Project
Data Science

Customer Churn Prediction and Segmentation for Gym Membership

Reza Aprillian Nugroho

Content

- Introduction
- Overview Project
- Main Project
- EDA
- Machine Learning
- Conclusion & Recommendation

Reza Aprillian Nugroho

Data Science

Hi,

A graduated in Informatics Engineering from Yogyakarta University of Technology. He has work experience in data entry, including at Bank Exim, Iron Mountain, and as a document scanner operator at Halsindo Jaya. Currently, Reza is interested in Data Science and Data Analytics. He has mastered various supporting skills, such as Microsoft Office, Python, SQL, Power BI, and Tableau, and possesses teamwork, data analysis, and public speaking skills. With this background, Reza is confident he can contribute positively to the team and support the achievement of the company's goals. Reza expressed his appreciation for the attention and opportunity given and looked forward to further discussions regarding his contributions.

Linkedin/RezaAprillianNugroho

Education

Universitas Teknologi Yogyakarta
Informatics
2020 - 2024

Relevant Coursework :
Software Engineering, Web Development, Mobile Application Development, Big Data and Analytics, Algorithms and Data Structures, Discrete Mathematics, Machine Learning, IoT.

- Final Project Electronic Tour Guide On the Relief of the Plaosan Temple Using Mobile Based Augmented Reality
- Volleyball Student Activity Unit
- Mobile and Web Applications
- Web Projects

DiBimbing.id
Full-Stack Data Science Bootcamp
Mar 2024 - Oct 2024

- Learn about statistics, data analytics, machine learning,data visualization, and do projects related to the application of science
- Technical Skills: Python, SQL, Tableau, Google Data Studio, Streamlit.
- Techniques: Hypothesis Testing, Exploratory Data Analysis, Data Visualization, Linear Regression, Ridge Regression, Lasso Regression, K-Nearest Neighbors,Decision Tree, Random Forest, Gradient Boosted Trees, Logistic Regression, Factor Analysis, K-Means Clustering

Skills

Data Analysis

- Experienced in exploring and analyzing data to uncover patterns and business insights.
- Proficient in descriptive statistics, inferential statistics, and hypothesis testing.

Data Modeling

- Skilled in building predictive models using Machine Learning techniques such as regression, Random Forest, and XGBoost.
- Familiar with handling imbalanced data and model ptimization strategies.

Data Processing

- Proficient in cleaning, processing, and handling large datasets using Python or R.
- Expertise in data manipulation techniques using Pandas and SQL

