Purwadhika
Digital Technology School

Analisis Prediksi Churn Pelanggan E-Commerce

Machine Learning untuk
Deteksi Dini Churn Pelanggan

Muhammad Naufal Maahir
JCDS 2804016

Latar Belakang

Churn pelanggan memiliki dampak langsung terhadap bisnis terutama dalam segi revenue. Dengan bantuan teknologi, churn pelanggan dapat dideteksi sejak dini yang berguna dalam menyusun langkah intervensi dan preventif untuk mengurangi atau menghindari dampak dari churn. Salah satu teknologi yang dapat mendeteksi churn sejak dini adalah Machine Learning.

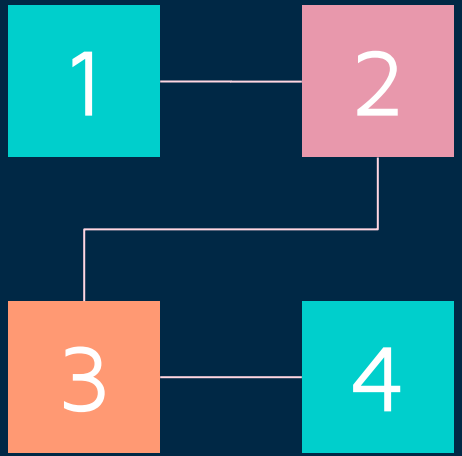
Machine Learning (ML) merupakan cabang dari kecerdasan buatan (*Artificial Intelligence*/AI) yang memungkinkan komputer untuk mempelajari pola-pola dari data yang dimasukan dan membuat keputusan atau prediksi tanpa harus diprogram secara eksplisit.

Tujuan Analisa

Purwadhika
Digital Technology School

Prediksi
Memprediksi churn berdasarkan perilaku pelanggan

Identifikasi
Identifikasi features penyebab churn



Evaluasi
Mengevaluasi model ML yang dibentuk

Insight
Memberikan insight strategis untuk retensi pelanggan

Purwadhika
Digital Technology School

Dataset

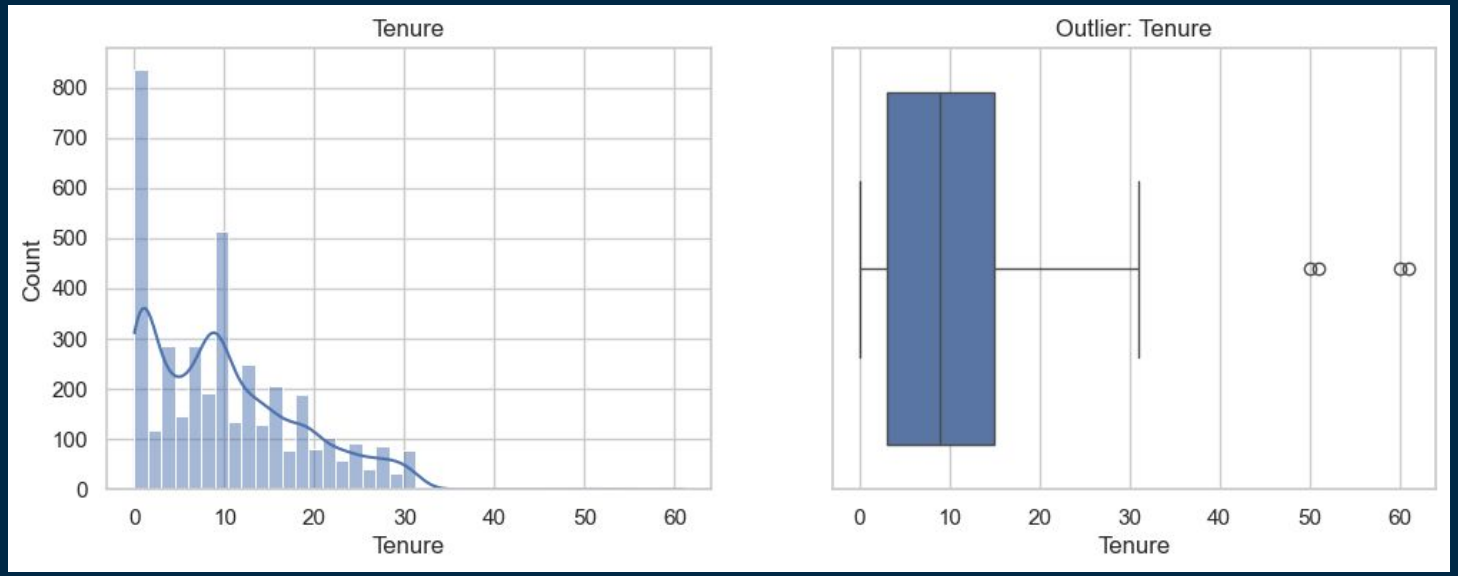
- Nama Dataset: data_ecommerce_customer_churn.csv
- Jumlah data: 3941 entri, 11 fitur
- Fitur Target: Churn
- Tipe data:
  - Numerik: Tenure, WarehouseToHome, NumberOfDeviceRegistered, SatisfactionScore, NumberofAddress, DaySinceLastOrder, CashackAmount
  - Kategorikal: PreferedOrderCat, MaritalStatus
  - Biner: Complain

