e-ISSN : 2621 – 5934
P-ISSN : 2621 – 7112

Analisis Prediksi Biaya Asuransi Kesehatan Menggunakan Metode Machine Learning

Rizki Hesananda[11], Azka Fasya[2], Muhammad Zulfariansyah[3]

[1]Jurusan Teknologi Informasi, Universitas Siber Indonesia, Jakarta
[2]Jurusan Teknik Informatika, Fakultas Teknik, Universitas Krisnadwipayana, Jakarta
[3]Jurusan Sistem Informasi, Universitas Mulawarman, Samarinda

Abstrak. Penelitian ini bertujuan untuk memprediksi biaya asuransi kesehatan berdasarkan data demografis dan kebiasaan individu menggunakan berbagai metode machine learning. Dataset yang digunakan berasal dari Kaggle yang berisi 1338 baris dan 7 kolom, yaitu usia (*age*), jenis kelamin (*sex*), indeks massa tubuh (*bmi*), jumlah anak (*children*), status perokok (smoker), wilayah (region), dan biaya asuransi (charges). Model Linear Regression, Ridge Regression, Random Forest, dan Support Vector Regression (SVR) diterapkan dan dievaluasi. Hasil menunjukkan bahwa model Random Forest memberikan performa terbaik dengan nilai R^2 sebesar 0.8667. Penelitian ini memberikan wawasan penting dalam pemodelan biaya asuransi kesehatan untuk meningkatkan akurasi prediksi dan pengambilan keputusan.

Kata kunci—*Machine Learning; Asuransi, Regresi; CRISP-DM.*

1. PENDAHULUAN

Prediksi biaya asuransi kesehatan merupakan aspek penting dalam industri asuransi karena berpengaruh langsung pada penetapan premi dan pengelolaan risiko. Akurasi prediksi biaya ini membantu perusahaan asuransi dalam mengoptimalkan keuntungan sekaligus memberikan premi yang adil kepada konsumen. Selain itu, prediksi yang tepat juga membantu individu dalam merencanakan keuangan mereka terkait kebutuhan kesehatan di masa depan [1].

Namun, prediksi biaya asuransi kesehatan tidaklah mudah karena melibatkan berbagai variabel yang saling berinteraksi, seperti usia, jenis kelamin, indeks massa tubuh, jumlah anak, kebiasaan merokok, dan wilayah tempat tinggal. Kompleksitas ini menimbulkan tantangan dalam menemukan model yang mampu menangkap pola-pola tersembunyi dan hubungan non-linier antar variabel [2], [3].

Dalam konteks ini, machine learning menawarkan solusi yang potensial dengan kemampuannya dalam menangani data besar dan kompleks serta menemukan pola yang tidak mudah terlihat oleh metode statistik konvensional [4], [5]. Berbagai algoritma machine learning dapat diaplikasikan untuk memodelkan biaya asuransi dengan pendekatan yang berbeda, mulai dari model regresi linier sederhana hingga model ensemble dan kernel-based [6].

Setiap algoritma memiliki keunggulan tersendiri; Linear Regression mudah diinterpretasi, Ridge Regression membantu mengurangi overfitting melalui regularisasi, Random Forest mampu menangani data non-linier dan interaksi fitur dengan baik, sementara Support Vector Regression (SVR) efektif dalam menangani data dengan distribusi yang kompleks. Oleh karena itu, penting untuk membandingkan kinerja berbagai metode tersebut dalam konteks prediksi biaya asuransi kesehatan [7], [8].

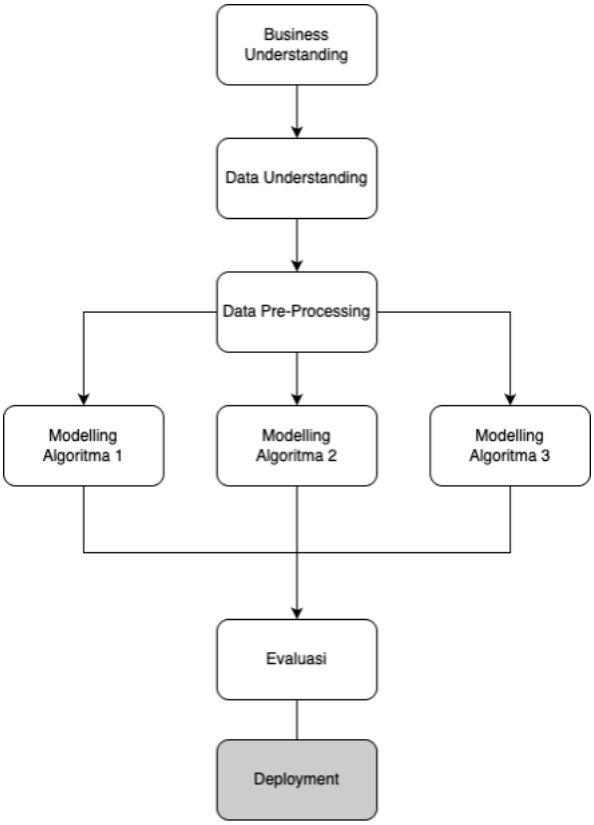
Penelitian ini bertujuan untuk mengembangkan dan membandingkan model prediksi biaya asuransi kesehatan menggunakan dataset dari Kaggle dengan menerapkan Linear Regression, Ridge Regression,

