KEYLA'S PROJECT

Churn Prediction ML Models Comparison

NAME

Keyla Faristha Rindani

INTRODUCTION

Project Overview

Problem

Banyak perusahaan kehilangan pelanggan tanpa mengetahui penyebabnya. Churn yang tidak terdeteksi membuat perusahaan sulit mempertahankan pelanggan, meningkatkan biaya operasional, dan mengurangi pendapatan.

Solution

Saya membangun model machine learning yang memprediksi apakah seorang pelanggan akan churn atau tidak.

Why It Matters

Model ini membantu perusahaan mengambil keputusan yang lebih proaktif dan data-driven.

WORKFLOW

Workflow

Pada tahap ini saya mulai memahami data. Insight dari EDA membantu saya menentukan preprocessing dan model mana yang paling cocok

Saya lakukan splitting data dengan membagi menjadi train dan test lalu saya lakukan encoding fitur kategorikal, data sudah siap digunakan oleh model machine learning

Saya membangun dan membandingkan beberapa model seperti Logistic Regression, Random Forest, dan K-Means untuk clustering.

Evaluasi dilakukan menggunakan classification report (precision, recall, f1-score) serta feature importance. Dari sini terlihat model mana yang paling akurat dan seberapa baik model mengenali customer yang berisiko churn

Dataset

Sumber: Dataset Churn dari Bank/Perusahaan Telekomunikasi (diberikan oleh dosen)

Jumlah Data: ±10.000 baris dengan 14 kolom

Isi Fitur: Rownumber, CustomerId, Surname, CreditScore, Geography, Gender, Age, Tenure, Balance, NumofProducts, HasCrCredit, IsActiveMember, EstimatedSalary, Exited

Insight Awal:

- Ada beberapa fitur numerik dan kategorikal

Fitur-fitur

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   RowNumber        10000 non-null  int64
 1   CustomerId       10000 non-null  int64
 2   Surname          10000 non-null  object
 3   CreditScore      10000 non-null  int64
 4   Geography        10000 non-null  object
 5   Gender           10000 non-null  object
 6   Age              10000 non-null  int64
 7   Tenure           10000 non-null  int64
 8   Balance          10000 non-null  float64
 9   NumOfProducts    10000 non-null  int64
 10  HasCrCard        10000 non-null  int64
 11  IsActiveMember   10000 non-null  int64
 12  EstimatedSalary  10000 non-null  float64
 13  Exited           10000 non-null  int64
dtypes: float64(2), int64(9), object(3)
memory usage: 1.1+ MB
```

EDA (Exploratory Data Analysis)

Mengecek Missing Value & Duplikat

	Missing Values	Missing Percentage	Data Type
RowNumber	0	0.0	int64
CustomerId	0	0.0	int64
Surname	0	0.0	object
CreditScore	0	0.0	int64
Geography	0	0.0	object
Gender	0	0.0	object
Age	0	0.0	int64
Tenure	0	0.0	int64
Balance	0	0.0	float64
NumOfProducts	0	0.0	int64
HasCrCard	0	0.0	int64
IsActiveMember	0	0.0	int64
EstimatedSalary	0	0.0	float64
Exited	0	0.0	int64

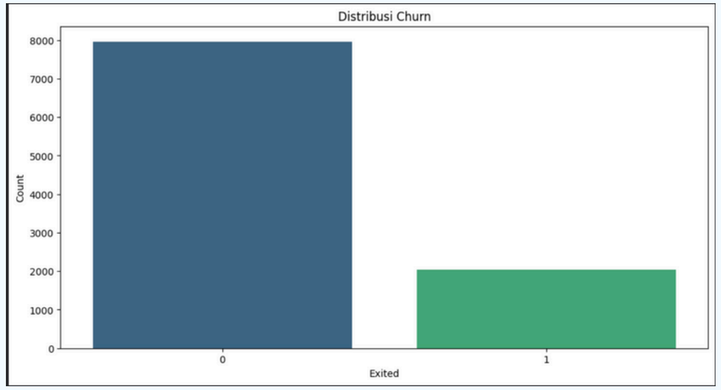
Duplicate Data Summary:	
	0
Duplicate Count	0
Duplicate Percentage	0.0
Data Type	int64

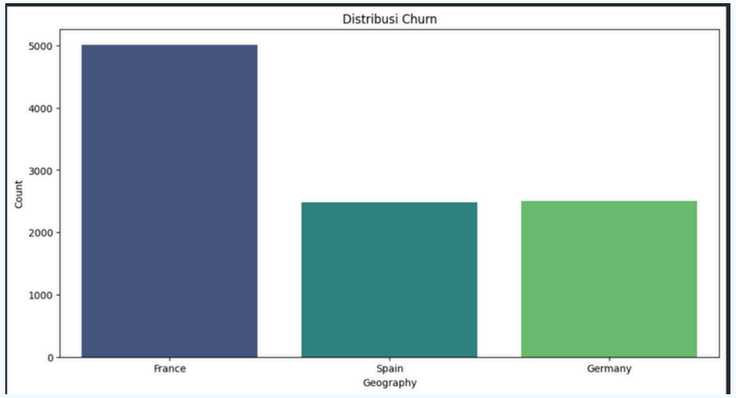
Mengecek Konsisten Data Kategorikal

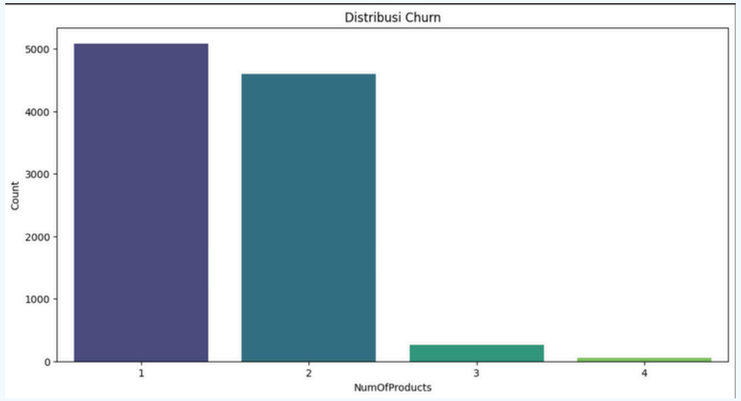
Geography	count
France	5014
Germany	2509
Spain	2477

Gender	count
Male	5457
Female	4543

Visualisasi Fitur Target & Fitur Numerikal







Data Preprocessing

Splitting Data

Feature Engineering (One Hot Encoding)

