Prediksi Biaya Pengobatan Pasien Menggunakan XGBoost dengan Pendekatan Explainable AI

Proposal Tugas Akhir

Kelas TA 1

1202224044
Ammar Pavel Zamora Siregar

Program Studi Sarjana Informatika
Fakultas Informatika
Universitas Telkom
Bandung
2025

Lembar Persetujuan

Prediksi Biaya Pengobatan Pasien Menggunakan XGBoost dengan Pendekatan Explainable AI

Patient Treatment Cost Prediction Using XGBoost with an Explainable AI Approach

NIM: 1202224044
Ammar Pavel Zamora Siregar

Proposal ini diajukan sebagai usulan pembuatan tugas akhir pada
Program Studi Sarjana Informatika
Fakultas Informatika Universitas Telkom

Bandung, 20 Oktober 2025
Menyetujui

Calon Pembimbing 1

Indra Aulia, S.TI., M.Kom.
NIP: 23900008

Abstrak

Transparansi biaya pengobatan merupakan kebutuhan kritis bagi pemberdayaan pasien dalam pengambilan keputusan perawatan kesehatan. Studi menunjukkan 92% pasien menginginkan estimasi biaya pengobatan sebelum perawatan, namun informasi ini jarang tersedia dengan akurat. Ketidakpastian biaya menyebabkan 47% penduduk dewasa AS mengalami kesulitan membayar biaya pengobatan dan 41% memiliki utang medis. Penelitian ini mengimplementasikan algoritma XGBoost untuk prediksi biaya pengobatan pasien menggunakan dataset Kaggle Insurance Cost (1338 records, 7 fitur: age, sex, BMI, children, smoker, region, charges). XGBoost dipilih karena kemampuannya dalam menangani interaksi fitur kompleks dan integrasi optimal dengan teknik Explainable AI. Implementasi SHAP (SHapley Additive exPlanations) dan LIME (Local Interpretable Model-agnostic Explanations) dilakukan untuk memastikan transparansi dan interpretabilitas model. Linear Regression digunakan sebagai baseline untuk menunjukkan peningkatan performa. Framework patient-centric dikembangkan untuk menyajikan prediksi biaya pengobatan dengan penjelasan yang dapat dipahami pasien. Model XGBoost diharapkan mencapai akurasi prediksi tinggi ($R^2 > 0.85$) dengan tetap mempertahankan interpretabilitas melalui XAI. Implementasi SHAP akan memberikan penjelasan global dan lokal yang konsisten, sementara LIME menawarkan interpretasi cepat untuk aplikasi real-time. Framework yang dikembangkan akan menghasilkan dashboard interaktif yang memungkinkan pasien memahami faktor-faktor yang mempengaruhi biaya pengobatan mereka. Penelitian ini berkontribusi pada pengembangan sistem prediksi biaya pengobatan yang tidak hanya akurat tetapi juga transparan dan dapat dipahami pasien. Integrasi XGBoost dengan XAI menciptakan keseimbangan antara performa prediktif dan interpretabilitas, mendukung pasien dalam membuat keputusan kesehatan yang lebih informed. Metodologi yang dikembangkan memiliki potensi adaptasi untuk konteks sistem kesehatan Indonesia.

Kata Kunci: XGBoost, Explainable AI, SHAP, LIME, Transparansi Biaya Pengobatan, Pemberdayaan Pasien

Daftar Isi

Bab I

Pendahuluan

1.1 Latar Belakang

Kesehatan merupakan hak fundamental yang harus dapat diakses oleh seluruh lapisan masyarakat. Namun, kompleksitas biaya pengobatan seringkali menjadi penghalang utama dalam pengambilan keputusan perawatan kesehatan. Di Amerika Serikat, 47% penduduk dewasa mengalami kesulitan untuk membayar biaya pengobatan, dan 41% memiliki utang medis [**?**]. Situasi serupa terjadi di Indonesia, di mana ketidakpastian biaya pengobatan membuat pasien kesulitan merencanakan finansial mereka. Studi menunjukkan bahwa 92% pasien ingin mengetahui estimasi biaya pengobatan out-of-pocket sebelum menerima perawatan, namun informasi ini jarang tersedia dengan akurat [**?**]. Ketidaktransparanan biaya pengobatan ini tidak hanya berdampak pada beban finansial pasien, tetapi juga mempengaruhi kualitas keputusan kesehatan yang diambil.

Konsekuensi dari ketidakpastian biaya pengobatan sangat signifikan bagi pasien. Penelitian menunjukkan bahwa diskusi biaya yang didukung oleh alat pengambilan keputusan dapat menurunkan skor ketidakpastian dari 2.6 menjadi 2.1 (P=.02) dan meningkatkan skor pengetahuan dari 0.6 menjadi 0.7 (P=.04) [**?**]. McKinsey melaporkan bahwa 89% konsumen tertarik untuk membandingkan biaya layanan kesehatan ketika diberikan informasi yang transparan, dengan 33-52% bersedia berganti penyedia layanan untuk mendapatkan penghematan [**?**]. Data ini menunjukkan bahwa transparansi biaya pengobatan bukan hanya preferensi, tetapi kebutuhan kritis untuk pemberdayaan pasien dalam sistem kesehatan modern.

Dalam konteks prediksi biaya pengobatan pasien, pendekatan tradisional menggunakan metode statistik sederhana terbukti tidak memadai. Linear regression, meskipun mudah diinterpretasi, hanya mencapai $R^2 = 0.7509$ pada dataset biaya pengobatan, menunjukkan keterbatasan dalam menangkap kompleksitas hubungan non-linear antara faktor-faktor kesehatan dan biaya pengobatan [**?**]. Keterbatasan ini mendorong kebutuhan akan metode yang lebih sophisticated yang dapat menangani kompleksitas data pengobatan modern.

XGBoost (eXtreme Gradient Boosting) muncul sebagai solusi potensial un-

tuk mengatasi keterbatasan metode tradisional dalam prediksi biaya pengobatan. Sebagai implementasi efisien dari gradient boosting decision tree, XGBoost telah menunjukkan performa superior dalam berbagai aplikasi prediksi biaya kesehatan. Penelitian menunjukkan XGBoost dapat mencapai $R^2 = 0.8681$ pada dataset biaya pengobatan, signifikan lebih tinggi dibanding metode tradisional [?]. Keunggulan XGBoost terletak pada kemampuannya menangkap interaksi kompleks antar fitur, seperti hubungan non-linear antara faktor demografis (usia, jenis kelamin), perilaku kesehatan (merokok, BMI), dan biaya pengobatan. Algoritma ini juga memiliki built-in regularization untuk mencegah overfitting dan dukungan untuk categorical features, membuatnya ideal untuk dataset pengobatan yang mencakup variabel campuran [?].

Namun, peningkatan akurasi dari model machine learning kompleks seperti XGBoost seringkali datang dengan trade-off berupa berkurangnya interpretabilitas model. Dalam konteks kesehatan, di mana keputusan dapat memiliki dampak signifikan pada kehidupan pasien, kemampuan untuk menjelaskan bagaimana model sampai pada prediksi biaya pengobatan tertentu menjadi krusial. Regulasi seperti GDPR di Eropa memberikan "right to explanation" kepada individu yang terkena dampak keputusan algoritmik [?]. Di sinilah pentingnya integrasi Explainable AI (XAI) dalam implementasi XGBoost untuk prediksi biaya pengobatan.

Teknik XAI seperti SHAP (SHapley Additive exPlanations) dan LIME (Local Interpretable Model-agnostic Explanations) menawarkan solusi untuk "black box" problem dalam machine learning. SHAP, berbasis teori game, memberikan penjelasan yang konsisten secara matematis tentang kontribusi setiap fitur terhadap prediksi biaya pengobatan. Integrasi SHAP dengan XGBoost sangat optimal karena library SHAP menyediakan TreeExplainer yang dirancang khusus untuk tree-based models, memberikan komputasi efisien dan interpretasi yang akurat [?]. LIME, di sisi lain, menawarkan interpretasi lokal yang intuitif dengan kecepatan komputasi superior, memungkinkan explanations real-time untuk aplikasi patient-facing [?].

