Predict Clicked Ads Customer Classification by using Machine Learning

Created by:
Muhammad Niko Dwi Putranto
nikoputranto@gmail.com
https://www.linkedin.com/in/niko-putranto/

"I am a fresh graduate from IPB University with a degree in Business and a deep passion for Data Analysis. Through my studies at IPB and my Data Science learning journey at Rakamin, I have gained a strong foundation in business strategy, data-driven decision-making, and analytical tools. I am eager to apply my skills in a dynamic and innovative environment, contributing to meaningful projects that drive organizational success and create a positive impact. I am excited to collaborate with forward-thinking teams and continue learning in the ever-evolving field of data analytics "

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Overview

“Sebuah perusahaan di Indonesia ingin mengetahui efektifitas sebuah iklan yang mereka tayangkan, hal ini penting bagi perusahaan agar dapat mengetahui seberapa besar ketercapainnya iklan yang dipasarkan sehingga dapat menarik customers untuk melihat iklan.

Dengan mengolah data historical advertisement serta menemukan insight serta pola yang terjadi, maka dapat membantu perusahaan dalam menentukan target marketing, fokus case ini adalah membuat model machine learning classification yang berfungsi menentukan target customers yang tepat ”

Customer Type and Behaviour Analysis on Advertisement

- Tulislah proses ***Exploration Data Analysis*** (EDA) yang mencakup ***Statistical analysis*** baik untuk data numerik maupun kategori, Selanjutnya buat visualisasi data untuk ***Univariate*** dan ***Bivariate analysis***, serta ***Multivariate analysis***

- Khusus untuk ***Bivariate analysis***, tunjukan hubungan antara kolom umur, daily internet usage, dan daily time spent on site.

- Tulislah juga **proses korelasi heatmap** untuk mengetahui tingkat korelasi antar kolom

- **Source code** yang sudah kamu buat, dapat ditampilkan dan berikan link untuk mengakses file tersebut. Contohnya seperti di pojok kanan bawah.

Untuk selengkapnya, dapat melihat notebook disini

Data Cleaning & Preprocessing

```
[29] data.duplicated().sum()
0
```

```
data.isnull().sum()
```

	0
Unnamed: 0	0
Daily Time Spent	13
Age	0
Area Income	13
DailyInternetUsage	11
Gender	3
Timestamp	0
ClickedOnAd	0
city	0
province	0
category	0

```
data['DailyTimeSpent'].fillna(data['DailyTimeSpent'].mean(), inplace=True)
data['DailyInternetUsage'].fillna(data['DailyInternetUsage'].mean(), inplace=True)
```

```
[13] data['Gender'].fillna(data['Gender'].mode()[0], inplace=True)
```

```
data.dropna(inplace=True)
```

- Pada tahap ***cleaning data***, ***null*** atau ***missing value*** serta ***duplicated value*** pada dataset, tidak ditemukan duplicated data pada dataset, terdapat 4 kolom dengan nilai null untuk penyelesaiannya pada kolom 'DailyTimeSpent' dan 'DailyInternetUsage' menggunakan rata-rata sedangkan pada kolom 'Gender' menggunakan modus serta untuk 'Area Income' karena tidak diketahui datanya dan tidak terlalu banyak data yang hilang dapat dihapus saja.

Untuk selengkapnya, dapat melihat jupyter notebook disini

```
data['Timestamp'] = pd.to_datetime(data['Timestamp'], errors='coerce')

# Extract year, month, week, and day into new columns
data['Year'] = data['Timestamp'].dt.year
data['Month'] = data['Timestamp'].dt.month
data['Week'] = data['Timestamp'].dt.isocalendar().week
data['Day'] = data['Timestamp'].dt.day

# Display the first few rows to verify the extraction
data[['Timestamp', 'Year', 'Month', 'Week', 'Day']].head()
```

