PRA-SKRIPSI

Perbandingan Performa XGBoost dan LightGBM dalam Prediksi Churn Pelanggan Netflix

HILDA DESFIANTY ARIFIN
NPM 22081010206

KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI
UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAWA TIMUR
FAKULTAS ILMU KOMPUTER
PROGRAM STUDI INFORMATIKA
SURABAYA
2025

BAB I
PENDAHULUAN

1.1. Latar Belakang

Di era digital, layanan berlangganan hiburan berbasis platform streaming telah mengubah pola konsumsi media dari format konvensional ke konsumsi on-demand yang dipersonalisasi. Literatur akademik menunjukkan perpindahan struktural ini dan implikasinya terhadap strategi konten serta loyalitas pengguna [1]. Customer churn atau keputusan pelanggan untuk menghentikan langganan menjadi masalah strategis utama bagi platform OTT over-the-top. Penelitian empiris pada OTT dan layanan sejenis menegaskan bahwa pengalaman pengguna, kepuasan terhadap konten, dan aspek penggunaan seperti kebiasaan menonton berkontribusi terhadap risiko churn sehingga memerlukan model prediktif yang sensitif terhadap fitur-fitur perilaku tersebut [2]. Ciri khas domain streaming adalah data perilaku yang sangat granular, yaitu durasi tontonan, jeda antar episode, pola binge-watching, dan preferensi genre. Penelitian statistik dan model perilaku menonton menunjukkan bahwa mereduksi kompleksitas data waktu nyata ke fitur-fitur ringkas dapat meningkatkan kemampuan prediktif untuk masalah churn. Oleh karena itu, konstruksi fitur perilaku menjadi aspek penting dalam studi churn pada platform streaming [3].

Studi tentang prediksi churn menegaskan bahwa teknik ensemble dan gradient boosting menunjukkan performa kuat pada masalah klasifikasi biner dengan banyak fitur sehingga menjadi pilihan populer pada literatur churn. Namun performa akhir sangat terpengaruh oleh penanganan class imbalance dan strategi tuning hyperparameter [4]. Masalah class imbalance dan kebutuhan resampling atau pengaturan bobot kelas sering menjadi pembeda utama dalam efektivitas model. Beberapa studi open-access memperlihatkan bagaimana teknik sampling dan strategi ensemble mengubah kualitas prediksi pada dataset churn. Oleh karena itu, penelitian pada konteks streaming harus memperhatikan teknik resampling dan evaluasi yang tepat [5].

Beberapa penelitian terbaru menyoroti bahwa pendekatan Bayesian Optimization seperti Optuna memberikan solusi tuning hyperparameter yang lebih efisien dan adaptif dibanding GridSearch atau RandomSearch pada model boosting. Hal ini relevan untuk memastikan bahwa perbandingan model dilakukan pada kondisi

parameter optimal. Meski demikian, terdapat gap penelitian karena sedikit studi open-access yang secara khusus melakukan perbandingan terkontrol antara XGBoost dan LightGBM pada dataset churn dari platform streaming video dengan fokus pada fitur perilaku watch-time [6]. Banyak perbandingan model di literatur berfokus pada domain telekomunikasi, perbankan, atau e-commerce sehingga aplikasinya ke konteks video on-demand dengan karakteristik watch-time dan binge patterns masih terbatas.

Berdasarkan gap ini, penelitian ini melakukan perbandingan komprehensif performa XGBoost dan LightGBM untuk prediksi churn pada dataset Netflix atau dataset streaming yang merepresentasikan perilaku pengguna. Penelitian ini mengekstraksi fitur perilaku yang relevan seperti durasi tontonan, jeda antar episode, dan genre, melakukan penanganan class imbalance yang sesuai, serta menggunakan Bayesian Optimization sebagai metode tuning untuk memastikan fair-comparison. Analisis mencakup metrik klasifikasi akurasi, presisi, recall, F1, AUC, serta pengukuran efisiensi pelatihan. Kontribusi penelitian berupa rekomendasi empiris untuk pemilihan model boosting di domain streaming dan panduan praktis untuk strategi retensi berbasis data.

1.2. Rumusan Masalah

Berdasarkan uraian latar belakang, rumusan masalah yang akan dibahas dalam penelitian ini Adalah:

1. Bagaimana perbandingan performa algoritma LightGBM dan XGBoost dalam memprediksi customer churn pada pelanggan Netflix?
2. Sejauh mana penerapan Bayesian Optimization dapat membantu menghasilkan perbandingan yang lebih objektif antara kedua model tersebut?
3. Model mana yang memberikan hasil terbaik berdasarkan metrik evaluasi seperti akurasi, presisi, recall, F1-score, AUC, serta efisiensi waktu pelatihan dalam konteks prediksi churn pelanggan Netflix?

1.3. Tujuan Penelitian

Adapun tujuan dari penelitian ini berdasarkan rumusan masalah Adalah sebagai berikut:

1. Menganalisis dan membandingkan performa model LightGBM dan XGBoost dalam memprediksi churn pelanggan Netflix.

2. Menggunakan Bayesian Optimization (Optuna) sebagai metode tuning untuk memastikan setiap model diuji dalam kondisi parameter terbaiknya.
3. Menentukan model terbaik berdasarkan hasil evaluasi komprehensif terhadap metrik performa (akurasi, presisi, recall, F1-score, AUC, dan waktu komputasi).

1.4. Manfaat Penelitian

Penelitian yang dilakukan mempunyai manfaat untuk masyarakat, baik masyarakat akademik maupun non akademik. Adapun beberapa manfaat dari penelitian ini adalah sebagai berikut:

1. Manfaat Akademis
    a. Menambah wawasan dan literatur ilmiah dalam bidang data science dan machine learning, khususnya terkait penerapan algoritma gradient boosting (LightGBM dan XGBoost) untuk kasus prediksi customer churn.
    b. Menjadi referensi akademik bagi penelitian lanjutan yang membahas studi komparatif performa model machine learning pada domain hiburan digital.
    c. Memberikan pemahaman empiris tentang bagaimana perbedaan arsitektur dan parameter dua model gradient boosting memengaruhi hasil prediksi pada dataset pelanggan Netflix.
    d. Dapat digunakan sebagai bahan ajar atau studi pustaka di bidang analisis prediksi pelanggan berbasis machine learning.
2. Manfaat Praktis
    a. Membantu perusahaan layanan streaming digital seperti Netflix dalam mengidentifikasi pelanggan berisiko tinggi churn secara lebih akurat dan efisien.
    b. Menjadi dasar dalam menyusun strategi retensi pelanggan berbasis data, seperti personalisasi konten, pemberian promosi, atau peningkatan pengalaman pengguna.
    c. Memberikan acuan bagi pengambil keputusan bisnis untuk mengoptimalkan pendapatan berulang (recurring revenue) melalui sistem prediksi churn yang handal.

d. Dapat diterapkan sebagai dasar pengembangan dashboard analytics churn bagi tim marketing dan analis data untuk pemantauan performa pelanggan secara real-time.

