

# **Predict Clicked Ads Customer Classification by using Machine Learning**

Bernadetta Quinta

# Agenda



**Created by:**

**Bernadetta Quinta P.**

quintapradiptaa@gmail.com

<https://www.linkedin.com/in/bernadettaquinta/>

**Resume:**

Lulusan Magister Kebijakan Publik dengan pengalaman lebih dari 3 tahun di bidang pendidikan, komunikasi, dan analisis kebijakan. Saat ini aktif sebagai policy analyst dan minat besar pada bidang data. Bernadetta memiliki perhatian tinggi terhadap detail, kemampuan komunikasi yang baik, dan cepat beradaptasi dalam menghadapi tantangan.

Overview

EDA & Insight

Data Pre-Processing

Data Modelling

Business Recommendation & Simulation

# Data Overview

Periode: -

Populasi: 1000 pelanggan

Testing: 20%

Modeling: 80%

Proporsi data: Tidak Klik: 500 dan Klik: 500

- Daily Internet Usage
- Daily Time Spent on Site
- Age
- Area Income
- Gender
- City
- Province
- Category
- Clicked On Ad

List of Variable



# Dataset Exploration

No	Variabel	Definisi
1	Daily Time Spent on Site	Total waktu (dalam menit) yang dihabiskan pengguna di situs per hari
2	Age	Usia pengguna dalam tahun
3	Area Income	Pendapatan rata-rata pengguna di area tempat tinggalnya
4	Daily Internet Usage	Total waktu (dalam menit) penggunaan internet oleh pengguna setiap hari.
5	Gender	Jenis kelamin pengguna
6	Timestamp	Waktu dan tanggal saat data dikumpulkan
7	Clicked on Ad	Menunjukkan apakah pengguna mengklik iklan atau tidak
8	City	Kota tempat tinggal pengguna
9	Province	Provinsi tempat tinggal pengguna
10	Category	Kategori iklan yang dilihat oleh pengguna

# Problem

Dalam dunia digital saat ini, perusahaan menghabiskan anggaran besar untuk iklan online tanpa mengetahui siapa saja yang benar-benar akan tertarik dan mengklik iklan tersebut. Tanpa pendekatan berbasis data, strategi pemasaran menjadi tidak efisien dan mahal.

Dengan memanfaatkan data historis iklan, dilakukan analisis untuk menggali pola dan insight dari perilaku pengguna. Dengan itu Data Scientist perlu mengidentifikasi pelanggan yang kemungkinan besar akan mengklik iklan, sehingga strategi pemasaran dapat menjadi lebih terarah dan efisien. Dalam iklan online, jumlah klik menunjukkan seberapa relevan iklan bagi pengguna. Meningkatkan CTR adalah cara efektif untuk menjaga pertumbuhan iklan online secara berkelanjutan. (*Jurnal Click-Through Rate Prediction in Online Advertising, Robinson et al., 2007; Rosales et al., 2012; Tan et al., 2020*)

# Problem

Perusahaan ingin memaksimalkan profit dari iklan online, namun biaya iklan dan jumlah pengguna yang ditargetkan sering kali tidak sebanding dengan hasil yang diperoleh. Salah satu kunci untuk mengoptimalkan pendapatan adalah memahami hubungan antara jumlah pengguna, conversion rate, dan potensi pendapatan.

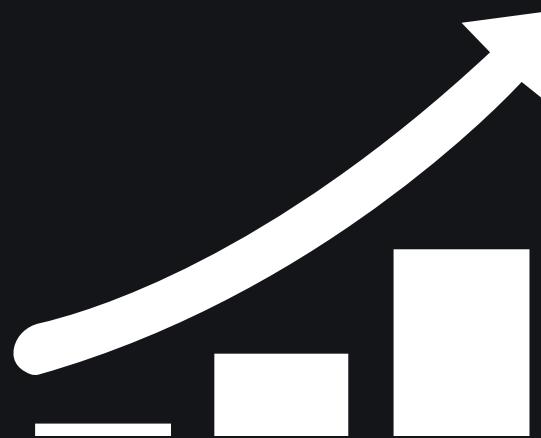
**Prediksi Klik Iklan = Jumlah User × Conversion Rate**

**Conversion Rate (%) = Jumlah Pengunjung / Jumlah Konversi × 100%**

**Revenue = Prediksi Klik Iklan × Revenue per user yang klik iklan**

Dengan mengetahui hubungan ini, perusahaan dapat menggunakan machine learning untuk menargetkan audiens yang lebih tepat, sehingga biaya iklan lebih efisien dan profit meningkat.