Tools

Programming Languages: Python, SQL

Framework: Scikit-Learn, Pandas

Visualization: Tableau, PowerBI, matplotlib, Seaborn

Tools : Google Collab, Jupiter Notebook, PostgreSQL, Ms.Excel

Work Experience

Sep 2023 – Oct 2023

Data Entry, Exim Bank

Apr 2024 – Jun 2024

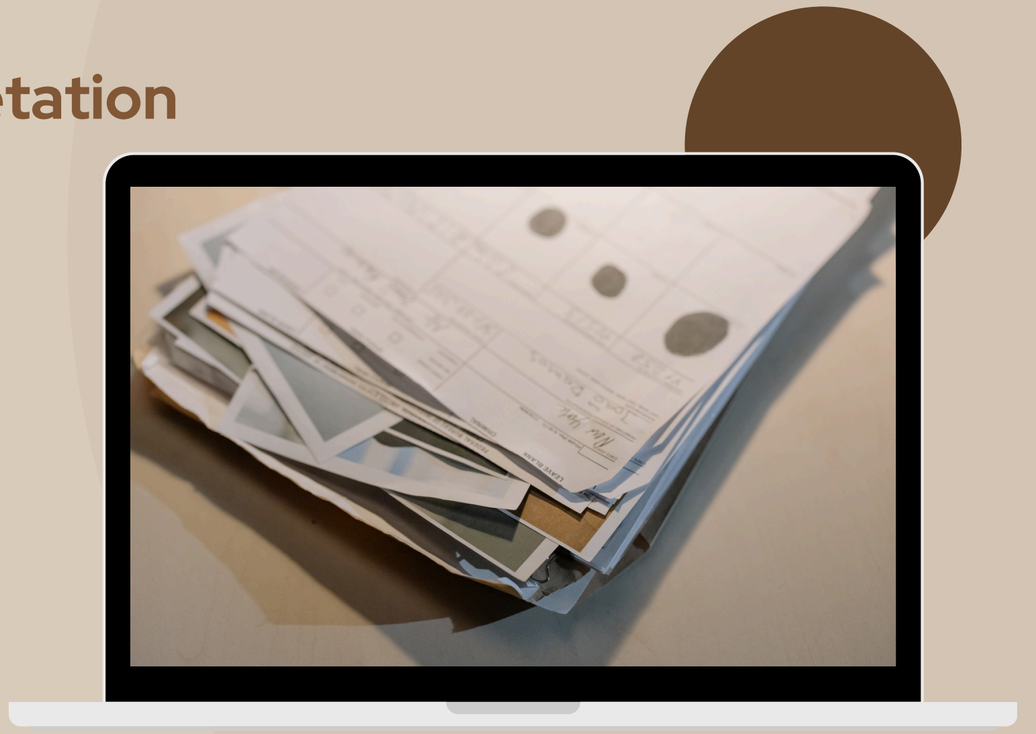
Data Entry, PT.Iron Mountain

Jul 2024 – Oct 2024

Scanner, PT.Halsindo Jaya

Overview Project

1. Fundamentals of Data Visualization
   LINK PENGERJAAN
2. Portofolio Building with Streamlit
   LINK PENGERJAAN
3. Feature Importance Analysis & Model Interpretation
   LINK PENGERJAAN
4. Exploring Clustering Techniques with Python
   LINK PENGERJAAN
5. Final Project Data Analyst
   LINK PENGERJAAN
6. Hyperparameter Tuning in Python
   LINK PENGERJAAN
7. Web Dashboard Development
   LINK PENGERJAAN



Main Project

Project Background

Industri kebugaran menghadapi tantangan besar dalam menjaga anggota. Rata-rata gym kehilangan 5–10% member tiap bulan, sehingga pendapatan turun. Tantangan terbesarnya adalah churn, sehingga strategi berbasis data sangat penting untuk menjaga pendapatan.

Dataset Gym Membership berisi data demografi, frekuensi kunjungan, layanan tambahan, dan kontribusi finansial anggota. Data ini membantu analisis retensi dan segmentasi, sehingga manajemen bisa mengenali anggota yang berisiko churn maupun yang bernilai tinggi bagi bisnis.

⚠️ Problem Statement

Tantangan utama industri kebugaran adalah tingginya churn. Banyak anggota berhenti setelah beberapa bulan, sehingga:

- Pendapatan tidak stabil.
- Biaya mencari anggota baru lebih besar daripada mempertahankan yang lama.
- Strategi promosi kurang personal karena perilaku anggota belum dipahami.

Padahal, gym sudah punya data demografi, kunjungan, dan layanan tambahan, tetapi belum dimanfaatkan maksimal untuk analisis retensi dan strategi berbasis segmen pelanggan.

🎯 Tujuan Utama

- Mengetahui faktor yang memengaruhi retensi dan churn anggota.
- Membagi anggota ke dalam segmen berdasarkan perilaku, kunjungan, dan nilai.
- Memberikan strategi retensi untuk meningkatkan loyalitas dan keuntungan.

👥 Pihak yang Diuntungkan

- Manajemen Gym → bisa membuat strategi retensi yang lebih efektif.
- Tim Marketing → bisa merancang promosi sesuai segmen pelanggan.
- Anggota Gym → mendapat layanan lebih personal dan sesuai kebutuhan.

📌 Hasil yang Diharapkan

- Pemetaan perilaku pelanggan dari data demografi, aktivitas, dan finansial.
- Segmentasi pelanggan untuk mendukung keputusan manajemen.
- Insight tentang pola kunjungan dan faktor utama churn.
- Rekomendasi strategi retensi dan promosi yang lebih personal.

📊 Metrik Keberhasilan

- Akurasi segmentasi pelanggan (contoh: Silhouette Score, Davies-Bouldin Index).
- Kualitas insight churn, yaitu seberapa jelas faktor penyebab churn teridentifikasi.
- Relevansi hasil bagi bisnis, apakah bisa dipakai untuk strategi marketing dan retensi.

📊 Data Understanding

1.Sumber Data

Dataset Gym Membership (CSV) berisi 1000 anggota dengan 23 variabel demografi, pola kunjungan, layanan, dan finansial.