3. Manfaat Sosial dan Ekonomi
    a. Mendorong penerapan teknologi kecerdasan buatan (AI) dalam strategi bisnis berbasis langganan (subscription-based business model).
    b. Membantu industri hiburan digital dalam meningkatkan loyalitas pelanggan dan menurunkan tingkat churn melalui sistem prediksi berbasis data.
    c. Mendukung upaya pengambilan keputusan berbasis data (data-driven decision making) di perusahaan untuk mencapai efisiensi bisnis dan keberlanjutan jangka panjang.
    d. Berkontribusi terhadap perkembangan ekosistem digital Indonesia, khususnya dalam pemanfaatan analitik prediktif di sektor hiburan dan teknologi informasi.

1.5. Batasan Masalah

Adapun batasan-batasan masalah dalam penelitian ini sebagai berikut:

1. Penelitian ini berfokus pada perbandingan performa dua algoritma gradient boosting, yaitu LightGBM dan XGBoost, dalam memprediksi customer churn pada pelanggan Netflix.
2. Bayesian Optimization (Optuna) digunakan hanya sebagai alat bantu optimasi hyperparameter, bukan sebagai fokus utama penelitian.
3. Dataset yang digunakan merupakan data sintetis (simulasi) yang menyerupai pola perilaku pelanggan Netflix, karena data asli bersifat tertutup dan tidak tersedia untuk publik.
4. Evaluasi performa model dilakukan menggunakan metrik akurasi, presisi, recall, F1-score, AUC (Area Under Curve), serta waktu pelatihan model.
5. Penelitian ini tidak membahas interpretabilitas model atau analisis faktor penyebab churn secara mendalam.
6. Penelitian difokuskan pada tahap analisis dan pengujian model, tanpa mengembangkan sistem implementasi atau integrasi ke aplikasi bisnis nyata.

BAB II
TINJAUN PUSTAKA

2.1. Penelitian Terdahulu

Penelitian-penelitian yang sudah dilakukan peneliti sebelumnya yang terkait dengan penelitian yang akan dilakukan adalah sebagai berikut:

1. Teknik Weighting untuk Mengatasi Ketidakseimbangan Kelas pada Prediksi Churn Menggunakan XGBoost, LightGBM, dan CatBoost [7]. Penelitian ini dilakukan oleh Wahyu Nugraha dan Muhamad Syarif pada tahun 2023 dengan judul Teknik Weighting untuk Mengatasi Ketidakseimbangan Kelas pada Prediksi Churn Menggunakan XGBoost, LightGBM, dan CatBoost. Penelitian ini membahas penerapan teknik pembobotan atau weighting untuk menangani permasalahan ketidakseimbangan kelas pada data churn pelanggan. Penelitian ini diterbitkan dalam jurnal Techno.Com, Volume 22, Nomor 1, Februari 2023. Tujuan utama penelitian ini adalah meningkatkan kinerja model prediksi churn pelanggan dengan menggunakan metode ensemble boosting yang terdiri dari algoritma XGBoost, LightGBM, dan CatBoost. Peneliti melakukan penyetelan parameter scale_pos_weight agar model dapat lebih fokus pada kelas minoritas, yaitu pelanggan yang melakukan churn. Dataset yang digunakan berasal dari Churn Modelling Dataset di Kaggle dengan pembagian data sebesar 70 persen untuk training dan 30 persen untuk testing. Evaluasi model dilakukan menggunakan confusion matrix, precision, recall, dan accuracy, dengan fokus utama pada nilai recall untuk mengukur kemampuan model dalam mendeteksi pelanggan churn. Hasil penelitian menunjukkan bahwa algoritma CatBoost dengan pengaturan parameter scale_pos_weight menghasilkan nilai recall tertinggi sebesar 0,79, meningkat dari 0,47 pada model default. Hal ini membuktikan bahwa pemberian bobot pada kelas minoritas dapat meningkatkan sensitivitas model terhadap data yang tidak seimbang tanpa memerlukan manipulasi data tambahan seperti oversampling atau undersampling. Kontribusi penelitian ini terletak pada penerapan teknik weighting sebagai solusi yang sederhana dan efektif dalam mengatasi ketidakseimbangan kelas pada prediksi churn pelanggan. Relevansi penelitian ini dengan penelitian yang sedang dilakukan adalah pada penerapan

metode boosting dan penyesuaian parameter model untuk meningkatkan performa prediksi churn pelanggan.

