



HOME CREDIT SCORECARD

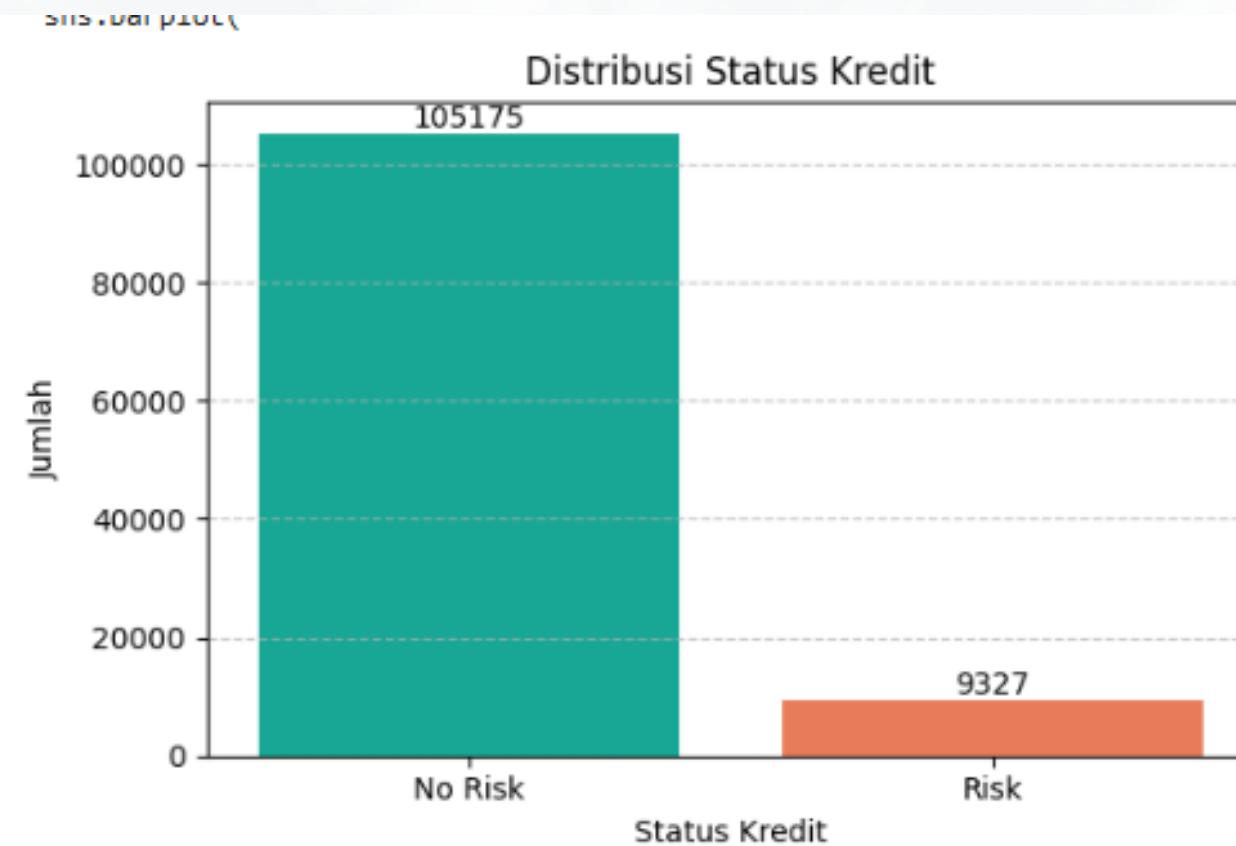
Project-Based Internship: Data Scientist Home Credit
Indonesia x Rakamin Academy

Presented by
Haryadi Tri Nugroho

About Company

Home Credit Indonesia didirikan sebagai perusahaan pembiayaan berbasis teknologi yang beroperasi di Indonesia. Dengan pengalaman di sektor multifinance, Home Credit telah melayani jutaan pelanggan di seluruh negeri, khususnya di segmen ritel dan pembiayaan barang konsumen. Perusahaan ini menawarkan layanan pembiayaan yang berfokus pada pemanfaatan teknologi dan analisis data untuk pengambilan keputusan kredit. Dengan menganalisis perilaku nasabah, Home Credit bertujuan membangun model prediktif yang akurat untuk membedakan nasabah berisiko dan tidak berisiko. Kombinasi antara teknologi, analisis data, dan fokus pada manajemen risiko ini menjadikan Home Credit sebagai mitra pembiayaan yang inovatif, dengan tujuan memastikan pelanggan yang mampu membayar tidak ditolak dan risiko gagal bayar dapat diminimalkan.

Data Understanding



- Dataset yang digunakan pada proyek ini adalah application_train.csv.
- Kolom TARGET menunjukkan label masalah, dengan 0 = nasabah tidak berisiko (No Risk) dan 1 = nasabah berisiko gagal bayar (Risk).
- Dataset terdiri dari 307,511 baris dengan ID unik pinjaman pada kolom SK_ID_CURR.
- Dataset memiliki 122 kolom fitur terkait demografi dan informasi aplikasi kredit.
- Terdapat 106 fitur numerik dan 16 fitur kategorikal.