Missing Values

Fitur	Missing Values
Tenure	194
WarehouseToHome	169
DaySinceLastOrder	213

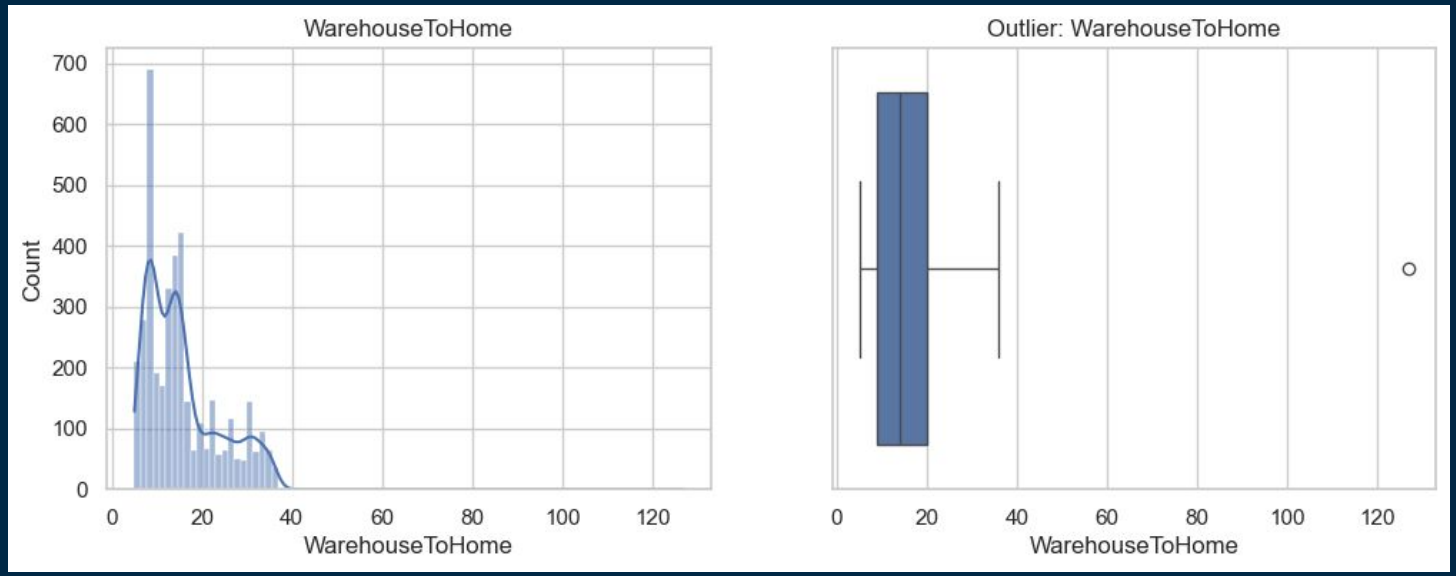
Missing Values ditangani dengan imputasi median untuk jenis data Numerik dan imputasi modus untuk jenis data Kategorikal

EDA Tenure



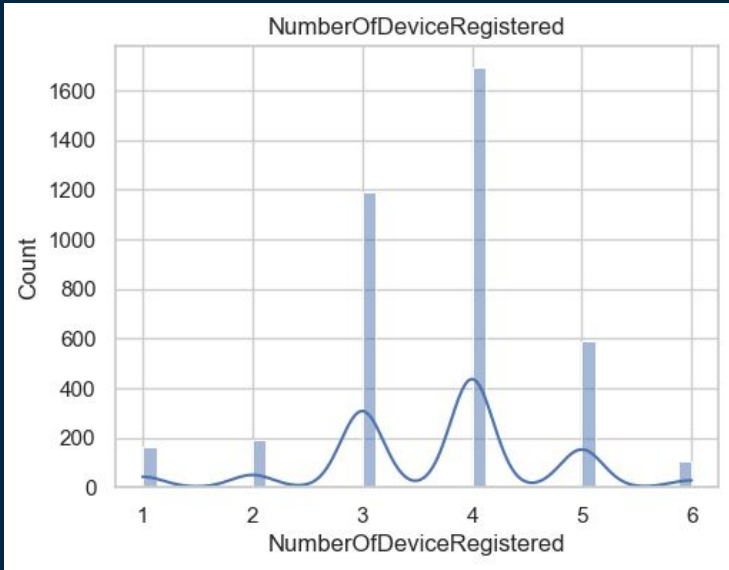
`Tenure` didominasi pelanggan dengan masa berlangganan 10–30 bulan, outlier di bawah 5

