Jurnal TRANSFORMATIKA
Vol.23, No.1, Juli 2025, pp. 118 - 125
P-ISSN: 1693-3656, E-ISSN: 2460-6731
https://journals.usm.ac.id/index.php/transformatika/

Komparasi Algoritma Random Forest dan XGBoost dalam Prediksi Premi Asuransi Kesehatan

Windy Karuniasari[1], Rastri Prathivi[2]
[1]Fakultas Teknologi Informasi dan Komunikasi, Universitas Semarang
Jl. Soekarno Hatta, Semarang, telp:024-6702757, e-mail: windykaruniasari26@gmail.com
[2] Fakultas Teknologi Informasi dan Komunikasi, Universitas Semarang
Jl. Soekarno Hatta, Semarang, telp:024-6702757, e-mail: vivi@usm.ac.id

ARTICLE INFO

History of the article :
Received 23 April 2025
Received in revised form 25 April 2025
Accepted 30 Juni 2025
Available online 31 Juli 2025

Keywords:

Asuransi kesehatan; Komparasi; Random Forest; XGBoost

*** Correspondece:**

Telepon:
+62 81910726485

E-mail:
windykaruniasari26@gmail.com

ABSTRACT

Asuransi kesehatan saat ini menjadi salah satu hal yang banyak orang persiapkan dikarenakan adanya ketidakpastian risiko kesehatan dan biaya layanan kesehatan yang semakin naik. Perhitungan premi tiap individu dapat berbeda dikarenakan terdapat perbedaan profil kesehatan seperti usia, BMI maupun gaya hidup seperti merokok yang membuat perusahaan asuransi harus memperhitungkan premi dengan akurat agar tidak menimbulkan kerugian finansial dan sesuai dengan tingkat risiko terjadinya klaim. Adapaun tujuan dari penelitian ini adalah melakukan komparasi antara algoritma Random Forest dan XGBoost dalam memprediksi premi asuransi kesehatan berdasarkan beberapa faktor yang sulit dihitung secara manual. Evaluasi dilihat berdasarkan metrik regresi yaitu MAE, MSE, RMSE, dan R2. Pada penelitian ini, algoritma Random Forest berhasil memprediksi premi asuransi kesehatan lebih baik dari XGBoost dengan nilai MAE 2.573, MSE 24199792,43 RMSE 4919, 33 dan R2 sebesar 84.04%.

INTRODUCTION

Tingginya biaya layanan kesehatan serta meningkatnya ketidakpastian risiko kesehatan, semakin menyadarkan individu pentingnya asuransi kesehatan. Asuransi kesehatan adalah upaya untuk mengatasi resiko ketidakpastian akibat sakit dan biaya-biaya yang ditimbulkannya [1]. Perusahaan asuransi dalam hal ini berperan sebagai lembaga yang menyediakan perlindungan finansial kepada individu atau keluarga apabila terjadi keadaan darurat medis yang tidak dapat diprediksi. Perusahaan asuransi kesehatan semakin berfokus pada pencegahan primer untuk mengurangi biaya kesehatan jangka panjang, terutama dalam menangani penyakit kronis yang disebabkan oleh gaya hidup [2].

Dengan mengumpulkan risiko dari banyak individu, perusahaan asuransi dapat menilai probabilitas terjadinya klaim dan menetapkan premi yang sesuai [3]. Keakuratan dalam penentuan premi membantu perusahaan asuransi menjaga stabilitas finansial dengan memastikan bahwa premi mencerminkan risiko yang sebenarnya. Hal ini penting untuk menghindari kerugian finansial yang

