Predict Clicked Ads Customer Classification by using Machine Learning

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

Rakamin Academy

Created by:
Devia Febyanti
devia.febyanti99@gmail.com
https://www.linkedin.com/in/devia-febyanti/

Experienced data operator with proven work history in the field of education. Currently motivated to become a Data Scientist to help driving financial inclusion through the use of big data and machine learning.

Bachelor of Geophysics
2018-2022

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

Rakamin Academy

DATASET

```
Data columns (total 11 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Unnamed: 0                1000 non-null   int64
 1   Daily Time Spent on Site  987 non-null    float64
 2   Age                       1000 non-null   int64
 3   Area Income               987 non-null    float64
 4   Daily Internet Usage      989 non-null    float64
 5   Gender                    997 non-null    object
 6   Timestamp                 1000 non-null   object
 7   Clicked on Ad             1000 non-null   object
 8   City                      1000 non-null   object
 9   Province                  1000 non-null   object
 10  Category                  1000 non-null   object
dtypes: float64(3), int64(2), object(6)
```

- Unnamed: ID Pelanggan
- Daily Time Spent on Site: Waktu yang dihabiskan pelanggan di situs (menit)
- Age: Usia pelanggan (tahun)
- Area Income: Pendapatan rata-rata wilayah geografis pelanggan
- Daily Internet Usage: Waktu yang dihabiskan pelanggan di internet dalam satu hari (menit)
- Gender: Jenis kelamin pelanggan
- Timestamp: Waktu pelanggan mengklik iklan atau menutup jendela
- Clicked on Ad: Apakah pengguna mengklik iklan atau tidak
- City: Kota pelanggan
- Province: Provinsi pelanggan
- Category: Kategori iklan

Shape

- 1000 data rows
- 11 features

Data Type

- Float64 : 3 feature
- Int64 : 2 features
- Object : 6 features

Missing Value

- Daily Time Spent on Site
- Area Income
- Daily Internet Usage
- Gender

Duplicated

None

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Type and Behaviour Analysis on Advertisement

Rakamin Academy

EXPLORATORY DATA ANALYSIS – STATISTIC DESCRIPTIVE

Fitur Numerik

	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage
count	987.000000	1000.000000	9.870000e+02	989.000000
mean	64.929524	36.009000	3.848647e+08	179.863620
std	15.844699	8.785562	9.407999e+07	43.870142
min	32.600000	19.000000	9.797550e+07	104.780000
25%	51.270000	29.000000	3.286330e+08	138.710000
50%	68.110000	35.000000	3.990683e+08	182.650000
75%	78.460000	42.000000	4.583554e+08	218.790000
max	91.430000	61.000000	5.563936e+08	267.010000

- Tidak ada nilai yang tidak valid di antara fitur-fitur yang digunakan.
- Berdasarkan nilai mean dan median, fitur **Area Income** memiliki distribusi yang condong ke kanan (positively skewed).
- Berdasarkan nilai minimum, fitur **Area Income** memiliki outliers.

Fitur Kategorik

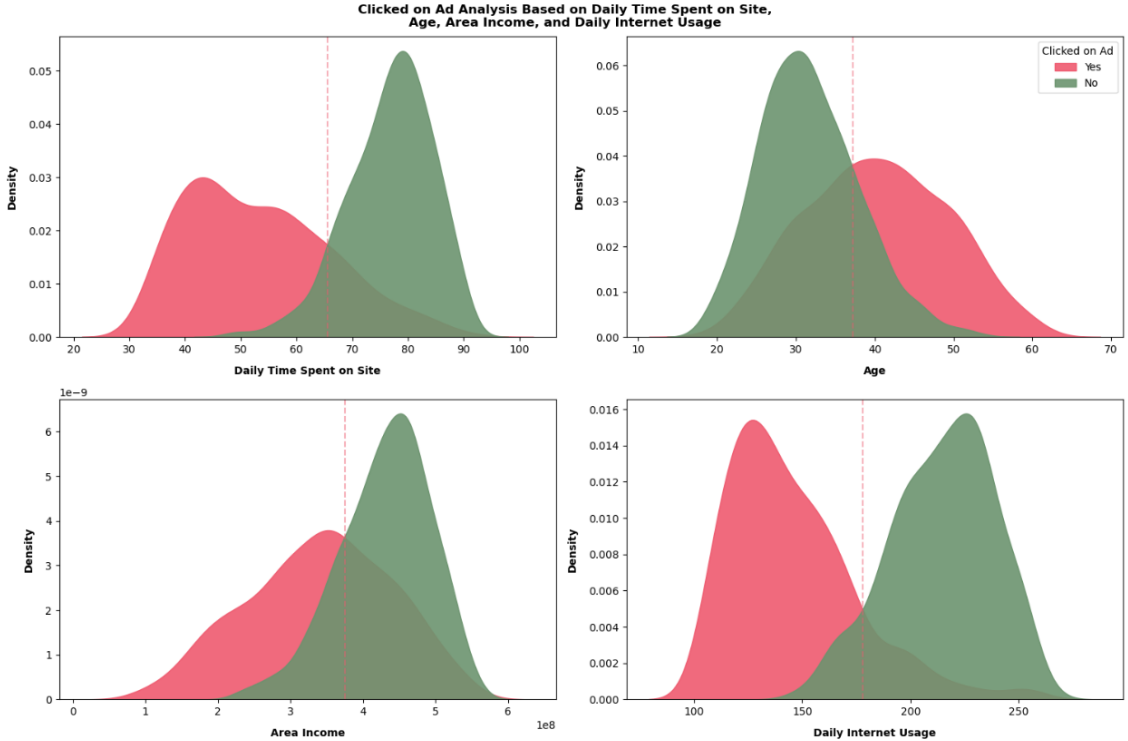
	Gender	Clicked on Ad	City	Province	Category
count	997	1000	1000	1000	1000
unique	2	2	30	16	10
top	Perempuan	No	Surabaya	Daerah Khusus Ibukota Jakarta	Otomotif
freq	518	500	64	253	112