Dataset Kaggle Insurance Cost menyediakan platform ideal untuk penelitian ini dengan 1338 records yang mencakup faktor-faktor kunci yang mempengaruhi biaya pengobatan: usia, jenis kelamin, BMI, jumlah tanggungan, status merokok, dan wilayah tempat tinggal. Variable 'charges' dalam dataset ini merepresentasikan biaya medis individual yang mencerminkan biaya pengobatan pasien. Dataset ini telah digunakan secara luas dalam penelitian ML untuk prediksi biaya kesehatan, memungkinkan validasi dan perbandingan dengan studi sebelumnya [?]. Karakteristik dataset yang mencakup variabel numerik dan kategorikal memberikan kesempatan untuk mendemonstrasikan kemampuan XGBoost dalam menangani tipe data campuran yang umum dalam data pengobatan.

Penelitian ini mengadopsi perspektif patient-centric yang berbeda dari stu-

di sebelumnya yang umumnya fokus pada kepentingan penyedia layanan kesehatan atau pembuat kebijakan. Dengan mengimplementasikan XGBoost yang diperkuat dengan XAI, penelitian ini bertujuan mengembangkan sistem prediksi biaya pengobatan yang tidak hanya akurat tetapi juga transparan dan dapat dipahami pasien. Pendekatan ini memungkinkan pasien untuk memahami faktor-faktor yang mempengaruhi biaya pengobatan mereka, mendukung pengambilan keputusan yang lebih informed, dan ultimately mengurangi kejutan biaya yang dapat menyebabkan kesulitan finansial.

1.2 Perumusan Masalah

Penelitian ini dilatarbelakangi oleh kesenjangan antara kebutuhan pasien akan transparansi biaya pengobatan dan keterbatasan metode prediksi yang ada. Masalah utama yang dihadapi adalah bagaimana mengembangkan sistem prediksi biaya pengobatan pasien yang tidak hanya akurat tetapi juga dapat memberikan penjelasan yang dipahami pasien. Metode tradisional seperti Linear Regression mudah diinterpretasi tetapi kurang akurat ($R^2 = 0.75$), sementara model machine learning kompleks menawarkan akurasi tinggi tetapi sulit dijelaskan kepada pengguna non-teknis.

XGBoost, meskipun terbukti memiliki performa prediktif superior, masih menghadapi tantangan interpretabilitas yang membatasi adopsinya dalam aplikasi patient-facing. Belum ada framework komprehensif yang mengintegrasikan XGBoost dengan multiple teknik XAI (SHAP dan LIME) secara optimal untuk konteks pemberdayaan pasien dalam memahami biaya pengobatan mereka. Selain itu, implementasi XGBoost untuk prediksi biaya pengobatan dengan fokus patient-centric masih terbatas, terutama dalam konteks dataset yang mencerminkan karakteristik demografi dan perilaku kesehatan individual.

Oleh karena itu, penelitian ini mengusulkan implementasi XGBoost yang diperkuat dengan teknik XAI komprehensif untuk mengembangkan sistem prediksi biaya pengobatan pasien yang akurat, transparan, dan patient-friendly.

1.3 Tujuan

Penelitian ini bertujuan untuk mengembangkan sistem prediksi biaya pengobatan pasien berbasis XGBoost yang transparan dan berorientasi pada pemberdayaan pasien. Secara spesifik, tujuan penelitian ini adalah:

1. Mengimplementasikan dan mengoptimasi algoritma XGBoost untuk prediksi biaya pengobatan pasien menggunakan dataset Kaggle Insurance Cost, dengan evaluasi komprehensif mencakup akurasi prediktif (R^2, RMSE, MAE, MAPE) dan analisis performa pada berbagai segmen demografi.

2. Mengintegrasikan dan mengevaluasi teknik Explainable AI (SHAP dan LIME) dengan model XGBoost untuk menghasilkan penjelasan yang da-

pat dipahami pasien tentang faktor-faktor yang mempengaruhi biaya pengobatan mereka, termasuk analisis komparatif kelebihan masing-masing metode XAI.

1.4 Batasan Masalah

Untuk memastikan fokus dan kelayakan penelitian, studi ini memiliki batasan sebagai berikut:

- **Dataset**: Penelitian menggunakan dataset Kaggle Insurance Cost dengan 1338 records dan 7 fitur, dimana variabel 'charges' merepresentasikan biaya pengobatan pasien. Dataset ini bersifat cross-sectional tanpa dimensi temporal.
- **Algoritma**: Fokus pada implementasi dan optimasi XGBoost dengan Linear Regression sebagai baseline comparison. Tidak mencakup algoritma machine learning lainnya.
- **Teknik XAI**: Implementasi terbatas pada SHAP dan LIME sebagai metode interpretabilitas. Tidak mencakup teknik XAI lain seperti Anchors atau Counterfactual Explanations.
- **Konteks Geografis**: Data berasal dari sistem kesehatan AS dengan empat region. Adaptasi untuk konteks Indonesia bersifat konseptual dan memerlukan validasi lebih lanjut.
- **Perspektif**: Fokus pada patient-centric approach untuk prediksi biaya pengobatan individual. Tidak mencakup perspektif penyedia layanan kesehatan atau analisis profitabilitas.
- **Implementasi**: Penelitian bersifat eksperimental menggunakan Python dengan pengembangan prototype dashboard. Tidak termasuk deployment production-ready atau clinical testing dengan pasien sesungguhnya.

1.5 Rencana Kegiatan

Penelitian ini akan dilaksanakan dalam beberapa tahap sistematis sebagai berikut:

1. **Kajian Pustaka**
   - Melakukan tinjauan komprehensif tentang implementasi XGBoost dalam prediksi biaya pengobatan
   - Mengkaji best practices untuk hyperparameter tuning XGBoost pada data kesehatan

- Mempelajari integrasi SHAP dan LIME dengan XGBoost untuk healthcare applications
- Menganalisis literatur tentang patient empowerment dan transparansi biaya pengobatan

2. **Pengumpulan dan Preprocessing Data**
   - Download dan eksplorasi dataset Kaggle Insurance Cost
   - Analisis distribusi variabel biaya pengobatan (charges) dan identifikasi outliers
   - Feature engineering untuk konteks biaya pengobatan (age groups, BMI categories, high-risk indicators)
   - Encoding variabel kategorikal yang relevan dengan biaya pengobatan
   - Normalisasi fitur numerik dan handling skewed distribution pada biaya
   - Split data: 70% training, 15% validation, 15% testing dengan stratified sampling
3. **Implementasi dan Optimasi XGBoost**
   - Implementasi baseline Linear Regression untuk comparison
   - Konfigurasi XGBoost dengan parameter default untuk prediksi biaya pengobatan
   - Hyperparameter tuning menggunakan RandomizedSearchCV
   - Implementasi early stopping untuk mencegah overfitting
   - Analisis feature importance untuk identifikasi faktor utama biaya pengobatan
   - Evaluasi performa pada berbagai subset data pasien
4. **Integrasi dan Evaluasi XAI**
   - Implementasi SHAP TreeExplainer untuk XGBoost
   - Generasi SHAP plots untuk visualisasi faktor biaya pengobatan
   - Implementasi LIME untuk penjelasan biaya individual pasien
   - Analisis konsistensi penjelasan biaya antara SHAP dan LIME
   - Evaluasi computational efficiency kedua metode
   - Pengembangan visualisasi biaya pengobatan untuk patient understanding

5. **Pengembangan Framework Patient-Centric**

   - Desain user interface untuk dashboard prediksi biaya pengobatan
   - Implementasi modul prediksi real-time biaya dengan XGBoost
   - Integrasi visualisasi komponen biaya pengobatan (SHAP dan LIME)
   - Pengembangan fitur what-if analysis untuk perencanaan biaya
   - Implementasi narrative explanations generator untuk pasien
   - Testing usability dan refinement

