

# Predict Clicked Ads Customer Classification by using Machine Learning

Supported by:  
**Rakamin Academy**  
Career Acceleration School  
[www.rakamin.com](http://www.rakamin.com)



**Created by:**  
**Candraditya Dwaya Putra**

Email : [candradityaputra1@gmail.com](mailto:candradityaputra1@gmail.com)

linkedin:  
<https://www.linkedin.com/in/candraditya-dwaya-putra-77719a66>

“Have more than 11 years professional career in GIS and Management geodatabase. Good knowledge WebGIS Development, with strong experience Geospatial data in Forestry, environmental and regional planning. Skilled at data collection and analysis that elicits accurate and valuable information utilizing technical principles and theories. Technical proficiencies include SQL, Python, Machine Learning, MS Office, ESRI GIS, QGIS, HTML, CSS/Bootstrap, Javascript, geomorphology and Geospatial Software.”

# 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 ”

# Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype  
 ---  -- 
 0   Daily Time Spent on Site    987 non-null   float64
 1   Age                  1000 non-null   int64   
 2   Area Income          987 non-null   float64
 3   Daily Internet Usage  989 non-null   float64
 4   Male                 997 non-null   object  
 5   Timestamp            1000 non-null   object  
 6   Clicked on Ad        1000 non-null   object  
 7   city                 1000 non-null   object  
 8   province             1000 non-null   object  
 9   category              1000 non-null   object  
dtypes: float64(3), int64(1), object(6)
memory usage: 78.2+ KB
```

## Description

Dataset mengandung kebiasaan customer melihat iklan

## Shape

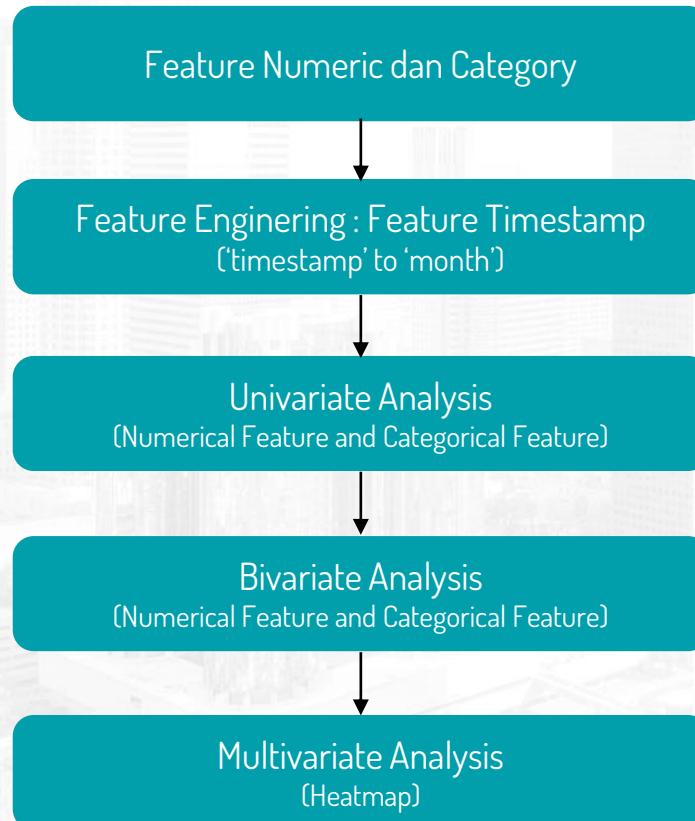
1000 Row, 10 Columns

## Dtype

Float64 (3 features), int64(1 features), Object (6 features)

## Missing Value

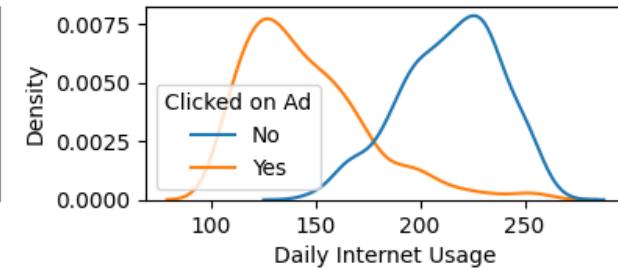