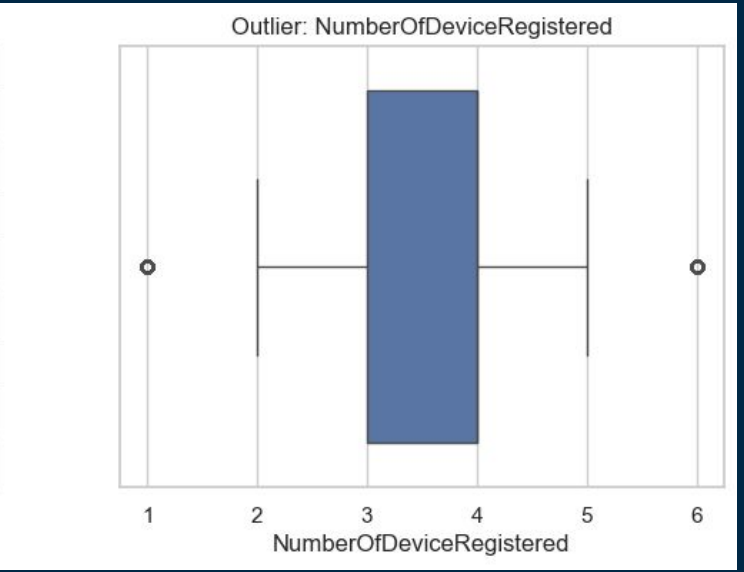
EDA WarehouseToHome



- Sebagian besar pelanggan memiliki jarak 0–20 unit dari gudang ke rumah.
- Outlier terlihat jelas di atas 40 unit, dengan satu pelanggan memiliki jarak lebih dari 120 unit.

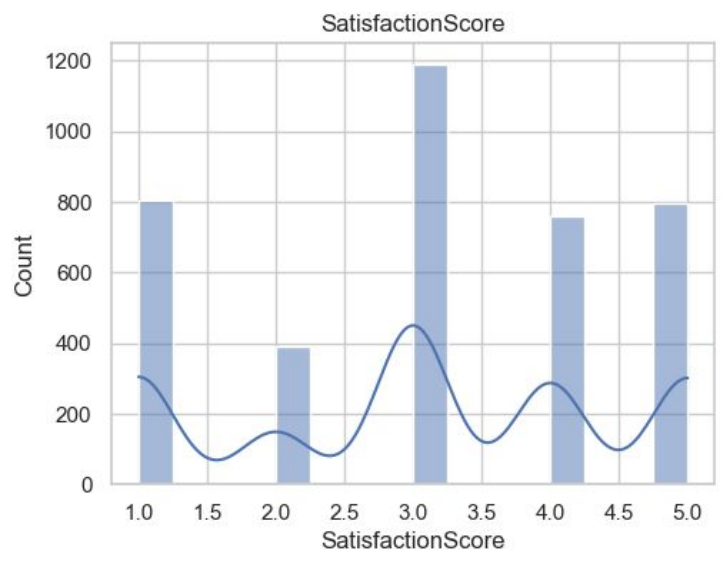
EDA Registered Device

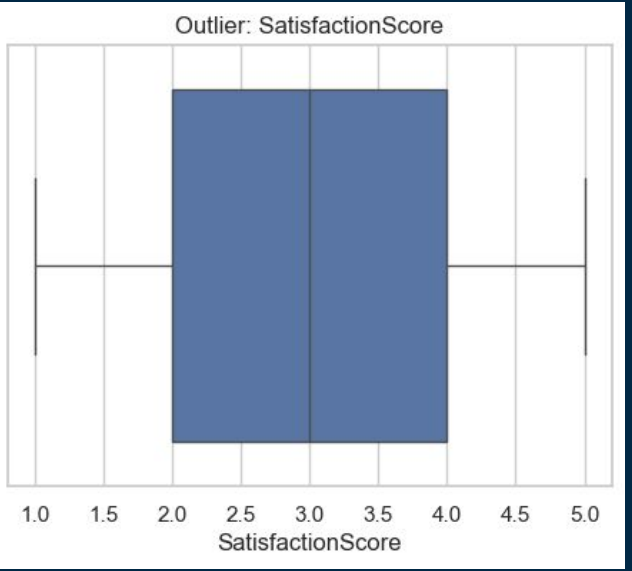




- Sebagian besar pelanggan memiliki 3 atau 4 perangkat terdaftar, yang merupakan puncak distribusi.
- Outlier terlihat pada pelanggan dengan 1 perangkat (di bawah batas bawah) dan 6 perangkat (di atas batas atas).