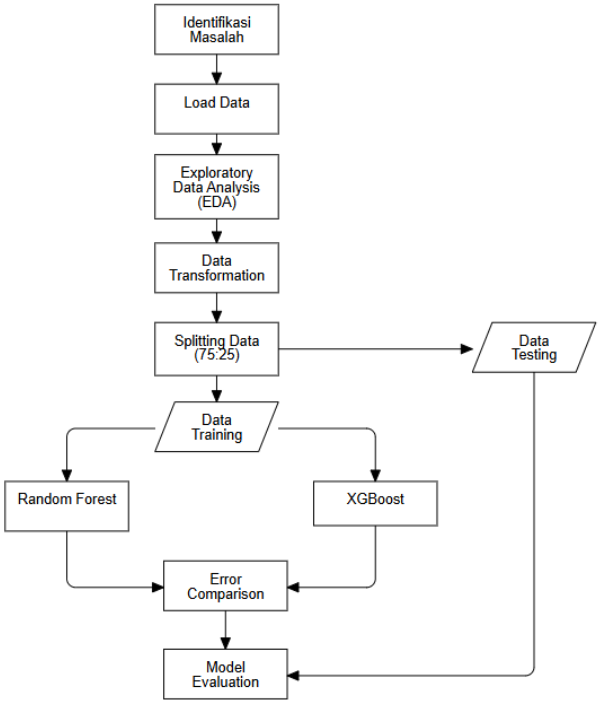
dapat terjadi jika premi ditetapkan terlalu rendah [4][5]. Penentuan premi tentunya tidak sama pada tiap individu. Faktor seperti usia, indeks massa tubuh (BMI), kebiasaan merokok, dan lokasi geografis adalah faktor utama yang mempengaruhi premi asuransi kesehatan [6][7][8][9]. Dikarenakan terdapat perbedaan faktor tersebut, prediksi akurat mengenai premi asuransi kesehatan sangat penting bagi perusahaan asuransi dan pemegang polis, karena dapat mengoptimalkan strategi penetapan harga, meningkatkan penilaian risiko, dan memperbaiki pengambilan keputusan di sektor asuransi kesehatan [10].

Penggunaan Machine Learning dalam prediksi premi memungkinkan perusahaan asuransi untuk menghasilkan estimasi premi yang lebih akurat, dan transparan dalam proses pengambilan keputusan. Algoritma pembelajaran mesin dapat mengidentifikasi pola dalam data, yang kemudian digunakan untuk membuat keputusan atau prediksi tanpa memerlukan pemrograman eksplisit untuk setiap tugas [11]. Algoritma XGBoost dan Random Forest adalah algoritma pembelajaran mesin yang termasuk dalam kategori *ensemble learning*. *Ensemble learning* menggabungkan beberapa model pembelajaran mesin untuk menghasilkan prediksi yang lebih akurat dan stabil dibandingkan dengan model tunggal [12]

Kedua algoritma ini memungkinkan analisis fitur penting yang mempengaruhi prediksi, seperti kebiasaan merokok, indeks massa tubuh (BMI), dan tingkat tekanan darah, yang penting dalam menentukan kebijakan dan strategi penetapan harga asuransi [13]. Penelitian [14] menunjukkan bahwa random forest dapat memprediksi biaya asuransi mendekati nilai aktual. Sementara itu menurut penelitian [15], extreme gradient boosting juga menunjukkan kinerja yang lebih baik secara keseluruhan dibandingkan model lain seperti random forest, meskipun menggunakan lebih banyak sumber daya komputasi. Oleh karena itu, penelitian ini akan membandingkan kinerja kedua algoritma tersebut untuk memprediksi premi asuransi kesehatan.

RESEARCH METHODS

Secara garis besar alur metode yang digunakan pada penelitian ini adalah sebagai berikut :