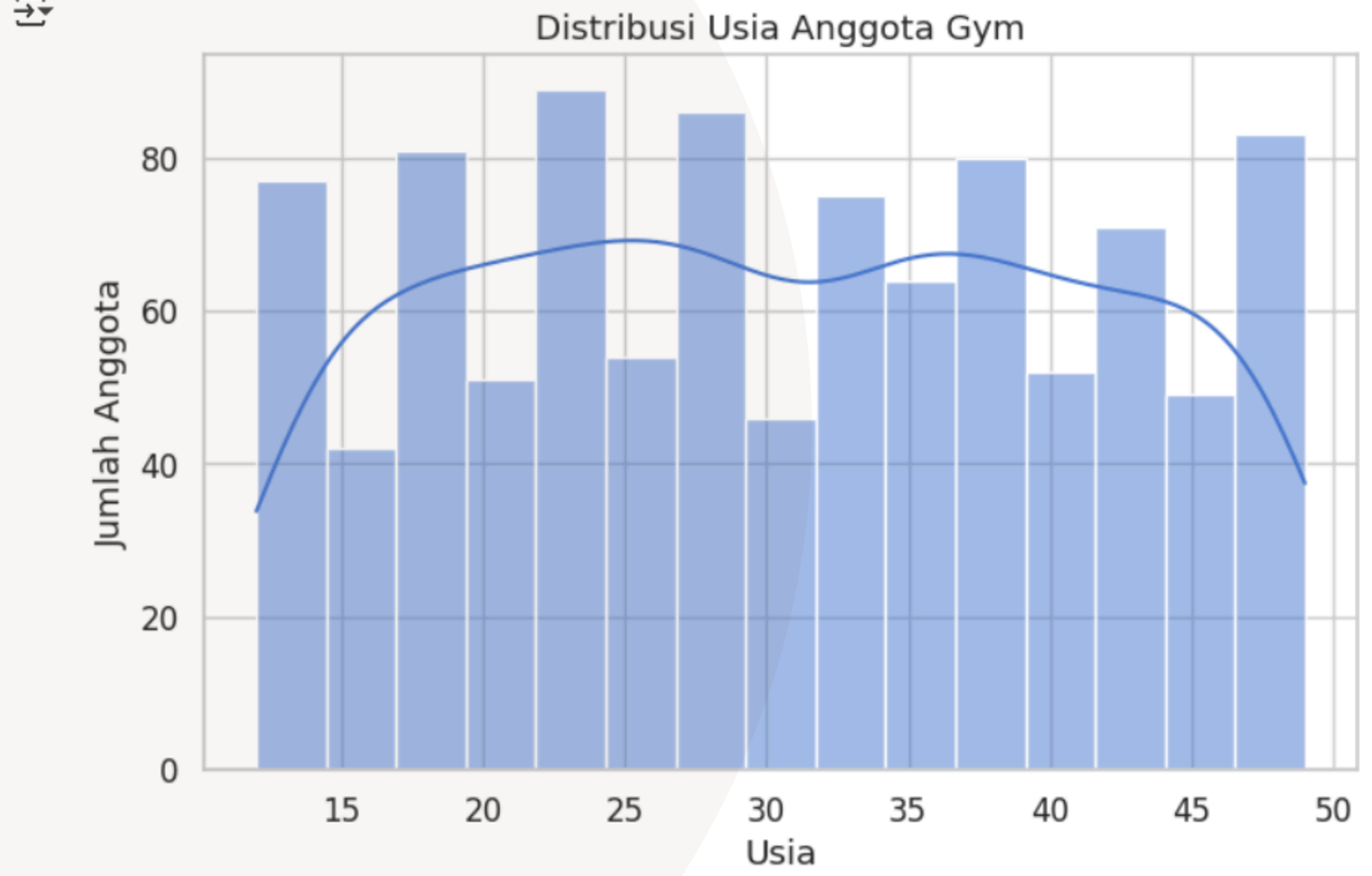
2.Jenis & Struktur Data

- Numerik: age, visit_per_week, avg_time_in_gym, recency, monetary
- Kategorikal: gender, abonoment_type, days_per_week, age_category
- Boolean: attend_group_lesson, personal_training, drink_abo, uses_sauna
- Tanggal/Waktu: birthday, last_visit_date, avg_time_check_in/out
- Metadata: id, number_of_members