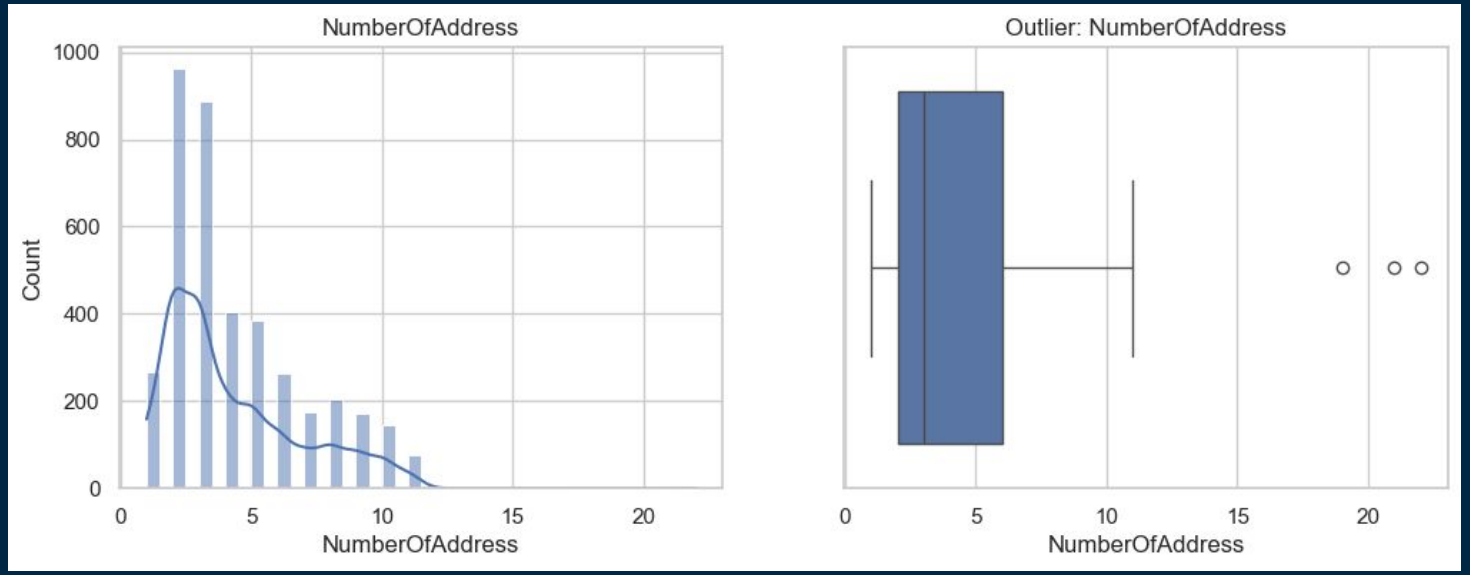
EDA Satisfaction Score





- Skor kepuasan 3 adalah yang paling umum, dengan jumlah pelanggan tertinggi.
- Tidak ada outlier yang signifikan, karena semua skor berada dalam rentang 1 hingga 5.