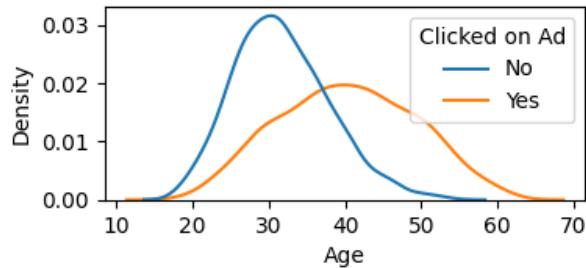
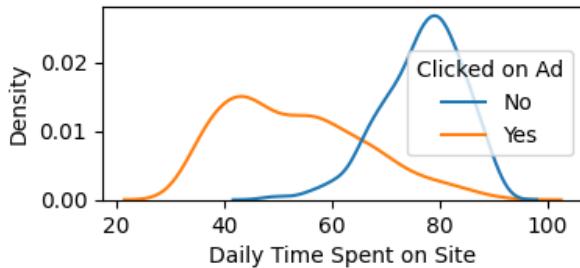
'Daily Time Spent on Site', 'Area Income', 'Daily Internet Usage', 'Male'



- Memisahkan feature numbering dan category
- Feature timestamp yang digunakan adalah format bulan ‘month’
- Selanjutnya dilakukan univariate analysis dan bivariate analysis pada feature numerical dan categorical
- Multivariate analysis

# Customer Type and Behaviour Analysis on Advertisement

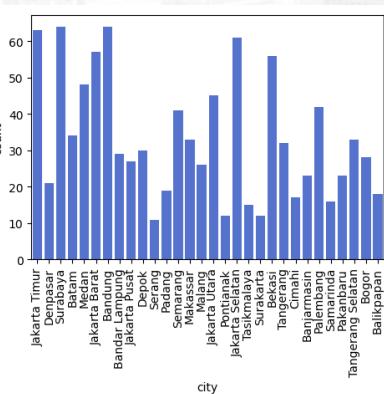
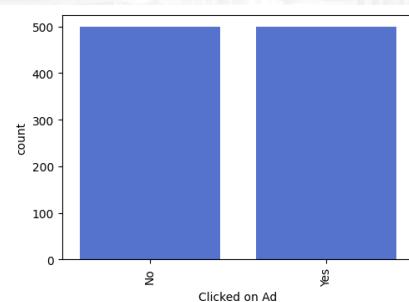
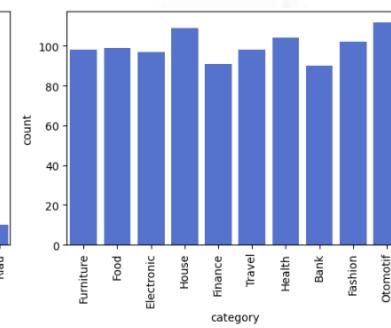
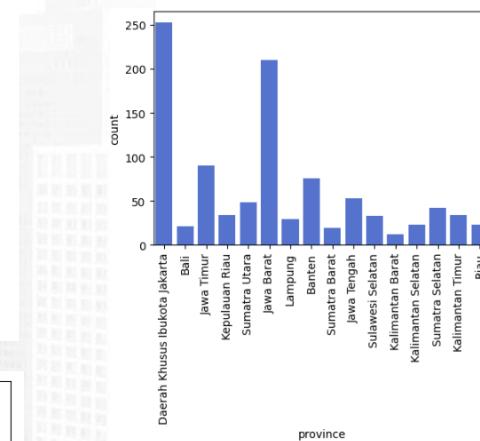
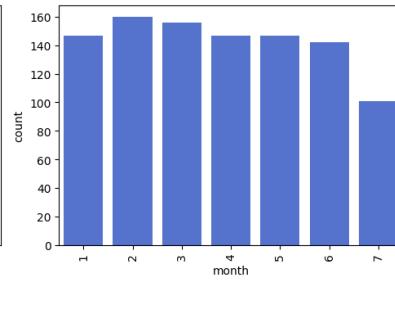
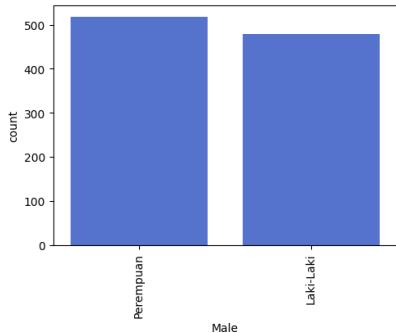
## Univariate Analysis (numeric)



- User yang mengklik Ads adalah user dengan 'Daily Time Spend on Site' sekitar 40-45 menit. Sedangkan, user yang tidak mengklik Ads adalah user dengan 'Daily Time Spend on Site' sekitar 75-80 menit.
- User yang mengklik Ads rata-rata ada pada usia(Age) 40 tahun. Sedangkan, user yang tidak mengklik Ads sebagian besar ada pada usia(Age) 30 tahun.
- User dengan 'Daily Internet Usage' sekitar 100-150 cenderung mengklik Ads. Sedangkan, user dengan 'Daily Internet Usage' sekitar 200-250 cenderung tidak mengklik Ads.

# Customer Type and Behaviour Analysis on Advertisement

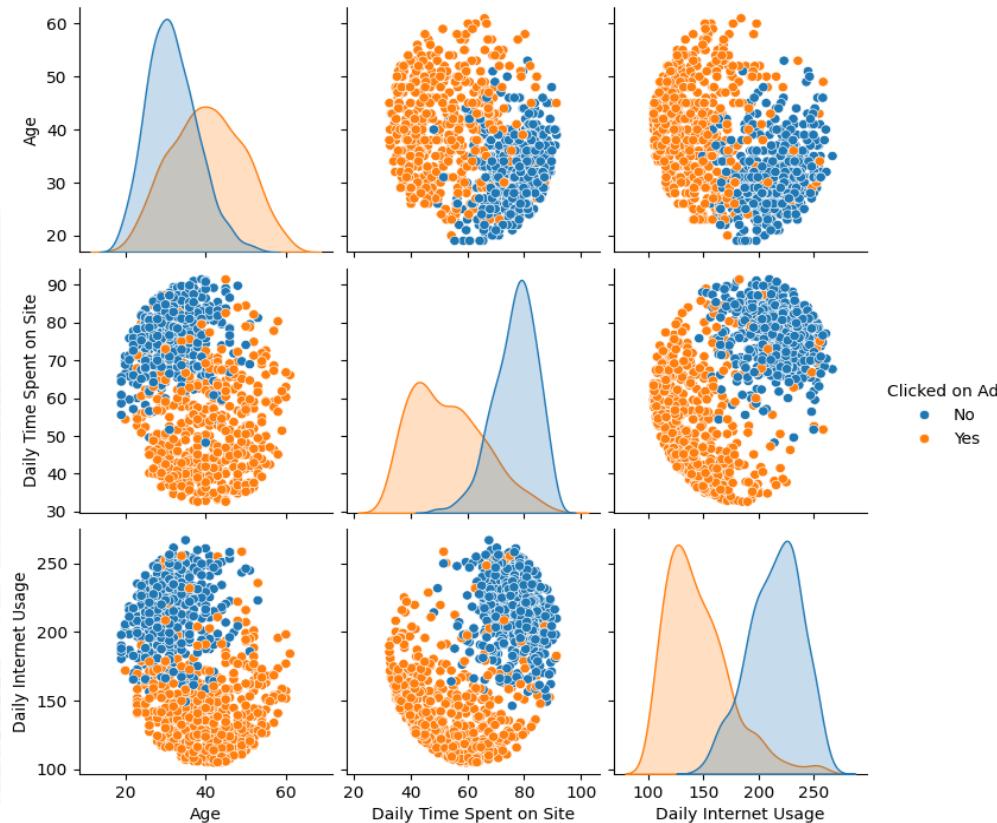
## Univariate Analysis (category)



- Tidak ada ketimpangan antara Label `Perempuan` dan `Laki-Laki` pada feature `Male`
- Iklan yang Diklik memiliki distribusi yang merata antara `Yes` dan `No` pada feature `Clicked on Ad`
- Feature province didominasi oleh 2 nilai yaitu **DKI Jakarta** dan **Jawa Barat**
- Categori hampir merata pada semua nilai

# Customer Type and Behaviour Analysis on Advertisement

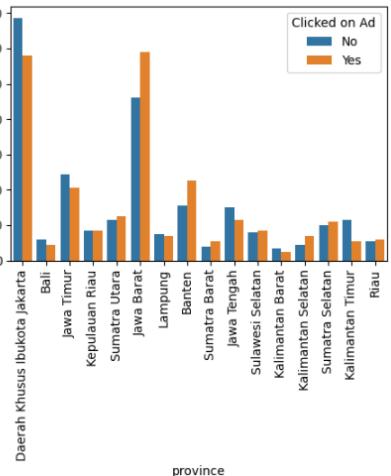
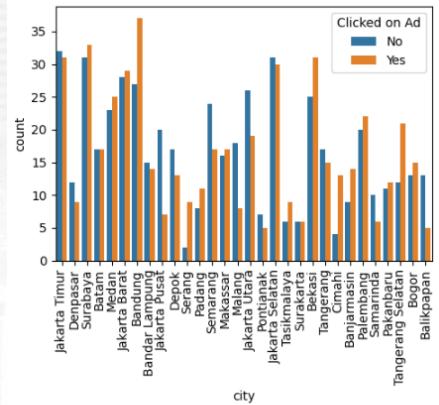