- Fitur **Clicked on Ad** (variabel target) memiliki kelas yang seimbang (No=500 dan Yes=500), sehingga tidak memerlukan perlakuan khusus.
- Fitur **City, Province, dan Category** memiliki kardinalitas tinggi atau memiliki banyak nilai unik, sehingga kemungkinan besar akan dihapus.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Type and Behaviour Analysis on Advertisement

EXPLORATORY DATA ANALYSIS – UNIVARIATE ANALYSIS

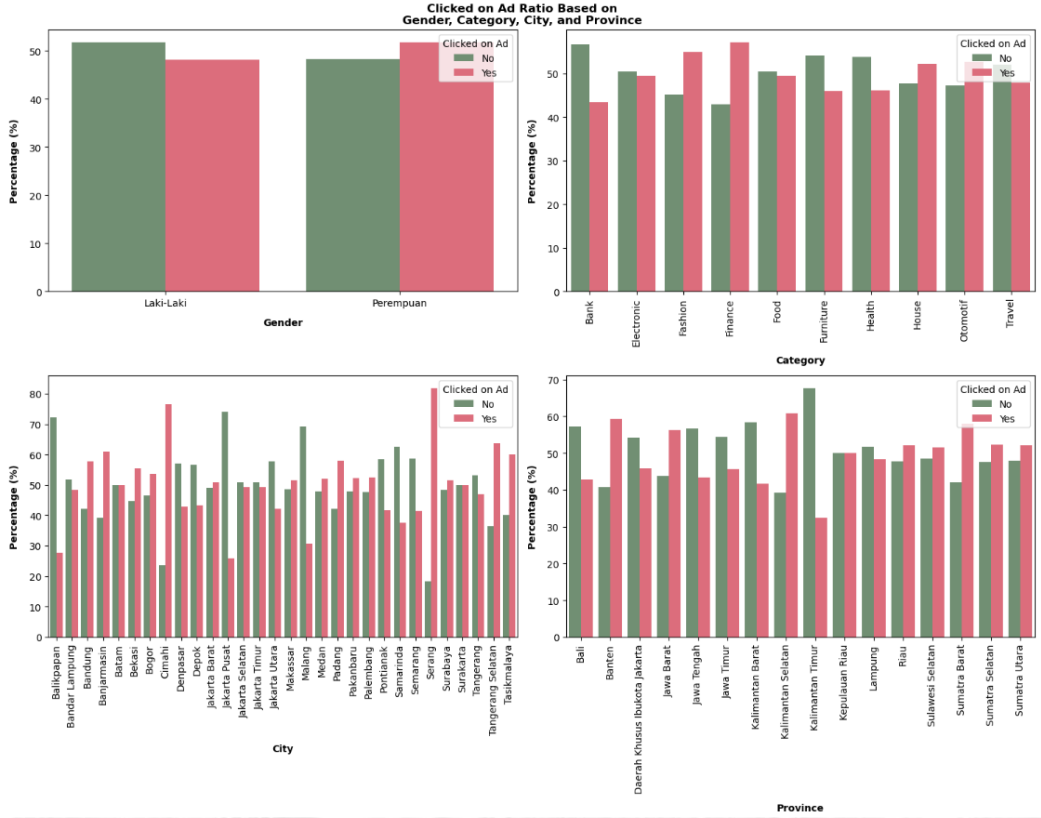


- Semakin banyak waktu yang dihabiskan di situs atau internet (**Daily Time Spent on Site**), semakin kecil kemungkinan pelanggan mengklik iklan (Clicked on Ad).
- Semakin tua usia pelanggan (**Age**), semakin besar kemungkinan pelanggan mengklik iklan (Clicked on Ad).
- Semakin tinggi pendapatan wilayah pelanggan (**Area Income**), semakin kecil kemungkinan pelanggan mengklik iklan (Clicked on Ad).
- Semakin tinggi penggunaan internet harian (**Daily Internet Usage**), semakin kecil kemungkinan pelanggan akan mengklik iklan (Clicked on Ad).