```
One-Hot Encoded Training Data:
      CreditScore  Gender  Age  Tenure    Balance  NumOfProducts  HasCrCard  \
5866          735  Female   53       8  123845.36              2          0
1938          518    Male   38       3   90957.81              1          0
4194          572  Female   54       9   97382.53              1          1
6332          619  Female   35       4   90413.12              1          1
1             608  Female   41       1   83807.86              1          0

      IsActiveMember  EstimatedSalary  Geography_France  Geography_Germany  \
5866               1        170454.93                 1                  0
1938               1        162304.59                 1                  0
4194               1        195771.95                 0                  1
6332               1         20555.21                 1                  0
1                  1        112542.58                 0                  0

      Geography_Spain
5866                0
1938                0
4194                0
6332                0
1                   1

One-Hot Encoded Testing Data:
      CreditScore  Gender  Age  Tenure    Balance  NumOfProducts  HasCrCard  \
2605          635  Female   34       5   98683.47              2          1
9717          757    Male   30       3  145396.49              1          0
68            661  Female   35       5  150725.53              2          0
9397          766  Female   52       7   92510.90              2          0
4004          530  Female   41       4       0.00              2          0

      IsActiveMember  EstimatedSalary  Geography_France  Geography_Germany  \
2605               0         15733.19                 0                  0
9717               1        198341.15                 0                  0
68                 1        113656.85                 0                  1
9397               1         66193.61                 1                  0
4004               1        147606.71                 0                  0

      Geography_Spain
2605                1
9717                1
68                  0
9397                0
4004                1
```

Feature Engineering (Binary Encoding)

```
Training Data with Binary Encoded Gender:
      Gender
5866       0
1938       1
4194       0
6332       0
1          0

