Pendekatan Cost–Sensitive Classification dalam Penentuan Income

untuk Mendukung Risk-Based Lending

Presented By :
Aulia Aorama

Latar Belakang & Permasalahan Utama

Latar Belakang

- Bank A mengalami peningkatan rasio NPL dalam dua tahun terakhir.
- Strategi saat ini lebih konservatif dan fokus pada penguatan manajemen risiko.
- Tidak semua calon nasabah memiliki data pendapatan formal (self-employed, UMKM, informal).
- Dibutuhkan pendekatan berbasis data untuk mengestimasi kemampuan finansial sebagai bagian dari proses underwriting.

Permasalahan Utama

Model digunakan untuk mengklasifikasikan income >50K sebagai proxy kemampuan bayar.
Risiko kesalahan klasifikasi:

- False Positive (FP)
- Nasabah diprediksi berincome tinggi padahal tidak → risiko pemberian kredit ke segmen berisiko.
- False Negative (FN)
- Nasabah diprediksi berincome rendah padahal sebenarnya tinggi → kehilangan potensi bisnis.

Dalam kondisi pengetatan risiko, pengendalian False Positive menjadi prioritas utama.

Goals & Analytic Approach

Tujuan Analisis

- Mengidentifikasi faktor demografis dan sosial-ekonomi yang berkorelasi dengan income >50K.
- Mengembangkan model klasifikasi sebagai proxy kemampuan bayar.
- Mendukung proses underwriting berbasis risiko (risk-based lending).

Pendekatan Analitik

- Data Understanding Exploratory Data Analysis (EDA) untuk memahami pola income.
- Preprocessing & feature engineering untuk meningkatkan kualitas model.
- Pengembangan model klasifikasi dengan cross validation.
- Evaluasi menggunakan F0.5-score untuk memprioritaskan precision (mengurangi False Positive).
- Optimasi threshold probabilitas untuk menghasilkan keputusan yang lebih konservatif dan selaras dengan strategi manajemen risiko Bank A.

Data Understanding

Nama Variabel	Deskripsi
Age	Usia individu dalam tahun.
Workclass	Jenis atau sektor tempat individu bekerja (swasta, pemerintah, wiraswasta, dll.).
fnlwgt (Final Weight)	Bobot sampel dari sensus yang merepresentasikan jumlah populasi yang diwakili oleh individu tersebut.
Education	Tingkat pendidikan terakhir yang ditempuh individu.
EducationNum	Representasi numerik dari tingkat pendidikan (versi ordinal dari Education).
Marital Status	Status pernikahan individu.
Occupation	Jenis atau bidang pekerjaan individu.
Relationship	Peran individu dalam rumah tangga (misalnya suami, istri, anak, dll.).
Race	Ras atau latar belakang etnis individu.
Gender	Jenis kelamin individu.
Capital Gain	Keuntungan yang diperoleh dari investasi atau penjualan aset.
Capital Loss	Kerugian yang dialami dari investasi atau penjualan aset.
Hours per Week	Jumlah jam kerja individu dalam satu minggu.
Native Country	Negara asal individu.
Income	Kategori pendapatan tahunan individu (≤50K atau >50K).

Jumlah kolom: 15 sumber data: Kaggle
Jumlah baris: 48841
Data Duplikat : 29 data (di drop)
Data Nan: 0 data
Categori: 9
Numerik: 6

