

FINAL PROJECT DIGITAL SKOLA DATA SCIENCE BATCH 54

DEMOGRAPHIC PATTERN ANALYSIS: ALGORITHMIC MODELING FOR CLASSIFYING HIGH-INCOME PROFILES



Shabiha Rahma Fauziah · Aulia Aorama
· Farras Zihan Harmany ·
Rifqi Permadi · Leonard Ari Raharja

OUTLINES

Chapter 1 - Data & Business Understandings

Chapter 2 - Preprocessing

Chapter 3 - Modelling & Interpretation

Chapter 4 - Conclusions & Reccomendations

CHAPTER 1



**DATA & BUSINESS
UNDERSTANDING**



BUSINESS UNDERSTANDING

Proyek ini berfokus pada prediksi tingkat pendapatan individu (di atas atau di bawah \$50K USD per tahun) berdasarkan data sensus. Memahami pola pendapatan sangat penting bagi berbagai pemangku kepentingan dalam membuat keputusan berbasis data di bidang keuangan, pemasaran, dan kebijakan.

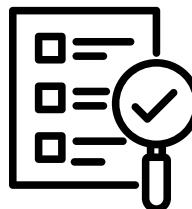
Tujuan utama proyek ini adalah:

- Mengidentifikasi faktor-faktor yang mempengaruhi tingkat pendapatan, termasuk karakteristik demografis, pendidikan, pekerjaan, dan finansial.
- Memberikan wawasan yang dapat ditindaklanjuti untuk pemangku kepentingan yang ditargetkan, meliputi:
 - Institusi Keuangan: Untuk penilaian risiko kredit dan segmentasi pelanggan
 - Tim Pemasaran: Untuk menargetkan kampanye produk premium
 - Pemerintah/Peneliti Sosial: Untuk analisis kesenjangan pendapatan dan perencanaan kebijakan
- Menyiapkan dataset yang bersih dan terstruktur dengan baik untuk tahap pemodelan prediktif guna memungkinkan klasifikasi kategori pendapatan yang akurat.

DATA UNDERSTANDING

Census Income Analysis and Modeling

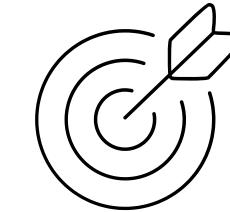
<https://www.kaggle.com/code/tawfikelmetwally/census-income-analysis-and-modeling/input>



48.841 entries data



6.464 missing values
dan 29 duplicates



Target : Kategori pendapatan
tahunan individu

6 Numerical Data

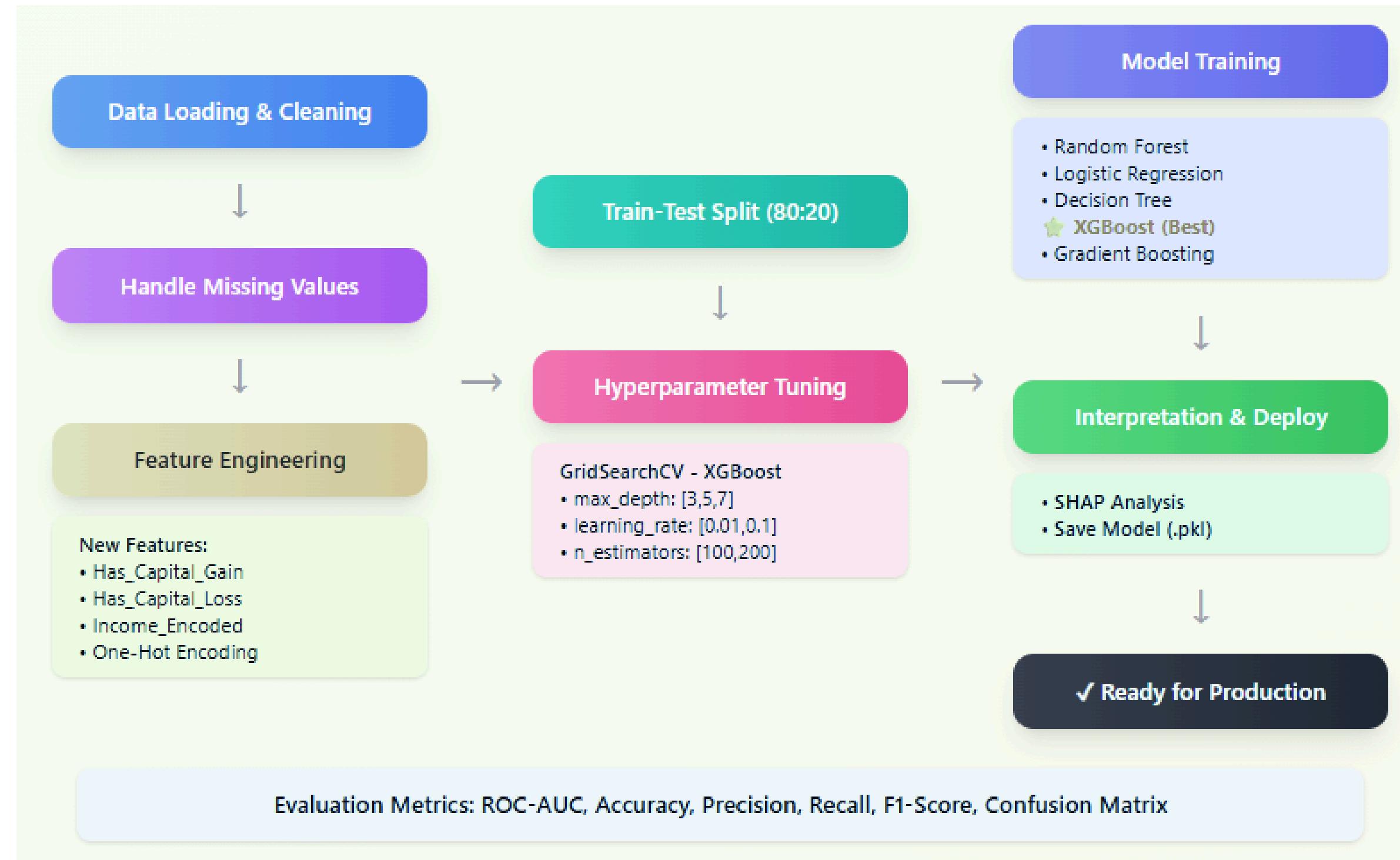
- **Age** - Usia
- **Final Weight** - Bobot sampel dari sensus
- **EducationNum** - Representasi numerik dari tingkat pendidikan
- **Capital Gain** - Pendapatan tambahan dari investasi atau aset
- **Capital Loss** - Kerugian finansial dari investasi atau aset
- **Hours per Week** - Rata - rata jam kerja dalam satu minggu

9 Categorical Data

- **Workclass** - Sektor pekerjaan
- **Education** - Pendidikan terakhir
- **Marital Status** - Status pernikahan
- **Occupation** - Jenis pekerjaan/jabatan
- **Relationship** - Hubungan dengan kepala rumah tangga
- **Race** - Ras
- **Gender** - Jenis kelamin
- **Native Country** - Negara asal
- **Income** - Kategori pendapatan tahunan

PIPELINE

PROJECT PIPELINE



CHAPTER 2

PREPROCESSING & EXPLORATORY DATA ANALYSIS

DATA PRE-PROCESSING

Unique Findings

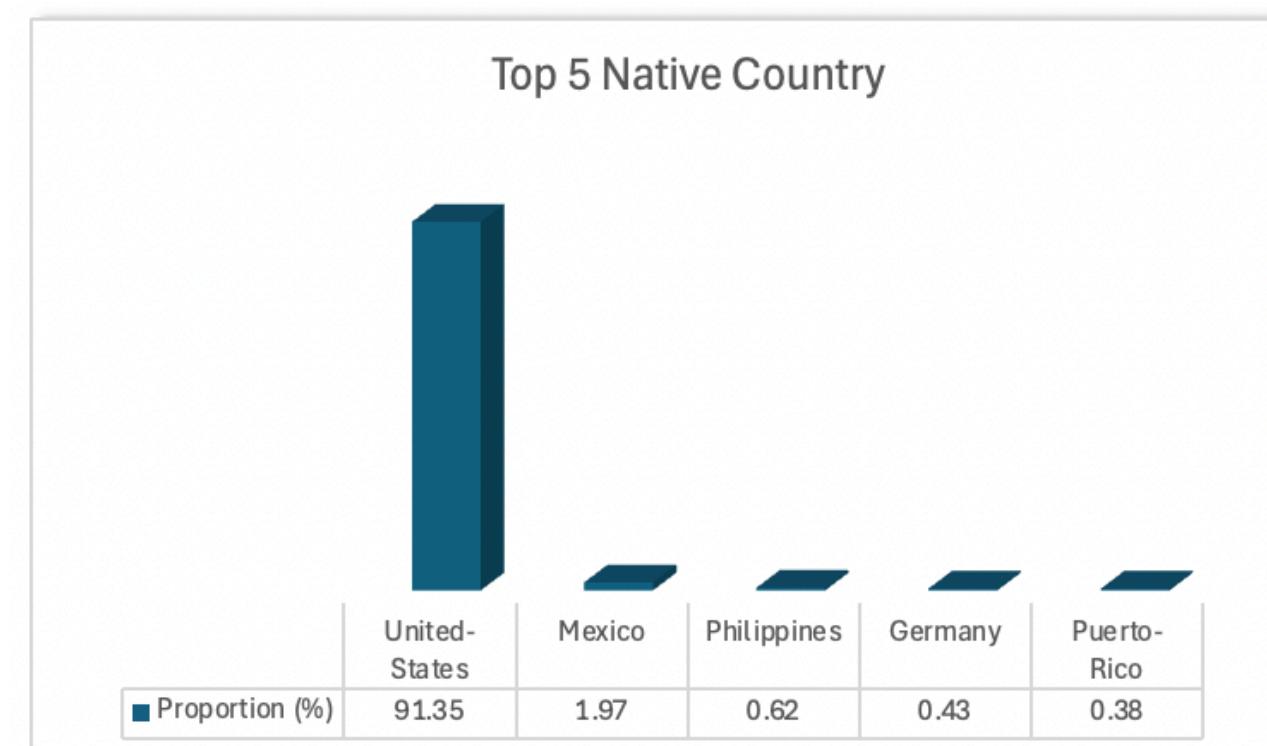
Handling Missing Values

Category	Count
Workclass	2,799
Occupation	2,809
Native Country	85

10 Data Selisih antara Workclass dan Occupation

- Workclass : Never-worked
- Occupation : NaN

Data NaN diubah menjadi No-Occupation



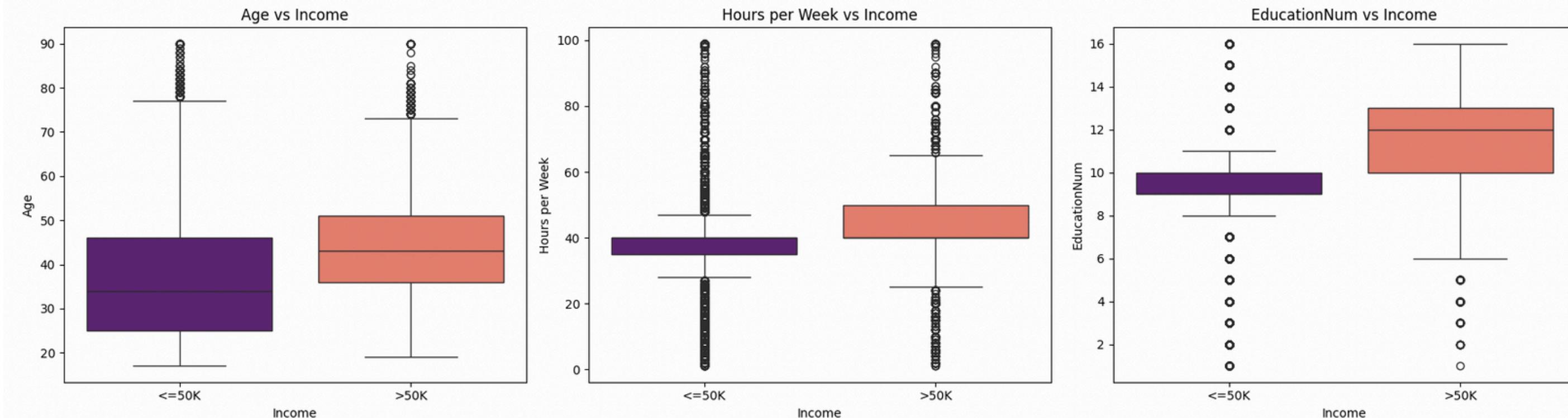
Karena **Workclass dan Occupation sama-sama kosong (NaN)** dan **proporsinya cukup besar sekitar 5%**, maka diisi "**Unknown**" karena tidak ada dasar yang valid untuk menentukan kategorinya.