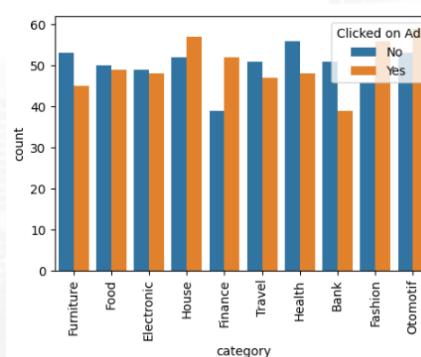
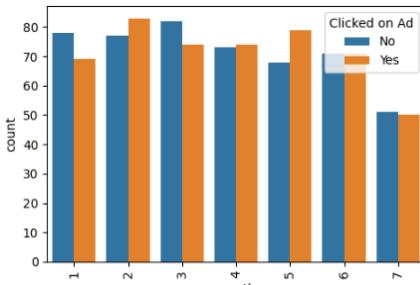
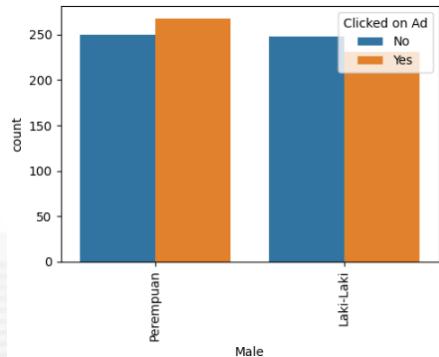
## Bivariate Analysis (Numeric)



- Semakin tua usia (`Age`) user serta semakin sedikit `Daily Internet Usage` dan `Daily Time Spent on Site` maka seorang user cenderung mengklik Ads.
- Semakin sedikit `Daily Internet Usage` dan `Daily Time Spent on Site` maka seorang user cenderung mengklik Ads.

# Customer Type and Behaviour Analysis on Advertisement

## Bivariate Analysis (Categoric)



- User perempuan lebih banyak mengklik Ads dibandingkan dengan user laki-laki berdasarkan feature male
- Setiap bulannya, perbandingan jumlah antara user yang mengklik atau tidak mengklik Ads hampir sebanding. Perbedaan yang cukup signifikan terdapat pada bulan ke 12, dimana jumlah user yang mengklik Ads 2 kali lebih banyak dibandingkan yang tidak mengklik Ads.
- 13 dari 30 kota yang ada pada feature city, Jumlah user yang mengklik Ad lebih banyak dibandingkan dengan yang tidak mengklik Ads.
- Pada feature province, user yang berasal dari Jawa Barat mengklik Ads lebih banyak dibandingkan dengan yang tidak mengklik Ads.
- Pada feature category,user lebih banyak mengklik Ads dengan category House, Finance, Fashion dan Otomotif jika dibandingkan dengan user yang tidak mengklik Ads.

# Customer Type and Behaviour Analysis on Advertisement

## Multivariate Analysis



- Feature `Daily Time Spent on Site` berkorelasi positif cukup kuat dengan `Daily Internet Usage`
- Feature `Age` berkorelasi negatif lemah dengan feature `Daily Time Spent on Site`, `Area Income`, dan `Daily Internet Usage`
- Feature `Area Income` berkorelasi positif dengan feature `Daily Time Spent on Site` dan `Daily Internet Usage` dan berkorelasi negatif dengan feature `Age`

# Data Cleaning & Preprocessing

## Missing Value dan Duplicate