- Kolom Timestamp dilakukan ekstraksi data menjadi tahun, bulan, minggu, dan hari seperti yang dapat dilihat pada gambar disamping.

```
target_column = 'ClickedOnAd'

# Split data
X = data.drop(columns=[target_column])
y = data[target_column]

# Display the first few rows of X and y to confirm split
X.head(), y.head()
```

- Split data dilakukan pada target x dan target y (‘ClickedOnAd’)

- Proses feature encoding dibagi menjadi dua yaitu label encoding dan one hot encoding. Label encoding dilakukan pada fitur Gender dan ClickOnAd karena terdapat dua kategori, sedangkan one hot encoding dilakukan pada fitur city, province dan category.

```
from sklearn.preprocessing import LabelEncoder

label_encoder = LabelEncoder()
data['Gender'] = label_encoder.fit_transform(data['Gender'])
data['ClickedOnAd'] = label_encoder.fit_transform(data['ClickedOnAd'])

# Apply One-Hot Encoding to columns with multiple categories
data = pd.get_dummies(data, columns=['city', 'province', 'category'], drop_first=True)
```

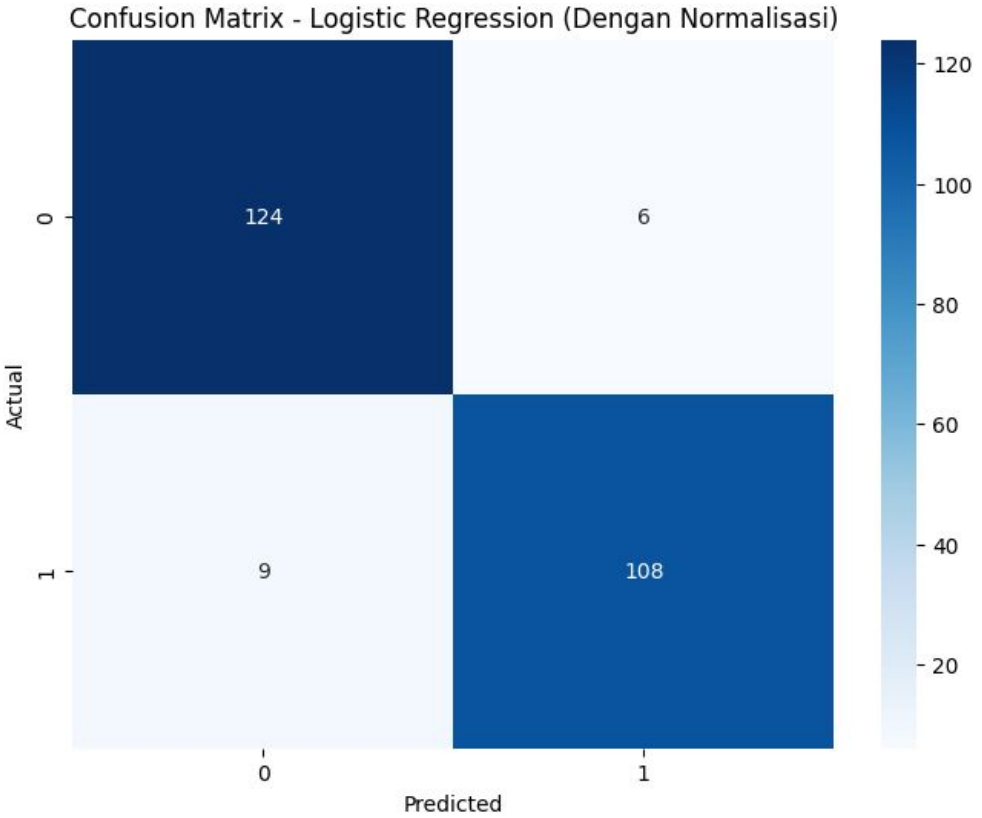
Tanpa Normalisasi

```
Logistic Regression without normalization
Accuracy: 0.8906882591093117
              precision    recall  f1-score   support

           0       0.90      0.89      0.90       130
           1       0.88      0.89      0.89       117

    accuracy                           0.89       247
   macro avg       0.89      0.89      0.89       247
weighted avg       0.89      0.89      0.89       247


K-Nearest Neighbors without normalization
Accuracy: 0.6639676113360324
              precision    recall  f1-score   support

           0       0.66      0.74      0.70       130
           1       0.67      0.58      0.62       117

    accuracy                           0.66       247
   macro avg       0.66      0.66      0.66       247
weighted avg       0.66      0.66      0.66       247
```

