STRATEGI *DATA-DRIVEN* UNTUK MENGURANGI RISIKO GAGAL BAYAR PINJAMAN PRIBADI

Nama Tim **:** Menang Gak Ya

Anggota Tim :

1. Aulia Mirfah Setyo Ayu Damayanti
2. Erlin Shofiana
3. Fitri Hartanti

DATA SCIENCE COMPETITION

ICONICIT UNIVERSITAS SILIWANGI

2024

I. PENDAHULUAN

A. Latar Belakang

Dalam dunia perbankan modern, salah satu tantangan terbesar adalah menyeimbangkan antara pertumbuhan bisnis dan pengelolaan risiko. Bank dan lembaga keuangan menghadapi tantangan dalam menilai kelayakan kredit calon peminjam dengan akurat. Proses ini sangat penting, terutama dalam persetujuan pinjaman pribadi, di mana keputusan yang salah bisa berakibat pada kerugian finansial yang signifikan. Oleh karena itu, lembaga keuangan semakin mengandalkan pendekatan berbasis data untuk memprediksi kemampuan pemohon dalam melunasi pinjaman. Sistem penilaian ini harus mempertimbangkan berbagai aspek dari profil peminjam, seperti riwayat keuangan, stabilitas pekerjaan, dan kemampuan membayar kembali pinjaman.

Seiring dengan perkembangan teknologi, bank mulai mengadopsi teknik analisis data yang lebih canggih untuk membuat keputusan yang lebih baik dan mengurangi risiko kredit macet. Mereka tidak hanya bergantung pada data tradisional seperti catatan perbankan dan laporan kredit, tetapi juga memanfaatkan data tidak terstruktur seperti pola perilaku digital dan aktivitas transaksi yang lebih luas. Dengan menggabungkan berbagai sumber data ini, bank dapat membangun model prediksi yang lebih akurat dan komprehensif, yang mampu menilai risiko secara lebih mendalam dan spesifik.

Dalam konteks ini, analisis data menjadi alat yang sangat penting untuk pengambilan keputusan yang tepat di dunia ekonomi yang semakin kompleks. Salah satu pendekatan yang sering digunakan adalah metode klasifikasi, yang bertujuan untuk memprediksi kategori atau kelas dari suatu entitas berdasarkan karakteristiknya. Penggunaan metode klasifikasi dalam penyelesaian masalah ekonomi dapat membantu berbagai pihak, termasuk lembaga keuangan, dalam membuat keputusan yang lebih akurat dan berbasis data.

Penelitian ini menggunakan dua dataset utama untuk membangun model klasifikasi yang mampu memprediksi kemampuan individu dalam membayar kredit. Dataset pertama berisi informasi demografis dan keuangan individu, seperti jenis kelamin, status kepemilikan mobil dan properti, jumlah anak, pendapatan tahunan, tingkat pendidikan, status pernikahan, jenis tempat tinggal, serta data mengenai pekerjaan dan akses ke teknologi komunikasi. Dataset kedua hanya berisi dua kolom, yaitu identifikasi unik individu dan label target. Label target ini merupakan klasifikasi biner di mana nilai 0 menunjukkan bahwa individu dapat membayar kredit, sedangkan nilai 1 menunjukkan bahwa individu gagal membayar kredit.

Dengan menggunakan Python sebagai alat analisis, penelitian ini akan menerapkan berbagai teknik klasifikasi untuk membangun model prediktif yang dapat mendukung pengambilan keputusan di berbagai bidang ekonomi, khususnya dalam penilaian risiko kredit. Model ini diharapkan dapat memberikan wawasan yang berharga dalam mengelola risiko kredit secara lebih efektif dan efisien.

B. Tujuan

Tujuan utama dari proyek ini adalah untuk meningkatkan keakuratan prediksi persetujuan pinjaman pribadi dengan mengklasifikasikan risiko gagal bayar peminjam berdasarkan label: '1' untuk risiko tidak bisa membayar dan '0' untuk risiko bisa membayar. Kami berfokus pada data yang tersedia yang lebih mendekati informasi pribadi peminjam, seperti jenis kelamin, kepemilikan kendaraan dan properti, jumlah anak, pendapatan tahunan, jenis pendapatan, tingkat pendidikan, status perkawinan, jenis tempat tinggal, usia, lama bekerja, dan informasi kontak seperti telepon dan email. Tantangan utamanya adalah mengintegrasikan berbagai jenis data ini ke dalam satu model klasifikasi yang efektif untuk mengevaluasi risiko kredit secara akurat.

Pendekatan ini tidak hanya bertujuan untuk mengurangi potensi kerugian akibat kredit macet, tetapi juga meningkatkan profitabilitas bank dengan cara memberikan penilaian yang lebih tepat terhadap risiko kredit. Dengan menggunakan teknik analisis data seperti pembelajaran mesin, kami berusaha memahami pola-pola yang tersembunyi dalam data yang tersedia, untuk mengidentifikasi peminjam yang memiliki risiko tinggi dan memastikan keputusan kredit yang lebih bijaksana dan tepat waktu.

Proyek ini juga bertujuan untuk mengeksplorasi penggunaan data yang lebih dalam, seperti status kepemilikan properti atau jumlah anggota keluarga, yang dapat memberikan wawasan lebih lanjut tentang stabilitas finansial peminjam. Dengan pemahaman yang lebih baik tentang faktor-faktor ini, kami dapat mengembangkan model prediksi yang lebih tangguh dan mampu menangkap dinamika risiko yang kompleks, sekaligus mendukung upaya bank untuk memberikan layanan kredit yang lebih aman dan efisien.

C. Rencana Penelitian

Proses analisis untuk mengembangkan model penilaian kredit dalam penelitian ini akan disusun berdasarkan kerangka kerja CRISP-DM (Cross-Industry Standard Process for Data Mining) dengan langkah-langkah sebagai berikut:

Gambar 1. Alur Analisis

II. METODOLOGI

A. Data Understanding

Pada analisis ini digunakan 2 dataset diantaranya:

- Dataset 1: Berisi informasi mengenai demografis dan keuangan dari 1548 individu, yang dijelaskan dari 17 fitur atau variabel. Keterangan dari tiap fitur adalah sebagai berikut.