[1] Corresponding author: hessananda@cyber-univ.ac.id

Random Forest, dan SVR. Evaluasi model dilakukan untuk menentukan metode yang paling akurat dan dapat diandalkan dalam konteks ini [9], [10].

2. METODE

Penelitian ini menggunakan kerangka kerja CRISP-DM (*Cross-Industry Standard Process for Data Mining*) sebagai pendekatan metodologi [11], [12], [13]. Alur penelitian dapat dilihat pada gambar berikut:



Gambar 1. Alur penelitian CRISP-DM

CRISP-DM terdiri dari enam tahapan utama:

a. *Business Understanding*: Memahami tujuan bisnis dan permasalahan utama yang ingin diselesaikan, serta kebutuhan pemangku kepentingan.

b. *Data Understanding*: Mengeksplorasi data untuk memahami karakteristik, kualitas, dan pola dasar yang ada dalam dataset.

c. *Data Preparation*: Melakukan pra-pemrosesan data seperti encoding, scaling, dan pembagian data agar siap digunakan oleh model machine learning.

d. *Modelling*: Menerapkan dan membandingkan beberapa algoritma machine learning untuk membangun model prediksi.

e. *Evaluation*: Mengevaluasi performa model menggunakan metrik yang sesuai untuk memastikan akurasi dan generalisasi.

f. *Deployment*: Menyiapkan model untuk penerapan nyata dan memberikan wawasan bagi pengambilan keputusan bisnis.

Setiap tahapan tersebut saling terkait dan dapat dilakukan secara iteratif untuk memperoleh hasil yang optimal dalam proses data mining [14], [15].

3. HASIL DAN PEMBAHASAN

a. Business Understanding

Prediksi biaya asuransi kesehatan merupakan hal penting dalam industri asuransi untuk mendukung penetapan premi yang tepat dan pengelolaan risiko. Dengan prediksi yang akurat, perusahaan asuransi dapat meningkatkan efisiensi operasional dan memberikan penawaran premi yang adil bagi konsumen.

Pada konteks bisnis, variabel target adalah *charges* (biaya asuransi individu) dan variabel penentu meliputi usia, indeks massa tubuh (BMI), jumlah anak tanggungan, kebiasaan merokok, jenis kelamin, dan wilayah domisili. Stakeholder utama adalah aktuaria dan tim pricing yang membutuhkan estimasi akurat untuk menetapkan premi yang adil namun berkelanjutan, serta manajemen risiko yang harus menjaga kestabilan *loss ratio* portofolio. Dengan asumsi kebijakan underwriting konstan, model prediktif digunakan untuk memperkirakan biaya yang diharapkan per profil pelanggan.

