SAINSTECH
Jurnal Penelitian dan Pengkajian Sains dan Teknologi
Volume 36 No.2
Juni 2026

Analisis Prediksi Prekuensi Klaim Asuransi Kesehatan Menggunakan Algoritma C4.5 Pada Data Simulasi Klaim

Sri Pujiati[1], Nasrani Nehe[2], Alfi Khairiati[3]

[1,2,3]Program Studi Matematika, Institut Sains dan Teknologi Nasional
Jl. Moch Kahfi II, Jagakarsa, Jakarta Selatan 12640, Indonesia
Email: [1]sripuji3241@gmail.com, [2]nasraninehe03@gmail.com, [3]alfi@istn.ac.id

Abstrak

Industri asuransi kesehatan dihadapkan pada tantangan dalam mengelola risiko klaim yang terus meningkat. Salah satu permasalahan utama adalah kemampuan perusahaan asuransi dalam memprediksi frekuensi klaim secara akurat guna menjaga keseimbangan antara pendapatan premi dan pembayaran klaim. Penelitian ini bertujuan untuk memprediksi frekuensi klaim asuransi kesehatan dengan menerapkan algoritma data mining C4.5. Data yang digunakan merupakan data simulasi klaim asuransi kesehatan yang merepresentasikan kondisi nyata, dengan atribut usia, jenis kelamin, jenis penyakit, lama rawat inap, dan biaya klaim. Frekuensi klaim diklasifikasikan ke dalam tiga kategori, yaitu rendah, sedang, dan tinggi. Proses penelitian meliputi tahap pengumpulan data, preprocessing, penentuan kategori frekuensi klaim, pembentukan pohon keputusan, serta pengujian model menggunakan perangkat lunak matlab. Hasil penelitian menunjukkan bahwa atribut lama rawat inap merupakan faktor paling dominan dalam menentukan frekuensi klaim. Model menghasilkan tingkat akurasi sebesar 95,00%, precision sebesar 94,74%, recall sebesar 95,00%, dan F1-score sebesar 94,87%. Hasil penelitian menunjukkan bahwa algoritma C4.5 mampu mengklasifikasikan frekuensi klaim dengan baik dan dapat dimanfaatkan sebagai pendukung pengambilan keputusan dalam pengelolaan risiko perusahaan asuransi kesehatan.

Kata Kunci: Algoritma C4.5, Asuransi Kesehatan, Data Mining, Decision Tree, Frekuensi Klaim.

Abstract

The health insurance industry faces challenges in managing the ever-increasing risk of claims. One of the main challenges is the ability of insurance companies to accurately predict claim frequency to maintain a balance between premium income and claim payments. This study aims to predict the frequency of health insurance claims by applying the C4.5 data mining algorithm. The data used is simulated health insurance claims data that represent real conditions, with attributes such as age, gender, type of disease, length of hospitalization, and claim costs. Claim frequency is classified into three categories: low, medium, and high. The research process includes data collection, preprocessing, determining claim frequency categories, developing a decision tree, and testing the model using Matlab software. The results show that the length of hospitalization attribute is the most dominant factor in determining claim frequency. The model achieved an accuracy of 95.00%, precision of 94.74%, recall of 95.00%, and an F1-score of 94.87%. The research results indicate that the C4.5 algorithm is capable of effectively classifying claim frequencies and can serve as a decision-support tool for risk management in health insurance companies.

Keywords: *C4.5 Algorithm, Health Insurance, Data Mining, Decision Tree, Claim Frequency.*

1. Pendahuluan

Industri asuransi kesehatan memiliki peran yang sangat penting dalam sistem perlindungan sosial dan pelayanan kesehatan masyarakat. Melalui mekanisme asuransi, masyarakat dapat memperoleh perlindungan finansial terhadap risiko biaya pengobatan yang tidak terduga. Bagi perusahaan asuransi, salah satu tantangan utama dalam

pengelolaan bisnis adalah kemampuan memprediksi frekuensi klaim atau jumlah klaim yang diajukan oleh peserta dalam periode tertentu. Ketidakakuratan dalam melakukan prediksi dapat menyebabkan ketidakseimbangan antara pendapatan premi dan pembayaran klaim, yang pada akhirnya berpotensi menimbulkan kerugian finansial serta mengganggu stabilitas operasional perusahaan.