Problem dan strategi Data

Tabel Strategi Data Cleaning

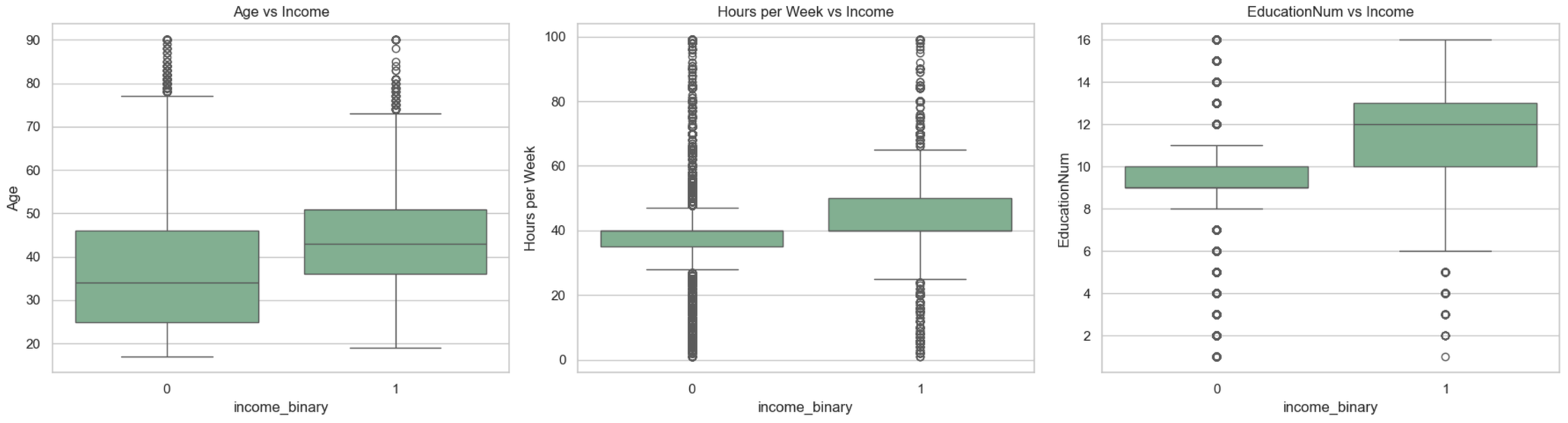
Permasalahan	Perbaikan	Alasan
Nilai NaN pada `Occupation` ketika `Workclass = Never-worked`	Mengganti dengan kategori **"No Occupation"**	Secara logis individu yang tidak pernah bekerja memang tidak memiliki pekerjaan. Ini bukan missing murni, melainkan kondisi struktural.
Nilai NaN simultan pada `Workclass` dan `Occupation`	Mengganti dengan kategori **"Unknown"**	Tidak tersedia informasi untuk inferensi. Menggunakan kategori khusus lebih aman dibanding imputasi spekulatif.
Distribusi `Native Country` sangat tidak seimbang (±91% United-States)	Recategorisasi menjadi **United-States** dan **Non-United-States**	Mengurangi sparsity, menurunkan dimensionalitas encoding, dan mencegah overfitting akibat kategori minoritas sangat kecil.
Inkonsistensi antara `Relationship` dan `Gender` (misal Wife–Male, Husband–Female)	Menghapus baris inkonsisten	Tidak ada ground truth untuk menentukan kolom mana yang salah, sehingga penghapusan lebih metodologis dibanding koreksi asumtif.
Distribusi `Capital Gain` dan `Capital Loss` sangat skewed dan didominasi nilai 0	Melakukan **binning** (0 vs >0)	Mengurangi efek outlier ekstrem, menstabilkan model, serta meningkatkan interpretabilitas fitur finansial.
Kategori `Education` terlalu banyak (16 kategori)	Recategorisasi menjadi tingkat pendidikan utama (Preschool/SD/SMP/SMA/Bachelor/Master/Doktor)	Mengurangi sparse category, meningkatkan interpretabilitas bisnis, dan menyederhanakan model.
`EducationNum` redundan dengan `Education`	Menghapus kolom `EducationNum`	Tidak menambah informasi baru dan berpotensi menimbulkan multicollinearity jika digunakan bersamaan.
Kolom `Final Weight` untuk estimasi populasi jadi tidak relevan dengan klasifikasi income	menghapus kolom `Final Weight`	Merupakan sampling weight sensus, bukan karakteristik individu, sehingga tidak relevan untuk modeling income

Data Cleaning

	Column Name	Data Type	Null Values	Number of Unique	Unique Sample
0	Age	int64	0	74	[39, 50, 38, 53, 28, 37, 49, 52, 31, 42, 30, 23, 32, 40, 34, 25, 43, 54, 35, 59, 56, 19, 20, 45, 22, 48, 21, 24, 57, 44, 41, 29, 18, 47, 46, 36, 79, 27, 67, 33, 76, 17, 55, 61, 70, 64, 71, 68, 66, 51, 58, 26, 60, 90, 75, 65, 77, 62, 63, 80, 72, 74, 69, 73, 81, 78, 88, 82, 83, 84, 85, 86, 87, 89]
1	Workclass	object	0	9	[State-gov, Self-emp-not-inc, Private, Federal-gov, Local-gov, Unknown, Self-emp-inc, Without-pay, Never-worked]
2	Education	object	0	8	[Bachelor, SMA, Master, SMP, Diploma, Doctor, SD, Preschool]
3	Marital Status	object	0	7	[Never-married, Married-civ-spouse, Divorced, Married-spouse-absent, Separated, Married-AF-spouse, Widowed]
4	Occupation	object	0	16	[Adm-clerical, Exec-managerial, Handlers-cleaners, Prof-specialty, Other-service, Sales, Craft-repair, Transport-moving, Farming-fishing, Machine-op-inspct, Tech-support, Unknown, Protective-serv, Armed-Forces, Priv-house-serv, No Occupation]
5	Relationship	object	0	6	[Not-in-family, Husband, Wife, Own-child, Unmarried, Other-relative]
6	Race	object	0	5	[White, Black, Asian-Pac-Islander, Amer-Indian-Eskimo, Other]
7	Gender	object	0	2	[Male, Female]
8	Hours per Week	int64	0	96	[40, 13, 16, 45, 50, 80, 30, 35, 60, 20, 52, 44, 15, 25, 38, 43, 55, 48, 58, 32, 70, 2, 22, 56, 41, 28, 36, 24, 46, 42, 12, 65, 1, 10, 34, 75, 98, 33, 54, 8, 6, 64, 19, 18, 72, 5, 9, 47, 37, 21, 26, 14, 4, 59, 7, 99, 53, 39, 62, 57, 78, 90, 66, 11, 49, 84, 3, 17, 68, 27, 85, 31, 51, 77, 63, 23, 87, 88, 73, 89, 97, 94, 29, 96, 67, 82, 86, 91, 81, 76, 92, 61, 74, 95, 79, 69]
9	Native Country	object	0	2	[United-States, Non-United-States]
10	income_binary	int64	0	2	[0, 1]
11	capital_gain_bin	object	0	2	[Has Gain, No Gain]
12	capital_loss_bin	object	0	2	[No Loss, Has Loss]