3.Kualitas Data

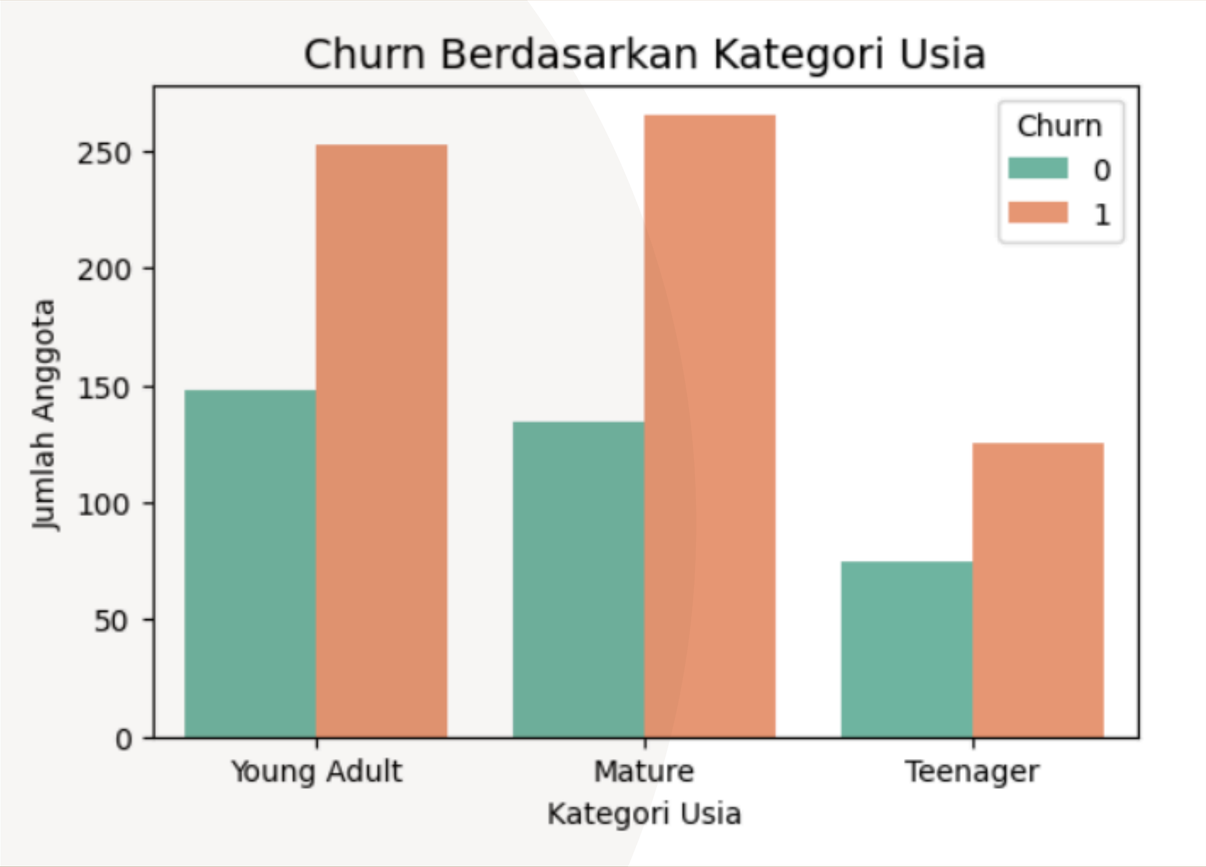
- Tidak ada missing values atau duplikasi.
- Tipe data konsisten, outlier wajar (usia 12–49 tahun, recency 0–89 hari).
- Beberapa kolom seperti number_of_members kurang relevan untuk analisis individu.

Explanatory Data Analysis and Visualisations



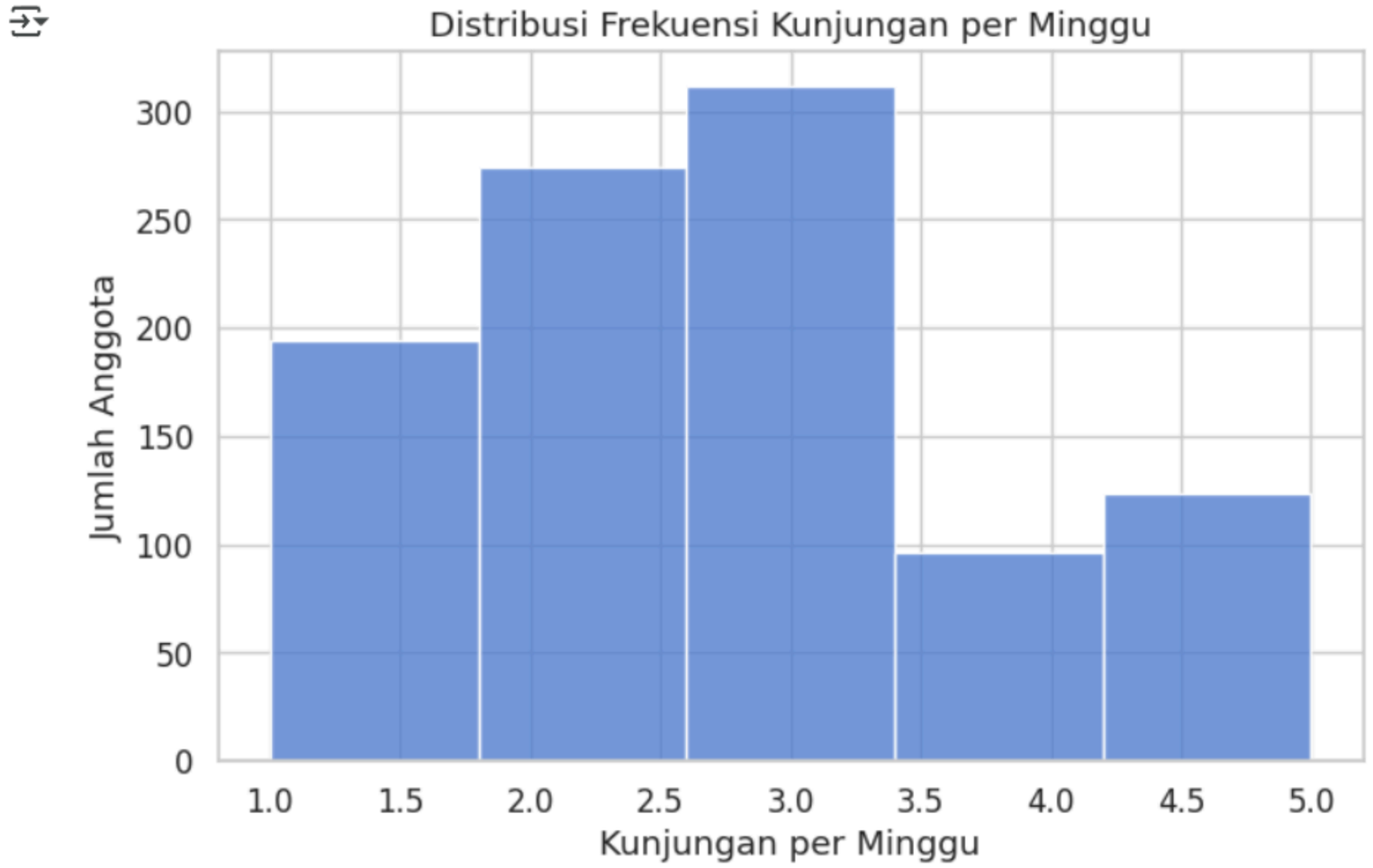
Mayoritas anggota berada pada rentang 20–35 tahun, segmen usia muda lebih dominan.