Dengan Normalisasi

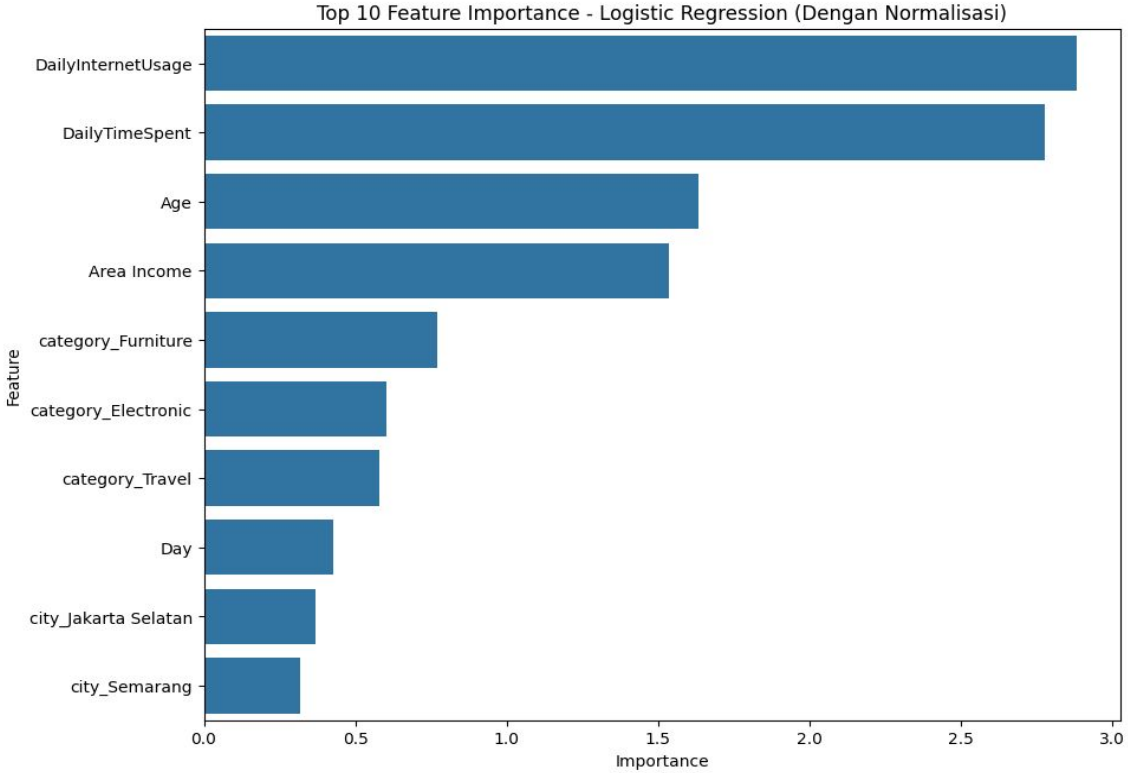
```
Logistic Regression with normalization
Accuracy: 0.9392712550607287
              precision    recall  f1-score   support

           0       0.93      0.95      0.94       130
           1       0.95      0.92      0.94       117

    accuracy                           0.94       247
   macro avg       0.94      0.94      0.94       247
weighted avg       0.94      0.94      0.94       247


K-Nearest Neighbors with normalization
Accuracy: 0.7894736842105263
              precision    recall  f1-score   support

           0       0.79      0.82      0.80       130
           1       0.79      0.75      0.77       117

    accuracy                           0.79       247
   macro avg       0.79      0.79      0.79       247
weighted avg       0.79      0.79      0.79       247
```

Hasil dari eksperimen pertama tanpa normalisasi memberikan performa yang lebih buruk dari eksperimen menggunakan normalisasi. Eksperimen dilakukan dengan dua model yaitu logistic regression dan KNN. Keduanya menunjukkan hasil yang lebih baik ketika data dilakukan normalisasi.