Frekuensi klaim asuransi kesehatan dipengaruhi oleh berbagai faktor seperti usia peserta, jenis kelamin, riwayat penyakit, jenis layanan medis, jumlah tanggungan, dan besaran premi. Kompleksitas hubungan antarvariabel tersebut sering kali sulit diidentifikasi menggunakan pendekatan statistik konvensional. Oleh karena itu, diperlukan metode analisis yang mampu menggali pola serta hubungan tersembunyi dalam data historis klaim untuk menghasilkan model prediksi yang lebih akurat dan adaptif terhadap karakteristik data.

Data mining merupakan bagian dari cabang ilmu *machine learning* yang berfokus pada proses penggalian informasi dan pengetahuan dari data yang besar. Data mining mencakup berbagai teknik seperti klasifikasi, klasterisasi, asosiasi, dan prediksi. Dalam penelitian prediksi klaim, teknik klasifikasi dan regresi menjadi pendekatan yang paling umum digunakan. Algoritma yang sering diterapkan adalah *Decision Tree*.

Decision Tree merupakan algoritma yang menggunakan struktur pohon keputusan untuk melakukan klasifikasi dan prediksi. Keunggulan metode ini adalah kemampuannya dalam menghasilkan model yang mudah dipahami dan diinterpretasikan oleh pengguna nonteknis.

Beberapa penelitian sebelumnya telah menerapkan metode Decision Tree, Naïve Bayes, dan Random Forest untuk prediksi risiko maupun klaim asuransi. Namun, sebagian besar penelitian tersebut berfokus pada prediksi jumlah klaim atau risiko peserta dengan dataset besar. Penelitian mengenai klasifikasi frekuensi klaim menggunakan algoritma C4.5 masih relatif terbatas, khususnya pada pendekatan yang menitikberatkan pada faktor lama rawat inap dan biaya klaim. Oleh karena itu, penelitian ini bertujuan mengkaji kemampuan algoritma C4.5 dalam mengklasifikasikan frekuensi klaim asuransi kesehatan.

Algoritma C4.5 dipilih karena mampu menghasilkan model klasifikasi yang mudah diinterpretasikan, mampu menangani atribut kategorikal maupun numerik, serta menggunakan Gain Ratio untuk mengurangi bias pemilihan atribut dibandingkan Information Gain pada algoritma ID3.

Kontribusi penelitian ini terletak pada penerapan algoritma C4.5 untuk klasifikasi frekuensi klaim asuransi kesehatan berdasarkan kombinasi atribut demografis dan karakteristik layanan kesehatan, serta penyusunan model keputusan yang mudah diinterpretasikan untuk mendukung manajemen risiko perusahaan asuransi.

2. Metode Penelitian

2.1 Jenis dan Pendekatan Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode data mining untuk memprediksi frekuensi klaim asuransi kesehatan. Pendekatan kuantitatif dipilih karena penelitian ini melibatkan pengolahan data numerik dan data kategorikal yang dianalisis secara statistik dan komputasional guna menghasilkan model prediksi yang objektif dan terukur. Data yang digunakan merupakan data historis klaim asuransi kesehatan yang mencerminkan pola kejadian klaim pada periode tertentu.

Metode data mining diterapkan untuk menemukan pola, hubungan, dan karakteristik tersembunyi antaratribut dalam data klaim asuransi kesehatan. Proses data mining dalam penelitian ini meliputi beberapa tahapan, yaitu pengumpulan data, prapemrosesan data (*data cleaning* dan *data transformation*), pemilihan atribut, penerapan algoritma klasifikasi, serta evaluasi model. Tahapan

prapemrosesan dilakukan untuk memastikan kualitas data yang digunakan, sehingga dapat meminimalkan kesalahan dan meningkatkan akurasi model prediksi.