Gambar 1. Alur Penelitian

DOI : https://doi.org/10.26623/transformatika.v23i1.11914

1. **Identifikasi masalah**
   Adanya perbedaan profil risiko kesehatan pada tiap individu membuat adanya kompleksitas pada prediksi biaya asuransi kesehatan. Data ini umumnya terdiri dari variabel yang bersifat kuantitatif maupun kualitatif. Variabel-variabel ini dapat saling berpengaruh terhadap satu sama lain, sehingga sulit untuk dianalisis dengan cara manual. *Machine learning* dapat mengidentifikasi pola dalam data sehingga dapat memberikan manfaat bagi bisnis dan organisasi dalam membuat keputusan yang lebih baik.
2. **Pengumpulan Data**
   Pada penelitian ini, data yang digunakan merupakan data sekunder yang didapat dari www.kaggle.com. Adapun data ini merupakan data histori premi asuransi kesehatan.
3. **Exploratory Data Analysis (EDA)**
   EDA merupakan tahap awal untuk menganalisa data, EDA digunakan untuk mengetahui statistik data, seperti informasi mengenai nilai rata-rata, standar deviasi dan kuartil selain itu EDA dapat digunakan untuk mengetahui distribusi data, pola data dan hubungan antar data dengan menggunakan visualisasi.
4. **Data Tranformation**
   Data *transformation* merupakan tahap pemrosesan data untuk mengubah data kategorikal menjadi numerikal agar dapat diproses pada model *machine learning*.
5. **Splitting Data**
   Splitting data dilakukan untuk membagi data menjadi 2 bagian yaitu data training dan data testing. Pada penelitian ini menggunakan skala (75:25) yaitu 75% data training, 25% data testing.
6. **Error Comparison**
   Pada proses ini dilakukan evaluasi performa kedua model menggunakan data latih dengan cross-validation. Pada tahap ini akan dilakukan perbandingan antara Random Forest dan XGBoost dengan metrik *Root Mean Squared Error* (RMSE).
7. **Model Evaluation**
   Tahap terakhir merupakan evaluasi kedua model menggunakan data uji dengan metrik penilaian seperti MAE, MSE, RMSE, dan R2.

RESULTS

Hasil dan pembahasan dari penelitian ini memiliki tahapan sebagai berikut :

1. **Analisa Data**
   Dalam penelitian ini, terdapat 1338 data dengan 7 total kolom yaitu age, sex, BMI, children, smoker, region dan charges. Statistik data dapat dilihat pada tabel berikut :

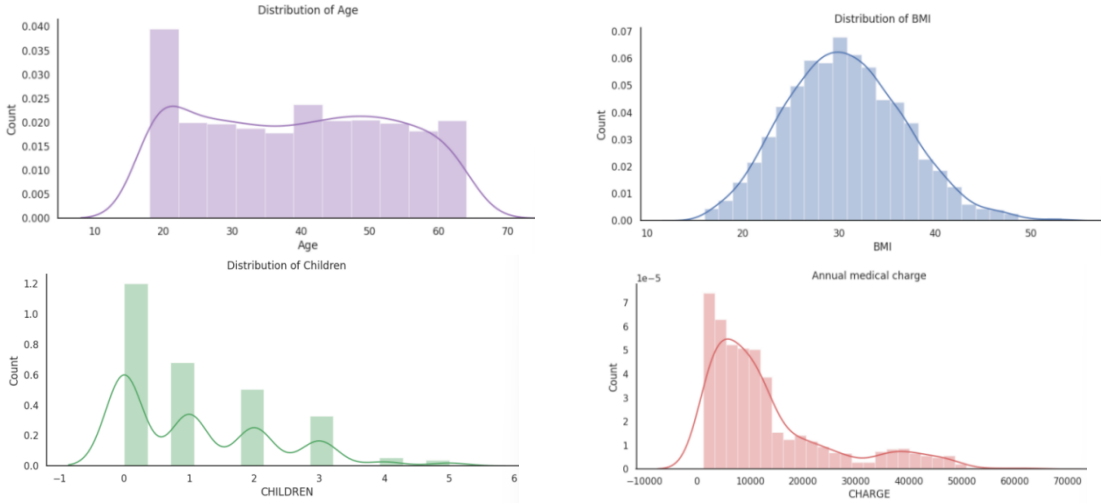
Table 1. Statistik Data

	Age	BMI	Children	Charges
Count	1338	1338	1338	1338
Mean	39.207025	30.663397	1.094918	13270.422265
Std	14.09960	6.098187	1.205493	12110.011237
Min	18	15.96	0	1121.873900
25%	27	26.29625	0	4740.287150
50%	39	30.4	1	9382.033000
75%	51	34.69375	2	16639.912515
Max	64	15.96	5	63770.428010