Data Preprocessing

```
# === 3) Jumlah kolom yang punya missing (hasil sama) ===  
jumlah_missing = df.isnull().any().sum()  
print(jumlah_missing)
```

“Terdapat 96 kolom yang memiliki missing values. Missing values ditangani dengan menghapus kolom yang memiliki missing > 50% dan mengimputasi sisa nilai hilang menggunakan modus (kategorikal) dan median (numerik).”

```
df.isnull().any().sum()
```

```
df.duplicated().sum()
```

Tidak ada duplikat pada dataset

```
df = df.drop(columns=high_card_cols, errors='ignore')  
df = df.drop(columns=['SK_ID_CURR'], errors='ignore')  
  
# Drop semua kolom yang mengandung substring 'FLAG_DOCUMENT' (hasil sama)  
flag_document_cols = df.filter(like='FLAG_DOCUMENT').columns  
df = df.drop(columns=flag_document_cols)
```

Menghapus kolom yang tidak diperlukan

Data Preprocessing

Kolom yang mengandung 'XNA' atau 'Unknown':

`['CODE_GENDER', 'NAME_FAMILY_STATUS', 'ORGANIZATION_TYPE']`

Kolom dengan lebih dari 10 nilai unik (high cardinality):

`['OCCUPATION_TYPE', 'ORGANIZATION_TYPE']`

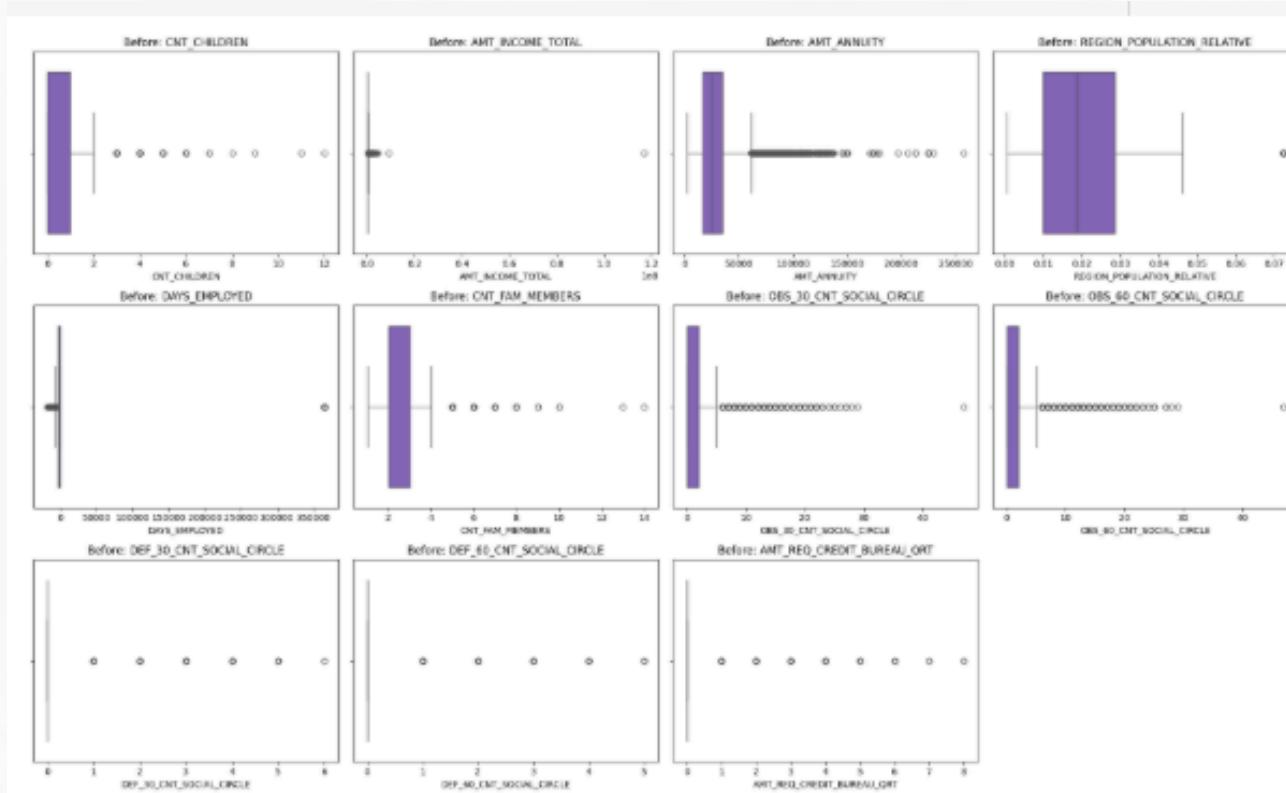
96

Terdapat 3 kolom yang mengandung 'XNA' atau 'Unknown' pada data yang ditangani menggunakan modus

