# Identify Customer Churn with Machine Learning

Author: Wahyu Eki Sepriansyah

## Project Overview

Proyek ini bertujuan membangun model machine learning untuk memprediksi customer churn pada sebuah perusahaan e-commerce. Customer churn adalah kondisi ketika pelanggan berhenti menggunakan layanan. Kehilangan pelanggan lama lebih mahal daripada mendapatkan pelanggan baru, sehingga penting bagi perusahaan untuk bisa mengidentifikasi pelanggan berisiko churn secara dini.

## Goals

### Business Goals

- Mengurangi tingkat churn dengan intervensi yang tepat waktu.  
- Meningkatkan efektivitas kampanye retensi.  
- Mengoptimalkan alokasi anggaran dan sumber daya.

### Analytical Goals

- Membangun model klasifikasi churn dengan performa terbaik.  
- Mengidentifikasi fitur yang paling berpengaruh terhadap churn.  
- Menentukan metrik evaluasi yang sesuai untuk kebutuhan bisnis.

## Dataset

- Sumber data: data\_ecommerce\_customer\_churn.csv  
- Isi data: demografi pelanggan, perilaku transaksi, serta interaksi layanan.  
- Target: 0 = Tidak Churn, 1 = Churn

## Methodology

1. Business Understanding – mendefinisikan masalah dan tujuan.  
2. Data Understanding – eksplorasi data, statistik deskriptif, visualisasi.  
3. Data Cleaning & Preparation – menangani missing values, duplikasi, encoding, scaling.  
4. Feature Engineering – transformasi variabel dan seleksi fitur.  
5. Modeling – mencoba berbagai algoritma:  
 - Logistic Regression  
 - KNN  
 - Decision Tree  
 - Random Forest  
 - Gradient Boosting  
 - XGBoost  
6. Handling Imbalance – oversampling (SMOTE), undersampling (NearMiss), dll.  
7. Model Evaluation – menggunakan metrik: Accuracy, Recall, Precision, F1, ROC-AUC.  
8. Hyperparameter Tuning – RandomizedSearchCV.

## Evaluation Metric

- Precision: mengukur ketepatan prediksi churn.  
- Recall: fokus utama (agar pelanggan berisiko tidak terlewat).  
- F1-score: keseimbangan precision & recall.  
- ROC-AUC: mengukur kemampuan model membedakan churn vs non-churn.  
  
Error Types:  
- Type I (False Positive) → biaya retensi ke pelanggan yang sebenarnya loyal.  
- Type II (False Negative) → kehilangan pelanggan berharga (lebih merugikan).

## Business Insight

- Faktor dominan churn: lama berlangganan (tenure), jumlah komplain, dan cashback.  
- Model dengan recall tinggi lebih cocok untuk strategi retensi (meskipun precision sedikit turun).  
- Insight ini dapat membantu tim marketing membuat program retensi yang lebih tepat sasaran.

## Deployment

Model dapat di-deploy menggunakan Streamlit untuk membuat aplikasi prediksi customer churn berbasis web.  
  
Contoh penggunaan:  
streamlit run app.py

## Requirements

- Python 3.10+  
- Library: pandas, numpy, scikit-learn, imbalanced-learn, xgboost, seaborn, matplotlib, streamlit  
  
Install dependencies:  
pip install -r requirements.txt

## Next Step

- Menambahkan interpretabilitas model (SHAP/feature importance).  
- Menghubungkan aplikasi dengan database real-time pelanggan.  
- Integrasi ke sistem CRM untuk otomatisasi kampanye retensi.