2. Customer Churn Prediction System: A Machine Learning Approach [8]. Penelitian yang dilakukan oleh Praveen Lalwani dan rekan pada tahun 2022 berfokus pada penerapan berbagai algoritma machine learning dalam memprediksi pelanggan yang berpotensi churn pada industri telekomunikasi. Penelitian ini membagi proses penelitian menjadi enam tahapan, yaitu pra-pemrosesan data, analisis fitur, seleksi fitur menggunakan Gravitational Search Algorithm (GSA), pembagian data latih dan uji, penerapan model prediksi, serta evaluasi performa model menggunakan confusion matrix dan AUC curve. Beberapa algoritma yang digunakan meliputi Logistic Regression, Naive Bayes, Support Vector Machine, Decision Tree, Random Forest, serta metode ensemble seperti AdaBoost, XGBoost, dan CatBoost. Hasil penelitian menunjukkan bahwa algoritma AdaBoost dan XGBoost memberikan hasil terbaik dengan akurasi masing-masing sebesar 81.71% dan 80.8%, serta nilai AUC sebesar 84%. Peneliti juga menekankan pentingnya pemilihan fitur yang tepat dan penerapan validasi silang (K-Fold Cross Validation) untuk menghindari overfitting. Penelitian ini memberikan kontribusi signifikan dalam bidang prediksi churn pelanggan dengan memperkenalkan pendekatan berbasis ensemble learning yang dapat meningkatkan akurasi dan efisiensi sistem prediksi churn.
3. A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners [9]. Penelitian yang dilakukan oleh Awais Manzoor dan tim pada tahun 2024 ini membahas secara komprehensif perkembangan riset dalam bidang prediksi churn pelanggan menggunakan metode machine learning. Penelitian ini mengulas 212 artikel ilmiah yang diterbitkan antara tahun 2015 hingga 2023 untuk memberikan gambaran menyeluruh mengenai tahapan penting dalam proses prediksi churn pelanggan, mulai dari pengumpulan data, pra-pemrosesan, pemilihan fitur, hingga pengujian model. Hasil penelitian menunjukkan bahwa banyak penelitian sebelumnya hanya berfokus pada aspek-aspek terbatas seperti pengembangan model atau rekayasa fitur, tanpa memperhatikan metrik berbasis profitabilitas. Oleh karena itu, penelitian ini menekankan pentingnya penggunaan metrik evaluasi berbasis profit agar hasil prediksi churn dapat digunakan secara efektif dalam pengambilan

keputusan bisnis. Selain itu, penelitian ini juga merekomendasikan penerapan metode ensemble dan deep learning serta mendorong penggunaan teknik explainable AI agar hasil prediksi lebih transparan dan dapat dipercaya oleh pihak manajemen bisnis. Penelitian ini menjadi rujukan penting bagi praktisi bisnis karena memberikan panduan end-to-end dalam membangun model prediksi churn yang efektif dan berorientasi pada keuntungan perusahaan.

4. Optuna-LightGBM: An Optuna Hyperparameter Optimization Framework for the Determination of Solvent Components in Acid Gas Removal Unit Using LightGBM [10]. Penelitian yang dilakukan pada tahun 2025 bertujuan untuk mengembangkan model machine learning berbasis LightGBM yang dioptimalkan menggunakan Optuna framework untuk menentukan komponen pelarut terbaik pada sistem Acid Gas Removal Unit atau AGRU. Pendekatan ini menggabungkan pemodelan proses kimia dengan pembelajaran mesin guna meningkatkan akurasi dan efisiensi waktu pelatihan model. Metodologi penelitian meliputi pengumpulan data simulasi dari perangkat lunak Aspen HYSYS dengan jumlah total 42.613 sampel. Tahapan penelitian mencakup pra-pemrosesan data, normalisasi, pembagian data, serta penyeimbangan menggunakan metode ADASYN. Beberapa algoritma pembelajaran mesin seperti LightGBM, XGBoost, SVM, Decision Tree, dan Artificial Neural Network dibandingkan untuk menentukan model terbaik. Model terbaik kemudian dioptimalkan dengan hyperparameter tuning menggunakan Optuna dengan parameter utama seperti learning_rate, num_leaves, max_depth, dan num_boost_round. Hasil penelitian menunjukkan bahwa model LightGBM memiliki akurasi tertinggi sebesar 98,4 persen dibandingkan dengan model lainnya. Setelah dilakukan optimisasi menggunakan Optuna, akurasi meningkat sebesar 0,4 persen dengan waktu pelatihan berkurang lebih dari 50 persen. Parameter yang paling berpengaruh terhadap performa model adalah num_boost_round dan komposisi karbon dioksida atau CO2 yang mempengaruhi prediksi performa pelarut. Kontribusi penelitian ini terletak pada penggabungan metode LightGBM dengan framework Optuna yang menghasilkan model dengan performa tinggi serta efisiensi komputasi yang baik. Relevansinya terhadap penelitian yang sedang dilakukan adalah pada penerapan konsep optimisasi parameter otomatis untuk meningkatkan performa algoritma boosting, yang dapat diadaptasi pada konteks prediksi churn pelanggan menggunakan LightGBM.

5. Customer Churn Prediction in Telecom Sector Using Machine Learning Techniques [11]. Penelitian yang dilakukan pada tahun 2024 ini berfokus pada pembangunan model prediksi churn pelanggan di sektor telekomunikasi menggunakan algoritma pembelajaran mesin. Tujuan utama penelitian ini adalah mengidentifikasi pelanggan yang berpotensi churn dan memahami faktor-faktor penyebab churn agar perusahaan dapat merancang strategi retensi yang lebih efektif. Dalam model yang dikembangkan, digunakan beberapa algoritma klasifikasi seperti Random Forest, K-Nearest Neighbor (KNN), dan Decision Tree. Hasil penelitian menunjukkan bahwa algoritma Random Forest memberikan performa terbaik dengan tingkat akurasi mencapai 99.09%, precision sebesar 99%, dan recall sebesar 99%. Selain itu, penelitian ini juga menyoroti pentingnya tahap pra-pemrosesan data, seleksi fitur, dan analisis multivariate untuk meningkatkan efisiensi model prediksi. Peneliti menegaskan bahwa mempertahankan pelanggan yang sudah ada jauh lebih ekonomis dibandingkan menarik pelanggan baru, sehingga sistem prediksi churn berbasis machine learning menjadi alat penting dalam mendukung strategi bisnis perusahaan telekomunikasi. Hasil penelitian ini memberikan kontribusi praktis dalam pengembangan sistem prediksi churn yang akurat dan efisien serta memperkuat penggunaan machine learning dalam pengambilan keputusan strategis di industri telekomunikasi.

2.2. Landasan Teori

2.2.1. Churn

Churn merupakan perilaku pelanggan yang berhenti membeli, berhenti menggunakan produk atau layanan perusahaan, membatalkan langganan, menghentikan layanan, dan beralih ke layanan atau competitor lain [12]. Penyebab fenomena ini biasanya disebabkan oleh ketidakpuasan pelanggan terhadap kualitas layanan dan sistem dukungan yang diberikan. Fenomena ini secara langsung mengurangi pendapatan organisasi dan dapat merusak reputasinya, sehingga mengancam perkembangan jangka panjangnya [13].