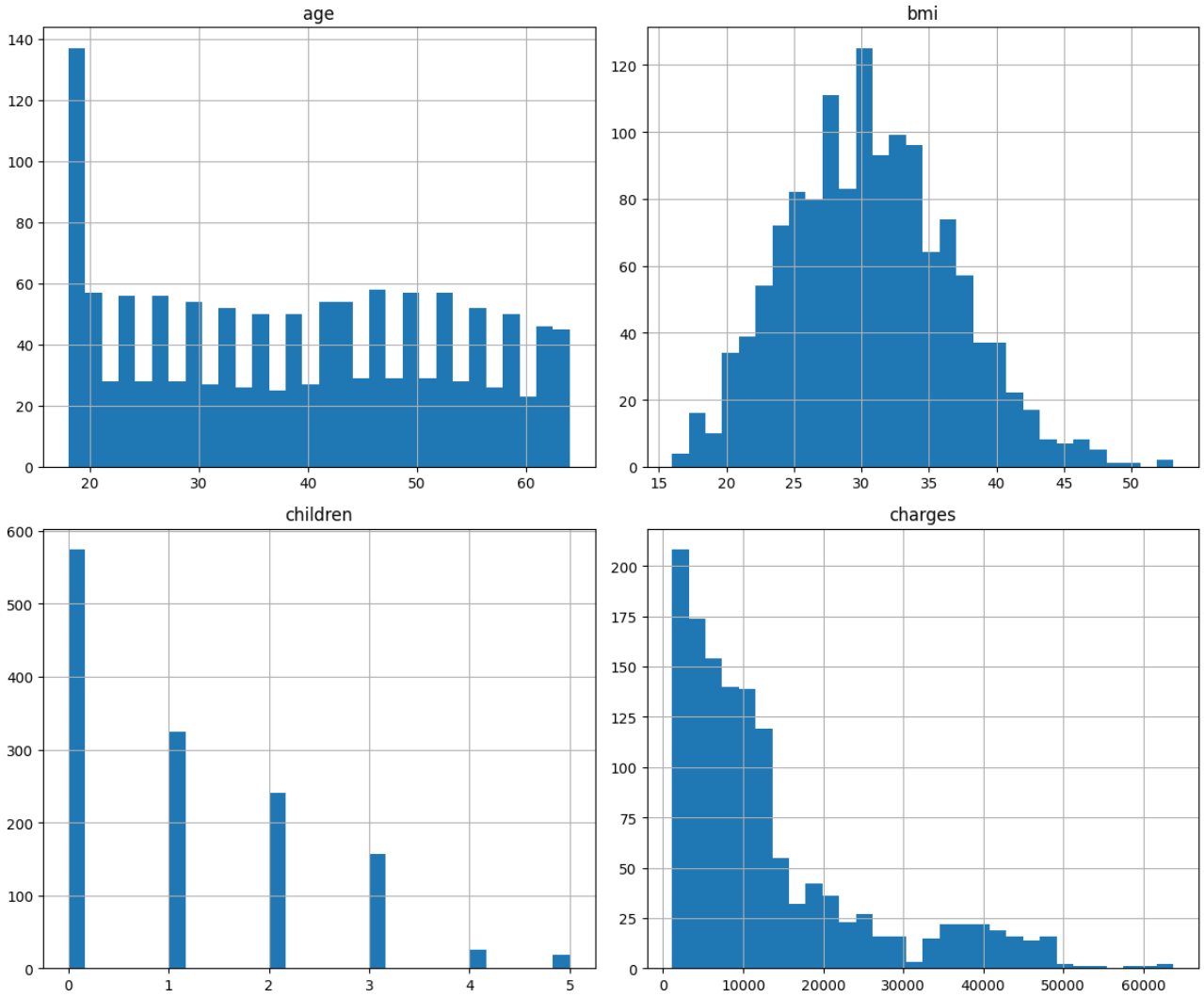
Objektif pemodelan diformalkan sebagai minimisasi kesalahan prediksi pada data yang tidak terlihat, dengan batasan: interpretabilitas memadai untuk audit, tidak ada *data leakage*, dan waktu inferensi wajar. Kriteria sukses praktis ditetapkan sebelum eksperimen: $R^2 \geq 0{,}80$ agar variasi yang dijelaskan oleh model cukup tinggi, serta MAE ≤ 3.000 agar galat absolut rata-rata tetap dalam kisaran yang operasional. Ambang ini diambil agar output model relevan bagi keputusan pricing pada level individu.

Dari sudut pandang dampak bisnis, peningkatan akurasi tidak hanya menekan risiko mispricing, tetapi juga mengurangi volatilitas cadangan klaim dan meningkatkan kepuasan pelanggan. Selain itu, pemahaman faktor pendorong biaya misalnya perokok, usia, dan BMI—mendukung kebijakan promosi kesehatan yang berbasis data. Hal-hal terkait etika (misalnya keadilan antarkelompok) dicatat sebagai perhatian lanjutan ketika model digunakan di luar lingkungan eksperimen.

b. Data Understanding

Dataset yang digunakan terdiri dari 1338 baris data dengan 7 fitur utama. Analisis distribusi fitur dapat dilihat pada Gambar 1, sedangkan hubungan antar fitur divisualisasikan menggunakan heatmap pada Gambar 2. Informasi ini membantu memahami karakteristik data dan hubungan antar variabel yang akan digunakan dalam pemodelan.