91.3% data Native Country adalah United-States, sehingga dibuat simplifikasi menjadi **2 kategori** saja **United-States dan Non US**

EXPLORATORY DATA ANALYSIS

Unique Findings

Numerical Data



- 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.

Handling Outliers

- Outlier ≠ noise**= Pada data pendapatan, nilai ekstrem sering merepresentasikan individu bernilai ekonomi tinggi.
- Tidak dihapus, tetapi dikelola**= Outlier dipertahankan agar sinyal bisnis penting tidak hilang.
- Model yang robust**= Digunakan model tree-based yang secara alami tahan terhadap outlier.
- Transformasi fitur ekstrem**= Fitur sangat skewed diubah menjadi biner untuk menangkap keberadaan sinyal ekonomi.
- Fokus makna bisnis**= Model diarahkan pada sinyal ekonomi utama, bukan besarnya nilai ekstrem.

EXPLORATORY DATA ANALYSIS

Unique Findings

Numerical Data

Correlation Heatmap of Numerical Features



Secara umum, **tidak terdapat korelasi yang kuat antar fitur numerik.**

EducationNum

Memiliki **korelasi positif tertinggi** (meskipun lemah) dengan:

- **Hours per Week (0.14)**
- **Capital Gain (0.13)**

Hal ini menunjukkan bahwa individu dengan **tingkat pendidikan lebih tinggi** cenderung:

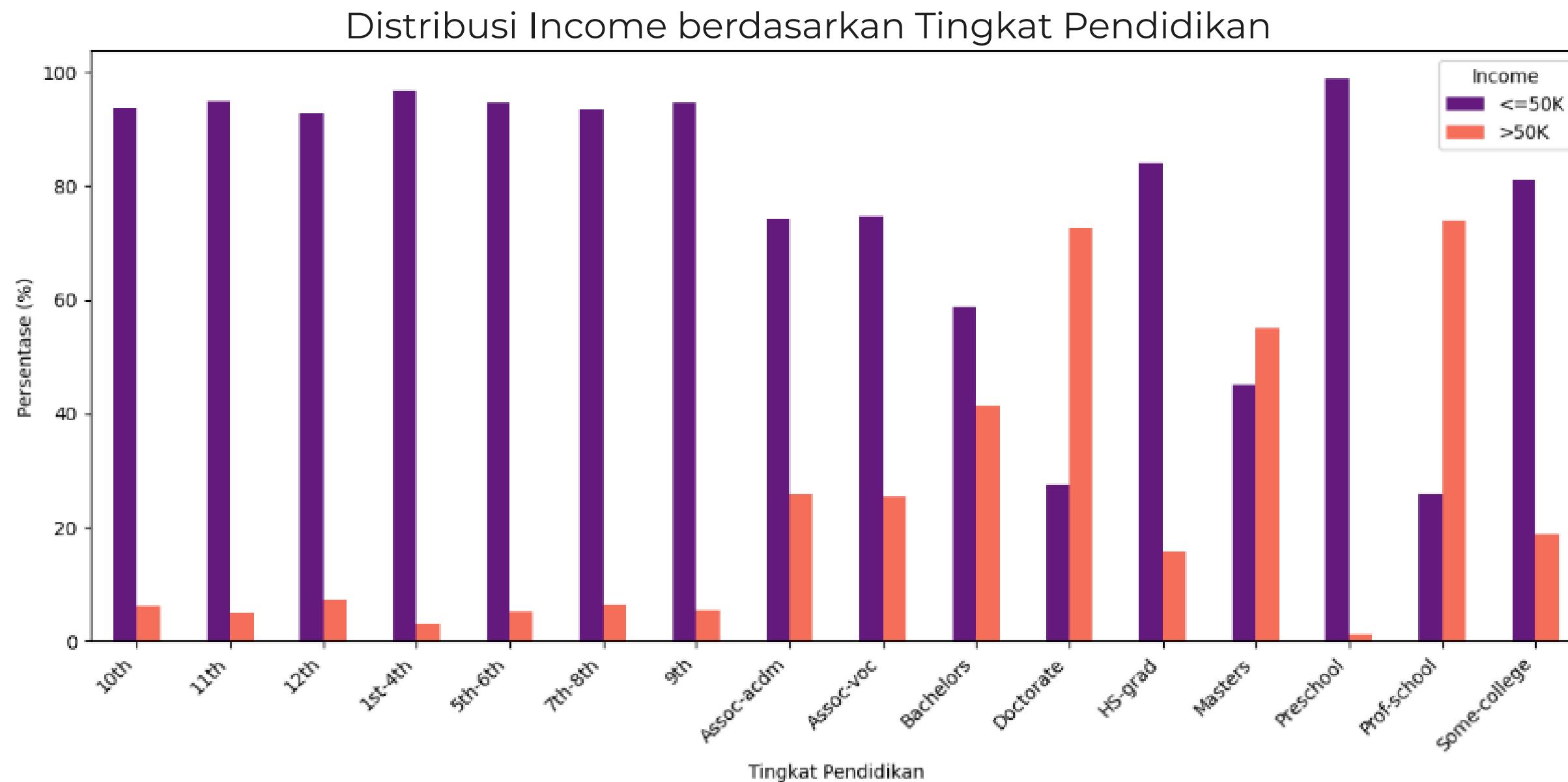
- **Bekerja sedikit lebih lama**
- **Memiliki peluang memperoleh capital gain**

EXPLORATORY DATA ANALYSIS

Unique Findings

Categorical Data

PENDIDIKAN



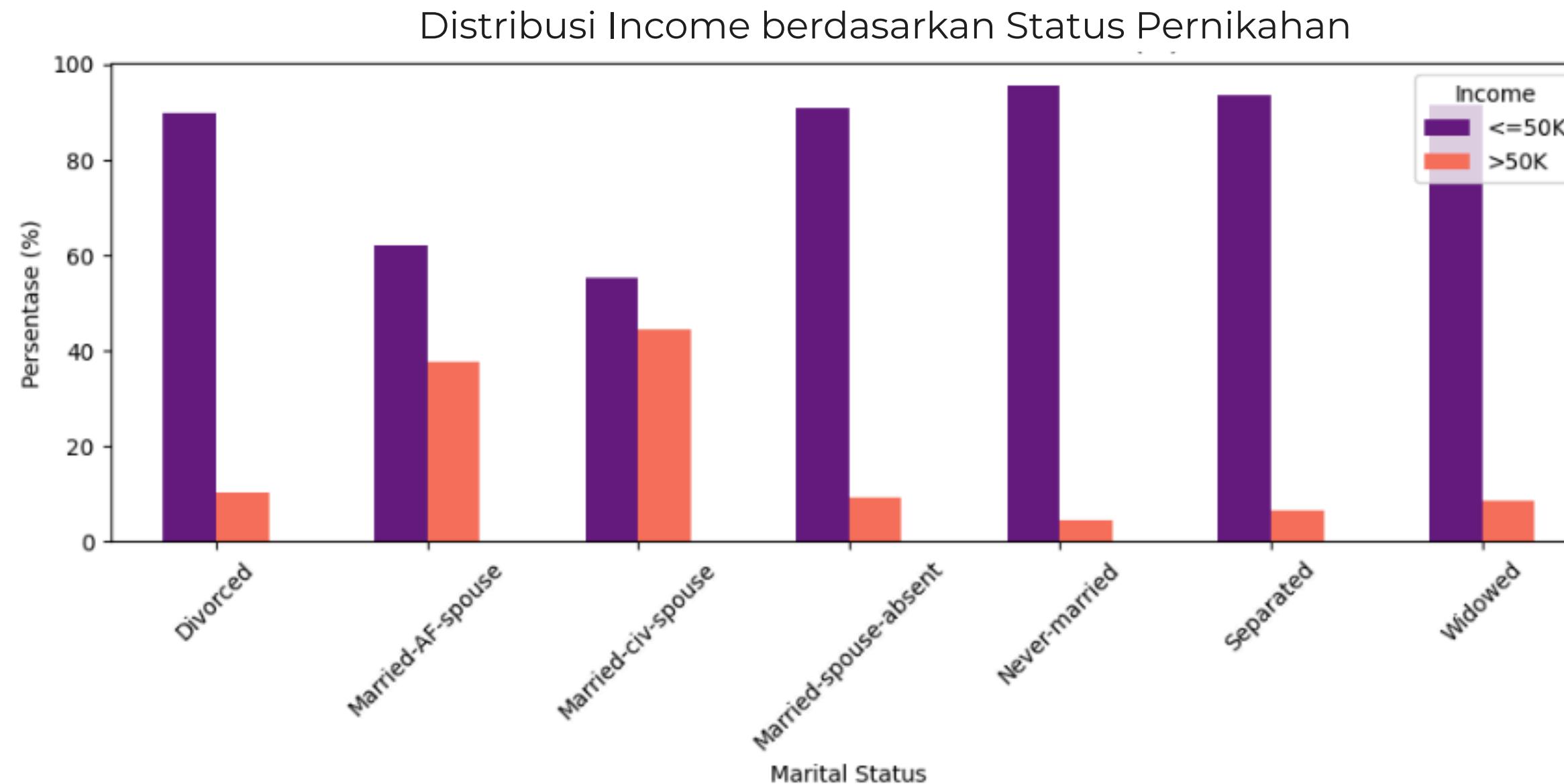
- Semakin **tinggi** **tingkat** **pendidikan**, semakin **besar** proporsi **individu** dengan pendapatan **di atas 50K**.

