Predict Clicked Ads Customer Classification by using Machine Learning

Created by:
Yehezkiel Novianto Aryasena
hezkyaryasena@gmail.com
www.linkedin.com/in/yehezkielnov

"**Saya Yehezkiel Novianto Aryasena**. Saya merupakan fresh graduate dari Institut Teknologi Sepuluh Nopember Surabaya. Saat menjalani masa perkuliahan, saya memiliki pengalaman organisasi dan kepanitiaan yang membuat saya mampu bekerja mandiri maupun dalam tim. Saya memiliki ketertarikan untuk mempelajari hal baru terutama pada bidang data science dan saat ini sedang mendalami pengetahuan saya dalam hal tersebut dengan mengikuti course yang diselenggarakan oleh Rakamin. "

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 jupyter notebook disini
https://colab.research.google.com/drive/15SHMTPKUUGJg7FojdHdC1VWMcxFzkSyG?usp=sharing

EDA I
Statisctic Descriptive

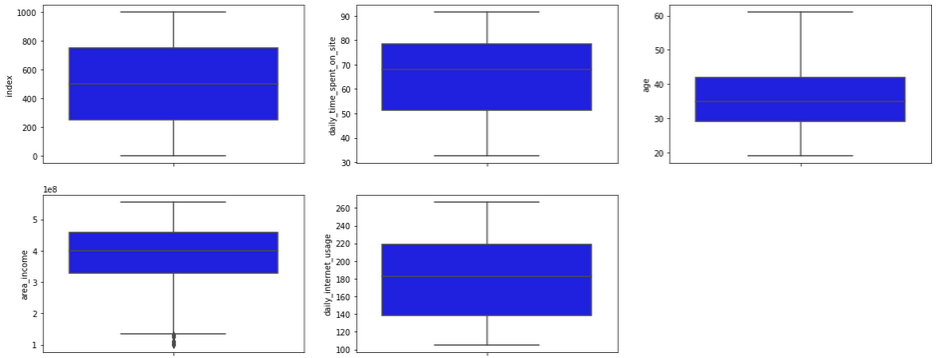
- Terlihat bahwa tipe data tidak ada yang aneh atau semuanya sesuai dengan feature (Terdapat kesalahan nama feature diawal Male yang sudah dibah menjadi gender)
- Distribusi data juga terlihat cukup normal

```
Data columns (total 11 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   index                     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
```

	index	daily_time_spent_on_site	age	area_income	daily_internet_usage
count	1000.000000	987.000000	1000.000000	9.870000e+02	989.000000
mean	499.500000	64.929524	36.009000	3.848647e+08	179.863620
std	288.819436	15.844699	8.785562	9.407999e+07	43.870142
min	0.000000	32.600000	19.000000	9.797550e+07	104.780000
25%	249.750000	51.270000	29.000000	3.286330e+08	138.710000
50%	499.500000	68.110000	35.000000	3.990683e+08	182.650000
75%	749.250000	78.460000	42.000000	4.583554e+08	218.790000
max	999.000000	91.430000	61.000000	5.563936e+08	267.010000

EDA II
Univariate Analysis (BoxPlot)

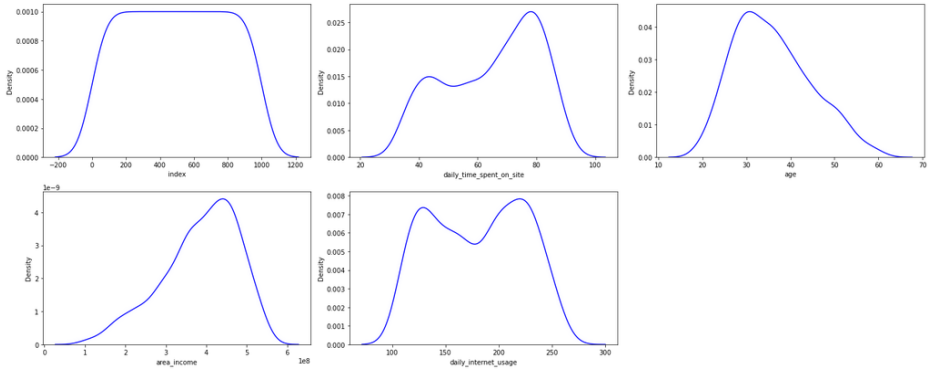
Berdasarkan visualisasi boxplot feature numerikal yang ada dalam dataset dapat diketahui bahwa hampir semua feature terbebas dari outlier. Hanya terdapat satu feature yang memiliki outlier yakni area income. Untuk feature area_income ini akan diberi penanganan dengan metode IQR (Interquartile Range).