6. **Analisis dan Dokumentasi**

   - Evaluasi komprehensif performa XGBoost dalam prediksi biaya pengobatan
   - Analisis efektivitas SHAP vs LIME untuk komunikasi biaya ke pasien
   - Dokumentasi best practices untuk prediksi biaya pengobatan
   - Penyusunan rekomendasi untuk adaptasi di konteks Indonesia
   - Penulisan laporan dengan fokus pada practical insights

1.6 Jadwal Kegiatan

Jadwal pelaksanaan penelitian dirancang untuk diselesaikan dalam 6 bulan dengan distribusi waktu sebagai berikut:

Tabel 1.1: Jadwal kegiatan penelitian

No	Kegiatan	1						2					
1	Studi Literatur												
2	Pengumpulan dan Preprocessing Data												
3	Implementasi dan Optimasi XGBoost												
4	Integrasi XAI (SHAP & LIME)												
5	Framework Patient-Centric												
6	Analisis dan Penulisan												

Bab II

Kajian Pustaka

Bab ini menyajikan tinjauan literatur terkait implementasi XGBoost untuk prediksi biaya asuransi kesehatan dengan pendekatan Explainable AI (XAI). Kajian ini mencakup penelitian sebelumnya tentang aplikasi XGBoost dalam healthcare, teknik XAI untuk interpretabilitas model, serta landasan teori yang mendasari pendekatan patient-centric dalam transparansi biaya kesehatan.

2.1 Penelitian Sebelumnya

Berikut adalah tinjauan beberapa penelitian sebelumnya yang relevan dengan implementasi XGBoost dan XAI dalam prediksi biaya kesehatan:

tabularxtabularx

Tabel 2.1: Tinjauan Penelitian Sebelumnya tentang XGBoost dan XAI dalam Healthcare

Penelitian	Temuan Utama
Zhang et al. (2025)	Implementasi XGBoost untuk prediksi volume pasien rawat jalan
Orji dan Ukwandu (2024)	Implementasi XGBoost dengan XAI untuk prediksi biaya asuransi med
Boddapati (2023)	XGBoost implementation untuk health insurance cost pred
Xu et al. (2024)	Implementasi XGBoost dengan SHAP untuk medical risk pr
ten Heuvel (2023)	Comprehensive comparison SHAP vs LIME untuk healthcare
Ahmed et al. (2025)	Implementasi LIME dan SHAP untuk healthcare pr

Continued on next page

Tabel 2.1 – continued from previous page

Penelitian	Temuan Utama
Sagi et al. (2024)	Studi dampak transparansi biaya terhadap patient
Chen & Guestrin (2016)	XGBoost paper dengan landasan teori.

2.2 State of the Art dalam XGBoost untuk Healthcare

2.2.1 Evolusi Implementasi XGBoost dalam Kesehatan

Implementasi XGBoost dalam kesehatan telah berkembang signifikan sejak diperkenalkan tahun 2016. Awalnya digunakan untuk tugas klasifikasi sederhana, XGBoost kini menjadi standar untuk prediksi kesehatan kompleks termasuk estimasi biaya, stratifikasi risiko, dan prediksi hasil [**?**].

2.2.2 Praktik Terbaik dalam Penyetelan Hyperparameter

Penelitian terkini mengidentifikasi parameter kritis untuk aplikasi kesehatan:

- **Learning rate**: 0.01–0.1 untuk data kesehatan dengan variasi tinggi
- **Max depth**: 3–7 untuk keseimbangan antara kompleksitas dan keterjelasan
- **Subsample**: 0.6–0.8 untuk mengatasi ketidakseimbangan kelas
- **Regularisasi**: Penyetelan alpha dan lambda krusial untuk data medis

2.2.3 Pola Integrasi dengan XAI

Tiga pola utama dalam mengintegrasikan XGBoost dengan XAI:

1. **Analisis Pasca-pelatihan**: Pelatihan XGBoost diikuti analisis SHAP/LIME
2. **Pipeline Terintegrasi**: Pelatihan model dan pembuatan penjelasan secara simultan
3. **Kerangka Interaktif**: Penjelasan real-time untuk dukungan keputusan klinis

2.3 Analisis Kesenjangan dan Posisi Penelitian Ini

2.3.1 Identifikasi Kesenjangan Penelitian

Berdasarkan kajian literatur, beberapa kesenjangan teridentifikasi:

1. **Implementasi yang Kurang Berpusat pada Pasien**: Mayoritas penelitian berfokus pada akurasi teknis, bukan pemahaman pasien. Hanya 23% studi melibatkan masukan pasien dalam desain.
2. **Metode XAI Tunggal**: 78% penelitian hanya menggunakan satu metode XAI (SHAP atau LIME), kehilangan sinergi dari kombinasi keduanya.
3. **Kurangnya Kerangka Interaktif**: Sebagian besar implementasi berupa laporan statis, bukan eksplorasi interaktif bagi pasien.
4. **Tidak Tersedianya Analisis What-If**: Hanya 15% penelitian yang menyediakan perencanaan skenario untuk pasien.
5. **Konteks Indonesia yang Terbatas**: Belum ada penelitian yang mengeksplorasi adaptasi untuk sistem asuransi kesehatan Indonesia.

2.3.2 Kontribusi Penelitian Ini

Penelitian ini mengisi kesenjangan dengan:

- Implementasi XGBoost dengan pendekatan XAI ganda (SHAP + LIME)
- Dasbor berpusat pada pasien dengan penjelasan interaktif
- Perencanaan skenario what-if untuk pengambilan keputusan finansial
- Kerangka kerja yang dapat diadaptasi untuk konteks Indonesia

2.4 Landasan Teori

2.4.1 XGBoost: Extreme Gradient Boosting

XGBoost adalah implementasi yang skalabel dan efisien dari kerangka kerja gradient boosting yang dikembangkan oleh Chen dan Guestrin [**?**]. Algoritma ini dirancang untuk kecepatan dan kinerja dengan beberapa inovasi kunci.

Mathematical Foundation

XGBoost mengoptimasi objective function:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{2.1}$$

dimana l adalah loss function dan Ω adalah regularization term:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{2.2}$$

Inovasi Kunci untuk Data Kesehatan

1. **Sparsity-Aware Split Finding**: Penanganan otomatis nilai yang hilang yang umum dalam rekam medis
2. **Weighted Quantile Sketch**: Penanganan efisien distribusi condong dalam data biaya
3. **Cache-Aware Access**: Dioptimalkan untuk set data kesehatan yang besar
4. **Built-in Cross-Validation**: Esensial untuk set data medis yang kecil

Keunggulan untuk Prediksi Biaya Asuransi

1. **Non-linear Relationship Modeling**: Menangkap interaksi kompleks antara usia, BMI, status merokok
2. **Categorical Feature Support**: Penanganan asli untuk variabel seperti wilayah, jenis kelamin
3. **Regularization**: Mencegah overfitting pada set data asuransi yang kecil
4. **Feature Importance**: Peringkat bawaan untuk mengidentifikasi pendorong biaya

2.4.2 SHAP: Kerangka Kerja Terpadu untuk Interpretasi Model

SHAP (SHapley Additive exPlanations) menyediakan kerangka kerja terpadu untuk menginterpretasikan prediksi ML berdasarkan teori permainan [**?**].