Tabel 1. Deskripsi fitur dalam Dataset 1

Variable Name	Role	Description
Ind_ID	*Feature*	Identifikasi unik untuk setiap individu.
Gender	*Feature*	Jenis kelamin individu (M untuk laki-laki, F untuk perempuan).
Car_Owner	*Feature*	Status kepemilikan mobil (Y untuk ya, N untuk tidak).
Propert_Owner	*Feature*	Status kepemilikan properti (Y untuk ya, N untuk tidak).
Children	*Feature*	Jumlah anak yang dimiliki individu.
Annual_Income	*Feature*	Pendapatan tahunan individu.
Type_Income	*Feature*	Sumber pendapatan individu (misalnya, Pensiunan, Pegawai Komersial).
Education	*Feature*	Tingkat pendidikan individu.
Marital_status	*Feature*	Status pernikahan individu.
Housing_type	*Feature*	Jenis tempat tinggal (misalnya, Rumah atau apartemen).
Birthday_count	*Feature*	Jumlah usia dalam hari.
Employed_days	*Feature*	Jumlah hari individu telah bekerja (nilai negatif menunjukkan jumlah hari bekerja di pekerjaan terakhir).
Mobile_phone	*Feature*	Apakah individu memiliki ponsel (1 untuk ya, 0 untuk tidak).
Work_Phone	*Feature*	Apakah individu memiliki telepon kerja (1 untuk ya, 0 untuk tidak)
Phone	*Feature*	Apakah individu memiliki telepon rumah (1 untuk ya, 0 untuk tidak).
Email_ID	*Feature*	Apakah individu memiliki email (1 untuk ya, 0 untuk tidak)
Type_Occupation	*Feature*	Jenis pekerjaan individu (banyak nilai yang hilang/NaN).
Family_Members	*Feature*	Jumlah anggota keluarga.

- Dataset 2: Berisi label klasifikasi untuk setiap individu, menunjukan sanggup (0) dan tidaknya (1) seorang peminjam dalam memenuhi kewajibannya melunasi pinjaman dengan tepat waktu.

i. Data Cleaning

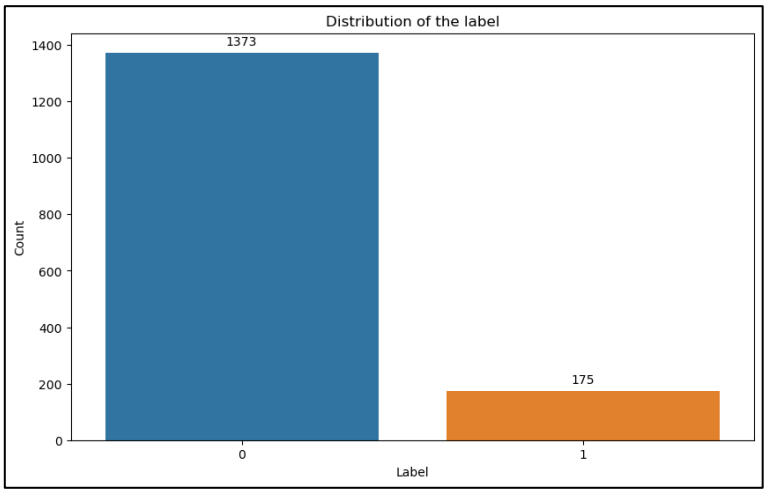
Sebelum dilakukan analisis lebih lanjut, dilakukan proses *data cleaning* dengan tujuan memperbaiki dataset yang tidak konsisten dan tidak akurat. Berikut beberapa proses *data cleaning* yang dilakukan:

1. Memeriksa keberadaan data duplikat dan hasilnya menunjukkan bahwa tidak ada data duplikat pada dataset.
2. Mengubah variabel kategorik yakni Mobile_phone, Work_phone, Phone, dan EMAIL_ID menjadi bertipe data *object*.
3. Membentuk fitur Age yang diperoleh dengan mengekstrak nilai pada fitur Birthday_count menjadi tahun
   Contoh:
   $$Age = \frac{Birthday\_count}{-365}$$
4. Membentuk fitur Employed_years yang diperoleh dengan mengekstrak nilai pada fitur Employed_days menjadi tahun. Namun, pada fitur Employed_days awalnya terdapat nilai 365243, setelah diidentifikasi semua individu dengan nilai tersebut pasti juga seorang *pensioner*, sehingga individu dengan nilai 365243 pada Employed_days akan memiliki nilai 0 pada Employed_years yang berarti individu tersebut sudah tidak bekerja.
   Contoh:
   $$Employed\_years = \frac{Employed\_days\ /\ -365}{-365}$$
   jika Employed_days $\neq$ 365243
   $$Employed\_years = 0$$
   jika Employed_days $=$ 365243

***ii. Exploratory Data Analysis* (EDA)**

Selanjutnya, dilakukan eksplorasi data untuk mendapatkan gambaran awal dan mengidentifikasi masalah yang terdapat pada data sehingga dapat diatasi pada bagian berikutnya sebelum memasuki tahapan pemodelan.

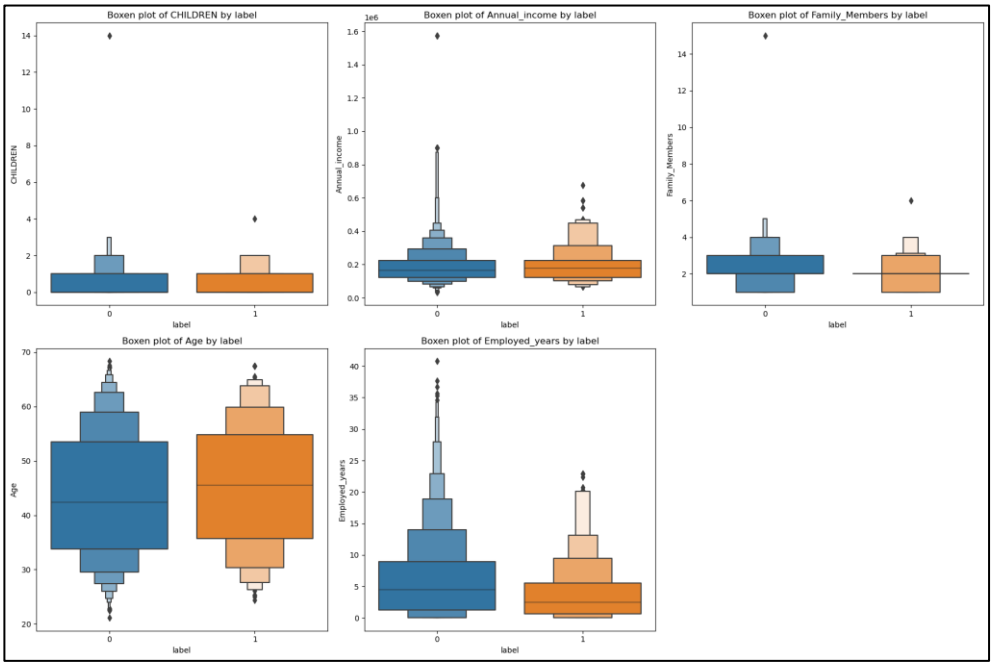
1. Distribusi Variabel Target



Gambar 2. Distribusi Variabel Target (Status Gagal Bayar)