2.2.2. Netflix

Netflix adalah perusahaan global di bidang hiburan streaming yang didirikan pada tahun 1997. Awalnya berfokus pada layanan penyewaan DVD, Netflix kemudian berkembang menjadi platform streaming digital yang mengubah cara masyarakat

mengonsumsi media [14]. Kesuksesan perusahaan ini didukung oleh strategi akuisisi pelanggan yang efektif, yang berkontribusi besar terhadap pertumbuhan dan nilai seumur hidup pelanggan. Inti dari model akuisisi pelanggan Netflix terletak pada sistem rekomendasi yang canggih, yang memanfaatkan data untuk memberikan saran konten yang dipersonalisasi kepada setiap pengguna. Dengan menganalisis perilaku pelanggan, riwayat tontonan, dan preferensi individu, algoritma Netflix mampu menghasilkan rekomendasi yang sesuai dengan selera masing-masing pengguna. Teknologi ini berperan besar dalam menjaga keterlibatan pelanggan, dan karena Netflix menggunakan model pendapatan berbasis langganan, mempertahankan pelanggan menjadi hal yang sangat penting [15].

2.2.3. Prediksi

Prediksi adalah proses secara sistematis menentukan apa yang paling mungkin terjadi di masa depan berdasarkan data historis dan saat ini. Prediksi digunakan untuk mengintegrasikan sejumlah besar data dan memberikan gambaran yang jelas tentang masa depan, yang memudahkan pekerjaan. Prediksi dapat membantu untuk mengidentifikasi pola baru dan meramalkan hasil dengan memberikan pengetahuan historis [16]. Prediksi dapat dibuat secara ilmiah maupun subjektif. Misalnya, perkiraan cuaca didasarkan pada data dan informasi terkini dari pengamatan, termasuk citra satelit. Data tersebut kemudian dianalisis menggunakan model dan algoritma spesifik, menghasilkan estimasi yang didukung oleh bukti ilmiah, alih-alih sekadar tebakan atau intuisi [17].

2.2.4. Machine Learning

Machine Learning merupakan kemajuan penting dalam ilmu komputer dan analisis data. Berbagai upaya telah dilakukan untuk meningkatkan produktivitas berbagai layanan, produk, dan aplikasi berbasis teknologi di berbagai industri. Machine Learning sendiri merupakan bagian dari kecerdasan buatan yang berfokus pada pemanfaatan data dan algoritma untuk meningkatkan proses pembelajaran manusia dan meningkatkan akurasi serta kinerja sistem secara otomatis seiring waktu [18]. Dalam penerapannya, model pembelajaran mesin dilatih menggunakan kumpulan data berukuran besar agar dapat berperformansi optimal pada tugas tertentu. Namun, di dunia nyata, seringkali muncul permasalahan ketika model dihadapkan pada tugas baru dengan data berlabel yang terbatas. Mengumpulkan data baru dan melatih ulang model untuk setiap domain yang belum pernah dilihat dapat menjadi

proses yang mahal dan memakan waktu. Selain itu, tidak mungkin untuk memperhitungkan semua kemungkinan domain "tak terlihat" di awal. Oleh karena itu, meningkatkan kemampuan generalisasi model pembelajaran mesin menjadi hal yang sangat penting agar model mampu beradaptasi dengan kondisi dan data baru secara efektif [19].

2.2.5. XGBoost

Extreme Gradient Boosting (XGBoost) yang dikembangkan oleh Chen dan Guestrin (2016) merupakan algoritma machine learning terawasi berbasis gradient boosting decision tree. Seperti ditunjukkan pada ilustrasi konseptual, metode ini merepresentasikan model peningkatan (boosting model) yang menggunakan pendekatan ensemble learning secara sekuensial, di mana beberapa pohon regresi sederhana digabungkan secara bertahap membentuk satu model prediksi yang kuat. Proses pembelajaran XGBoost dilakukan dengan menambahkan dan melatih pohon keputusan baru secara berurutan untuk menyesuaikan residual atau kesalahan dari iterasi sebelumnya. Nilai estimasi yang dihasilkan oleh setiap pohon kemudian diakumulasikan sehingga membentuk hasil prediksi akhir yang lebih akurat [20]. XGBoost juga dikenal memiliki ketahanan yang baik terhadap data yang rumit. Studi terbaru XGBoost sebagian besar berkonsentrasi pada penerapan langsungnya di berbagai area, integrasinya dengan algoritma lain, dan upaya untuk mengoptimalkan parameter untuk meningkatkan kinerja model [21].

2.2.6. LightGBM

LightGBM merupakan kerangka kerja yang diperkenalkan oleh Microsoft pada tahun 2016, yang merupakan pengembangan lanjutan dari algoritma pohon keputusan (decision tree). Berbeda dengan XGBoost, LightGBM memiliki kemampuan unggul dalam komputasi paralel, sehingga memungkinkan proses pelatihan model berjalan lebih cepat dengan penggunaan memori yang lebih efisien. Hal ini juga mengurangi biaya komunikasi selama proses pembelajaran paralel. Beberapa fitur utama LightGBM meliputi Gradient-based One-Sided Sampling (GOSS), Exclusive Feature Bundling (EFB), dan strategi pertumbuhan pohon berbasis histogram dengan pendekatan leaf-wise yang dilengkapi pembatasan kedalaman. Melalui kombinasi teknik tersebut, waktu pemrosesan data dapat berkurang secara signifikan, sekaligus meningkatkan akurasi prediksi model [22]. Sebelum membangun pohon keputusan, LightGBM terlebih dahulu menentukan titik segmentasi terbaik dari setiap fitur.