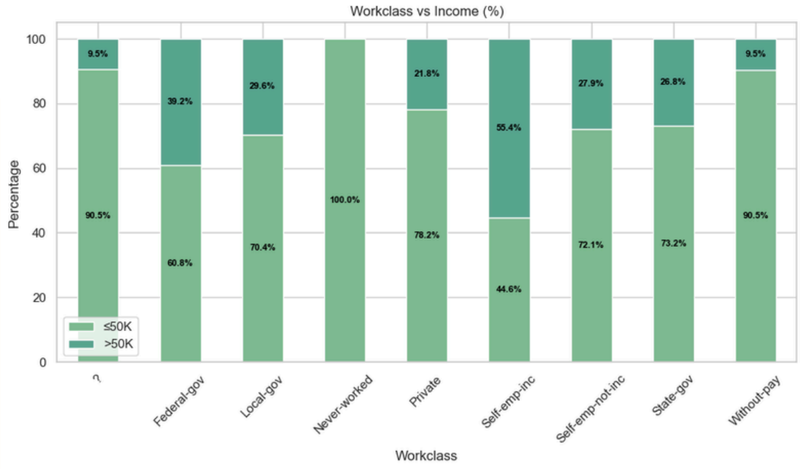
Setelah di cleaning
Jumlah kolom: 13
Jumlah baris: 48808
Data Duplikat : 0 data
Data Nan: 0 data
Categori: 10
Numerik: 3

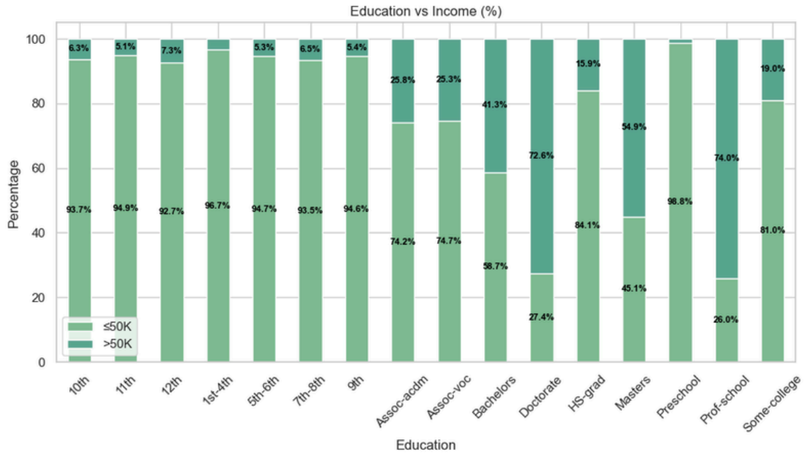
EDA

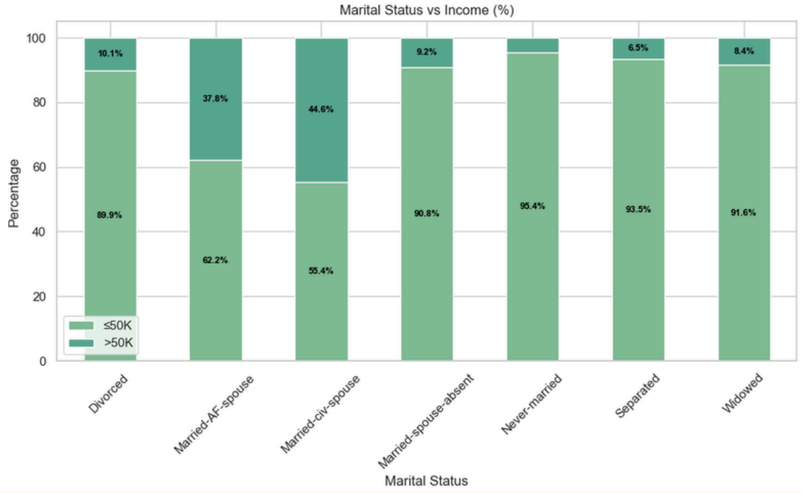


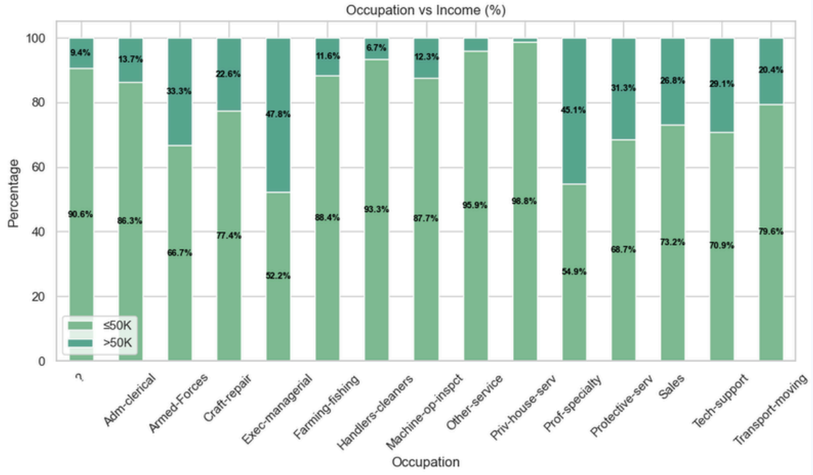
Individu dengan pendapatan lebih dari 50K cenderung berusia lebih tua, bekerja lebih lama, dan memiliki tingkat pendidikan yang lebih tinggi. Dari ketiga variabel yang dianalisis, **tingkat pendidikan merupakan faktor yang paling dominan dalam membedakan kelompok pendapatan**