```
Daily Time Spent on Site    13
Age                           0
Area Income                   13
Daily Internet Usage         11
Male                          3
Timestamp                     0
Clicked on Ad                0
city                          0
province                      0
category                      0
dtype: int64
```

- Pada feature numerik nilai yang kosong diisi dengan nilai median dari masing-masing feature. Penggunaan nilai median sebagai nilai yang diinput untuk nilai yang kosong karena distribusi data cenderung skewed.
- Sedangkan, Pada feature kategorik nilai kosong diisi dengan modus dari feature tersebut.

```
→ <ipython-input-39-45d31d8b8ea3>:2: Future
      df1.fillna(df1.median(), inplace=True)
      Daily Time Spent on Site    0
      Age                         0
      Area Income                 0
      Daily Internet Usage       0
      Male                        0
      Timestamp                   0
      Clicked on Ad              0
      city                        0
      province                     0
      category                     0
      dtype: int64
```

- Hasil dari penyelesaian
- Tidak ada Duplicate

# Data Cleaning & Preprocessing

## Proses Extract date time

### Feature Engineering : Feature Timestamp

Ekstraksi pada kolom yang berhubungan dengan waktu  
(mengekstraksi data waktu menjadi tahun, bulan, pekan, dan hari)

### Mengubah Data To Timestamp

### Menambah kolom baru

Year = Tahun

Month = Bulan

Week = Minggu

Day = Hari

	Daily Time Spent on Site		Age	Area Income	Daily Internet Usage	Male	Timestamp	Clicked on Ad	city	province	category	year	month	week	day
0	68.95		35	432837300.0	256.09	Perempuan	2016-03-27 00:53:00	No	Jakarta Timur	Daerah Khusus Ibukota Jakarta	Furniture	2016	3	12	27
1	80.23		31	479092950.0	193.77	Laki-Laki	2016-04-04 01:39:00	No	Denpasar	Bali	Food	2016	4	14	4
2	69.47		26	418501580.0	236.50	Perempuan	2016-03-13 20:35:00	No	Surabaya	Jawa Timur	Electronic	2016	3	10	13
3	74.15		29	383643260.0	245.89	Laki-Laki	2016-01-10 02:31:00	No	Batam	Kepulauan Riau	House	2016	1	1	10
4	68.37		35	517229930.0	225.58	Perempuan	2016-06-03 03:36:00	No	Medan	Sumatra Utara	Finance	2016	6	22	3

# Data Cleaning & Preprocessing

## Before Split data

```
[62] df.shape
```

```
(1000, 37)
```

```
[63] df.columns
```

```
Index(['Daily Time Spent on Site', 'Age', 'Area Income',
       'Daily Internet Usage', 'Male', 'Clicked on Ad', 'city', 'province',
       'category', 'month', 'week', 'province_Bali', 'province_Banten',
       'province_DKI Jakarta', 'province_Jawa Barat', 'province_Jawa Tengah',
       'province_Jawa Timur', 'province_Kalimantan Barat',
       'province_Kalimantan Selatan', 'province_Kalimantan Timur',
       'province_Kepulauan Riau', 'province_Lampung', 'province_Riau',
       'province_Sulawesi Selatan', 'province_Sumatra Barat',
       'province_Sumatra Selatan', 'province_Sumatra Utara', 'category_Bank',
       'category_Electronic', 'category_Fashion', 'category_Finance',
       'category_Food', 'category_Furniture', 'category_Health',
       'category_House', 'category_Otomotif', 'category_Travel'],
      dtype='object')
```

# Data Cleaning & Preprocessing

## Feature Encoding

**Label Jenis Kelamin**  
Male = 1, Perempuan = 0

**Label Click on add**  
Yes = 1, No = 0

**One Hot Encoding (OHE)**  
Province

# Data Modeling Experiment 1

- Logistic Regression

```
Accuracy (Train Set): 0.51
Accuracy (Test Set): 0.49
Precision (Train Set): 0.00
Precision (Test Set): 0.00
Recall (Train Set): 0.00
Recall (Test Set): 0.00
F1-Score (Train Set): 0.00
F1-Score (Test Set): 0.00
AUC (Train Set): 0.50
AUC (Test Set): 0.50
```

- Decission Tree

```
Accuracy (Train Set): 1.00
Accuracy (Test Set): 0.94
Precision (Train Set): 1.00
Precision (Test Set): 0.96
Recall (Train Set): 1.00
Recall (Test Set): 0.92
F1-Score (Train Set): 1.00
F1-Score (Test Set): 0.94
AUC (Train Set): 1.00
AUC (Test Set): 0.94
```

- Random Forests

```
Accuracy (Train Set): 1.00
Accuracy (Test Set): 0.95
Precision (Train Set): 1.00
Precision (Test Set): 0.99
Recall (Train Set): 1.00
Recall (Test Set): 0.92
F1-Score (Train Set): 1.00
F1-Score (Test Set): 0.95
AUC (Train Set): 1.00
AUC (Test Set): 0.95
```

# Data Modeling Experiment 1

- Logistic Regression

```
Accuracy (Train Set): 0.51
Accuracy (Test Set): 0.49
Precision (Train Set): 0.00
Precision (Test Set): 0.00
Recall (Train Set): 0.00
Recall (Test Set): 0.00
F1-Score (Train Set): 0.00
F1-Score (Test Set): 0.00
AUC (Train Set): 0.50
AUC (Test Set): 0.50
```

- Decission Tree