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Type and Behaviour Analysis on Advertisement

EXPLORATORY DATA ANALYSIS – UNIVARIATE ANALYSIS

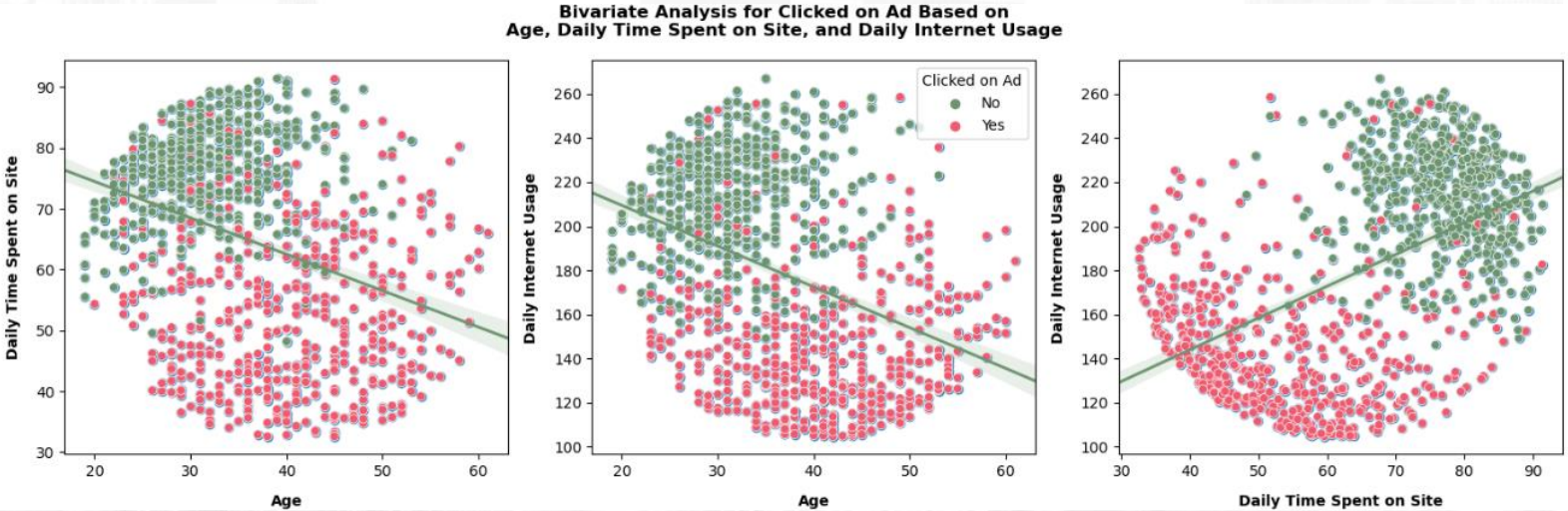


- **Perempuan** memiliki sedikit lebih tinggi kemungkinan untuk mengklik iklan (Click on Ad) dibandingkan Laki-laki.
- Kategori iklan tertinggi yang diklik pengguna adalah **Finance** dan yang terendah adalah Bank.
- Kota dengan rasio klik tertinggi adalah **Serang** dan yang terendah adalah Jakarta Pusat.
- Tiga provinsi dengan rasio klik tertinggi adalah **Kalimantan Selatan, Banten, dan Sumatra Barat**.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Type and Behaviour Analysis on Advertisement