EDA









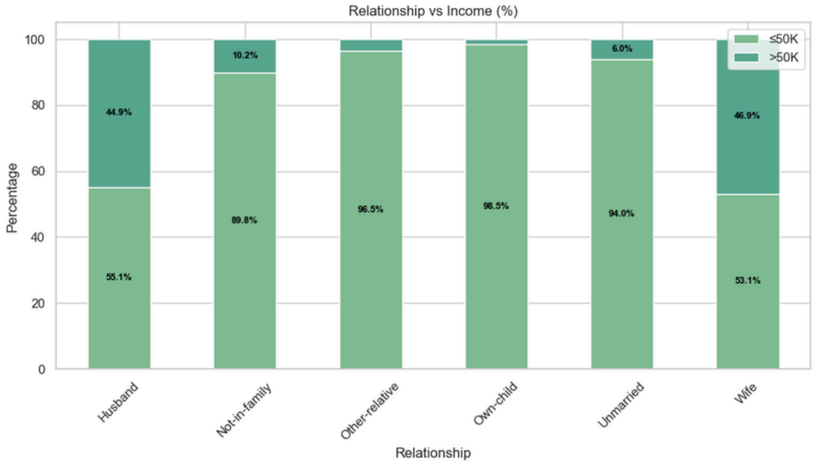
Workclass
Self-emp-inc dan sektor pemerintah → proporsi income >50K lebih tinggi.
Private, Never-worked, Without-pay → didominasi income ≤50K.
→ Workclass cukup informatif untuk indikasi awal kemampuan bayar.

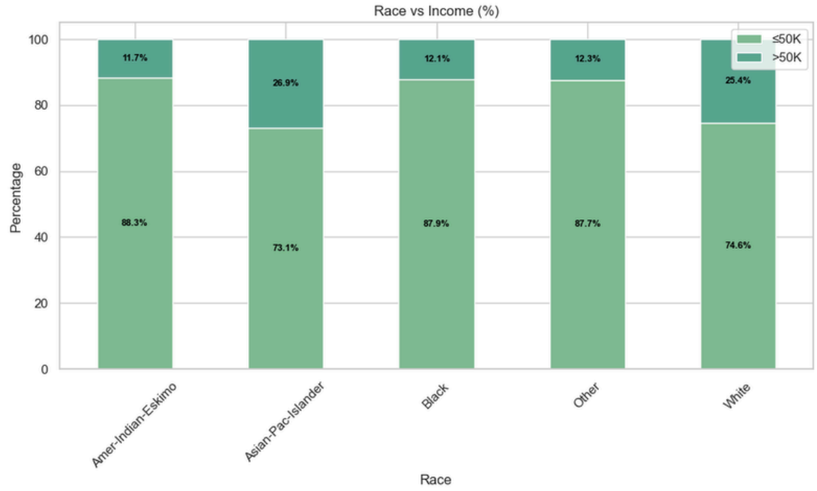
Education
Semakin tinggi pendidikan → semakin tinggi proporsi income >50K.
HS-grad ke bawah → mayoritas ≤50K.
→ Pendidikan merupakan salah satu prediktor terkuat income.

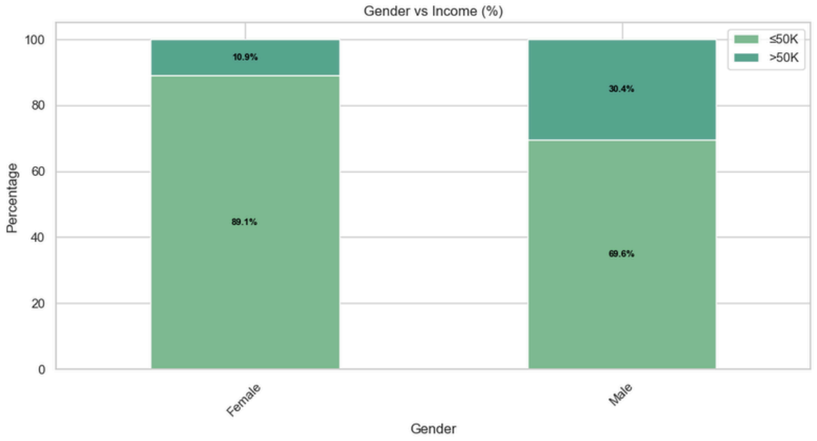
Marital Status
Married-civ-spouse → proporsi >50K tertinggi.
Never-married, Divorced, Widowed → didominasi ≤50K.
→ Status pernikahan berkorelasi dengan stabilitas ekonomi.