```
Accuracy (Train Set): 1.00
Accuracy (Test Set): 0.94
Precision (Train Set): 1.00
Precision (Test Set): 0.96
Recall (Train Set): 1.00
Recall (Test Set): 0.92
F1-Score (Train Set): 1.00
F1-Score (Test Set): 0.94
AUC (Train Set): 1.00
AUC (Test Set): 0.94
```

- Random Forests

```
Accuracy (Train Set): 1.00
Accuracy (Test Set): 0.95
Precision (Train Set): 1.00
Precision (Test Set): 0.99
Recall (Train Set): 1.00
Recall (Test Set): 0.92
F1-Score (Train Set): 1.00
F1-Score (Test Set): 0.95
AUC (Train Set): 1.00
AUC (Test Set): 0.95
```

# Data Modeling

## Experiment 2

- Logistic Regression

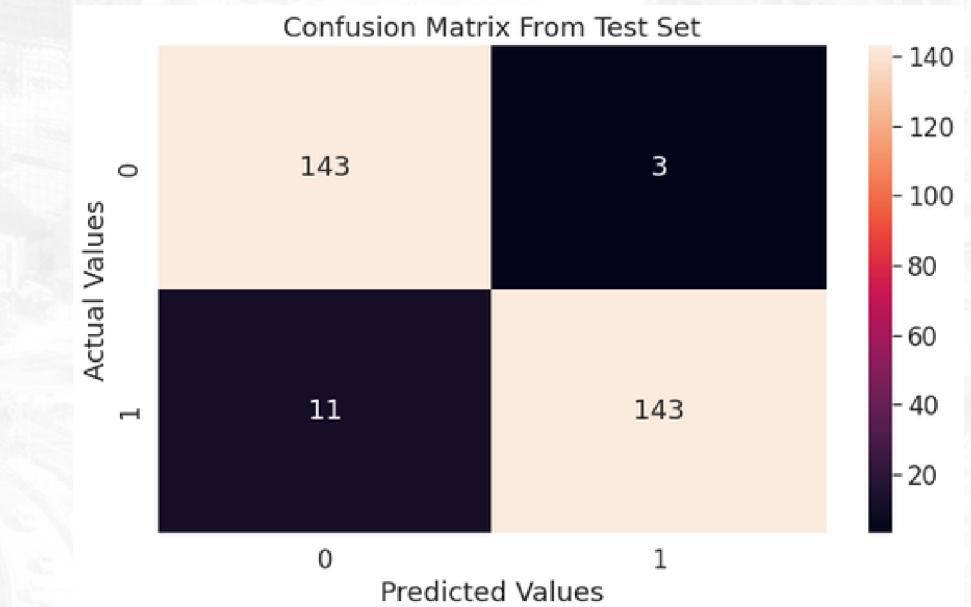
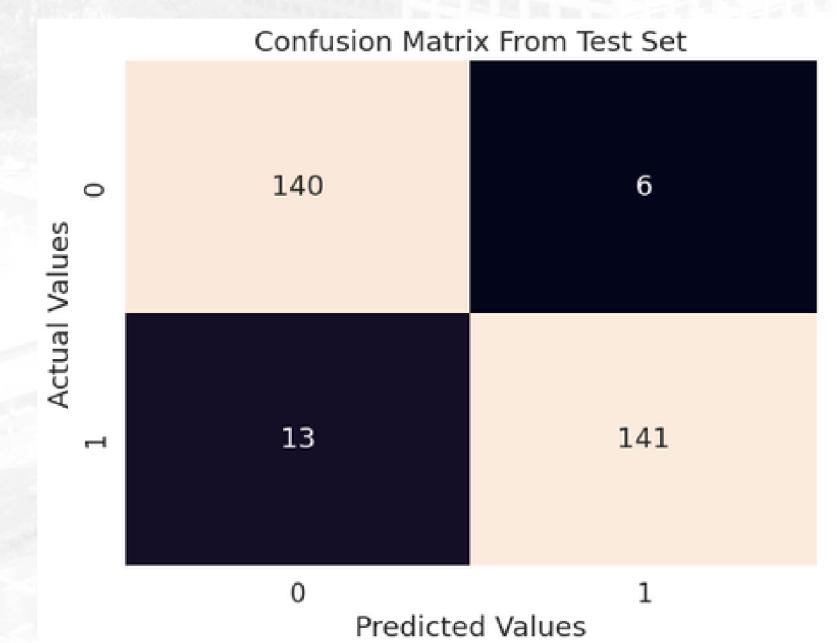
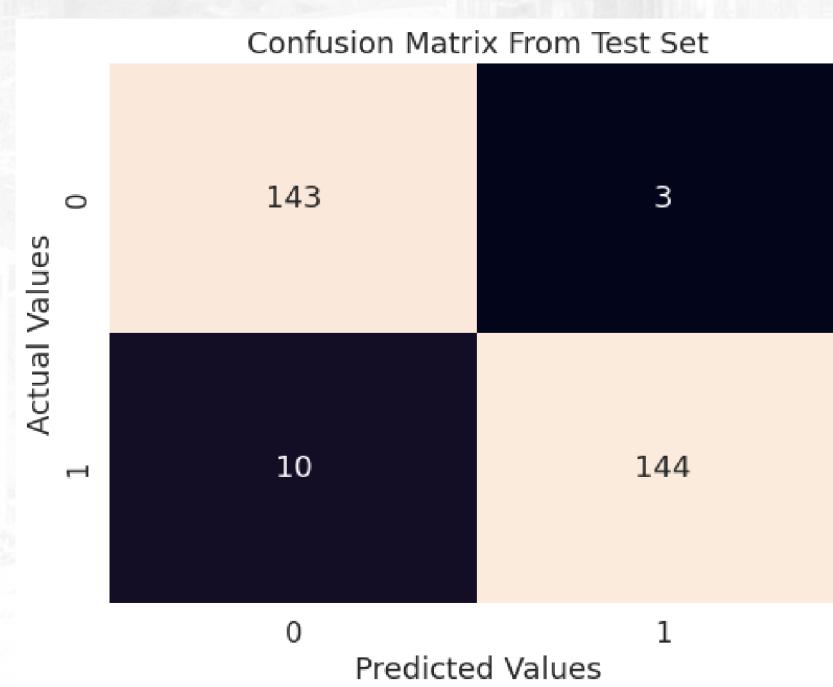
```
Accuracy (Train Set): 0.98
Accuracy (Test Set): 0.96
Precision (Train Set): 0.99
Precision (Test Set): 0.98
Recall (Train Set): 0.98
Recall (Test Set): 0.94
F1-Score (Train Set): 0.98
F1-Score (Test Set): 0.96
AUC (Train Set): 0.98
AUC (Test Set): 0.96
```

- Decission Tree