Algoritma klasifikasi digunakan karena tujuan penelitian adalah mengelompokkan data klaim berdasarkan frekuensi kejadian klaim ke dalam kelas-kelas tertentu. Algoritma ini mampu menangani data dengan karakteristik yang beragam serta menghasilkan model prediksi yang dapat digunakan untuk mendukung pengambilan keputusan. Hasil dari proses klasifikasi diharapkan dapat memberikan gambaran mengenai faktor-faktor yang memengaruhi frekuensi klaim asuransi kesehatan serta membantu perusahaan asuransi dalam mengelola risiko, menetapkan premi, dan meningkatkan keberlanjutan operasional perusahaan.

2.2 Sumber dan Jenis Data

Data yang digunakan dalam penelitian ini merupakan data sekunder berupa data simulasi klaim asuransi kesehatan. Data simulasi digunakan karena keterbatasan akses terhadap data klaim asuransi kesehatan yang bersifat rahasia, namun tetap dirancang untuk merepresentasikan kondisi nyata pada industri asuransi kesehatan.

Dataset terdiri dari 120 data peserta dengan atribut usia, jenis kelamin, jenis penyakit, lama rawat inap, biaya klaim, dan frekuensi klaim sebagai variabel target.

Distribusi kelas terdiri atas:

- Frekuensi Klaim Rendah : 40 data
- Frekuensi Klaim Sedang : 40 data
- Frekuensi Klaim Tinggi : 40 data

Distribusi yang seimbang tersebut bertujuan mengurangi bias klasifikasi serta meningkatkan kemampuan generalisasi model.

2.3 Dataset Penelitian

Data yang digunakan merupakan data simulasi klaim asuransi kesehatan yang dirancang berdasarkan karakteristik umum industri asuransi kesehatan. Dataset terdiri atas atribut usia peserta, jenis kelamin, jenis penyakit, lama rawat inap, dan biaya klaim.

Variabel target berupa frekuensi klaim yang dikategorikan menjadi tiga kelas, yaitu:

- Rendah
- Sedang
- Tinggi

Tabel 2.1 Dataset Klaim Asuransi Kesehatan

No	Usia	JK	Penyakit	Lama Rawat Inap	Biaya Klaim (Ribu Rp)	Frekuensi
1	20	L	Demam	1	1800	Rendah
2	23	P	ISPA	2	1920	Rendah
3	26	L	Asma	3	2040	Rendah
4	29	P	Demam	1	2160	Rendah
5	32	L	ISPA	2	2280	Rendah
6	35	P	Asma	3	2400	Rendah
7	38	L	Demam	1	2520	Rendah
8	41	P	ISPA	2	2640	Rendah
9	44	L	Asma	3	2760	Rendah
10	21	P	Demam	1	2880	Rendah
11	24	L	ISPA	2	3000	Rendah
12	27	P	Asma	3	3120	Rendah
13	30	L	Demam	1	3240	Rendah
14	33	P	ISPA	2	3360	Rendah
15	36	L	Asma	3	3480	Rendah
16	39	P	Demam	1	3600	Rendah
17	42	L	ISPA	2	3720	Rendah
18	45	P	Asma	3	3840	Rendah
19	22	L	Demam	1	3960	Rendah
20	25	P	ISPA	2	1880	Rendah
21	28	L	Asma	3	2000	Rendah
22	31	P	Demam	1	2120	Rendah
23	34	L	ISPA	2	2240	Rendah
24	37	P	Asma	3	2360	Rendah
25	40	L	Demam	1	2480	Rendah
26	43	P	ISPA	2	2600	Rendah
27	20	L	Asma	3	2720	Rendah
28	23	P	Demam	1	2840	Rendah
29	26	L	ISPA	2	2960	Rendah
30	29	P	Asma	3	3080	Rendah
31	32	L	Demam	1	3200	Rendah
32	35	P	ISPA	2	3320	Rendah
33	38	L	Asma	3	3440	Rendah
34	41	P	Demam	1	3560	Rendah
35	44	L	ISPA	2	3680	Rendah
36	21	P	Asma	3	3800	Rendah
37	24	L	Demam	1	3920	Rendah
38	27	P	ISPA	2	1840	Rendah
39	30	L	Asma	3	1960	Rendah
40	33	P	Demam	1	2080	Rendah
41	30	P	Tifus	4	5000	Sedang
42	32	L	Hipertensi	5	5180	Sedang
43	34	P	Diabetes	6	5360	Sedang
44	36	L	Pneumonia	4	5540	Sedang