Explanatory Data Analysis and Visualisations



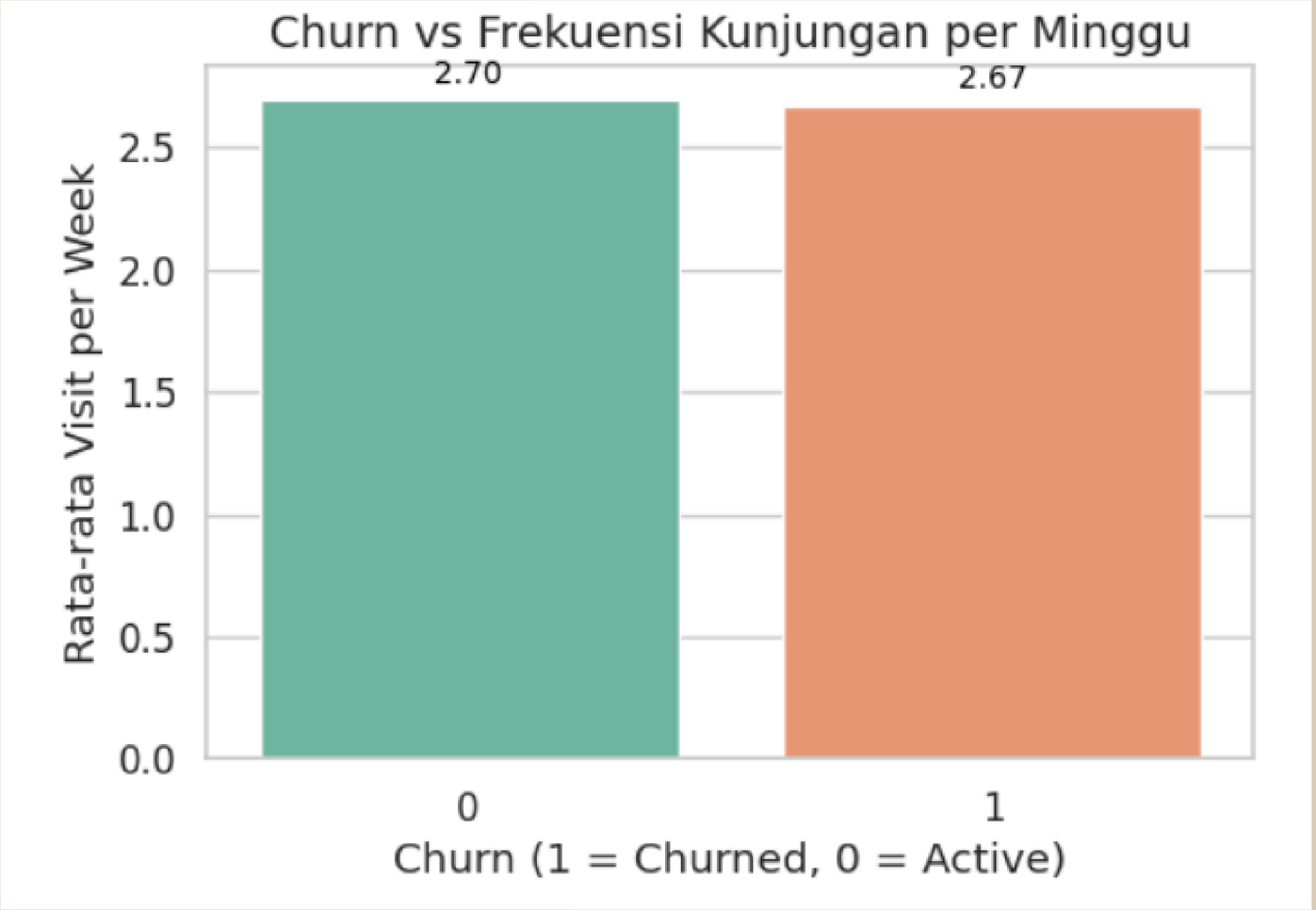
Tingkat churn lebih tinggi pada anggota usia muda dibandingkan kelompok usia lainnya, sementara usia produktif cenderung lebih stabil dan loyal, serta anggota usia lebih tua relatif jarang churn. Hal ini menandakan bahwa strategi retensi perlu difokuskan pada anggota muda yang masih kurang konsisten dalam berolahraga di gym.