Tabel diatas merupakan statistik data numerik. Statistik diatas menunjukan informasi mengenai jumlah data, nilai rata-rata, nilai terbesar dan terkecil hingga nilai kuartil. Pada data age statistik data menunjukan penyebaran usia yang cukup besar dengan nilai std 14.05, peserta rata-rata berusia 39 tahun, termuda 18 tahun, tertua 64 tahun. Pada data BMI rata-rata peserta memiliki BMI *overweight* dengan mean 30.66, sebaran data cukup luas dengan std 6.10, dan median 30.4 cukup dekat dengan nilai mean yang artinya distribusi normal. Pada data children terdapat peserta yang tidak memiliki anak dan paling banyak peserta memiliki 5 anak. Pada data charges terlihat memiliki standar deviasi yang besar artinya biaya sangat bervariasi, nilai mediannya 9.382 jauh dibawah nilai mean yang memungkinkan adanya *right skewed*. Visualisasi dibutuhkan untuk memperkuat temuan ini.

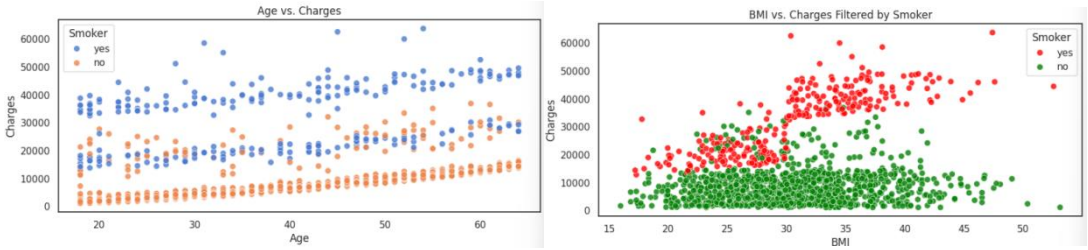
2. Visualisasi Data

Visualisasi data digunakan untuk mempelajari pola dan distribusi data sehingga berguna dalam menentukan proses pengolahan data. Dibawah ini merupakan grafik persebaran data :



Gambar 2. Distribusi Data

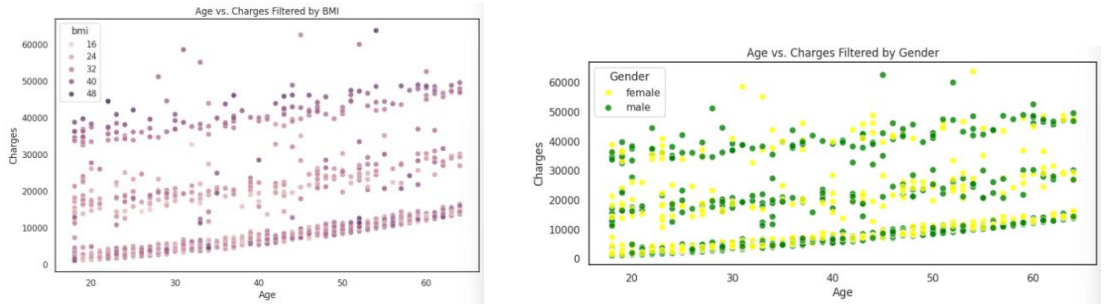
Dari grafik diatas terlihat bentuk pola grafik yang berbeda. Pada Age terdapat 2 puncak yang memungkinkan data usia didominasi oleh dua kelompok umur yang berbeda, Pada BMI distribusi mendekati normal (*bell curve*), rata rata orang memiliki nilai BMI sekitar 30. Data children merupakan data diskrit rata-rata orang tidak memiliki anak hingga 2 anak, hanya sedikit yang memiliki lebih dari 3 anak. Distribusi charge terlihat *right skewed*, sebagian besar oang memiliki premi dibawah 15.000 namun ada juga yang memiliki premi hampir 70.000, terdapat outlier yang signifikan yang mungkin dapat disebabkan adanya kondisi khusus seperti faktor usia, BMI atau perokok. Untuk itu perlu dilakukan visualisasi untuk mengetahui seberapa besar faktor-faktor tersebut dalam mempengaruhi besaran premi asuransi kesehatan.