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

EDA III
Univariate Analysis (DistPlot)

Berdasarkan visualisasi distribusi plot disamping, terlihat bahwa semua feature numerikal yang ada memiliki distribusi yang cukup normal (mendekati normal) atau tidak ada yang terlalu terdistribusi positif atau negatif yang ekstrem (Right/left Skewness).

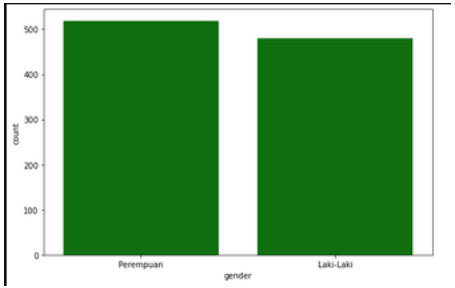


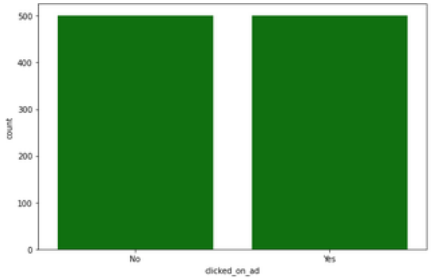
Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

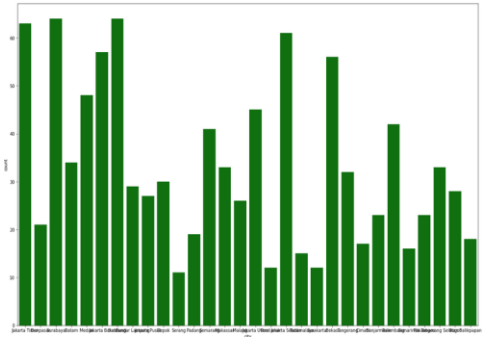
EDA IV
Univariate Analysis (BarPlot)

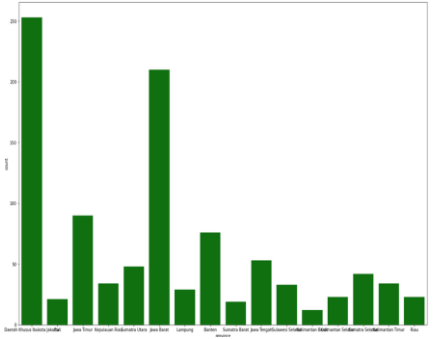
Dari visualisasi bar plot disamping, beberapa insight yang didapat dari feature kategorikal yang ada pada dataset adalah sebagai berikut.

1. Mayoritas customer berjenis kelamin perempuan
2. Tidak terdapat perbedaan antara jumlah customer yang mengklik pada iklan dan yang tidak mengklik pada iklan
3. Mayoritas customer berdomisili di kota Surabaya namun apabila dilihat dari tingkat provinsi maka domisili customer paling dominan adalah DKI Jakarta







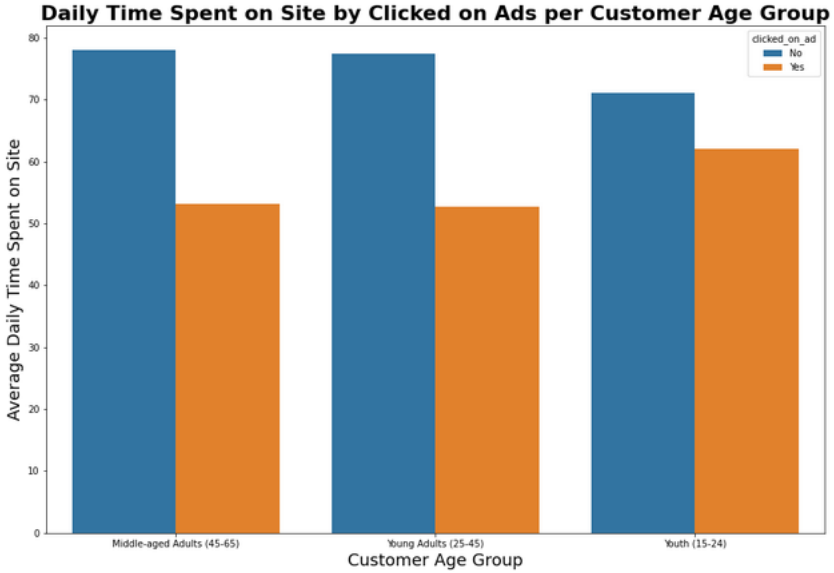


EDA V
Bivariate Analysis (1)

Menambah kolom customer_age_group untuk memudahkan EDA dan klasifikasi kedepannya.

Insight:

1. Persentase customer yang mengklik iklan lebih banyak pada customer dengan rentang umur 15-24 tahun (Youth)
2. Semakin lama customer menghabiskan waktunya pada situs maka tingkat kemungkinan customer akan mengklik pada iklan juga akan semakin kecil

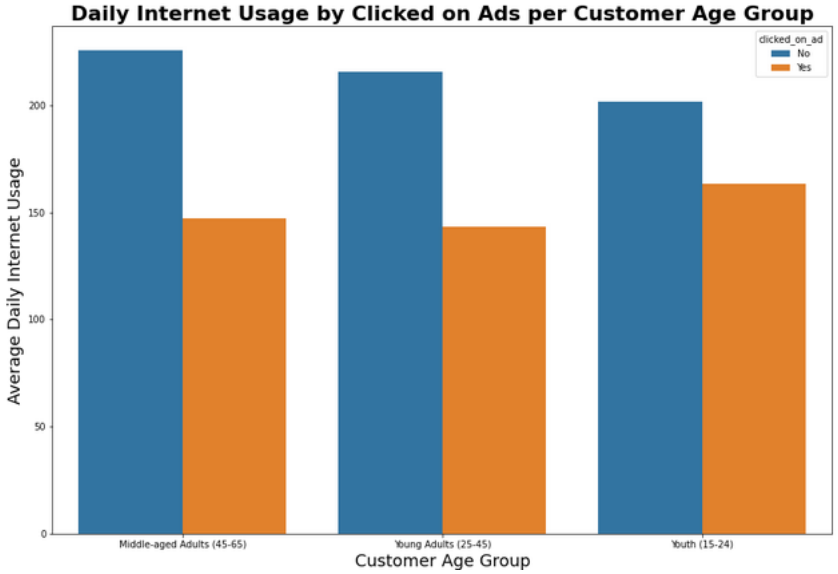


Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

EDA V
Bivariate Analysis (2)

Insight:

1. Dari dataset ini kita mendapati bahwa customer yang lebih banyak menghabiskan waktunya menggunakan internet akan berpeluang lebih tinggi untuk tidak mengklik pada iklan
2. Persentase customer yang menklik iklan lebih banyak pada customer dengan rentang umur 15-24 tahun (Youth)



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

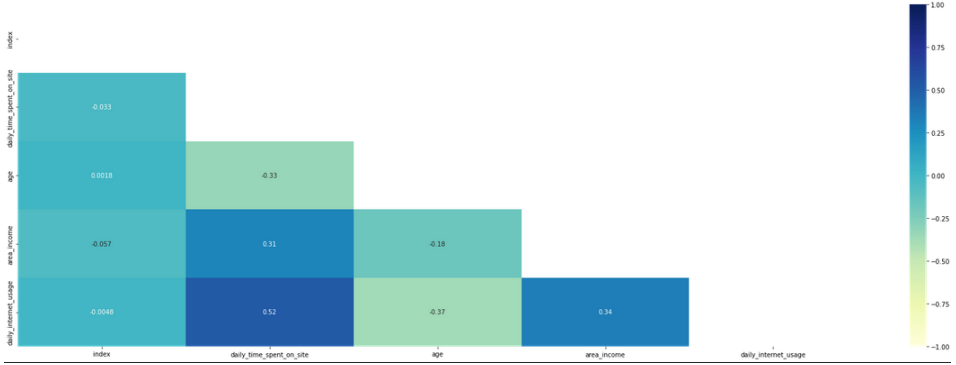
EDA VI
Multivariate Analysis (HeatMap)

Insight:

1. Tidak terdapat feature yang memiliki korelasi yang kuat atau positif dengan nilai korelasi diatas 0.7.

2. Nilai korelasi feature yang paling mendekati dengan 0.7 yaitu berada diantara feature daily_time_spent_on_site dengan feature daily_internet_usage dengan nilai korelasi 0.5.