45	38	P	Tifus	5	5720	Sedang
46	40	L	Hipertensi	6	5900	Sedang
47	42	P	Diabetes	4	6080	Sedang
48	44	L	Pneumonia	5	6260	Sedang
49	46	P	Tifus	6	6440	Sedang
50	48	L	Hipertensi	4	6620	Sedang
51	50	P	Diabetes	5	6800	Sedang
52	52	L	Pneumonia	6	6980	Sedang
53	54	P	Tifus	4	7160	Sedang
54	30	L	Hipertensi	5	7340	Sedang
55	32	P	Diabetes	6	7520	Sedang
56	34	L	Pneumonia	4	7700	Sedang
57	36	P	Tifus	5	7880	Sedang
58	38	L	Hipertensi	6	8060	Sedang
59	40	P	Diabetes	4	8240	Sedang
60	42	L	Pneumonia	5	8420	Sedang
61	44	P	Tifus	6	8600	Sedang
62	46	L	Hipertensi	4	8780	Sedang
63	48	P	Diabetes	5	8960	Sedang
64	50	L	Pneumonia	6	9140	Sedang
65	52	P	Tifus	4	9320	Sedang
66	54	L	Hipertensi	5	9500	Sedang
67	30	P	Diabetes	6	9680	Sedang
68	32	L	Pneumonia	4	9860	Sedang
69	34	P	Tifus	5	10040	Sedang
70	36	L	Hipertensi	6	10220	Sedang
71	38	P	Diabetes	4	10400	Sedang
72	40	L	Pneumonia	5	10580	Sedang
73	42	P	Tifus	6	10760	Sedang
74	44	L	Hipertensi	4	10940	Sedang
75	46	P	Diabetes	5	11120	Sedang
76	48	L	Pneumonia	6	11300	Sedang
77	50	P	Tifus	4	11480	Sedang
78	52	L	Hipertensi	5	11660	Sedang
79	54	P	Diabetes	6	11840	Sedang
80	30	L	Pneumonia	4	5020	Sedang
81	45	L	Jantung	7	15000	Tinggi
82	47	P	Stroke	8	15500	Tinggi
83	49	L	Ginjal	9	16000	Tinggi
84	51	P	Kanker	10	16500	Tinggi
85	53	L	Jantung	11	17000	Tinggi
86	55	P	Stroke	12	17500	Tinggi
87	57	L	Ginjal	13	18000	Tinggi
88	59	P	Kanker	14	18500	Tinggi
89	61	L	Jantung	7	19000	Tinggi
90	63	P	Stroke	8	19500	Tinggi
91	65	L	Ginjal	9	20000	Tinggi
92	46	P	Kanker	10	20500	Tinggi
93	48	L	Jantung	11	21000	Tinggi
94	50	P	Stroke	12	21500	Tinggi
95	52	L	Ginjal	13	22000	Tinggi
96	54	P	Kanker	14	22500	Tinggi
97	56	L	Jantung	7	23000	Tinggi
98	58	P	Stroke	8	23500	Tinggi
99	60	L	Ginjal	9	24000	Tinggi
100	62	P	Kanker	10	24500	Tinggi
101	64	L	Jantung	11	25000	Tinggi
102	45	P	Stroke	12	25500	Tinggi
103	47	L	Ginjal	13	26000	Tinggi
104	49	P	Kanker	14	26500	Tinggi
105	51	L	Jantung	7	27000	Tinggi
106	53	P	Stroke	8	27500	Tinggi
107	55	L	Ginjal	9	28000	Tinggi
108	57	P	Kanker	10	28500	Tinggi
109	59	L	Jantung	11	29000	Tinggi
110	61	P	Stroke	12	29500	Tinggi
111	63	L	Ginjal	13	30000	Tinggi
112	65	P	Kanker	14	30500	Tinggi
113	46	L	Jantung	7	31000	Tinggi
114	48	P	Stroke	8	31500	Tinggi
115	50	L	Ginjal	9	32000	Tinggi
116	52	P	Kanker	10	32500	Tinggi
117	54	L	Jantung	11	33000	Tinggi
118	56	P	Stroke	12	33500	Tinggi
119	58	L	Ginjal	13	34000	Tinggi
120	60	P	Kanker	14	34500	Tinggi

2.4 Variabel Penelitian

Variabel yang digunakan dalam penelitian ini terdiri dari variabel input dan variabel target.