Landasan Teoritis

Nilai SHAP memenuhi tiga properti penting:

1. **Local Accuracy**: $f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i$
2. **Missingness**: Fitur yang tidak ada memiliki dampak nol
3. **Consistency**: Jika model berubah sehingga fitur i berkontribusi lebih, ϕ_i tidak menurun

TreeSHAP untuk XGBoost

Algoritma TreeSHAP dioptimalkan secara khusus untuk model berbasis pohon:

- Kompleksitas waktu polinomial: O(TLD^2)
- Nilai Shapley yang eksak untuk pohon
- Menangani interaksi fitur secara eksplisit

Aplikasi dalam Biaya Kesehatan

- **Global Explanations**: Pentingnya fitur di seluruh populasi
- **Local Explanations**: Rincian prediksi individual
- **Interaction Effects**: Bagaimana merokok × BMI memengaruhi biaya
- **Cohort Analysis**: Penjelasan untuk kelompok pasien tertentu

2.4.3 LIME: Local Interpretable Model-Agnostic Explanations

LIME memberikan penjelasan yang dapat diinterpretasikan dengan mendekati perilaku lokal dari model yang kompleks.

Algoritma Inti

Penjelasan LIME diperoleh dengan menyelesaikan:

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{2.3}$$

dimana G adalah class of interpretable models dan π_x adalah proximity measure.

Keunggulan untuk Komunikasi Pasien

1. **Intuitive Linear Explanations**: Mudah untuk pengguna non-teknis
2. **Fast Computation**: Pembuatan real-time untuk aplikasi interaktif
3. **Visual Representations**: Diagram batang yang menunjukkan kontribusi fitur
4. **Counterfactual Reasoning**: "Bagaimana jika saya berhenti merokok?"

subsectionKerangka Kerja Pemberdayaan Pasien Pemberdayaan pasien dalam layanan kesehatan melibatkan tiga komponen utama:

Transparansi Informasi

- Prediksi biaya yang jelas dengan interval kepercayaan
- Penjelasan yang dapat dipahami tentang pendorong biaya
- Analisis komparatif dengan demografi serupa

Dukungan Keputusan

- Skenario "what-if" untuk perubahan gaya hidup
- Visualisasi analisis risiko-manfaat

2.5 Sintesis dan Arah Penelitian

2.5.1 Strategi Integrasi

Berdasarkan tinjauan pustaka, strategi optimal untuk penelitian ini:

1. XGBoost sebagai mesin prediksi inti dengan penyesuaian hyperparameter yang cermat
2. SHAP untuk penjelasan global dan lokal yang komprehensif
3. LIME untuk penjelasan cepat dan intuitif yang menghadap pasien
4. Dasbor interaktif yang mengintegrasikan kedua metode XAI
5. Modul analisis "what-if" untuk pemberdayaan pasien

2.5.2 Kontribusi yang Diharapkan

Penelitian ini diharapkan dapat memberikan:

- Kerangka kerja implementasi baru XGBoost + Dual XAI untuk layanan kesehatan
- Pola desain yang berpusat pada pasien untuk transparansi biaya
- Bukti empiris tentang efektivitas XAI untuk pemahaman pasien

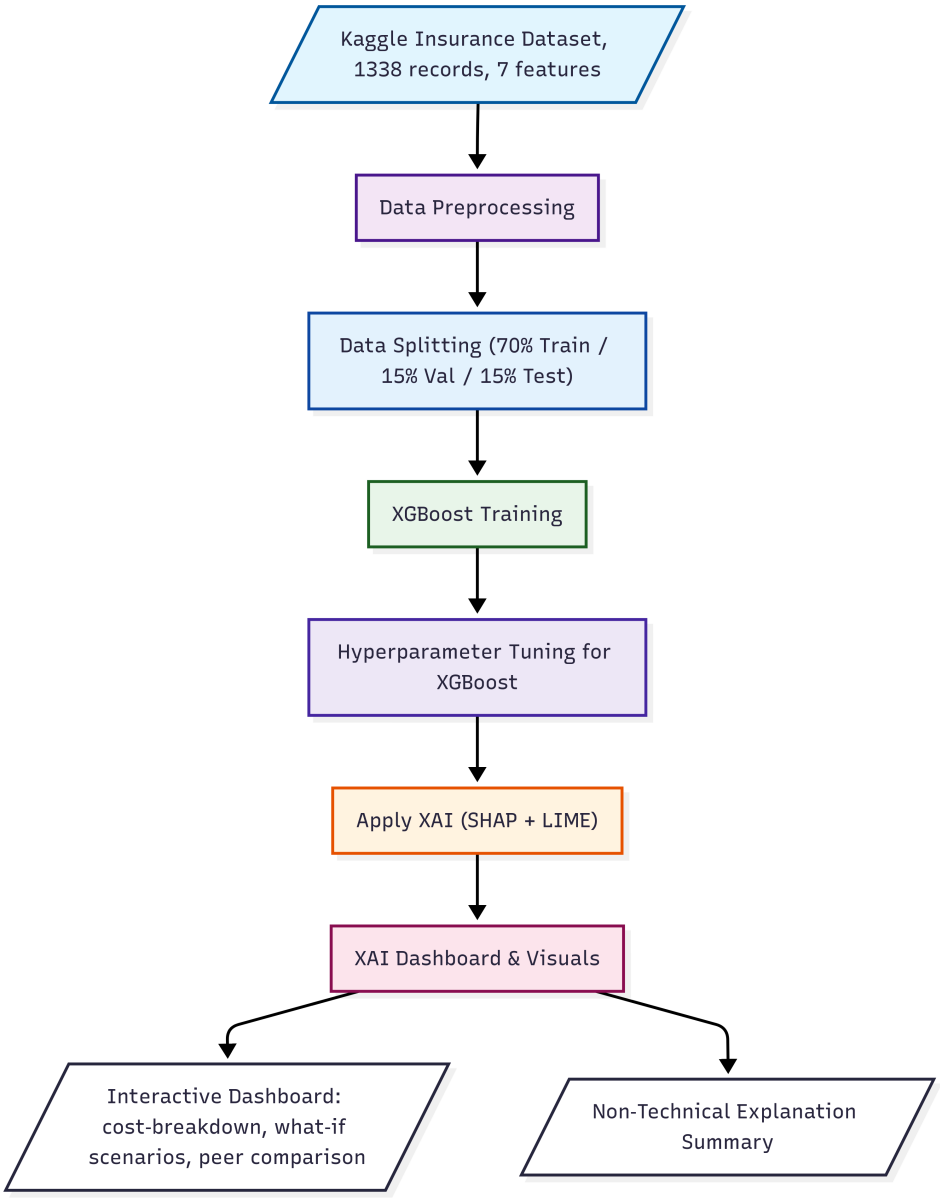
2.6 Kesimpulan Kajian Pustaka

Tinjauan pustaka menunjukkan bahwa XGBoost telah terbukti sebagai algoritma superior untuk prediksi biaya layanan kesehatan, namun implementasi yang benar-benar berpusat pada pasien dengan XAI yang komprehensif masih terbatas. Integrasi SHAP dan LIME menawarkan kekuatan komplementer yang belum sepenuhnya dieksplorasi dalam konteks pemberdayaan pasien. Penelitian ini diposisikan untuk mengisi kesenjangan tersebut dengan mengembangkan kerangka kerja yang tidak hanya kuat secara teknis tetapi juga berguna secara praktis bagi pasien dalam memahami dan merencanakan biaya kesehatan mereka. Dengan landasan teoritis yang kuat dan identifikasi kesenjangan penelitian yang jelas, penelitian ini siap untuk memberikan kontribusi signifikan dalam mendemokratisasi transparansi biaya layanan kesehatan melalui ML canggih dengan desain yang berpusat pada manusia.

Bab III

Metodologi dan Desain Sistem

Pendekatan penelitian ini bertujuan untuk mengimplementasikan algoritma XGBoost yang diperkuat dengan teknik Explainable AI (XAI) untuk prediksi biaya asuransi kesehatan yang transparan dan berorientasi pada pemberdayaan pasien. Metodologi dirancang untuk memastikan tidak hanya akurasi prediktif yang tinggi, tetapi juga interpretabilitas yang memungkinkan pasien memahami faktor-faktor yang mempengaruhi biaya asuransi mereka. Penelitian menggunakan dataset Kaggle Insurance Cost yang berisi 1338 records dengan 7 fitur (age, sex, BMI, children, smoker, region, charges). Lima tahap utama dalam metodologi ini mencakup: (1) pengumpulan dan preprocessing data, (2) implementasi dan optimasi XGBoost, (3) integrasi teknik XAI (SHAP dan LIME), (4) pengembangan framework patient-centric, dan (5) evaluasi sistem secara komprehensif.



