BAB I
PENDAHULUAN

1.1 Latar Belakang

Industri layanan streaming musik telah mengalami pertumbuhan yang pesat dan revolusioner dalam beberapa tahun terakhir, mengubah secara signifikan cara masyarakat mengakses dan mengonsumsi musik. Dengan meningkatnya akses terhadap internet dan penggunaan perangkat pintar, layanan streaming musik kini telah menjadi bagian tak terpisahkan dari kehidupan sehari-hari. Platform seperti Spotify, Apple Music, dan YouTube Music tidak hanya menggantikan peran radio konvensional, tetapi juga menciptakan ekosistem musik digital yang lebih fleksibel, personal, dan sesuai dengan preferensi pengguna.

Menurut data MIDiA Research (2024), pada kuartal ketiga tahun 2023, jumlah pelanggan musik berbayar secara global mencapai 713,4 juta, meningkat sebanyak 90 juta pelanggan dibandingkan tahun sebelumnya. Dari jumlah tersebut, Spotify memimpin pasar dengan pangsa sebesar 31,7%, diikuti oleh Apple Music (12,6%), Amazon Music (11,1%), dan YouTube Music (9,7%) [1]. Fakta ini menunjukkan bahwa layanan berbasis internet semakin mendominasi industri musik global. Pengguna kini lebih memilih fleksibilitas dalam menentukan kapan dan bagaimana mereka menikmati musik dibandingkan dengan media tradisional atau pembelian musik fisik.

Namun, di tengah pertumbuhan ini, persaingan antar platform juga semakin ketat. Setiap layanan berlomba-lomba menawarkan katalog konten yang luas serta pengalaman pengguna yang lebih baik guna mempertahankan loyalitas pelanggan. Dalam konteks ini, muncul permasalahan churn pelanggan, yaitu kondisi ketika pelanggan berhenti berlangganan atau berpindah ke platform layanan lain. Jika tidak diantisipasi dengan baik, churn dapat berdampak serius terhadap perusahaan, seperti penurunan pendapatan dan meningkatnya biaya akuisisi pelanggan baru [2].

Untuk meminimalkan risiko churn, diperlukan pendekatan yang mampu memprediksi perilaku pelanggan secara akurat. Salah satu pendekatan yang

relevan adalah penerapan teknologi machine learning, yang mampu mengidentifikasi pola-pola kompleks dalam data pelanggan dan memprediksi kemungkinan mereka akan berhenti menggunakan layanan. Di antara berbagai algoritma yang tersedia, LightGBM dan CatBoost merupakan dua algoritma berbasis boosting yang dikenal efektif untuk tugas klasifikasi seperti prediksi churn.

LightGBM (Light Gradient Boosting Machine) merupakan kerangka kerja pembelajaran mesin berbasis peningkatan gradien yang mengimplementasikan algoritma pohon keputusan. Dibandingkan metode boosting konvensional, LightGBM menawarkan kecepatan pelatihan dan efisiensi memori yang lebih tinggi. Model ini membangun pohon keputusan secara bertahap, di mana setiap pohon bertugas mengoreksi kesalahan dari pohon sebelumnya. Karena kemampuannya dalam menangani data berskala besar, akurasi prediksi yang tinggi, serta interpretabilitasnya yang baik, LightGBM banyak digunakan dalam aplikasi klasifikasi dan regresi [3].

CatBoost, di sisi lain, adalah algoritma boosting yang dikembangkan untuk menyederhanakan pemrosesan data kategorikal. Tidak seperti algoritma lain yang memerlukan proses pra-pemrosesan kompleks seperti one-hot encoding, CatBoost dapat menangani fitur kategorikal secara langsung. Sebagai bagian dari keluarga Gradient Boosting, CatBoost memiliki stabilitas yang lebih tinggi, performa lebih cepat, dan ketahanan terhadap overfitting melalui pendekatan ordered boosting [4].

Penelitian-penelitian sebelumnya menunjukkan efektivitas kedua algoritma ini dalam memprediksi churn pelanggan di berbagai industri. Penelitian oleh [5], misalnya, menunjukkan bahwa CatBoost unggul dalam mengolah fitur kategorikal tanpa langkah pra-pemrosesan tambahan, menghasilkan akurasi sebesar 94% dan recall sebesar 97%. Hal ini menjadikan CatBoost sangat efektif dalam mengidentifikasi pelanggan yang berisiko churn, dengan performa lebih cepat dan stabil dibanding algoritma lain seperti Stochastic Gradient Boosting dan XGBoost.