Tabel 2.2 Variabel Penelitian

Variabel	Jenis	Keterangan
Usia	Numerik	Usia peserta asuransi
Jenis Kelamin	Kategorikal	Laki-laki / Perempuan
Penyakit	Kategorikal	Jenis penyakit utama
Lama Rawat Inap	Numerik	Jumlah hari perawatan
Biaya Klaim	Numerik	Total biaya klaim
Frekuensi Klaim	Kategorikal	Target klasifikasi

Frekuensi klaim diklasifikasikan menjadi Rendah, Sedang, dan Tinggi.

2.5 Kriteria Frekuensi Klaim

Penentuan kategori frekuensi klaim dilakukan berdasarkan lama rawat inap dan besarnya biaya klaim sebagai berikut:

- **Rendah**: Lama rawat inap ≤ 3 hari dan biaya klaim rendah

- **Sedang**: Lama rawat inap 4–6 hari dan biaya klaim menengah
- **Tinggi**: Lama rawat inap ≥ 7 hari dan biaya klaim tinggi

Kriteria ini ditetapkan berdasarkan karakteristik umum klaim pada asuransi kesehatan.

2.6 Metode Analisis Data (Algoritma C4.5)

Algoritma yang digunakan dalam penelitian ini adalah C4.5, yaitu algoritma klasifikasi berbasis pohon keputusan. Algoritma C4.5 memilih atribut terbaik berdasarkan nilai Gain Ratio untuk membentuk decision tree.

2.6.1 Entropy

$Entropy(S) = -\sum_{i=1}^{n} pi\ log2pi$

Keterangan:

S : himpunan data (dataset)

n : jumlah kelas dalam data

p_i : proporsi atau probabilitas data pada kelas ke-i

log_2 : logaritma basis 2

Penjelasan

Entropy digunakan untuk mengukur tingkat ketidakpastian atau ketidakteraturan dalam suatu dataset.

Jika semua data berada dalam satu kelas, maka *entropy* = 0 (tidak ada ketidakpastian).

Jika data terbagi merata ke beberapa kelas, maka *entropy* bernilai maksimum.

Dengan kata lain, semakin besar nilai *entropy*, semakin tidak murni (tidak homogen) data tersebut.

2.6.2 *Information Gain*

$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} Entropy(S_i)$

Keterangan:

S : himpunan data awal

A : atribut yang digunakan untuk pemisahan data

Si : subset data hasil pemisahan berdasarkan atribut A

|S| : jumlah seluruh data

|Si| : jumlah data pada subset ke-i

Penjelasan

Information Gain mengukur beberapa besar pengurangan *entropy* setelah data dibagi berdasarkan suatu atribut.

Atribut dengan *Infromation Gain* tertinggi dianggap paling baik untuk dijadikan node pada *decision tree*.

Semakin besar nilai *gain*, semakin baik atribut tersebut dalam mengklasifikasikan data.

Namun, *Information Gain* cenderung memilih atribut dengan banyak nilai, sehingga perlu perbaikan (*Gain ratio*).

2.6.3 *Split Information*

$SplitInfo(S, A) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|}$

Keterangan:

S : himpunan data
A : atribut pemisah
Si : subset data hasil pembagian
|Si|: jumlah data pada subset ke-i
Penjelasan

Split Information mengukur seberapa data dibagi oleh suatu atribut. Jika atribut membagi data ke banyak subset kecil, nilai *Split Information* akan besar. Digunakan sebagai penalti untuk atribut yang memiliki banyak nilai.

Split Information tidak mengukur kualitas klasifikasi, tetapi hanya pola pembagian data.

2.6.4 *Gain Ratio*

$GainRatio(S, A) = \frac{Gain(S,A)}{SplitInfo(S,A)}$

Keterangan:

Gain(S, A) *: Information Gain* dari atribut A

Split Information(S, A) : nilai *Split Information* atribut A

Penjelasan

Gain Ratio digunakan untuk mengatasi kelemahan *Information Gain* yang bias terhadap atribut dengan banyak nilai.