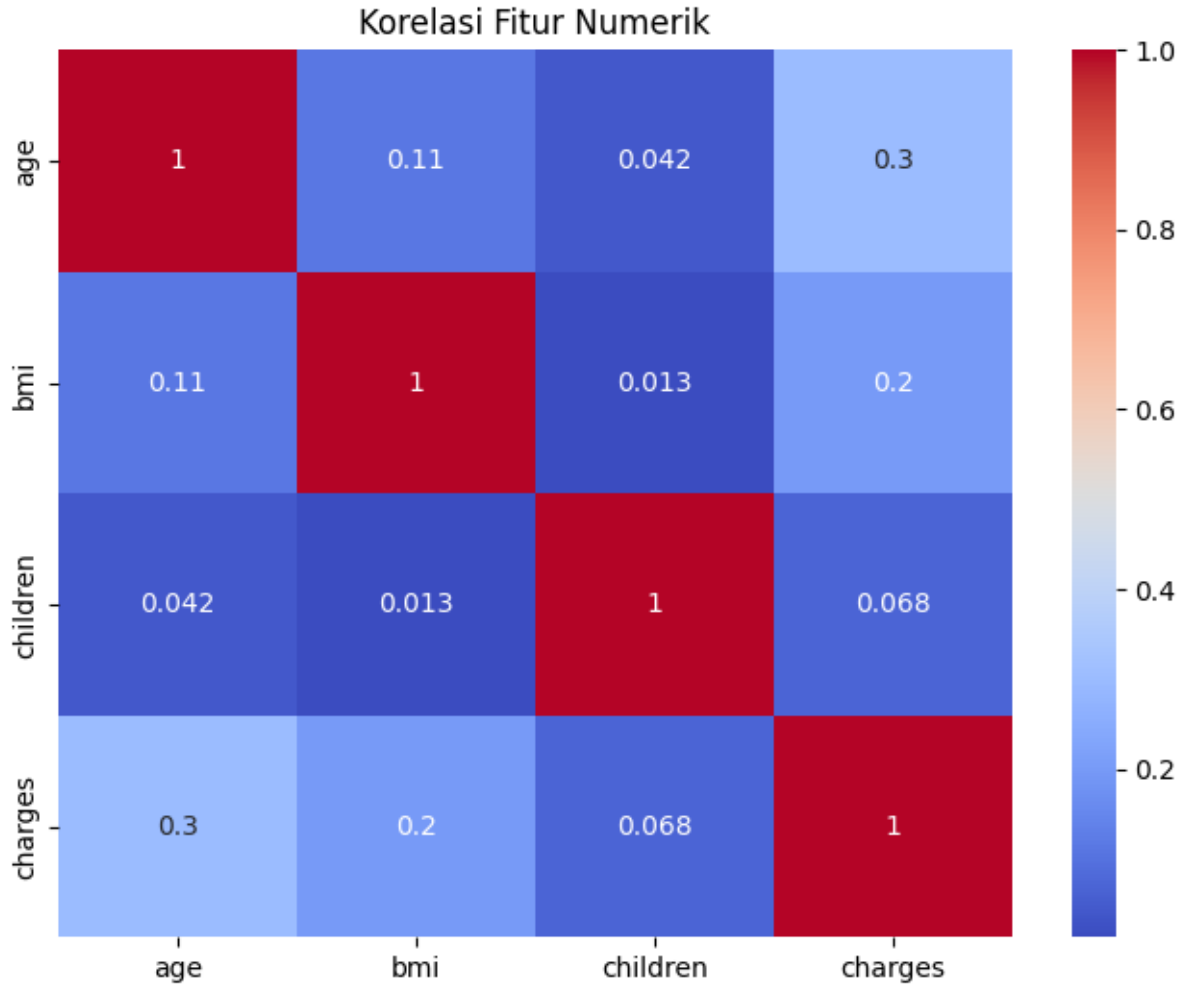
e-ISSN : 2621 – 5934
P-ISSN : 2621 – 7112



Gambar 2. Distribusi Fitur (Histogram)

Gambar 2 memperlihatkan bentuk distribusi fitur inti. Secara umum, *age* menyebar pada rentang dewasa produktif, *bmi* cenderung unimodal dengan ekor kanan yang menandai individu dengan BMI tinggi, sementara *charges* biasanya bersifat skewed (miring ke kanan) karena sebagian kecil individu menanggung biaya jauh di atas median. Temuan ini mengindikasikan bahwa transformasi non-linier atau model non-linier berpotensi lebih sesuai dibandingkan asumsi garis lurus murni.