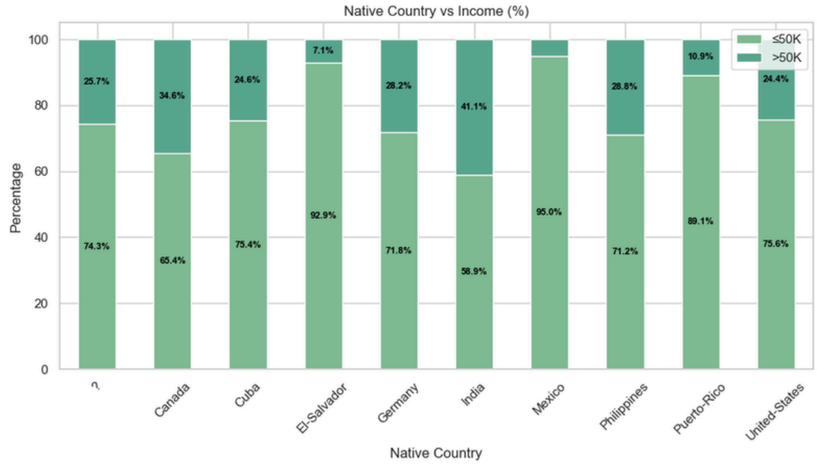
Occupation
Exec-managerial & Prof-specialty → income >50K tertinggi.
Pekerjaan manual/service → didominasi ≤50K.
→ Jenis pekerjaan sangat membedakan level pendapatan.

EDA









Relationship
Husband & Wife → proporsi income >50K tertinggi.
Own-child & Other-relative → mayoritas ≤50K.
→ Menggambarkan posisi ekonomi dalam rumah tangga (berpotensi overlap dengan Marital Status).

Race
Terdapat variasi proporsi income >50K antar ras.
Asian-Pac-Islander relatif lebih tinggi.
White dominan secara jumlah (dipengaruhi ukuran sampel).
→ Perlu hati-hati karena distribusi tidak seimbang.

Gender
Terdapat perbedaan signifikan income >50K antara Male dan Female.
→ Variabel informatif namun termasuk fitur sensitif (perlu pertimbangan etis/regulasi).

Native Country
Beberapa negara (India, Germany) menunjukkan proporsi >50K lebih tinggi.
Distribusi sangat tidak merata (didominasi United States).
→ Kategori granular dan sparse berpotensi menimbulkan instabilitas model.

Data Preprocessing

Tabel Strategi Data Preprocessing

Jenis Variabel	Kolom	Metode Transformasi	Detail Implementasi	Tujuan / Alasan
Numerik	Age, Hours per Week	RobustScaler	Scaling berbasis median dan IQR	Mengurangi pengaruh outlier dan menjaga stabilitas distribusi
Ordinal	Education	OrdinalEncoder	Urutan kategori: Preschool → SD → SMP → SMA → Diploma → Bachelor → Master → Doctor	Mempertahankan struktur hierarkis tingkat pendidikan
Kategorikal (High Cardinality)	Workclass, Marital Status, Occupation, Relationship	BinaryEncoder	Encoding berbasis representasi biner	Mengurangi dimensionalitas dan sparsity dibanding One-Hot
Kategorikal (Low Cardinality)	Gender, Race, Native Country, capital_gain_bin, capital_loss_bin	OneHotEncoder	drop='first', handle_unknown='ignore'	Menghindari dummy variable trap dan menjaga stabilitas pipeline
Integrasi Transformasi	Seluruh kolom	ColumnTransformer	Transformasi terpisah per tipe variabel	Menghindari data leakage dan menjaga konsistensi pipeline
Fitur Lainnya	Kolom di luar daftar	Passthrough	remainder='passthrough'	Mempertahankan fitur yang tidak perlu ditransformasi

Define X dan Y

Tabel Rancangan Define X dan Y

Tahap	Deskripsi	Implementasi	Tujuan
Define Target (Y)	Menentukan variabel dependen	y = df['income_binary']	Menjadikan income_binary sebagai target klasifikasi (0 = ≤50K, 1 = >50K)
Define Feature (X)	Menghapus kolom target dari dataset	X = df.drop(columns=['income_binary'])	Mencegah data leakage dan memastikan model hanya belajar dari fitur prediktor
Validasi Fitur	Mengecek kolom yang digunakan sebagai X	print(X.columns)	Memastikan seluruh fitur relevan masuk ke tahap modeling