```
Accuracy (Train Set): 1.00
Accuracy (Test Set): 0.94
Precision (Train Set): 1.00
Precision (Test Set): 0.96
Recall (Train Set): 1.00
Recall (Test Set): 0.92
F1-Score (Train Set): 1.00
F1-Score (Test Set): 0.94
AUC (Train Set): 1.00
AUC (Test Set): 0.94
```

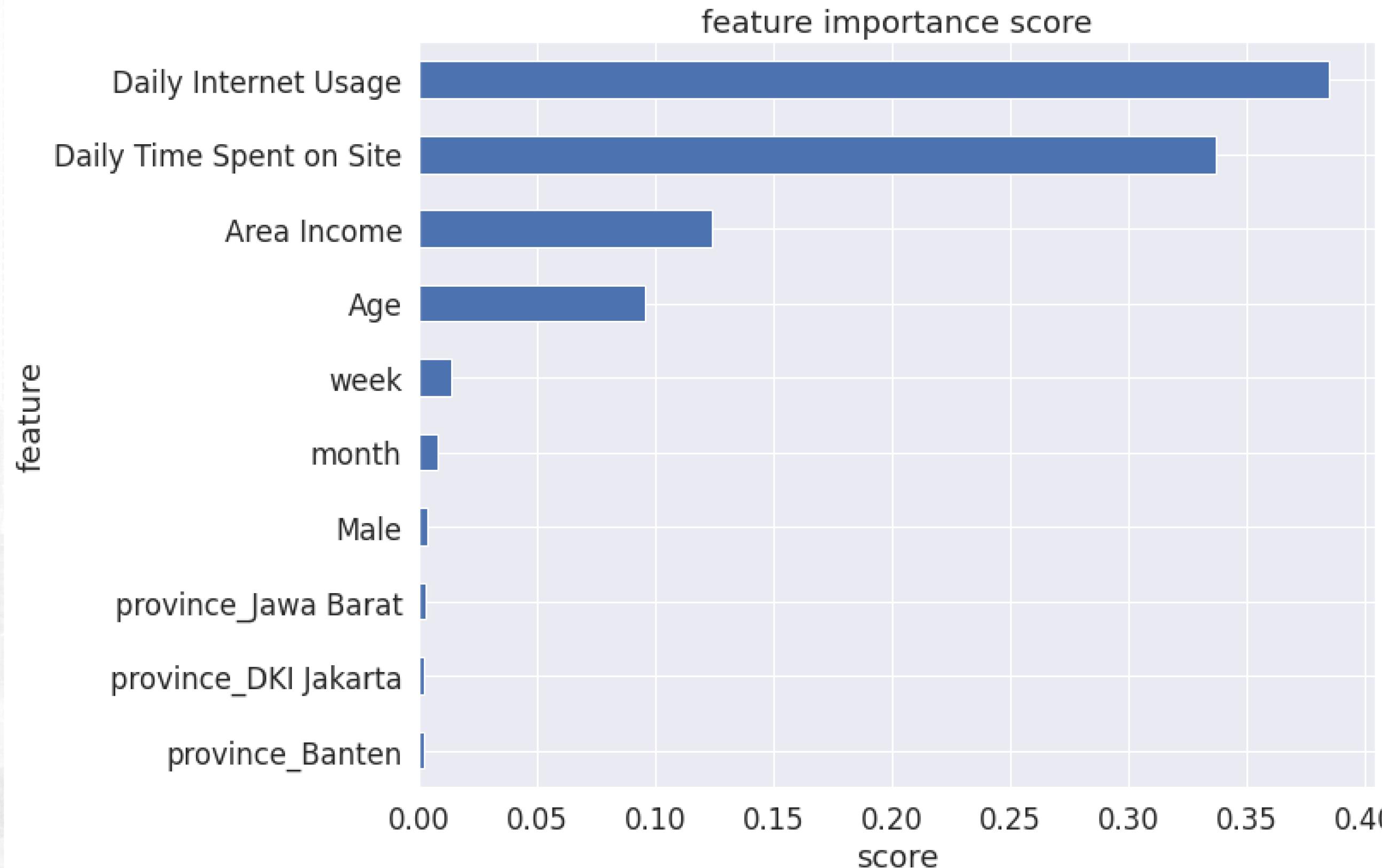
- Random Foresst