# Goals and Objectives



## Goals:

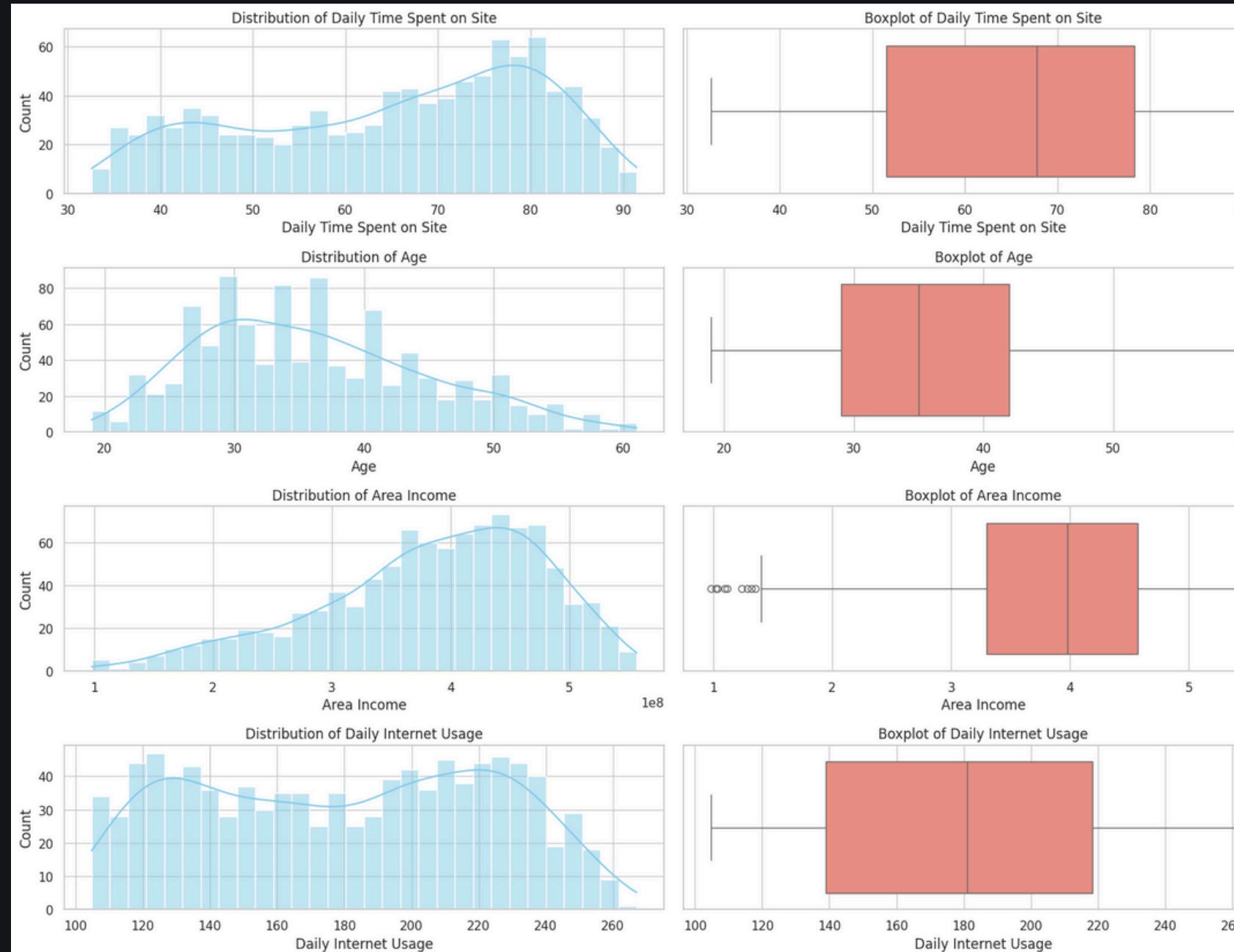
- Memprediksi variable apa saja yang mempengaruhi pengguna pada Clicked Ad
- Meningkatkan efektivitas kampanye pemasaran
- Mengoptimalkan anggaran iklan untuk hasil maksimal

## Objectives:

- Membangun model machine learning untuk memprediksi jumlah Clicked Ad dan yang tidak Clicked Ad
- Melakukan simulasi data untuk mengetahui strategi yang tepat bagi perusahaan

# Data Understanding

# Univariate Analysis



## Data Numerik

### 1. Daily Time Spent on Site

- Distribusi: Terdistribusi hampir normal dengan sedikit kemiringan ke kiri. Dengan rentang sekitar 30-90 menit. Pada boxplot ini tidak ada outlier signifikan.
- Mayoritas pengguna menghabiskan waktu antara 60–80 menit per hari di situs, menunjukkan engagement yang cukup tinggi.

### 2. Age

- Distribusi: Positively skewed. Dengan rentang sekitar 20-60 tahun. Pada boxplot ini tidak terlihat outlier ekstrem, usia mayoritas berada di sekitar 30–40 tahun.
- Mayoritas pengguna berada di usia produktif.