Perbandingan Train test split : 80% train 20% test

Cross Validation

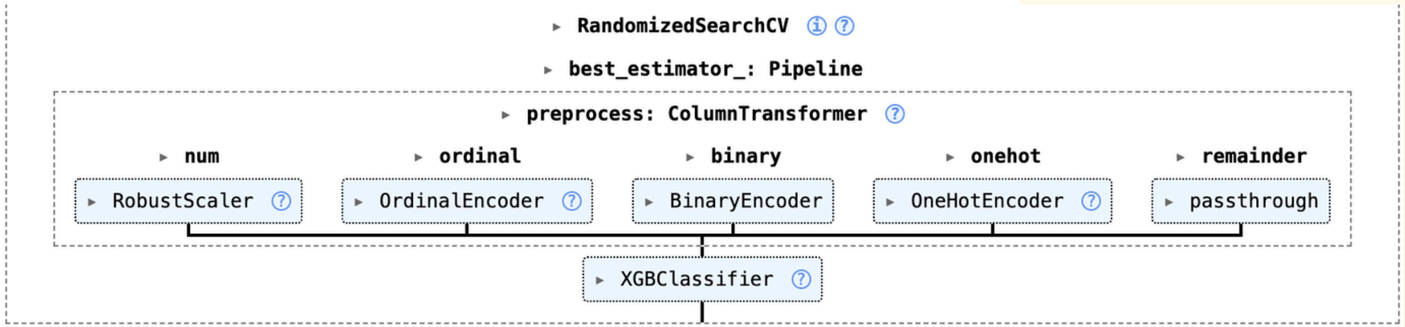
	Resampling	Model	Mean_F0.5	Std_F0.5
0	No_Resampling	XGBoost	0.692995	0.003808
7	SMOTE	StackingClassifier	0.636793	0.001854
10	RandomOverSampler	StackingClassifier	0.623809	0.004054
20	RandomUnderSampler	XGBoost	0.590956	0.005308
43	NearMiss	AdaBoost	0.543104	0.009997

Model Terbaik: XGBoost tanpa Resampling

- Mean F0.5: 0.693
- Std Dev: 0.0038 → performa stabil dan konsisten
- Resampling (SMOTE, UnderSampling, NearMiss) tidak meningkatkan performa
- UnderSampling justru menurunkan F0.5 secara signifikan

Kesimpulan:
XGBoost sudah mampu menangani ketidakseimbangan kelas secara intrinsik.
Resampling tidak diperlukan pada kasus ini.

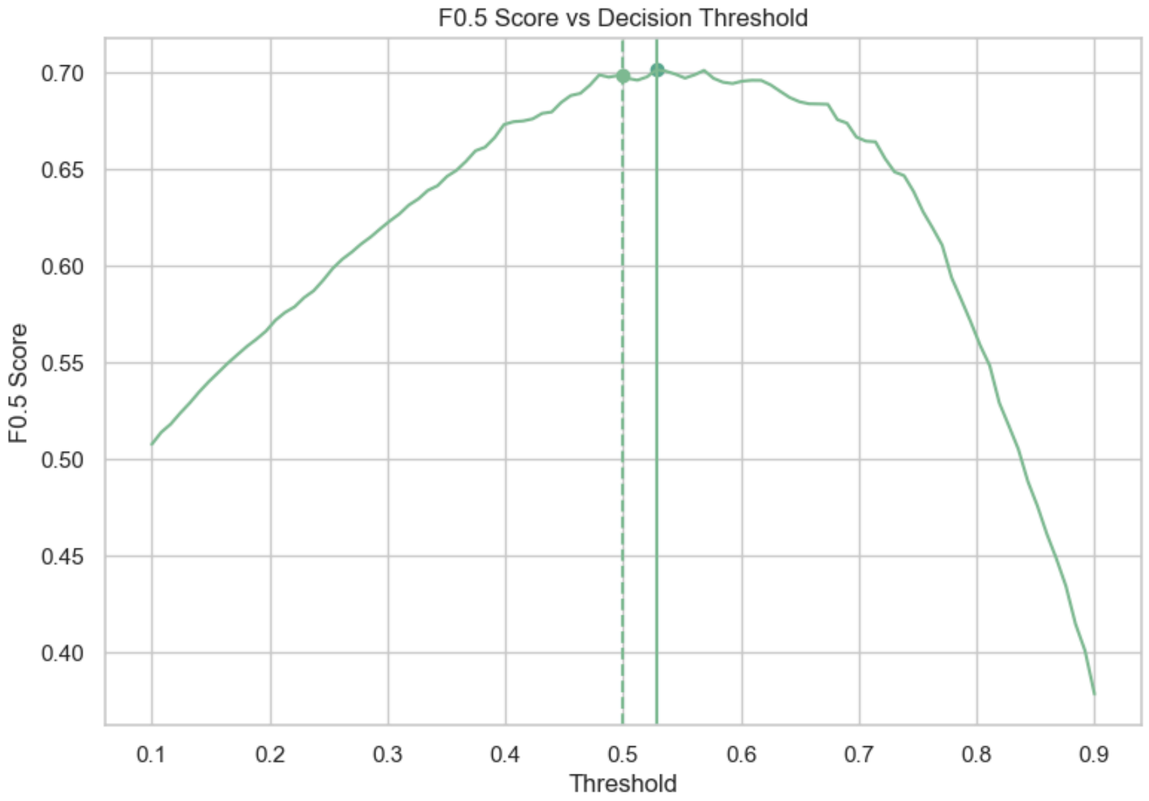
Hyperparameter Tunning



Best Model: XGboost
Best F0.5 CV Score: 0.696
Parameter Optimal:

- max_depth = 5
- min_child_weight = 5
- learning_rate = 0.0595
- n_estimators = 185
- gamma = 2.098
- subsample = 0.606
- colsample_bytree = 0.845

Evaluasi Model



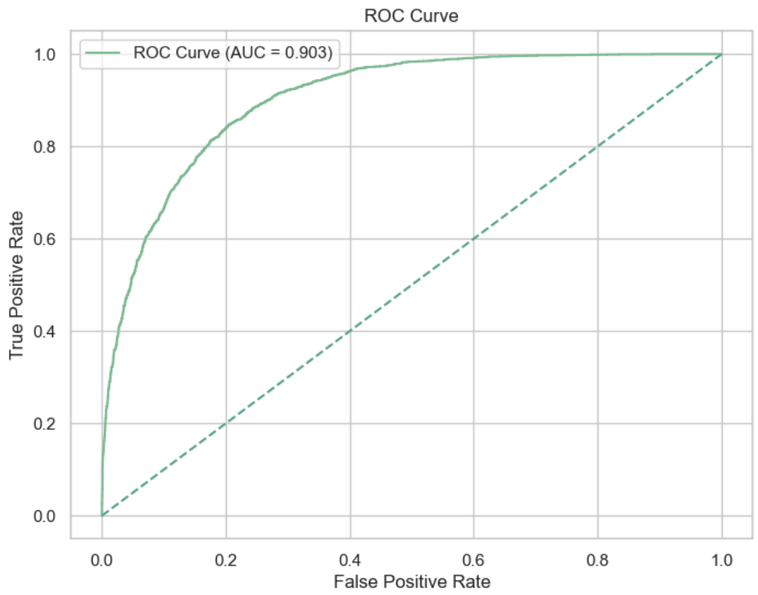
Default Threshold (0.5) F0.5: 0.6985
Optimal Threshold: 0.5283
Optimal F0.5: 0.7017

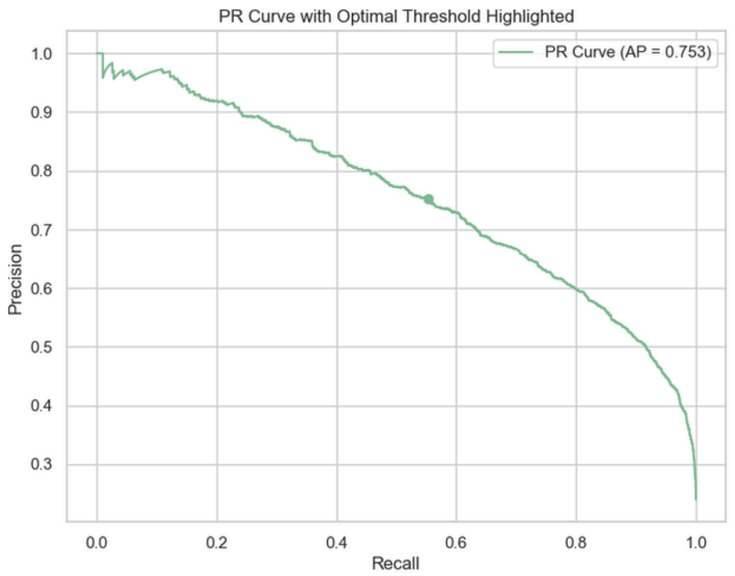
```
Confusion Matrix (Optimal Threshold)
[[6998  427]
 [1043 1294]]
```

```
F0.5 Test Score: 0.6984878098961013
[[6937  488]
 [ 979 1358]]
```

model menjadi lebih konservatif dalam mengklasifikasikan income >50K. Penurunan False Positive menunjukkan bahwa model lebih berhati-hati dalam menyatakan seseorang sebagai income tinggi, dengan konsekuensi meningkatnya False Negative.