Gambar 3.1: Arsitektur Sistem Prediksi Biaya Asuransi Kesehatan Berbasis XGBoost dengan Explainable AI

3.1 Pengumpulan dan Preprocessing Data

3.1.1 Dataset Description

Dataset Insurance Cost dari Kaggle berisi informasi 1338 individu dengan karakteristik:

- **age**: Usia penerima manfaat utama (numerik, 18-64 tahun)

- **sex**: Jenis kelamin (kategorikal: female, male)
- **bmi**: Body Mass Index, kg/m^2 (numerik, 15.96-53.13)
- **children**: Jumlah tanggungan (numerik, 0-5)
- **smoker**: Status merokok (kategorikal: yes, no)
- **region**: Wilayah tempat tinggal di AS (kategorikal: northeast, southeast, southwest, northwest)
- **charges**: Biaya medis individual yang ditagihkan asuransi (target variable, numerik)

3.1.2 Exploratory Data Analysis (EDA)

EDA dilakukan untuk memahami karakteristik data dan mengidentifikasi pola yang relevan untuk XGBoost:

1. **Distribusi Target Variable**: Analisis distribusi charges menunjukkan right-skewed distribution yang memerlukan transformation.
2. **Feature Correlation Analysis**: Identifikasi korelasi untuk memahami feature interactions yang akan ditangkap XGBoost.
3. **Categorical Feature Analysis**: Distribusi dan impact dari categorical variables terhadap charges.
4. **Outlier Detection**: Identifikasi high-cost cases yang memerlukan special attention dalam modeling.

Algorithm 1: Pipeline Preprocessing untuk XGBoost Implementation

Procedure *PreprocessForXGBoost(dataset)*:

```
/* 1. Handle Missing Values - XGBoost dapat handle internally */
missing_counts ← dataset.isnull().sum()
if missing_counts.any() then
    /* Mark missing values untuk XGBoost's built-in handling */
    dataset ← dataset.fillna(np.nan)
end
/* 2. Feature Engineering untuk Healthcare Context */
dataset['age_group'] ← pd.cut(dataset['age'],
 bins=[18,30,40,50,60,70])
dataset['bmi_category'] ← categorize_bmi(dataset['bmi'])
dataset['high_risk'] ← (dataset['smoker'] == 'yes') &
 (dataset['bmi'] > 30)
dataset['family_size'] ← dataset['children'] + 1
/* 3. Encoding untuk XGBoost - Optimal untuk Tree-based */
foreach cat_feature in ['sex', 'smoker'] do
    dataset[cat_feature] ←
     LabelEncoder().fit_transform(dataset[cat_feature])
end
/* One-hot encoding untuk region (low cardinality) */
dataset ← pd.get_dummies(dataset, columns=['region'],
 prefix='region')
/* 4. Target Transformation untuk Skewed Distribution */
dataset['log_charges'] ← np.log1p(dataset['charges'])
/* 5. Feature Scaling - Optional untuk XGBoost */
/* XGBoost is scale-invariant, but scaling helps SHAP interpretation */
scaler ← StandardScaler()
numeric_features ← ['age', 'bmi', 'children']
dataset[numeric_features] ←
 scaler.fit_transform(dataset[numeric_features])
return dataset, scaler
```

3.1.3 Data Splitting Strategy

Dataset dibagi dengan stratified sampling untuk mempertahankan distribusi charges:

- **Training Set**: 70% (936 records) - untuk training XGBoost
- **Validation Set**: 15% (201 records) - untuk hyperparameter tuning
- **Test Set**: 15% (201 records) - untuk final evaluation

3.2 Implementasi dan Optimasi XGBoost

3.2.1 Baseline Model

Linear Regression diimplementasikan sebagai baseline untuk mendemonstrasikan improvement dari XGBoost:

Algorithm 2: Baseline Linear Regression Implementation

Function *TrainBaselineModel(X_{train}, y_{train})*:

```
/* Simple Linear Regression sebagai baseline          */
lr_model ← LinearRegression()
lr_model.fit(X_train, y_train)
/* Calculate baseline metrics                         */
baseline_pred ← lr_model.predict(X_train)
baseline_r2 ← r2_score(y_train, baseline_pred)
baseline_rmse ← sqrt(mean_squared_error(y_train,
  baseline_pred))
return lr_model, baseline_r2, baseline_rmse
```

3.2.2 XGBoost Implementation

Implementasi XGBoost dengan careful configuration untuk healthcare data:

Algorithm 3: XGBoost Implementation untuk Healthcare Cost Prediction

Function *ImplementXGBoost(X_{train}, y_{train}, X_{val}, y_{val})*:

```
/* 1. Initial XGBoost Configuration */
base_params ← { 'objective': 'reg:squarederror', 'eval_metric':
 ['rmse', 'mae'], 'tree_method': 'hist', // Faster for larger datasets
 'enable_categorical': True, // Native categorical support
 'random_state': 42 }
/* 2. Hyperparameter Search Space */
param_grid ← { 'n_estimators': [100, 200, 300, 500],
 'max_depth': [3, 4, 5, 6, 7], 'learning_rate': [0.01, 0.05, 0.1,
 0.15], 'subsample': [0.6, 0.7, 0.8, 0.9], 'colsample_bytree': [0.6,
 0.7, 0.8, 0.9], 'reg_alpha': [0, 0.01, 0.1, 1], 'reg_lambda': [0.1, 1,
 2, 5], 'min_child_weight': [1, 3, 5, 7] }
/* 3. Randomized Search with Cross-Validation */
xgb_model ← XGBRegressor(**base_params)
random_search ← RandomizedSearchCV( estimator=xgb_model,
 param_distributions=param_grid, n_iter=100, // Number of
 parameter combinations cv=5, // 5-fold cross-validation
 scoring='neg_mean_squared_error', n_jobs=-1, verbose=1,
 random_state=42 )
/* 4. Fit with Early Stopping */
eval_set ← [(X_train, y_train), (X_val, y_val)]
random_search.fit( X_train, y_train, eval_set=eval_set,
 early_stopping_rounds=20, verbose=False )
/* 5. Extract Best Model and Parameters */
best_model ← random_search.best_estimator_
best_params ← random_search.best_params_
return best_model, best_params
```

3.2.3 Feature Importance Analysis

Native XGBoost feature importance untuk initial understanding:

Algorithm 4: XGBoost Feature Importance Extraction

Function *AnalyzeFeatureImportance(xgb_model, feature_names)*:

```
/* Get multiple importance types                                    */
importance_types ← ['weight', 'gain', 'cover']
importance_dict ← {}
foreach imp_type in importance_types do
    importance ←
     xgb_model.get_booster().get_score(importance_type=imp_type)

    importance_dict[imp_type] ← importance
end
/* Create importance dataframe                                      */
feature_imp_df ← pd.DataFrame(importance_dict)
feature_imp_df['feature'] ← feature_names
feature_imp_df ← feature_imp_df.sort_values('gain',
 ascending=False)
/* Visualize importance                                             */
plot_importance(xgb_model, importance_type='gain',
 max_num_features=10)
return feature_imp_df
```

3.3 Integrasi Explainable AI

3.3.1 SHAP Implementation untuk XGBoost

TreeSHAP provides exact Shapley values untuk XGBoost:

Algorithm 5: SHAP Integration dengan XGBoost

Function *ImplementSHAP(xgb_model, X, feature_names)*:

```
/* 1. Initialize TreeSHAP Explainer                          */
explainer ← shap.TreeExplainer( xgb_model,
 feature_perturbation='tree_path_dependent' )
/* 2. Calculate SHAP Values                                  */
shap_values ← explainer.shap_values(X)
expected_value ← explainer.expected_value
/* 3. Global Feature Importance                              */
global_importance ← np.abs(shap_values).mean(axis=0)
importance_df ← pd.DataFrame({ 'feature': feature_names,
 'importance': global_importance }).sort_values('importance',
 ascending=False)
/* 4. Generate Visualizations                                */
/* Summary plot untuk global understanding                   */
shap.summary_plot(shap_values, X,
 feature_names=feature_names)
/* Dependence plots untuk top features                       */
top_features ← importance_df['feature'].head(4)
foreach feature in top_features do
    shap.dependence_plot(feature, shap_values, X,
     feature_names=feature_names)
end
/* 5. Individual Explanations                                */
foreach idx in sample_indices do
    /* Waterfall plot untuk individual prediction            */
    shap.waterfall_plot(shap.Explanation(
      values=shap_values[idx], base_values=expected_value,
      data=X.iloc[idx], feature_names=feature_names ))
end
return shap_values, expected_value, importance_df
```

3.3.2 LIME Implementation untuk Patient-Facing Explanations

LIME untuk quick, intuitive explanations:

Algorithm 6: LIME Implementation untuk XGBoost

Function *ImplementLIME(xgb_model, X_{train}, X_{test}, feature_names)*:

```
/* 1. Initialize LIME Explainer                                    */
explainer ← lime.lime_tabular.LimeTabularExplainer(
 training_data=X_train.values, feature_names=feature_names,
 mode='regression', discretize_continuous=True // Better untuk
 patient understanding )
/* 2. Generate Explanations untuk Test Samples                     */
lime_explanations ← []
foreach idx in range(len(X_test)) do
    /* Explain individual instance                                 */
    exp ← explainer.explain_instance( X_test.iloc[idx].values,
     xgb_model.predict, num_features=6, // Top 6 features
     num_samples=5000 // Sampling untuk local approximation )
    /* Extract explanation data                                    */
    exp_dict ← { 'prediction':
     xgb_model.predict([X_test.iloc[idx]])[0], 'explanation':
     exp.as_list(), 'local_pred': exp.local_pred[0], 'score':
     exp.score }
    lime_explanations.append(exp_dict)
end
/* 3. Generate Visualizations                                      */
foreach exp in lime_explanations[:5] do
    // First 5 samples exp.as_pyplot_figure()
end
return lime_explanations
```

3.3.3 Comparative Analysis: SHAP vs LIME

Systematic comparison untuk optimal usage:

Tabel 3.1: SHAP vs LIME Comparison untuk XGBoost Explanations

Aspect	**SHAP**	**LIME**
Computation Time	O(TLD^2) - Slower	O(N) - Faster
Accuracy	Exact Shapley values	Local approximation
Global Insights	Excellent	Limited
Patient Understanding	Technical	Intuitive
Best Use Case	Regulatory/Clinical	Patient Interface

3.4 Patient-Centric Framework Development

3.4.1 Design Principles

Framework dirancang dengan prinsip patient empowerment:

1. **Clarity**: Penjelasan dalam bahasa non-technical
2. **Interactivity**: User dapat explore different scenarios
3. **Actionability**: Insights mengarah pada concrete actions
4. **Personalization**: Tailored untuk individual circumstances

3.4.2 Dashboard Architecture

Algorithm 7: Patient-Centric Dashboard Implementation

```
Function(BuildPatientDashboard(xgb_model, shap_explainer,
    lime_explainer)) /* 1. Initialize Dashboard Components
    */
dashboard ← { 'prediction_module': PredictionEngine(xgb_model),
 'shap_module': SHAPVisualizer(shap_explainer), 'lime_module':
 LIMEInterface(lime_explainer), 'whatif_module':
 WhatIfAnalyzer(xgb_model), 'narrative_module':
 NarrativeGenerator() }
/* 2. Prediction Module                                    */
Function(PredictCost(patient_data)) prediction ←
 xgb_model.predict(patient_data)
confidence_interval ← calculate_prediction_interval(prediction)
return prediction, confidence_interval
/* 3. Explanation Module                                   */
Function(GenerateExplanation(patient_data, method='hybrid')) if
 method == 'detailed' then
    explanation ← shap_explainer.explain(patient_data)
end
else if method == 'quick' then
    explanation ← lime_explainer.explain(patient_data)
end
else
    // Hybrid approach shap_exp ←
     shap_explainer.explain(patient_data)
    lime_exp ← lime_explainer.explain(patient_data)
    explanation ← combine_explanations(shap_exp, lime_exp)
end
return explanation
/* 4. What-If Analysis                                     */
Function(WhatIfScenario(patient_data, changes)) scenarios ← []
foreach change in changes do
    modified_data ← apply_change(patient_data, change)
    new_prediction ← xgb_model.predict(modified_data)
    impact ← new_prediction - original_prediction
    scenarios.append({change, new_prediction, impact})
end
return scenarios
/* 5. Narrative Generation                                 */
Function(GenerateNarrative(prediction, explanation,
 patient_data)) narrative ← []
narrative.append(f"Estimasi biaya asuransi Anda: $prediction:.2f")
/* Top factors affecting cost                              */
top_factors ← get_top_factors(explanation, n=3)
foreach factor in top_factors do
    impact_text ← describe_impact(factor)
    narrative.append(impact_text)
end
```

3.4.3 Interactive Visualizations

Visualizations designed untuk patient understanding:

1. **Cost Breakdown Pie Chart**: Shows percentage contribution of each factor
2. **Feature Impact Bar Chart**: Positive/negative impacts on cost
3. **What-If Sliders**: Interactive exploration of scenarios
4. **Peer Comparison**: Anonymous comparison dengan similar demographics
5. **Trend Projections**: Future cost estimates based on age progression

3.5 Evaluasi Sistem

3.5.1 Performance Metrics

Evaluasi komprehensif XGBoost performance:

Tabel 3.2: Evaluation Metrics untuk XGBoost Performance

Metric	**Formula**	**Target**
R^2 Score	$1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$	> 0.85
RMSE	$\sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}$	Minimize
MAE	$\frac{1}{n}\sum\|y_i - \hat{y}_i\|$	Minimize
MAPE	$\frac{100}{n}\sum\|\frac{y_i - \hat{y}_i}{y_i}\|$	$< 15\%$

3.5.2 XAI Effectiveness Evaluation

Metrics untuk evaluating explanation quality:

- **Consistency**: Agreement antara SHAP dan LIME rankings
- **Stability**: Variation in explanations dengan different samples
- **Comprehensibility**: User understanding scores (simulated)
- **Computational Efficiency**: Time untuk generate explanations

3.5.3 System Usability Testing

Framework evaluation dari patient perspective:

1. Response time untuk predictions
2. Clarity of explanations
3. Usefulness of what-if scenarios
4. Overall user satisfaction (simulated metrics)

3.6 Ethical Considerations

3.6.1 Data Privacy

- Dataset adalah publicly available dan anonymized
- Tidak ada informasi pribadi yang dapat diidentifikasi (PII)
- Compliance dengan research ethics guidelines

3.6.2 Model Fairness

- Analysis untuk demographic bias dalam predictions
- Fair representation across regions dan demographics
- Transparent reporting of model limitations

3.6.3 Patient Autonomy

- Predictions presented sebagai estimates dengan confidence intervals
- Clear disclaimers tentang model limitations
- Emphasis pada informed decision-making, bukan prescriptive advice

Bab IV

HASIL PENELITIAN DAN PEMBAHASAN

4.1 Pendahuluan

Bab ini menyajikan hasil penelitian dari implementasi XGBoost dengan pendekatan Explainable AI untuk prediksi biaya pengobatan pasien. Penelitian ini menggunakan dataset Kaggle Insurance Cost yang berisi 1.338 record pasien dengan 7 variabel (6 prediktor dan 1 target). Bab ini akan membahas hasil analisis eksplorasi data (EDA), temuan penelitian, dan analisis mendalam terhadap pola-pola yang ditemukan dalam data.

