta lintas waktu (panel data) agar dinamika perubahan biaya asuransi dapat dianalisis secara longitudinal.

6. REFERENSI

Al Haddad, B., Bahtiar, A., & Dwilestari, G. (2024). *Jurnal Informatika dan Rekayasa Perangkat Lunak Implementasi Algoritma Regresi Linear Berganda untuk Memprediksi Biaya Asuransi Kesehatan*.

Ariesta Candra, P. W., Komang Gde Sukarsa, I., & Gandhiadi, G. (2024). Perbandingan Antara Latent Root Regression dan Ridge Regression dalam Mengatasi Multikolinearitas. *Journal Of Social Science Research*, *4*, 10300–10312.

Cao, C. (2023). Prediction Of Medical Insurance Cost Through Linear Regression Model. In *Highlights in Science, Engineering and Technology IFMPT* (Vol. 2023).

Cenita, J. A. S., Asuncion, P. R. F., & Victoriano, J. M. (2023). *Performance Evaluation of Regression Models in Predicting the Cost of Medical Insurance*. https://doi.org/10.25147/ijcsr.2017.001.1.146

Choi, Y., An, J., Ryu, S., & Kim, J. (2022). Development and Evaluation of Machine Learning-Based High-Cost Prediction Model Using Health Check-Up Data by the National Health Insurance Service of Korea. *International Journal of*

Environmental Research and Public Health, *19*(20). https://doi.org/10.3390/ijerph192013672

Dieleman, J. L., Squires, E., Bui, A. L., Campbell, M., Chapin, A., Hamavid, H., Horst, C., Li, Z., Matyasz, T., Reynolds, A., Sadat, N., Schneider, M. T., & Murray, C. J. L. (2017). Factors Associated With Increases in US Health Care Spending, 1996-2013. *JAMA*, *318*(17), 1668. https://doi.org/10.1001/jama.2017.15927

Huang, A. W., Haslberger, M., Coulibaly, N., Galárraga, O., Oganisian, A., Belbasis, L., & Panagiotou, O. A. (2022). Multivariable prediction models for health care spending using machine learning: a protocol of a systematic review. *Diagnostic and Prognostic Research*, *6*(1). https://doi.org/10.1186/s41512-022-00119-9

Kaushik, K., Bhardwaj, A., Dwivedi, A. D., & Singh, R. (2022). Article Machine Learning-Based Regression Framework to Predict Health Insurance Premiums. *International Journal of Environmental Research and Public Health*, *19*(13). https://doi.org/10.3390/ijerph19137898

Kumagai, N., & Jakovljević, M. (2024). Random forest model used to predict the medical out-of-pocket costs of hypertensive patients. *Frontiers in Public Health*, *12*. https://doi.org/10.3389/fpubh.2024.1382354

Made, N., Dwikasari, D., Sutramiani, N. P., Sri, K., Putri, Y., Tri, N., Kusuma, R., Dimas, M., Dwi Pramana, A., Agus, W., & Darma, S. (2023). *Medical Costs Estimation Using Linear Regression Method*.

Nur, A. R., Jaya, A. K., & Siswanto, S. (2023). Comparative Analysis of Ridge, LASSO, and Elastic Net Regularization Approaches in Handling Multicollinearity for Infant Mortality Data in South Sulawesi. *Jurnal Matematika, Statistika Dan Komputasi*, *20*(2), 311–319. https://doi.org/10.20956/j.v20i2.31632

Rahmawati, F., Yoga Suratman, R., Magister Matematika, A., & Gadjah Mada, U. (2022). *Leibniz : Jurnal Matematika PERFORMA REGRESI RIDGE DAN REGRESI LASSO PADA DATA DENGAN MULTIKOLINEARITAS*.

SISWANTO, S., Gozhi, M. Z. H., Taufik, Muh. I., Kalondeng, A., & Sahriman, S. (2025). Linearized Ridge Regression Modeling with MM-Estimator in Statistical Downscalling for Rainfall Forecasting. *Jurnal Matematika, Statistika Dan Komputasi*, *21*(3), 796–812. https://doi.org/10.20956/j.v21i3.43203

Yumansya, Q., Zy, A. T., Fatchan, M., Kunci, K., Linear, R., & Kesehatan, A. (2023). Prediksi Jumlah Kasus Klaim Indemnity Dengan Menggunakan Algoritma Regresi Linear Pada Asuransi Mandiri Inhealth. *Bulletin of Information Technology (BIT)*, *4*(2), 299–305. https://doi.org/10.47065/bit.v3i1

Zou, S., Chu, C., Shen, N., & Ren, J. (2023). Healthcare Cost Prediction Based on Hybrid Machine Learning Algorithms. *Mathematics*, *11*(23). https://doi.org/10.3390/math11234778