

# PERBANDINGAN PERFORMA XGBOOST DAN LIGHTGBM DALAM PREDIKSI CHURN PELANGGAN NETFLIX

HILDA DESFIANTY A - 22081010206



# RUMUSAN MASALAH



## MASALAH 1

Bagaimana perbandingan performa algoritma LightGBM dan XGBoost dalam memprediksi customer churn pada pelanggan Netflix?



## MASALAH 2

Sejauh mana penerapan Bayesian Optimization dapat membantu menghasilkan perbandingan yang lebih objektif antara kedua model tersebut?



## MASALAH 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?



# TUJUAN PENELITIAN

## > TUJUAN 01

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

## > TUJUAN 02

Menggunakan Bayesian Optimization (Optuna) sebagai metode tuning untuk memastikan setiap model diuji dalam kondisi parameter terbaiknya.

## > TUJUAN 02

Menentukan model terbaik berdasarkan hasil evaluasi komprehensif terhadap metrik performa (akurasi, presisi, recall, F1-score, AUC, dan waktu komputasi).



# RESEARCH GAP

## > RESERACH GAP 01

Studi prediksi churn lebih banyak dilakukan pada telekomunikasi, perbankan, atau e-commerce. Analisis spesifik pada platform streaming video seperti Netflix masih sedikit