EXPLORATORY DATA ANALYSIS

Unique Findings

Categorical Data

STATUS PERNIKAHAN



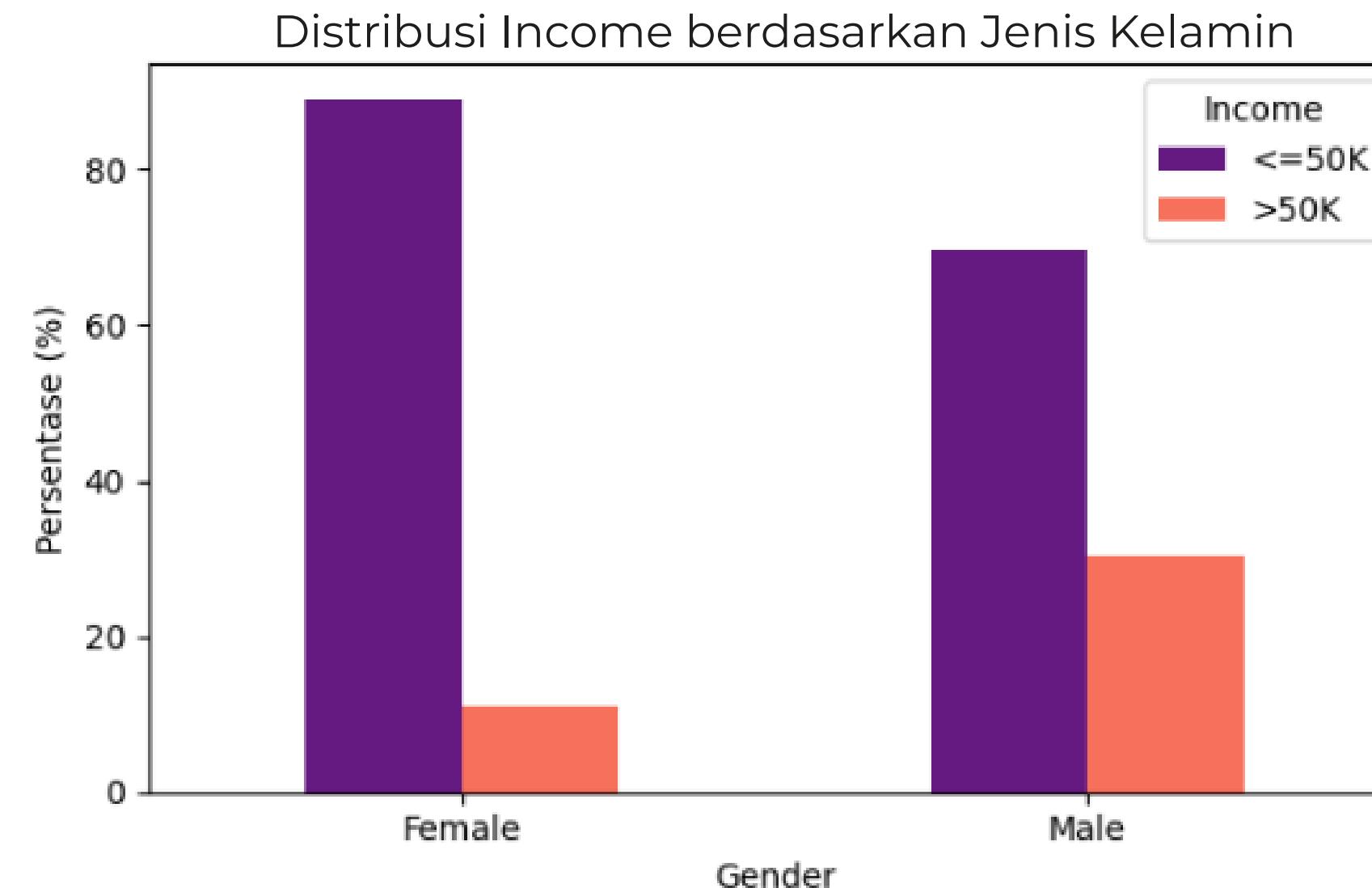
- Status pernikahan memiliki hubungan yang jelas dengan tingkat pendapatan.
- **Individu yang menikah**, khususnya **married-civ-spouse**, memiliki **peluang lebih besar** untuk memperoleh **pendapatan di atas 50K dibandingkan individu yang belum menikah** atau berada dalam **kondisi pernikahan yang tidak stabil**.

EXPLORATORY DATA ANALYSIS

Unique Findings

Categorical Data

JENIS KELAMIN



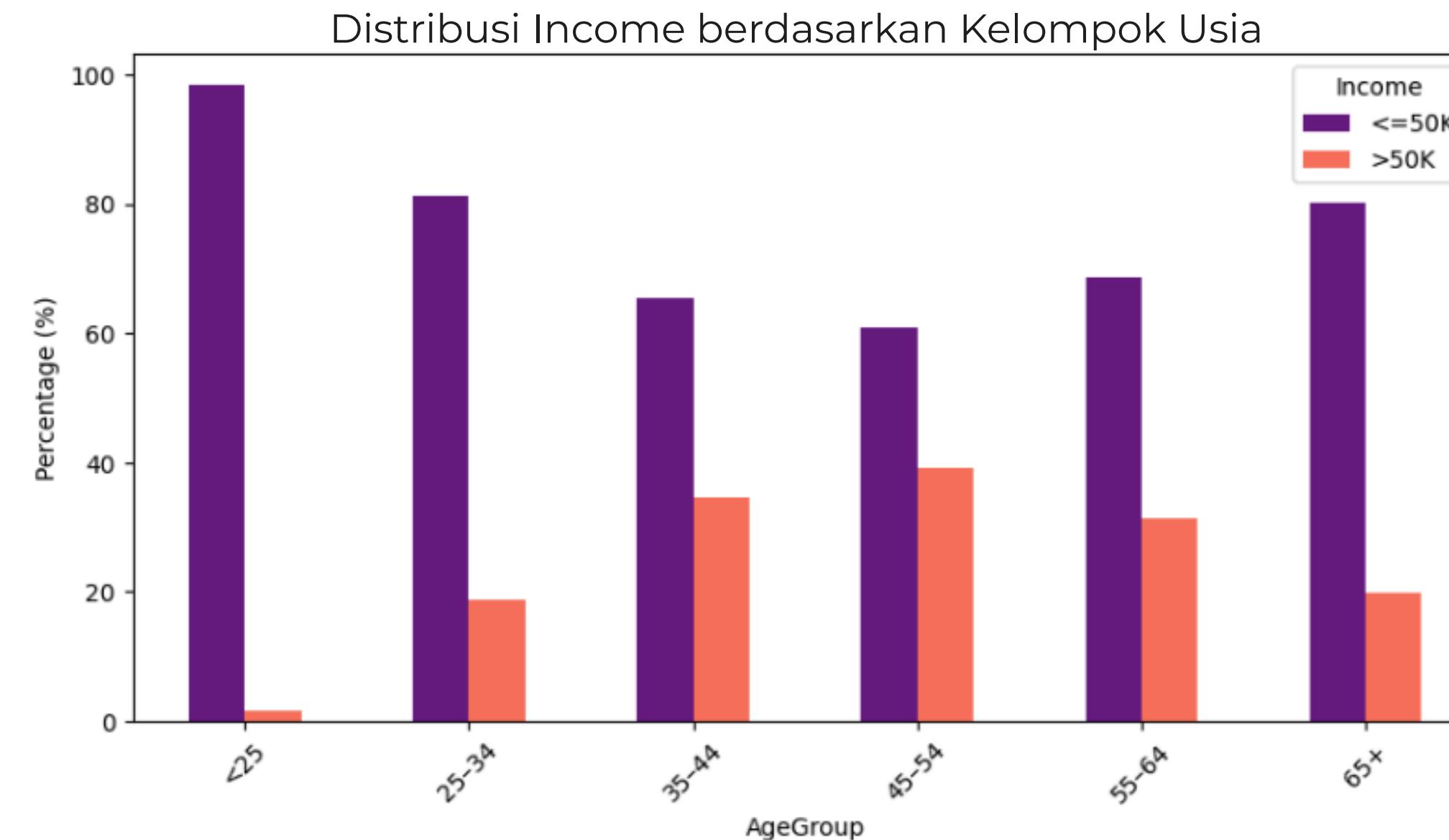
- **Laki-laki** memiliki **peluang** yang jauh **lebih besar** untuk memperoleh **income >50K dibandingkan perempuan**
- Hal ini menunjukkan adanya gender income gap dalam dataset ini.

EXPLORATORY DATA ANALYSIS

Unique Findings

Categorical Data

KELOMPOK USIA



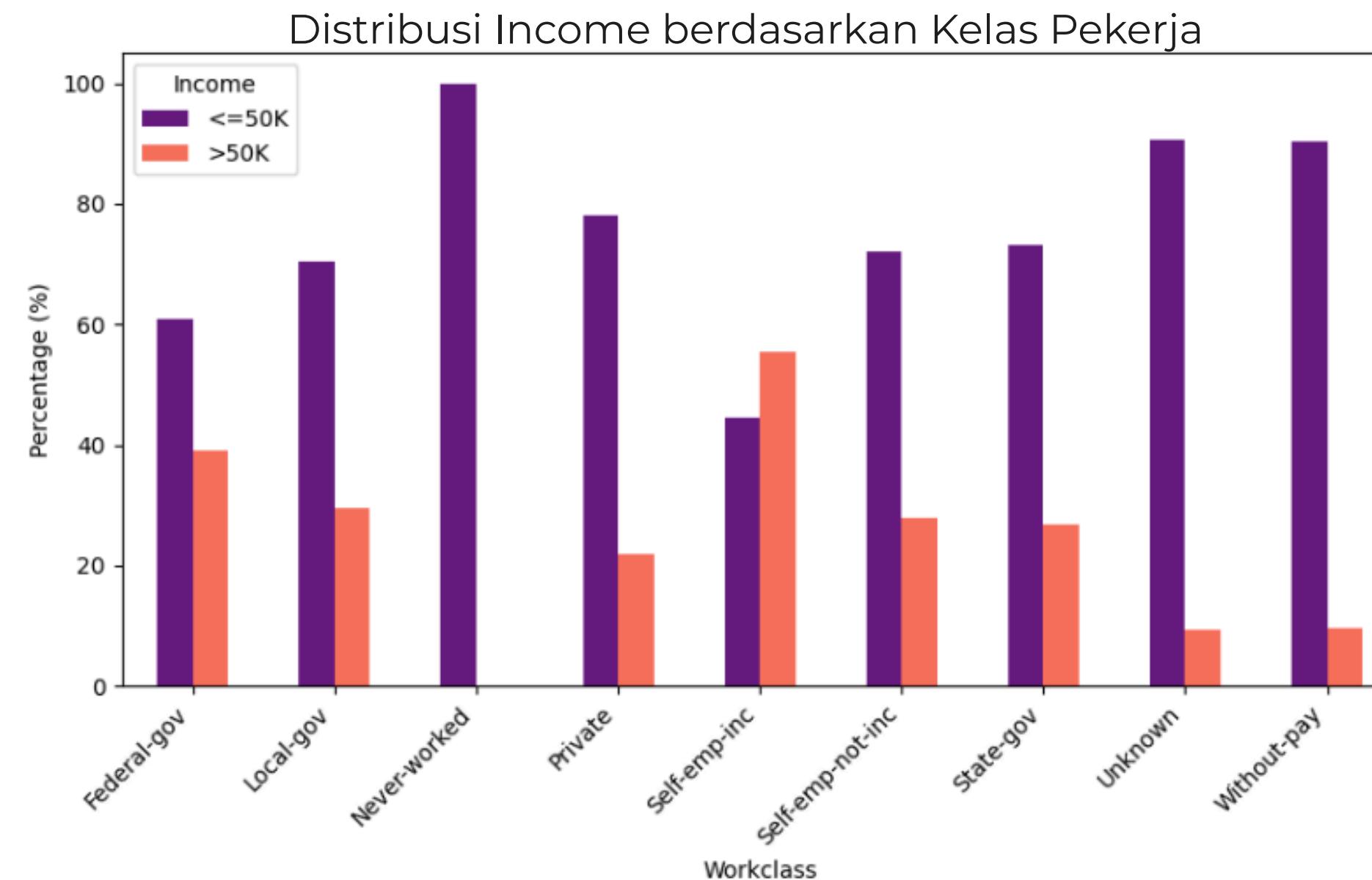
- Peluang memperoleh **income lebih dari 50K** meningkat **seiring bertambahnya usia**, mencapai **puncaknya** pada kelompok usia **45-54 tahun**, dan kemudian cenderung **menurun** pada **usia lanjut**.
- Kelompok **usia muda (<25 tahun)** didominasi oleh income **kurang dari 50K**.

EXPLORATORY DATA ANALYSIS

Unique Findings

Categorical Data

KELAS PEKERJA



- Individu yang bekerja sebagai **self-employed** dengan badan usaha atau di sektor pemerintahan memiliki **peluang lebih besar** untuk memperoleh **income lebih dari 50K dibandingkan pekerja sektor swasta atau informal.**

FEATURE ENGINEERING

Generating New Features

Has_Capital_Gain

Fitur biner yang menandai individu yang memiliki Capital Gain (nilai > 0), mengindikasikan aktivitas investasi atau aset finansial.

Has_Capital_Loss

Fitur biner yang menandai individu yang memiliki Capital Loss (nilai > 0), mengindikasikan kerugian finansial dari investasi.

Income_Encoded

Transformasi variabel target dari kategorikal (' $\leq 50K$ ', ' $>50K$ ') menjadi numerik biner (0, 1) untuk keperluan pemodelan klasifikasi.