EDA Number of Address

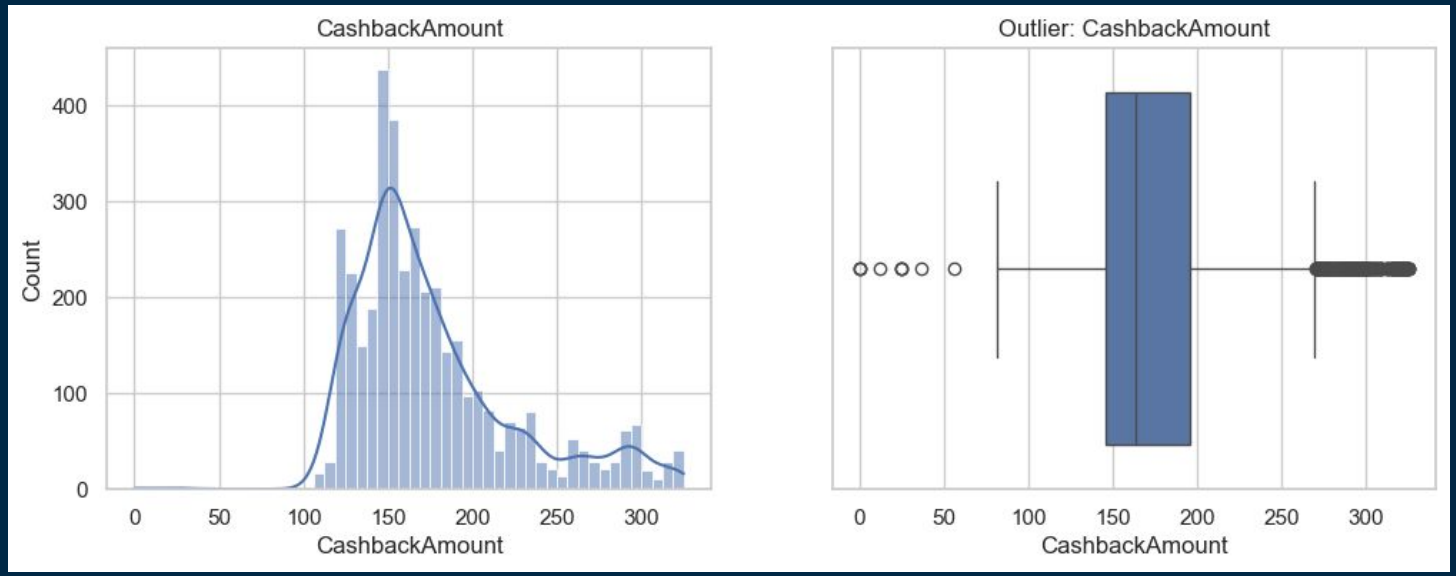


- Sebagian besar pelanggan memiliki 1 hingga 5 alamat terdaftar, dengan puncak distribusi pada 2 hingga 3 alamat.
- Outlier terlihat pada pelanggan dengan lebih dari 10 alamat terdaftar, dengan beberapa kasus ekstrem hingga 20 alamat.

EDA Order Terakhir

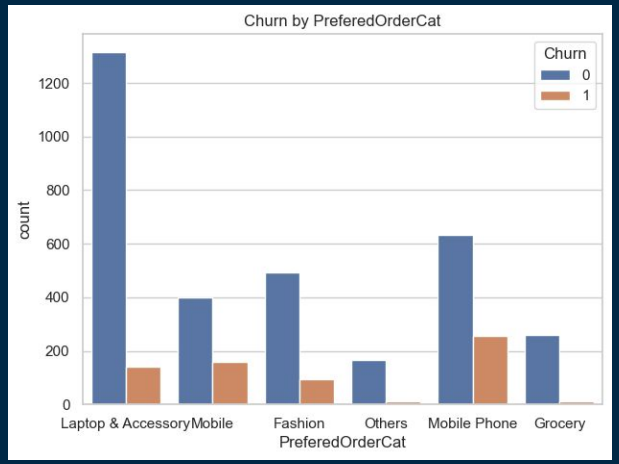
- Sebagian besar pelanggan melakukan pemesanan terakhir mereka dalam 0 hingga 10 hari.
- Outlier terlihat pada pelanggan dengan waktu lebih dari 15 hari, dengan beberapa kasus ekstrem hingga lebih dari 40 hari.

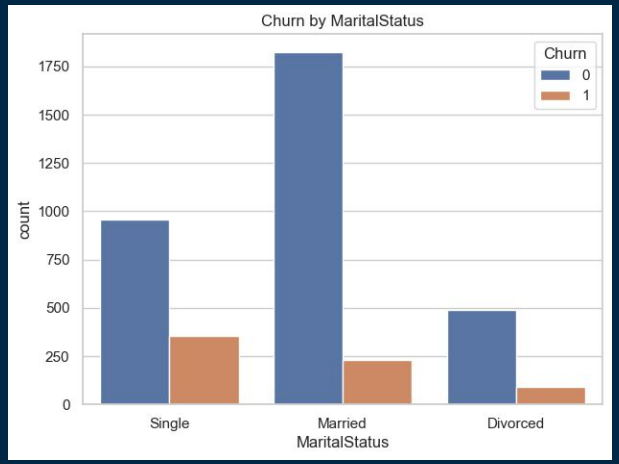
EDA Cashback

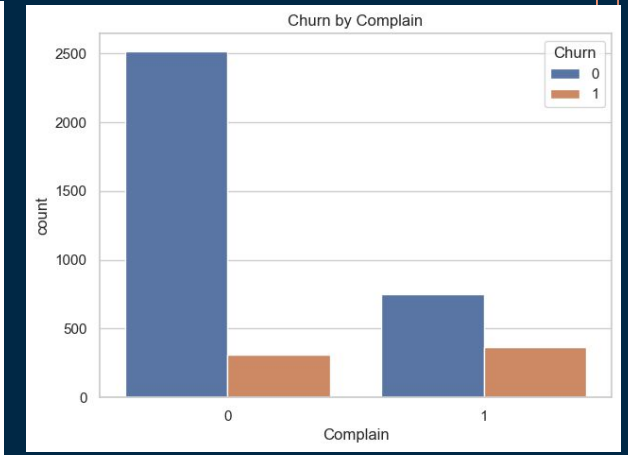


- Sebagian besar pelanggan menerima cashback di kisaran 100 hingga 200.
- Outlier terlihat pada pelanggan yang menerima cashback di bawah 100 atau di atas 250.

