Jurnal Informatika dan Rekayasa Perangkat Lunak ISSN 2656-2855
Vol. 6, No. 1, Maret 2024, Hal. 142-149 e-ISSN 2685-5518

Implementasi Algoritma Regresi Linear Berganda untuk Memprediksi Biaya Asuransi Kesehatan

Bagas Al Haddad[1*], **Agus Bahtiar**[2], **Gifthera Dwilestari**[3]
[1]Program Studi Teknik Informatika, STMIK IKMI Cirebon
[2,3]Program Studi Sistem Informasi, STMIK IKMI Cirebon
[*]Email: bagasalhaddad105@gmail.com

Abstrak

Perkembangan teknologi seperti *telemedicine*, dan analisis data besar (big data) memberikan dampak yang signifikan terhadap industri asuransi kesehatan. Sangat sulit untuk membuat keputusan yang bijak jika pelanggan tidak memahami biaya asuransi. Usia, jenis kelamin, riwayat medis, wilayah, perokok, dan indeks massa tubuh (BMI) adalah sejumlah variabel yang digunakan untuk menentukan variabel yang berkontribusi pada biaya asuransi kesehatan. Regresi linear berganda digunakan untuk mengidentifikasi variabel-variabel yang berkontribusi untuk memprediksi biaya asuransi kesehatan secara relatif. Analisis regresi linear berganda, juga dikenal sebagai analisis regresi ganda, adalah suatu model regresi yang melibatkan lebih dari satu variabel independen. Ini ditentukan dengan menggunakan perangkat lunak statistik untuk menentukan variabel independen yang memiliki pengaruh yang signifikan terhadap variabel dependen. Nilai dari penggunaan regresi linear berganda terutama terkait dengan kebutuhan akan prediksi biaya asuransi. Dalam *tools* RapidMiner, operator linear regression digunakan untuk melakukan penghitungan regresi linear. Dari total 1338 dataset, data dibagi menjadi dua bagian. 90% digunakan sebagai data pelatihan (dengan jumlah 1204 data) dan 10% digunakan sebagai data uji (dengan jumlah 134 data). Hasil analisis menunjukkan bahwa faktor independen seperti status perokok, usia, dan indeks massa tubuh memiliki korelasi yang signifikan dengan biaya premi asuransi. Nilai 5891.019 dihasilkan dari evaluasi model yang menggunakan Root Mean Squared Error (RMSE). Korelasi kuat antara status perokok dan biaya premi, bersama dengan korelasi positif dengan usia dan indeks massa tubuh (BMI), menunjukkan bahwa biaya premi meningkat seiring bertambahnya usia dan kategori berat badan.

***Kata kunci**:* asuransi, data mining, regresi linear berganda

Abstract

Technological developments such as telemedicine and big data analysis have had a significant impact on the health insurance industry. It is very difficult to make wise decisions if customers do not understand the cost of insurance. Age, gender, medical history, region, smoking, and body mass index (BMI) are a number of variables used to determine the variables that contribute to health insurance costs. Multiple linear regression was used to identify variables that contribute to predicting relative health insurance costs. Multiple linear regression analysis, also known as multiple regression analysis, is a regression model that involves more than one independent variable. This is determined by using statistical software to determine which independent variables have a significant influence on the dependent variable. The value of using multiple linear regression is primarily related to the need for prediction of insurance costs. In the RapidMiner tool, the linear regression operator is used to perform linear regression calculations. From a total of 1338 datasets, the data is divided into two parts. 90% is used as training data (with a total of 1204 data) and 10% is used as test data (with a total of 134 data). The results of the analysis show that independent factors such as smoking status, age, and body mass index have a significant correlation with insurance premium costs. The value 5891.019 was generated from model evaluation using Root Mean Squared Error (RMSE). The strong correlation between smoking status and premium costs, along with positive correlations with age and body mass index (BMI), suggests that premium costs increase with increasing age and weight category.