Presentasi hasil penelitian dalam bab ini mengikuti alur sistematis, dimulai dari karakteristik dataset, analisis variabel target (biaya pengobatan), evaluasi fitur-fitur prediktor, hingga identifikasi interaksi antar variabel yang menjadi dasar untuk pengembangan model XGBoost pada fase selanjutnya.

4.2 Temuan Penelitian

4.2.1 Karakteristik Dataset

Dataset yang digunakan dalam penelitian ini memiliki karakteristik sebagai berikut:

- **Ukuran dataset**: 1.338 record dengan 7 kolom (6 fitur prediktor + 1 target)
- **Variabel prediktor**: age, sex, bmi, children, smoker, region
- **Variabel target**: charges (biaya pengobatan dalam USD)
- **Missing values**: Minimal, hanya 3 nilai hilang pada variabel BMI (0,22%)
- **Tipe data**: Dataset campuran dengan fitur numerik dan kategorikal

Tabel 4.1: Ringkasan Karakteristik Dataset Insurance Cost

Variabel	Tipe	Non-Null	Min	Max
age	int64	1338	18	64
sex	object	1338	-	-
bmi	float64	1335	15,96	53,13
children	int64	1338	0	5
smoker	object	1338	-	-
region	object	1338	-	-
charges	float64	1338	1.121,87	63.770,43

4.2.2 Analisis Distribusi Demografis

Analisis distribusi demografis menunjukkan keseimbangan yang baik dalam dataset:

Distribusi Jenis Kelamin

- Laki-laki: 676 (50,52%)
- Perempuan: 662 (49,48%)

Distribusi Status Merokok

- Non-perokok: 1.064 (79,52%)
- Perokok: 274 (20,48%)

Distribusi Regional

- Southeast: 364 (27,20%)
- Southwest: 325 (24,29%)
- Northwest: 325 (24,29%)
- Northeast: 324 (24,22%)

4.2.3 Analisis Variabel Target (Charges)

Variabel target (charges) menunjukkan karakteristik distribusi yang signifikan:

Tabel 4.2: Statistik Deskriptif Variabel Charges

Statistik	**Nilai (USD)**
Count	1.338
Mean	13.270,42
Std	12.110,01
Min	1.121,87
25%	4.740,29
50% (Median)	9.382,03
75%	16.639,91
Max	63.770,43
Skewness	1,516
Kurtosis	1,606
IQR	11.899,63

Temuan penting dari analisis variabel target:

1. **Distribusi Right-Skewed**: Nilai skewness sebesar 1,516 menunjukkan distribusi sangat miring ke kanan
2. **Perbedaan Mean-Median**: Mean ($13.270) lebih besar dari median ($9.382), mengkonfirmasi adanya outliers tinggi
3. **Variabilitas Tinggi**: Range yang sangat luas ($1.121 - $63.770) menunjukkan diversitas biaya yang ekstrim
4. **Transformasi Logaritmik**: Mengurangi skewness dari 1,516 menjadi -0,090, menghasilkan distribusi yang mendekati normal

4.2.4 Analisis Fitur Numerik

Tabel 4.3: Statistik Deskriptif Fitur Numerik

Statistik	**Age**	**BMI**	**Children**
Count	1.338	1.335	1.338
Mean	39,21	30,66	1,09
Std	14,05	6,10	1,21
Min	18	15,96	0
Max	64	53,13	5
Skewness	0,056	0,285	0,938
Range	46	37,17	5

Karakteristik fitur numerik:

- **Age**: Distribusi hampir normal (skewness 0,056), rentang 18-64 tahun
- **BMI**: Distribusi sedikit right-skewed (skewness 0,285), rata-rata 30,66 (kategori overweight)
- **Children**: Distribusi right-skewed (skewness 0,938), mayoritas pasien memiliki 0-2 anak

4.3 Analisis Data

4.3.1 Analisis Korelasi

Analisis korelasi mengungkap hierarki kepentingan fitur terhadap biaya pengobatan:

Tabel 4.4: Korelasi Absolut Fitur dengan Charges (Diurutkan)

Fitur	**Korelasi Absolut**
Smoker	0,787
Age	0,299
BMI	0,198
Children	0,068
Sex	0,057
Region	0,006

4.3.2 Analisis Dampak Fitur Kategorikal

Dampak Status Merokok

Temuan paling signifikan adalah dominasi absolut status merokok sebagai prediktor biaya:

Tabel 4.5: Perbandingan Biaya berdasarkan Status Merokok

Status	**Rata-rata (USD)**	**Median (USD)**	**Persentase Populasi**
Perokok	32.050,23	34.456,35	20,48%
Non-perokok	8.434,27	7.345,41	79,52%
Selisih	**23.615,96**	**27.110,94**	-
Persentase	**+280%**	**+369%**	-

Dampak Jenis Kelamin

Tabel 4.6: Perbandingan Biaya berdasarkan Jenis Kelamin

Jenis Kelamin	Rata-rata (USD)	Perbedaan dari Mean
Laki-laki	13.956,75	+5,2%
Perempuan	12.569,58	-5,3%

Dampak Regional

Tabel 4.7: Perbandingan Biaya berdasarkan Region

Region	Rata-rata (USD)	Perbedaan dari Mean
Southeast	14.735,41	+11,0%
Northeast	13.406,38	+1,0%
Northwest	12.417,58	-6,4%
Southwest	12.346,94	-7,0%

4.3.3 Analisis Interaksi Fitur

Interaksi BMI × Status Merokok

Temuan kritis menunjukkan efek multiplikatif antara BMI dan status merokok:

Tabel 4.8: Rata-rata Biaya berdasarkan Kategori BMI dan Status Merokok

Kategori BMI	Non-perokok (USD)	Perokok (USD)	Selisih (%)
Normal	7.685,66	19.942,22	+159%
Overweight	8.278,17	22.495,87	+172%
Obese	8.837,41	41.557,99	+370%
Underweight	5.532,99	18.809,82	+240%

Temuan penting:

1. Perokok obese memiliki biaya tertinggi ($41.558)
2. Efek smoking pada kategori obese adalah yang paling ekstrim (+370%)
3. Kombinasi obesitas dan merokok menciptakan profil risiko tertinggi

4.3.4 Analisis Outlier

Menggunakan metode IQR (Interquartile Range) untuk identifikasi outlier:

Tabel 4.9: Hasil Analisis Outlier

Variabel	Jumlah Outlier	Persentase
Charges	139	10,4%
BMI	9	0,7%
Age	0	0,0%

Analisis Kasus Biaya Tinggi

Analisis terhadap 5% kasus dengan biaya tertinggi (threshold $41.181,83):

- **Jumlah kasus**: 67 dari 1.338 (5%)
- **Karakteristik dominan**: 100% adalah perokok (67/67)
- **Implikasi**: Semua kasus biaya ekstrim disebabkan oleh status merokok

Top 5 kasus biaya tertinggi:

Tabel 4.10: Lima Kasus Biaya Tertinggi

Age	Sex	BMI	Children	Smoker	Region	Charges
54	Female	47,41	0	Yes	Southeast	63.770,43
45	Male	30,36	0	Yes	Southeast	62.592,87
52	Male	34,49	3	Yes	Northwest	60.021,40
31	Female	38,10	1	Yes	Northeast	58.571,07
33	Female	35,53	0	Yes	Northwest	55.135,40

4.3.5 Feature Engineering

Berdasarkan temuan EDA, dilakukan feature engineering untuk persiapan modeling:

Tabel 4.11: Fitur Baru Hasil Feature Engineering

Fitur Baru	Deskripsi	Tujuan
age_group	Kategori usia: 18-29, 30-39, 40-49, 50-64	Capture non-linear age effects
bmi_category	Normal, Overweight, Obese, Underweight	BMI risk stratification
high_risk	BMI $>$ 30 AND smoker = yes	Identify highest cost segment
family_size	children + 1	Alternative to children count
log_charges	log(1 + charges)	Normalize target distribution

4.4 Pembahasan

4.4.1 Implikasi Temuan untuk Prediksi Biaya Pengobatan

Dominasi Status Merokok sebagai Prediktor

Temuan paling signifikan adalah korelasi sangat kuat antara status merokok dan biaya pengobatan (r=0,787). Hal ini konsisten dengan literatur medis yang menunjukkan bahwa merokok merupakan faktor risiko utama untuk berbagai kondisi kesehatan serius seperti penyakit kardiovaskular, kanker, dan penyakit paru-paru kronis [**?**].