EDA Churn to Fitur



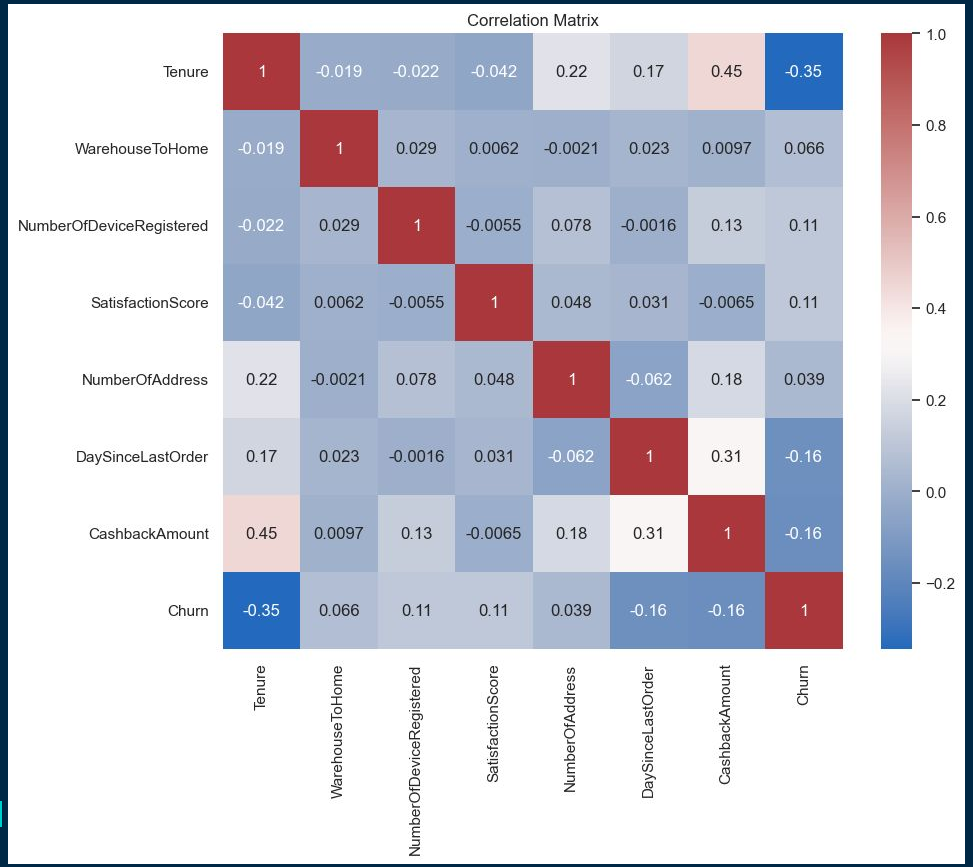




- `PreferedOrderCat`: Variasi rasio churn signifikan antar kategori.
- `Complain`: Pelanggan yang pernah komplain lebih rentan churn.
- `MaritalStatus`: Perbedaan rasio churn antar status relatif kecil namun tetap terlihat.