Explanatory Data Analysis and Visualisations



Kebanyakan anggota datang 2–3 kali seminggu, sangat sedikit yang berkunjung >5 kali.

Explanatory Data Analysis and Visualisations



Intensitas kunjungan menjadi salah satu faktor kunci churn. Gym perlu menjaga agar anggota tetap aktif hadir supaya retensi meningkat.

Machine Learning

1. Problem Definition
   - Dengan memprediksi risiko churn, gym bisa menjaga stabilitas pendapatan, meningkatkan loyalitas anggota, dan mengambil keputusan strategis yang lebih tepat berbasis data.

2. Data Preprocessing
   - Setelah dilakukan pengecekan menggunakan perintah df.isnull().sum(), dataset tidak memiliki missing values sehingga tidak diperlukan proses imputasi maupun penghapusan data.
   - Feature Engineering :
     - visit_consistency → konsistensi kunjungan anggota
     - premium_score → skor berdasarkan jenis membership
     - recency_bucket → kategori lama tidak datang ke gym
   - Encoding variabel kategori → mengubah gender, type, dan periode check-in ke bentuk numerik.
   - Scaling fitur numerik → standarisasi pada age, monetary, avg_time_in_gym agar setara skala.
   - Train -Test Split → data dibagi 80% untuk training, 20% untuk testing.

Train-Test Split

- Jumlah data training: 800
- Jumlah data testing : 200
- Total data : 1000

Data dibagi 80:20 untuk menjaga keseimbangan antara pelatihan model dan evaluasi kinerja. Training set digunakan untuk membangun model, sedangkan testing set untuk mengukur generalisasi.

3. Model Selection

- Baseline → Logistic Regression
- Tree-based models → Random Forest

4. Training & Evaluation

- Training → model dilatih menggunakan data training.
- Evaluation → model diuji pada data testing dengan metrik:
  - Accuracy → seberapa tepat model secara keseluruhan
  - Precision → seberapa tepat model dalam memprediksi anggota churn
  - Recall → seberapa banyak anggota churn yang berhasil dideteksi
  - F1-score → keseimbangan antara precision dan recall
  - ROC-AUC → kemampuan model membedakan churn vs non-churn

5. Results

```
=== Logistic Regression ===
Accuracy: 0.69
ROC-AUC: 0.4922861150070126
              precision    recall  f1-score   support

           0       0.69      1.00      0.82       138
           1       0.00      0.00      0.00        62

    accuracy                           0.69       200
   macro avg       0.34      0.50      0.41       200
weighted avg       0.48      0.69      0.56       200
```