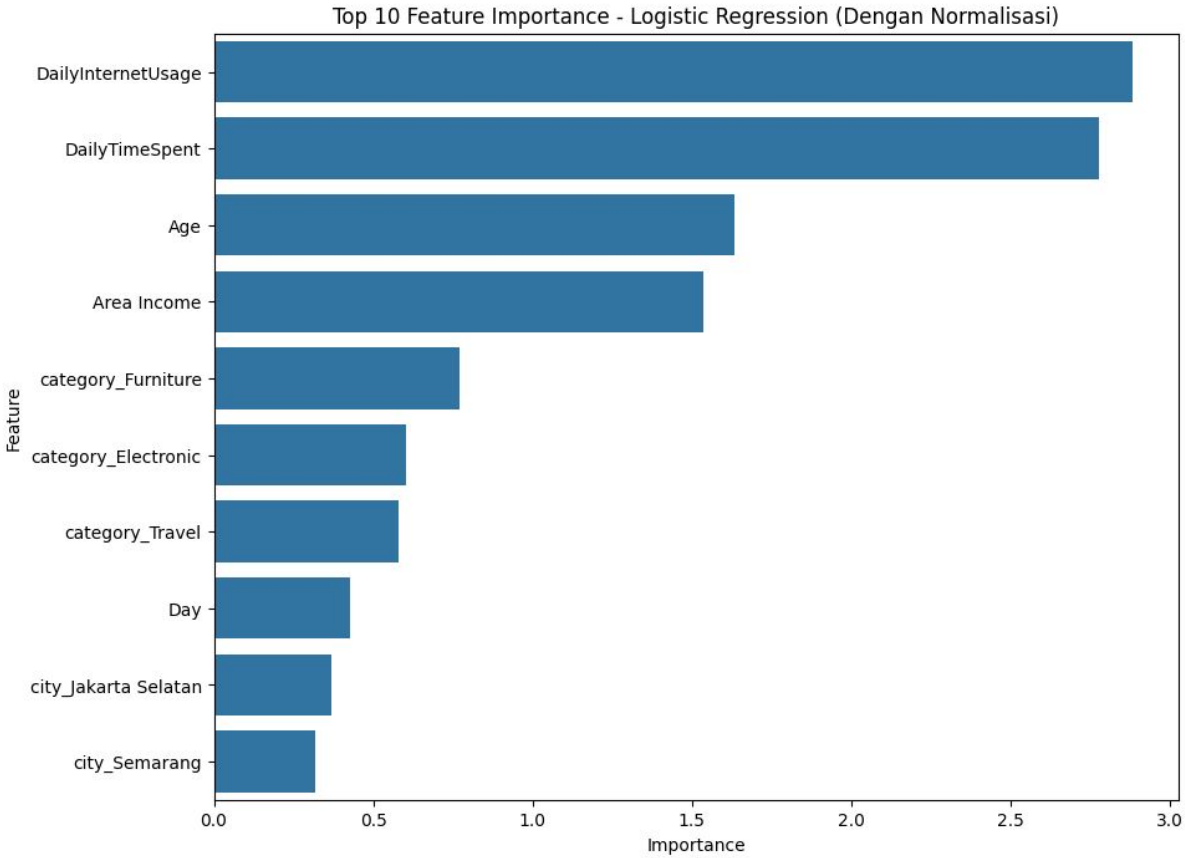
Hasil dari confusion matrix disamping menunjukkan bahwa True Negative sebanyak 124 artinya sebanyak 124 pengguna yang sebenarnya tidak mengklik iklan (Clicked on Ad = No) berhasil diprediksi dengan benar oleh model. False Positives sebanyak 6 pengguna yang sebenarnya tidak mengklik iklan (Clicked on Ad = No) justru diprediksi akan mengklik iklan. False Negatives sebanyak 9 pengguna yang sebenarnya mengklik iklan (Clicked on Ad = Yes) gagal diprediksi oleh model. True Positives sebanyak 108 pengguna yang benar-benar mengklik iklan (Clicked on Ad = Yes) berhasil diprediksi dengan benar oleh model.