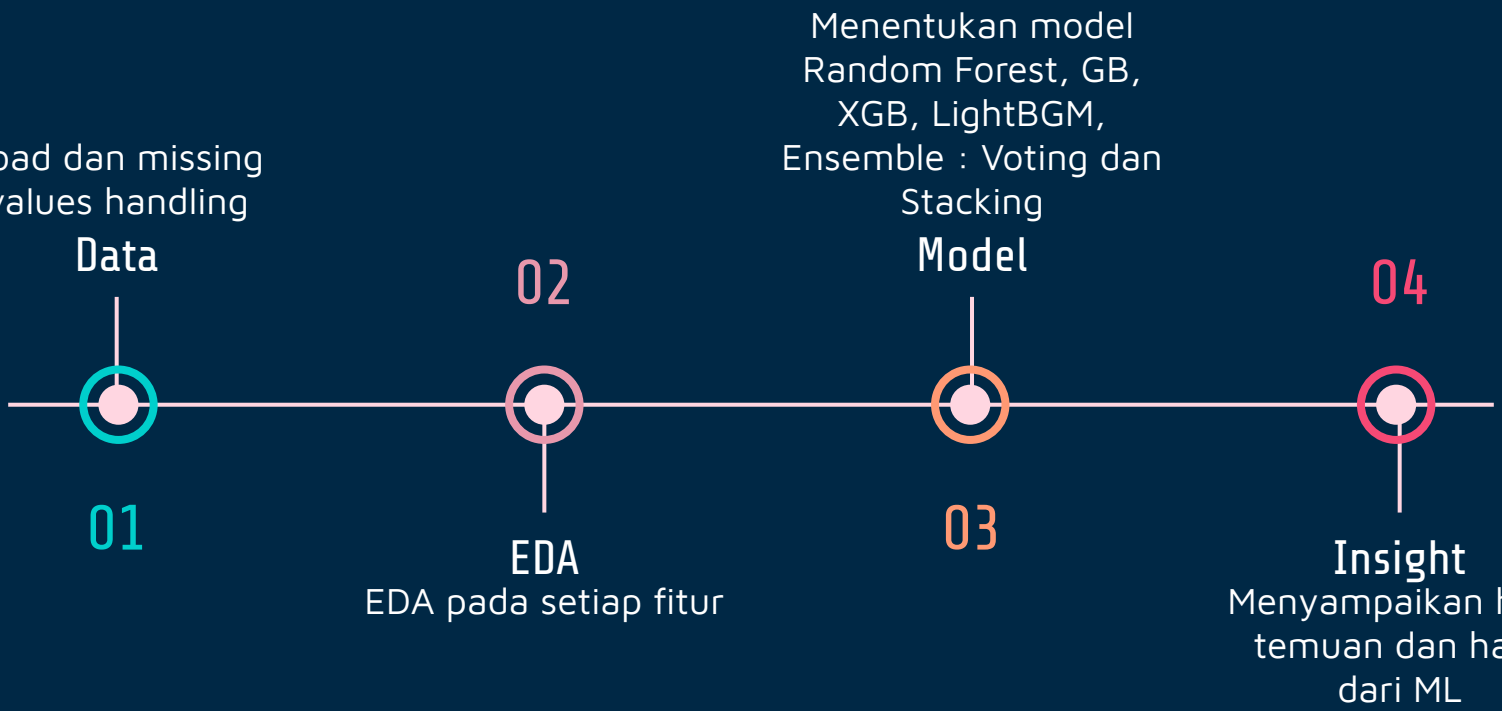
Purwadhika
Digital Technology School

Korelasi Fitur



Fitur	Korelasi terhadap Churn
Tenure	**-0.35**
CashbackAmount	**-0.16**
DaySinceLastOrder	**-0.16**
NumberOfDeviceRegistered	+0.11
SatisfactionScore	+0.11
WarehouseToHome	+0.066
NumberOfAddress	+0.039

PROCESS

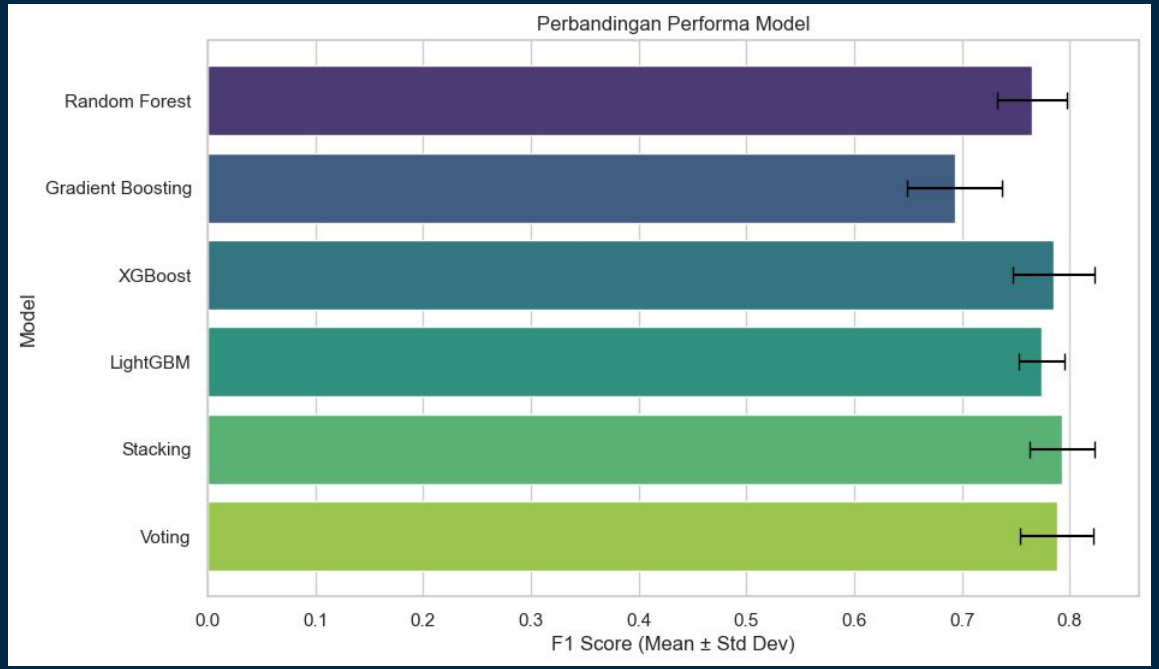

Load dan missing values handling
Data
01
02
EDA
EDA pada setiap fitur
Menentukan model Random Forest, GB, XGB, LightBGM, Ensemble : Voting dan Stacking
Model
03
04
Insight
Menyampaikan hasil temuan dan hasil dari ML

Model diuji

Purwadhika
Digital Technology School

Model	Justifikasi
Random Forest	Model baseline yang andal, tetapi kalah dalam ketajaman prediksi dibanding boosting.
Gradient Boosting	Model boosting klasik dengan performa baik dan stabil.
XGBoost	Boosting cepat dan efisien, cocok untuk data tabular.
LightGBM	Performa terbaik, efisien, dan cocok untuk fitur campuran.
Voting Classifier	Gabungan beberapa model untuk stabilitas prediksi.
Stacking Classifier	Menggabungkan output model dengan meta-model yang fleksibel.