Logistic Regression

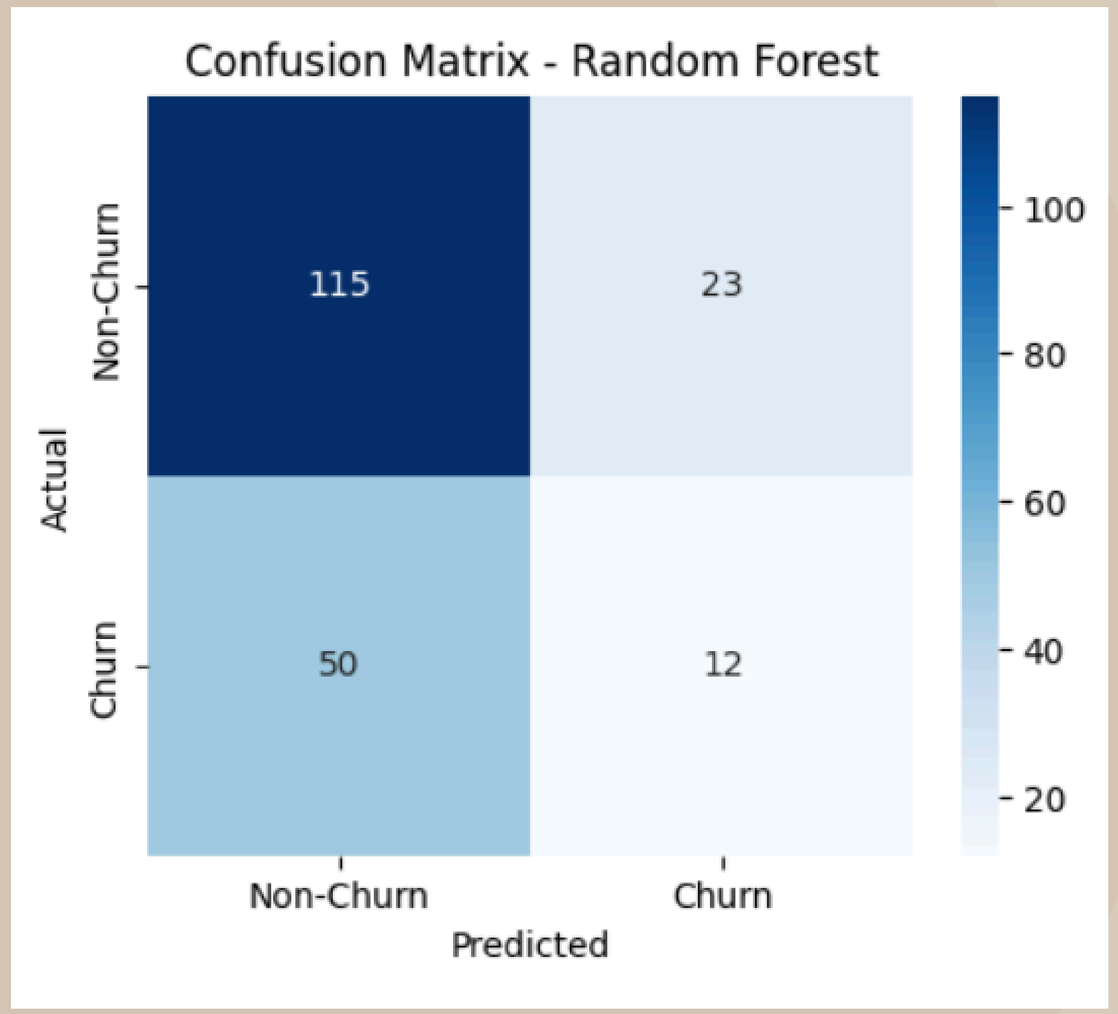
- Accuracy: 0.69 (69%)
- ROC-AUC: 0.49
- Kelemahan: model gagal mendeteksi churn (class 1), precision dan recall = 0.
- Model hanya fokus pada non-churn (class 0).

Random Forest

- Accuracy: 0.64 (64%)
- ROC-AUC: 0.55
- Lebih baik dibanding Logistic Regression karena masih mampu menangkap sebagian churn meskipun recall churn masih rendah (0.19).