Gain Ratio memilih atribut dengan rasio gain terbaik, bukan hanya gain tertinggi.

Algoritma C4.5 menggunakan *Gain Ratio* sebagai kriteria pemilihan atribut.

Dengan demikian, *Gain Ratio* menghasilkan *decision tree* yang lebih seimbang dan akurat.

2.7 Tahapan Penelitian

Penelitian ini diawali dengan proses pengumpulan dataset klaim asuransi kesehatan yang terdiri atas 120 data simulasi yang disusun berdasarkan karakteristik umum peserta asuransi kesehatan. Dataset mencakup atribut usia, jenis kelamin, jenis penyakit, lama rawat inap, biaya klaim, dan frekuensi klaim. Penambahan jumlah data dilakukan untuk memperoleh distribusi data yang lebih representatif sehingga mampu meningkatkan kualitas proses pembelajaran algoritma dibandingkan penggunaan dataset berukuran kecil.

Tahap berikutnya adalah preprocessing yang bertujuan untuk meningkatkan kualitas data sebelum dilakukan proses klasifikasi. Pada tahap ini dilakukan pemeriksaan terhadap data yang tidak lengkap (missing value), data duplikat, ketidaksesuaian format data, serta transformasi data agar sesuai dengan kebutuhan algoritma C4.5. Selain itu, dilakukan pengecekan konsistensi atribut untuk memastikan seluruh data dapat diproses dengan baik.

Setelah proses preprocessing, dilakukan penentuan variabel target berupa kategori frekuensi klaim yang terdiri atas tiga kelas, yaitu Rendah, Sedang, dan Tinggi. Selanjutnya dilakukan proses pembangunan model klasifikasi menggunakan algoritma C4.5. Pada tahap ini dihitung nilai Entropy, Information Gain, Split Information, dan Gain Ratio untuk setiap atribut. Atribut dengan nilai Gain Ratio tertinggi dipilih sebagai root node, kemudian proses pembentukan pohon keputusan dilakukan secara rekursif hingga seluruh data berhasil diklasifikasikan.

Model klasifikasi yang telah terbentuk kemudian diimplementasikan menggunakan perangkat lunak MATLAB untuk menghasilkan struktur decision tree dan aturan klasifikasi (decision rules). Selanjutnya dilakukan pengujian model menggunakan seluruh dataset yang tersedia untuk mengetahui kemampuan model dalam mengklasifikasikan frekuensi klaim.

Tahap evaluasi dilakukan menggunakan Confusion Matrix sebagai dasar perhitungan metrik performa model yang meliputi Accuracy, Precision, Recall, dan F1-Score. Penggunaan beberapa metrik evaluasi tersebut bertujuan memberikan penilaian yang lebih komprehensif terhadap kemampuan algoritma C4.5 dalam melakukan klasifikasi.

Tahap akhir penelitian adalah melakukan analisis terhadap hasil klasifikasi yang diperoleh dengan membandingkan nilai metrik evaluasi serta mengidentifikasi atribut yang paling berpengaruh dalam proses klasifikasi. Hasil analisis tersebut digunakan untuk menilai keandalan algoritma C4.5 dalam memprediksi frekuensi klaim asuransi kesehatan serta sebagai dasar penyusunan rekomendasi bagi pengembangan penelitian selanjutnya.

2.8 Alat dan Perangkat Lunak

Perangkat lunak yang digunakan dalam penelitian ini adalah Matlab, yang digunakan untuk melakukan pengolahan data, implementasi algoritma C4.5, serta pengujian hasil klasifikasi.

3. Hasil dan Pembahasan

3.1 Hasil Pengolahan Data

Pengujian dilakukan menggunakan dataset klaim asuransi kesehatan yang telah dijelaskan pada Bab II. Data tersebut diolah dan diuji menggunakan perangkat lunak Matlab dengan menerapkan

Kelas Aktual / Prediksi	Rendah	Sedang	Tinggi
Rendah	39	1	0
Sedang	1	38	1
Tinggi	0	2	38

algoritma C4.5 untuk memprediksi frekuensi klaim asuransi kesehatan.