One-Hot Encoded Features

Konversi semua variabel kategorikal (Workclass, Marital Status, Occupation, Relationship, Race, Gender, Native Country) menjadi format numerik biner menggunakan dummy variables dengan `drop_first=True` untuk menghindari multikolinearitas.

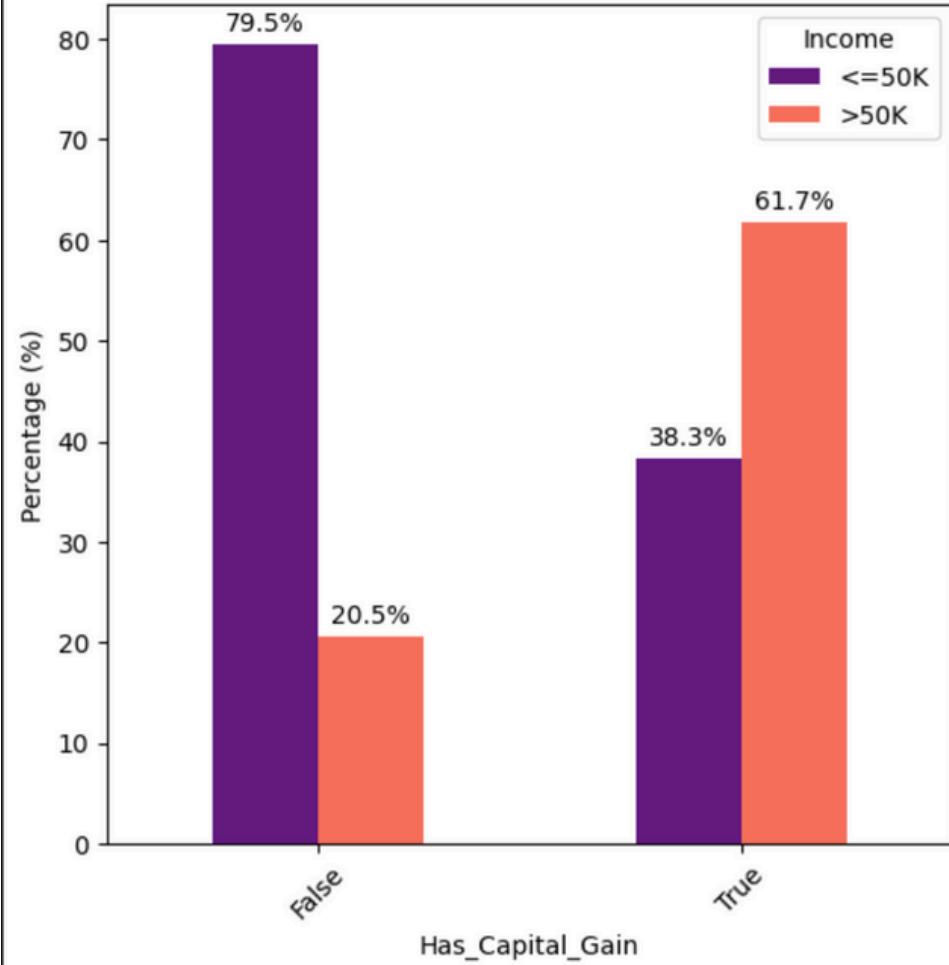
EXPLORATORY DATA ANALYSIS

Unique Findings

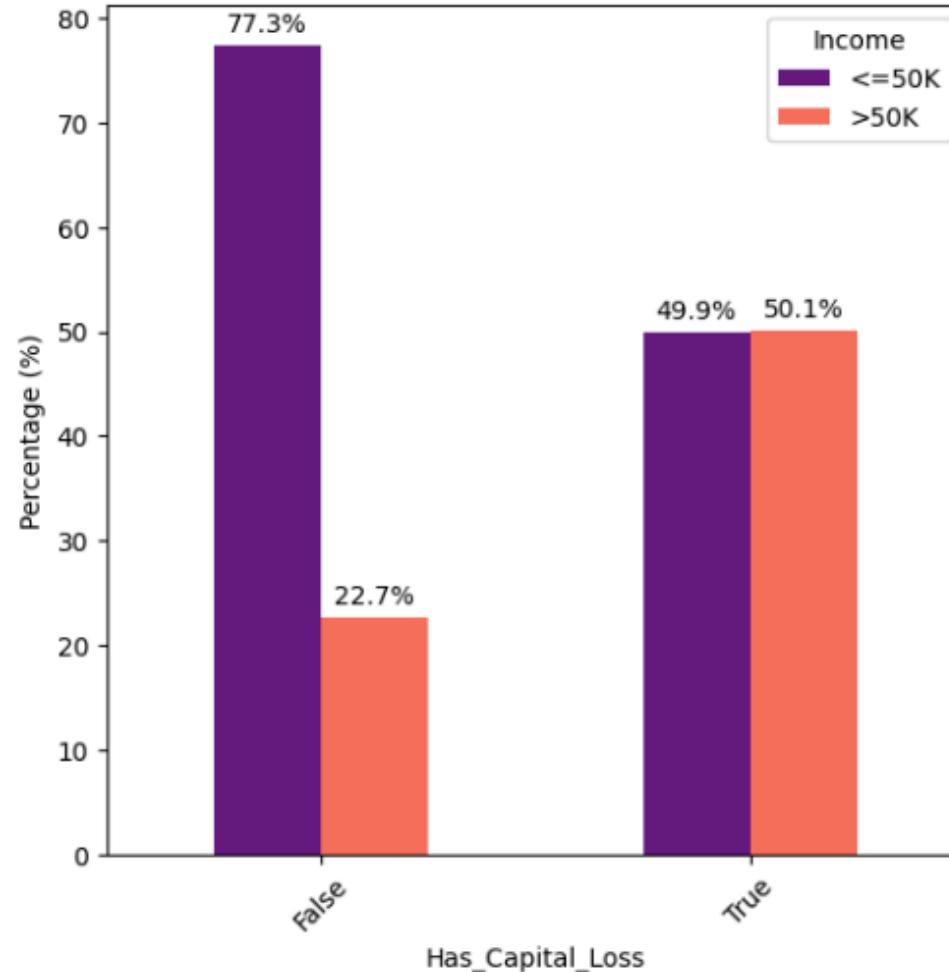
Categorical Data

CAPITAL GAIN & LOSS

Distribusi Income berdasarkan Capital Gain



Distribusi Income berdasarkan Capital Loss

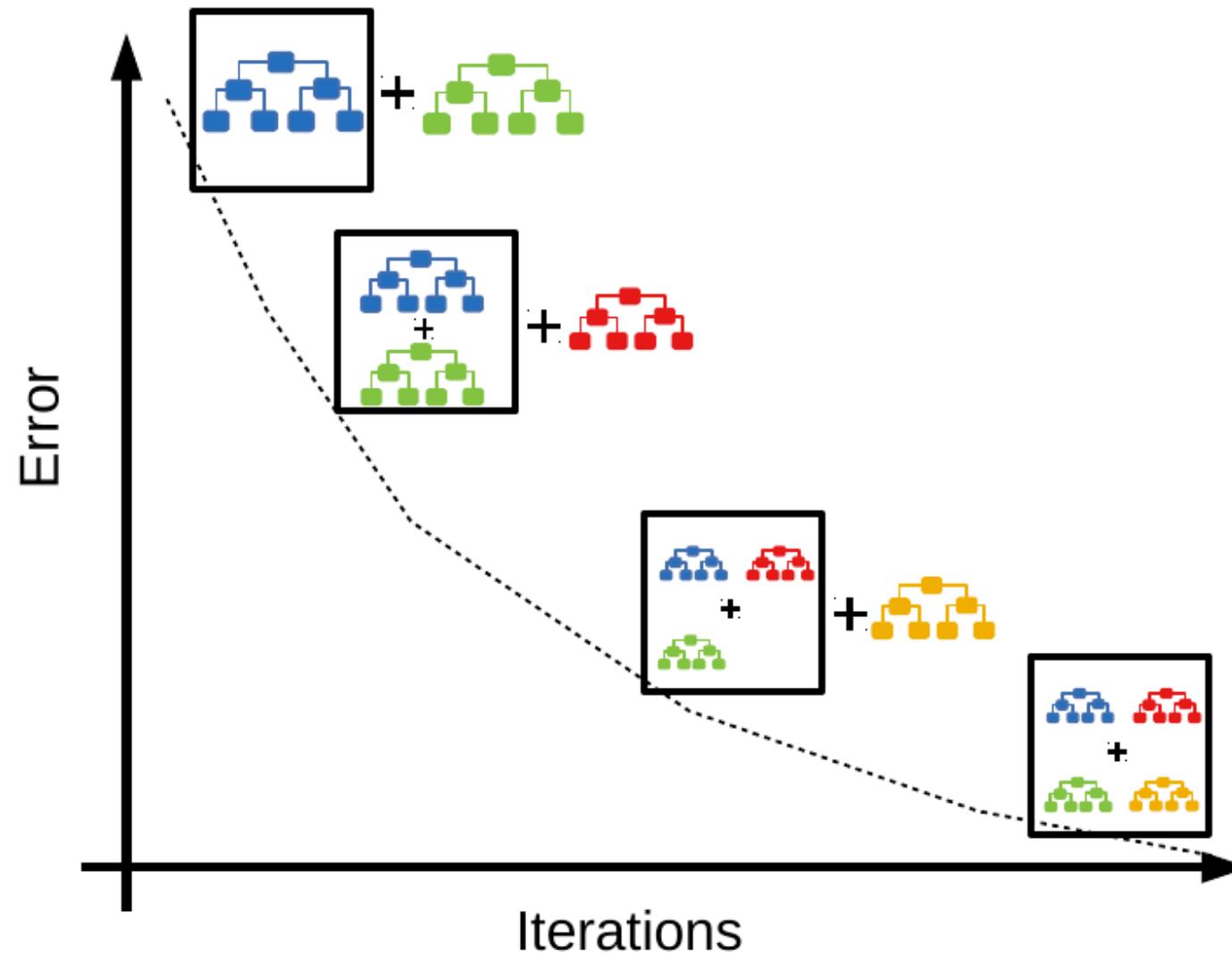


- **Capital Gain:** Individu yang memiliki capital gain memiliki **peluang hampir 3 kali lipat lebih besar** untuk memperoleh **income lebih dari 50K** dibandingkan yang tidak.
- **Capital Loss:** **Keberadaan capital loss** meningkatkan peluang individu masuk ke **kelompok pendapatan lebih dari 50K**, capital loss mencerminkan kapasitas ekonomi dan aktivitas investasi, bukan kondisi finansial buruk.