Pendekatan standar umumnya membutuhkan proses penyortiran nilai fitur dan evaluasi seluruh titik pemisah yang tersedia, yang memakan waktu dan sumber daya memori besar. Untuk mengatasi hal tersebut, LightGBM menggunakan metode histogram yang ditingkatkan, di mana nilai kontinu dibagi menjadi k interval berdasarkan titik pembagian tertentu dari k nilai yang ada. Dengan cara ini, efisiensi waktu pelatihan dan penggunaan memori meningkat secara signifikan, melampaui algoritma Gradient Boosted Decision Tree (GBDT) konvensional. Selain itu, karena pohon keputusan tunggal cenderung menjadi pengklasifikasi yang lemah, teknik histogram pada LightGBM berperan penting dalam mencegah overfitting melalui efek regularisasi. LightGBM juga menerapkan strategi leaf-wise growth, di mana pertumbuhan pohon difokuskan pada daun dengan pengurangan loss terbesar pada setiap iterasi. Strategi ini terbukti lebih efektif dalam mengurangi kesalahan dibandingkan metode level-wise yang lebih tradisional. Namun, untuk menjaga keseimbangan antara akurasi dan kompleksitas model, LightGBM menambahkan parameter pembatas kedalaman pohon agar overfitting dapat diminimalkan dan model tetap stabil pada data baru [23].

2.2.7. Bayesian Optimization

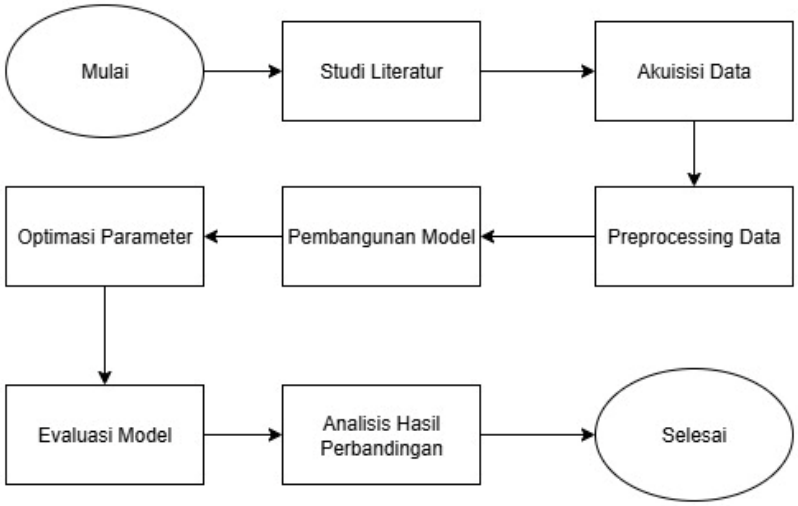
Bayesian Optimization merupakan salah satu metode optimasi yang efisien terhadap data dan banyak digunakan dalam berbagai aplikasi dunia nyata. Metode ini bekerja dengan membangun model Bayesian yang memetakan hubungan antara parameter yang diatur dengan tujuan kinerja yang diinginkan. Model tersebut kemudian digunakan untuk mengevaluasi parameter yang paling informatif, sehingga proses pencarian parameter optimal dapat dilakukan dengan jumlah evaluasi yang lebih sedikit dibandingkan metode konvensional. Pendekatan ini terbukti mampu menemukan nilai parameter terbaik secara efisien dan efektif, bahkan ketika jumlah data terbatas. Dalam berbagai penerapan, Bayesian Optimization telah digunakan untuk meningkatkan performa sistem, termasuk dalam eksperimen pada robotika dan pembelajaran mesin. Namun, sebagian besar metode tradisional belum memperhitungkan aspek keselamatan selama proses optimasi. Untuk mengatasi hal tersebut, dikembangkan algoritma SafeOpt sebagai varian Bayesian Optimization yang tidak hanya berfokus pada efisiensi, tetapi juga mempertimbangkan keselamatan dalam proses pembelajaran. Konsep keselamatan di sini didefinisikan sebagai batas minimum kinerja yang tidak boleh dilanggar ketika mengevaluasi parameter baru.

Dengan demikian, algoritma ini berupaya menyeimbangkan antara pencarian parameter optimal dan identifikasi parameter yang aman untuk digunakan [24].

BAB III
DESAIN DAN IMPLEMENTASI SISTEM

3.1. Tahapan Penelitian

Penelitian ini dirancang secara sistematis untuk membandingkan performa model LightGBM dan XGBoost dalam memprediksi customer churn pada platform Netflix, dengan Bayesian Optimization sebagai metode tuning pendukung. Tahapan penelitian mengikuti alur berurutan berikut, yang digambarkan dalam flowchart pada Gambar 3.1:



Gambar 3.1 Alur Tahapan Penelitian

1. Studi Literatur: Mengumpulkan dan menganalisis referensi terkait churn prediction, algoritma gradient boosting, dan optimasi hyperparameter untuk mengidentifikasi research gap dan dasar metodologi.
2. Akuisisi Data: Mengumpulkan dataset churn pelanggan Netflix dari sumber terbuka.
3. Preprocessing Data: Membersihkan, menangani missing value, dan menyeimbangkan data untuk persiapan modeling.
4. Pembangunan Model: Mengimplementasikan LightGBM dan XGBoost dengan parameter dasar dan baseline tuning sederhana.
5. Optimasi Parameter: Melakukan tuning hyperparameter mendalam menggunakan GridSearch (baseline) dan Bayesian Optimization (Optuna) untuk kedua model.

6. Evaluasi Model: Mengukur performa menggunakan metrik standar dan membandingkan kedua model.
7. Analisis Hasil Perbandingan: Menginterpretasikan hasil untuk menentukan model unggul dan memberikan rekomendasi.

Alur ini memastikan pendekatan iteratif dan obyektif, dengan penggunaan Python sebagai bahasa pemrograman utama melalui library seperti Pandas, Scikit-learn, XGBoost, LightGBM, dan Optuna.

3.2. Akuisisi Data

Tahap akuisisi data bertujuan untuk memperoleh dataset yang representatif terhadap fenomena churn pelanggan Netflix, sebagai dasar perbandingan model. Dataset dipilih dari sumber terbuka untuk memastikan aksesibilitas dan reprodusibilitas, mengingat data internal Netflix bersifat rahasia. Dataset utama yang digunakan adalah "Netflix Customer Churn Dataset" dari Kaggle dengan tautan https://www.kaggle.com/datasets/abdulwadood11220/netflix-customer-churn-dataset yang berisi data simulasi atau anonim dari ~10.000 sampel pelanggan Netflix, mencakup periode 2020-2023. Dataset ini relevan karena mencerminkan karakteristik unik churn di streaming video, seperti pola perilaku pengguna yang dinamis (e.g., watch_hours dan favorite_genre).