3. Hal ini menandakan bahwa semakin lama customer menghabiskan waktunya pada website maka semakin lama pula penggunaan internet customer tersebut.



Data Cleaning & Preprocessing

- Pada tahap ***cleaning data***, tunjukan ***null*** atau ***missing value*** serta ***duplicated value*** pada dataset, serta cara penyelesaiannya.

- Tulislah pula proses **extract datetime data** sebelum dilakukan model machine learning.

- Tunjukan **Split Data** sebelum melakukan model machine learning

- Tulislah proses **feature encoding** pada tahap ini (gunakan get_dummy)

- **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 jupyter notebook disini
https://colab.research.google.com/drive/15SHMTPKUUGJg7FojdHdC1VWMcxFzkSyG?usp=sharing

Data Pre-Processing I (Missing Value)

Treatment:

1. Untuk feature dengan persentase missing value diatas 0% dan dibawah 1% maka row yang memiliki missing value tersebut akan didrop
2. Untuk feature dengan persentase missing value diatas 1% akan diisi dengan metode interpolation data. Hal ini dikarenakan apabila diisi nilai aggregate akan mengubah sebaran data

1. Bersihkan dataset dari missing value dan duplicated value (Hint: Perhatikan null/missing value, data yang kosong dapat dibuang atau diisi dengan nilai yang tidak mengubah sebaran data)

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

Rakamin Academy

Missing Value

```
df_clean.isnull().mean() * 100
```

```
index                      0.0
daily_time_spent_on_site   1.3
age                        0.0
area_income                1.3
daily_internet_usage       1.1
gender                     0.3
timestamp                  0.0
clicked_on_ad              0.0
city                       0.0
province                   0.0
category                   0.0
customer_age_group         0.0
dtype: float64
```

```
df_clean.interpolate(method ='linear', limit_direction ='both',inplace=True)
```

```
df_clean.isnull().mean() * 100
```

```
index                      0.0
daily_time_spent_on_site   0.0
age                        0.0
area_income                0.0
daily_internet_usage       0.0
gender                     0.0
timestamp                  0.0
clicked_on_ad              0.0
city                       0.0
province                   0.0
category                   0.0
customer_age_group         0.0
dtype: float64
```

Data Pre-Processing II (Duplicated Value)

Tidak ditemukan data yang duplikat

Duplicated Value

```
df_clean.duplicated().sum()
```

```
0
```

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

Data Pre-Processing III (Extract Timestamp)

Treatment:

1. Menggunakan bantuan pd.to_datetime untuk mengubah timestamp dari object menjadi datetime64
2. Mengekstrak baik tahun, bulan, pekan, dan hari mengunakan pd.DatetimeIndex[df[col]].year/moth/week/day

```
df_clean['timestamp']=pd.to_datetime(df['timestamp'])
df_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 997 entries, 0 to 999
Data columns (total 12 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   index                     997 non-null    int64
 1   daily_time_spent_on_site  997 non-null    float64
 2   age                       997 non-null    int64
 3   area_income               997 non-null    float64
 4   daily_internet_usage      997 non-null    float64
 5   gender                    997 non-null    object
 6   timestamp                 997 non-null    datetime64[ns]
 7   clicked_on_ad             997 non-null    object
 8   city                      997 non-null    object
 9   province                  997 non-null    object
 10  category                  997 non-null    object
 11  customer_age_group        997 non-null    object
dtypes: datetime64[ns](1), float64(3), int64(2), object(6)
memory usage: 101.3+ KB
```

```
df_clean['timestamp']
```

```
0     2016-03-27 00:53:00
1     2016-04-04 01:39:00
2     2016-03-13 20:35:00
3     2016-01-10 02:31:00
4     2016-06-03 03:36:00
              ...
995   2016-02-11 21:49:00
996   2016-04-22 02:07:00
997   2016-02-01 17:24:00
998   2016-03-24 02:35:00
999   2016-06-03 21:43:00
Name: timestamp, Length: 997, dtype: datetime64[ns]
```

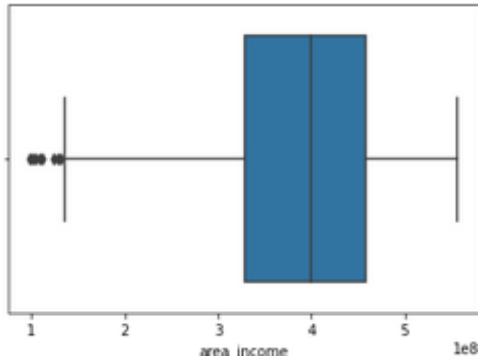
tahun	bulan	pekan	hari
2016	3	12	27
2016	4	14	4
2016	3	10	13
2016	1	1	10
2016	6	22	3
...	...	...	...
2016	2	6	11
2016	4	16	22
2016	2	5	1
2016	3	12	24
2016	6	22	3

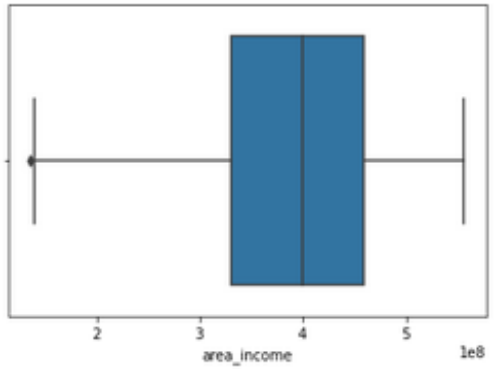
Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Data Pre-Processing IV (Handle Outlier)

Treatment:

1. Menggunakan metode IQR pada feature yang masih memiliki outlier yakni area_income





Data Pre-Processing V (Feature Encoding)

Treatment:

1. Untuk feature gender & clicked on ads = Label Encoding karena urutan yang cukup jelas
2. Untuk feature city, province, category, customer_age_group = OHE
3. Men-drop kolom asli dari OHE

```
mapping_gender = {
    'Perempuan' : 0,
    'Laki-Laki' : 1
}
df_clean['gender']=df_clean['gender'].map(mapping_gender)

mapping_clicked = {
    'No' : 0,
    'Yes' : 1
}
df_clean['clicked_on_ad']=df_clean['clicked_on_ad'].map(mapping_clicked)
```

```
kota=pd.get_dummies(df_clean['city'], prefix='kota')
provinsi=pd.get_dummies(df_clean['province'], prefix='provinsi')
kategori=pd.get_dummies(df_clean['category'], prefix='category')
umur=pd.get_dummies(df_clean['customer_age_group'], prefix='golumur')
```

```
df_clean=pd.concat([df_clean,kota,provinsi,kategori,umur], axis=1)
```

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

Data Pre-Processing VI (Split Target & Feature)

Treatment:

1. Mendefinisikan y sebagai target dan x sebagai feature dimana target adalah kolom clicked_on_ad dan feature adalah kolom selain target.

```
x = df_real.drop(columns=['clicked_on_ad'], axis=1)
y = df_real['clicked_on_ad']
```

```
x = df_real_norm.drop(columns=['clicked_on_ad'], axis=1)
y = df_real_norm['clicked_on_ad']
```

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

Data Modeling

- Tulislah proses model machine learning terdiri dari
  a. **hasil *experiment 1*** (sebelum normalisasi/standardisasi),
  b. **hasil *experiment 2*** (setelah normalisasi/standardisasi).
  c. hasil tabel **confusion matrix** dari model tersebut.
  d. Daftar ***Feature Important***.