Testing Data with Binary Encoded Gender:
      Gender
2605       0
9717       1
68         0
9397       0
4004       0
```

Modelling (K-Means)

Analisis K-Means Clustering (3 Cluster) pada Data Churn Bank

Cluster 1 (warna biru kehijauan)

- Pelanggan berusia sekitar 30–40 tahun
- Memiliki saldo rekening yang cukup tinggi
- Pelanggan di cluster ini cenderung lebih stabil secara finansial, sehingga kemungkinan churn lebih rendah

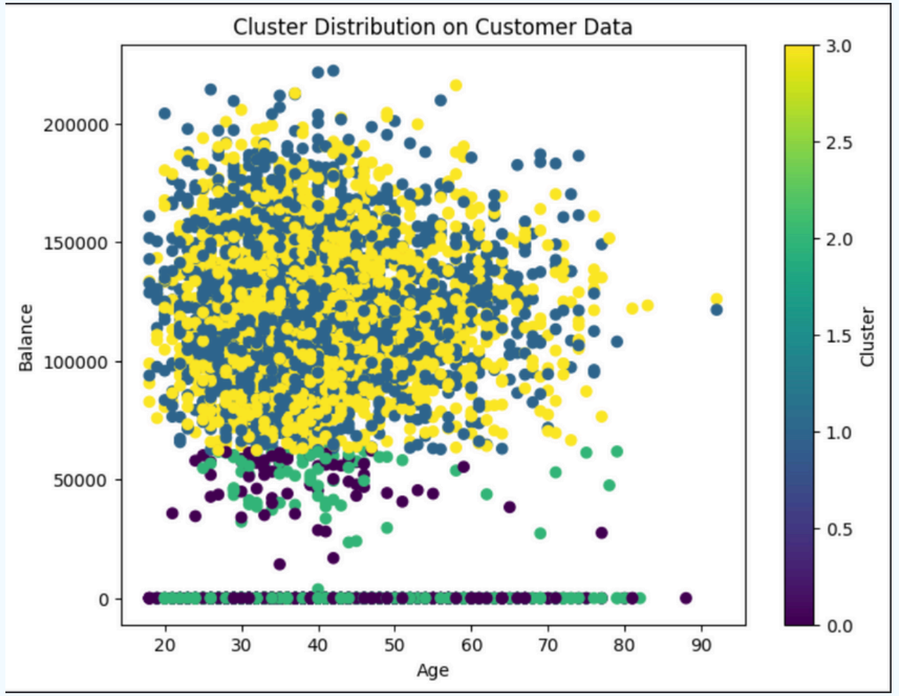
Cluster 2 (warna kuning)

- Sebaran usia luas, tapi cenderung dua puluh akhir sampai empat puluh
- Saldo banyak yang tinggi
- Ini kelompok paling aktif secara finansial dan sering dianggap pelanggan yang paling “bernilai”

Cluster 3 (warna ungu gelap)

- Banyak usia 30–60 tahun
- Saldo paling rendah (banyak yang berada di saldo = 0)
- Cluster ini umumnya pelanggan dengan aktivitas finansial rendah, sehingga sering dikaitkan dengan risiko churn lebih tinggi.

K-Means



x

Modelling (Random Forest)

Performa Utama Model

- Akurasi total: 84% → secara umum model cukup baik memprediksi apakah nasabah akan churn atau tidak.

- Precision
  - Class 0 (Stay): 84% → artinya kalau model bilang “tidak churn”, 84% benar
  - Class 1 (Exited): 93% → kalau model bilang “churn”, 92% benar.

- Recall
  - Class 0 (Stay): 99% → model sangat bagus mengenali nasabah yang tidak churn.
  - Class 1 (Exited): 25% → model jelek dalam mengenali churn, banyak churn yang salah diprediksi sebagai stay.

- F1-score
  - Class 0: 91% → model sangat kuat di mayoritas class tidak churn
  - Class 1: 39% → model lemah di minoritas class churn

Classification Report RF

Classification Report

	precision	recall	f1-score	support
0	0.84	0.99	0.91	1991
1	0.93	0.25	0.39	509
accuracy			0.84	2500
macro avg	0.88	0.62	0.65	2500
weighted avg	0.86	0.84	0.80	2500

x

Grid Search for Tuning Random Forest

Setelah dilakukan GridSearchCV, performa model meningkat dibanding sebelum tuning.

Class 0 – Stayed

- Precision: 87% → Saat model memprediksi “tidak churn”, 87% benar.
- Recall: 97% → Hampir semua nasabah yang benar-benar bertahan berhasil dikenali.
- F1-score: 92% → Kinerjanya sangat baik di mayoritas class.

Class 1 – Exited

- Precision: 82% → Saat model bilang “churn”, 82% benar.
- Recall: 45% → Naik dari sebelumnya 24% (sebelum tuning).
- F1-score: 58% → Masih belum perfect, tapi jauh lebih baik dibanding sebelum tuning.

Accuracy keseluruhan: 87%
Model menjadi lebih seimbang dan lebih bagus dalam mendeteksi nasabah churn.

Classification Report After Tuning RF

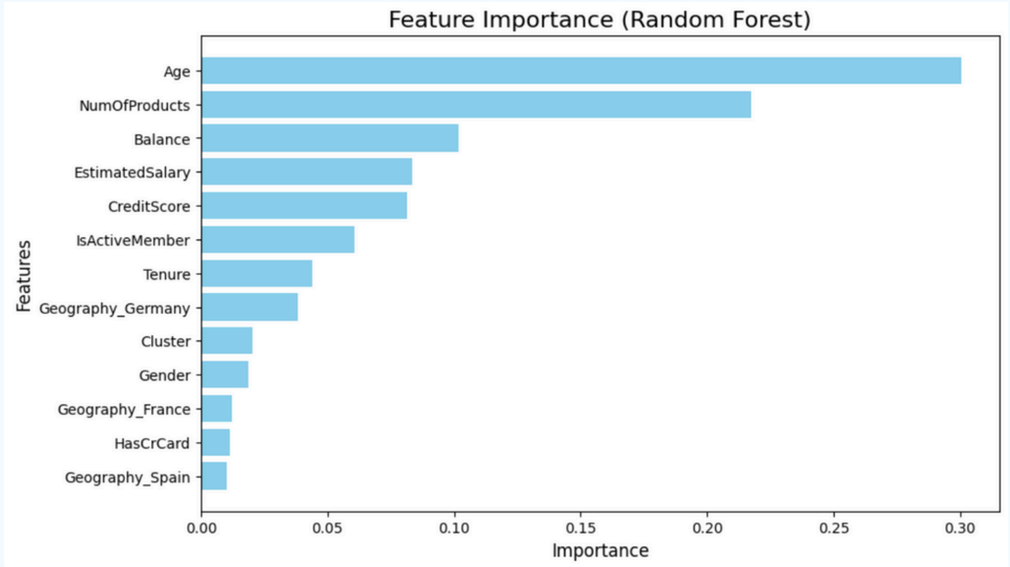
Random Forest Classification Report

	precision	recall	f1-score	support
Stayed (0)	0.87	0.97	0.92	1991
Exited (1)	0.82	0.45	0.58	509
accuracy			0.87	2500
macro avg	0.85	0.71	0.75	2500
weighted avg	0.86	0.87	0.85	2500

Feature Importance Random Forest

- Prediksi churn lebih dipengaruhi oleh usia, jumlah produk, dan saldo.