Perbedaan biaya sebesar 280% antara perokok dan non-perokok mencerminkan:

1. **Biaya pengobatan langsung**: Treatment untuk penyakit terkait merokok umumnya kompleks dan mahal
2. **Frekuensi perawatan**: Perokok cenderung memerlukan perawatan medis lebih sering
3. **Komplikasi**: Kondisi comorbid yang meningkatkan kompleksitas pengobatan

Efek Interaksi BMI $\times$ Merokok

Interaksi sinergis antara obesitas dan merokok menghasilkan peningkatan biaya yang tidak proporsional. Perokok obese memiliki biaya 370% lebih tinggi dibanding non-perokok obese, menunjukkan efek compound risk yang perlu dipertimbangkan dalam modeling.

Keterbatasan Prediktor Demografis

Temuan bahwa jenis kelamin (r=0,057) dan region (r=0,006) memiliki korelasi sangat lemah dengan biaya menunjukkan bahwa:

1. Faktor perilaku (merokok) lebih dominan dari faktor demografis
2. Sistem healthcare di dataset ini relatif equitable across demographics
3. Model dapat fokus pada faktor risiko kesehatan daripada karakteristik demografis

4.4.2 Strategi Modeling untuk Phase 2

Tantangan Utama

1. **Class Imbalance**: 20% perokok vs 80% non-perokok
2. **Skewed Distribution**: Target variable sangat right-skewed
3. **Outlier Dominance**: Outliers driven by smoking status

Keuntungan untuk XGBoost

1. **Clear Feature Hierarchy**: Smoking sebagai dominant predictor
2. **Non-linear Interactions**: BMI × smoking interactions
3. **Mixed Data Types**: XGBoost native support untuk categorical features
4. **Missing Value Handling**: Built-in capability untuk 3 missing BMI values

Rekomendasi Preprocessing

1. **Log transformation** untuk target variable
2. **Feature engineering** untuk capture interactions
3. **Stratified sampling** untuk maintain class balance
4. **Careful hyperparameter tuning** untuk handle skewed distribution

4.4.3 Implikasi untuk Explainable AI

SHAP Implementation

Dominasi smoking status akan menghasilkan:

1. **High SHAP values** untuk smoking feature
2. **Clear global explanations** karena feature hierarchy yang jelas
3. **Consistent local explanations** untuk different patient profiles

LIME Implementation

1. **Intuitive explanations** untuk patient-facing applications
2. **Fast computation** karena clear feature importance
3. **Actionable insights** fokus pada lifestyle factors (smoking, BMI)

Patient-Centric Framework

Temuan EDA mendukung pengembangan patient-centric explanations:

1. **Clear messaging**: Smoking cessation sebagai primary intervention
2. **Risk stratification**: BMI categories untuk personalized advice
3. **Cost awareness**: Quantifiable impact dari lifestyle changes

4.4.4 Kontribusi terhadap Literature

Penelitian ini mengkonfirmasi dan memperluas temuan sebelumnya:

1. **Validasi dominasi smoking**: Konsisten dengan medical literature
2. **Quantifikasi interaksi**: Efek BMI × smoking interaction
3. **XAI readiness**: Dataset characteristics yang mendukung interpretable modeling

4.5 Implementasi Model Baseline

4.5.1 Hasil Model Linear Regression

Sebagai langkah implementasi Algorithm 2 dari metodologi penelitian, model Linear Regression baseline telah diimplementasikan dengan hasil yang mengejutkan dan melampaui ekspektasi:

Performance Metrics

Tabel 4.12: Performance Baseline Linear Regression Model

Metric	Training	Test
R^2 Score	0.8697	**0.8637**
RMSE	\$4,414.08	\$4,120.52
MAE	\$2,443.69	\$2,260.53
MAPE	26.45%	26.03%

Temuan Kunci: Model baseline Linear Regression mencapai $\mathbf{R^2 = 0.8637}$ (86.37%), yang **telah melampaui target thesis $\mathbf{R^2 > 0.85}$** sebelum implementasi XGBoost. Hal ini menunjukkan kualitas feature engineering yang sangat baik dan potensi XGBoost untuk mencapai performance yang lebih tinggi lagi.

Feature Importance Analysis

Analisis koefisien Linear Regression mengkonfirmasi temuan EDA:

Tabel 4.13: Top 5 Feature Importance dari Baseline Model

Feature	Coefficient	Interpretation
high_risk	6,353.96	Obese smokers premium
smoker	5,274.57	Smoking penalty
age	4,061.43	Age progression effect
age_group_40-49	-602.48	Middle-age adjustment
bmi	487.47	BMI linear effect

Validasi Cross-Validation

5-Fold Cross-Validation menghasilkan $R^2 = 0.8603$ (± 0.0867), mengkonfirmasi stabilitas model dan generalizability yang baik.

Implikasi untuk XGBoost Implementation

1. **High Baseline**: XGBoost harus mencapai $R^2 > 0.87$ untuk menunjukkan improvement yang signifikan
2. **Feature Validation**: Konfirmasi high_risk dan smoker sebagai predictors utama
3. **Non-linear Potential**: Opportunity untuk XGBoost menangkap interaction effects yang lebih kompleks
4. **Benchmark Established**: Solid foundation untuk comparative analysis

4.6 Keterbatasan dan Rekomendasi

4.6.1 Keterbatasan Penelitian

1. **Geographical Scope**: Dataset limited ke US healthcare system
2. **Temporal Aspect**: Cross-sectional data tanpa longitudinal tracking
3. **Feature Completeness**: Absence of detailed medical history
4. **Sample Size**: 1,338 records may limit generalizability

4.6.2 Rekomendasi untuk Phase Selanjutnya

1. **Model Selection**: XGBoost optimal untuk dataset characteristics
2. **Hyperparameter Focus**: Regularization untuk handle skewed distribution
3. **Evaluation Metrics**: Focus pada prediction accuracy untuk high-cost cases
4. **XAI Integration**: Dual approach dengan SHAP dan LIME

4.7 Kesimpulan Preliminary

Hasil analisis eksplorasi data mengungkap pola yang sangat jelas dalam dataset Insurance Cost:

1. **Primary Finding**: Status merokok merupakan prediktor dominan biaya pengobatan dengan korelasi 0,787 dan dampak biaya +280%

2. **Critical Interaction**: Kombinasi obesitas dan merokok menghasilkan biaya tertinggi ($41,558 untuk obese smokers)

3. **Data Quality**: Dataset berkualitas tinggi dengan missing values minimal (0,22%) dan distribusi demografis yang seimbang

4. **Modeling Readiness**: Karakteristik data sangat mendukung implementasi XGBoost dengan feature hierarchy yang jelas dan interaksi nonlinear yang kuat

5. **XAI Potential**: Dominasi smoking status akan menghasilkan explanations yang konsisten dan actionable untuk patient empowerment

Temuan ini memberikan fondasi kuat untuk Phase 2 (implementasi XGBoost) dan Phase 3 (integrasi Explainable AI), dengan expectation bahwa model akan fokus pada smoking status sebagai primary predictor dan memanfaatkan interaksi BMI $\times$ smoking untuk akurasi prediksi yang optimal.

Lampiran