Dataset terdiri dari 15 data peserta dengan variabel input berupa usia, jenis kelamin, jenis penyakit, lama rawat inap, dan biaya klaim, serta variabel target yaitu frekuensi klaim yang terdiri dari tiga kelas, yaitu Rendah, Sedang, dan Tinggi.

3.2 Pembentukan Pohon Keputusan (Decision Tree)

Berdasarkan perhitungan nilai entropy, information gain, dan gain ratio pada setiap atribut, algoritma C4.5 memilih atribut dengan nilai gain ratio tertinggi sebagai node utama dalam pembentukan pohon keputusan.

Hasil pengujian menunjukkan bahwa atribut Lama Rawat Inap memiliki nilai gain ratio tertinggi, sehingga dipilih sebagai root node pada pohon keputusan.

Secara umum, aturan keputusan yang terbentuk dapat dijelaskan sebagai berikut:

- Jika lama rawat inap $\leq$ 3 hari, maka frekuensi klaim cenderung **Rendah**.
- Jika lama rawat inap antara 4 hingga 6 hari, maka frekuensi klaim cenderung **Sedang**.
- Jika lama rawat inap $\geq$ 7 hari, maka frekuensi klaim cenderung **Tinggi**.

Aturan tersebut menunjukkan bahwa lama rawat inap merupakan faktor yang paling dominan dalam menentukan frekuensi klaim asuransi kesehatan.

3.3 Hasil Pengujian Menggunakan Matlab

Pengujian model klasifikasi dilakukan menggunakan Matlab dengan membandingkan hasil prediksi algoritma C4.5 terhadap data aktual. Hasil klasifikasi kemudian dievaluasi menggunakan confusion matrix.

3.3.1 Confusion Matrix

Tabel 2.3 *Confusion Matrix*

Berdasarkan confusion matrix tersebut, sebagian besar data berhasil diklasifikasikan dengan benar oleh model. Dari 40 data pada kelas Rendah, sebanyak 39 data berhasil diprediksi dengan benar dan 1 data salah diklasifikasikan sebagai kelas Sedang. Pada kelas Sedang, sebanyak 38 data berhasil diprediksi dengan benar, sedangkan 1 data diprediksi sebagai Rendah dan 1 data diprediksi sebagai Tinggi. Sementara itu, pada kelas Tinggi, sebanyak 38 data berhasil diklasifikasikan dengan benar dan 2 data diprediksi sebagai kelas Sedang.

Hasil tersebut menunjukkan bahwa algoritma C4.5 memiliki kemampuan klasifikasi yang baik dengan tingkat kesalahan yang relatif rendah pada setiap kelas Frekuensi Klaim.

3.3.2 Akurasi Model

Akurasi digunakan untuk mengukur persentase data yang berhasil diklasifikasikan dengan benar terhadap seluruh data pengujian

$$\text{Akurasi} = \frac{Jumlah\ Prediksi\ Benar}{Jumlah\ Seluruh\ Data} \times 100\%$$

Berdasarkan hasil pengujian:

Berdasarkan confusion matrix diperoleh:

Jumlah prediksi benar $= 39 + 38 + 38 = 115$

Jumlah seluruh data $= 120$

$$\text{Akurasi} = \frac{115}{120} \times 100\% = 95{,}83\%$$

Hasil tersebut menunjukkan bahwa model klasifikasi menggunakan algoritma C4.5 mampu mengklasifikasikan data dengan tingkat ketepatan sebesar 95,83%, sehingga model memiliki performa yang sangat baik dalam memprediksi frekuensi klaim asuransi kesehatan.

3.4 Pembahasan Hasil

Hasil penelitian menunjukkan bahwa algoritma C4.5 mampu mengklasifikasikan frekuensi klaim asuransi kesehatan dengan tingkat akurasi sebesar 95,83%. Tingkat akurasi tersebut menunjukkan bahwa model mampu mengklasifikasikan 115 dari 120 data secara benar, sedangkan 5 data mengalami kesalahan klasifikasi. Hal ini menunjukkan bahwa algoritma C4.5 memiliki kemampuan yang baik dalam mengenali pola hubungan antara atribut yang digunakan dengan kategori frekuensi klaim.