Model ini cukup efektif dalam memprediksi perilaku klik iklan berdasarkan fitur seperti waktu di situs, usia, dan penggunaan internet.



Dari visualisasi disamping fitur yang memilki pengaruh terbesar dalam memprediksi adalah Daily Internet Usage, Daily Time Spent dan Age.

Feature Important



Rekomendasi Bisnis

1. Daily Internet Usage: Terlihat bahwa pengguna dengan penggunaan internet harian yang lebih rendah cenderung lebih sering mengklik iklan. Sementara itu, pengguna dengan penggunaan internet harian lebih tinggi sebagian besar tidak mengklik iklan.
   - Fokuskan iklan pada pengguna dengan penggunaan internet rendah hingga sedang (120-160 menit), yang masih menunjukkan kecenderungan lebih tinggi untuk mengklik iklan.
2. Daily Time Spent: Pengguna dengan waktu yang lebih sedikit di situs tampak lebih mungkin mengklik iklan dibandingkan mereka yang menghabiskan lebih banyak waktu di situs. Distribusi ini mengindikasikan bahwa pengguna yang hanya mengunjungi sebentar lebih rentan terhadap klik iklan.
   - Karena pengguna yang menghabiskan waktu singkat lebih rentan terhadap klik iklan, tempatkan iklan di area yang terlihat langsung, seperti bagian atas atau samping halaman utama.
   - Gunakan iklan yang lebih mencolok atau dengan konten langsung seperti diskon, penawaran cepat, atau produk populer.
3. Age: Grafik menunjukkan bahwa pengguna yang lebih tua cenderung lebih sering mengklik iklan dibandingkan pengguna yang lebih muda. Terdapat distribusi yang lebih padat di rentang usia tertentu untuk pengguna yang tidak mengklik iklan.
   - Gunakan iklan yang menyasar kebutuhan spesifik, seperti investasi, layanan kesehatan, atau produk rumah tangga.
   - Pengguna Muda (20-30 tahun): Gunakan strategi konten yang interaktif, seperti kuis, game, atau video pendek, untuk menarik perhatian mereka.

Simulasi Cost, Revenue, dan Profit Tanpa Model

Jumlah Pengguna Data Test : 247

Biaya Pemasaran : $10/user

Pendapatan rata-rata per konversi = $12

Total Cost = 247 * Biaya Pemasaran = 2470

Conversion Rate = TP + FN/Total Users = 117/247 = 47,36%

User Converted = jumlah pengguna*Conversion Rate = 117

Revenue = 117 * 12 = $1404

Profit = 1404 - 2470 = -$1066

Berdasarkan simulasi di atas tanpa machine learning maka akan mendapatkan potential loss sebesar $1066

Simulasi Cost, Revenue, dan Profit dengan Model

Target Potensial = 114

User Converted = 108

Biaya Pemasaran : $10/user

Pendapatan rata-rata per konversi = $12

Total Cost = 114 * Biaya Pemasaran = $1140

Conversion Rate = TP/Total Users = 108/114 = 94.73%

Revenue = 108 * 12 = $1296

Profit = 1296 - 1140 = $156

Berdasarkan simulasi di atas tanpa machine learning maka akan mendapatkan profit sebesar $156

Kesimpulan

Model Machine Learning dapat menangkap hasil yang lebih baik serta mendapatkan profit.