Gambar 3. Hubungan Usia dan BMI terhadap Charge Berdasarkan Status Perokok

DOI : https://doi.org/10.26623/transformatika.v23i1.11914

Grafik diatas menunjukan hubungan antara usia dan biaya dengan BMI dan biaya dilihat dari faktor perokok. Pada hubungan antara usia dan biaya di semua rentang usia, perokok memiliki biaya medis lebih tinggi dibanding non perokok. Pada perokok biaya cenderung meningkat seiring dengan bertambahnya usia. Pada non perokok biaya medis cenderung stabil, walaupun tetap sedikit meningkat seiring dengan bertambahnya usia. Pada hubungan antara BMI dan biaya yang dipengaruhi faktor perokok, perokok dengan nilai BMI tinggi memiliki biaya medis yang sangat melambung tinggi ditandai dengan titik-titik merah. Sementara non perokok memiliki biaya yang relatif rendah meskipun nilai BMI meningkat ditandai titik-titik hijau.



Gambar 3. Hubungan Usia dan Biaya Berdasarkan BMI dan Jenis Kelamin

Grafik diatas menunjukan perbandingan antara usia dan biaya dipengaruhi faktor BMI dan jenis kelamin. Pada grafik kiri perbandingan antara usia dan biaya dipengaruhi oleh BMI menunjukan seiring dengan bertambahnya usia, biaya cenderung meningkat. Namun faktor BMI tidak terlalu berpengaruh terhadap biaya dikarenakan titik titik yang berwarna gelap berada di beberapa posisi dan cenderung menyebar. Pada grafik kanan menunjukan tidak ada perbedaan yang signifikan antara *Male* dan *Female* dalam hal biaya berdasarkan usia, dapat disimpulkan bahwa jenis kelamin tidak terlalu mempengaruhi besaran biaya asuransi. Dibawah ini merupakan tabel nilai korelasi untuk mengetahui lebih detail mengenai seberapa kuat hubungan antar variabel dengan nilai antara -1 hingga 1.

Table 2. Nilai Korelasi

	age	**sex**	**bmi**	**children**	**smoker**	**region**	**charges**
Age	1	-0.02	0.11	0.042	-0.026	0.0016	0.3
Sex	-0.02	1	0.046	0.018	0.077	0.0049	0.058
Bmi	0.11	0.046	1	0.013	0.0037	0.16	0.2
Children	0.042	0.018	0.013	1	0.0073	0.016	0.067
Smoker	-0.026	0.077	0.0037	0.0073	1	-0.0024	0.79
Region	0.0016	0.0049	0.16	0.016	-0.0024	1	-0.0065
Charges	0.3	0.058	0.2	0.067	0.79	-0.0065	1

Dari informasi tabel diatas menunjukan perokok memiliki korelasi yang paling kuat dengan nilai 0.79. Hal ini selaras dengan visualisasi sebelumnya yang memperlihatkan Smoker menjadi faktor yang sangat berpengaruh terhadap tingginya biaya asuransi kesehatan. Setelah Smoker, faktor Age dan BMI juga cukup berpengaruh namun tidak sekuat smoker. Sementara faktor sex, region dan children memiliki pengaruh yang kecil terhadap charges. Sehingga urutan fitur terpenting secara berurutan yaitu smoker, age dan BMI.

3. Preprocessing Data

Proses *preprocessing* meliputi data *cleaning*, data *selection* dan *data transformation*. Pada tahap analisis data tidak ditemukan adanya *missing value* namun ditemukan 1 *duplicat*e data sehingga perlu dilakukan penghapusan data *duplicate*. Semua fitur akan dipakai untuk memberikan kompleksitas pada model. Transformasi data menggunakan Label Encoding untuk merubah variabel kategorikal menjadi numerik agar dapat diproses oleh model. Dalam hal ini, perubahan dilakukan untuk data region, smoker dan sex yang memiliki tipe kategorikal.