CHAPTER 3

MODELLING AND
INTERPRETATION

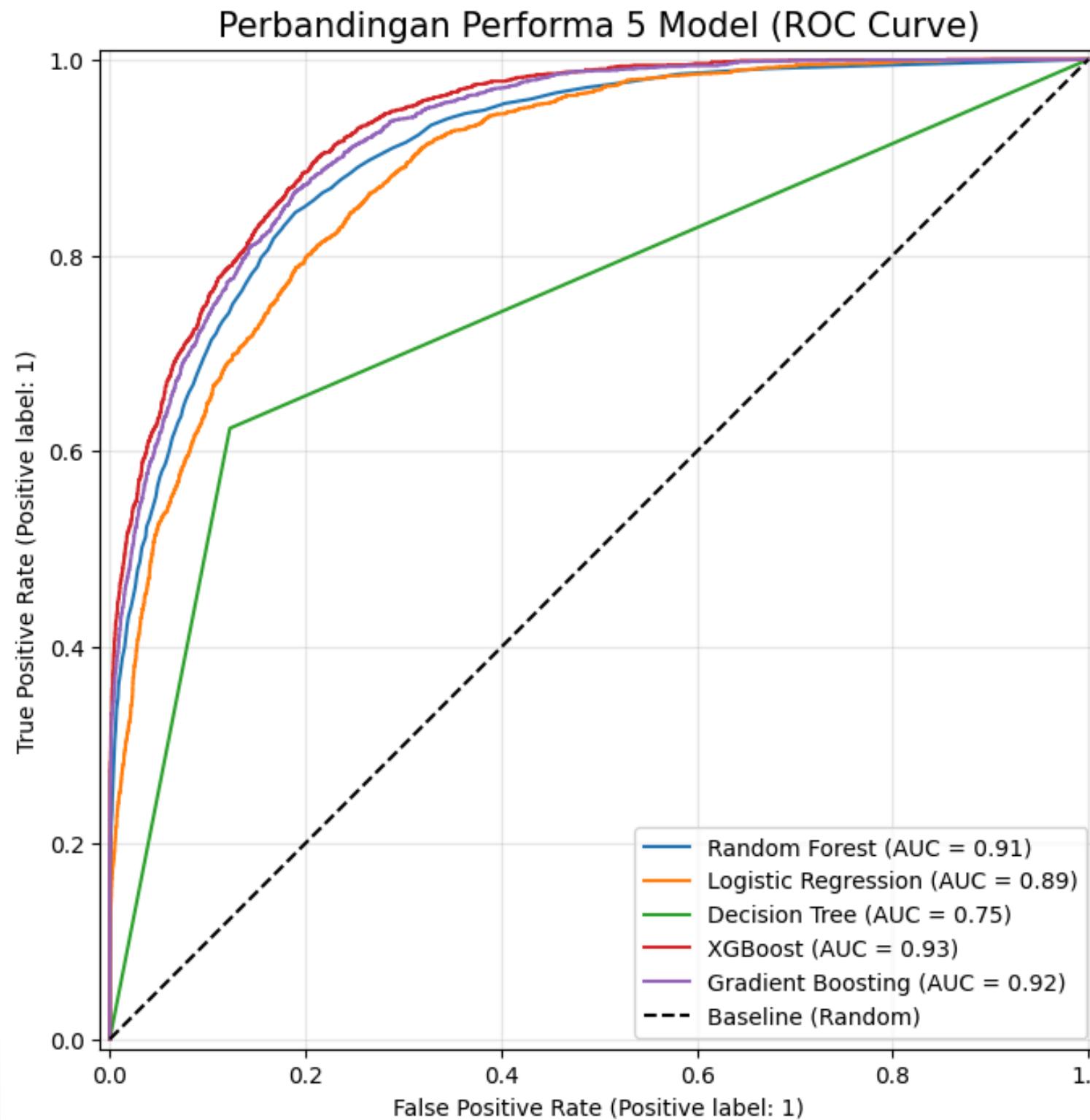
STRATEGI PEMODELAN & EKSPERIMENTASI



Mengapa Melakukan Perbandingan Model?

- **Kompleksitas Non-Linear:** Hubungan fitur demografi terhadap pendapatan sangat kompleks dan tidak selalu dapat ditangkap oleh model linear.
- **No Free Lunch Theorem:** Tidak ada algoritma tunggal yang unggul secara universal.
- **Stabilitas & Presisi:** Mencari model terbaik untuk menangani ketimpangan data (Class Imbalance) yang akurat.
- **Validasi Generalisasi:** Pengujian melalui Cross-Validation menjamin model memiliki kemampuan prediksi yang stabil pada data baru (avoid overfitting).

ROC CURVE



1. Perbandingan ROC-AUC

- 🏆 XGBoost — 0.93 → Performa terbaik & paling presisi
- 🥈 Gradient Boosting — 0.92 → Akurat, namun lebih lambat
- 🥉 Random Forest — 0.91 → Stabil & robust

2. Keunggulan Utama (XGBoost)

Anti-overfitting: Regularisasi menjaga konsistensi performa
Efisiensi tinggi: Pemisahan kelas (TPR-FPR) paling optimal

3. Metodologi & Evaluasi

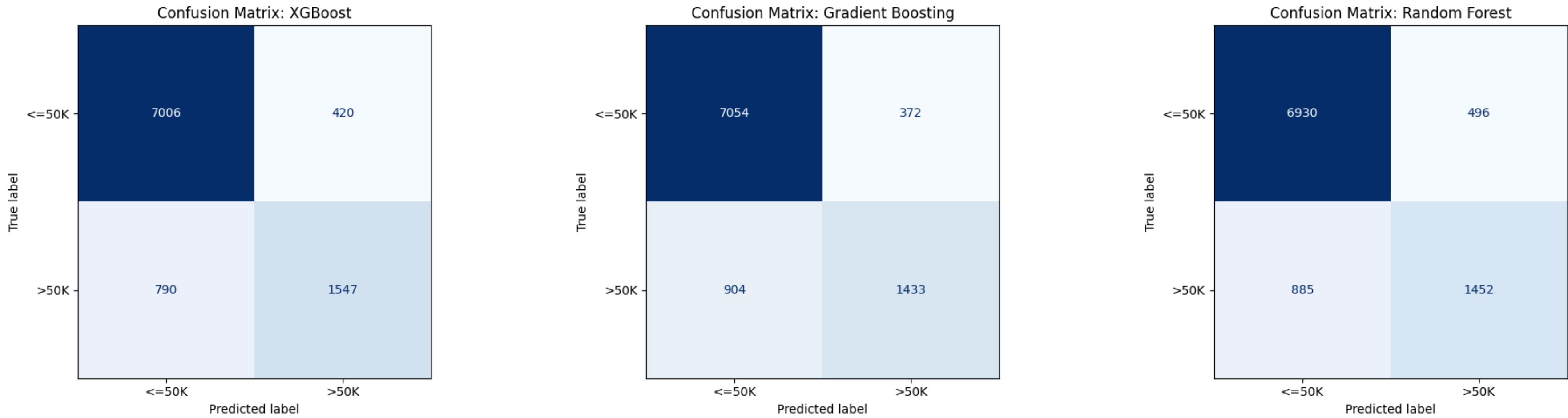
Validasi: 5-Fold Cross-Validation
Metrik utama: ROC-AUC (tahan ketimpangan data)
Fitur: 48 fitur hasil One-Hot Encoding

4. Threshold & Optimasi Bisnis

Default threshold: 0.5 (fleksibel)
Presisi ↑ : Threshold > 0.5
Recall ↑ : Threshold < 0.5

XGBoost menunjukkan **Performa terbaik** dalam **membedakan** antara **kedua kelas** pendapatan.

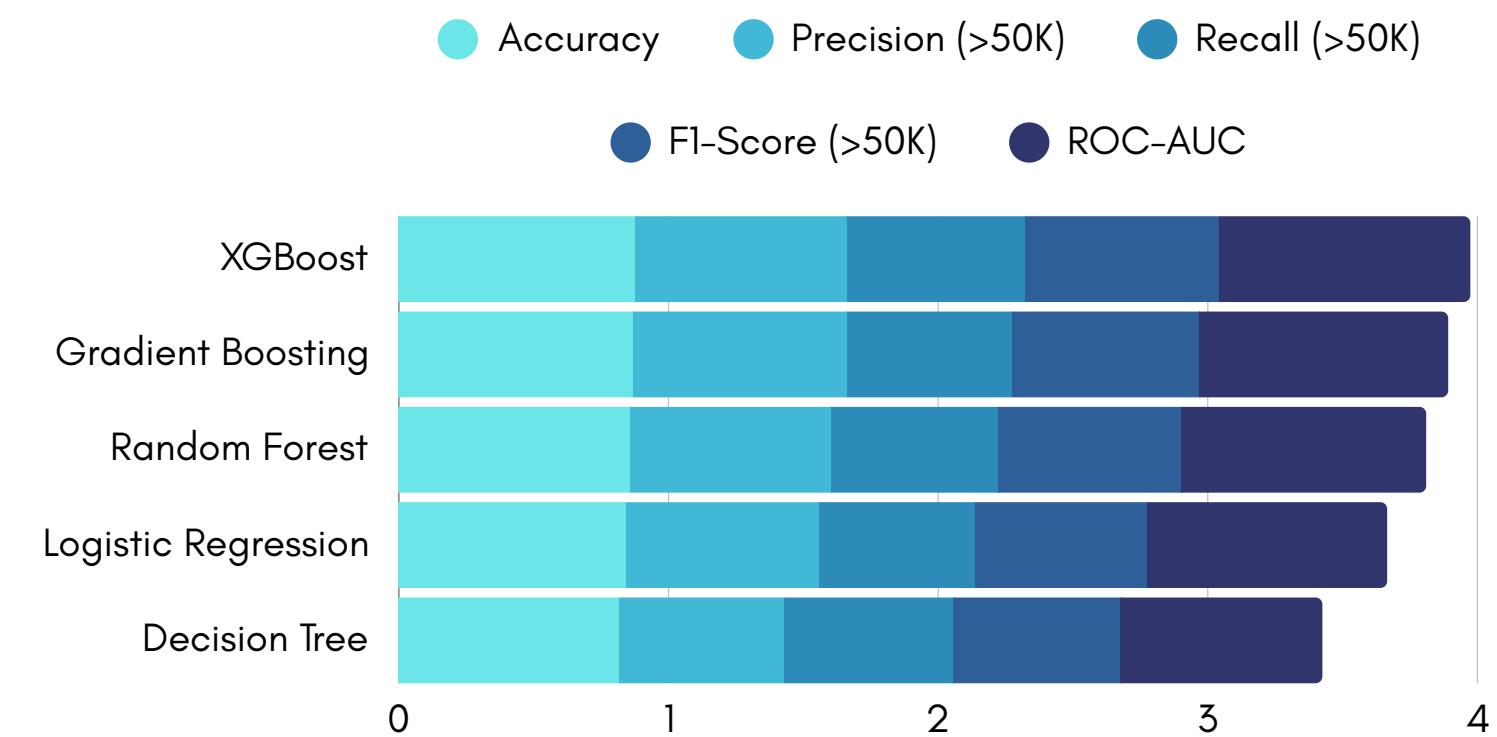
CONFUSION MATRIX



- **XGBoost** adalah **model terbaik**: paling banyak mendekripsi income >50K dengan kesalahan paling rendah.
- **Gradient Boosting** lebih **konservatif**: jarang salah memprediksi income tinggi, tapi lebih banyak **melewatkannya**.
- **Random Forest** seimbang: **performa stabil**, namun **sedikit lebih agresif** dan menghasilkan lebih banyak false positive.