EXPLORATORY DATA ANALYSIS – BIVARIATE ANALYSIS

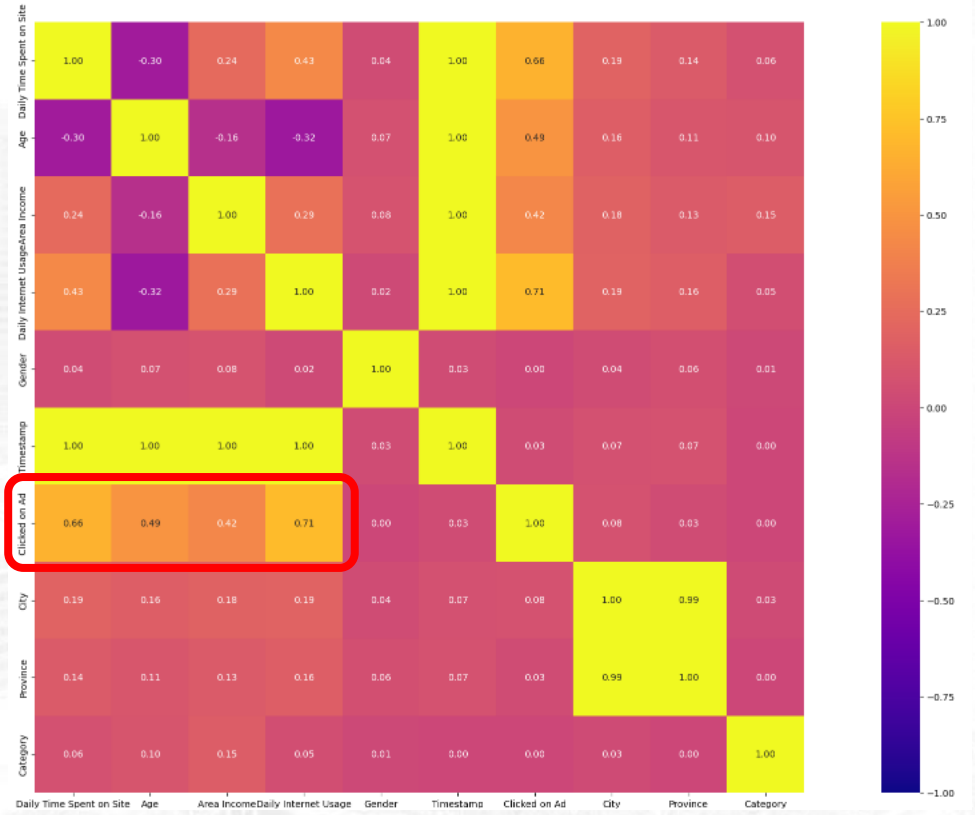


- Usia (**Age**) dengan **Daily Time Spent on Site** memiliki **korelasi negatif**. Semakin tua usia pelanggan, maka menunjukkan semakin sedikit waktu yang mereka habiskan di situs.
- Begitu pula dengan usia (**Age**) dengan **Daily Internet Usage** yang juga memiliki **korelasi negatif**. Semakin tua usia pelanggan, maka menunjukkan semakin sedikit waktu pemakaian internet.
- Sementara itu, **Daily Time Spent on Site** dan **Daily Internet Usage** memiliki **korelasi positif**. Semakin banyak waktu yang dihabiskan pelanggan di situs, maka menunjukkan semakin banyak pemakaian internet juga.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Type and Behaviour Analysis on Advertisement

EXPLORATORY DATA ANALYSIS – MULTIVARIATE ANALYSIS



Berdasarkan heatmap di atas, fitur-fitur yang berhubungan dengan variabel target (Clicked on Ad) dan akan digunakan untuk pemodelan antara lain: **Age, Area Income, Daily Internet Usage, dan Daily Time Spent on Site** karena memiliki nilai korelasi yang cukup tinggi dengan variabel target (Click on Ad).

Untuk selengkapnya, dapat melihat jupyter notebook disini

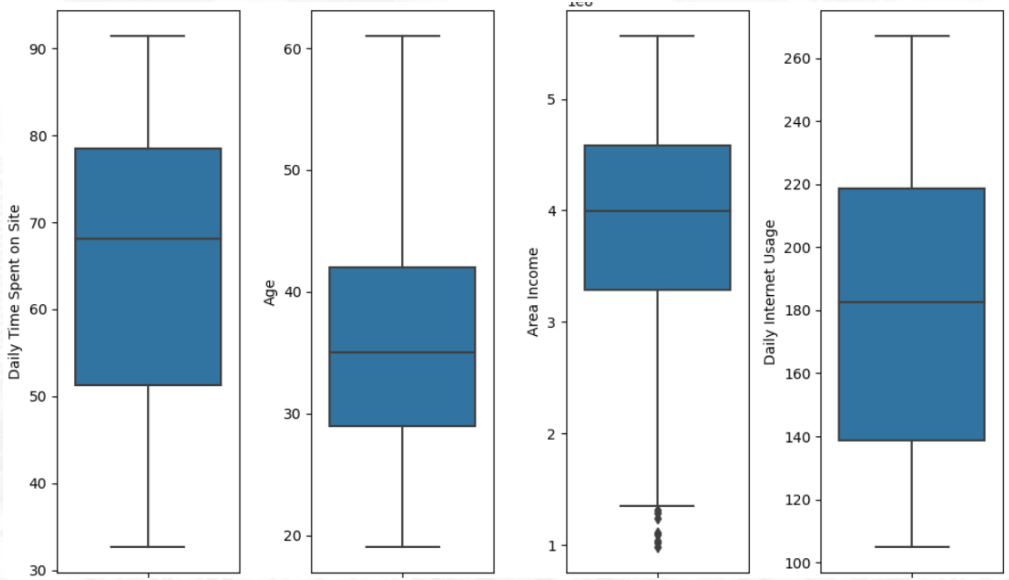
Data Cleaning & Preprocessing

HANDLING MISSING VALUE

The features that has missing value:

Daily Time Spent on Site	13
Age	0
Area Income	13
Daily Internet Usage	11
Gender	3
Timestamp	0
Clicked on Ad	0
City	0
Province	0
Category	0

Checking the outliers:



- 13 row data has missing value on **Daily Time Spent on Site**
- Fill it with mean (no outlier detected)

- 13 row data has missing value on **Area Income**
- Fill it with median (outliers detected)

- 11 row data has missing value on **Daily Internet Usage**
- Fill it with mean (no outlier detected)

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Cleaning & Preprocessing

HANDLING DUPLICATED DATA

- 0 duplicated data

FEATURE ENCODING

Feature: Clicked on Ad
0 for No
1 for Yes

FEATURE EXTRACTION

new features

	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Gender	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	0	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	0	Denpasar	Bali	Food	2016	4	14	4
2	69.47	26	418501580.0	236.50	Perempuan	2016-03-13 20:35:00	0	Surabaya	Jawa Timur	Electronic	2016	3	10	13
3	74.15	29	383643260.0	245.89	Laki-Laki	2016-01-10 02:31:00	0	Batam	Kepulauan Riau	House	2016	1	1	10
4	68.37	35	517229930.0	225.58	Perempuan	2016-06-03 03:36:00	0	Medan	Sumatra Utara	Finance	2016	6	22	3

Untuk selengkapnya, dapat melihat jupyter notebook disini

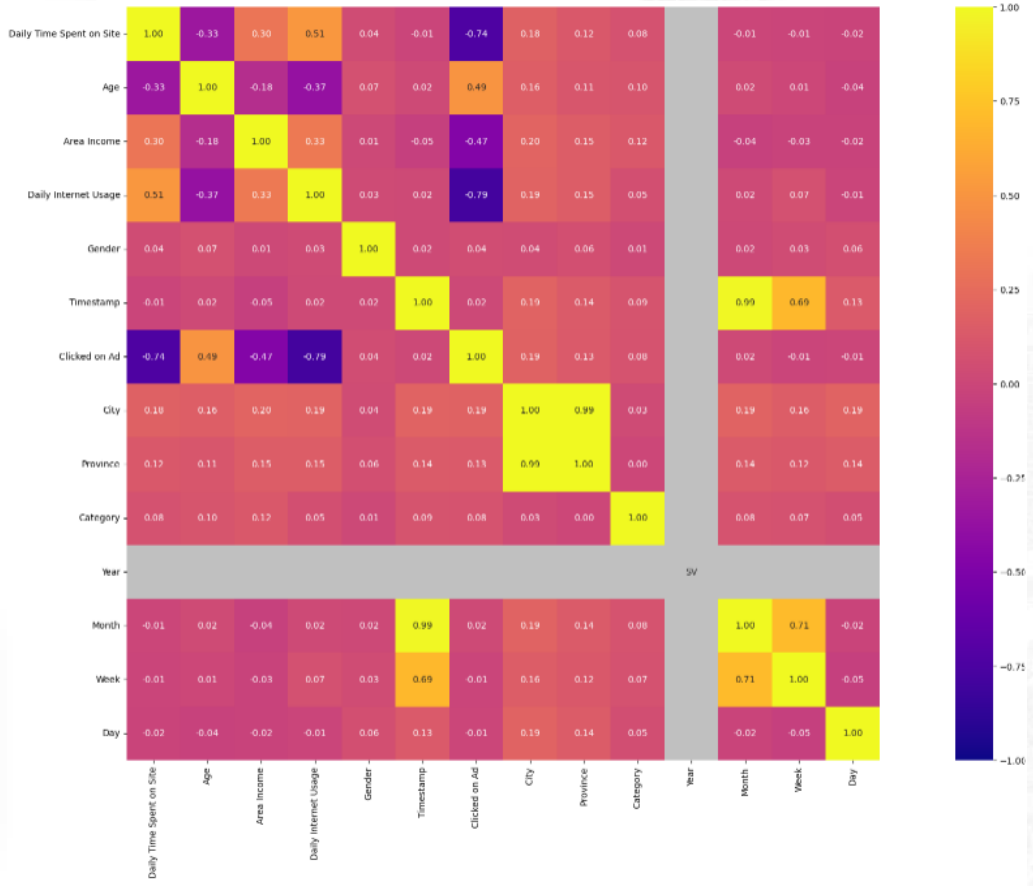
Data Cleaning & Preprocessing

FEATURE SELECTION

Berdasarkan value correlation, tidak ada korelasi antara variabel target (Clicked on Ad) dengan fitur Year, Month, Week, maupun Day.

Sehingga fitur yang digunakan dalam pemodelan tetap hanya **Clicked on Ad, Age, Daily Time Spent on Site, Area Income, dan Daily Internet Usage.**

	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Clicked on Ad
0	68.95	35	432837300.0	256.09	0
1	80.23	31	479092950.0	193.77	0
2	69.47	26	418501580.0	236.50	0
3	74.15	29	383643260.0	245.89	0
4	68.37	35	517229930.0	225.58	0



Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Cleaning & Preprocessing

SPLIT DATA

- Data train : Data test = 7 : 3
- Data train = 700 rows
- Data test = 300 rows

HANDLING OUTLIER

- Menggunakan metode IQR
- Handling outlier pada Data Train
- N_rows before handling = 700
- N_rows after handling = 689

FEATURE TRANSFORMATION

Normalization (MinMaxScaler)

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Modeling

Jenis model Machine Learning yang digunakan:

1. Decision Tree
2. AdaBoost
3. Logistic Regression
4. Logistic Regression with Elasticnet
5. KNNeighbors

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Modeling

Rakamin Academy

Non-Normalization Dataset

	Model	Accuracy	Precision	Recall	Time Elapsed
0	Decision Tree	0.953333	0.972973	0.935065	5.460287
1	AdaBoost	0.943333	0.941935	0.948052	3.199852
2	KNNeighbors	0.680000	0.750000	0.564935	2091.753802
3	Logistic Regression	0.486667	0.000000	0.000000	0.683123
4	Logistic Regression (Elasticnet)	0.486667	0.000000	0.000000	0.046180

- Model Decision Tree memiliki akurasi tertinggi.
- Model Logistic Regression baik tanpa maupun menggunakan elasticnet memiliki akurasi terendah, namun keduanya memiliki waktu eksekusi tercepat.
- Model KNNeighbors membutuhkan waktu eksekusi yang paling lama.
- Jadi, berdasarkan accuracy value, model yang memiliki performa terbaik adalah **Decision Tree** dan **AdaBoost**.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Modeling

Rakamin Academy

Normalization Dataset

	Model	Accuracy (Normalized)	Precision (Normalized)	Recall (Normalized)	Time Elapsed (Normalized)
0	KNNeighbors	0.956667	0.986207	0.928571	2735.483490
1	Decision Tree	0.953333	0.972973	0.935065	5.304255
2	AdaBoost	0.943333	0.941935	0.948052	2.343205
3	Logistic Regression	0.926667	1.000000	0.857143	0.121019
4	Logistic Regression (Elasticnet)	0.513333	0.513333	1.000000	0.047342

- Setelah dataset dinormalisasi, terjadi perubahan signifikan dalam kinerja model berbasis jarak.
- Model dengan performa terbaik berubah menjadi model **KNNeighbors**.
- Model dengan performa terburuk tetap Logistic Regression, baik tanpa maupun dengan elasticnet, meskipun kinerjanya meningkat secara signifikan.
- Waktu terlama yang diperlukan masih terjadi pada model KNNeighbors.
- Jadi, berdasarkan accuracy value setelah dataset dinormalisasi, model yang memiliki performa terbaik adalah **KNNeighbors** dan **Decision Tree**.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Comparing Non-Normalization and Normalization Dataset

	Model	Accuracy	Accuracy (Normalized)	Δ Accuracy	Precision	Precision (Normalized)	Δ Precision	Recall	Recall (Normalized)	Δ Recall	Time Elapsed	Time Elapsed (Normalized)	Δ Time Elapsed
0	KNNeighbors	0.680000	0.956667	0.276667	0.750000	0.986207	0.236207	0.564935	0.928571	0.363636	2091.753802	2735.483490	643.729689
1	Decision Tree	0.953333	0.953333	0.000000	0.972973	0.972973	0.000000	0.935065	0.935065	0.000000	5.460287	5.304255	-0.156031
2	AdaBoost	0.943333	0.943333	0.000000	0.941935	0.941935	0.000000	0.948052	0.948052	0.000000	3.199852	2.343205	-0.856647
3	Logistic Regression	0.486667	0.926667	0.440000	0.000000	1.000000	1.000000	0.000000	0.857143	0.857143	0.683123	0.121019	-0.562104
4	Logistic Regression (Elasticnet)	0.486667	0.513333	0.026667	0.000000	0.513333	0.513333	0.000000	1.000000	1.000000	0.046180	0.047342	0.001162

- Secara keseluruhan, semua model berkinerja lebih baik setelah dataset **dinormalisasi** berdasarkan evaluasi metrik dan juga waktu yang diperlukan lebih singkat.
- Model yang dipilih adalah model **Decision Tree** karena memiliki salah satu akurasi tertinggi dan proses komputasi yang cepat.

Untuk selengkapnya, dapat melihat jupyter notebook disini

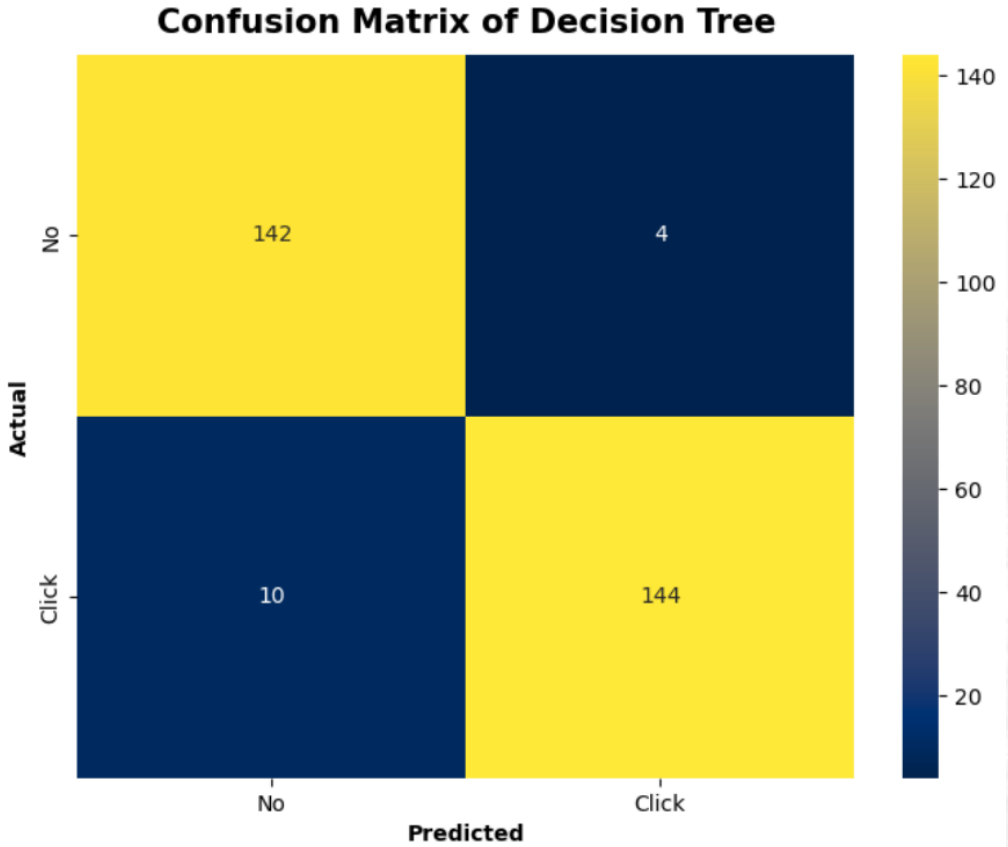
Confusion Matrix

Parameter terbaik pada model Decision Tree, antara lain:

- Maximal depth = 6
- Maximal features = sqrt
- Minimal samples leaf = 1
- Minimal samples split = 5

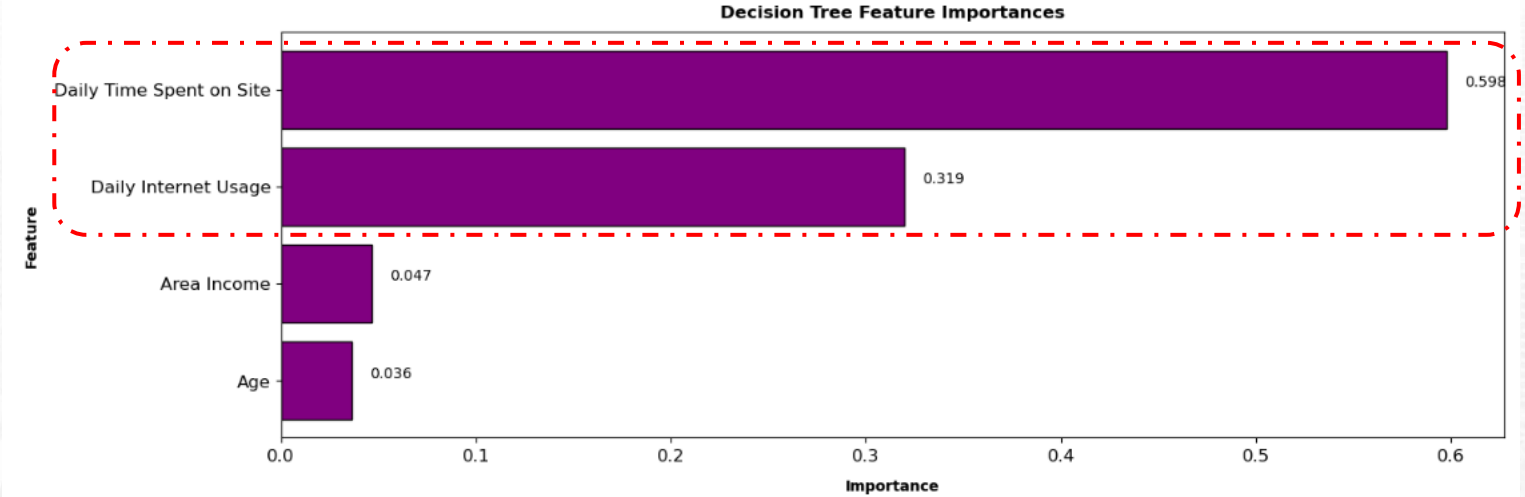
Dengan menggunakan hasil tuning hyperparameter untuk model Decision Tree, kita melatih model kembali untuk mendapatkan confusion matrix seperti yang ditunjukkan di atas, dengan hasil sebagai berikut:

- True Positive (TP): Diprediksi akan mengklik iklan dan ternyata benar sebanyak 144 kali.
- True Negative (TN): Diprediksi tidak akan mengklik iklan dan ternyata benar sebanyak 142 kali.
- False Positive (FP): Diprediksi akan mengklik iklan dan ternyata salah sebanyak 4 kali.
- False Negative (FN): Diprediksi tidak akan mengklik iklan dan ternyata salah sebanyak 10 kali.