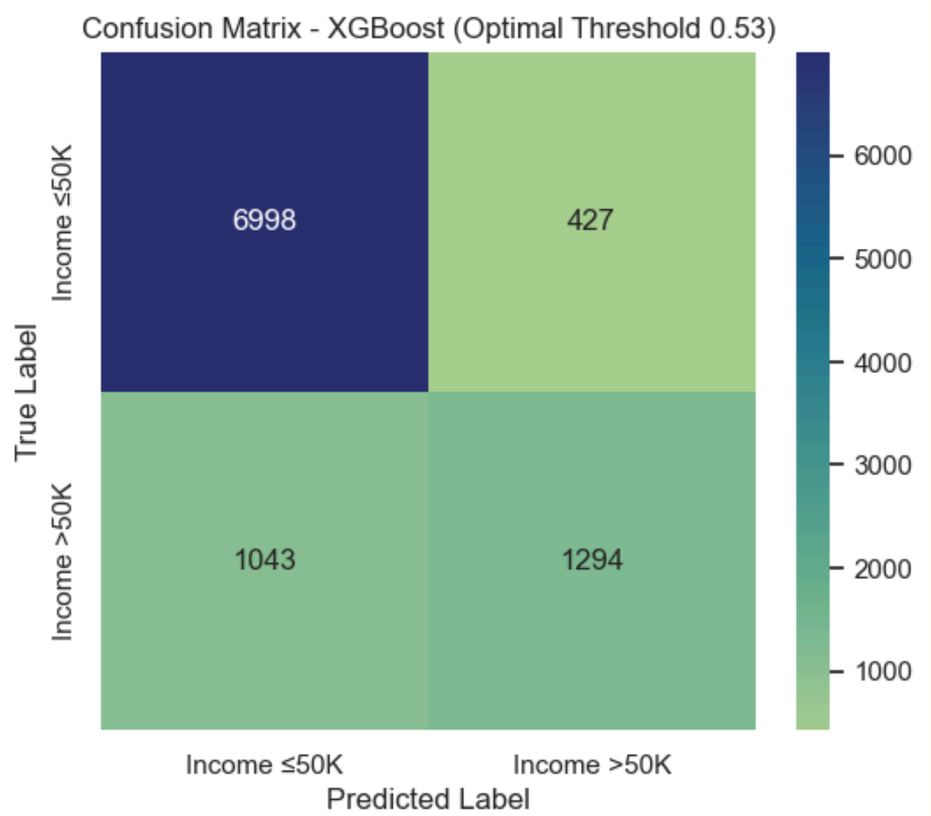
ROC dan PR Cruve





- ROC-AUC = 0.903 → Kemampuan diskriminasi sangat kuat dalam membedakan income >50K vs ≤50K.
- PR-AUC = 0.753 → Model efektif mendeteksi kelas >50K meskipun data imbalanced (~24% kelas positif).
- PR-AUC jauh di atas baseline (0.24), menunjukkan performa nyata, bukan bias kelas mayoritas.

Confusion Metrix



- False Positive (FP) = 427
- False Negative (FN) = 1.043
- Cost per FP = Rp36.000.000
- Cost per FN = Rp14.400.000

1. **Kerugian akibat False Positive**

427×36.000.000= Rp15.372.000.000 ≈ Rp15,37 miliar

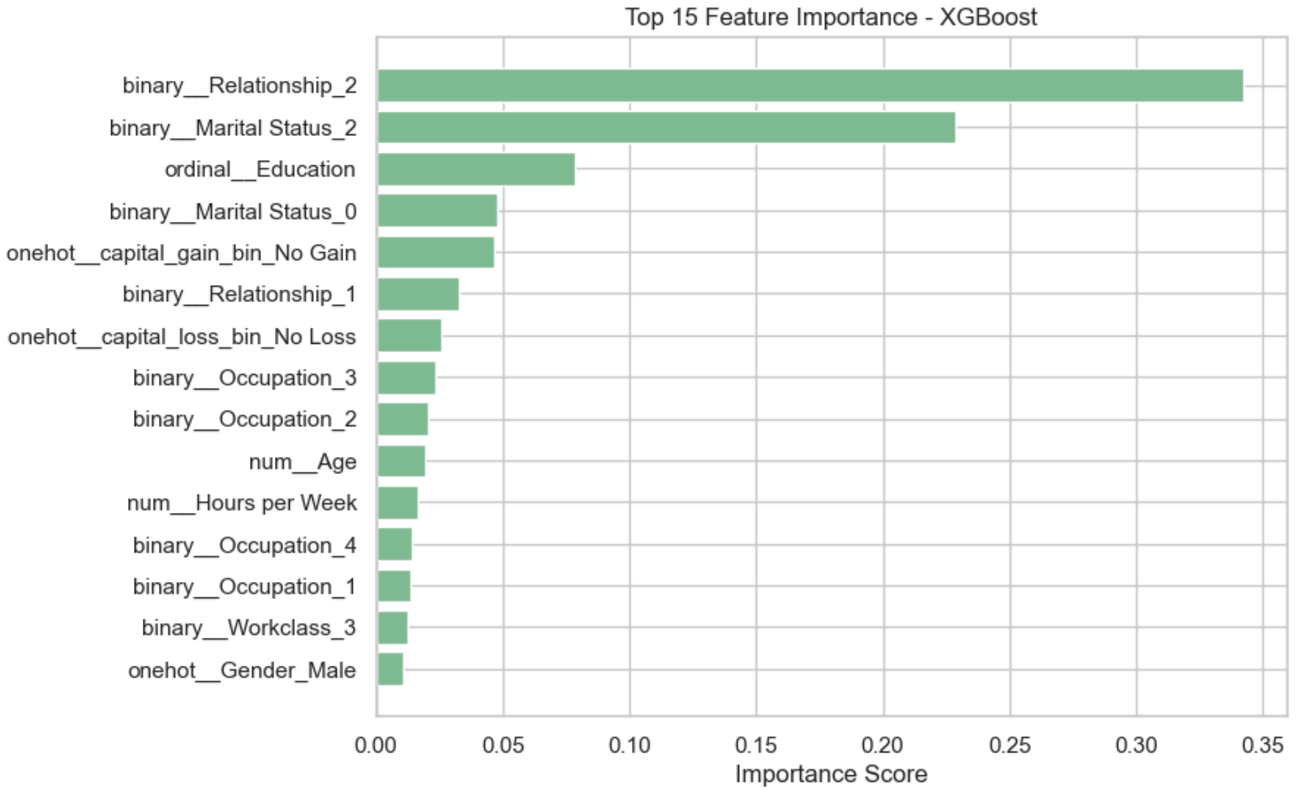
2. **Kerugian akibat False Negative**

1.043×14.400.000= Rp15.019.200.000 ≈ Rp15,02 miliar

3. **Total Expected Loss**

15.372.000.000+15.019.200.0001= Rp30.391.200.000≈ Rp30,39 miliar

feature importance



Berdasarkan hasil feature importance dari model XGBoost, terlihat bahwa variabel dengan kontribusi terbesar dalam memprediksi income >50K adalah fitur terkait Relationship dan Marital Status

Thank You

Aulia Aorama
Aoramaaulia@gmail.com