Table 3. Hasil Encoding

	age	sex	bmi	children	smoker	region	Charges
0	19	0	27.9	0	1	3	16884.924
1	18	1	33.77	1	0	2	1725.5523
2	28	1	33	3	0	2	4449.462
3	33	1	22.705	0	0	1	21984.47061
4	32	1	28.88	0	0	1	3866.85520

Tabel diatas menunjukan hasil dari tranformasi data kategorikal ke format numerikal. Contohnya seperti *male* dan *female* setelah di transformasi menjadi 0 untuk *female* dan 1 untuk *male*, begitu pula untuk smoker dan region di tranformasi menjadi bentuk angka. Setelah itu dilakukan pembagian dataset, dengan rasio 75:25, yaitu 75% untuk data *training* dan 25% untuk data *testing*.

4. Evaluasi Model

Penelitian ini menggunakan model Random Forest dan XGBoost. Pada proses ini pipeline dibuat untuk menyatukan standarisasi dan *modeling*. Standarisasi fitur dilakukan menggunakan standardscaler**,** setelah melakukan standarisasi dilakukan pelatihan model untuk setiap pipeline menggunakan data pelatihan (X_train, y_train).

Cross-validation membantu mengetahui seberapa baik model akan bekerja pada data yang belum pernah dikerjakan sebelumnya**.** Fungsi cross_val_score dari sklearn.model_selection dilakukan untuk mengevaluasi model dengan teknik k-fold cross-validation. Kedua model dievaluasi menggunakan 10-fold cross-validation dengan metrik *Root Mean Squared Error* (RMSE) pada data latih (X_train, y_train).

Table 4. Hasil nilai RMSE

	RMSE
Random Forest	-4721.599240
XGBoost	-5084.077836

Nilai RMSE yang dihasilkan mencerminkan seberapa jauh prediksi model menyimpang dari nilai sebenarnya. Dapat dilihat bahwa Random Forest memiliki performa yang lebih baik dibandingkan XGBoost pada penelitian ini, karena menghasilkan RMSE yang lebih rendah yang artinya prediksi model lebih dekat ke nilai aktual. Evaluasi random forest dan XGBoost pada data uji (X_test) dengan metrik penilaian menggunakan R2, MAE, MSE, RMSE menghasilkan nilai sebagai berikut :

DOI : https://doi.org/10.26623/transformatika.v23i1.11914

Table 5. Perbandingan Random Forest dan XGBoost

	MAE	MSE	RMSE	R2
Random Forest	2573.0668612189	24199792.432581577	4919.328453415	0.8403996629767
XGBoost	3082.5743728189	30852273.00583208	5554.482244622	0.7965219427074

Berdasarkan evaluasi yang ditunjukan tabel diatas menggunakan metrik regresi (MAE, MSE, RMSE, dan R^2), model Random Forest memberikan performa lebih baik dibandingkan Random Forest, dengan MAE atau rata-rata kesalahan antara nilai prediksi dengan nilai asli sebesar 2.573,07 tergolong kecil (20% dari nilai rata-rata), R^2 sebesar 84.04%, nilai MSE 24.199.792,43 dan RMSE 4.919, 33. Hal ini menunjukkan model Random Forest mampu memprediksi premi asuransi lebih baik dengan rata-rata kesalahan yang relatif kecil.

CONCLUSIONS AND RECOMMENDATIONS

Setelah dilakukan pengujian model Random Forest dan XGBoost pada dataset premi asuransi kesehatan, telah didapatkan hasil perbandingan performa kedua model. Kedua model menunjukan hasil yang cukup baik dengan nilai error lebih kecil pada model Random Forest. Evaluasi model dilakukan dengan melihat hasil metrik regresi seperti MAE, MSE, RMSE dan R2. Pada XGBoost mungkin perlu dilakukan tuning hyperparameter lebih lanjut seperti learning_rate, max_depth, n_estimators untuk menemukan kombinasi faktor terbaik agar model dapat bekerja lebih optimal.

REFERENCES

[1] M. Siti Rifani Wulandari, B. Hilaliyah, F. Dwi Elnora, I. Nurbidari, and M. Simanjuntak, "Edukasi Alokasi Keuangan untuk Asuransi Kesehatan pada Mahasiswa (Education of Financial Allocation for Health Insurance of College Students)," *Agrokreatif Maret 2023*, vol. 9, no. 1.