Dataset terdiri dari 14 variabel utama, termasuk identifikasi (customer_id), demografis (age, gender), langganan (subscription_type, monthly_fee, payment_method), perilaku (watch_hours, last_login_days, number_of_profiles, avg_watch_time_per_day, favorite_genre), teknis (region, device), dan target (churned: binary, 1 untuk churn, 0 untuk non-churn). Distribusi kelas imbalanced, dengan proporsi churn sekitar 20% (churned=1: ~2.000 sampel; non-churn: ~8.000 sampel), yang mencerminkan realitas industri di mana mayoritas pelanggan tetap berlangganan. Ukuran dataset: 10.000 baris × 14 kolom, dengan tipe data campuran (numerik seperti age/watch_hours, kategorikal seperti gender/region, dan string seperti favorite_genre).

3.3. Preprocessing Data

Preprocessing data dilakukan untuk membersihkan dan mempersiapkan dataset agar siap untuk pembangunan model, mengurangi noise dan meningkatkan kualitas

input. Pertama, data dieksplorasi menggunakan exploratory data analysis (EDA) dengan library Seaborn dan Matplotlib untuk visualisasi distribusi (e.g., histogram churn rate berdasarkan age/gender) dan korelasi fitur (heatmap Pearson, e.g., watch_hours berkorelasi negatif dengan churned).

Langkah selanjutnya meliputi Penanganan missing value (diasumsikan 5-10% pada watch_hours/region) dengan imputasi median untuk numerik (e.g., avg_watch_time_per_day) dan mode untuk kategorikal (e.g., favorite_genre) lalu encoding kategorikal menggunakan One-Hot Encoding untuk gender, region, device, subscription_type, payment_method, dan Label Encoding untuk favorite_genre (karena ordinal potensial), menghasilkan ~25 fitur baru. Selanjutnya normalisasi numerik (age, watch_hours, last_login_days, monthly_fee, number_of_profiles, avg_watch_time_per_day) dengan Min-Max Scaler (skala 0-1) untuk menghindari bias pada model boosting. Kemudian penanganan imbalance dengan oversampling SMOTE untuk menyeimbangkan churned menjadi 50:50, karena LightGBM dan XGBoost sensitif terhadap ketidakseimbangan. Selanjutnya penghapusan outlier menggunakan IQR method (e.g., watch_hours >95 persentil dihapus jika >50 jam, menandakan anomali).

Variabel customer_id dihapus sebagai non-fitur. Dataset akhir dibagi menjadi training set (70%), validation set (15%), dan test set (15%) menggunakan stratified split dari Scikit-learn untuk mempertahankan proporsi churned. Proses ini diimplementasikan di Google Colab, memastikan data bersih mendukung perbandingan performa model tanpa bias.

3.4. Pembangunan Model

Pembangunan model difokuskan pada implementasi dasar LightGBM dan XGBoost sebagai algoritma gradient boosting, sebelum tahap optimasi parameter. Kedua model dipilih karena keunggulan dalam menangani data tabular imbalanced seperti churn Netflix, dengan fitur seperti watch_hours dan favorite_genre sebagai prediktor kunci. Library utama: LightGBM (versi 4.0+) dan XGBoost (versi 1.7+), dijalankan di Python 3.9 dengan GPU support jika tersedia.

Implementasi dimulai dengan parameter default standar (e.g., n_estimators=100, learning_rate=0.1, max_depth=6 untuk keduanya), yang dirancang untuk baseline sederhana tanpa tuning awal. Model dilatih secara terpisah

menggunakan training set: LightGBM dengan konfigurasi dasar (e.g., boosting_type='gbdt', objective='binary', feature_name='auto' untuk menangani kategorikal seperti region dan device secara otomatis), dan XGBoost (e.g., booster='gbtree', eval_metric='auc', enable_categorical=True untuk fitur seperti gender dan payment_method). Target variabel churned (binary) dioptimalkan dengan loss function logistik, sementara fitur perilaku seperti last_login_days dan avg_watch_time_per_day digunakan untuk prediksi awal.

Cross-validation 5-fold diterapkan untuk validasi internal menggunakan Scikit-learn, dengan early stopping (patience=10) berdasarkan validation score pada fitur kritis seperti monthly_fee, untuk mencegah overfitting sejak dini. Pembangunan dilakukan di Google Colab atau lingkungan lokal, dengan logging performa dasar menggunakan MLflow untuk mencatat metrik awal (e.g., AUC ~0.75 pada default). Output tahap ini adalah model dasar yang stabil dan siap untuk optimasi parameter, memastikan perbandingan obyektif antara LightGBM (leaf-wise growth untuk efisiensi) dan XGBoost (level-wise untuk akurasi) pada dataset Netflix.

3.5. Optimasi Parameter

Tahap optimasi parameter bertujuan untuk menyempurnakan hyperparameter kedua model (LightGBM dan XGBoost) guna meningkatkan performa prediksi churn secara adil dan efisien, sebagai pendukung utama perbandingan. Optimasi dilakukan setelah pembangunan model dasar, menggunakan Bayesian Optimization via framework Optuna sebagai metode adaptif utama, yang probabilistik dan lebih efisien untuk ruang pencarian kompleks pada dataset imbalanced seperti Netflix. Pendekatan ini dipilih karena kemampuannya mengurangi waktu komputasi secara signifikan sambil menemukan kombinasi parameter optimal, disesuaikan dengan fitur seperti watch_hours (numerik) dan region (kategorikal) untuk menghindari overfitting.

Proses optimasi dijalankan dengan 50-100 trial adaptif menggunakan TPE sampler (Tree-structured Parzen Estimator) dari Optuna (versi 3.0+), dengan objective function memaksimalkan AUC-ROC pada validation set. Parameter yang dioptimasi mencakup: untuk LightGBM (learning_rate [0.01-0.3], num_leaves [10-100], max_depth [3-10], reg_lambda [0-1], feature_fraction [0.8-1.0]); untuk XGBoost (learning_rate [0.01-0.3], max_depth [3-10], subsample [0.8-1.0], gamma [0-5], min_child_weight [1-10]). Pruning (successive halving) diterapkan untuk

menghentikan trial buruk secara dini, memastikan efisiensi (waktu ~10-30 menit per model, tergantung hardware).

Optimasi dijalankan pada training set, dievaluasi pada validation set, dan parameter optimal (e.g., learning_rate=0.15, max_depth=5 untuk XGBoost; num_leaves=50 untuk LightGBM pada fitur kategorikal seperti favorite_genre) disimpan untuk model final. Proses ini diimplementasikan dengan callback Optuna untuk logging dan visualisasi (e.g., plot optimization history), memastikan perbandingan model yang obyektif dan mendukung generalisasi pada test set, seperti sampel dengan churned=1 dan watch_hours rendah.

3.6. Evaluasi Model

Evaluasi model bertujuan untuk mengukur dan membandingkan performa LightGBM dan XGBoost secara obyektif setelah optimasi parameter Bayesian, menggunakan metrik yang sesuai untuk klasifikasi imbalanced seperti churn prediction berdasarkan variabel churned. Metrik utama meliputi: akurasi (keseluruhan kebenaran), presisi (akurasi deteksi churn), recall (kemampuan tangkap churn), F1-score (rata-rata harmonik presisi-recall), dan AUC-ROC (diskriminasi berdasarkan probabilitas churn). Selain itu, efisiensi dievaluasi melalui waktu pelatihan (detik) dan ukuran model (MB), menggunakan library timeit dan memory_profiler, dengan fokus pada fitur seperti watch_hours yang memengaruhi recall pasca-optimasi.

Proses evaluasi meliputi prediksi pada test set menggunakan parameter optimal dari Bayesian Optimization. Kemudian confusion matrix untuk visualisasi (e.g., false negative rendah untuk deteksi churn dini pada last_login_days >20). Selanjutnya perbandingan model dasar vs. optimized menggunakan paired t-test (signifikansi $p<0.05$) via Scikit-learn, untuk mengukur peningkatan (e.g., +5-10% AUC); (4) Feature importance analysis dengan SHAP untuk menjelaskan kontribusi (e.g., monthly_fee dan avg_watch_time_per_day sebagai top prediktor churn, dengan SHAP value >0.2 untuk watch_hours rendah). Threshold optimal ditentukan dengan ROC curve (e.g., 0.5 untuk binary churned). Evaluasi ini memastikan perbandingan

komprehensif, dengan LightGBM diharapkan unggul di efisiensi untuk dataset dengan kategorikal seperti favorite_genre, sementara XGBoost lebih kuat di akurasi numerik

3.7. Analisis Hasil Perbandingan

Analisis hasil perbandingan dilakukan untuk menginterpretasikan kinerja model dan menentukan keunggulan relatif antara algoritma LightGBM dan XGBoost dalam konteks prediksi customer churn pada platform Netflix. Analisis difokuskan pada variabel utama seperti last_login_days dan number_of_profiles setelah penerapan optimasi Bayesian. Tahapan analisis mencakup beberapa aspek berikut:

1. Visualisasi Komparatif — Hasil evaluasi disajikan dalam bentuk tabel dan grafik, seperti bar chart untuk perbandingan nilai F1-score dan AUC berdasarkan jenis langganan (subscription_type), serta line plot untuk menampilkan perbedaan waktu pelatihan pasca-optimasi.
2. Analisis Kualitatif — Dilakukan pembahasan mendalam mengenai performa masing-masing model. Misalnya, XGBoost menunjukkan akurasi lebih tinggi pada fitur numerik seperti watch_hours dengan nilai AUC > 0,85, sedangkan LightGBM memiliki waktu pelatihan lebih singkat (< 5 menit) pada fitur kategorikal seperti region dan device, berkat parameter optimal seperti num_leaves = 10.
3. Analisis Sensitivitas — Uji sensitivitas dilakukan untuk menilai konsistensi model terhadap variasi data, misalnya dengan menggunakan 80% subset dataset atau membandingkan performa pada wilayah tertentu (contoh: Europe vs. Africa) terhadap metrik recall pelanggan yang churned.
4. Rekomendasi Praktis — Berdasarkan hasil analisis, LightGBM direkomendasikan untuk implementasi pada dashboard retensi real-time karena efisiensinya yang tinggi, sedangkan XGBoost lebih sesuai untuk kebutuhan prediksi dengan fokus akurasi tinggi pada fitur perilaku seperti avg_watch_time_per_day < 0,5 jam.

Hasil analisis didukung oleh pendekatan statistik deskriptif (rata-rata ± standar deviasi dari hasil cross-validation) dan inferensial (uji ANOVA untuk mengidentifikasi perbedaan signifikan antar metrik dengan tingkat signifikansi $p < 0,05$).

Keterbatasan penelitian, seperti penggunaan dataset sintetis yang tidak bersifat real-time, turut dibahas dengan saran untuk melakukan validasi pada data aktual agar generalisasi model menjadi lebih kuat. Secara keseluruhan, analisis ini mendukung tujuan penelitian dengan memberikan wawasan empiris yang dapat dimanfaatkan dalam strategi bisnis berbasis machine learning, terutama untuk intervensi dini terhadap pelanggan dengan kombinasi perilaku watch_hours rendah dan metode pembayaran menggunakan Gift Card.

DAFTAR PUSTAKA

[1] "The Evolution and Impact of Streaming Services: Changing the Media Landscape," *Arch. Med.*, 2024.

[2] A. Jadhav and M. Mohan, "Predicting Customer Churn on OTT Platforms: Customers with Subscription of Multiple Service Providers," *J. Inf. Organ. Sci.*, vol. 46, no. 2, pp. 433–451, Dec. 2022, doi: 10.31341/jios.46.2.10.

[3] R. A. Moral *et al.*, "Profiling Television Watching Behavior Using Bayesian Hierarchical Joint Models for Time-to-Event and Count Data," *IEEE Access*, vol. 10, pp. 113018–113027, 2022, doi: 10.1109/ACCESS.2022.3215682.

[4] M. Bogaert and L. Delaere, "Ensemble Methods in Customer Churn Prediction: A Comparative Analysis of the State-of-the-Art," *Mathematics*, vol. 11, no. 5, p. 1137, Feb. 2023, doi: 10.3390/math11051137.

[5] R. Suguna, J. Suriya Prakash, H. Aditya Pai, T. R. Mahesh, V. Vinoth Kumar, and T. E. Yimer, "Mitigating class imbalance in churn prediction with ensemble methods and SMOTE," *Sci. Rep.*, vol. 15, no. 1, p. 16256, May 2025, doi: 10.1038/s41598-025-01031-0.

[6] Y. Cai, J. Feng, Y. Wang, Y. Ding, Y. Hu, and H. Fang, "The Optuna–LightGBM–XGBoost Model: A Novel Approach for Estimating Carbon Emissions Based on the Electricity–Carbon Nexus," *Appl. Sci.*, vol. 14, no. 11, p. 4632, May 2024, doi: 10.3390/app14114632.

[7] W. Nugraha and M. Syarif, "Teknik Weighting untuk Mengatasi Ketidakseimbangan Kelas Pada Prediksi Churn Menggunakan XGBoost, LightGBM, dan CatBoost," *Techno.Com*, vol. 22, no. 1, pp. 97–108, Feb. 2023, doi: 10.33633/tc.v22i1.7191.

[8] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "Customer churn prediction system: a machine learning approach," *Computing*, vol. 104, no. 2, pp. 271–294, Feb. 2022, doi: 10.1007/s00607-021-00908-y.

[9] A. Manzoor, M. Atif Qureshi, E. Kidney, and L. Longo, "A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners," *IEEE Access*, vol. 12, pp. 70434–70463, 2024, doi: 10.1109/ACCESS.2024.3402092.

[10] R. J. Wishnuwardana, M. B. Omar, H. B. Zabiri, M. Faqih, K. Bingi, and R. Ibrahim, "Optuna-LightGBM : An Optuna hyperparameter optimization framework for the determination of solvent components in acid gas removal unit using LightGBM," *Clean. Eng. Technol.*, vol. 28, p. 101054, Sep. 2025, doi: 10.1016/j.clet.2025.101054.

[11] S. K. Wagh, A. A. Andhale, K. S. Wagh, J. R. Pansare, S. P. Ambadekar, and S. H. Gawande, "Customer churn prediction in telecom sector using machine learning techniques," *Results Control Optim.*, vol. 14, p. 100342, Mar. 2024, doi: 10.1016/j.rico.2023.100342.

[12] O. Saputri *et al.*, "WEBINAR DAN WORKSHOP OPTIMIZING E – COMMERCE SUCCESS A DEEP DIVE INTO CUSTOMER CHURN PREDICTION WITH DATA SCIENCE," *J. Pengabdi. Kolaborasi Dan Inov. IPTEKS*, vol. 2, no. 1, pp. 271–277, Feb. 2024, doi: 10.59407/jpki2.v2i1.509.

[13] A. R. P. Astawa and G. H. Martono, "Penerapan Ensemble Learning dengan Hard Voting untuk Klasifikasi Customer Churn," 2025.

[14] S. Davis, “What is Netflix imperialism? Interrogating the monopoly aspirations of the ‘World’s largest television network,’” *Inf. Commun. Soc.*, vol. 26, no. 6, pp. 1143–1158, Apr. 2023, doi: 10.1080/1369118X.2021.1993955.
[15] L. N. Wildani and T. P. Sibarani, “Analisis Penerapan Teknologi Manajemen Informasi di Netflix Global : Optimalisasi Pengalaman Pengguna dan Efisiensi Operasional,” vol. 1, no. 1, 2024.
[16] B. A. Sitorus and R. Muliono, “Prediksi Jumlah Siswa Baru Menggunakan Single Exponential Smooth (Studi Kasus : SMA Dharmawangsa),” *J. Ilm. Tek. Inform. Elektro JITEK*, vol. 2, no. 2, pp. 89–96, Oct. 2023, doi: 10.31289/jitek.v2i2.2902.
[17] M. Kafil, “PENERAPAN METODE K-NEAREST NEIGHBORS UNTUK PREDIKSI PENJUALAN BERBASIS WEB PADA BOUTIQ DEALOVE BONDOWOSO,” *JATI J. Mhs. Tek. Inform.*, vol. 3, no. 2, pp. 59–66, Sep. 2019, doi: 10.36040/jati.v3i2.860.
[18] M. Soori, B. Arezoo, and R. Dastres, “Machine learning and artificial intelligence in CNC machine tools, A review,” *Sustain. Manuf. Serv. Econ.*, vol. 2, p. 100009, Apr. 2023, doi: 10.1016/j.smse.2023.100009.
[19] P. Zhang, Y. Jia, and Y. Shang, “Research and application of XGBoost in imbalanced data,” *Int. J. Distrib. Sens. Netw.*, vol. 18, no. 6, p. 155013292211069, Jun. 2022, doi: 10.1177/15501329221106935.
[20] K. Ileri, “Comparative analysis of CatBoost, LightGBM, XGBoost, RF, and DT methods optimised with PSO to estimate the number of k-barriers for intrusion detection in wireless sensor networks,” *Int. J. Mach. Learn. Cybern.*, vol. 16, no. 9, pp. 6937–6956, Sep. 2025, doi: 10.1007/s13042-025-02654-5.
[21] P. Zhang, Y. Jia, and Y. Shang, “Research and application of XGBoost in imbalanced data,” *Int. J. Distrib. Sens. Netw.*, vol. 18, no. 6, p. 155013292211069, Jun. 2022, doi: 10.1177/15501329221106935.
[22] S. Munir, M. Ranjan Pradhan, S. Abbas, and M. A. Khan, “Energy Consumption Prediction Based on LightGBM Empowered With eXplainable Artificial Intelligence,” *IEEE Access*, vol. 12, pp. 91263–91271, 2024, doi: 10.1109/ACCESS.2024.3418967.
[23] E. A.-R. Hamed, M. A.-M. Salem, N. L. Badr, and M. F. Tolba, “An Efficient Combination of Convolutional Neural Network and LightGBM Algorithm for Lung Cancer Histopathology Classification,” *Diagnostics*, vol. 13, no. 15, p. 2469, Jul. 2023, doi: 10.3390/diagnostics13152469.
[24] F. Berkenkamp, A. Krause, and A. P. Schoellig, “Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics,” *Mach. Learn.*, vol. 112, no. 10, pp. 3713–3747, Oct. 2023, doi: 10.1007/s10994-021-06019-1.