***Keywords**: insurance, data mining, multiple linear regression.*

PENDAHULUAN

"Big data" dalam industri asuransi kesehatan mengacu pada jumlah data yang sangat besar. Perusahaan asuransi kesehatan sekarang dapat mengelola, menyimpan, dan menganalisis sejumlah besar data yang kompleks berkat kemajuan teknologi informasi dan komunikasi(Pratama *et al.*, 2023). Usia, jenis kelamin, riwayat kesehatan, lokasi, dan faktor lain adalah komponen dari data asuransi kesehatan. Kemajuan dalam penyimpanan data telah memungkinkan perusahaan asuransi kesehatan untuk mengelola, menyimpan, dan menganalisis lebih banyak data kompleks, karena asuransi kesehatan dianggap sebagai hak yang diberikan kepada setiap orang(Sholeh, Suraya and Andayati, 2022). Kemungkinan terkena penyakit tetap tidak dapat diprediksi, meskipun hak setiap orang untuk mendapatkan asuransi kesehatan dianggap sesuai dengan keinginan umum untuk menjaga kesehatan mereka(Jannah *et al.*, 2022).

Baik saat ini maupun di masa depan, keuntungan dari asuransi kesehatan menjadi sangat penting untuk kesehatan dan kesejahteraan keluarga. Bagi mereka yang tidak memiliki dana yang cukup untuk membayar perawatan kesehatan yang mahal, manfaat ini sangat penting(Hidayatullah *et al.*, 2021). Industri asuransi kesehatan sangat memperhatikan aspek pelayanannya. Salah satu elemen penting dari pelayanan rumah sakit adalah penyelesaian kasus, yang diatur oleh Perjanjian Tingkat Layanan (SLA) yang disetujui antara pihak tertanggung dan perusahaan asuransi(Cenora and Hermawan, 2022). Ketidaktahuan tentang biaya asuransi kesehatan dapat menyebabkan ketidakpahaman yang signifikan tentang berapa banyak yang harus dibayar secara mandiri dan bagaimana membayarnya, yang dapat mempengaruhi kemampuan pelanggan untuk membuat keputusan yang tepat tentang jenis asuransi kesehatan yang sesuai dengan kebutuhan mereka(Yumansya, Zy and Fatchan, 2023). Pada akhirnya, menggunakan teknologi informasi dan komunikasi membuat menjalankan berbagai tugas dan aktivitas sehari-hari menjadi lebih mudah. Sebagai contoh, analisis yang dilakukan untuk memproyeksikan biaya asuransi kesehatan di masa mendatang dapat digunakan untuk memprediksi biaya asuransi kesehatan. Ini dapat dilakukan dengan menggunakan teknik data mining seperti regresi linear berganda (Pratama *et al.*, 2023).

Pada penelitian sebelumnya yang berkaitan dengan analisis biaya asuransi kesehatan menggunakan algoritma regresi linear berganda dilakukan oleh (Sholeh, Suraya and Andayati, 2022) yang berjudul *Machine Linear* untuk Analisis Regresi Linier Biaya Asuransi Kesehatan dengan Menggunakan Python Jupyter Notebook. *Machine learning* atau pembelajaran mesin merupakan bagian integral dari kecerdasan buatan, dan salah satu algoritma populer yang digunakan adalah regresi linear untuk melakukan prediksi. penerapan prediksi biaya asuransi kesehatan yang dipengaruhi oleh berbagai faktor seperti umur, jenis kelamin, indeks massa tubuh (BMI), jumlah anak, status perokok, dan wilayah.

Sementara itu penelitian lainnya mengenai analisis regresi linear berganda yang dilakukan oleh(Puteri and Silvanie, 2020) yang berjudul *Machine Learning* Untuk Model Prediksi Harga Sembako Dengan Metode Regresi Linier Berganda. Harga sembilan bahan pokok (sembako) dapat mengalami fluktuasi setiap saat. Oleh karena itu, diperlukan peramalan harian untuk harga sembako dalam beberapa waktu ke depan. Salah satu metode yang dapat digunakan untuk meramalkan harga yang memiliki nilai numerik kontinu adalah menggunakan metode regres.

Penelitian selanjutnya yang dilakukan oleh (Yumansya, Zy and Fatchan, 2023) yang berjudul Prediksi Jumlah Kasus Klaim Indemnity Dengan Menggunakan Algoritma Regresi Linear Pada Asuransi Mandiri Inhealth. Dengan menggunakan algoritma regresi linear sederhana, proses prediksi dapat dilakukan. Hasilnya menawarkan perspektif baru tentang kebutuhan prediksi untuk data klaim.Setelah dibandingkan hasil perhitungan secara manual dan dengan aplikasi Rapid Miner, model persamaan regresi linear sederhana secara umum menunjukkan data yang sama. Pengujian dengan Rapid Miner menghasilkan performa yang relevan dengan skenario yang dimodelkan. Selain itu, evaluasi performa model yang diterapkan menghasilkan nilai RMSE sebesar 0,273, dengan standar deviasi +- 0,0.