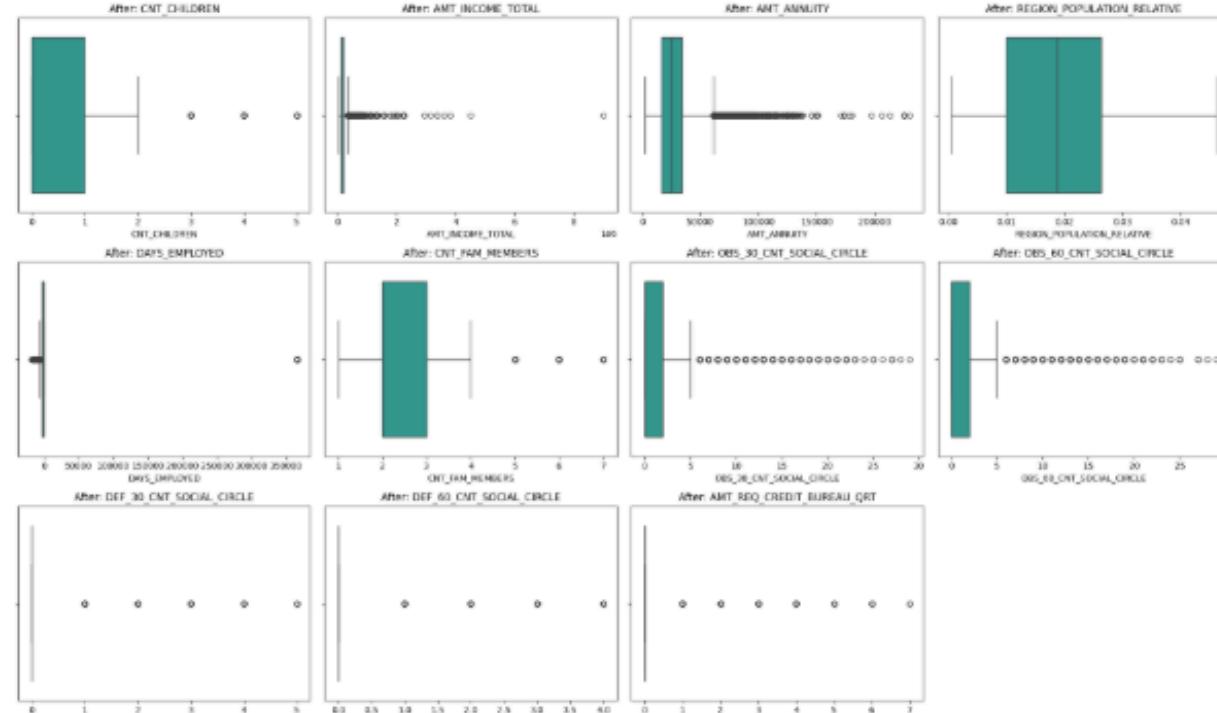
Menghapus kolom yang memiliki lebih dari 10 nilai unik.