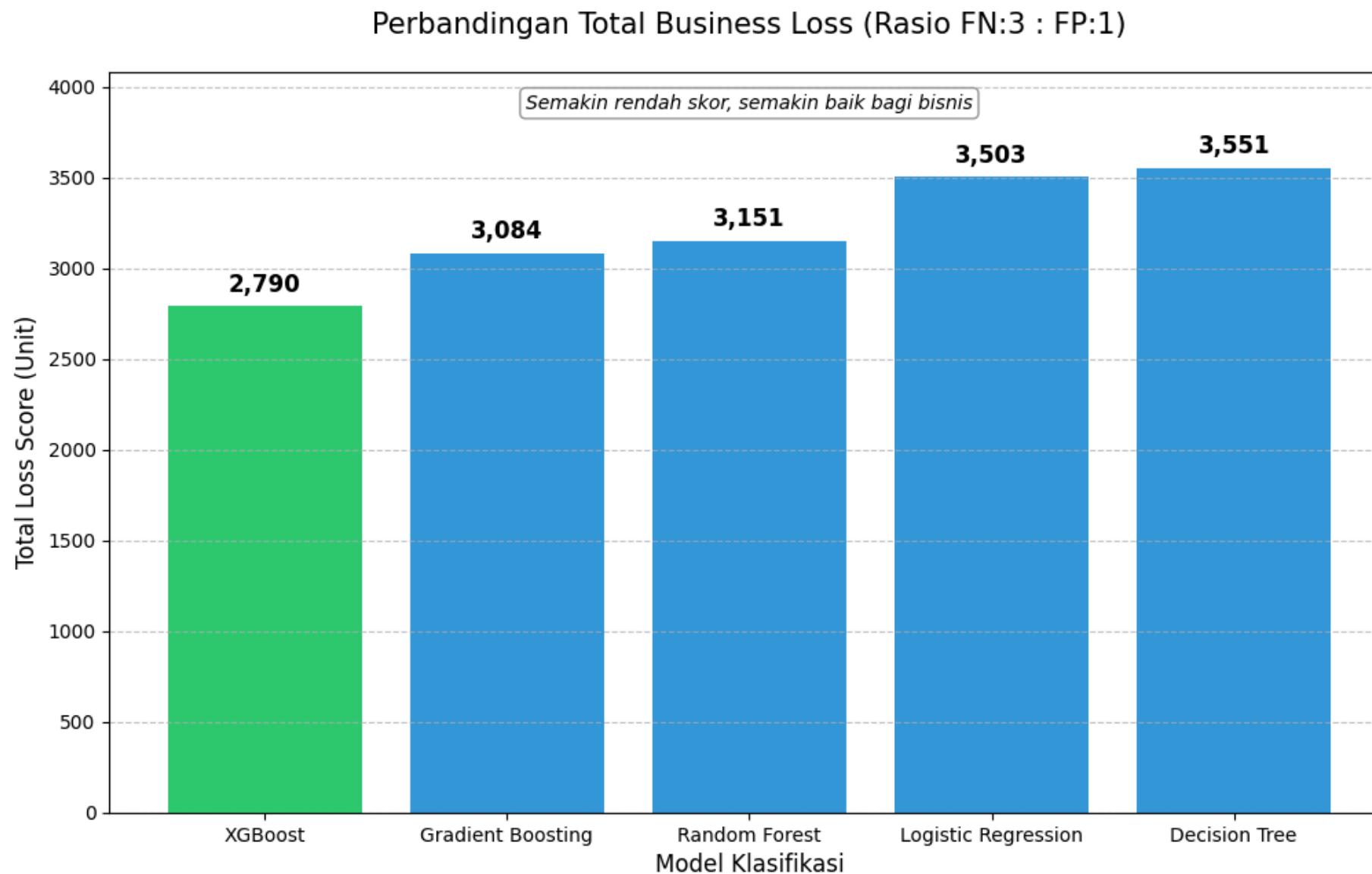
COMPARATIVE EVALUATION OF MODEL TRAINING

Model	Accuracy	Precision (>50K)	Recall (>50K)	F1-Score (>50K)	ROC-AUC
XGBoost	0.876063	0.786477	0.66196	0.718866	0.930432
Gradient Boosting	0.869302	0.793906	0.613179	0.691936	0.923371
Random Forest	0.858548	0.74538	0.621309	0.677713	0.907077
Logistic Regression	0.844208	0.717484	0.575952	0.638975	0.888714
Decision Tree	0.816552	0.615287	0.623449	0.619341	0.750386



- **MODEL TERPILIH: XGBoost**
- XGBoost dipilih bukan hanya berdasarkan akurasi, tetapi karena kemampuannya mengurangi business loss dengan meminimalkan False Positive dan False Negative secara seimbang.
- **Optimalisasi Recall (XGBoost 0.66):** Efektif meminimalkan False Negative untuk mereduksi Opportunity Loss dalam menjaring calon pembeli berpenghasilan tinggi (>50K).

STRATEGIC BUSINESS IMPACT & MODEL OPTIMIZATION



Expected Value Framework

$$E[X] = \sum x_i p(x_i)$$

Rumus Optimasi Bisnis

$$\text{Total Loss} = (FP \times \text{Cost}_F) + (FN \times \text{Cost}_N)$$

Asumsi Biaya Rasio (3:1)

$$\text{Total Loss} = (FP \times 1) + (FN \times 3)$$

Strategi bisnis: Recall-oriented

Asumsi biaya: False Negative lebih mahal daripada False Positive kita buat rasio 3:1

Metodologi: Rasio Risiko 3:1

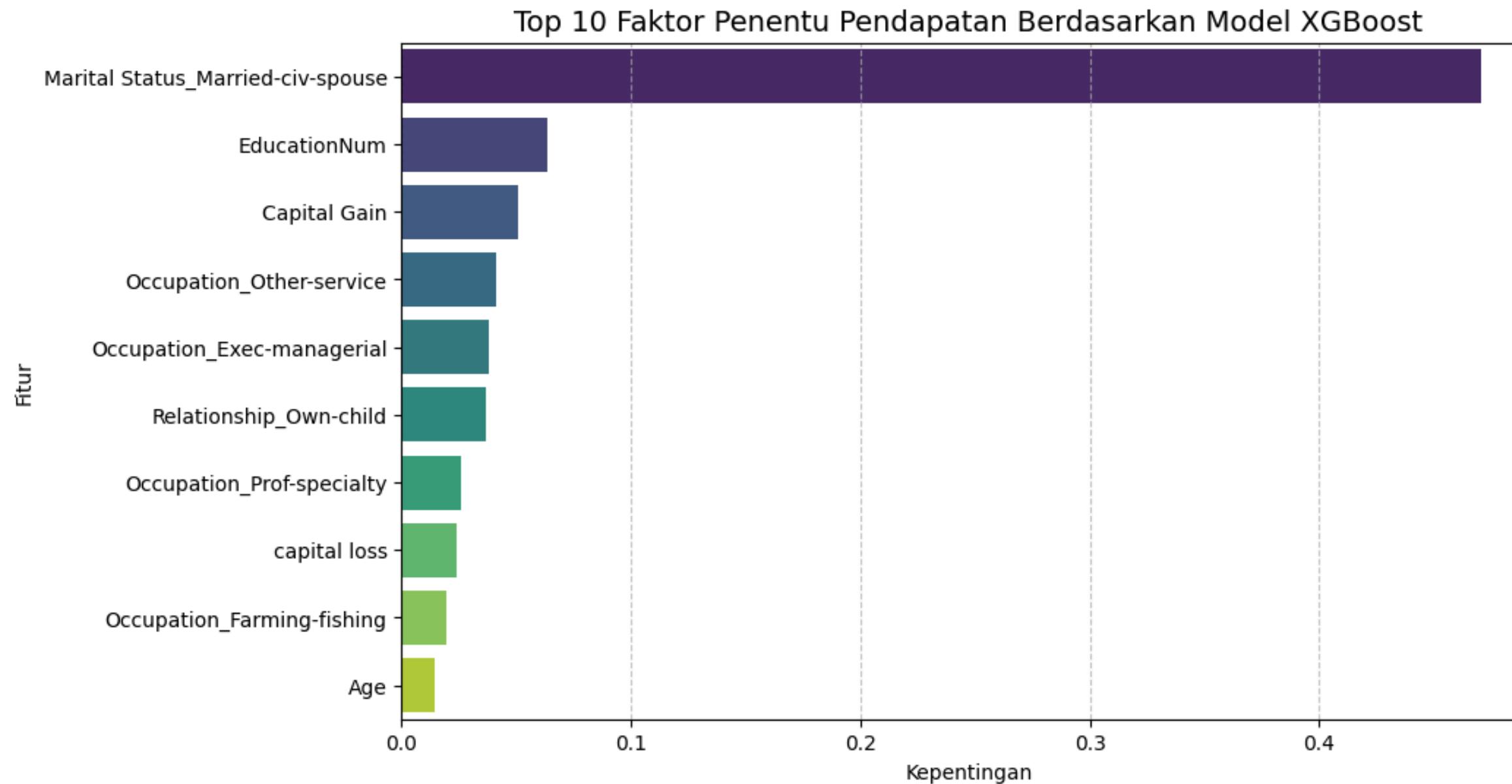
- Prioritas:** Menghindari kehilangan potensi pendapatan besar (False Negative dibobot 3x lebih berat).
- Fokus:** Menyeimbangkan biaya operasional vs. efektivitas deteksi.

Hasil Utama:

- XGBoost (Juara):** Menghasilkan Skor Kerugian Terendah dibandingkan 4 model lainnya.
- Efisiensi:** Memangkas risiko kerugian hingga >20% dibanding model standar.
- Korelasi:** Akurasi teknis (AUC 0.93) terbukti berbanding lurus dengan efisiensi ekonomi.

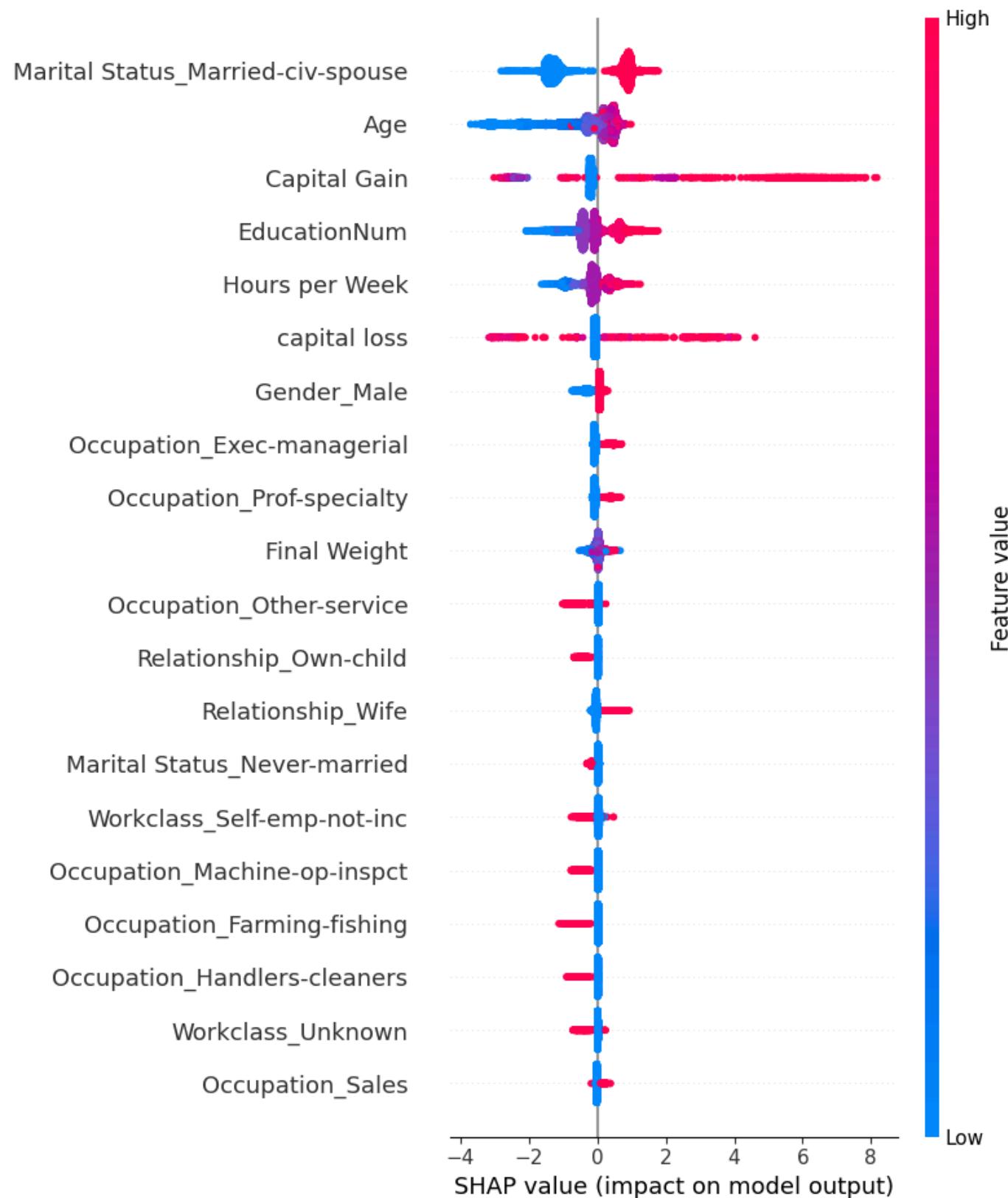
Kesimpulan: XGBoost adalah model paling aman dan paling menguntungkan bagi bisnis.