Hasil Evaluasi Model



Penjelasan:

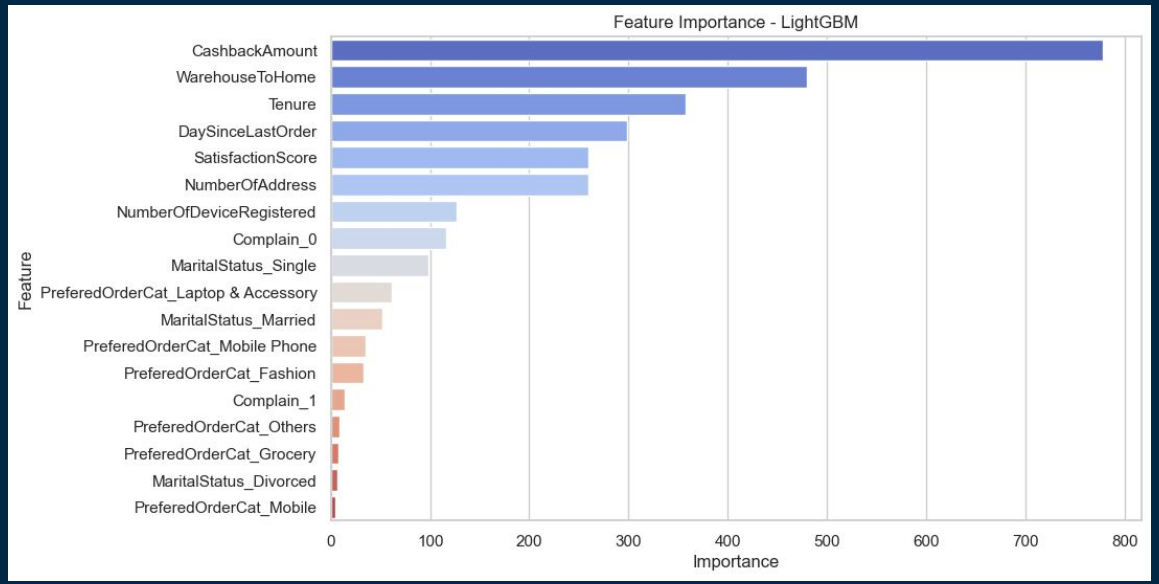
Stacking memiliki mean F1 tertinggi (0.793) dan stabilitas baik.

XGBoost unggul dalam mean, namun std dev (0.038) menunjukkan fluktuasi antar subset.

LightGBM pilihan optimal jika mengutamakan konsistensi di seluruh fold.

Gradient Boosting memerlukan hyper-tuning lebih lanjut karena mean rendah dan std tinggi.

Feature Importance



Penjelasan:

CashbackAmount adalah pendorong utama retensi—semakin besar cashback, semakin kecil churn.

WarehouseToHome & Tenure menunjukkan pentingnya kedekatan & loyalitas waktu lama.

DaySinceLastOrder dan SatisfactionScore kritikal untuk intervensi cepat: pelanggan yang lama tidak order atau kurang puas harus diprioritaskan.

Fitur kategorikal (`PreferedOrderCat`, `MaritalStatus`, `Complain`) memiliki peran minor namun tetap memberikan sinyal tambahan.

Kesimpulan dan Rekomendasi

Kesimpulan	Rekomendasi
Ensemble Stacking memberikan performa terbaik secara rata-rata (F1 = 0.793).	Gunakan LightGBM untuk kestabilan, atau Stacking Classifier untuk performa puncak.
XGBoost unggul dalam skor rata-rata, tetapi ada fluktuasi yang perlu dikendalikan.	Optimasi Hyperparameter
LightGBM menawarkan kinerja yang konsisten dan cocok untuk produksi.	Cashback strategis dan segmentasi jarak
Feature penting untuk prediksi churn adalah `CashbackAmount`, `WarehouseToHome`, `Tenure`, `DaySinceLastOrder`, dan `SatisfactionScore`.	Tambahkan fitur Customer Behavior
Model sudah memiliki precision tinggi dan recall memadai, namun peluang peningkatan terutama pada deteksi false negatives masih ada.	Deploy model dengan monitoring real-time dan update model berkala berdasarkan data terbaru dan evaluasi drift

THANK YOU

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik