Predict Clicked Ads Customer Classification by using Machine Learning

Created by: Dimas Jabbar Rosul
Your Name
DimasJRosul@gmail.com
https://www.linkedin.com/in/dimas-jabbar-rosul/

A Fresh graduate majored in Mathematics from University of Indonesia (UI). A highly motivated, Collaborative, and technically-minded person who would like to make a high impact on society. Wholeheartedly interested in data science, machine learning, business intelligence, and math. He is a self-motivated, committed, and determined person in achieving his goals. He also has demonstrated organizing skills in leading a team. He has a great curiosity about new things and likes the learning process.

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 ”

Dataset

Deskripsi Data
Dataset yang digunakan terdiri dari fitur-fitur behaviour customer pada platform

Shape
1000 row dan 10 fitur

DTYPE
Int64 (1 fitur), Float64 (3 fitur), object (6 fitur).

Missing value
Fitur yang mempunyai nilai null: ‘Daily Time Spent on Site’,’Area Income’,’Daily Internet Usage’,’Male’.

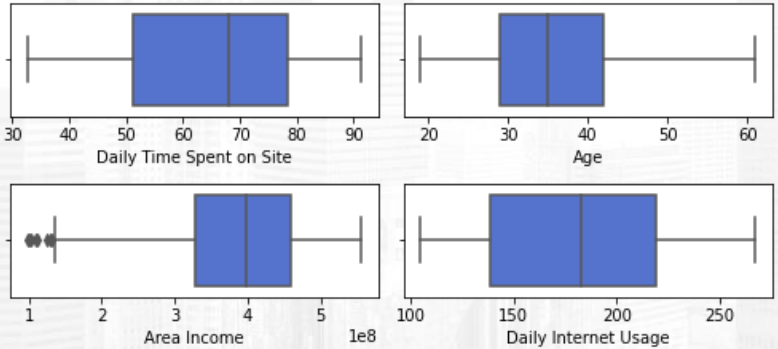
Duplicated data
0 data rows

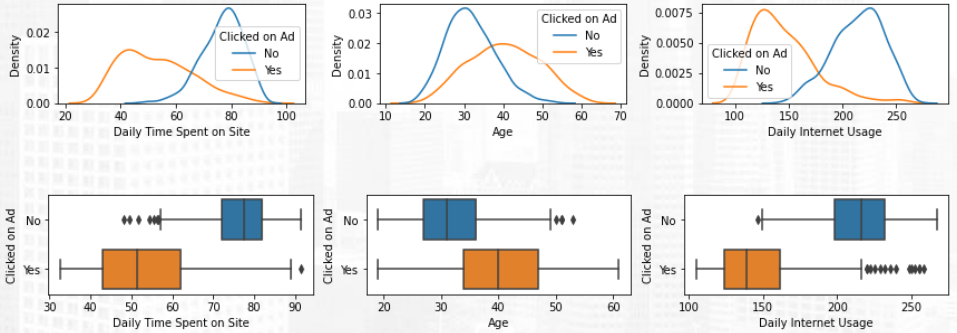
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 10 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Daily Time Spent on Site  987 non-null    float64
 1   Age                       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
```

Exploratory Data Analysis (EDA)

Univariate Analysis

Numerical Features





Observasi :

- Outliers hanya ada pada feature Area Income
- Feature Daily Time Spent on Site, Age, dan Area Income distribusinya sedikit skewed
- Feature Daily Internet Usage distribusinya mendeketi normal

Observasi :

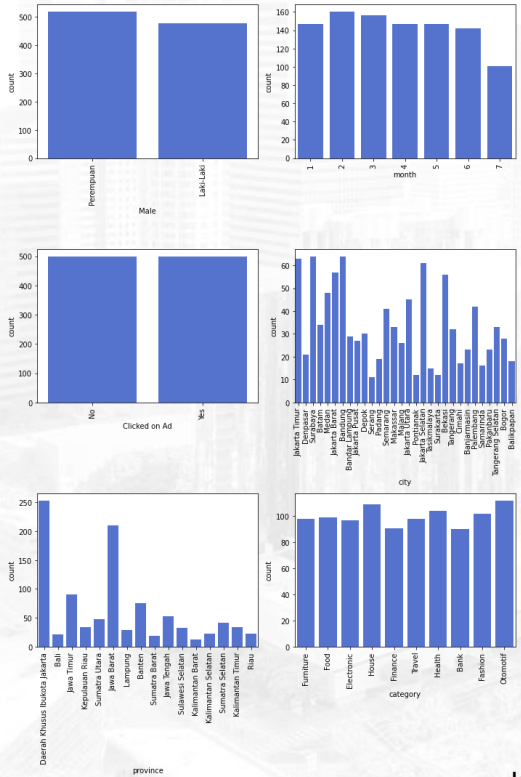
- User yang mengklik Ads adalah user dengan Daily Time Spend on Site sekitar 40-45 menit. Sedangkan, user yang tidak mengklik Ads adalah user dengan Daily Time Spend on Site sekitar 75-80 menit.
- User yang mengklik Ads rata-rata ada pada usia(Age) 40 tahun. Sedangkan, user yang tidak mengklik Ads sebagian 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.

https://colab.research.google.com/drive/1oztWUw3e3Md3k6PuDHiwpkNmNKvsYKkD?usp=sharing

Exploratory Data Analysis (EDA)

Univariate Analysis

Categorical Features



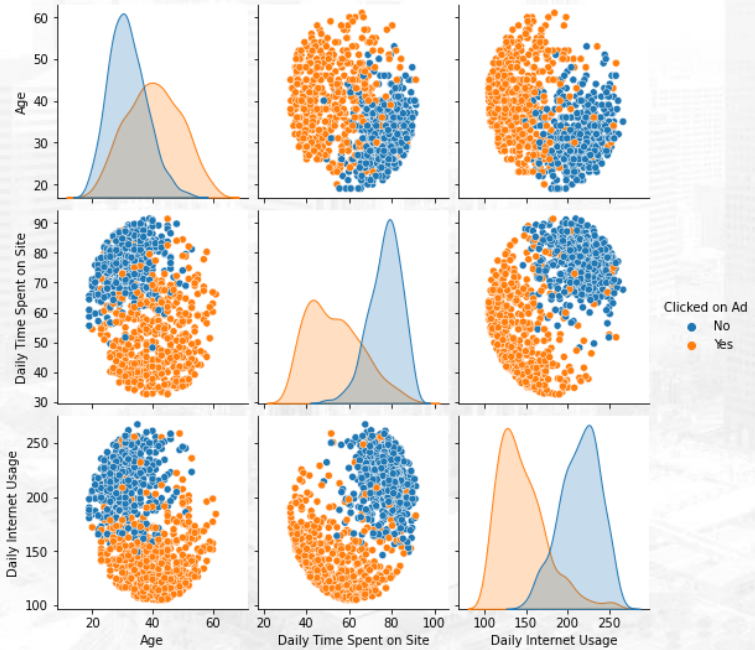
Observasi :

- Label Perempuan dan Laki-Laki pada feature Male tidak terlalu timpang
- Label Yes dan No pada feature Clicked on Ad balance
- Feature province didominasi oleh 2 nilai

https://colab.research.google.com/drive/1oztWUw3e3Md3k6PuDHiwpkNmNKvsYKkD?usp=sharing

Bivariate Analysis

Numerical Features



Observasi :

- 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.

https://colab.research.google.com/drive/1oztWUw3e3Md3k6PuDHiwpkNmNKvsYKkD?usp=sharing

Exploratory Data Analysis (EDA)

Bivariate Analysis

Categorical Features

Observasi :

•Pada feature Male, user perempuan lebih banyak mengklik Ads dibandingkan dengan user laki-laki

•Setiap bulannya, perbandingan jumlah antara user yang mengklik/tidak Ads hampir sebanding. Perbedaaan yang cukup signifikan terdapat pada bulan ke 12, dimana jumlah user yang mengklik Ads 2 kali lebih banyak dibandingkan yang tidak mengklik Ads.

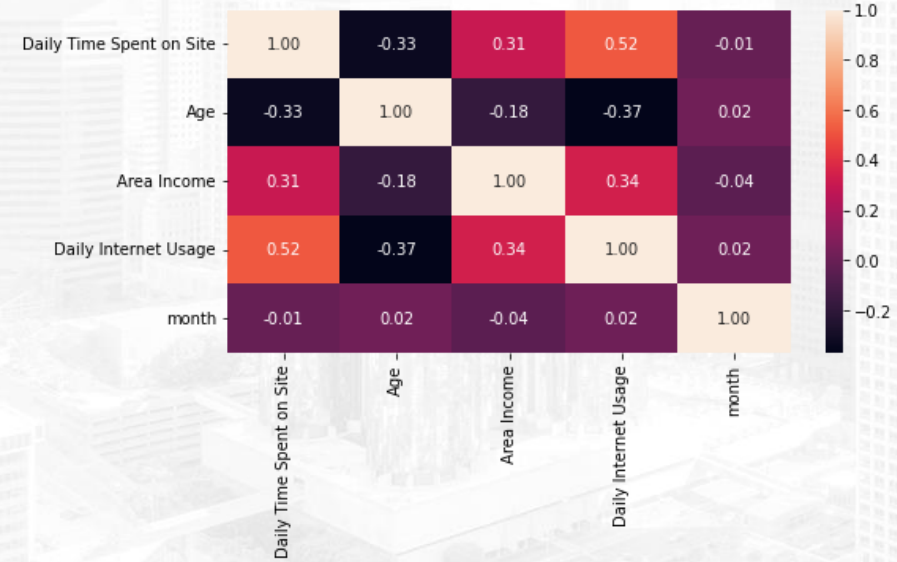
•13 dari 30 kota yang ada pada feature city, Jumlah user yang mengklik Ad lebih banyak dibandingkan dengan yang tidak mengklik Ads.

•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.

https://colab.research.google.com/drive/1oztWUw3e3Md3k6PuDHiwpkNmNKvsYKkD?usp=sharing

Multivariate Analysis



Observasi :

- 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

https://colab.research.google.com/drive/1oztWUw3e3Md3k6PuDHiwpkNmNKvsYKkD?usp=sharing

Data Cleaning & Preprocessing

Handling Missing Value

• Fitur numerik yang mempunyai missing value adalah: Daily Time Spent on Site, Area Income, Daily Internet Usage. Fitur yang kosong ini diisi dengan “median” dari masing-masing fitur, peggunaan median digunakan karena distribusi data yang cenderung skewed.

•Fitur kategorik yang mempunyai missing value adalah fitur ‘male’, Fitur yang kodong ini diisi dengan modus dari fitur ini.

Check Duplicated Data

•Dataset ini tidak mempunyai nilai duplikat

Extract datetime data

•Dilakukan ektraksi fitur pada timestamp, sehingga menghasilkan beberapa kolom baru yaitu: day, week, month, year.

Split data

Split data pada project ini menggunakan split:
• 80:20, dimana 80% untuk data train dan 20% untuk data testing.
•70:30, dimana 70% untuk data train dan 30% untuk data testing.

Feature Encoding

•Feature encoding yang digunakan adalah Teknik “One Hot Encoding” untuk data kategorikal seperti: city, province, category.

https://colab.research.google.com/drive/1oztWUw3e3Md3k6PuDHiwpkNmNKvsYKkD?usp=sharing

Data Modeling

Rakamin Academy

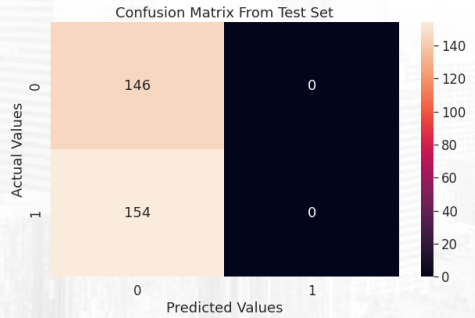
Experiment 1

70:30

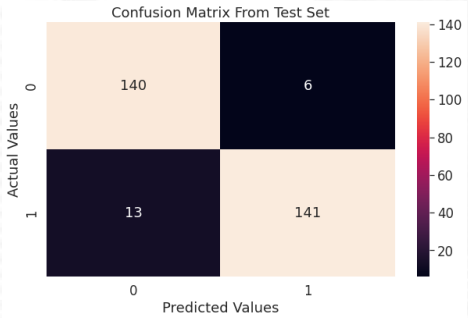
Note:

- **sebelum** normalisasi/standardisasi

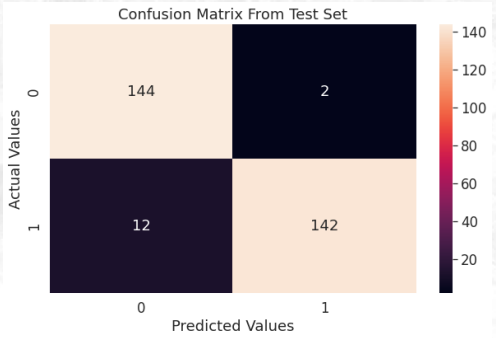
Logistic Regression



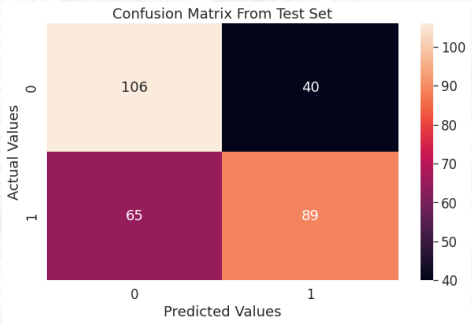
Decision Tree



Random Forest



KNN



Data Modeling

Rakamin Academy

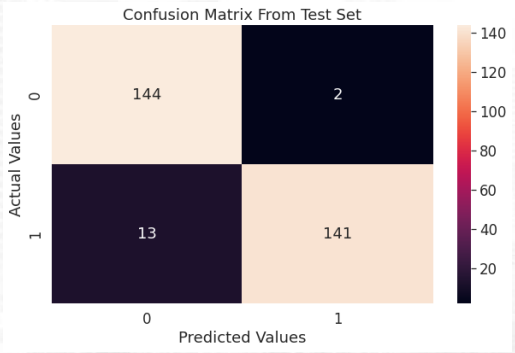
Experiment 1

70:30

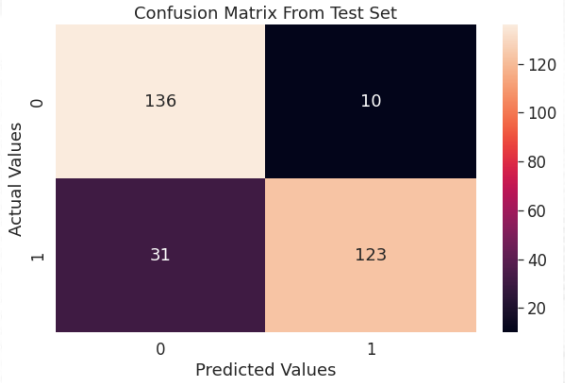
Note:

- **sebelum** normalisasi/standardisasi
- **Hyperparameter Tuning**

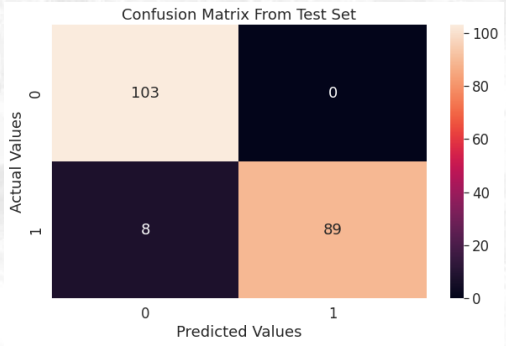
Logistic Regression



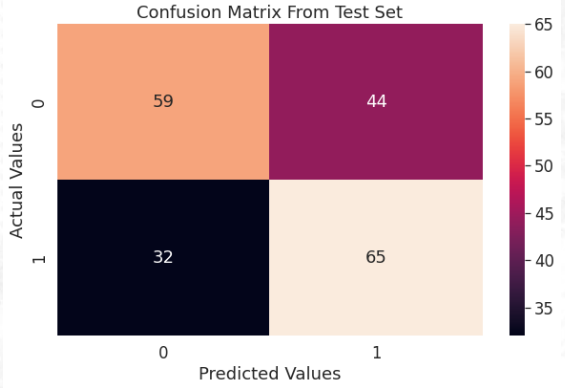
Decision Tree



Random Forest



KNN



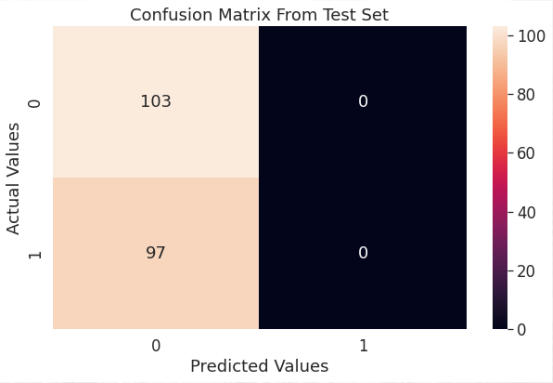
Rakamin Academy

Experiment 1

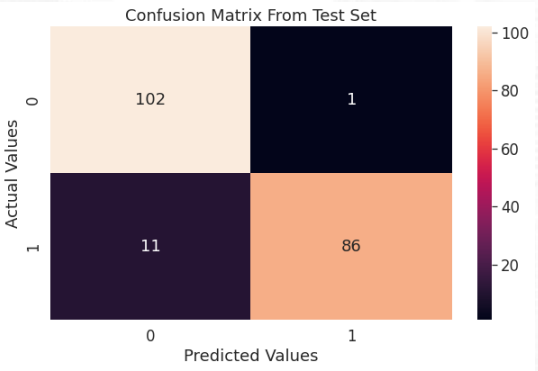
Note:
- **sebelum** normalisasi/standardisasi

80:20

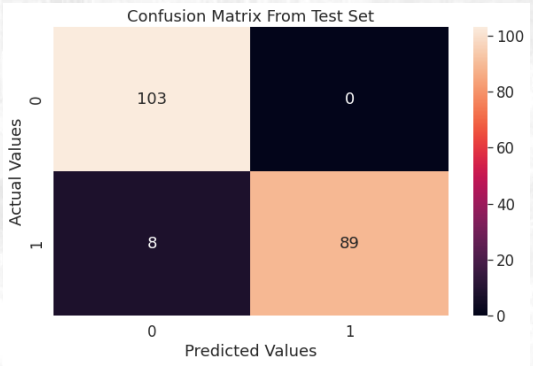
Logistic Regression



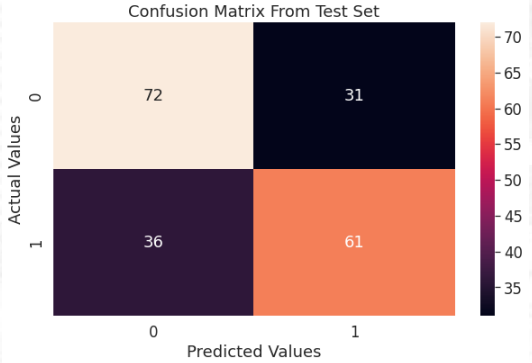
Decision Tree



Random Forest



KNN



Data Modeling

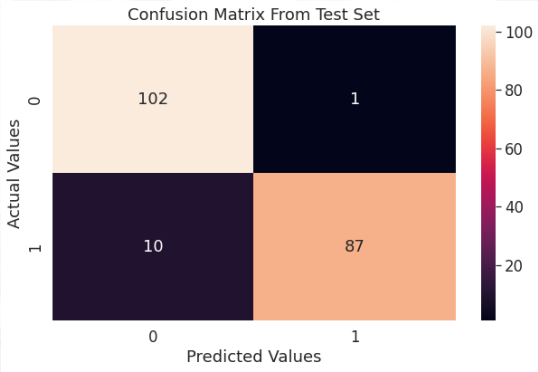
Rakamin Academy

Experiment 1

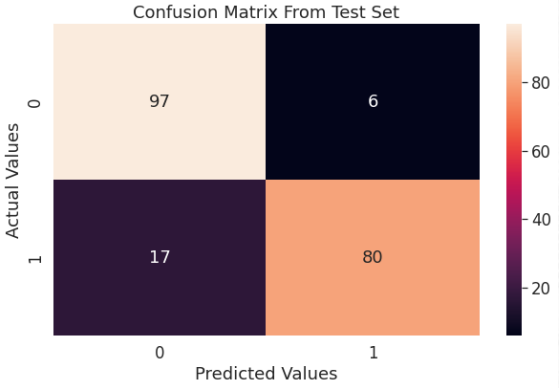
80:20

Note:
- **sebelum** normalisasi/standardisasi
- **Hyperparameter Tuning**

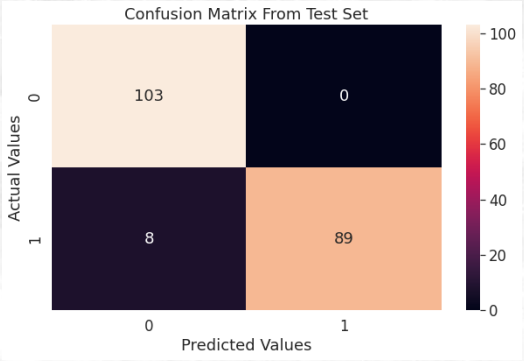
Logistic Regression



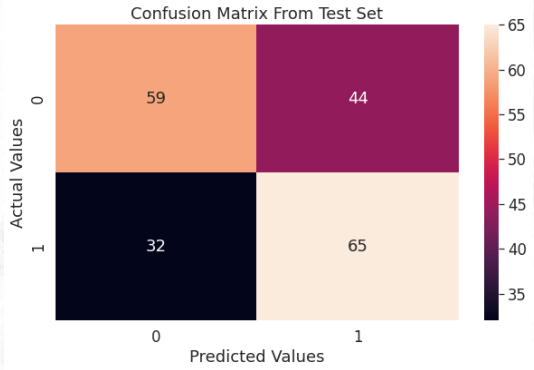
Decision Tree



Random Forest



KNN



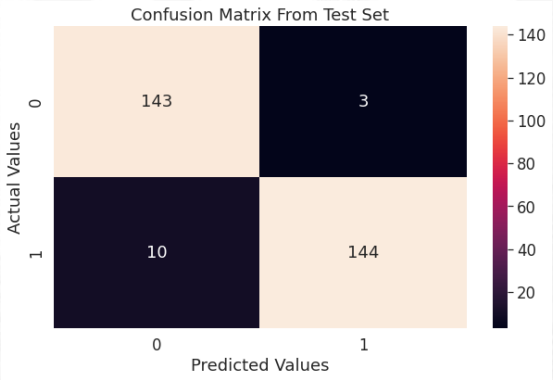
Rakamin Academy

Experiment 2

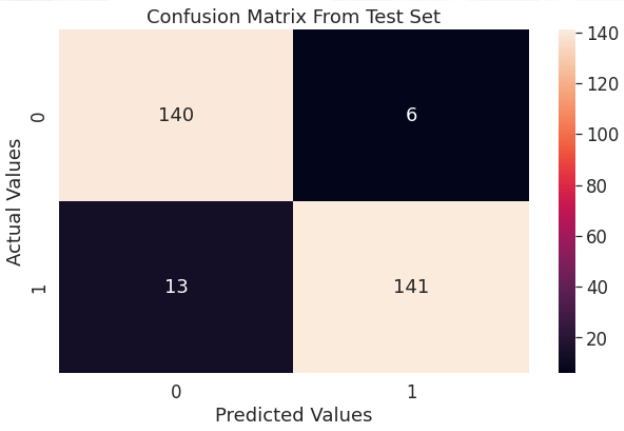
Note:
- **dengan** normalisasi/standardisasi

70:30

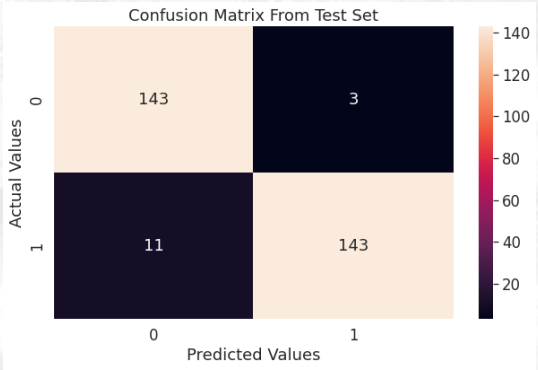
Logistic Regression



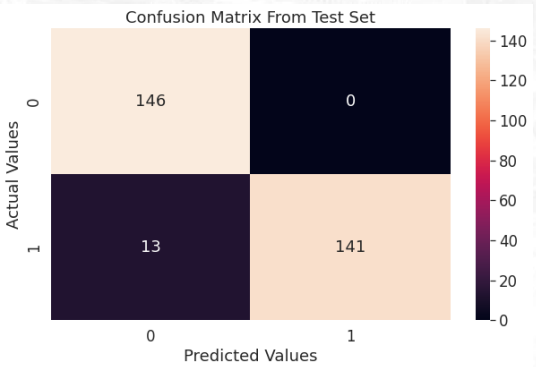
Decision Tree



Random Forest



KNN



Data Modeling

Rakamin Academy

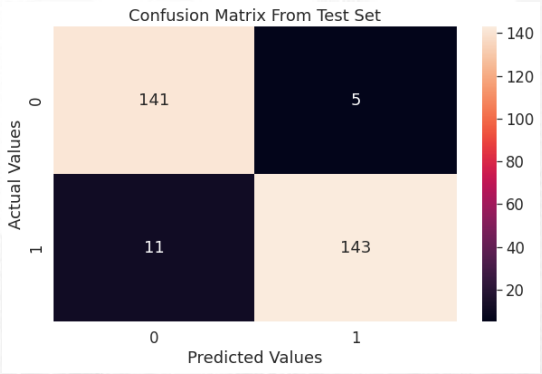
Experiment 2

70:30

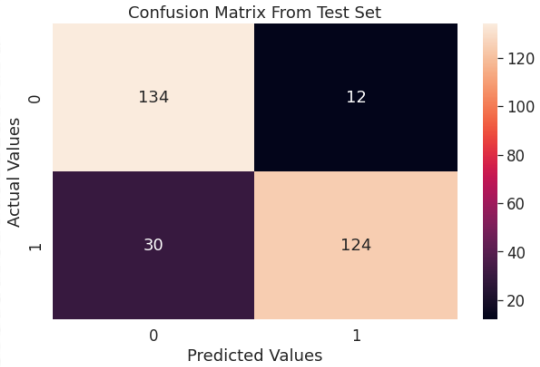
Note:

- **dengan** normalisasi/standardisasi
- **Hyperparameter Tuning**

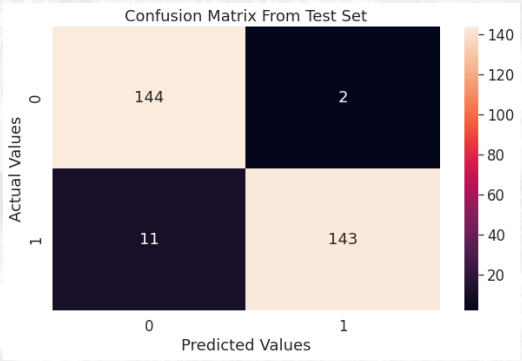
Logistic Regression



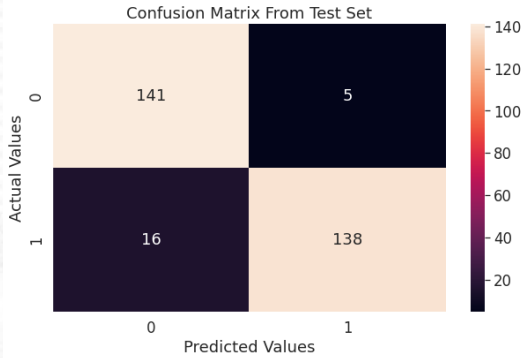
Decision Tree



Random Forest



KNN



Data Modeling

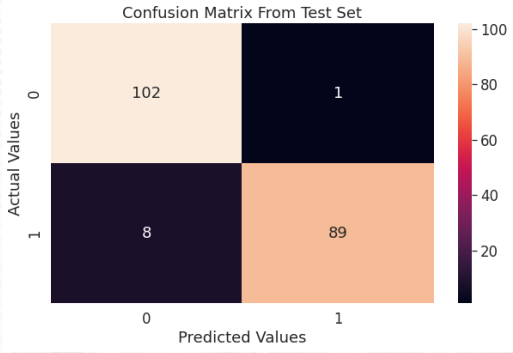
Rakamin Academy

Experiment 2

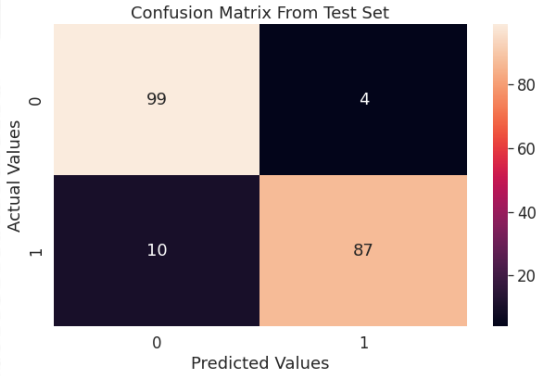
Note:
- **dengan** normalisasi/standardisasi

80:20

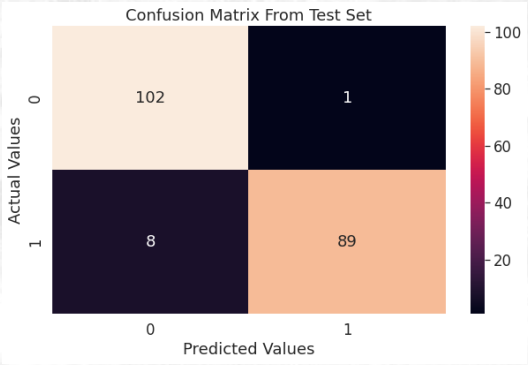
Logistic Regression



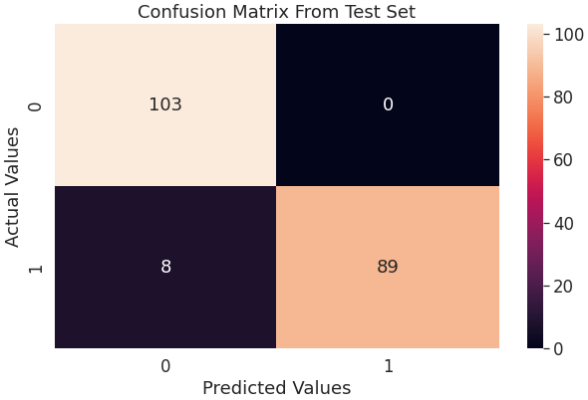
Decision Tree



Random Forest



KNN



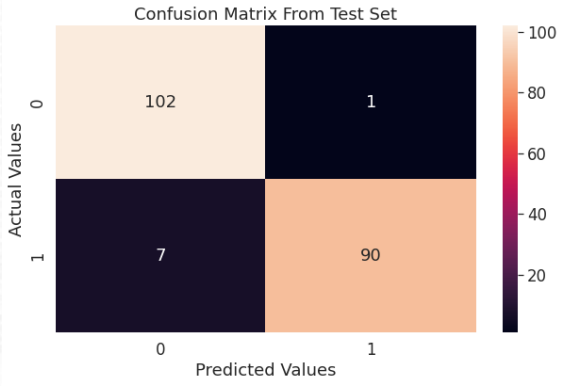
Rakamin Academy

Experiment 2

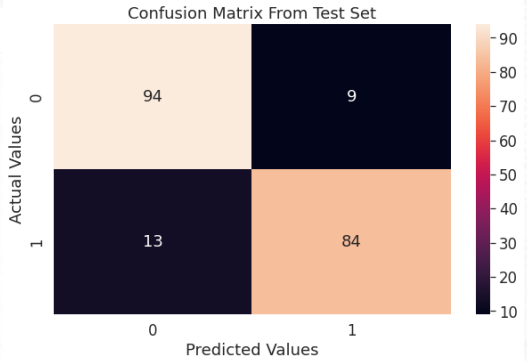
80:20

Note:
- **dengan** normalisasi/standardisasi
- **Hyperparameter Tuning**

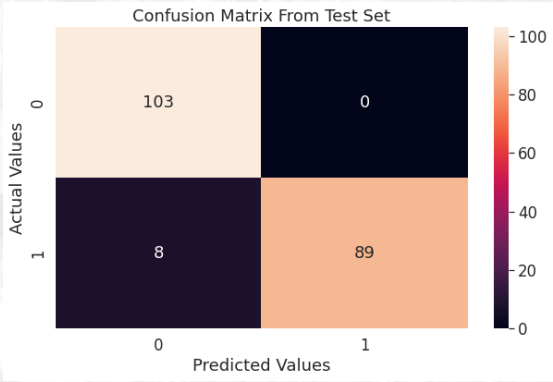
Logistic Regression



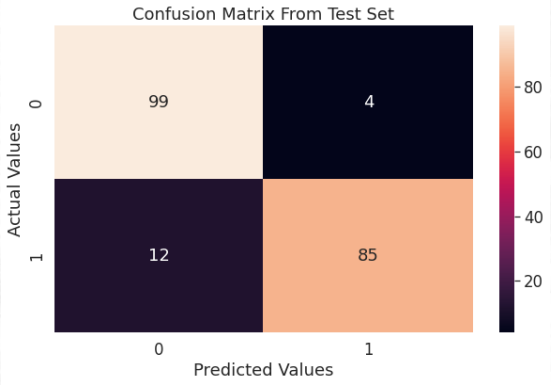
Decision Tree



Random Forest

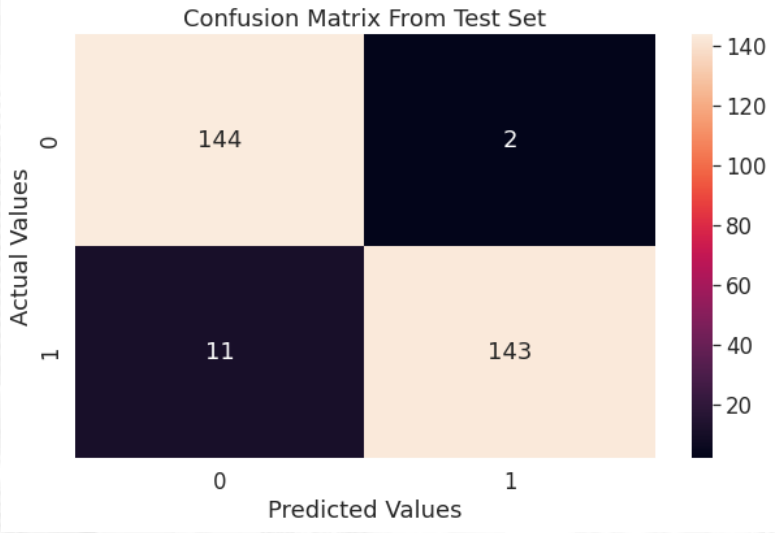


KNN



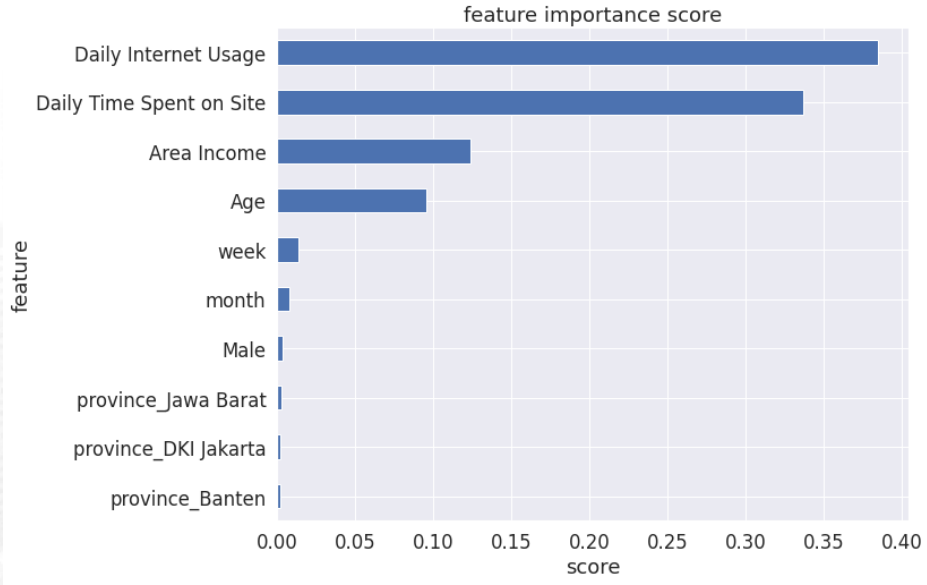
Model Selection

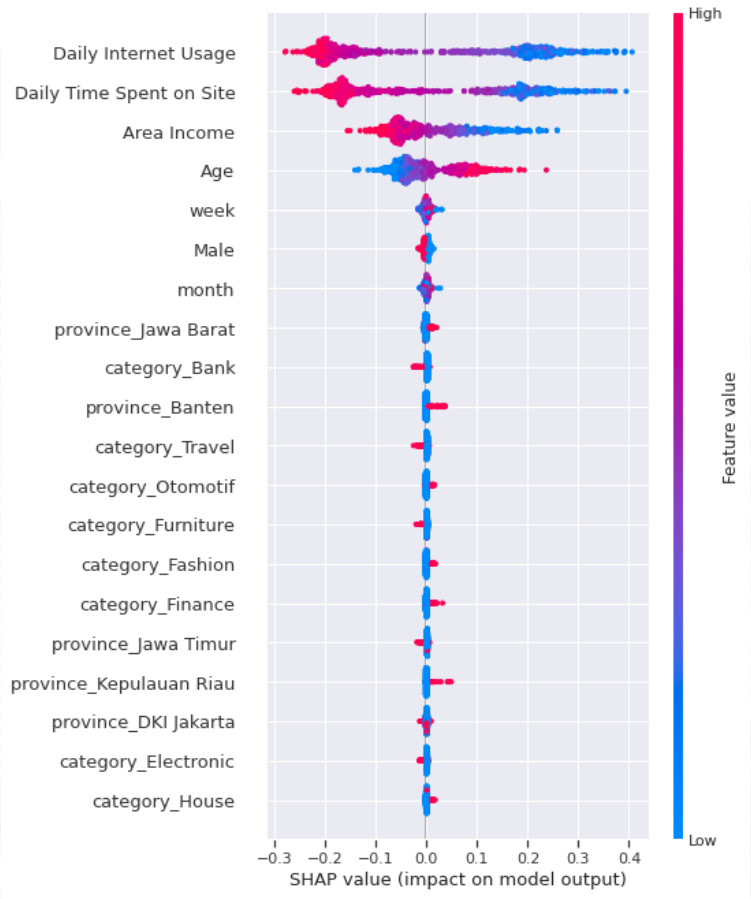
Model yang dipilih adalah **Random Forest** pada experiment 2 (dengan normalisasi data), split 70:30, dengan hyperparameter tuning.



Accuracy (Train Set): 0.99
Accuracy (Test Set): 0.96
Precision (Train Set): 0.99
Precision (Test Set): 0.99
Recall (Train Set): 0.99
Recall (Test Set): 0.93
F1-Score (Train Set): 0.99
F1-Score (Test Set): 0.96
AUC (Train Set): 0.99
AUC (Test Set): 0.96

Feature Importance

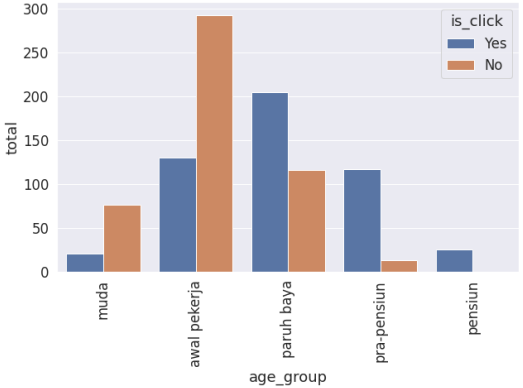




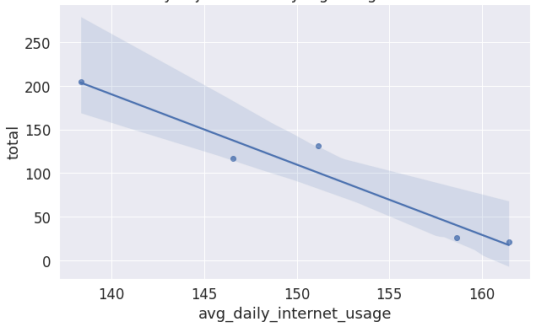
Dari model yang dipilih, yakni random forest. Kita dapat mengetahui bahwa faktor/fitur yang paling mempengaruhi prediksi pada model adalah ‘Daily Internet Usage’, ‘Daily Time Spent on Site’,’Area Icome’,’Age’.

Business Recommendation

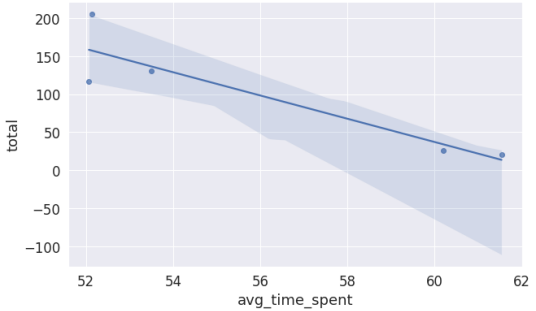
User usia paruh baya yang mengklik iklan jumlahnya 1.7x lebih banyak dibandingkan dengan yang tidak mengklik iklan
User usia pra-pensiun yang mengklik iklan jumlahnya 8x lebih banyak dibandingkan dengan yang tidak mengklik iklan
Seluruh user yang ada pada usia pensiun mengklik iklan tetapi jumlahnya sedikit



Berdasarkan User yang mengklik Ads
Semakin sedikit rata-rata penggunaan internet harian
maka semakin banyak jumlah user yang mengklik Ads



Berdasarkan User yang mengklik Ads
Semakin sedikit rata-rata waktu yang dihabiskan di website
maka semakin banyak jumlah user yang mengklik Ads



Rekomendasi:

- Berdasarkan EDA, user paruh baya 1.7 kali lebih banyak melakukan klik daripada yang tidak. Sementara untuk pra prensiun 8x lebih banyak melakukan klik dibandingkan yang tidak. Jika dibandingkan dengan kategori lain, kategori paruh baya dan pra-pension mempunyai potensi yang target marketing yg lebih menjanjikan. Sementara total kedua kategori terpapar iklan tidak lebih banyak dari kategori awal pekerja, oleh karena itu direkomendasikan untuk menargetkan campaign ke 2 kategori tersebut.
- Diketahui juga user yang menghabiskan banyak waktu di website semakin sedikit pula klik nya, oleh karena itu, direkomendasikan untuk ada inisiasi pembuatan semacam reminder atau UI yang menarik agar calon customer dapat klik campaign tersebut Ketika melebihi berapa waktu (threshhold) tertentu.