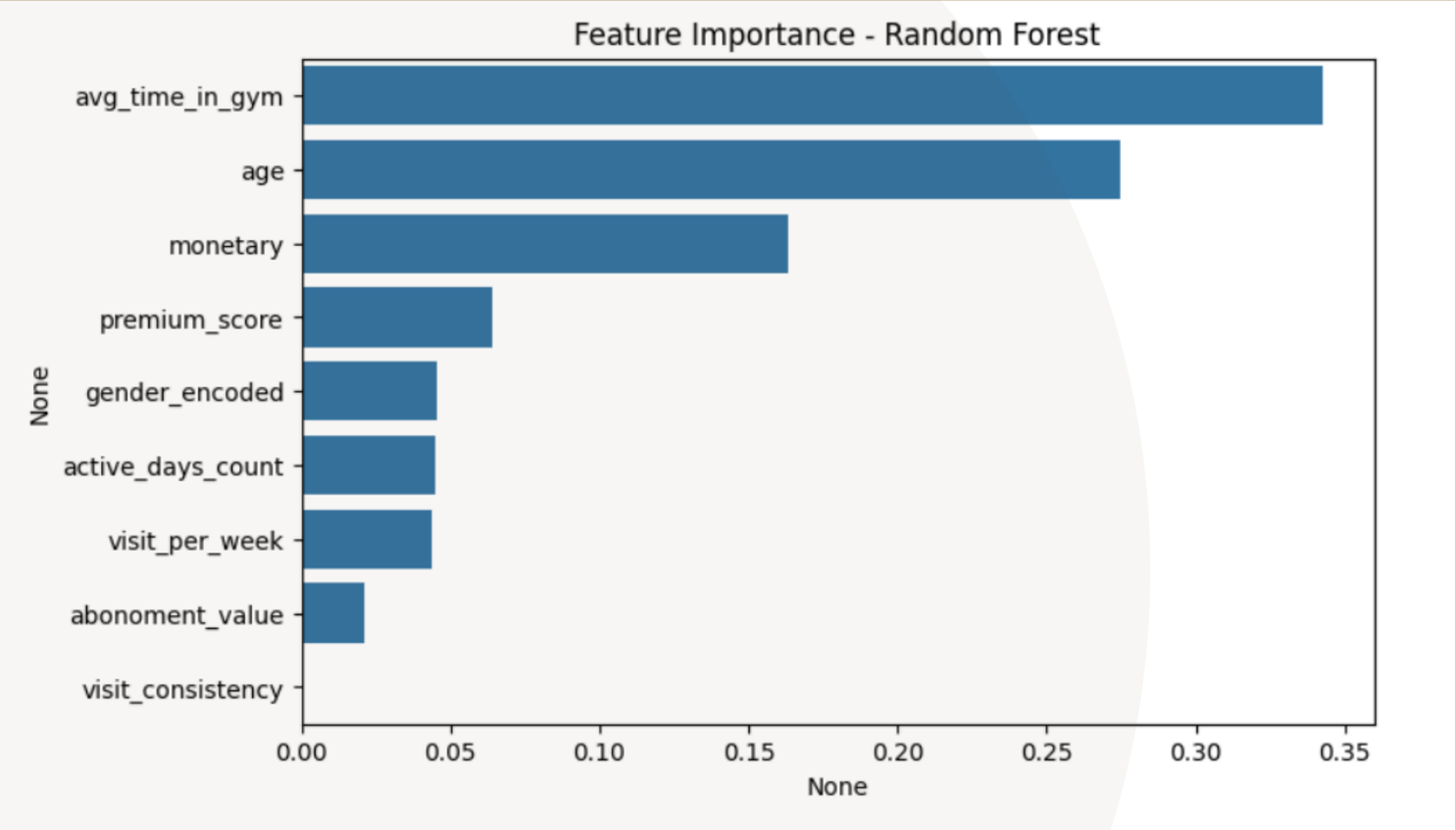
```
=== Random Forest ===
Accuracy: 0.635
ROC-AUC: 0.5459911173445535
              precision    recall  f1-score   support

           0       0.70      0.83      0.76       138
           1       0.34      0.19      0.25        62

    accuracy                           0.64       200
   macro avg       0.52      0.51      0.50       200
weighted avg       0.59      0.64      0.60       200
```

Random Forest memiliki F1-score terbaik, artinya model ini lebih seimbang dalam mendeteksi anggota churn.



- True Negative (115): anggota non-churn yang benar diprediksi non-churn.
- False Positive (23): anggota non-churn yang salah diprediksi churn.
- False Negative (50): anggota churn yang salah diprediksi non-churn.
- True Positive (12): anggota churn yang benar diprediksi churn.

Model mampu mengenali anggota non-churn dengan baik (115 benar vs 23 salah), tetapi lemah dalam mendeteksi churn karena hanya 12 yang terdeteksi dan 50 terlewat, sehingga recall churn rendah dan perlu perbaikan model agar lebih sensitif terhadap anggota yang benar-benar berisiko churn.



Hasil Feature Importance Random Forest menunjukkan bahwa faktor paling berpengaruh terhadap churn adalah rata-rata waktu di gym (avg_time_in_gym), diikuti oleh usia (age) dan monetary (total pengeluaran).

Artinya, semakin singkat waktu yang dihabiskan di gym, usia tertentu, serta rendahnya pengeluaran berhubungan erat dengan risiko churn. Sementara fitur lain seperti premium_score, frekuensi kunjungan, dan konsistensi kunjungan memiliki pengaruh lebih kecil.

7. Conclusion

- Random Forest lebih unggul dibanding Logistic Regression dalam mendeteksi churn, meskipun performanya masih perlu ditingkatkan. Random Forest mampu mengenali sebagian anggota yang benar-benar churn, sementara Logistic Regression cenderung hanya mengklasifikasikan anggota sebagai non-churn.

8. Recommendation

- Target At-Risk Members (Recency 31–60 hari): kirim reminder, promo, atau personal training.
- Tingkatkan Layanan Tambahan: dorong penggunaan kelas, PT, fasilitas premium (mis. bundling / trial gratis).
- Deploy Model Churn Prediction: jalankan monitoring rutin (bulanan) untuk deteksi dini anggota berisiko.
- Personalisasi Retensi: reminder untuk anggota jarang hadir, diskon upgrade untuk anggota dengan monetary rendah.
- Perbaikan Model ML: gunakan SMOTE, hyperparameter tuning, dan tambah data historis untuk akurasi lebih baik.

THANKS YOU

I’m open for discussion and feedback

Let’s Collaborate and Connect

+6285926094075

rezaaprillian0202@gmail.com

Linkedin/RezaAprillianNugroho