Feature Importance



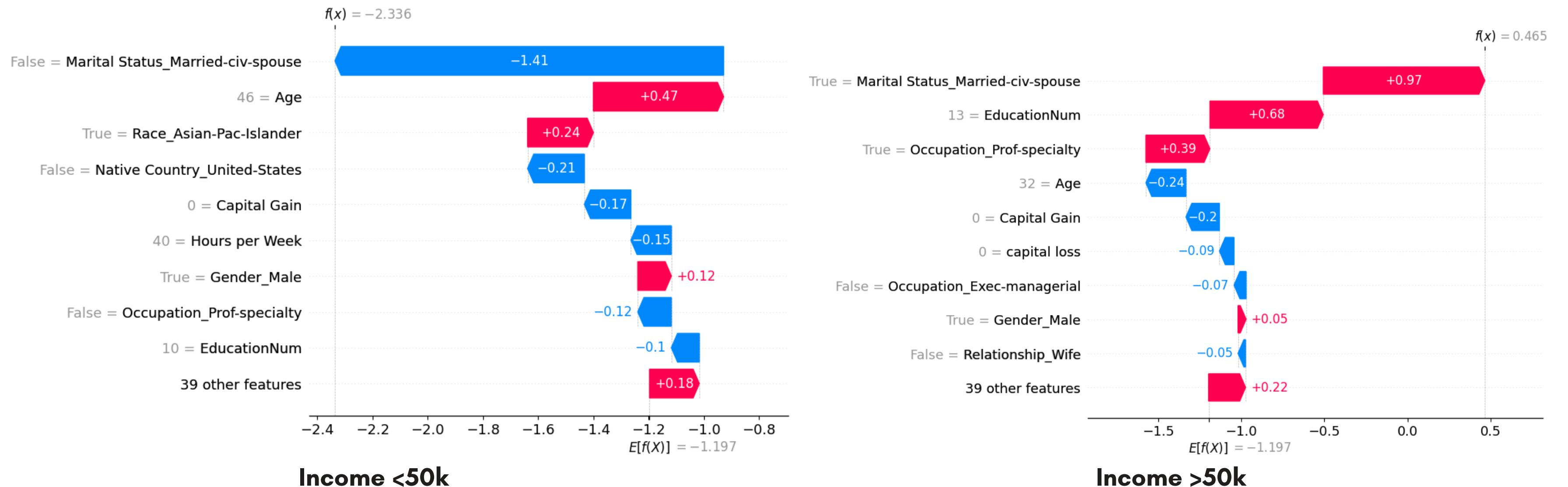
Model **XGBoost** menilai status pernikahan, pendidikan, dan kondisi ekonomi (capital gain) sebagai penentu utama peluang income tinggi.

SHAP



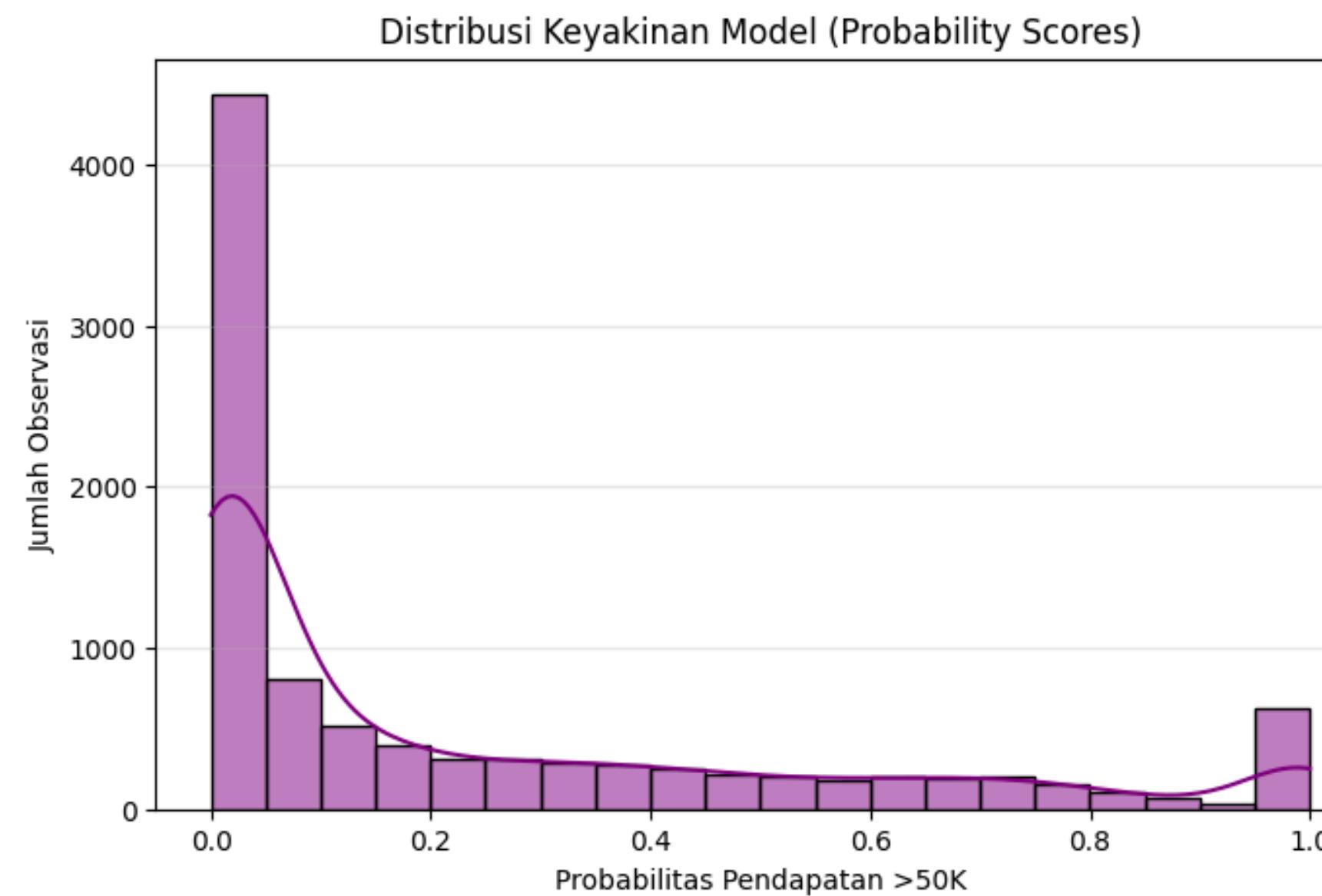
Marital Status (Married) → faktor paling kuat menaikkan peluang income >50K
Age, EducationNum, Capital Gain → semakin tinggi nilainya, semakin besar peluang income tinggi
Jam kerja & capital loss sedikit berpengaruh
Beberapa jenis pekerjaan cenderung **menurunkan peluang income tinggi**

SHAP



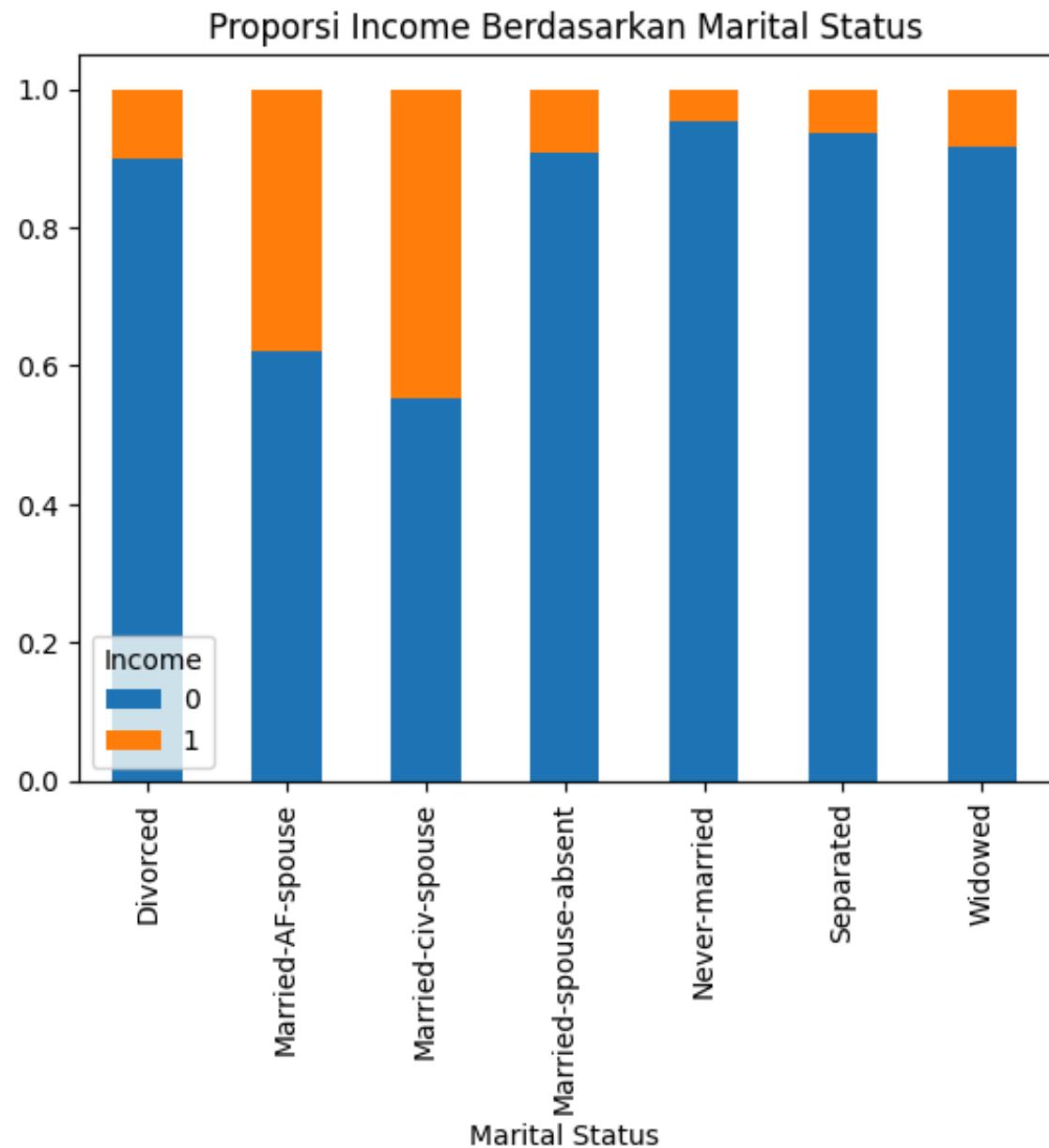
- **Income tinggi (>50K)**: didorong oleh menikah, pendidikan tinggi, dan pekerjaan profesional.
- **Income rendah ($\leq 50K$)**: dipengaruhi oleh tidak menikah, tidak ada capital gain, dan faktor pekerjaan tertentu.

Distribusi Probabilitas



- **Mayoritas data memiliki probabilitas rendah** (<0.2) → diprediksi berincome $\leqslant 50K$.
- Area 0.4–0.6 jarang muncul → **ambiguitas prediksi rendah**.
- **Kelompok kecil dengan probabilitas tinggi** ($\geqslant 0.8$) → model sangat yakin income $> 50K$

Analisis Proporsi



- Proporsi pendapatan $\leq 50K$ (biru) dan $> 50K$ (orange).
- **Married-civ-spouse** memiliki proporsi **income $> 50K$** paling **tinggi** dibanding status lain.
- **Married-AF-spouse** juga menunjukkan **peluang income $> 50K$ yang relatif tinggi**, meski jumlah datanya lebih sedikit.
- **Never-married, Divorced, Separated, Widowed** didominasi oleh **income $\leq 50K$** .
- **Married-spouse-absent** cenderung memiliki proporsi **income $> 50K$ yang rendah**

Model Limitation

Bias Kontekstual (US-Centric)

- Model dilatih menggunakan data Sensus Amerika Serikat tahun 1994.
- Hasil prediksi sangat bergantung pada standar ekonomi, mata uang, dan sistem pendidikan AS di masa tersebut, sehingga keterbatasan konteks wilayah dan waktu.