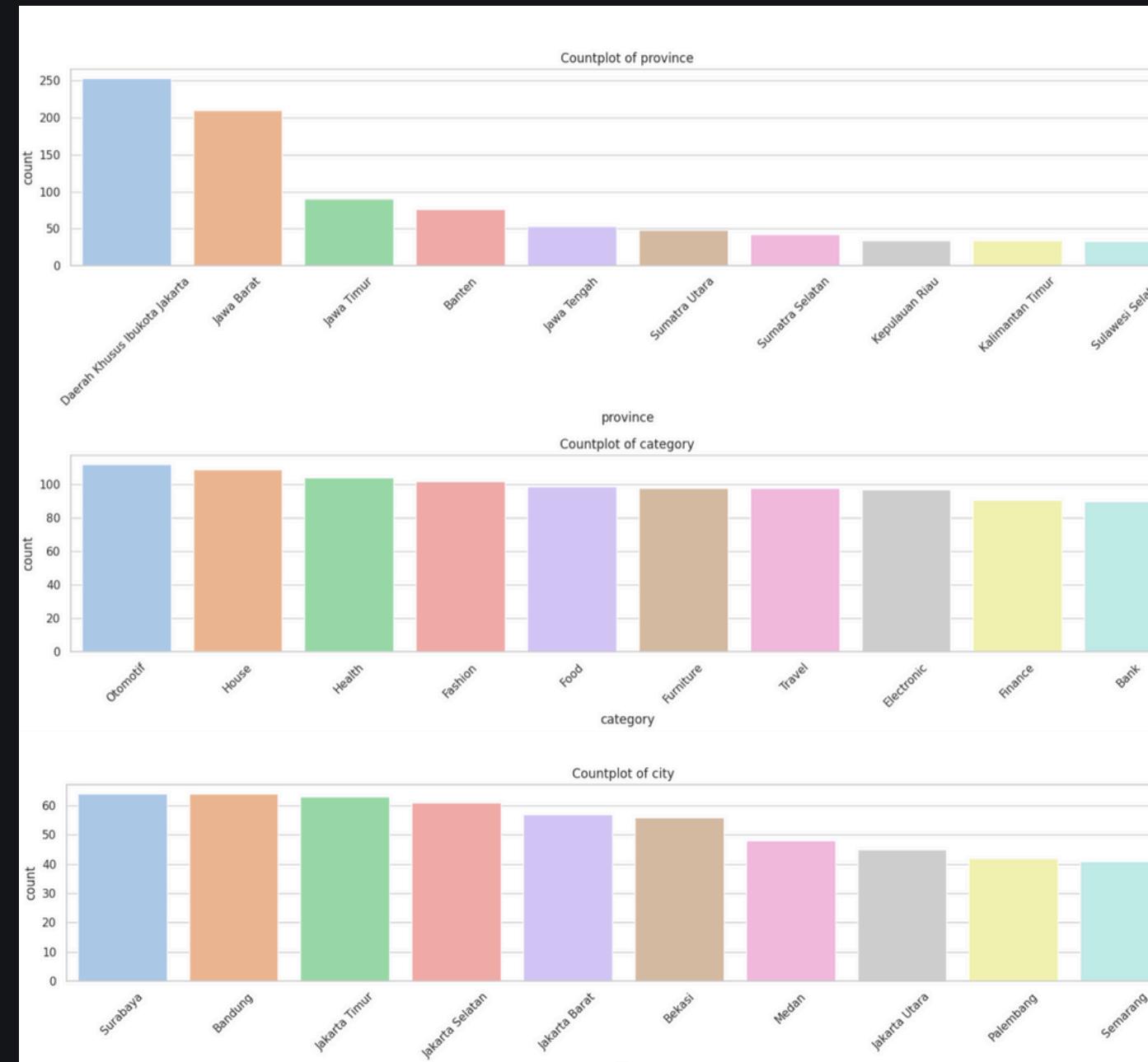
### 3. Area Income

- Distribusi: Negatively skewed. Dengan rentang sekitar 40-80 juta. Pada boxplot terdapat outlier di pendapatan tinggi.
- Pendapatan mayoritas berada di kisaran menengah, tetapi terdapat beberapa pengguna dari area dengan daya beli sangat tinggi.

### 4. Daily Internet Usage

- Distribusi: Agak merata namun dengan puncak di sekitar 200 menit. Dengan rentang sekitar 100-260 menit. Pada boxplot tidak ada outlier ekstrem.
- Mayoritas pengguna sangat aktif di internet (>150 menit/hari), ini membuka peluang besar untuk strategi digital marketing.

# Univariate Analysis

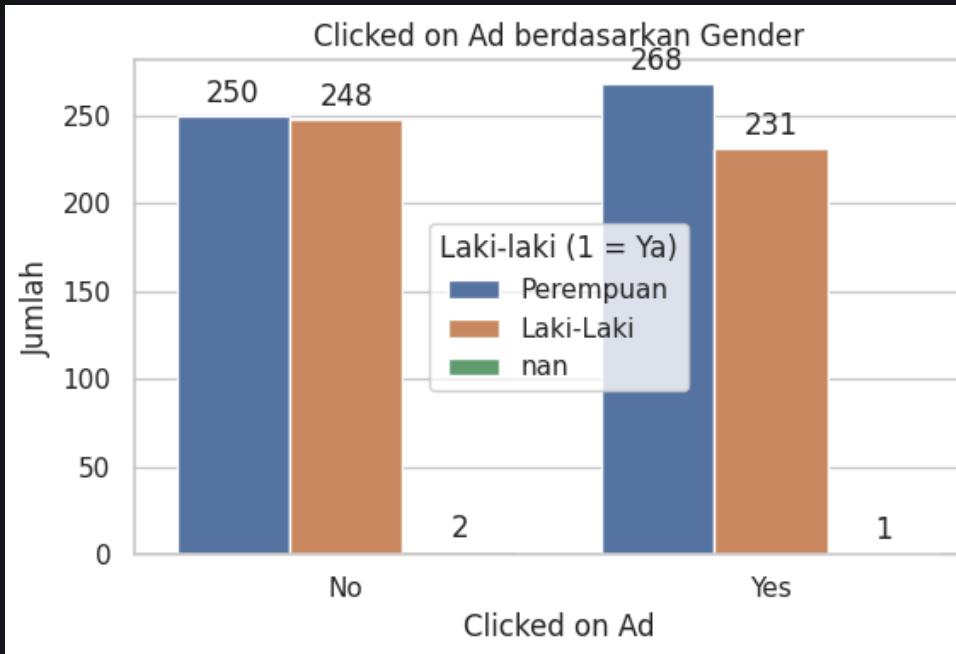


## Data Kategorik

- Province:** Didominasi oleh pengguna dari DKI Jakarta dan Jawa Barat. Dua provinsi ini merupakan wilayah dengan populasi dan aktivitas digital tinggi.
- Category:** Hampir merata antar kategori, namun kategori seperti Otomotif, House, dan Health memiliki sedikit lebih banyak representasi.
- City:** Kota-kota seperti Surabaya, Bandung, Jakarta Timur, dan Jakarta Selatan mendominasi. Wilayah ini memiliki tingkat aktivitas pengguna yang tinggi dapat dijadikan target utama dalam kampanye iklan.
- Gender:** Terlihat cukup seimbang antara perempuan dan laki-laki, dengan sedikit lebih banyak perempuan.