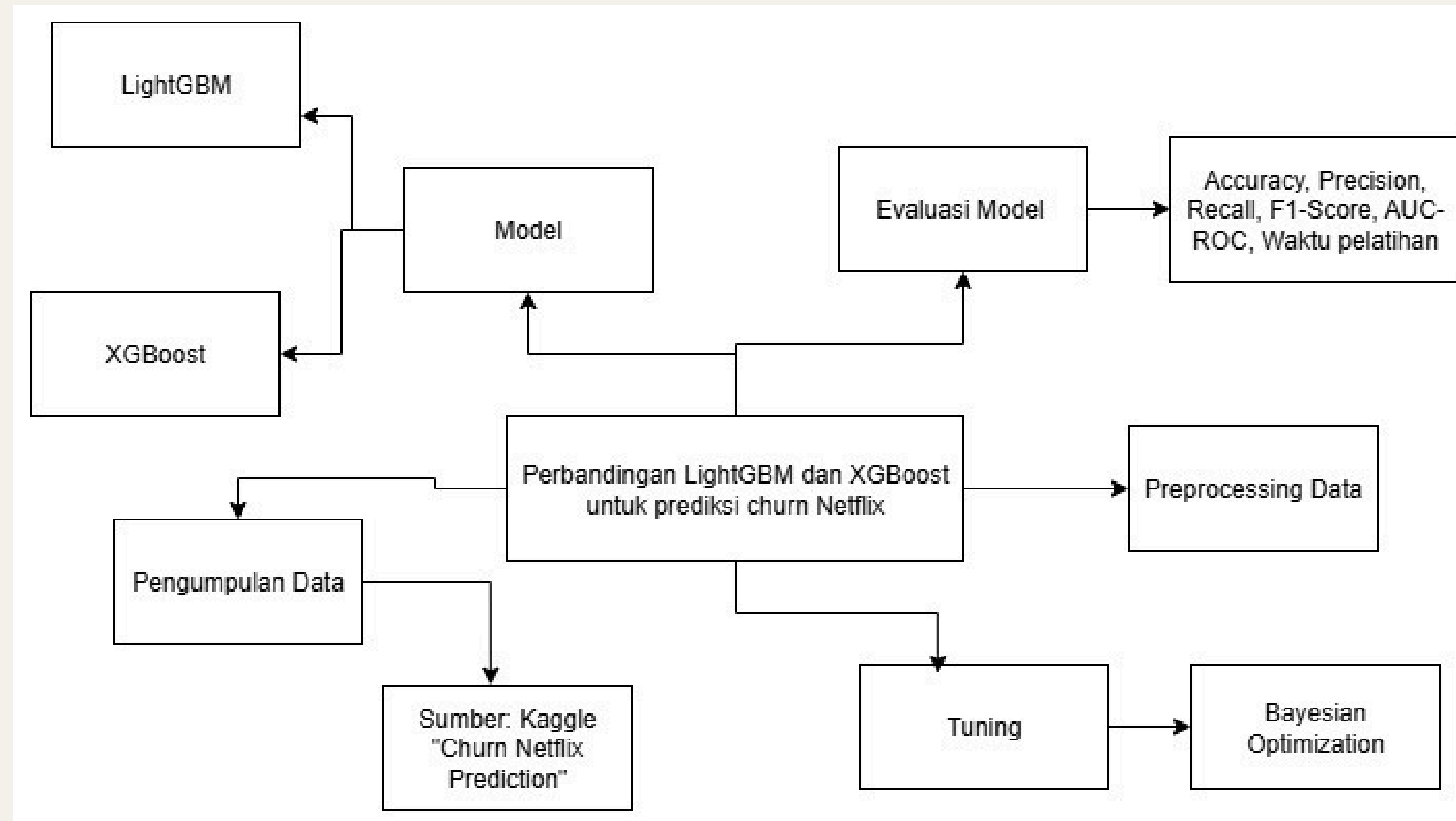
## > RESERACH GAP 02

Penelitian yang menunjukkan perbandingan LightGBM dan XGBoost lebih banyak diterapkan pada dataset lain, sehingga belum jelas model mana yang lebih unggul untuk churn Netflix.

## > RESERACH GAP 03

Tuning hyperparameter dengan Bayesian Optimization masih jarang digunakan pada studi churn di platform streaming, sehingga evaluasi perbandingan model belum optimal.

# MIND MAPPING



# METODOLOGI PENELITIAN



## AKUISISI DATA

Dataset yang digunakan berasal dari Netflix Customer Churn Dataset di Kaggle, berisi sekitar 10.000 sampel pelanggan dengan 14 variabel utama seperti usia, jenis langganan, metode pembayaran, durasi menonton, dan genre favorit. Dataset ini dipilih karena merepresentasikan perilaku pelanggan layanan streaming video. Proporsi data churn sekitar 20%, mencerminkan kondisi nyata industri OTT. Data bersifat terbuka dan anonim untuk menjaga reproduktibilitas penelitian.



## PREPROCESSING DATA

Data dibersihkan melalui imputasi missing value, encoding kategorikal, dan normalisasi numerik. Ketidakseimbangan kelas ditangani dengan SMOTE, sedangkan outlier dihapus dengan metode IQR. Dataset dibagi menjadi train, validation, dan test set menggunakan stratified split 70:15:15.

# METODOLOGI PENELITIAN



## PEMBANGUNAN MODEL

Model dasar dibangun menggunakan dua algoritma gradient boosting, yaitu LightGBM dan XGBoost, dengan parameter default untuk menghasilkan baseline awal. Kedua model dilatih menggunakan data latih dan divalidasi dengan 5-fold cross-validation serta early stopping untuk mencegah overfitting. Pembangunan dilakukan di Google Colab dengan pencatatan performa awal melalui MLflow sebagai dasar evaluasi sebelum optimasi parameter.



## OPTIMASI PARAMETER

Optimasi dilakukan menggunakan Bayesian Optimization melalui framework Optuna untuk mencari kombinasi parameter terbaik secara efisien. Proses ini menggunakan 50-100 percobaan (trials) dengan Tree-structured Parzen Estimator (TPE) sebagai sampler dan pruning otomatis untuk menghentikan percobaan yang tidak optimal. Parameter penting seperti learning rate, max depth, dan num leaves disesuaikan untuk memaksimalkan AUC pada data validasi.

# METODOLOGI PENELITIAN



## EVALUASI MODEL

Evaluasi dilakukan untuk membandingkan performa kedua model berdasarkan metrik akurasi, presisi, recall, F1-score, AUC-ROC, serta waktu pelatihan. Selain itu, dilakukan analisis feature importance menggunakan SHAP untuk mengetahui pengaruh setiap variabel terhadap prediksi churn. Hasil evaluasi menunjukkan peningkatan signifikan setelah optimasi dengan Bayesian Optimization dibanding baseline awal.



## HASIL PERBANDINGAN

Tahap ini menganalisis perbedaan kinerja LightGBM dan XGBoost secara kuantitatif dan kualitatif. LightGBM menunjukkan efisiensi waktu pelatihan yang lebih tinggi, sedangkan XGBoost unggul dalam akurasi pada fitur numerik. Visualisasi hasil disajikan dalam bentuk grafik AUC, F1-score, dan waktu pelatihan. Analisis juga mencakup rekomendasi penerapan model, di mana LightGBM cocok untuk sistem real-time, sementara XGBoost lebih sesuai untuk analisis mendalam.



# TERIMA KASIH BANYAK!

**HILDA DESFIANTY A - 22081010206**