Ketidakseimbangan Data (Imbalanced Bias)

- Mayoritas data adalah kelas pendapatan $\leq 50K$.
- Model cenderung lebih akurat menebak orang berpendapatan rendah, namun berisiko tinggi salah memprediksi orang kaya sebagai "tidak kaya" (False Negative).

Masalah Transparansi (Black-Box)

XGBoost adalah model kompleks yang sulit menjelaskan secara intuitif mengapa keputusan tertentu diambil pada level individu.

Model Limitation

Data Statis & Inflasi

Model tidak mengenal inflasi. Ambang batas \$50,000 di tahun 1994 memiliki nilai ekonomi yang jauh berbeda dengan saat ini, sehingga model tidak mencerminkan daya beli modern.

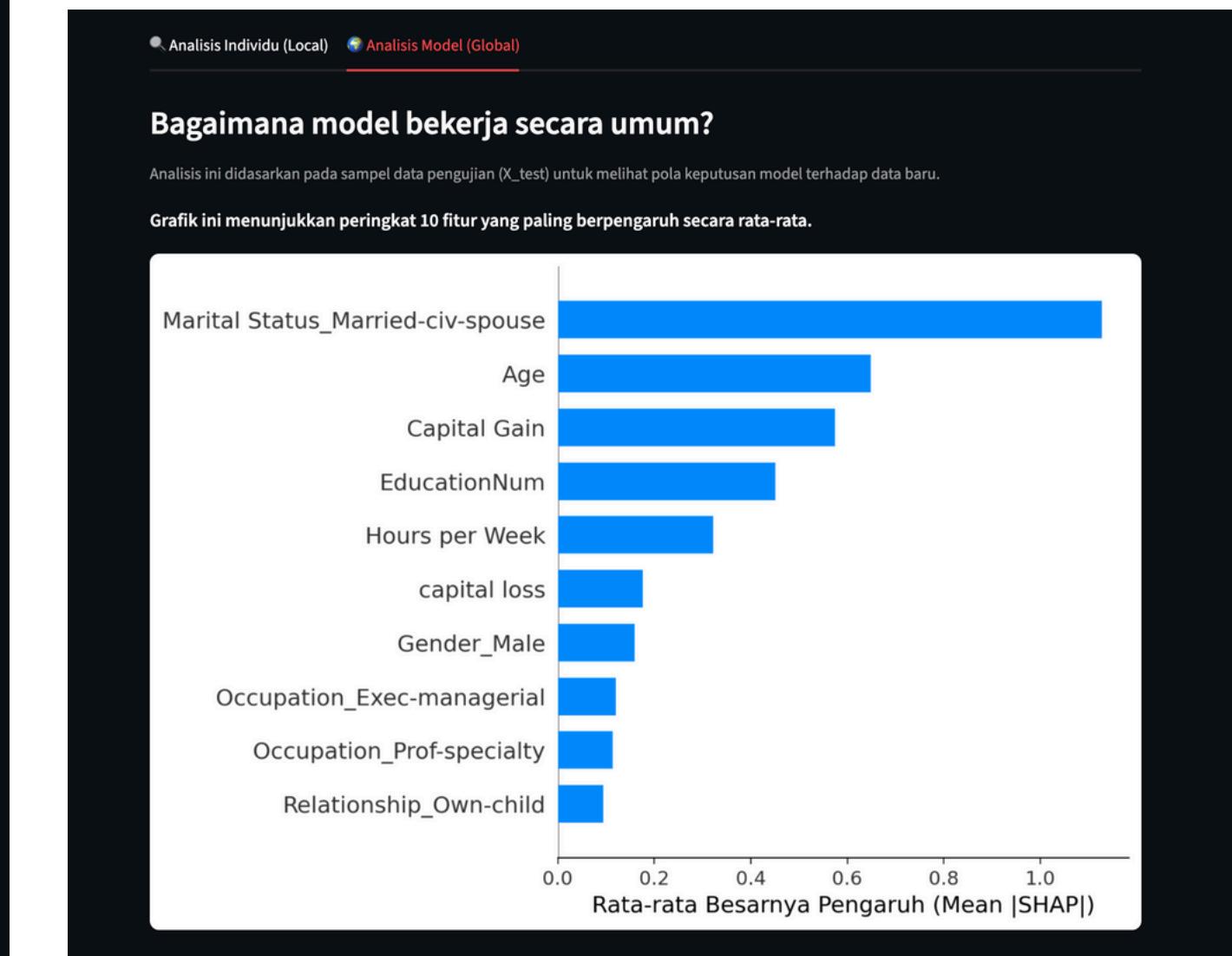
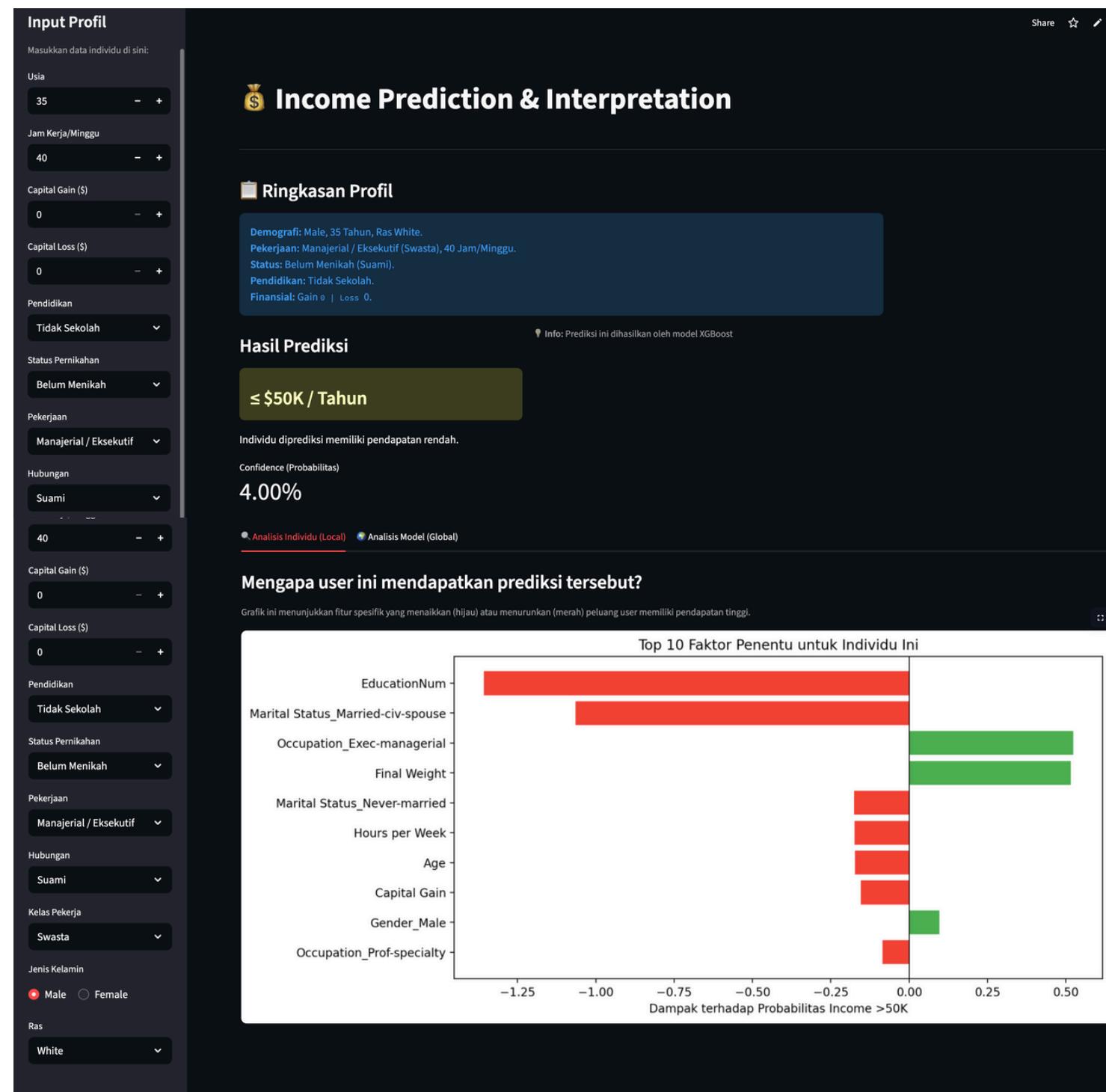
Kekakuan Struktur Input

Sangat sensitif terhadap format data. Kesalahan kecil pada nama fitur atau kolom yang hilang akan membuat sistem gagal total (crash), sehingga membutuhkan preprocessing yang sangat ketat di sisi aplikasi.

Keterbatasan SHAP

Meskipun memberikan penjelasan fitur, **SHAP hanya menunjukkan korelasi statistik**, bukan hubungan sebab-akibat (causality) yang pasti secara sosiologis.

MODEL DEPLOYMENT ON STREAMLIT



INCOME PREDICTOR ANALYSIS

THE MODEL CAN BE ASSESSED IN: [HTTPS://INCOME-PREDICTION-FINAL-PROJECT-RBHLDGDX4QPVE2ATJF7EWH.STREAMLIT.APP/](https://income-prediction-final-project-rbhldgdx4qpve2atjf7ewh.streamlit.app/)

CHAPTER 4

CONCLUSIONS & RECOMMENDATIONS

KESIMPULAN

MODEL TERBAIK

XGBoost terpilih sebagai model paling optimal karena akurasinya yang tinggi dan performa yang stabil dalam memprediksi kategori pendapatan.

FAKTOR PENENTU PENDAPATAN

Variabel EducationNum (tingkat pendidikan), Age (usia), dan Hours per Week (jam kerja) diidentifikasi sebagai fitur paling berpengaruh terhadap peluang pendapatan di atas \$50K.

KUALITAS DATA

Dataset mencakup 48.842 individu dengan distribusi fitur yang representatif. Setelah pembersihan dan preprocessing, data terbebas dari duplikasi dan missing values. Korelasi antar fitur rendah (<0.15), menunjukkan independensi variabel yang baik untuk pemodelan prediktif.

REKOMENDASI

- **Optimasi Model:** Melakukan Hyperparameter Tuning lanjutan pada XGBoost (seperti menyesuaikan learning rate dan depth) untuk memeras performa maksimal.
- **Penanganan Ketimpangan Data:** Menerapkan teknik Oversampling (SMOTE) guna meningkatkan kemampuan model dalam mengenali kelompok berpendapatan tinggi yang jumlahnya lebih sedikit.
- **Analisis Bias & Keadilan:** Melakukan evaluasi tambahan untuk memastikan prediksi model tetap adil dan tidak bias terhadap variabel sensitif seperti jenis kelamin atau ras.
- **Pemeliharaan Model:** Melakukan pemantauan berkala terhadap data drift pasca-implementasi agar prediksi tetap akurat di tengah perubahan kondisi ekonomi.

KONTRIBUSI PENGERJAAN TUGAS FINAL PROJECT GROUP 1

Nama	Bagian Yang dikerjakan
Aulia Aorama	Data Pre-Processing, EDA, Model Deployment, Controlling Project, PPT
Farras Zihan Harmany	Data Pre-Processing, EDA, PPT
Leonard Ari Raharja	Feature Engineering, Model Training & Evaluating, Deployment, PPT
Shabiha Rahma Fauziah	Project Initiaition, PPT
Rifqi Permadi	Feature Engineering, Deployment, PPT

TERIMA KASIH

DEMOGRAPHIC PATTERN ANALYSIS: ALGORITHMIC MODELING FOR CLASSIFYING HIGH-INCOME PROFILES

Shabiha Rahma Fauziah · Aulia Aorama
· Farras Zihan Harmany ·
Rifqi Permadi · Leonard Ari Raharja