e-ISSN : 2621 – 5934
P-ISSN : 2621 – 7112

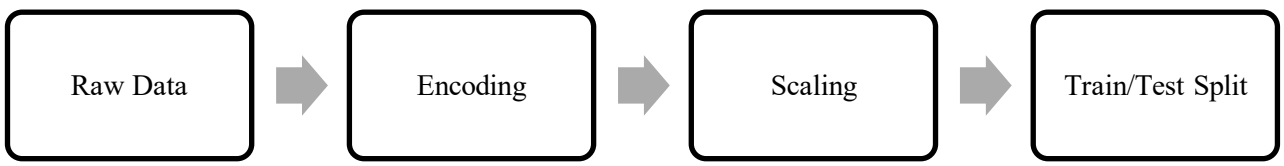


Gambar 3. Korelasi Fitur (Heatmap)

Gambar 3 menegaskan hubungan kuat antara *charges* dan indikator perilaku-kesehatan seperti *smoker*, serta korelasi positif dengan *age* dan *bmi*. Fitur seperti *children* cenderung menunjukkan korelasi moderat, sedangkan *region* dan *sex* berkontribusi lebih kecil secara linear. Pola ini menyiratkan kemungkinan adanya interaksi, misalnya dampak *smoker* yang meningkat pada usia tertentu, yang sulit ditangkap oleh model linear sederhana.

c. Data Preparation

Data diproses dengan melakukan encoding terhadap variabel kategorikal, scaling fitur numerik, serta pembagian data menjadi set pelatihan dan pengujian. Tahap ini memastikan data siap digunakan oleh model machine learning.



Gambar 4. *Pipeline* Pra-Pemrosesan Data

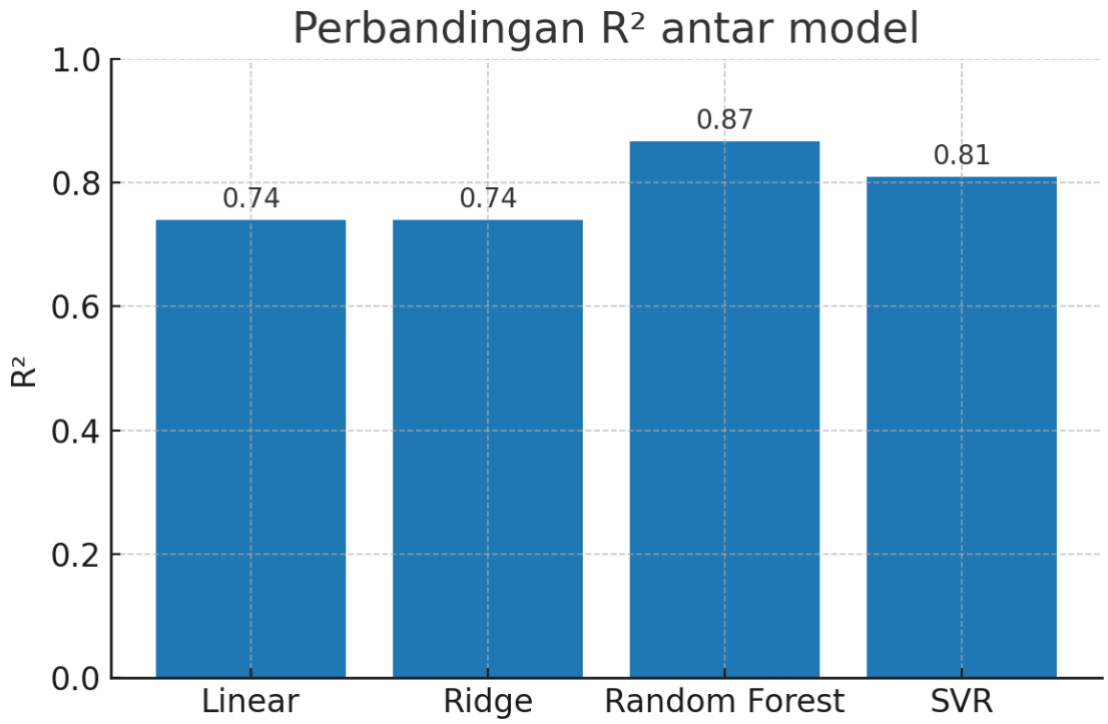
Diagram pada Gambar 3 (pipeline pra-pemrosesan) merangkum tiga langkah kunci. Pertama, *encoding* dilakukan untuk mengubah fitur kategorikal (*sex*, *smoker*, *region*) menjadi representasi numerik tanpa memaksakan urutan semu; skema *one-hot* dipilih agar tidak menyisipkan informasi ordinal yang keliru. Kedua, *scaling* diterapkan pada fitur numerik (*age*, *bmi*, *children*) untuk menormalkan skala sehingga algoritma sensitif skala tidak bias oleh perbedaan satuan. Ketiga, *train/test split* memastikan evaluasi pada data yang benar-benar tak terlihat.

Seluruh tahapan dibungkus dalam *pipeline* untuk menghindari *data leakage*, parameter transformasi misalnya rata-rata atau varian pada *scaler* dipelajari hanya dari *training set*, lalu diaplikasikan ke *test set*. Praktik ini menjaga objektivitas pengukuran performa. Selain itu, *random_state* dikunci untuk reproduktibilitas dan proporsi *split* dijaga agar cukup data tersedia bagi pelatihan sekaligus menyisakan sampel uji yang representatif.

Pemeriksaan kualitas data (missing values, nilai ekstrem) dilakukan sebelum pemodelan. Pada dataset ini tidak ditemukan *missing values* yang mengharuskan imputasi, namun ekor kanan pada *charges* dan *bmi* tetap diwaspadai. Upaya sederhana seperti transformasi atau *robust scaling* dapat dipertimbangkan di studi lanjutan untuk melihat dampaknya terhadap stabilitas galat.

d. Modelling

Beberapa model machine learning diuji untuk memprediksi biaya asuransi, yaitu Linear Regression, Ridge Regression, Random Forest, dan Support Vector Regression (SVR). Model-model ini dipilih untuk mengeksplorasi berbagai pendekatan dalam menangani data.



Gambar 5. Perbandingan R^2 Antar Model

Gambar 5 memperlihatkan perbandingan R^2 antar model. Dua model linear (Linear dan Ridge) berkisar di ~0,74 sehingga mampu menjelaskan sekitar 74% variasi *charges*—cukup baik, tetapi belum menangkap non-linearitas dan interaksi. Ridge membantu menstabilkan koefisien dibanding Linear murni, namun tidak banyak mengubah daya jelaskan karena struktur hubungan memang tidak sepenuhnya linear.

SVR menunjukkan R^2 ~0,81 setelah *scaling*, menandakan adanya peningkatan dibanding model linear. Meski demikian, kinerja SVR sangat sensitif terhadap pemilihan kernel dan hiperparameter (C, gamma, epsilon). Tanpa *tuning* yang agresif, SVR kerap kalah stabil di bawah variasi distribusi fitur dan *outliers* biaya.

Random Forest tampil terbaik (R^2 = 0,8667) karena kemampuannya memodelkan interaksi dan bentuk hubungan yang tidak linier, sekaligus relatif tahan terhadap *outliers*. Selain akurasi, model ini juga menyediakan *feature importance* yang bermanfaat untuk interpretasi tingkat tinggi, sebagaimana dibahas pada bagian *Deployment* (analitik pasca-model) meski penerapan produksi tidak dilakukan pada studi ini.

e. Evaluation

Performansi masing-masing model diukur menggunakan metrik Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), dan koefisien determinasi (R^2). Hasil evaluasi dapat dilihat pada Tabel 1. Secara umum, model Random Forest menunjukkan performa terbaik dengan MAE sebesar 2521.98, RMSE sebesar 4548.38, dan R^2 sebesar 0.8667, mengungguli model Linear Regression, Ridge Regression, dan SVR.

Tabel 1. Perbandingan Kinerja Model

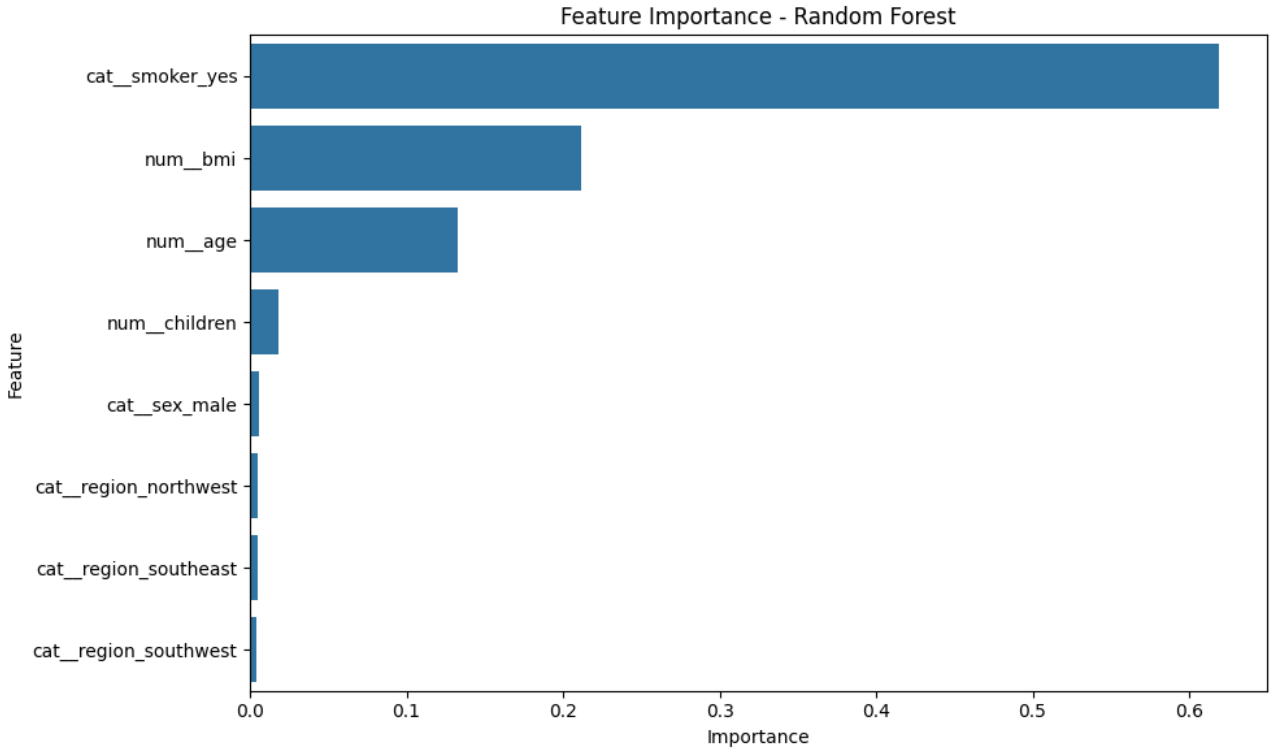
Model	MAE	RMSE
Linear Regression	3449.75	5937.03
Ridge Regression	3451.89	5927.88
Random Forest	2521.98	4548.38
Support Vector Regr.	2925.09	5173.45

Secara kuantitatif, Random Forest menurunkan RMSE dari 5.937,03 (Linear) menjadi 4.548,38, atau perbaikan relatif sekitar 23%. MAE turun dari 3.449,75 (Linear) menjadi 2.521,98—penurunan sekitar 27%. Kenaikan R^2 dari ~0,74 ke 0,8667 menunjukkan tambahan variasi *charges* yang terjelaskan oleh model non-linier. Dibanding SVR (RMSE 5.173,45; R^2 0,81), Random Forest masih unggul dengan selisih RMSE ~12% lebih rendah.

Interpretasi metrik: MAE menyatakan galat rata-rata yang diharapkan pada prediksi biaya individu; RMSE memberi penalti lebih besar pada galat besar sehingga sensitif pada ekor distribusi *charges*; R^2 mengukur proporsi variasi yang dijelaskan model. Kombinasi ketiganya memberikan pandangan seimbang antara akurasi rata-rata dan kinerja pada kasus mahal yang jarang namun penting secara finansial.

Validasi tambahan seperti *k*-fold cross-validation atau *bootstrapping* direkomendasikan untuk mengukur kestabilan performa dan menyusun *confidence interval* metrik. Analisis residual juga dapat menilai bias sistematis, misalnya apakah model cenderung *underpredict* pada kelompok perokok dengan BMI tinggisebagai dasar iterasi model berikutnya.

f. Feature Important



Gambar 6. *Feature Importance Model Random Forest*

Gambar 6 mengindikasikan bahwa *smoker*, *age*, dan *bmi* merupakan kontributor dominan terhadap variasi *charges*. Hal ini konsisten secara domain: kebiasaan merokok meningkatkan risiko kesehatan, usia lebih tua berkorelasi dengan utilisasi layanan, dan BMI tinggi terkait komorbiditas. Informasi ini berguna sebagai masukan kebijakan pencegahan (program berhenti merokok, edukasi gizi) maupun *risk adjustment* pada pricing.

KESIMPULAN

Penelitian ini membandingkan beberapa metode machine learning untuk memprediksi biaya asuransi kesehatan menggunakan dataset dari Kaggle. Model Random Forest terbukti menjadi yang terbaik dengan nilai R^2 sebesar 0.8667, MAE sebesar 2521.98, dan RMSE sebesar 4548.38, menunjukkan kemampuannya dalam menangkap pola kompleks dalam data. Linear Regression, Ridge Regression, dan SVR memberikan hasil yang lebih rendah dengan R^2 masing-masing sekitar 0.74 dan 0.81. Temuan ini memberikan dasar yang kuat untuk pengembangan sistem prediksi biaya asuransi yang lebih akurat dan efisien, yang dapat membantu perusahaan asuransi dan konsumen dalam pengambilan keputusan yang lebih baik.

DAFTAR PUSTAKA

[1] E. E. Silalahi, "Budaya organisasi perusahaan asuransi jiwa di Indonesia," *MIX: Jurnal Ilmiah Manajemen*, vol. VII, no. 1, pp. 48–58, 2017.

[2] P. DI Jumlah Produksi Padi Kabupaten Padang Lawas, S. Hidayat Nasution, N. Irsa Syahputri, and R. Aprilia, "Penerapan metode least square dalam prediksi jumlah produksi padi di Kabupaten Padang Lawas," *journal.ummat.ac.id*, vol. 7, no. 2, pp. 128–137, 2024, doi: 10.31764/justek.vXiY.ZZZ.

[3] U. Indonesia, "Prediksi preferensi pelanggan waralaba makanan cepat saji dengan menggunakan pendekatan data mining skripsi," 2011.

[4] Lindawati, "Data Mining Dengan Teknik Clustering Dalam Pengklasifikasian Data Mahasiswa Studi Kasus Prediksi Lama Studi Mahasiswa Universitas Bina Nusantara," *Universitas Stuttgart*, pp. 174–180, 2008.

[5] D. S. Saroso, H. Rudystira, and R. Hesananda, *Manajemen Perkreditan: Berbasis Pola Data Digital Anomali Aktivitas*. Penerbit NEM, 2024.

[6] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, “Statistical and Machine Learning forecasting methods: Concerns and ways forward,” *PLoS One*, vol. 13, no. 3, p. e0194889, 2018.

[7] R. Yanto, “Implementasi Data Mining Estimasi Ketersediaan Lahan Pembuangan Sampah menggunakan Algoritma Simple Linear Regression,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 2, no. 1, pp. 361–366, 2018, doi: 10.29207/resti.v2i1.282.

[8] R. Widyastuti and H. Basri, “Klasifikasi Ciri Fitur Kayu Merbau dan Kayu Jati menggunakan Support Vector Machine (SVM) Feature Classification of Merbau and Teak Wood using Support Vector Machine,” vol. 1, no. 2, pp. 66–71, 2020.

[9] M. Muchtar and R. A. Muchtar, “PERBANDINGAN METODE KNN DAN SVM DALAM KLASIFIKASI KEMATANGAN BUAH MANGGA BERDASARKAN CITRA HSV DAN FITUR STATISTIK,” *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 2, Apr. 2024, doi: 10.23960/jitet.v12i2.4010.

[10] A. Lestari, “PERBANDINGAN METODE REGRESI BIASA DENGAN GEOGRAPHICALLY WEIGHTED REGRESSION DALAM MEMODELKAN DATA COLUMBUS DI SOFTWARE R 2.6.1,” *Al-Khwarizmi*, vol. I, 2013.

[11] J. S. Saltz, “CRISP-DM for data science: strengths, weaknesses and potential next steps,” in *2021 IEEE International Conference on Big Data (Big Data)*, IEEE, 2021, pp. 2337–2344.

[12] S. Peker and Ö. Kart, “Transactional data-based customer segmentation applying CRISP-DM methodology: A systematic review,” *Journal of Data, Information and Management*, vol. 5, no. 1, pp. 1–21, 2023.

[13] U. Shafique and H. Qaiser, “A comparative study of data mining process models (KDD, CRISP-DM and SEMMA),” *International Journal of Innovation and Scientific Research*, vol. 12, no. 1, pp. 217–222, 2014.

[14] V. Plotnikova, M. Dumas, and F. P. Milani, “Applying the CRISP-DM data mining process in the financial services industry: Elicitation of adaptation requirements,” *Data Knowl Eng*, vol. 139, p. 102013, 2022.

[15] R. Wirth and J. Hipp, “CRISP-DM: Towards a standard process model for data mining,” in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, Springer-Verlag London, UK, 2000.