```
Accuracy (Train Set): 1.00
Accuracy (Test Set): 0.95
Precision (Train Set): 1.00
Precision (Test Set): 0.98
Recall (Train Set): 1.00
Recall (Test Set): 0.93
F1-Score (Train Set): 1.00
F1-Score (Test Set): 0.95
AUC (Train Set): 1.00
AUC (Test Set): 0.95
```



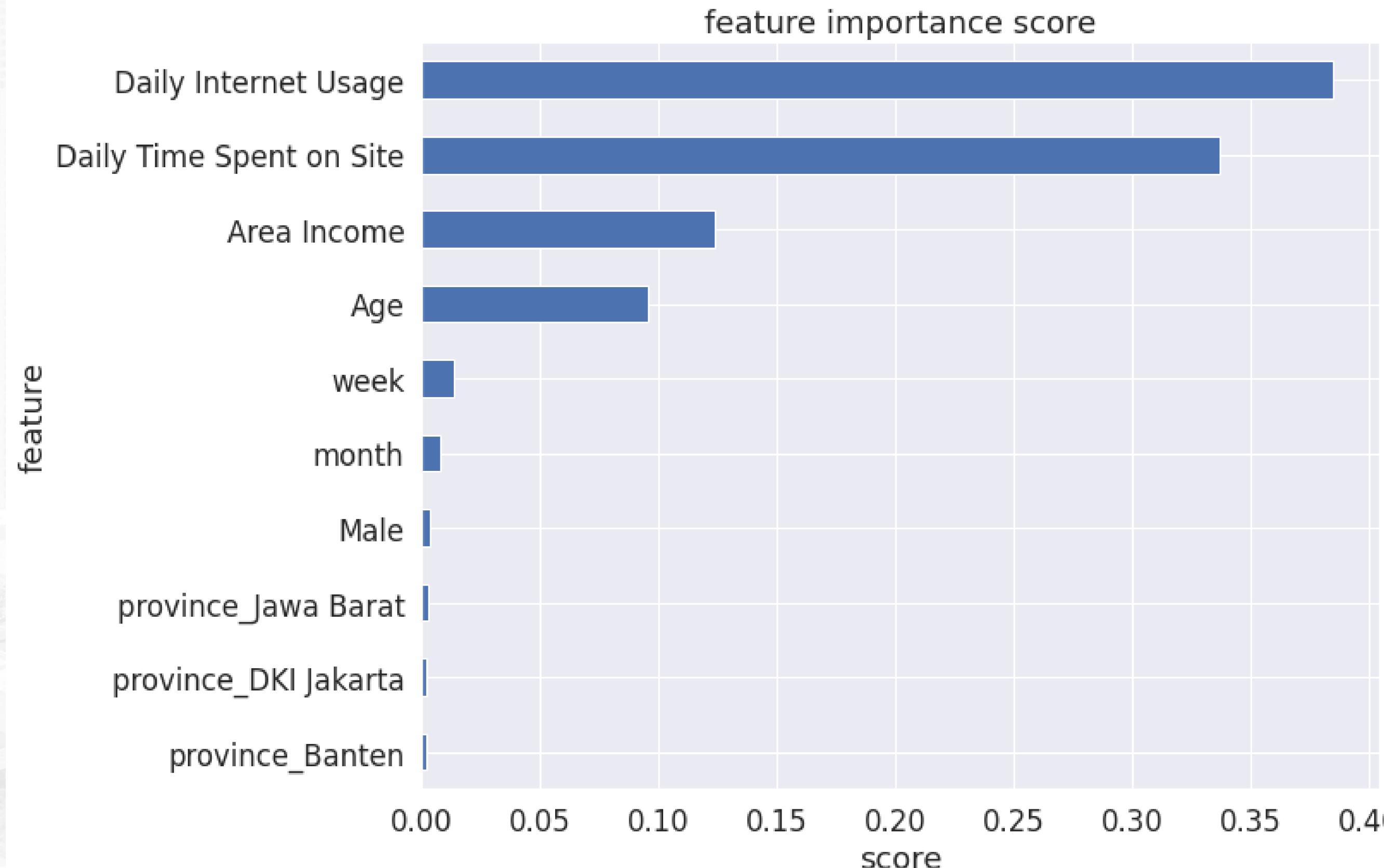
# Data Modeling

## Feature Important



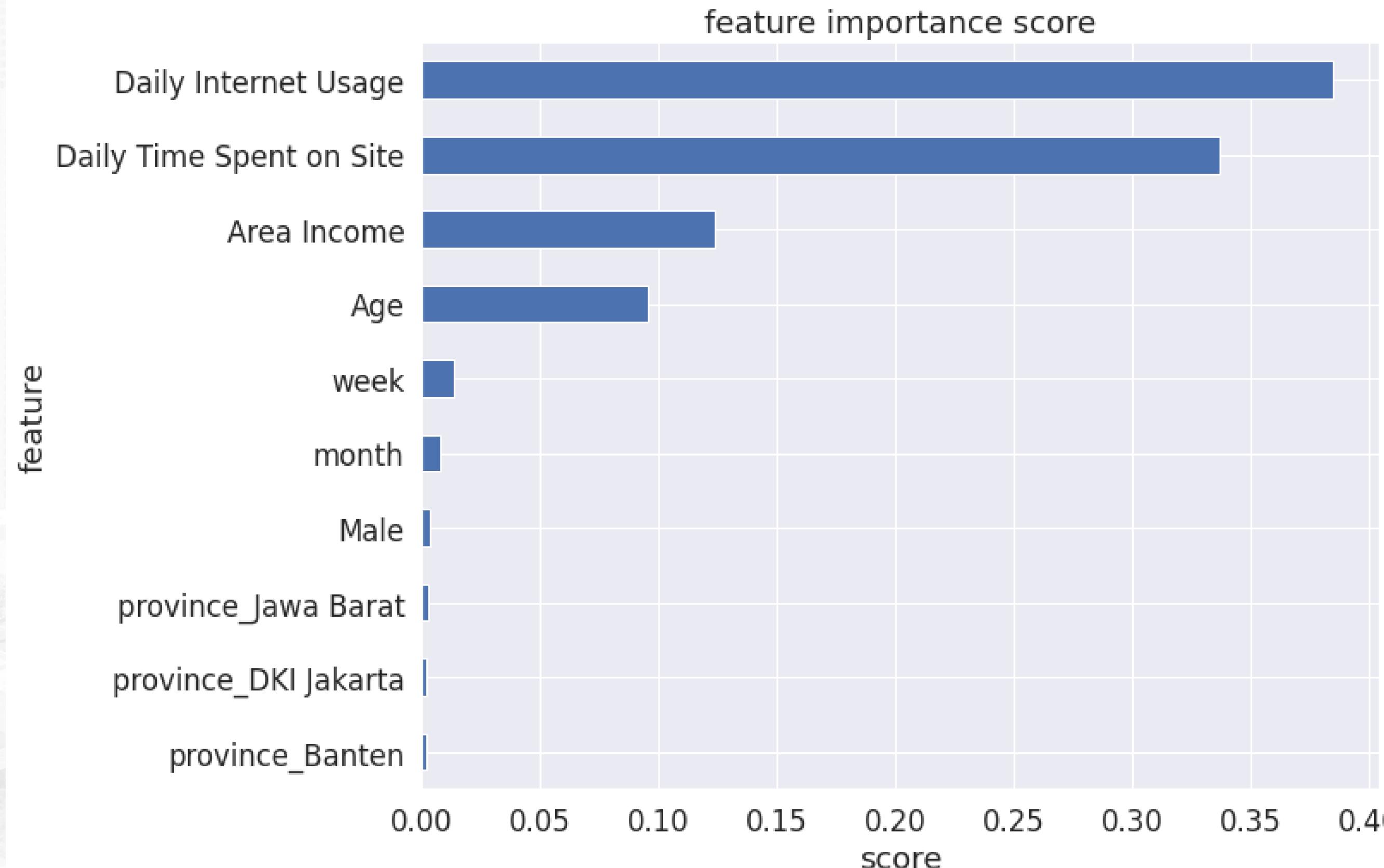
# Business Recommendation & Simulation

## Feature important



# Business Recommendation & Simulation

## Feature important



# Business Recommendation & Simulation

## Rekomendasi bisnis Berdasarkan EDA

**Berdasarkan EDA dan kepentingan fitur, dapat disimpulkan bahwa:**

- Data yang kami dapatkan mempunyai 2 segmen pengguna yaitu segmen pengguna aktif dan non aktif, dimana pengguna aktif mempunyai ciri-ciri pengguna yang sering menggunakan internet dan sering mengunjungi website suatu produk, selain itu juga memiliki pendapatan yang lebih tinggi seiring bertambahnya usia. kisaran 20–40.
- Sedangkan pengguna non aktif jarang sekali merupakan kebalikan dari pengguna aktif.
- Pengguna non-aktif cenderung lebih mudah tertarik untuk mengklik iklan produk pada iklan digital, dibandingkan pengguna aktif.
- Usia menengah merupakan pasar potensial bagi pasar digital. Titik aksi:
- Kita bisa mengubah cara mengiklankan produk seperti tidak menampilkan iklan terlalu banyak sehingga bisa menarik perhatian pengguna aktif.

