



Stage 2 - Preprocessing

Team:

1. Anel Fuad Abiyyu
2. M Febriyan
3. Yudha Prawira

Submission:

1. Report:  Preprocessing - NextDS
 2. Notebook:  Preprocessing - Next DS.ipynb
 3. Github : <https://github.com/nelebaebae/NextDS/>
-

1. Data Eksplorasi

- Deskriptif statistik dijalankan untuk kolom-kolom seperti Income, Age, Experience, CURRENT_JOB_YRS, dll.
 - Distribusi target (Risk_Flag) diperiksa untuk melihat ketidakseimbangan kelas.
 - Struktur data diperiksa menggunakan .info(), .describe(), serta jumlah nilai kosong (.isnull().sum()) dan duplikat (.duplicated().sum()).
-

2. Data Cleansing

- **Outlier Removal (Tidak Ada Outlier):**
 - Kolom numerik (Income, Age, Experience, CURRENT_JOB_YRS, CURRENT_HOUSE_YRS) diperiksa untuk outlier menggunakan boxplot.
 - Outlier dihapus menggunakan metode **Interquartile Range (IQR)**.
 - **Handling Missing Values (Tidak Ada Missing Values):**
 - Tidak ada langkah eksplisit yang menangani nilai kosong, artinya data diasumsikan tidak memiliki nilai null.
 - **Handling Duplicates (Tidak Ada Data Duplikat):**
 - Tidak ditemukan data duplikat.
-

4. Data Transformation

- **Scaling:**
 - Kolom Income di-scale menggunakan **Min-Max Scaler** menjadi Income_Scaled.

- **Encoding:**
 - Kolom kategorikal seperti Married/Single, House_Ownership, dan Car_Ownership diubah menjadi format numerik menggunakan **Label Encoding**.
-

5. Data Balancing

- Ketidakseimbangan kelas pada kolom target Risk_Flag diatasi menggunakan **oversampling** pada kelas minoritas dengan **resampling**.
 - Dataset diseimbangkan dengan jumlah sampel yang sama untuk kedua kelas.
-

6. Data Engineering

Beberapa fitur baru ditambahkan untuk memperkaya dataset:

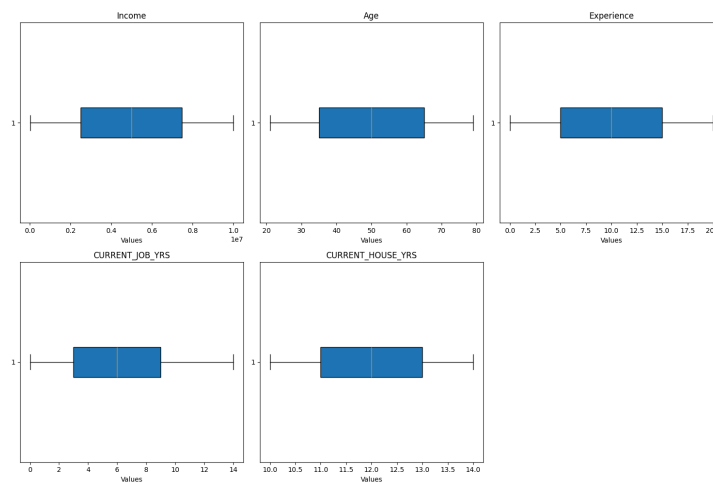
1. **Income_Category:**
 - Kategori berdasarkan kuartil Income:
 - Low: di bawah kuartil pertama.
 - Medium: antara kuartil pertama dan ketiga.
 - High: di atas kuartil ketiga.
 2. **Years_to_Retirement:**
 - Menghitung sisa tahun sebelum pensiun dengan asumsi usia pensiun adalah 60 tahun.
 3. **Ownership Stability Score (OSS):**
 - Skor dihitung berdasarkan:
 - Kategori pendapatan (Income_Category): High (3), Medium (2), Low (1).
 - Kepemilikan rumah (House_Ownership): Tidak punya (0), Sewa (1), Punya rumah (2).
 - Kepemilikan mobil (Car_Ownership): Tidak punya (0), Punya (1).
 4. **Early_Stability:**
 - Fitur biner yang menentukan apakah seseorang memiliki pekerjaan stabil:
 - 1 jika CURRENT_JOB_YRS > 1.
 - 0 jika sebaliknya.
-

7. Feature Importance

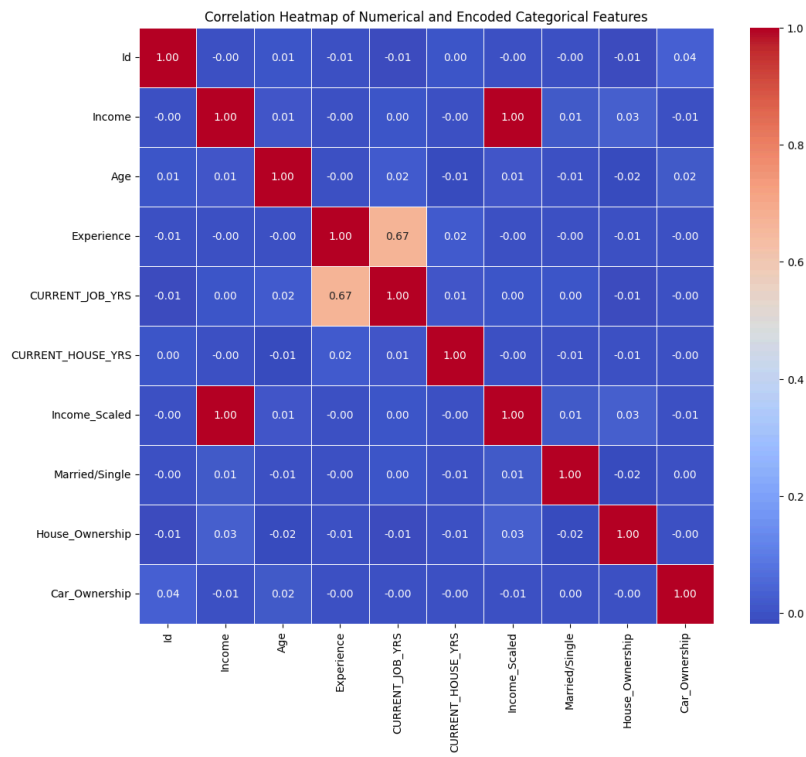
- Model **Random Forest** dilatih untuk menentukan tingkat kepentingan fitur.
- Hasil:
 - Menampilkan ranking fitur berdasarkan kepentingan prediktif terhadap target Risk_Flag.

8. Visualisasi

- **Outliers:**
 - Boxplot untuk mendeteksi outlier pada kolom numerik.



- **Correlation Heatmap:**
 - Visualisasi korelasi antar fitur numerik untuk memahami hubungan antara variabel.



- **Feature Importance:**

- Grafik batang untuk menunjukkan kepentingan fitur yang dihitung oleh model Random Forest.