# Bivariate Analysis



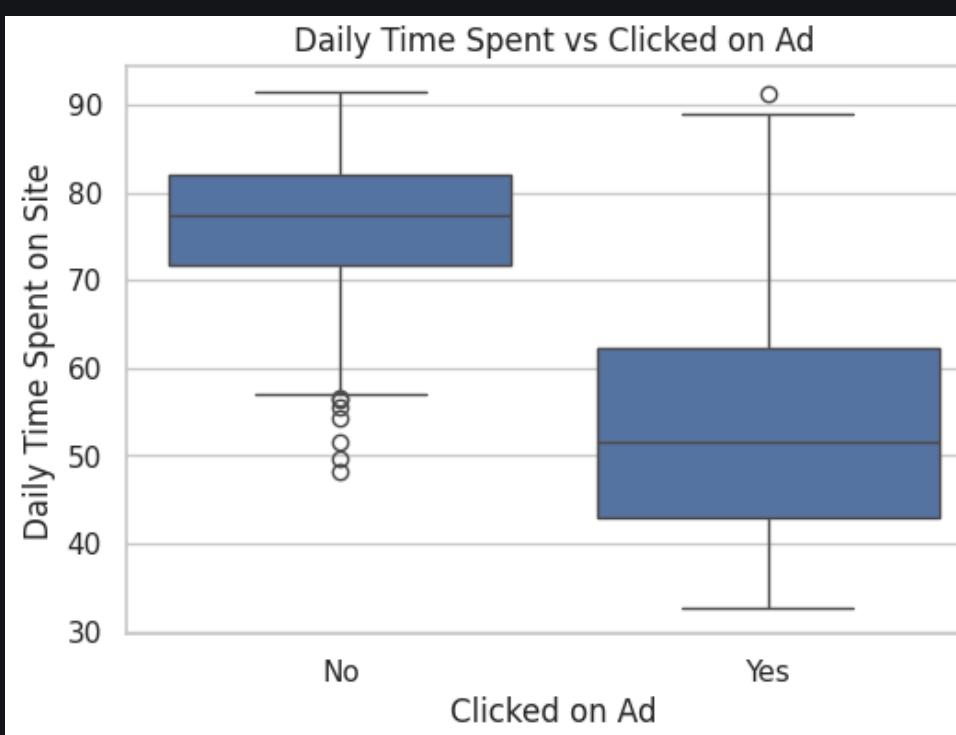
## 1. Clicked on Ad Berdasarkan Gender

- Perempuan lebih banyak melakukan klik pada iklan dibandingkan laki-laki: (268 perempuan klik iklan vs 231 laki-laki.)
- Padahal jumlah total keduanya tidak jauh berbeda (518 vs 479), ini menunjukkan bahwa perempuan cenderung lebih responsif terhadap iklan dalam data ini.

### Interpretasi:

- Segmentasi berdasarkan gender dapat membantu pengambilan keputusan iklan.

## 2. Daily Time Spent vs Clicked on Ad



### Interpretasi:

- Semakin lama seseorang berada di situs, semakin kecil kemungkinan mereka mengklik iklan karena pengguna dengan waktu lama lebih fokus pada konten utama situs.

# Multivariate Analysis



Fitur	Korelasi dengan Clicked on Ad	Interpretasi
Daily Internet Usage	<b>-0.79</b>	Korelasi negatif sangat kuat → semakin sering pengguna internet, semakin <b>tidak</b> klik iklan. Terbiasa dengan iklan dan mengabaikannya.
Daily Time Spent on Site	-0.74	Semakin lama waktu di situs, justru semakin <b>jarang</b> klik iklan. Bisa jadi hanya scrolling atau membaca.
Age	<b>0.49</b>	Korelasi positif sedang → semakin tua usianya, semakin <b>berpeluang</b> klik iklan. Strategi marketing bisa disesuaikan untuk segmen umur lebih tinggi.
Area Income	-0.47	Korelasi negatif sedang → pengguna dengan <b>pendapatan tinggi</b> cenderung <b>tidak</b> klik iklan. Mungkin mereka tidak tertarik dengan promosi.
Gender	-0.03	Korelasi sangat lemah → <b>gender tidak terlalu berpengaruh</b> terhadap klik iklan.
City / Province / Category	Sangat rendah ( $\pm 0.02$ )	Hampir tidak ada korelasi → lokasi dan kategori tidak terlalu menentukan.



# Data Preprocessing

# Data Preprocessing



- Data tidak ada duplicated. Tetapi terdapat missing value pada Daily Spent on Site, Area Income dan Daily Internet Usage dan Gender. Kemudian feature yang numerik diisi dengan mean dan yang kategorik diisi dengan mode sehingga missing value menjadi 0.
- Feature encoding dilakukan pada kolom kategorik dengan mengimport Label Encoder agar nantinya memudahkan proses modeling.
- Memilih fitur relevan: 'Daily Time Spent on Site', 'Age', 'Area Income', 'Daily Internet Usage', 'gender', 'city', 'province', 'category'
- Awalnya datetime memiliki tipe object lalu diubah menjadi tipe datetime. Selanjutnya feature ini diekstrak dengan pembagian year, month, day, hour dan minute.
- Menggunakan **StandardScaler** dari **scikit-learn**
- Tidak digunakan karena datasetnya sudah balance antara clicked on ads dan tidak.

# Data Modeling

Model	Normalisasi/ standardisasi	Accuracy	Precision	Recall	F1-Score
K-Nearest Neighbors	✗	0.63	0.63	0.63	0.63
Logistic Regression	✗	0.66	0.67	0.66	0.66
Decision Tree	✗	0.91	0.92	0.91	0.92
Random Forest	✗	0.94	0.95	0.94	0.94
K-Nearest Neighbors	✓	0.77	0.79	0.78	0.78
Logistic Regression	✓	0.95	0.96	0.96	0.95
Decision Tree	✓	0.91	0.92	0.91	0.92
Random Forest	✓	0.94	0.95	0.94	0.94

Pada tahap pemodelan, ada beberapa model machine learning yang dilakukan baik sebelum dan sesudah normalisasi/standardisasi. Dari keempat model ini, yang dipilih adalah K-Nearest Neighbors (KNN) dan Logistic Regression dibandingkan 2 model lainnya.