[2] W. Mekniran, J.-N. Kramer, and T. Kowatsch, "Reimagining Preventive Care and Digital Health: A Paradigm Shift in a Health Insurance's Role," in *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies*, SCITEPRESS - Science and Technology Publications, 2024, pp. 852–858. doi: 10.5220/0012400300003657.

[3] M. de Silva, "Editorial," *Proceedings of the Institution of Civil Engineers - Engineering Sustainability*, vol. 175, no. 2, pp. 55–56, Apr. 2022, doi: 10.1680/jensu.2022.175.2.55.

[4] A. Putri, "Estimation Model of Pure Health Insurance Premiums in Southeast America Using Generalized Linear Model (GLM) with Gamma Distribution," *International Journal of Mathematics, Statistics, and Computing*, vol. 3, no. 1, pp. 27–33, Feb. 2025, doi: 10.46336/ijmsc.v3i1.181.

[5] S. Dash, B. Sankar panigrahi, V. V. Bhaskar Reddy Sanikommu, B. K. Madhavi, and S. K. Sahoo, "A Comparative Analysis of Different Machine Learning Techniques For Medical Insurance Premium Prediction," in *2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU)*, IEEE, Mar. 2024, pp. 1–6. doi: 10.1109/IC-CGU58078.2024.10530731.

[6] M. Kapse, V. Sharma, R. Vidhale, and V. Vellanki, "Customization of health insurance premiums using machine learning and explainable AI," *International Journal of Information Management Data Insights*, p., 2025, doi: 10.1016/j.jjimei.2025.100328.

[7] M. Bader and M. Maalouf, "Evaluating Determinants of Health Insurance Premiums Using Advanced Multiple Linear Regression Techniques," in *2024 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, IEEE, Dec. 2024, pp. 440–444. doi: 10.1109/IEEM62345.2024.10857131.

[8] K. Kaushik, A. Bhardwaj, A. D. Dwivedi, and R. Singh, "Machine Learning-Based Regression Framework to Predict Health Insurance Premiums," *Int J Environ Res Public Health*, vol. 19, no. 13, p. 7898, Jun. 2022, doi: 10.3390/ijerph19137898.

[9] S. Chen, "Health Insurance Annual Premium Forecast Analysis," *Highlights in Business, Economics and Management*, vol. 28, pp. 343–349, Apr. 2024, doi: 10.54097/j1xsa139.
[10] Prof. M. S. Patil, Kulkarni Sanika, and Khurpe Sanjana, "MEDICAL INSURANCE PREMIUM PREDICTION WITH MACHINE LEARNING," *International Journal of Innovations in Engineering Research and Technology*, vol. 11, no. 5, pp. 5–11, May 2024, doi: 10.26662/ijiert.v11i5.pp5-11.
[11] S. Maesaroh *et al.*, *Pembelajaran Mesin dan Kecerdasan Buatan: Teori dan Aplikasi Praktis*. 2024.
[12] Intan Permata and Esther Sorta Mauli Nababan, "Application Of Game Theory In Determining Optimum Marketing Strategy In Marketplace," *JURNAL RISET RUMPUN MATEMATIKA DAN ILMU PENGETAHUAN ALAM*, vol. 2, no. 2, pp. 65–71, Jul. 2023, doi: 10.55606/jurrimipa.v2i2.1336.
[13] D. Li, H. Yu, and X. Zhu, "Research on UBI Claims Risk Probability Models Based on Machine Learning," in *2023 4th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, IEEE, Aug. 2023, pp. 354–358. doi: 10.1109/ICBASE59196.2023.10303111.
[14] D. I. Sumantiawan, "METODE ANALASIS MENGGUNAKAN ALGORITMA RANDOM FOREST UNTUK PREDIKSI BIAYA ASURANSI KESEHATAN," Februari. [Online]. Available: https://raw.githubusercontent.com/stedy/Machine-Learning-with-R-datasets/master/insurance.csv
[15] U. Orji and E. Ukwandu, "Machine learning for an explainable cost prediction of medical insurance," *Machine Learning with Applications*, vol. 15, p. 100516, Mar. 2024, doi: 10.1016/j.mlwa.2023.100516.