Berdasarkan visualisasi barplot di atas, dapat diketahui sebaran status gagal bayar dari setiap individu tidak seimbang. Yang mana 1373 individu memiliki label 0 (sanggup bayar) dan 175 individu memiliki label 1 (tidak sanggup bayar.

2. Variabel Numerik

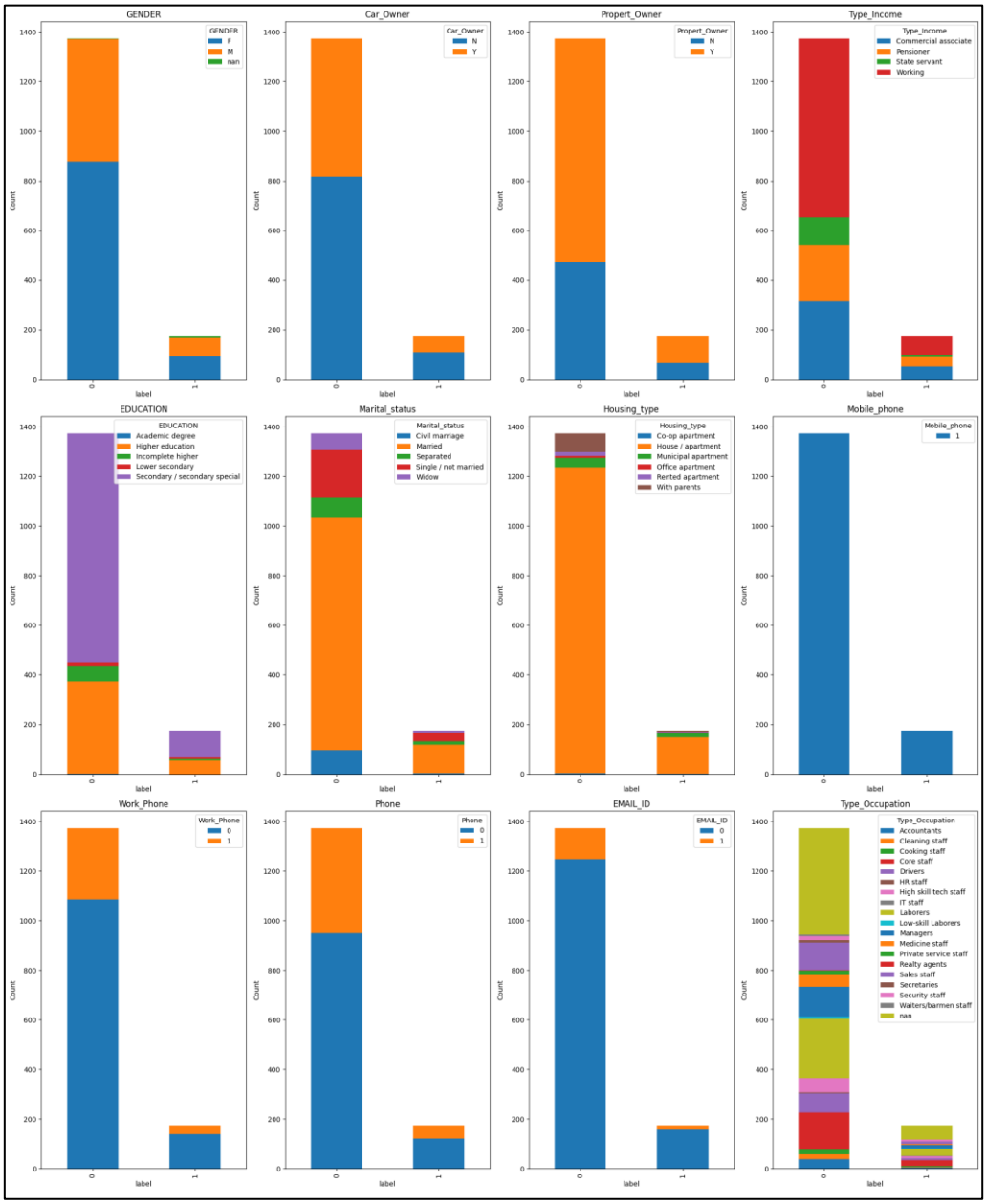


Gambar 3. Boxenplot Variabel Numerik terhadap Target

Visualisasi di atas menunjukkan perbedaan sebaran nilai tiap variabel numerik berdasarkan level kategori dari label target. Misalnya pada variabel

kelima yaitu Employed_years dapat dilihat bahwa sebagian besar individu yang berlabel 0 (sanggup bayar) memiliki nilai Employed_years yang cukup tinggi. Selain itu, melalui boxenplot, terungkap bahwa seluruh variabel numerik (CHILDREN, Annual_income, Age, Employed_years, Family_Members) tidak berdistribusi normal. Semua variabel menunjukkan distribusi skewed to the right, di mana nilai rendah lebih terpusat dan nilai tinggi lebih menyebar.

3. Variabel Kategorik



Gambar 4. Sebaran Variabel Kategorik Berdasarkan Kategori Target

Visualisasi di atas, menunjukkan sebaran nilai tiap variabel kategorik berdasarkan level kategori dari label targetnya, misalnya pada variabel pertama yaitu Gender, dapat dilihat bahwa terdapat dua jenis kelamin yaitu F (perempuan) dan M (laki-laki). Pada kedua label 0 dan 1, jenis kelamin perempuan jumlahnya lebih besar daripada jenis kelamin laki-laki. Selain itu, untuk variabel Work _Phone, Phone, dan Email_ID, mayoritas data baik untuk label 0 maupun 1, didominasi oleh nilai 0, yang menunjukkan tidak memiliki telepon atau email.

4. *Correlation Heat Map*

Gambar 5. Visualisasi Korelasi Antar Variabel

Berdasarkan heatmap di atas, dapat dilihat bahwa pada heatmap korelasi, variabel Mobile_Phone menunjukkan warna putih terhadap semua variabel numerik lainnya. Warna putih ini menandakan bahwa korelasi antara Mobile_Phone dengan variabel-variabel numerik lainnya adalah 1. Hal ini berarti bahwa variabel Mobile_Phone memiliki korelasi sempurna dengan variabel numerik lainnya di dataset. Korelasi sempurna seperti ini biasanya mengindikasikan bahwa variabel Mobile_Phone tidak membawa informasi tambahan yang berguna untuk model karena nilainya mungkin konstan atau sangat seragam dalam kaitannya dengan variabel lain. Oleh karena itu, variabel

ini akan dihapus dari dataset untuk mengurangi redundansi dan menyederhanakan model.

Selain itu, variabel CHILDREN dan Family Number menunjukkan korelasi yang sangat tinggi, yaitu sebesar 0,89. Nilai korelasi yang mendekati 1 ini menunjukkan hubungan linier yang sangat kuat antara kedua variabel. Korelasi sebesar 0,89 ini artinya jika jumlah CHILDREN (jumlah anak) meningkat, maka Family Number (jumlah anggota keluarga) juga cenderung meningkat secara proporsional. Hal ini logis karena semakin banyak jumlah anak dalam sebuah keluarga, maka jumlah total anggota keluarga juga akan bertambah.

B. Data Preparation

Pada bagian ini, akan dilakukan proses pembersihan data berdasarkan wawasan yang telah diperoleh dari tahap pemahaman data (*data understanding*). Tujuan dari langkah ini adalah untuk memastikan bahwa data yang digunakan dalam pemodelan berada dalam kondisi yang optimal, bebas dari anomali, dan siap untuk dianalisis lebih lanjut.

i. Reformat Categorical Variables

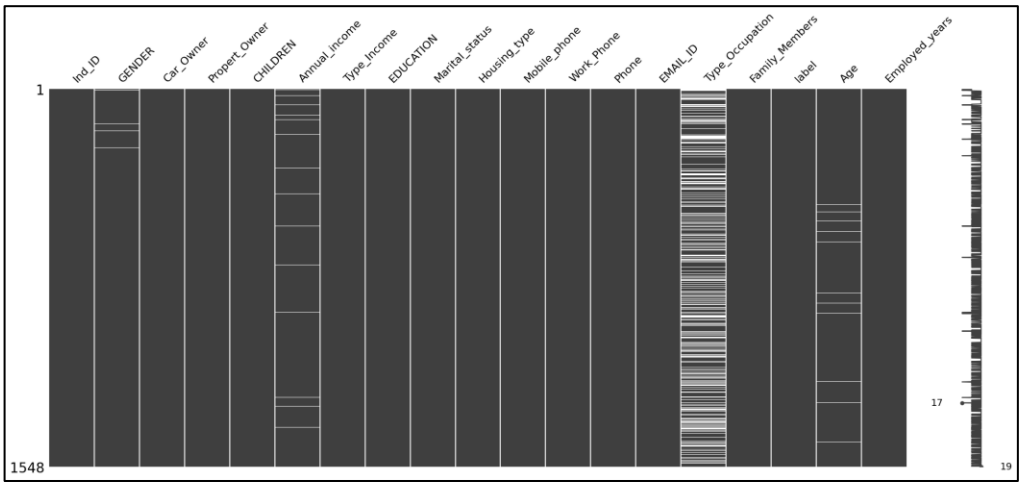
Reformat variabel kategorik bertujuan untuk meningkatkan konsistensi data, memudahkan analisis, dan mengurangi dimensionalitas data. Proses reformat categorical variables yang dilakukan meliputi:

- Pada variabel Marital_status, kategori widow dan separated digabung menjadi satu kategori separated karena keduanya serupa dan proporsinya sedikit. Selain itu, kategori lainnya dibiarkan tetap sama.
- Pada variabel Housing_type, kategori Co-op apartment dan office apartment digabung menjadi satu kategori office apartment karena keduanya serupa dan proporsinya sedikit. Kategori lainnya tetap dibiarkan seperti semula.
- Pemetaan kategori pada variabel type occupation dilakukan sebagai berikut: Core staff diubah menjadi Core Staff, High skill tech staff, Managers, IT staff, dan HR staff diubah menjadi Technical, Drivers, Security staff, dan Cleaning staff diubah menjadi Operational, Cooking staff, Sales staff, Waiters/barmen staff, dan Private service staff diubah menjadi Service, Accountants, Secretaries, dan Medicine staff diubah menjadi Administrative, serta Laborers dan Low-skill Laborers diubah menjadi Labor. Kategori Realty agents juga diubah menjadi Service.

ii. Splitting Data

Sebelum melakukan *modelling* dilakukan *splitting data* dengan proporsi 80% menjadi data train dan 20% menjadi data test.

iii. Penanganan *Missing Values*



Gambar 6. Plot Missing Value

Berdasarkan Gambar 6, terdapat *missing value* pada beberapa fitur dengan proporsi 0,45% untuk variabel Gender, 1,49% untuk Annual_income, 31,52% untuk variabel Type_Occopation, dan 1,42 % untuk Age. Untuk variabel Type_Occupation, kategori dengan Type_Income yang berstatus pensioner diisi dengan nilai unknown. Untuk variabel gender dan Sisa nilai yang hilang pada variabel Type_Occupation yang merupakan variabel kategorik dilakukan imputasi *missing value* dengan modus, selanjutnya untuk variabel Annual_income dan Age dilakukan penanganan dengan teknik *iterative imputer*. Teknik imputasi *missing value* tersebut dijalankan mengunakan operator *pipeline*.

iv. Categorical Data Encoding

Sebelum diproses ke dalam model *machine learning*, fitur dengan tipe data kategorik harus diubah ke numerik terlebih dahulu. Untuk tahap ini digunakan metode One Hot Encoding dengan operator *pipeline*.

v. Scaling Data

Proses scaling dilakukan pada fitur numerik dengan menggunakan teknik *standard scaling* dan diimplementasikan menggunakan operator *pipeline.*

vi. Penanganan Imbalance Data

Data train yang diperoleh dan telah dilakukan imptasi *missing value* kemudian dilakukan penanganan *imbalace* dengan membandingkan akurasi beberapa teknik *oversampling* yaitu SMOTE dan ADASYN.

III. PEMBENTUKAN MODEL DAN EVALUASI

A. Pembentukan Model

Pemodelan dilakukan untuk menentukan kemampuan bayar calon debitur berdasarkan karakteristik demografis dan keuangan setiap individu menggunakan 6 algoritma *linear-based model* dan *tree-based model*. Metriks evaluasi yang digunakan pada pemodelan ini adalah *recall* karena *recall* mengukur mengukur seberapa baik model dalam menangkap semua kasus "gagal bayar" yang sebenarnya terjadi di antara seluruh debitur yang benar-benar gagal bayar. Dengan fokus pada *recall*, model menjadi lebih sensitif dalam mendeteksi semua kemungkinan gagal bayar.

Tabel 2. Hasil Pembentukan Model

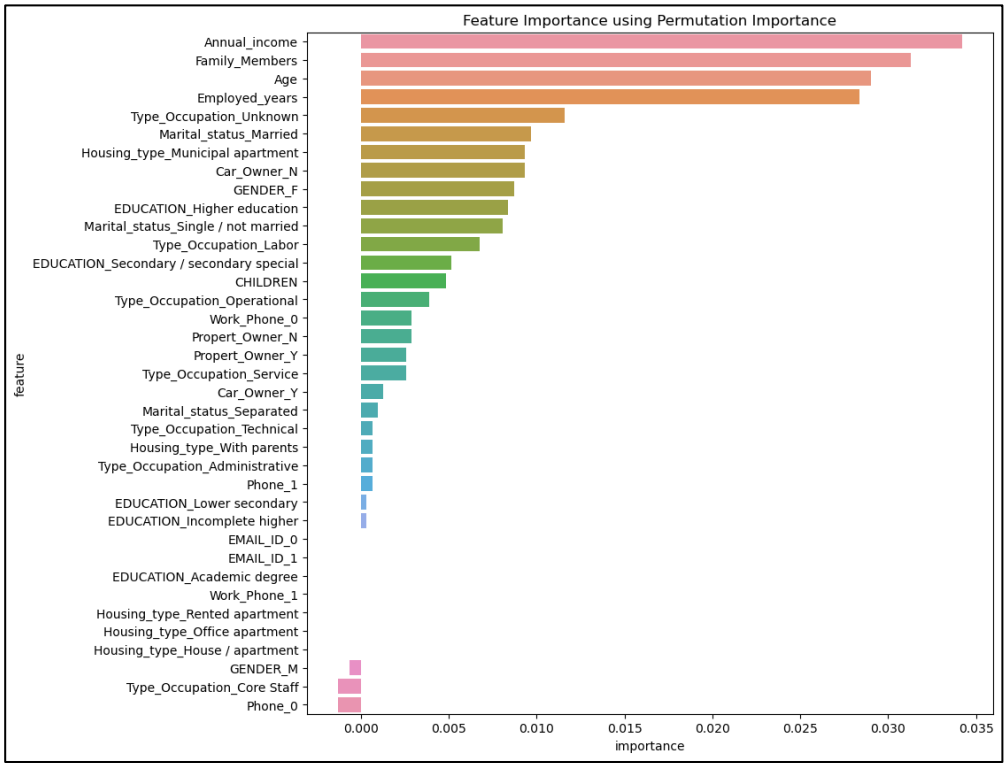
	SMOTE				ADASYN			
Algoritma	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy
Logistic Regression	0,847614	0,580645	0,656632	0,580645	0,844173	0,558065	0,637006	0.558065
Gradient Boosting Classifier	0,850912	0,870968	0,858997	0,870968	0,862071	0,877419	0,868299	0.877419
Hist Gradient Boosting Classifier	0,903433	0,912903	0,905638	0,912903	0,923243	**0,929032**	0,921042	0.929032
CatBoost Classifier	0,898911	0,909677	0,901298	0,909677	0,907844	0,916129	0.,909889	0.916129
XGBoost Classifier	0,895627	0,906452	0,898648	0,906452	0,914910	0,922581	0,915398	0.922581
Extra Tree Classifier	0,909677	0,909677	0,901298	0,909677	0,887821	0,900000	0.,891658	0.900000

Dari Tabel 2, diperoleh model terbaik adalah Hist Gradient Boosting dengan metode oversampling ADASYN yang memiliki nilai *recall* dan *f1 score* tertinggi. Dengan *recall* sebesar 0,929, ini berarti bahwa model berhasil menangkap 92,9% dari semua kasus positif yang sebenarnya. *Recall* yang tinggi menunjukkan bahwa model sangat efektif dalam menemukan kasus positif yang ada. Selanjutnya, dari model terbaik tersebut, akan dilakukan eksplanasi model untuk menentukan pengaruh fitur terhadap klasifikasi gagal bayar.

B. Evaluasi

Pada bagian evaluasi, digunakan **feature importance** dan **SHAP values** untuk memahami pengaruh setiap fitur terhadap model. Feature importance membantu mengidentifikasi fitur mana yang memiliki dampak terbesar pada prediksi model, sementara SHAP values memberikan wawasan yang lebih mendalam mengenai kontribusi masing-masing fitur terhadap keputusan model. Melalui evaluasi ini, kita dapat menilai kinerja model secara keseluruhan dan memahami faktor-faktor yang mempengaruhi hasil prediksi, sehingga dapat melakukan perbaikan dan optimasi model yang lebih efektif.

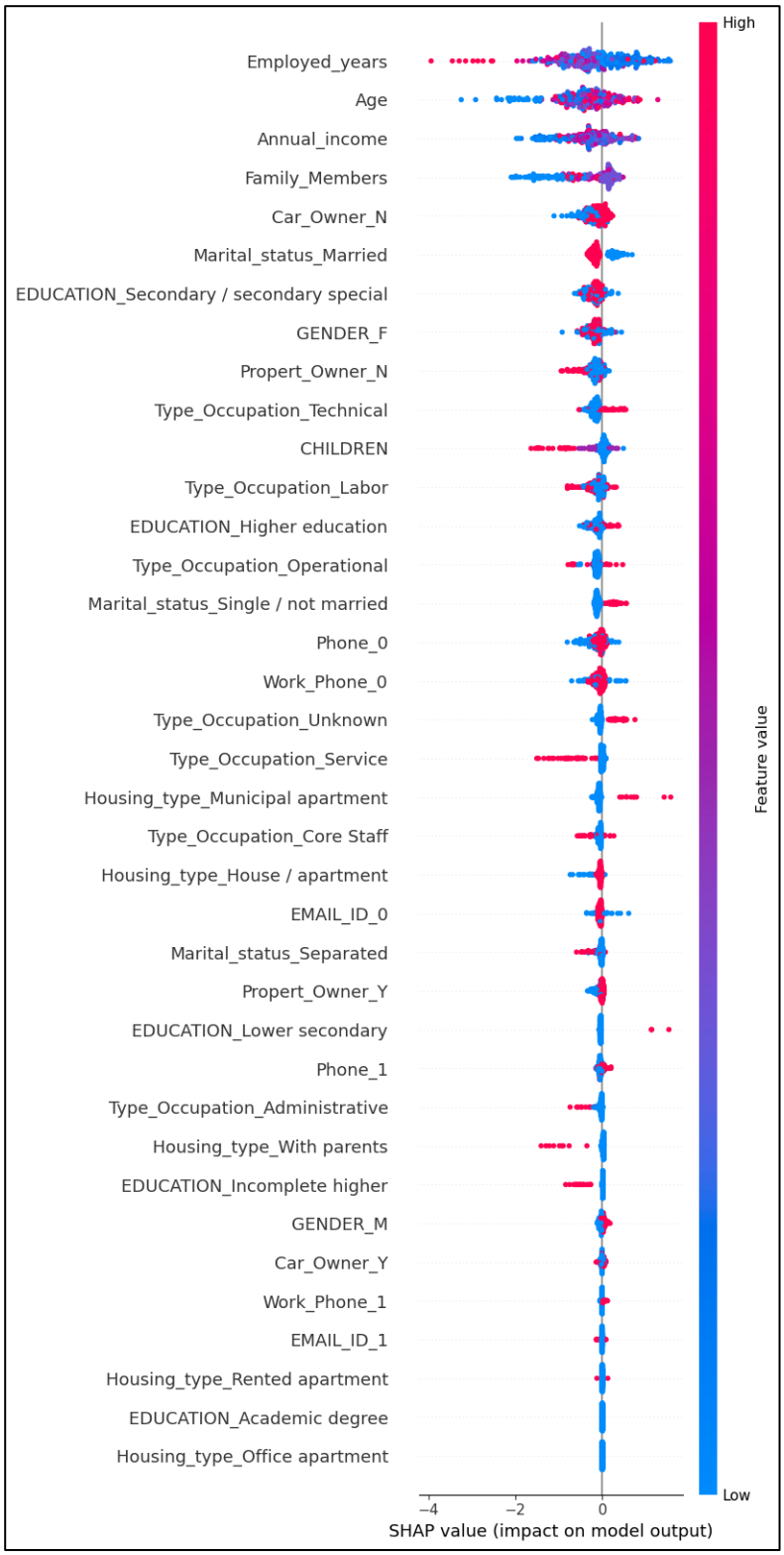
i. Feature Importance



Gambar 7. Feature Importance

Gambar 7 menunjukkan *feature imortance* berupa *permutation importance* dari model terbaik yaitu dengan algoritma Hist Gradient Boosting. Berdasarkan hasil *feature importance*, fitur yang memiliki pengaruh paling tinggi terhadap prediksi model adalah *annual income*, *family members*, *age*, dan *employed years*. Hal ini menunjukkan bahwa keempat fitur tersebut memberikan kontribusi terbesar dalam menentukan probabilitas gagal bayar pinjaman pribadi.

ii. SHAP Value Model

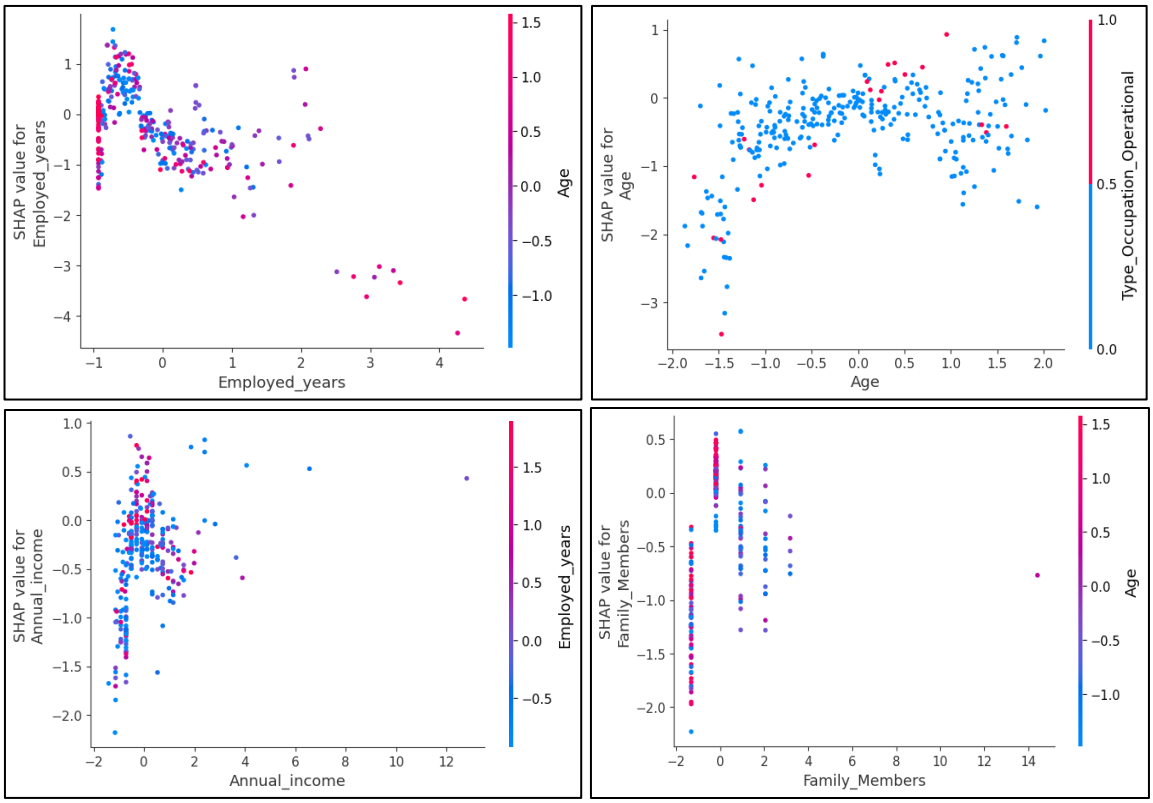


Gambar 8. Shap Values

Grafik pada Gambar 8 menunjukkan pengaruh dari masing-masing faktor terhadap probabilitas gagal bayar debitur, yang dapat dilihat di sepanjang sumbu horizontal. Selain itu, warna merah menunjukkan nilai yang tinggi pada suatu fitur. Berdasarkan grafik SHAP value yang ditunjukkan, berikut adalah interpretasi dari pengaruh fitur-fitur terhadap probabilitas gagal bayar pinjaman pribadi:

1. Employed_years:
   Titik-titik berwarna merah cenderung berada di sebelah kiri sumbu vertikal, yang berasosiasi dengan label 0 (tidak gagal bayar). Hal ini menunjukkan bahwa semakin lama seseorang telah bekerja (employed years yang lebih tinggi), semakin kecil kemungkinan mereka untuk gagal membayar pinjaman. Dengan kata lain, durasi pekerjaan yang lebih lama dikaitkan dengan ketepatan dalam membayar pinjaman.
2. Age:
   Titik-titik berwarna merah (nilai tinggi untuk usia) cenderung berada di sebelah kanan sumbu vertikal, yang berasosiasi dengan label 1 (gagal bayar). Ini menunjukkan bahwa semakin tua usia seseorang, semakin besar kemungkinan mereka untuk gagal membayar pinjaman pribadi. Faktor usia tua ini mungkin berhubungan dengan peningkatan risiko kesehatan atau pengurangan pendapatan.
3. Annual_income:
   Meskipun ada beberapa titik merah (nilai tinggi untuk pendapatan tahunan) yang berada di sebelah kanan, sebagian besar titik merah cenderung berada di sebelah kiri sumbu vertikal. Ini mengindikasikan bahwa pendapatan tahunan yang lebih tinggi (annual income yang tinggi) cenderung mengurangi risiko gagal bayar. Artinya, orang dengan pendapatan tahunan yang lebih besar lebih mungkin mampu membayar pinjaman mereka.
4. Family_Members:
   Titik-titik ungu (yang mewakili nilai menengah untuk jumlah anggota keluarga) cenderung berada di sebelah kanan (gagal bayar). Ini menunjukkan bahwa memiliki jumlah anggota keluarga yang moderat mungkin berkaitan dengan risiko gagal bayar yang lebih tinggi. Ini bisa jadi karena tanggungan finansial yang tidak terlalu besar maupun terlalu kecil mungkin memiliki risiko keuangan tertentu yang berkontribusi terhadap kemungkinan gagal bayar.

Lebih detail mengenai pengaruh dari tiap fitur dapat dilihat pada Gambar 9 di bawah yang memberikan visualisasi mendalam tentang bagaimana setiap fitur memengaruhi prediksi model secara individual.



Gambar 9. Shap Dependence Plot

iii. Optimalisasi Prediksi Label Resiko Gagal Bayar *(Threshold Moving Method)*

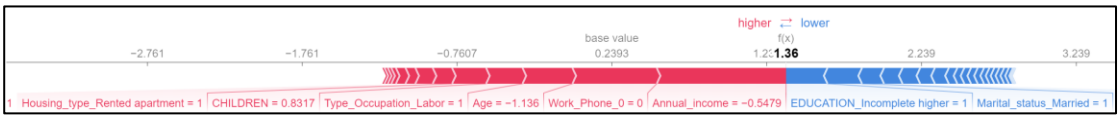
Dalam mengelola risiko gagal bayar pinjaman pribadi, sangat penting untuk memprediksi secara akurat siapa saja yang berisiko gagal bayar agar dapat memberikan rekomendasi dan tindakan preventif yang tepat. Pada model klasifikasi biner untuk memprediksi risiko gagal bayar, penting untuk dicatat bahwa data memiliki distribusi yang tidak seimbang, dengan label gagal bayar hanya mewakili sekitar 12,7% dari keseluruhan data. Oleh karena itu, diperlukan pendekatan khusus untuk meningkatkan akurasi prediksi, seperti penggunaan metode *Threshold Moving* berdasarkan probabilitas prediksi.

Metode *Threshold Moving* ini membantu menentukan ambang batas probabilitas tertentu di mana seseorang dianggap berisiko gagal bayar. Tujuan utamanya adalah untuk mengidentifikasi sebanyak mungkin individu yang benar-benar berisiko gagal bayar (diukur dengan metrik recall), sambil tetap menjaga keakuratan prediksi model (diukur dengan metrik precision). Setelah melakukan optimasi, ambang batas optimal yang ditemukan adalah 0,41. Artinya, setiap individu dengan probabilitas lebih dari 0,41 akan diklasifikasikan sebagai berisiko gagal bayar.

Dengan threshold ini, model berhasil mencapai nilai recall sebesar 92,9%, yang berarti mampu mengidentifikasi 92,9% dari semua individu yang berisiko gagal bayar. Selain itu, nilai precision sebesar 92,3% menunjukkan bahwa 92,3% dari individu yang

diprediksi berisiko gagal bayar oleh model memang benar-benar berisiko. Gabungan antara nilai recall dan precision yang tinggi ini menandakan bahwa model ini tidak hanya mampu mengidentifikasi banyak individu yang berisiko, tetapi juga akurat dalam prediksi tersebut, yang sangat penting untuk pengambilan keputusan dalam mengelola risiko pinjaman.

Untuk menentukan rekomendasi dan tindakan yang tepat bagi individu yang berisiko gagal bayar, nilai SHAP individu akan digunakan. Metode ini membantu menilai pengaruh setiap fitur terhadap prediksi model untuk observasi tertentu, memberikan dasar yang kuat untuk keputusan yang lebih terinformasi. Sebagai contoh, analisis dilakukan pada satu sampel, seperti yang diilustrasikan pada Gambar 10 di bawah ini. Individu tersebut diperkirakan memiliki nilai log-odds sebesar 1,36 dari model Gradient Boosting, yang dikonversi menjadi probabilitas sebesar 0,79—lebih tinggi dari ambang batas 0,4, sehingga diprediksi akan gagal bayar.



Gambar 10. Sample SHAP Values Individual

Faktor-faktor yang ditampilkan adalah nilai yang belum distandarisasi. Setelah dikembalikan ke skala asli dengan dikali standar deviasi dan ditambahkan mean, faktor utama dalam prediksi ini mencakup jenis tempat tinggal, yaitu "rented apartment," dan jumlah anak, yang adalah 1. Berdasarkan nilai SHAP, individu dengan lebih banyak anak cenderung memiliki peluang lebih baik untuk membayar, mungkin karena perencanaan keuangan yang lebih baik dan dukungan dari pendapatan tahunan yang lebih tinggi.

Faktor tambahan seperti pekerjaan sebagai "labor" (pekerja kasar), usia sekitar 30 tahun, tidak memiliki telepon kerja, dan pendapatan tahunan sebesar 129.392 juga berkontribusi pada peningkatan risiko gagal bayar. Status pendidikan yang "incomplete higher" dan status perkawinan "married" (menikah) menunjukkan penurunan risiko gagal bayar, kemungkinan karena stabilitas finansial yang lebih baik dalam rumah tangga yang sudah menikah.

Oleh karena itu, untuk nasabah ini, penting bagi bank untuk memberikan bantuan berupa saran keuangan dan opsi restrukturisasi utang. Selain itu, bank perlu menyediakan dukungan tambahan bagi nasabah dengan tanggungan besar atau pendapatan yang kurang stabil untuk mengurangi risiko gagal bayar.

IV. KESIMPULAN DAN SARAN

A. Kesimpulan

Penelitian ini bertujuan untuk mengklasifikasikan peminjam yang berpotensi gagal bayar dan tidak gagal bayar dengan menggunakan berbagai algoritma klasifikasi. Dari berbagai algoritma yang diuji, menggunakan metrik evaluasi *recall* dan *FI-score*, Hist Gradient Boosting dengan metode oversampling ADASYN terbukti memiliki peforma terbaik dengan nilai *recall* sebesar 0,929032 dan f1 score sebesar 0,921042.

Hasil menunjukkan bahwa Hist Gradient Boosting memiliki kemampuan yang sangat baik dalam mendeteksi peminjam yang berisiko gagal bayar, yang ditunjukkan oleh nilai recall yang tinggi. Hal ini penting karena dalam konteks manajemen risiko kredit, mendeteksi semua potensi gagal bayar merupakan aspek yang sangat penting. Selain itu, nilai *F1-score* yang tinggi mengindikasikan keseimbangan yang baik antara presisi dan recall, menjadikan Hist Gradient Boosting sebagai algoritma yang tidak hanya efektif dalam mengidentifikasi gagal bayar tetapi juga akurat dalam klasifikasi secara keseluruhan.

Berdasarkan hasil *feature importance*, fitur yang memiliki pengaruh paling tinggi terhadap prediksi model adalah *annual income*, *family members*, *age*, dan *employed years*. Hal ini menunjukkan bahwa keempat fitur tersebut memberikan kontribusi terbesar dalam menentukan probabilitas gagal bayar pinjaman pribadi. Selanjutnya berdasarkan hasil *shap values* mengungkapkan bahwa fitur *employed years, annual income, age, dan family members* memberikan kontribusi terbesar terhadap prediksi model. *Employed years* dan *annual income* yang lebih tinggi cenderung menurunkan risiko gagal bayar, sedangkan usia yang lebih tua dan jumlah anggota keluarga yang moderat meningkatkan risiko tersebut. Hasil dari penelitian ini menekankan pentingnya fitur-fitur tersebut dalam memprediksi gagal bayar pinjaman pribadi.

Selanjutnya, dengan menggunakan metode threshold moving, optimalisasi pada model Hist Gradient Boosting menghasilkan ambang batas terbaik sebesar 0,41. Hal ini menunjukkan bahwa nasabah dengan probabilitas di atas 0,41 akan diklasifikasikan sebagai gagal bayar. Berdasarkan ambang batas ini, model mencapai recall sebesar 92,9% dan precision sebesar 92,3%, yang berarti model mampu mengidentifikasi 92,9% dari individu yang berisiko gagal bayar. Dengan hasil ini, nilai SHAP individu dapat dimanfaatkan untuk memberikan rekomendasi dan tindakan yang tepat bagi setiap nasabah yang terancam gagal bayar.

B. Saran

Untuk menerapkan model ini dalam operasi sehari-hari di perbankan, ada beberapa langkah penting yang perlu diperhatikan. Pertama, kualitas data menjadi faktor utama

yang menentukan keberhasilan model, sehingga bank harus memastikan data yang digunakan selalu berkualitas tinggi, lengkap, dan diperbarui secara berkala. Proses integrasi model dengan sistem perbankan yang sudah ada juga menjadi tantangan teknis yang harus diatasi. Selain itu, pelatihan bagi staf yang terkait sangat diperlukan agar mereka dapat menggunakan dan menginterpretasikan hasil model dengan tepat, sehingga implementasi model dapat berjalan dengan lancar.

Selain itu, untuk mendukung penerapan model ini secara optimal, diperlukan investasi dalam infrastruktur teknologi yang memadai, seperti perangkat keras, perangkat lunak, dan sistem keamanan yang mampu menangani volume data yang besar dan kecepatan pemrosesan yang dibutuhkan. Kolaborasi yang erat antara tim teknis dan operasional juga sangat penting untuk memastikan bahwa setiap aspek implementasi berjalan dengan baik. Dengan perhatian terhadap aspek-aspek ini, model klasifikasi dapat diimplementasikan secara efektif dalam operasional perbankan, memberikan nilai tambah yang signifikan, dan mendukung pencapaian tujuan bisnis jangka panjang.