# Business Recommendation & Simulation

## Simulasi

Berdasarkan kinerja yang diharapkan, kami mendapatkan akurasi hasil pengujian sebesar 95%, sehingga jika diterapkan pada dataset awal, kami akan mendapatkan 950 pengguna yang melakukan konversi berdasarkan pengguna dengan karakteristik potensial mengklik produk iklan.

Dengan biaya iklan yang sama yaitu 1 juta  
Sedangkan tingkat konversi yang akan diperoleh adalah 95% (950 user convert)

Maka kita akan mendapat  $950 * \text{Rp.} 5000 = \text{Rp.} 4.750.000$   
Pendapatan = Rp. 4.750.000

Keuntungan = Rp. 4.750.000 – Rp. 1.000.000 = Rp3.750.000

Berdasarkan simulasi di atas, jika kita tidak menggunakan model pembelajaran mesin, maka kita akan mendapatkan pendapatan 1,5 juta dan dengan penggunaan ML pendapatan meningkat secara signifikan lebih dari dua kali lipat.

# Business Recommendation & Simulation

## Simulasi

Berdasarkan kinerja yang diharapkan, kami mendapatkan akurasi hasil pengujian sebesar 95%, sehingga jika diterapkan pada dataset awal, kami akan mendapatkan 950 pengguna yang melakukan konversi berdasarkan pengguna dengan karakteristik potensial mengklik produk iklan.

Dengan biaya iklan yang sama yaitu 1 juta  
Sedangkan tingkat konversi yang akan diperoleh adalah 95% (950 user convert)

Maka kita akan mendapat  $950 * \text{Rp.} 5000 = \text{Rp.} 4.750.000$   
Pendapatan = Rp. 4.750.000

Keuntungan = Rp. 4.750.000 – Rp. 1.000.000 = Rp3.750.000

Berdasarkan simulasi di atas, jika kita tidak menggunakan model pembelajaran mesin, maka kita akan mendapatkan pendapatan 1,5 juta dan dengan penggunaan ML pendapatan meningkat secara signifikan lebih dari dua kali lipat.