Karena, Model Decision Tree dan Random Forest tidak terpengaruh oleh skala fitur:

- Model ini bekerja dengan cara membagi data berdasarkan threshold pada setiap fitur, bukan berdasarkan perhitungan jarak.
- Hasil Decision Tree dan Random Forest tidak berubah sebelum dan sesudah normalisasi, karena kedua model tidak terpengaruh oleh skala fitur.

Model seperti k-Nearest Neighbors dan Logistic Regression sangat sensitif terhadap skala data, karena:

- KNN menghitung jarak antar titik data yakni skala fitur besar mendominasi hasil perhitungan.
- Logistic Regression menyesuaikan bobot fitur menggunakan proses bertahap sehingga hasil prediksinya lebih akurat.
- Hasil keduanya menunjukkan peningkatan performa setelah normalisasi.

Model	Normalisasi	Accuracy	Precision	Recall	F1-Score	Keterangan
K-Nearest Neighbors	✗	0.63	0.63	0.63	0.63	Performa rendah
Logistic Regression	✗	0.66	0.67	0.66	0.66	Lebih baik dari KNN, namun masih rendah.
K-Nearest Neighbors	✓	0.77	0.79	0.78	0.78	Performa meningkat drastis setelah normalisasi.
Logistic Regression	✓	0.95	0.96	0.96	0.95	Sangat tinggi, menunjukkan bahwa model sangat terbantu oleh scaling.

## Sebelum Normalisasi

KNN sangat bergantung pada jarak, sehingga performanya buruk tanpa normalisasi. Logistic Regression sedikit lebih stabil tanpa normalisasi, tapi tetap belum optimal.

## Sesudah Normalisasi

Normalisasi (scaling) membantu model memahami data secara lebih proporsional, terutama yang berbasis jarak (KNN) atau yang sensitif terhadap distribusi (LogReg). Hasil setelah normalisasi menunjukkan;

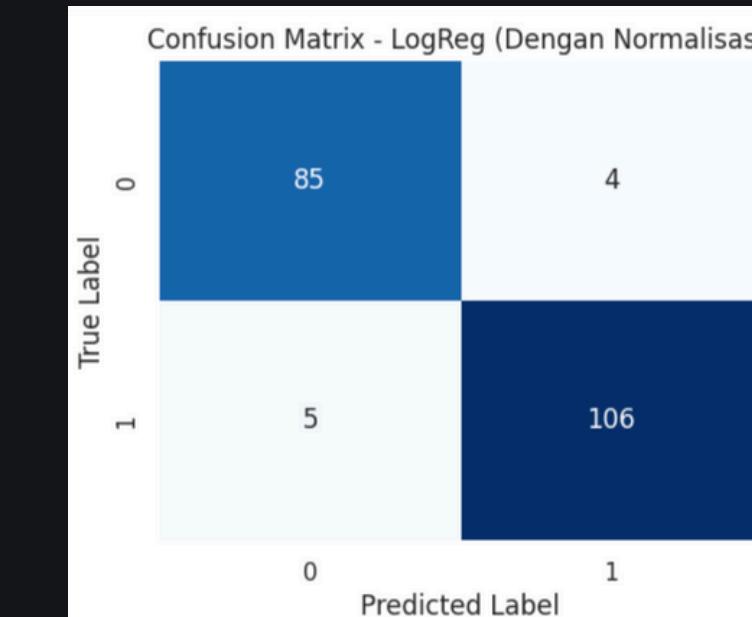
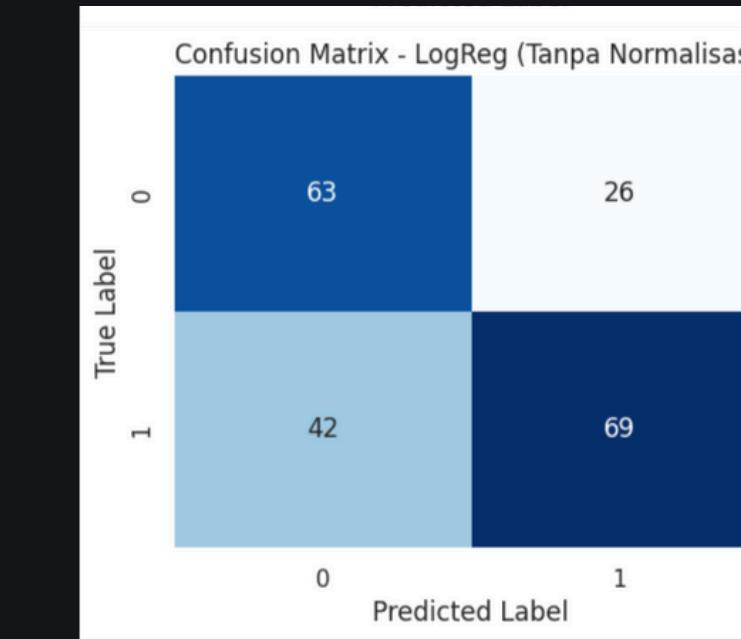
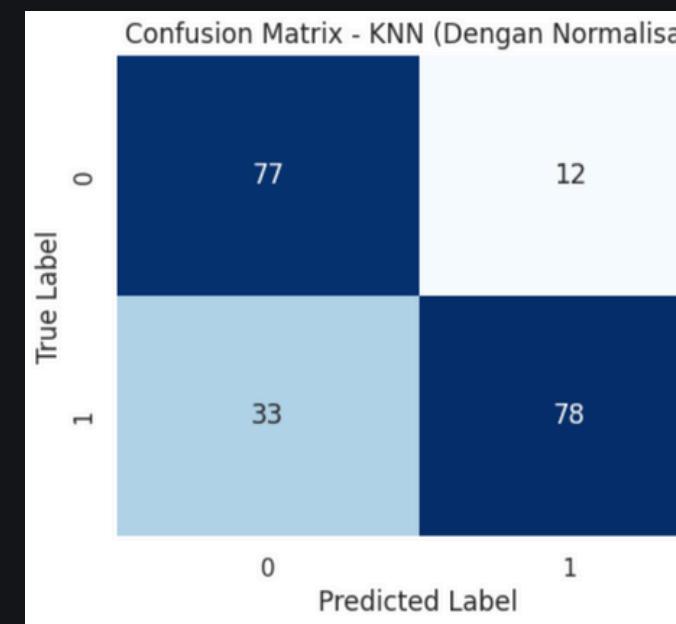
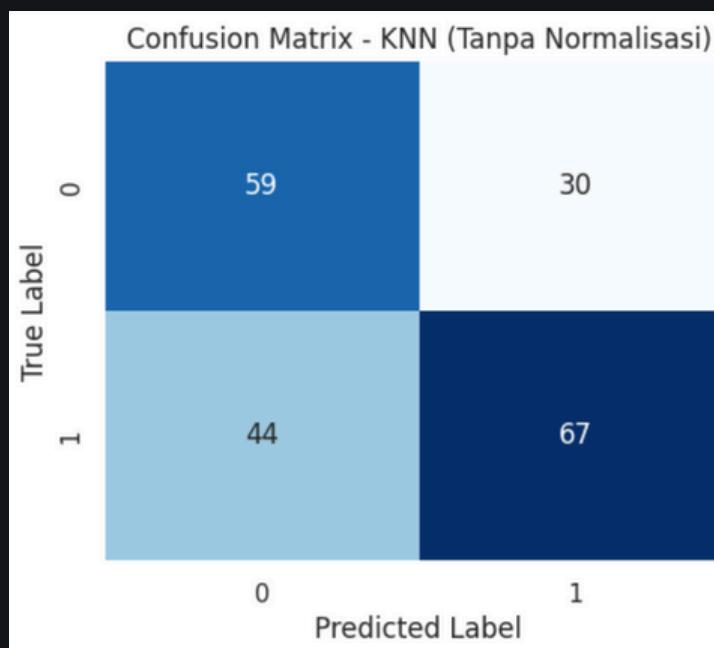
KNN mengalami peningkatan setelah normalisasi:

**Dari accuracy 0.63 → 0.77 dan f1-score 0.63 → 0.78**

Logistic Regression juga meningkat:

**Dari accuracy 0.66 → 0.95 dan f1-score 0.66 → 0.95**

# Confusion Matrix



Model	Normalisasi	True Negative (TN)	False Positive (FP)	False Negative (fN)	True Positive (TP)
KNN	✗	59	30	44	67
KNN	✓	77	12	33	78
LogReg	✗	63	26	42	69
LogReg	✓	85	4	5	106

Berdasarkan hasil evaluasi empat model klasifikasi, Logistic Regression dengan data yang telah dinormalisasi memberikan performa terbaik dengan akurasi tinggi dan jumlah kesalahan prediksi (false positive dan false negative) yang sangat rendah. Oleh karena itu, model ini dipilih sebagai model terbaik dalam memprediksi target pada dataset ini.

# Feature Importance

## 1. Fitur paling berpengaruh terhadap Clicked on Ad:

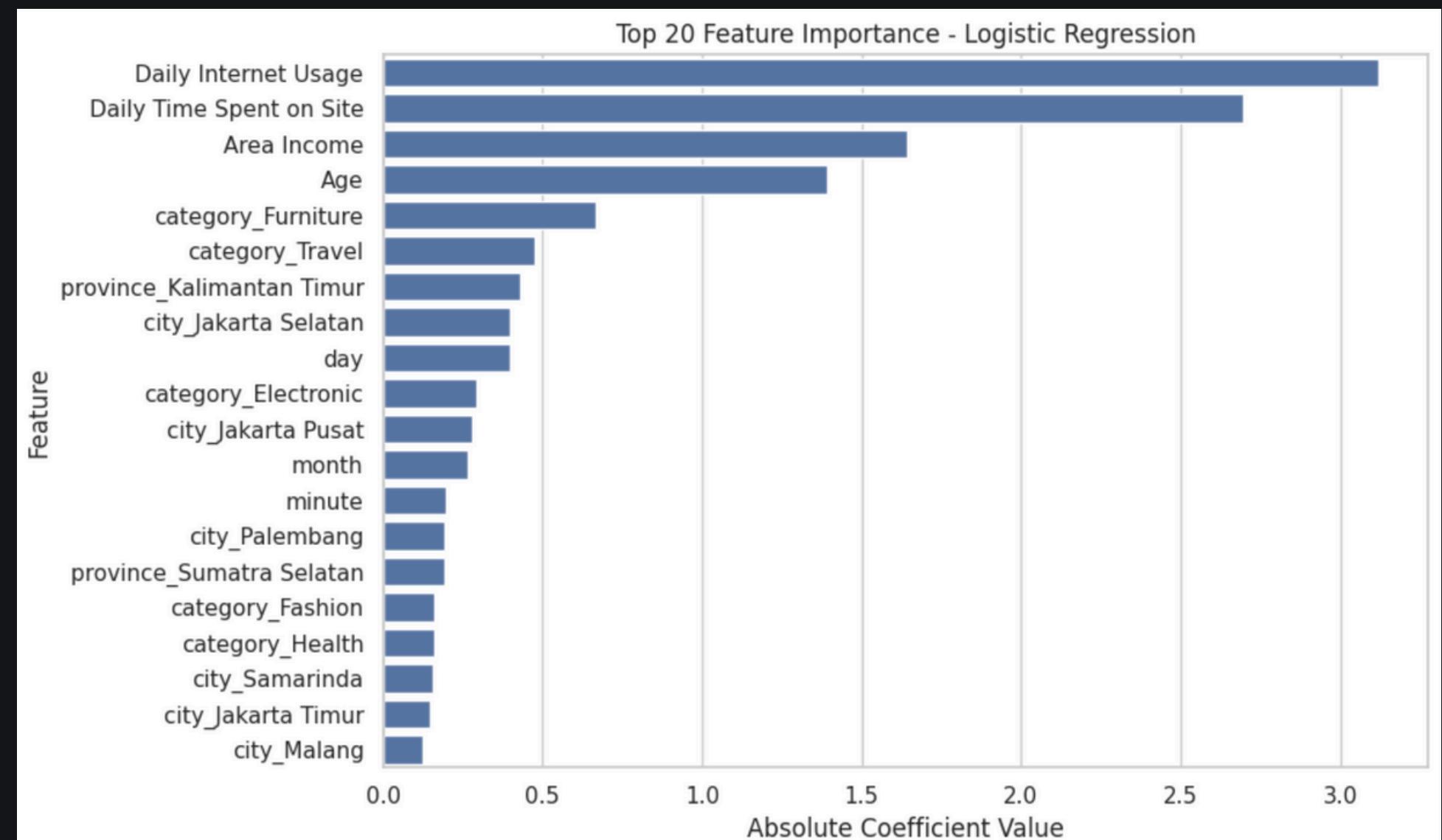
- **Daily Internet Usage** dan **Daily Time Spent on Site** adalah dua faktor terkuat: Semakin sering Pelanggan menggunakan internet dan berada di situs, makin besar kemungkinan mereka mengklik iklan.
- **Area Income** juga penting, namun arah pengaruhnya bisa positif atau negatif: Pengguna dengan penghasilan tertentu lebih responsif terhadap iklan.
- **Age** (Usia) ikut berperan: Ada kelompok usia tertentu yang lebih sering mengklik iklan (misalnya, orang dewasa muda yang lebih aktif online).

## 2. Kategori dan lokasi berpengaruh:

- **category\_Furniture**, **category\_Travel**, dan **category\_Electronic** cukup tinggi pengaruhnya: Pengguna yang melihat iklan pada kategori ini lebih tertarik/berpotensi untuk mengklik.
- Beberapa lokasi spesifik seperti **Jakarta Selatan**, **Jakarta Pusat**, dan **Samarinda** memiliki kontribusi: Ini menunjukkan asal kota pengguna dapat memengaruhi kecenderungan klik iklan, yang bisa jadi karena kebiasaan, kebutuhan lokal, atau tingkat digitalisasi daerah.

## 3. Waktu juga berpengaruh:

- **Hari (day)**, **bulan (month)**, **jam (minute/hour)** muncul meski pengaruhnya kecil. Menunjukkan waktu tertentu dalam hari atau bulan bisa memengaruhi performa iklan.