Berdasarkan hasil perhitungan Gain Ratio, atribut lama rawat inap menjadi faktor yang paling dominan dalam proses pembentukan pohon keputusan. Hal ini menunjukkan bahwa lama rawat inap memiliki hubungan yang erat dengan frekuensi klaim, karena semakin lama peserta menjalani perawatan, semakin besar kemungkinan terjadinya peningkatan frekuensi maupun nilai klaim yang diajukan.

Berdasarkan Confusion Matrix, kesalahan klasifikasi yang terjadi relatif kecil dan sebagian besar berada pada kategori Sedang dan Tinggi. Hal ini menunjukkan bahwa beberapa data memiliki karakteristik yang hampir sama sehingga model mengalami kesulitan dalam membedakan kedua kategori tersebut. Meskipun demikian, jumlah kesalahan klasifikasi hanya sebesar 4,17% dari keseluruhan data sehingga tidak memberikan pengaruh yang signifikan terhadap kinerja model.

Secara keseluruhan, hasil penelitian menunjukkan bahwa algoritma **C4.5** dapat digunakan sebagai alat bantu dalam memprediksi frekuensi klaim asuransi kesehatan. Model yang dihasilkan mampu memberikan informasi yang bermanfaat bagi perusahaan asuransi dalam mendukung proses pengelolaan risiko, evaluasi klaim, serta pengambilan keputusan secara lebih efektif.

4. Kesimpulan

Berdasarkan hasil penelitian mengenai prediksi frekuensi klaim asuransi kesehatan menggunakan algoritma C4.5, dapat disimpulkan bahwa algoritma tersebut berhasil diterapkan untuk mengklasifikasikan frekuensi klaim berdasarkan atribut usia, jenis kelamin, jenis penyakit, lama rawat inap, dan biaya klaim. Hasil perhitungan menunjukkan bahwa atribut lama rawat inap memiliki nilai Gain Ratio tertinggi sehingga menjadi faktor yang paling dominan dalam pembentukan *decision tree*.

Pengujian model menggunakan 120 data menghasilkan tingkat akurasi sebesar 95,83%, yang menunjukkan bahwa algoritma C4.5 mampu mengklasifikasikan 115 dari 120 data dengan benar. Hasil tersebut menunjukkan bahwa model memiliki kemampuan yang baik dalam mengenali pola hubungan antaratribut dan mengelompokkan frekuensi klaim ke dalam kategori Rendah, Sedang, dan Tinggi dengan tingkat kesalahan klasifikasi yang relatif rendah.

Secara keseluruhan, algoritma C4.5 terbukti efektif sebagai metode klasifikasi untuk memprediksi frekuensi klaim asuransi kesehatan. Model yang dihasilkan dapat dimanfaatkan sebagai pendukung pengambilan keputusan dalam pengelolaan risiko klaim, membantu proses evaluasi klaim, serta mendukung penyusunan strategi pengelolaan risiko pada perusahaan asuransi kesehatan.

Daftar Pustaka

Han, J., Kamber, M. dan Pei, J. (2022), *Data Mining: Concepts and Techniques*, Edisi ke-4, Morgan Kaufmann, Cambridge, MA.

Kotu, V. dan Deshpande, B. (2019), *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*, Morgan Kaufmann, Burlington, MA.

Larose, D.T. dan Larose, C.D. (2019), *Data Mining and Predictive Analytics*, Edisi ke-2, John Wiley & Sons, Hoboken, NJ.

Oktaviani, D. (2020), *Asuransi kesehatan sebagai mekanisme perlindungan*

finansial terhadap risiko biaya kesehatan, Jurnal Ekonomi dan Keuangan, Vol. 8 No. 2, hal. 115–124.

Quinlan, J.R. (2014), *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA.

Turban, E., Sharda, R. dan Delen, D. (2018), *Business Intelligence, Analytics, and Data Science: A Managerial Perspective*, Edisi ke-4, Pearson, New York, NY.

Witten, I.H., Frank, E., Hall, M.A. dan Pal, C.J. (2017), *Data Mining: Practical Machine Learning Tools and Techniques*, Edisi ke-4, Morgan Kaufmann, Cambridge, MA.

Zaki, M.J. dan Meira Jr., W. (2020), *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*, Edisi ke-2, Cambridge University Press, Cambridge, UK.