Sementara itu, penelitian oleh [6] menunjukkan bahwa LightGBM unggul dalam memprediksi churn pelanggan asuransi Lusitania Seguros dengan akurasi

mencapai 98%, mengungguli XGBoost. LightGBM juga menunjukkan performa yang baik dalam metrik recall, precision, dan F1-score pada kelas minoritas, serta mampu memberikan insight melalui analisis fitur yang relevan dengan perilaku pelanggan.

Selain itu, studi yang dilakukan oleh [7] membandingkan berbagai algoritma termasuk LightGBM dan CatBoost dalam memprediksi churn. Hasilnya, LightGBM menunjukkan performa terbaik, terutama setelah penerapan ADASYN untuk mengatasi ketidakseimbangan kelas, dengan AUC sebesar 0,95 dan F1-score sebesar 0,89. Meskipun selisihnya kecil, LightGBM tetap unggul dalam hal efisiensi dan kecepatan, sementara CatBoost lebih stabil dalam pengolahan data kategorikal.

Namun, sebagian besar penelitian belum membahas churn di layanan streaming musik, yang memiliki karakteristik unik. Penelitian ini fokus pada churn pelanggan streaming musik, mengombinasikan *Exploratory Data Analysis* (EDA) dan, jika diperlukan, teknik *Synthetic Minority Over-sampling Technique* (SMOTE) untuk mengatasi ketidakseimbangan data. Selain itu, pencarian parameter terbaik dilakukan menggunakan *GridSearch* untuk setiap metode dan setiap skenario pembagian data, kemudian dipilih kombinasi model dan parameter yang menghasilkan performa terbaik secara keseluruhan.

Dengan pendekatan ini, penelitian bertujuan menghasilkan model prediksi churn yang optimal menggunakan LightGBM dan CatBoost, sekaligus memberikan wawasan mendalam tentang faktor utama yang memengaruhi keputusan pelanggan dalam menggunakan layanan streaming musik.

1.2 Rumusan Masalah

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

1. Bagaimana performa model LightGBM dan CatBoost dalam memprediksi churn pelanggan berdasarkan data historis pada berbagai rasio pembagian data (60:40, 70:30, dan 80:20)?
2. Model prediktif manakah yang menunjukkan performa terbaik dalam memprediksi churn pelanggan berdasarkan pembagian data dan konfigurasi parameter terbaik pada model tersebut?

3. Apa saja faktor-faktor utama yang memengaruhi churn pelanggan berdasarkan hasil analisis fitur dari data historis?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang ditetapkan, penelitian ini bertujuan untuk:

1. Menganalisis performa model LightGBM dan CatBoost dalam memprediksi *churn* pelanggan berdasarkan data historis pada berbagai rasio pembagian data (60:40, 70:30, dan 80:20).
2. Menentukan model prediktif yang menunjukkan performa terbaik dalam memprediksi churn pelanggan berdasarkan pembagian data dan konfigurasi parameter terbaik pada model tersebut.
3. Mengidentifikasi faktor-faktor utama yang memengaruhi *churn* pelanggan berdasarkan hasil analisis fitur dari data historis.

1.4 Manfaat Penelitian

Penelitian ini memberikan manfaat bagi sejumlah pihak, antara lain:

1. Manfaat Akademis
   a. Menambah literatur ilmiah terkait perbandingan performa algoritma LightGBM dan CatBoost dalam kasus prediksi churn pelanggan layanan streaming musik.
   b. Memberikan wawasan mengenai penggunaan metrik evaluasi dan interpretabilitas model (seperti SHAP) dalam analisis data pelanggan.
2. Manfaat Praktis
   a. Memberikan gambaran kepada pelaku industri mengenai faktor-faktor penting yang memengaruhi churn berdasarkan hasil evaluasi model.
   b. Menyediakan rekomendasi awal mengenai algoritma yang potensial untuk digunakan dalam pengembangan sistem prediksi churn, jika ingin diimplementasikan lebih lanjut.
3. Manfaat Sosial dan Ekonomi
   a. Memberikan arah bagi pengembangan strategi retensi pelanggan di masa depan, berdasarkan hasil evaluasi model yang telah dilakukan.

b. Mendorong pemanfaatan analisis data sebagai dasar pengambilan keputusan yang lebih efektif dalam menghadapi churn pelanggan.

1.5 Batasan Masalah

Agar penelitian tetap terfokus pada isu yang dikaji dan tidak menyimpang dari topik utama, diperlukan penetapan batasan masalah sebagai berikut:

1. Hanya menggunakan data sekunder yang diunduh dari Kaggle.
2. Pembagian data yang digunakan untuk pelatihan dan pengujian adalah rasio 60:40, 70:30, dan 80:20.

Halaman ini sengaja dikosongkan