# Simulation

# Business Simulation



**Contoh Simulasi Marketing:**

Jumlah user = 1000

Biaya iklan per user = Rp 10.000

Revenue per user yang klik iklan = Rp 50.000

$$\text{Prediksi Klik Iklan} = \text{Jumlah User} \times \text{Conversion Rate}$$

Tanpa Machine Learning:

$$0,69 \times 1000 = 690 \text{ klik}$$

Dengan Machine Learning:

$$1,06 \times 1000 = 1060 \text{ klik}$$

Conversion Rate diambil dari TP LogReg setelah modeling karena ini memprediksi simulasi

$$\text{Revenue} = \text{Prediksi Klik Iklan} \times \text{Revenue per user yang klik iklan}$$

Tanpa Machine Learning:

$$\text{Revenue} = 690 \times \text{Rp } 50.000 = \text{Rp } 34.500.000$$

Dengan Machine Learning:

$$\text{Revenue} = 1060 \times \text{Rp } 50.000 = \text{Rp } 53.000.000$$

$$\text{Persentase Kenaikan} = \frac{\text{Profit dengan ML} - \text{Profit tanpa ML}}{\text{Profit tanpa ML}} \times 100\%$$

$$\text{Presentase Kenaikan} = (43.000.000 - 24.500.000) / 24.500.000 * 100\% = 75.5\%$$

Profit meningkat 75.5% dengan jumlah user & biaya iklan yang sama.

Model Logistic Regression membantu perusahaan menargetkan audience yang lebih tepat.

Simulasi	Jumlah User	Conversion Rate	Prediksi Klik Iklan	Cost (Rp)	Revenue (Rp)	Profit (Rp)
Tanpa Machine Learning	1000	69%	690	10.000.000	34.500.000	24.500.000
Dengan Machine Learning	1000	106%	1060	10.000.000	53.000.000	43.000.000

# Business Simulation



Parameter	Nilai
Jumlah User	1000
Biaya Iklan per User	Rp 10.000
Conversion Rate (ML)	75.5%
Prediksi Klik Iklan	1060
Total Cost	Rp 10.000.000
Revenue	Rp 53.000.000
Profit	Rp 43.000.000

## Mengurangi Pemborosan

Tanpa model, iklan disebar ke semua user tanpa mempertimbangkan siapa yang kemungkinan klik menyebabkan banyak biaya terbuang ke user yang tidak tertarik iklan.

Dengan model ML Logistic Regression hanya user dengan skor prediksi tinggi yang difokuskan, maka biaya iklan sama, tapi profit lebih besar.