- Faktor demografis (lokasi & gender) kurang signifikan.
- Hasil ini membantu bank menentukan strategi churn prevention, misalnya memberi penawaran untuk nasabah usia tertentu atau yang hanya memiliki sedikit produk.

Feature Importance Random Forest



Logistic Regression

Performa Model

- Akurasi total: 81%
- Kelas Stayed (0)
  - Precision: 83%
  - Recall: 96% → Model sangat bagus mengenali pelanggan yang tetap.
- Kelas Exited (1)
  - Precision: 60%
  - Recall: 21% → Model sangat lemah mendeteksi pelanggan yang churn.

Kesimpulan Utama

- Model cenderung bias ke majority class (Stayed) karena data tidak seimbang.
- Banyak pelanggan churn yang tidak berhasil dideteksi (false negatives).
- Logistic Regression kurang mampu menangkap pola yang lebih kompleks dalam data churn.

Classification Report on LR

Logistic Regression Classification Report

	precision	recall	f1-score	support
Stayed (0)	0.83	0.96	0.89	1991
Exited (1)	0.60	0.21	0.31	509
accuracy			0.81	2500
macro avg	0.71	0.59	0.60	2500
weighted avg	0.78	0.81	0.77	2500

Project Summary

Pada project ini saya membangun model Machine Learning untuk memprediksi apakah nasabah akan churn menggunakan data profil dan aktivitas keuangan mereka. Prosesnya dimulai dari EDA untuk memahami pola penting seperti umur, saldo, dan jumlah produk, lalu dilanjutkan dengan preprocessing dan pemodelan menggunakan Logistic Regression dan Random Forest. Setelah dilakukan tuning, Random Forest menjadi model terbaik dengan akurasi sekitar 87%, dan analisis feature importance menunjukkan bahwa Age, NumOfProducts, dan Balance adalah faktor utama yang mempengaruhi churn. Saya juga menambahkan segmentasi menggunakan K-Means untuk memahami kelompok nasabah berdasarkan karakteristiknya. Secara keseluruhan, project ini memberikan insight yang membantu perusahaan mengidentifikasi nasabah berisiko tinggi dan mendukung strategi retensi yang lebih efektif

CHALLENGES FACED

Challenge 1

Class Imbalance pada Label Churn

Data churn sangat tidak seimbang (nasabah stay jauh lebih banyak), sehingga model kesulitan mengenali nasabah yang benar-benar churn. Ini membuat recall untuk kelas “Exited” rendah pada beberapa model.

Challenge 2

Menentukan Model Terbaik & Tuning Parameter

Mencari kombinasi hyperparameter Random Forest yang optimal cukup menantang karena butuh banyak percobaan dan waktu komputasi. Namun setelah GridSearchCV, performanya meningkat cukup signifikan.

REFLECTIONS & GROWTH

What I Learned

Personal Skills Gained

- Saya belajar memahami pola perilaku pelanggan melalui EDA dan menemukan fitur-fitur yang paling berpengaruh terhadap churn, seperti Age, NumOfProducts, dan Balance.
- Saya juga belajar pentingnya menangani class imbalance dan bagaimana tuning model seperti Random Forest dapat meningkatkan kemampuan model dalam mendeteksi pelanggan yang berisiko churn.

The End

NAME

Keyla Faristha Rindani