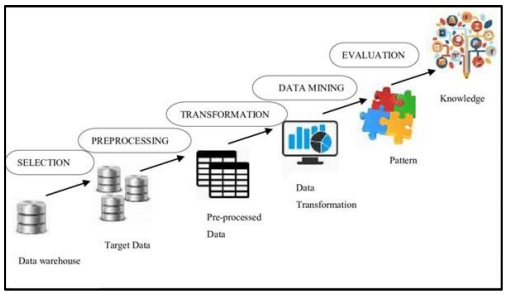
Analisis regresi linear berganda, juga dikenal sebagai analisis regresi ganda, adalah suatu model regresi yang melibatkan lebih dari satu variabel independen(Putri and Octova,

2022). Ini memanfaatkan perangkat lunak statistik untuk menemukan variabel independen yang memiliki pengaruh yang signifikan terhadap variabel dependen dan volume data yang besar(Sholeh, Suraya and Andayati, 2022).

Penelitian ini bertujuan untuk menganalisis biaya asuransi kesehatan dengan mengidentifikasi setiap variabel pada biaya asuransi(Maulita, 2023). Dalam analisis biaya asuransi kesehatan, Dataset ini telah diperbarui sekitar 25 November 2023. Pendekatan analisis yang diterapkan adalah regresi linear berganda, dan evaluasi hasilnya menggunakan metode RMSE. Dan metode yang digunakan adalah kdd *(Knowlegde Discovery in Database)*. *Software* yang digunakan dalam studi ini mencakup *Mendeley* dan *Rapidminer*, yang dijalankan pada sistem operasi Windows 10(Sholeh, Suraya and Andayati, 2022). Pada penelitian ini dataset yang digunakan yaitu tentang asuransi kesehatan, data yang diperoleh yaitu dari *kaggle* dengan nama dataset *Healthcare Insurance* sebanyak 1339 data. Terdiri dari 7 atribut *yaitu age, sex, bmi, children, smoker, region, charges.* Dari total 1339 data, data tersebut dipisahkan menjadi dua bagian, dimana 90% diantaranya digunakan sebagai data pelatihan dengan jumlah 1204 data, sementara sisanya, sebanyak 10%, digunakan sebagai data uji sebanyak 134 data.

METODE PENELITIAN

Knowledge Discovery In Database (KDD) adalah teknik analisis data yang digunakan dalam penelitian ini, yang menggunakan regresi linear berganda(Akmal, Faqih and Dikananda, 2023).



Gambar 1 KDD *(Knowledge Discovery in Database)*

Proses seleksi dan analisis data dalam konteks penentuan premi atau biaya asuransi kesehatan oleh peserta melibatkan evaluasi risiko kesehatan individu atau kelompok yang diasuransikan. Dalam proses ini, faktor-faktor seperti usia, riwayat kesehatan, dan lokasi geografis menjadi elemen kunci yang dievaluasi untuk menentukan sejauh mana kontribusi yang harus dibayarkan oleh peserta.

Berikut ini adalah penjelasan terperinci dari setiap langkah:

1. *Penentuan Dataset*

Pada langkah penelitian ini data yang di dapat dari *Healthcare Insurance | Kaggle.*

2. *Pre-Processing Data*

Dalam melakukan data cleaning, dilakukan proses penyaringan data yang akan digunakan dengan menghilangkan missing value, duplikasi data, serta melakukan pemeriksaan dan perbaikan kesalahan atau inkonsistensi data.

3. *Transformation*

Melakukan normalisasi data untuk menjaga konsistensi dalam skala, atau mengelompokkan kategori biaya asuransi untuk mempermudah analisis. informasi mengenai jenis klaim, riwayat penyakit, atau karakteristik khusus dari peserta asuransi kesehatan dapat diubah ke dalam format yang lebih cocok untuk analisis menyeluruh.

4. *Data Mining*

Mengidentifikasi pola atau informasi menarik dalam suatu kumpulan data dengan menggunakan metode dan algoritma. Pendekatan, metode, dan algoritma dalam data mining memiliki beragam variasi. Dalam konteks penelitian ini, penerapan data mining menggunakan algoritma regresi linear berganda karena melibatkan beberapa yaiut variabel independen, seperti, usia, jenis kelamin, berat badan, perokok, wilayah sementara variabel dependennya adalah jumlah biaya asuransi.

5. *Evaluation*

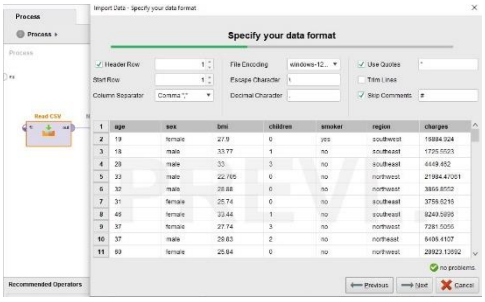
Evaluasi merupakan kesimpulan dari hasil data mining. Kesimpulan akhir dibuat dengan menggabungkan berbagai hipotesis yang dihasilkan dari data mining. Selanjutnya, dalam evaluasi akan diuraikan bagaimana hasil prediksi terkait jumlah biaya asuransi kesehatan.

HASIL DAN PEMBAHASAN

Hasil analisis biaya asuransi kesehatan yang dilakukan menggunakan algoritma regresi linear berganda dapat memberikan gambaran mendalam tentang bagaimana berbagai variabel independen, atau variabel tertentu, berhubungan dengan biaya asuransi kesehatan. Berikut adalah beberapa hal yang mungkin ditemukan dari penelitian ini:

1. Proses Data Selection

Pada *data selection* asuransi kesehatan merujuk pada proses pemilihan bagian data tertentu dari keseluruhan dataset untuk analisis dan pengolahan lebih lanjut, seperti yang ditunjukkan pada Gambar 2.



Gambar 2. Input data seleksi

Analisis data menunjukkan bahwa dataset yang digunakan adalah dataset yang diambil dari www.kaggle.com. Data dipilih menggunakan operator read csv pada rapid miner, yang menampilkan dataset awal sebanyak 15, seperti yang ditunjukkan pada Gambar 3.

Row No.	charges	age	sex	bmi	children	smoker	region
1	16884.924	19	female	27.900	0	yes	southwest
2	1725.552	18	male	33.770	1	no	southeast
3	4449.462	28	male	33	3	no	southeast
4	21984.471	33	male	22.705	0	no	northwest
5	3866.855	32	male	28.880	0	no	northwest
6	3756.622	31	female	25.740	0	no	southeast
7	8240.590	46	female	33.440	1	no	southeast
8	7281.506	37	female	27.740	3	no	northwest
9	6406.411	37	male	29.830	2	no	northeast
10	28923.137	60	female	25.840	0	no	northwest
11	2721.321	25	male	26.220	0	no	northeast
12	27808.725	62	female	26.290	0	yes	southeast
13	1826.843	23	male	34.400	0	no	southwest
14	11090.718	56	female	39.820	0	no	southeast
15	39611.758	27	male	42.130	0	yes	southeast

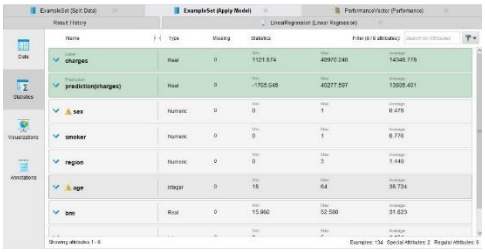
ExampleSet (1,338 examples, 1 special attribute, 6 regular attributes)

Gambar 3. Seleksi data

Hasil data asuransi terdiri dari 7 kolom. Data tersebut digunakan untuk melakukan analisis prediksi harga asuransi berdasarkan variabel-variabel berikut: usia , jenis kelamin, kategori berat badan (BMI), jumlah anak, perokok (perokok), dan lokasi tempat tinggal (wilayah).

2. Processing Data

Analisis data mencakup pemeriksaan data, seperti menemukan data yang tidak relevan, menghapus duplikat, dan mengubah data kategori menjadi format numerik. Gambar 4 berikut menunjukkan hasil dari dataset yang digunakan:



Gambar 4. Data statistika

Dalam tahap ini, informasi yang ada di dalam data digunakan untuk menentukannya. Sebelumnya, proses preprocessing, yang berarti membersihkan data yang hilang atau kosong, dilakukan sebelum proses selanjutnya. Karena data yang digunakan peneliti tidak memiliki data yang hilang atau kosong, penghapusan data yang kosong tidak dilakukan.

Row No.	charges	sex	smoker	region	age	bmi	children
1	16884.924	0	0	0	19	27.900	0
2	1725.552	1	1	1	18	33.770	1
3	4449.462	1	1	1	28	33	3
4	21984.471	1	1	2	33	22.705	0
5	3866.855	1	1	2	32	28.880	0
6	3756.622	0	1	1	31	25.740	0
7	8240.590	0	1	1	46	33.440	1
8	7281.506	0	1	2	37	27.740	3
9	6406.411	1	1	3	37	29.830	2
10	28923.137	0	1	2	60	25.840	0
11	2721.321	1	1	3	25	26.220	0
12	27808.725	0	0	1	62	26.290	0
13	1826.843	1	1	0	23	34.400	0
14	11090.718	0	1	1	56	39.820	0
15	39611.758	1	0	1	27	42.130	0

ExampleSet (1,338 examples, 1 special attribute, 6 regular attributes)

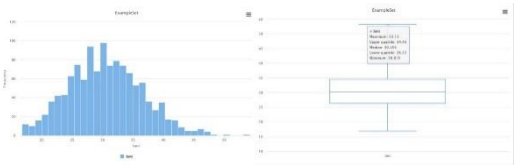
Gambar 5. Nominal ke Numerik

Hasil perubahan dari tipe nominal ke numerik terdapat pada d kolom sex (jenis kelamin), smoker (perokok), dan region (wilayah) telah diubah menjadi representasi numerik. Modifikasi ini dilakukan dengan tujuan untuk menyederhanakan proses regresi linear.

3. Transformation

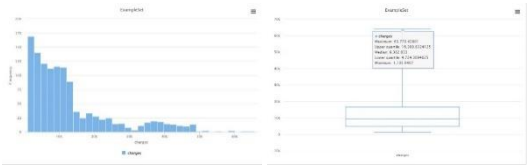
Visualisasi data adalah proses mengubah data mentah menjadi informasi yang ditunjukkan dalam bentuk grafik. Grafik seperti box plot, histogram, dan variasi lainnya digunakan dalam visualisasi data. Dalam visualisasi data, box plot secara statistik mencerminkan distribusi data melalui lima dimensi utama: nilai minimum, kuartil 1, kuartil 2 (median), dan kuartil 3. Untuk menilai kemungkinan adanya outlier dalam dataset, box plot sangat membantu. Selain itu, distribusi frekuensi dari kumpulan data numerik ditunjukkan dengan histogram, visualisasi data. Ada penggunaan box plot dan histogram untuk variabel usia, indeks massa tubuh, jumlah anak, dan biaya asuransi.

Gambar 6. histogram dan box plot Usia



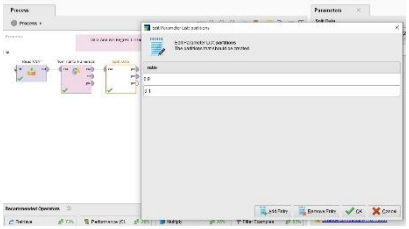
Gambar 7. histogram dan box plot (BMI)

Gambar 8. histogram dan box plot Jumlah Anak



Gambar 9. histogram dan box plot Biaya

Dalam tahapan *transformation* selanjutnya dilakukan split data ada beberapa perbandingan, tahapan pertama dilakukan perbandingan 9:1, tahapan ke dua 8:2 dan tahapan ke tiga dilakukan perbandingan 7:3. Berikut merupakan proses setiap perbandingan data :

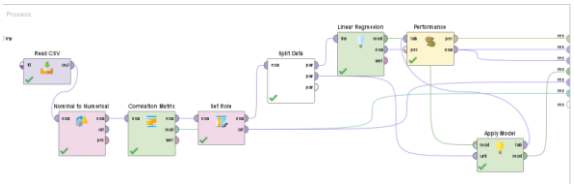


Gambar 10. Split data 9:1

Pembagian data dilakukan untuk membagi dataset menjadi set pelatihan (training set) dan set pengujian (testing set). Tujuan pembagian set pengujian adalah untuk melatih model pada sebagian kecil data dan menguji kinerja model pada sebagian kecil data yang tidak digunakan selama pelatihan. Ini membantu menentukan seberapa baik model dapat digeneralisasi pada data baru. Dari 1338 dataset yang ada, 90% adalah data pelatihan (1204 data) dan 10% adalah data pengujian (134 data).

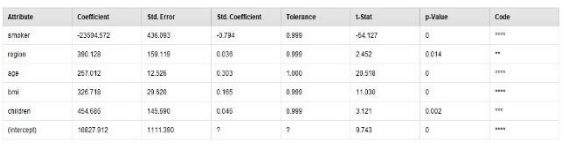
4. Data Mining

Proses data mining yang diterapkan dalam penelitian ini yaitu memprediksi menggunakan regresi linear berganda. Berdasarkan pengolahan data, pembuatan model regresi linear. Untuk menentukan pola hubungan antara variabel dependen dan variabel independen, dan nilai evaluasi dari hasil prediksi diukur dengan RMSE. Pada tahapan ini dilakukan perhitungan model regresi linear dari input read csv, lalu setelah itu dihubungan pada operator Nominal to Numerical berfungsi untuk merubah huruf menjadi angka, selanjutnya dihubungkan ke operator corelation matrix ini berfungsi untuk mencari nilai korelasi, setelah itu dihubungkan pada set role yaitu untuk pelabelan, selanjutnya dihubungkan ke split data bertujuan untuk membagi data antara data testing dan data training, lalu selanjutnya dihubungkan ke apply model, berfungsi untuk menerapkan model yang telah dilatih dari operator Linear Regression pada data testing dan training dari output split data, lalu terakhir dihubungkan pada port Result.



Gambar 11. Desain Pemodelan

Setelah Perhitungan regresi linear dilakukan dengan menggunakan operator linear regression. Dari total 1339 data, data tersebut dipisahkan menjadi dua bagian, dimana 90% diantaranya digunakan sebagai data pelatihan dengan jumlah 1204 data, sementara sisanya, sebanyak 10%, digunakan sebagai data uji sebanyak 134 data. dilakukan proses perhitungan nilai interception dan nilai masing-masing variabel independen.



Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
smoker	-23594.572	436.093	-0.794	0.999	-54.127	0	***
region	390.128	159.119	0.036	0.999	2.452	0.014	**
age	257.012	12.526	0.303	1.000	20.518	0	***
bmi	326.718	29.826	0.165	0.999	11.030	0	***
children	454.689	145.690	0.040	0.999	3.121	0.002	***
(intercept)	10627.912	1111.360	?	?	9.743	0	***

Gambar 12. Hasil Regresi Linear

Setelah hasil model diketahui, tahap selanjutnya melakukan proses perhitungan nilai interception dan nilai untuk semua variabel

independent. Hasil perhitungan dari data pada gambar 4.12, model multiple linear regression adalah Y = -10827.91 +257.01X1 +326.71X2 +454.68 X3 +23604.57 X4 +360.12 X5. Hasil model yang di dapat menunjukan dari 7 atribut yang ada hanya 5 atribut saja yang digunakan dalam memprediksi biaya yaitu smoker, region, age, bmi, children dan 1 atribut yang tidak di gunakan dalam memprediksi biaya yaitu sex.

Perhitungan korelasi diperlukan untuk mengetahui bagaimana setiap variabel independen berhubungan dengan variabel dependennya. Ini dilakukan pada baris 7, dan hasilnya ditunjukkan pada Gambar 13.

Attribut...	sex	smoker	region	age	bmi	children	charges
sex	1	-0.076	-0.005	-0.021	0.046	0.017	0.057
smoker	-0.076	1	-0.002	0.025	-0.004	-0.008	-0.787
region	-0.005	-0.002	1	-0.002	-0.158	-0.017	0.006
age	-0.021	0.025	-0.002	1	0.109	0.042	0.299
bmi	0.046	-0.004	-0.158	0.109	1	0.013	0.198
children	0.017	-0.008	-0.017	0.042	0.013	1	0.068
charges	0.057	-0.787	0.006	0.299	0.198	0.068	1

Gambar 13. Hasil Korelasi

Hasil korelasi pada Gambar 13 menunjukkan adanya hubungan yang kuat antara variabel smoker dengan charges (0,79), umur dengan charges (0.3), dan bmi dengan charges (0.2). Ini menyarankan bahwa individu yang merokok kemungkinan besar akan membayar premi asuransi yang lebih tinggi daripada non-perokok, dan terdapat hubungan yang relatif kuat antara usia dan biaya asuransi serta antara bmi dan biaya asuransi. Dengan demikian, berdasarkan prediksi dari korelasi antara umur dan bmi dengan biaya asuransi, biaya yang harus dibayarkan cenderung meningkat seiring dengan peningkatan usia atau indeks massa tubuh.

5. Evaluation

Pada tahap pengujian *corelation* yaitu untuk menilai sejauh mana variabel independen memengaruhi variabel dependen, dengan tujuan menentukan sejauh mana model regresi linier kesamaan dan cocok dengan data. Proses perhitungan uji *corelation* dan hasil perhitungan disajikan pada Tabel 1.

Tabel 1 Hasil Perhitungan

Nilai Korelasi	
Squared Correlation	0.81
RSME	5902.925
Squared Error	34844520.52

Berdasarkan hasil perhitungan yang tercantum dalam tabel, dapat disimpulkan bahwa hubungan antara variabel dependen dan variabel independen sangat kuat.

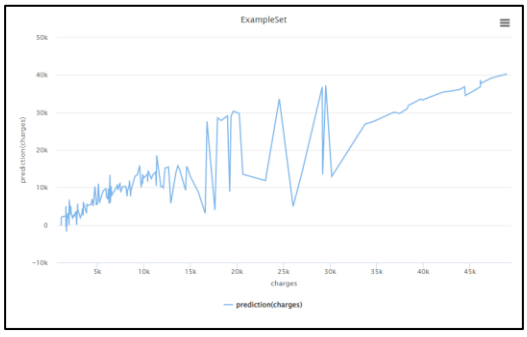
6. Pengujian Data Prediksi

Pengujian dilakukan untuk melakukan evaluasi terhadap perhitungan yang digunakan untuk memprediksi hasil dari regresi linear dengan menggunakan nilai aktual y. Sebanyak 134 data telah diuji dalam proses tersebut. Hasil pengujian pada data prediksi disajikan pada Gambar 14. Data yang ditampilkan sebanyak 13 dari 134 data.

Row No.	charges	prediction(charges)
1	4449.462	6955.547
2	36837.467	30071.407
3	2775.192	49.545
4	1625.434	-439.967
5	6272.477	5721.554
6	20630.284	13545.566
7	30166.618	12976.971
8	5920.104	9717.860
9	3947.413	5189.078
10	3877.304	5682.690
11	11381.325	18548.932
12	8601.329	9083.972
13	2166.732	4639.442

Gambar 14. Hasil Perbandingan Charges dan Prediction

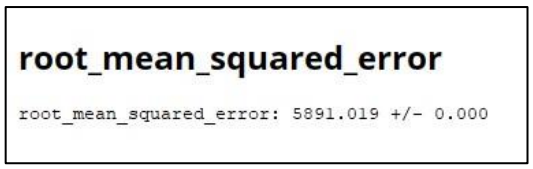
Hasil pada Gambar 14, antara data Y Charges dan Y Prediction ada perbedaan hasil dan hasil tampilan dalam bentuk grafik ditampilkan pada Gambar 15.



Gambar 15. Hasil Grafik prediction

Setelah pembangunan model selesai, dilakukan evaluasi kinerja model. Evaluasi dilakukan untuk menilai performa model dan menggunakan data pengujian yang telah ditentukan sebanyak 10% dari data testing.

Proses pengujian ini memanfaatkan operator performance. Hasil ditunjukan pada Gambar 16.



Gambar 16. Hasil Evaluasi RMSE

Berdasarkan pada Gambar 16, hasil nilai performa dengan menggunakan evaluasi root mean squared error menunjukkan nilai sebesar 5891. Dari hasil perhitungan dapat diketahui bahwa keterkaitan antara variabel dependen dan independen sangat kuat.