## Optimasi Marketing Campaign

Data feature importance (Daily Internet Usage, Time Spent on Site, Area Income, Age) membantu menentukan siapa target ideal. Ini membuat strategi marketing bisa diarahkan ke user dengan perilaku online yang sesuai, lokasi potensial, dan kategori produk yang relevan.

## Profit Naik Tanpa Tambahan Biaya

Conversion rate tinggi menghasilkan profit lebih besar tanpa menambah jumlah user atau biaya iklan yang langsung berdampak positif bagi perusahaan.

# Business Recommendation

# Business Recomendation

## Segmentasi Perilaku Online (Daily Internet Usage & Daily Time Spent on Site)

Segmentasikan kampanye marketing hanya ke user dengan rentang penggunaan internet & waktu di web yang sesuai, bukan semua user.

## Penyesuaian Konten Berdasarkan Pendapatan (Area Income)

Sesuaikan konten iklan berdasarkan income segment yaitu premium vs diskon/promo.

## Target Demografi Usia (Age)

Membuat konten iklan & copywriting yang spesifik ke kelompok usia dengan probabilitas klik tinggi.

## Kategori Produk

Alokasikan lebih banyak anggaran ke kategori produk dengan feature importance tinggi

## Geo-Targeting

Fokus kampanye di kota dengan kontribusi klik tinggi (Jakarta Selatan, Jakarta Pusat, Samarinda).

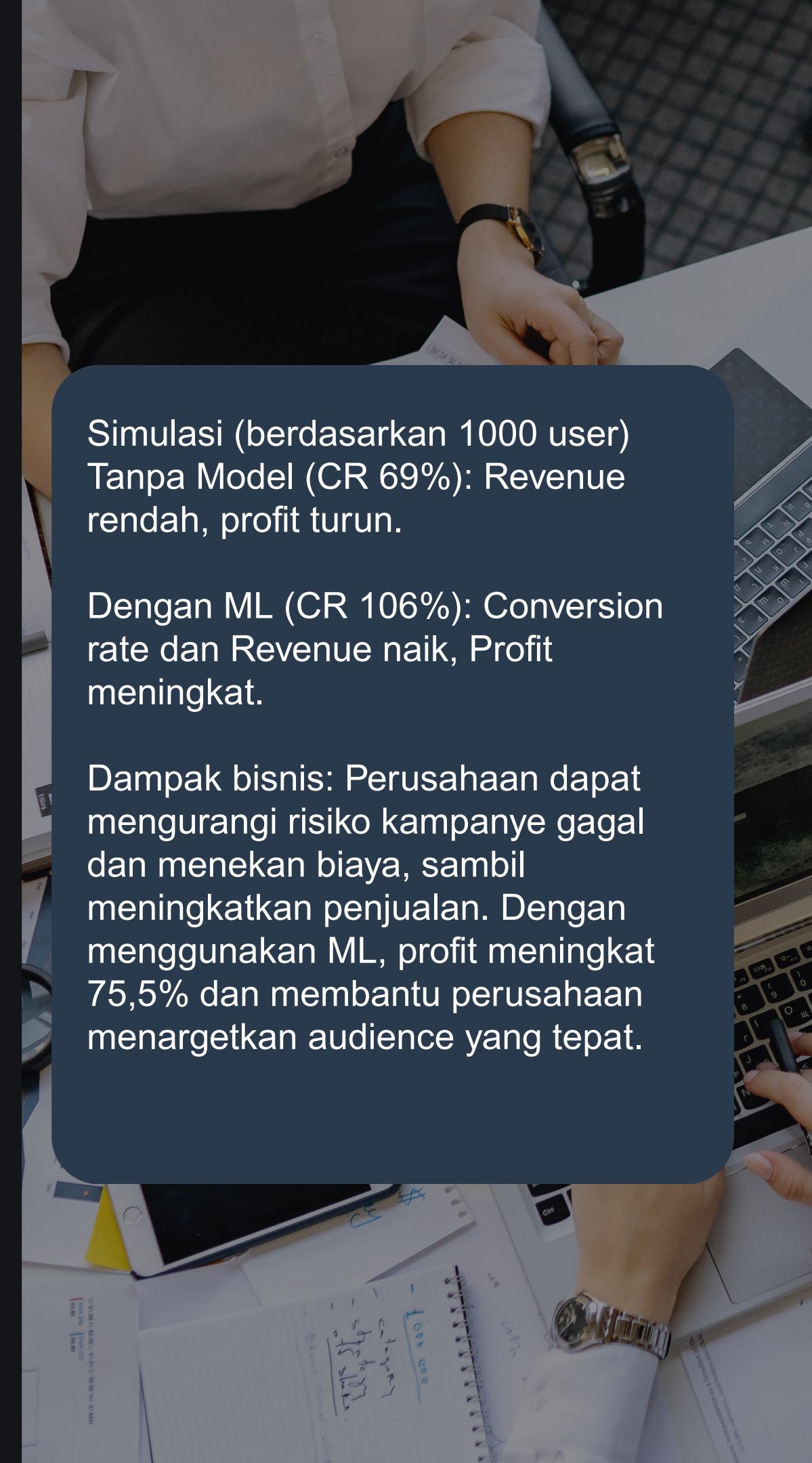
## Optimasi Waktu Tayang

Optimalkan jadwal iklan berdasarkan day, month, minute untuk CTR tertinggi.

Simulasi (berdasarkan 1000 user)  
Tanpa Model (CR 69%): Revenue rendah, profit turun.

Dengan ML (CR 106%): Conversion rate dan Revenue naik, Profit meningkat.

Dampak bisnis: Perusahaan dapat mengurangi risiko kampanye gagal dan menekan biaya, sambil meningkatkan penjualan. Dengan menggunakan ML, profit meningkat 75,5% dan membantu perusahaan menargetkan audience yang tepat.



# thank you