Untuk selengkapnya, dapat melihat jupyter notebook disini

Features Importance



Waktu yang dihabiskan setiap hari di situs web (**Daily Time Spent on Site**) adalah fitur yang **paling penting**, diikuti oleh penggunaan internet harian (**Daily Internet Usage**) di tempat kedua yang menentukan apakah pengguna mengklik iklan atau tidak.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Business Recommendation & Simulation

Business Recommendation

Berdasarkan insight dari Exploratory Data Analysis (EDA) dan feature importances, berikut adalah rekomendasi bisnis yang dapat diberikan:

Optimalisasi Konten

Karena semakin tinggi **Daily Time Spent on Site** dan **Daily Internet Usage**, semakin kecil kemungkinan pengguna akan mengklik iklan, maka kita perlu membuat konten iklan yang menarik dan relevan untuk pengguna target. Pastikan pesan dan visual iklan sesuai dengan minat dan kebutuhan pengguna. Yaitu dengan cara:

- Menggunakan gambar dan video yang menarik perhatian.
- Membuat headline yang kuat dan relevan.
- Menyediakan konten yang menawarkan nilai tambah bagi pengguna, seperti informasi yang berguna atau penawaran khusus.

Strategi Harga Terarah

Karena semakin rendah **Area Income**, semakin besar kemungkinan pengguna akan mengklik iklan, kita dapat menerapkan strategi harga yang sesuai dengan tingkat pendapatan target audiens. Yaitu dengan cara:

- Membuat tingkat harga khusus yang lebih terjangkau.
- Menawarkan diskon atau bundling produk/jasa.
- Mengembangkan dan mempromosikan produk atau jasa yang terjangkau untuk pengguna dengan pendapatan rendah.

Kampanye Pemasaran Berbasis Usia

Karena semakin tua (**Age**) pengguna, semakin besar kemungkinan mereka akan mengklik iklan, kita dapat mengembangkan kampanye pemasaran yang secara khusus dirancang untuk menarik demografi yang lebih tua. Yaitu dengan cara:

- Menciptakan pesan, visual, dan penawaran yang sesuai dengan preferensi dan minat pengguna yang lebih tua.
- Menggunakan saluran pemasaran yang lebih banyak digunakan oleh demografi ini, seperti email atau media sosial tertentu.
- Menyoroti manfaat produk/jasa yang lebih relevan dengan kebutuhan pengguna yang lebih tua.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Business Recommendation & Simulation

Business Simulation

Asumsi
Cost per Mille (CPM) = Rp100,000
Revenue per Ad Clicked = Rp3,000

Simulasi bisnis tanpa menggunakan hasil dari model machine learning

- **Jumlah pengguna yang ditargetkan**
  User = 1,000
- **Click-Through Rate (CTR):**
  CTR = 500/1,000 = 0.5
- **Total Cost**:
  CPM = Rp100,000
- **Total Revenue**:
  Total Revenue = CTR x Jumlah pengguna yang ditargetkan x Revenue per Ad Clicked = 0.5 x 1,000 x 3,000 = Rp1,500,000
- **Total Profit**:
  Total Profit = Total Revenue - Total Cost = Rp1,500,000 - Rp100,000 = Rp1,400,000

Simulasi bisnis dengan menggunakan hasil dari model machine learning

- **Jumlah pengguna yang ditargetkan**
  User = 1,000
- **Click-Through Rate (CTR):**
  Accuracy = 0.95
- **Total Cost**:
  CPM = Rp100,000
- **Total Revenue**:
  Total Revenue = CTR x Jumlah pengguna yang ditargetkan x Revenue per Ad Clicked = 0.95 x 1,000 x 3,000 = Rp2,850,000
- **Total Profit**:
  Total Profit = Total Revenue - Total Cost = Rp2,850,000 - Rp100,000 = Rp2,750,000

Untuk selengkapnya, dapat melihat jupyter notebook disini

Business Recommendation & Simulation

Kesimpulan

	Tanpa ML	Dengan ML
Jumlah Pengguna yang Ditargetkan	1000	1000
Revenue per Ad Clicked	Rp3,000	Rp3,000
CPM	Rp100,000	Rp100,000
CTR	0.5	0.95
Total Revenue	Rp1,500,000	Rp2,850,000
Total Profit	Rp1,400,000	Rp2,750,000

Dari hasil di atas, terlihat bahwa setelah **menggunakan model machine learning, performa iklan meningkat**. Click-Through Rate (CTR) meningkat dari 50% menjadi 95% dan total keuntungan meningkat sebesar 96,4% dari Rp1,400,000 menjadi Rp2,750,000.

Untuk selengkapnya, dapat melihat jupyter notebook disini