Data Outliers

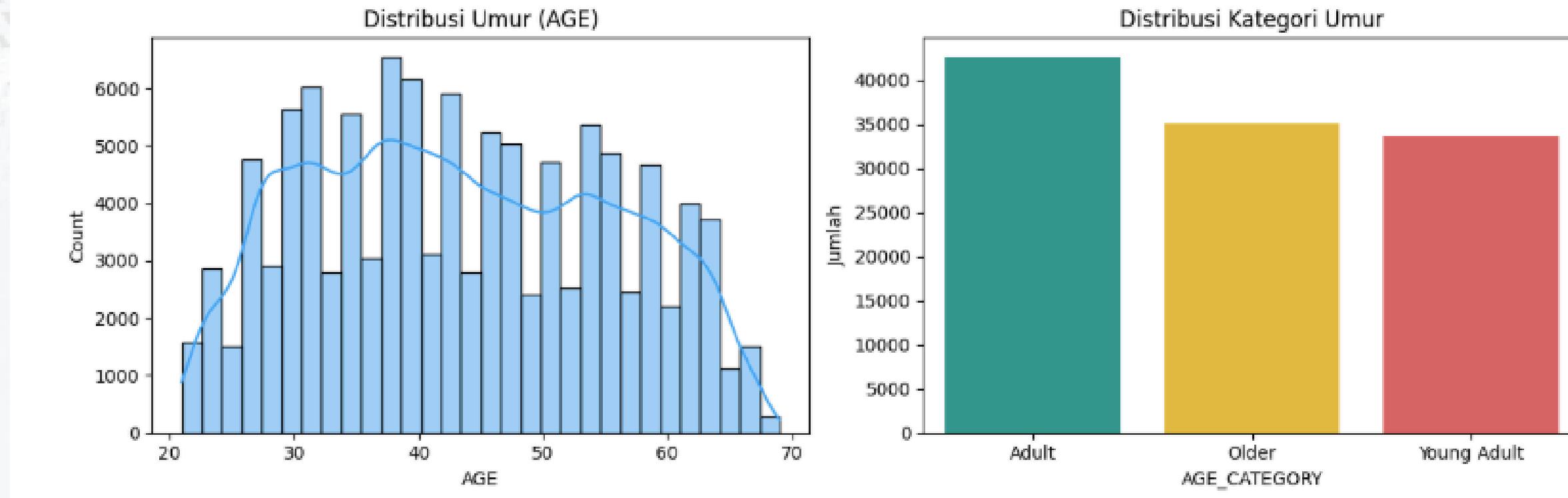
Sebelum



Sesudah

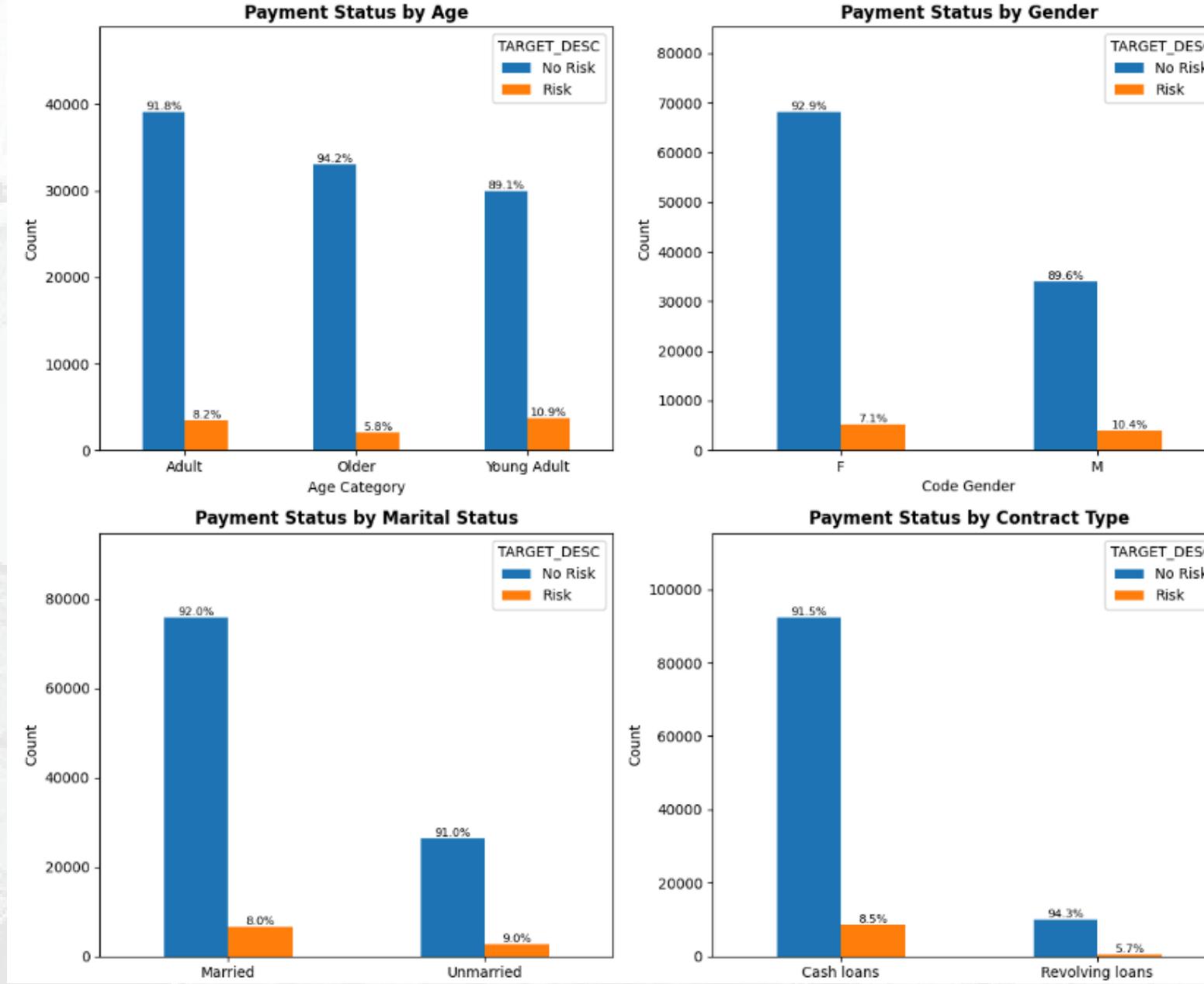


1. Exploratory Data Analysis



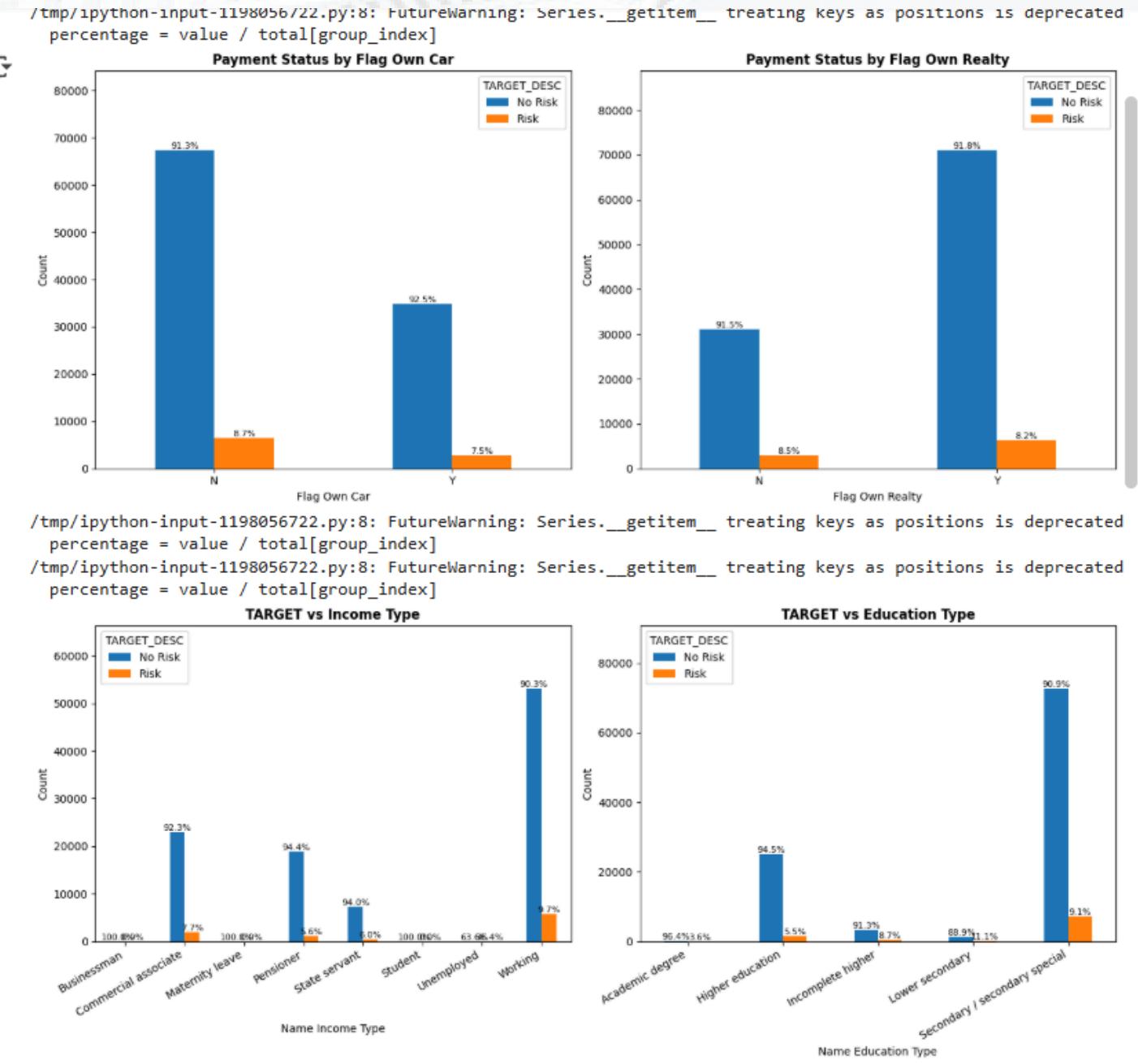
- Sebaran umur nasabah berkisar ~20-70 tahun, dengan puncak di kisaran 30-45 tahun; kurva KDE mulus (unimodal) → mayoritas berada pada usia produktif.
- Pada kategori umur, segmen Adult ($\approx 36-50$ tahun) merupakan yang paling dominan; segmen Young Adult (≤ 35) dan Older (>50) jumlahnya lebih kecil dan relatif berdekatan.
- Setelah pembersihan data (capping/penyaringan), tidak tampak lonjakan ekstrem pada distribusi umur; bentuk histogram tampak stabil dengan penurunan frekuensi setelah usia 50-an.

1. Exploratory Data Analysis



- **Young Adult (≤ 35 tahun) punya risiko tertinggi (10,9%), Older (>50) terendah (5,8%), sedangkan Adult (36–50) berada di tengah (8,2%).**
- **Secara umum, semakin muda usia, semakin tinggi kecenderungan risiko gagal bayar.**
- **Pria lebih berisiko (10,4%) dibanding wanita (7,1%).**
- **Unmarried sedikit lebih berisiko (9,0%) dibanding Married (8,0%).**
- **Cash loans menunjukkan risiko lebih tinggi (8,5%) dibanding Revolving loans (5,7%).**

1. Exploratory Data Analysis



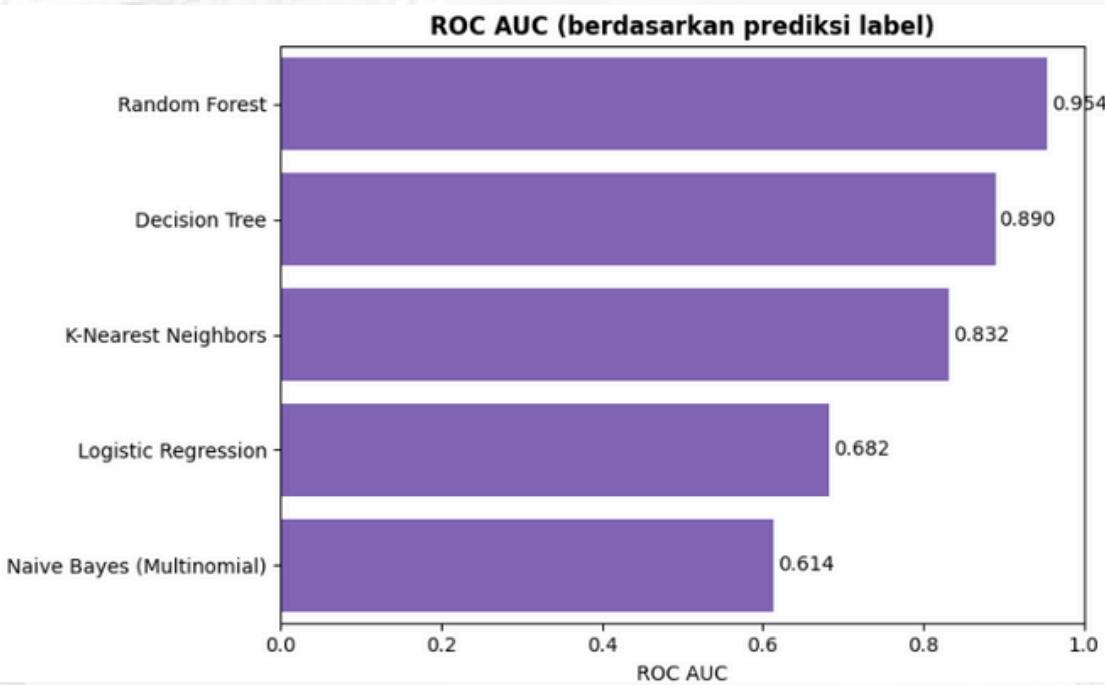
- **Flag Own Car:** nasabah yang memiliki mobil (Y) punya risiko lebih rendah (~7,5%) dibanding yang tidak memiliki (N) ~8,7%.
- **Flag Own Realty:** kepemilikan properti juga terkait risiko sedikit lebih rendah (Y ~8,2% vs N ~8,5%); selisihnya kecil.
- **Income Type:** pola jelas—**Unemployed** mencatat risiko tertinggi (~13,6%). Kelompok **Working/Businessman/Commercial associate/State servant** berada di rentang ~7–9%. **Pensioner** adalah yang terendah (~5,6%). (**Catatan:** kategori dengan jumlah kecil seperti **Student & Maternity leave** tetap ~6–7%.)
- **Education Type:** semakin tinggi pendidikan, semakin rendah risiko. **Academic degree** terendah (~3,5%), **Higher education** ~5%, **Incomplete higher** ~7,8%, **Secondary/secondary special** ~9,1%, dan **Lower secondary** tertinggi (~10,1%).

Feature Engineering

Label Encoder

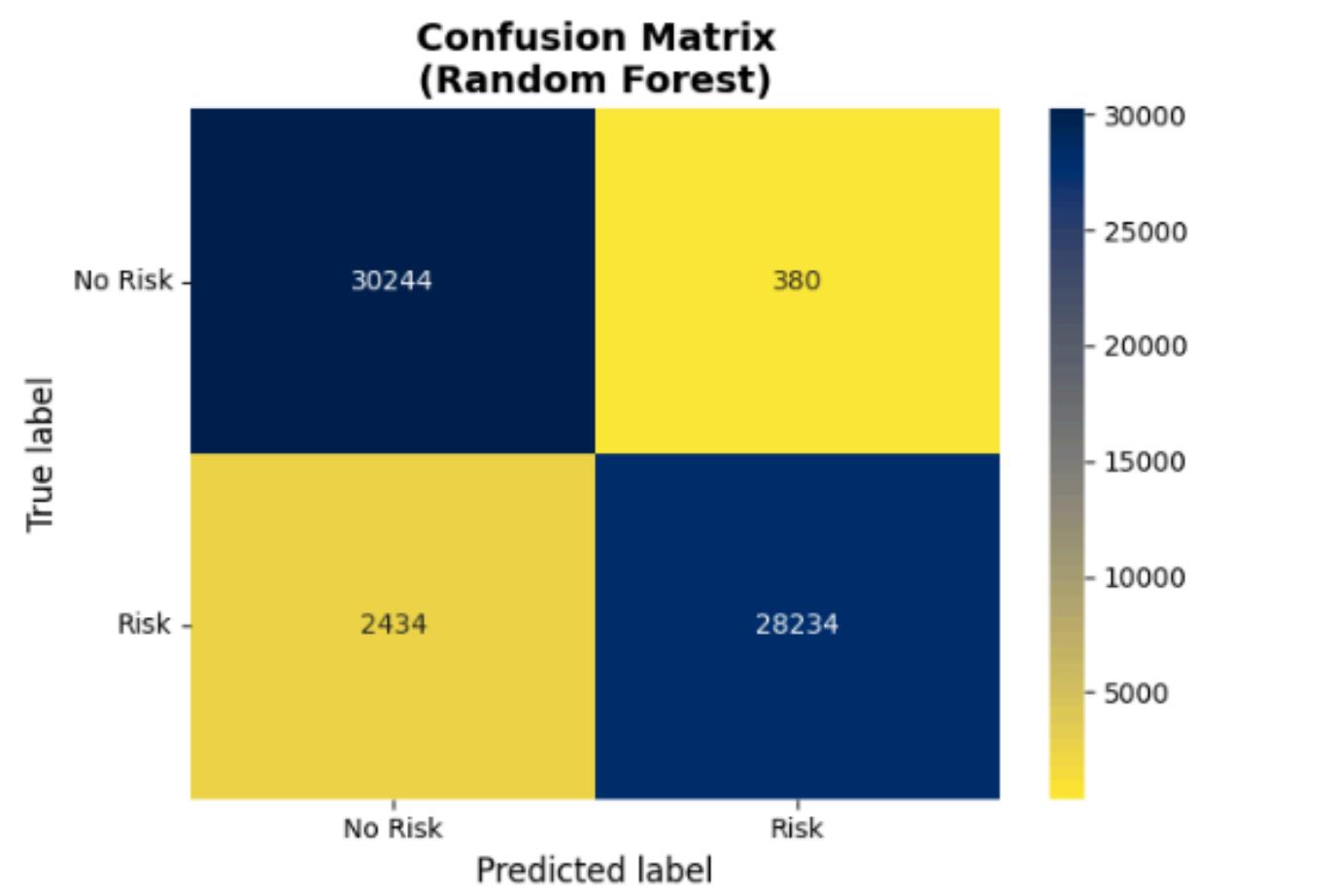
```
'NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY',  
'NAME_TYPE_SUITE', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE', 'NAME_HOUSING_TYPE',  
'WEEKDAY_APPR_PROCESS_START', 'EMERGENCYSTATE_MODE'
```

Modeling dan EVALUATION



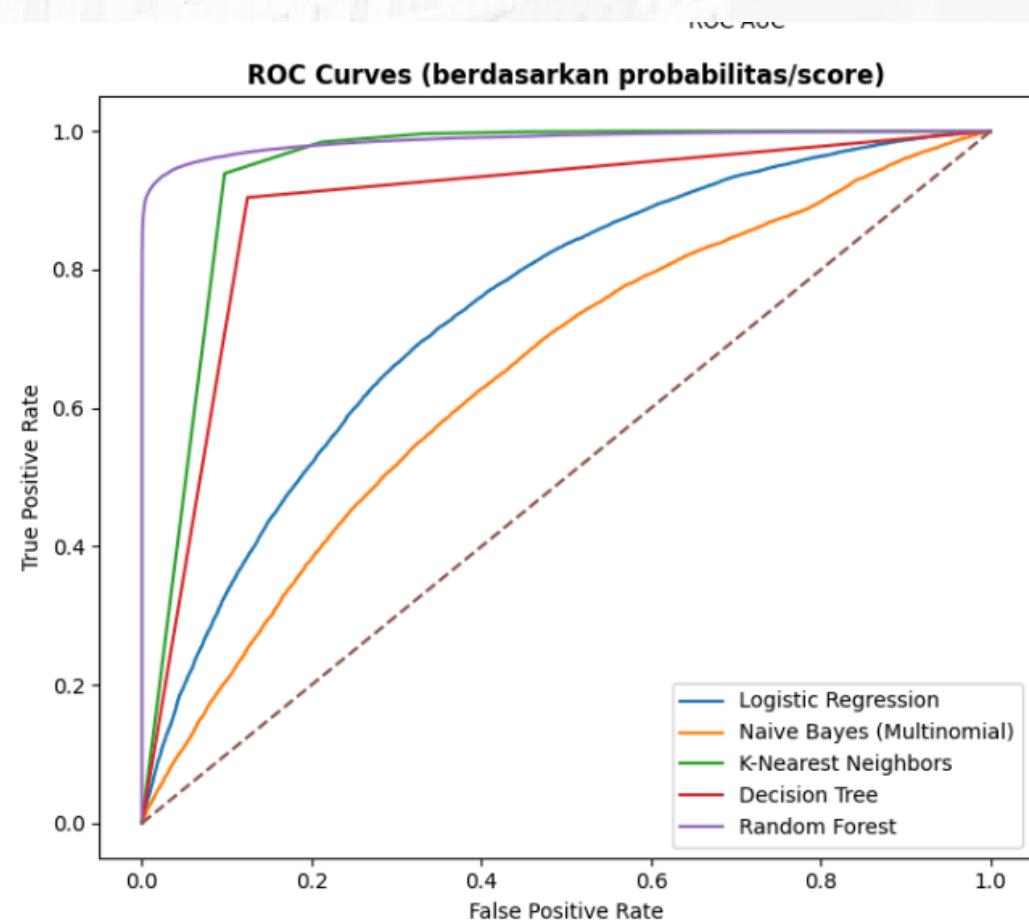
- **Random Forest jadi yang terbaik dengan ROC AUC 0,954, unggul ±0,064 poin dari Decision Tree.**
- **Decision Tree di posisi kedua (ROC AUC 0,890).**
- **K-Nearest Neighbors berada di tengah (ROC AUC 0,832), tertinggal dari model pohon tapi di atas model linear/naive.**
- **Logistic Regression memiliki ROC AUC 0,682.**
- **Naive Bayes (Multinomial) terendah dengan ROC AUC 0,614.**

Modeling dan EVALUATION



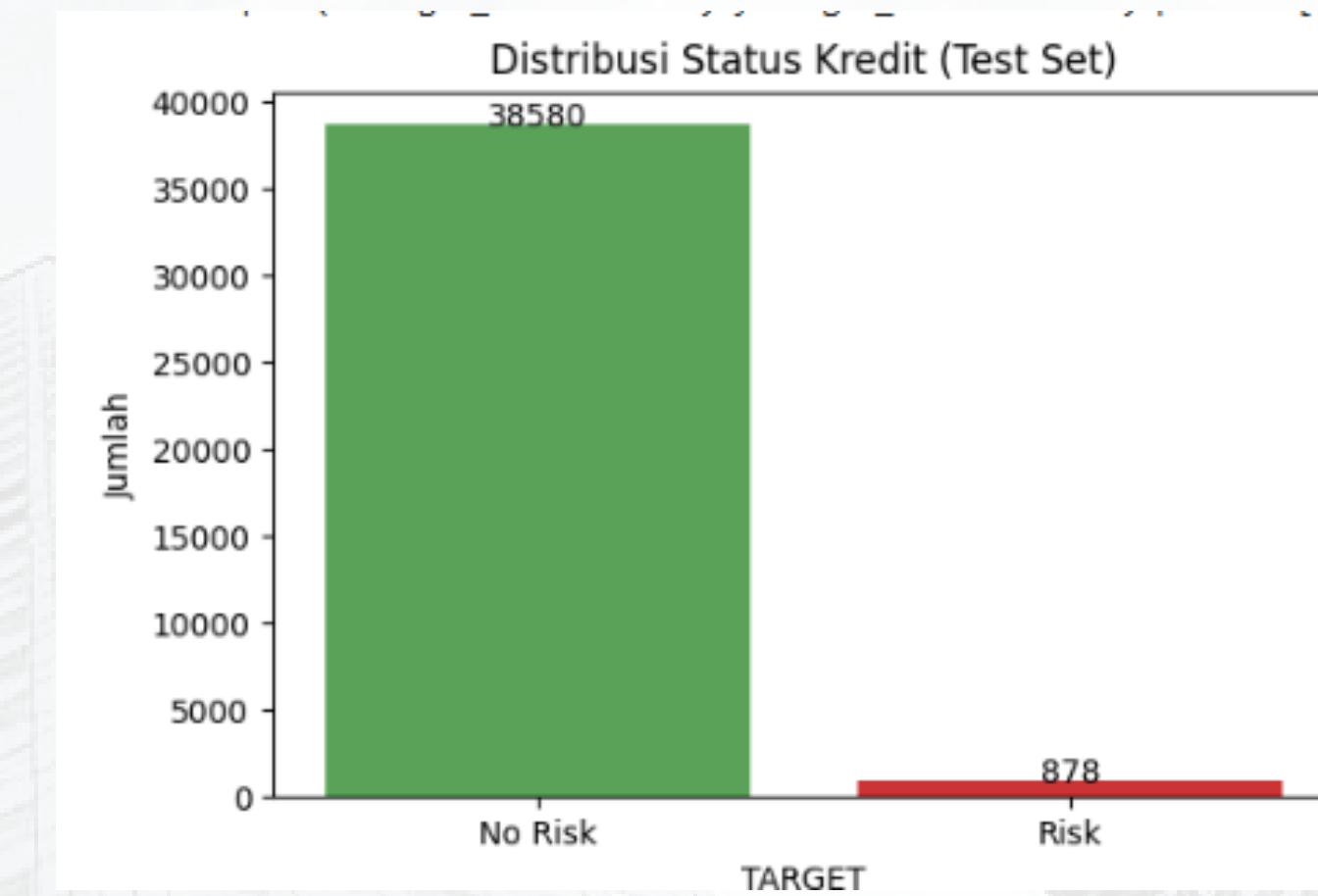
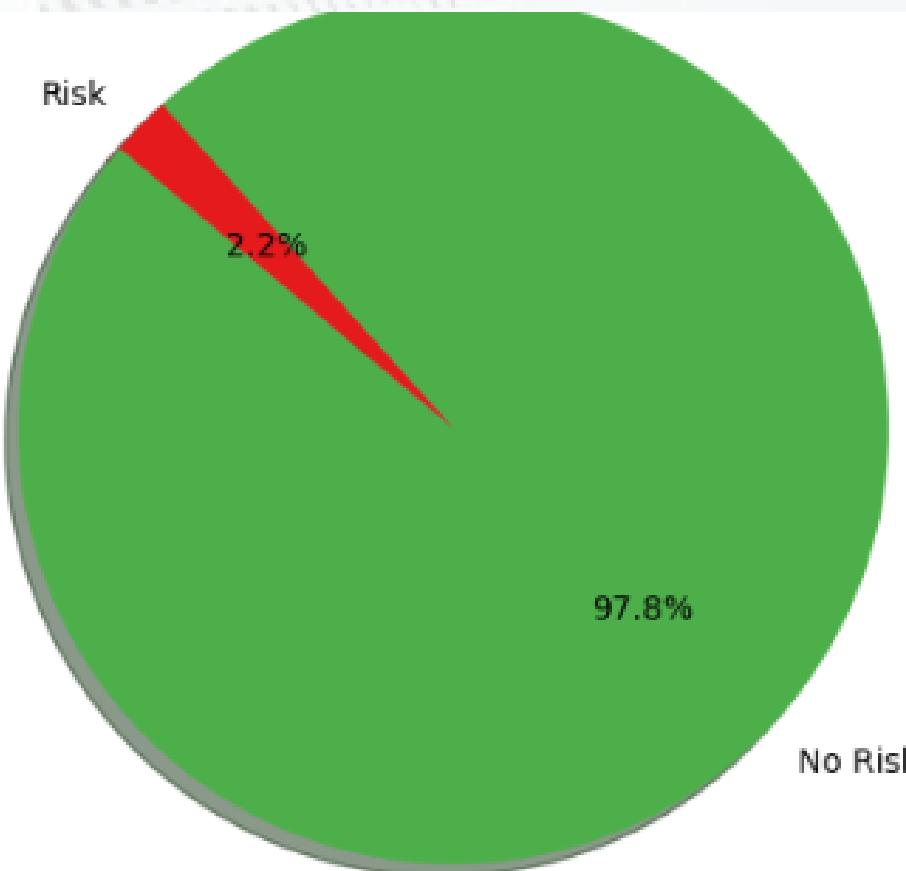
- Akurasi **0.95**; ROC AUC **0.954**.
- Confusion matrix (RF): **TN=30,244 | FP=380 | FN=2,434 | TP=28,234**.
- No Risk (0): precision **0.93**, recall **0.99**.
- Risk (1): precision **0.99**, recall **0.92** → ~8% kasus risk terlewat (**2.434 nasabah**).
- Implikasi: false positive sangat rendah; jika ingin menangkap lebih banyak Risk, pertimbangkan naikkan recall via penyesuaian threshold/class weight.

Modeling dan EVALUATION



- **Random Forest paling unggul: kurva menempel kiri-atas hampir sepanjang sumbu → AUC mendekati 1 dan TPR tinggi bahkan di FPR rendah.**
- **K-Nearest Neighbors menjadi runner-up, umumnya berada di atas Decision Tree pada FPR kecil-menengah; Decision Tree tetap kuat tapi sedikit di bawah KNN.**
- **Logistic Regression menengah; Multinomial Naive Bayes terlemah—kurva lebih dekat ke diagonal, menandakan separasi kelas kurang baik.**
- **Untuk skenario FPR rendah (<0.1), RF (dan KNN) memberi TPR paling tinggi → cocok bila false positive harus ditekan.**
- **Perbedaan dengan grafik AUC berbasis prediksi label wajar: memakai probabilitas memberi perbandingan yang lebih adil antar model; urutannya bisa sedikit bergeser ($KNN \gtrsim DT$).**

Modeling dan EVALUATION



- **Prediksi test set (Random Forest): No Risk 38.580 (97,8%), Risk 878 (2,2%).**
- **Distribusi sangat timpang ke kelas No Risk – terlihat konsisten pada pie chart dan bar chart.**
- **Implikasi: saat deployment, pertimbangkan threshold tuning, precision/recall (PR curve), atau class-weight/cost-sensitive agar deteksi Risk tetap optimal**

CONCLUSION

Kesimpulan

- Mayoritas No Risk 97,8% (Risk 2,2%).
- Kelompok lebih berisiko: Young Adult 10,9%, pria 10,4%, unmarried 9,0%, cash loans 8,5%, unemployed 13,6% (tertinggi).
- Random Forest terbaik: Acc 0,95, AUC 0,954, F1 \approx 0,95; masih ada FN \approx 2.434 (~8%).

Rekomendasi

- Gunakan Random Forest sebagai baseline (pipeline dipaku).
- Naikkan recall Risk: threshold tuning/kalibrasi atau class_weight.
- Terapkan kebijakan kredit lebih ketat pada segmen berisiko (limit kecil, verifikasi ekstra).
- Monitor & retrain berkala; tambah explainability (SHAP) dan cek fairness/compliance.

LINK Youtube

LINK Github

Thank You



X