SIMPULAN

Berdasarkan hasil analisis dengan menggunakan algoritma regresi linear berganda untuk memprediksi biaya asuransi kesehatan, dapat dinyatakan bahwa terdapat korelasi yang penting antara variabel independen, seperti status perokok, usia, dan indeks massa tubuh (BMI), dengan biaya premi asuransi. Tingginya korelasi antara status perokok dan biaya premi asuransi mengindikasikan bahwa kebiasaan merokok memberikan dampak yang substansial terhadap tingginya biaya asuransi kesehatan. Hasil uji korelasi antara variabel independen menunjukkan bahwa ada korelasi 0,79 antara biaya, atau biaya, dan status perokok, atau perokok. Hasilnya menunjukkan bahwa perokok memiliki korelasi yang kuat dengan biaya premi asuransi. Oleh karena itu, diperkirakan bahwa perokok akan membayar premi asuransi yang lebih tinggi daripada orang yang tidak merokok. Dengan cara yang sama, terdapat korelasi antara biaya asuransi dengan usia 0,3 dan indeks massa tubuh (BMI) 0,2, yang menunjukkan bahwa semakin tua usia dan kategori berat badan (BMI), semakin tinggi biaya asuransi. Simpulan memberikan gambaran tentang hasil penelitian ini menekankan analisis biaya asuransi kesehatan menggunakan regresi linear berganda.

Disarankan untuk memperkuat upaya pengelolaan dan pencegahan kebiasaan merokok karena memiliki dampak yang signifikan pada biaya premi asuransi kesehatan. Meningkatkan performa dengan melibatkan metode peningkatan model atau mempertimbangkan penggunaan metode analisis prediktif lain yang dapat memberikan hasil yang lebih baik. Pengkajian mungkin dapat dilakukan dengan mengeksplorasi kategori usia atau rentang BMI tertentu untuk mendapatkan wawasan lebih mendalam mengenai dampaknya terhadap biaya asuransi.

DAFTAR PUSTAKA

Akmal, K., Faqih, A. and Dikananda, F. (2023) 'Perbandingan Metode Algoritma Naïve Bayes Dan K-Nearest Neighbors Untuk Klasifikasi Penyakit Stroke', *JATI (Jurnal Mahasiswa Teknik Informatika)*, 7(1), pp. 470–477. Available at: https://doi.org/10.36040/jati.v7i1.6367.

Cenora, E. and Hermawan, D. (2022) 'Asuransi dan Pandemi Covid-19: Peran Persepsi Konsumen dalam Keputusan Pembelian', *Ekonomi, Keuangan, Investasi dan Syariah (EKUITAS)*, 3(3), pp. 386–394. Available at: https://doi.org/10.47065/ekuitas.v3i3.1033.

Hidayatullah, R.S. *et al.* (2021) 'Klasifikasi Loyalitas Calon Pengguna Jasa Asuransi Kesehatan Menggunakan Algoritma Decision Tree Loyalty Classification of Prospective Health Insurance Service Users Using the Decision Tree Algorithm', (December). Available at: https://doi.org/10.13140/RG.2.2.35727.71842.

Jannah, M. *et al.* (2022) 'Analisis Biaya Premi Asuransi Kesehatan Untuk Kasus Rawat Jalan Berdasarkan Tingkatan Usia', *MAp (Mathematics and Applications) Journal*, 4(1), pp. 40–49. Available at: https://doi.org/10.15548/map.v4i1.4195.

Maulita, M. (2023) 'Pendekatan Data Mining Untuk Analisa Curah Hujan Menggunakan Metode Regresi Linear Berganda (Studi Kasus: Kabupaten Aceh Utara)', 6, pp. 99–106.

Pratama, R. *et al.* (2023) 'Prediksi Customer Retention Perusahaan Asuransi Menggunakan Machine Learning', *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, 12(1), pp. 96–104. Available at: https://doi.org/10.32736/sisfokom.v12i1.1507.

Puteri, K. and Silvanie, A. (2020) 'Machine Learning untuk Model Prediksi Harga Sembako', *Jurnal Nasional Informatika*, 1(2), pp. 82–94.

Putri, M. and Octova, A. (2022) 'Analisis Regresi Linear Berganda Terhadap Losstime Untuk Mencapai Target Produksi Limestone Crusher Vi Pt Semen Padang',

Jurnal Pertambangan, 5(4), pp. 193–202. Available at: https://doi.org/10.36706/jp.v5i4.972.

Sholeh, M., Suraya, S. and Andayati, D. (2022) 'Machine Linear untuk Analisis Regresi Linier Biaya Asuransi Kesehatan dengan Menggunakan Python Jupyter Notebook', *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, 8(1), p. 20. Available at: https://doi.org/10.26418/jp.v8i1.48822.

Yumansya, Q., Zy, A.T. and Fatchan, M. (2023) 'Prediksi Jumlah Kasus Klaim Indemnity Dengan Menggunakan Algoritma Regresi Linear Pada Asuransi Mandiri Inhealth', 4(2), pp. 299–305.