- Tulislah hasil interpretasi dari model tersebut

Untuk selengkapnya, dapat melihat jupyter notebook disini
https://colab.research.google.com/drive/15SHMTPKUUGJg7FojdHdC1VWMcxFzkSyG?usp=sharing

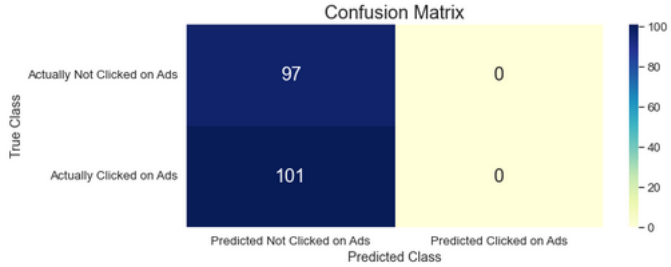
Hasil Modelling I (Tanpa Normalisasi)

1. Menggunakan 3 jenis algoritma permodelan untuk prediksi klasifikasi yakni Logistic Regression, K-Nearest Neighborhood (KNN), dan Decision Tree(DT).
2. Hasil berupa confusion matrix yang ada pada gambar disamping.

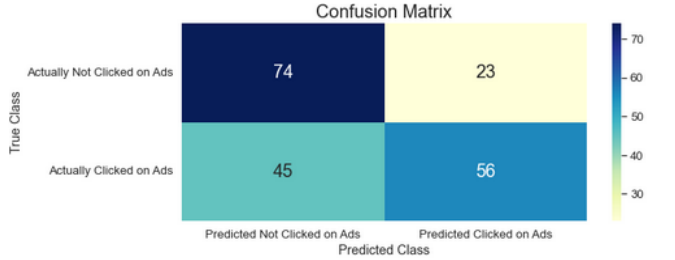
Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Rakamin Academy

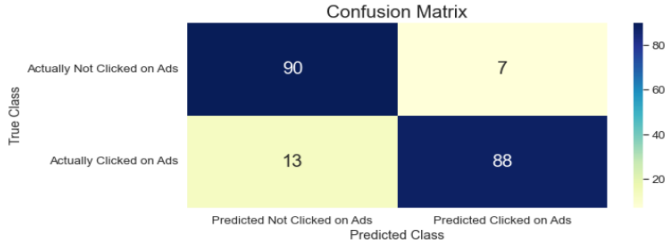
Logistic Regression



KNN



DT



Hasil Modelling I (Tanpa Normalisasi)

Intepretasi:

1. Karena konteksnya disini kita perlu memprediksi apakah customer akan mengklik pada ad/iklan atau tidak maka evaluasi yang diperlukan adalah accuracy.
2. Pada ketiga model nilai accuracy terbesar diperoleh pada model dengan algoritma jenis Decision Tree

	Log Res	KNN	Decision Tree
Accuracy	0.489899	0.656566	0.898990
Recall	0.000000	0.554455	0.871287
Precision	0.000000	0.708861	0.926316
F1 Score	0.000000	0.622222	0.897959

	Train	Test
Log Res	0.506953	0.489899
KNN	0.750948	0.656566
Decision Tree	1.000000	0.898990

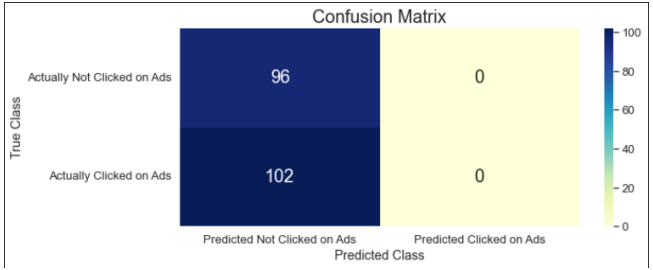
Hasil Modelling II (Dengan Normalisasi)

1. Menggunakan 3 jenis algoritma permodelan untuk prediksi klasifikasi yakni Logistic Regression, K-Nearest Neighborhood (KNN), dan Decision Tree(DT).
2. Hasil berupa confusion matrix yang ada pada gambar disamping.

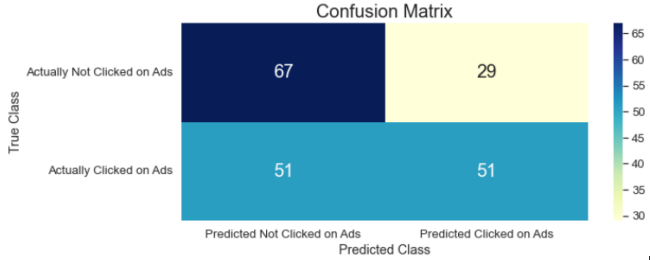
Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Rakamin Academy

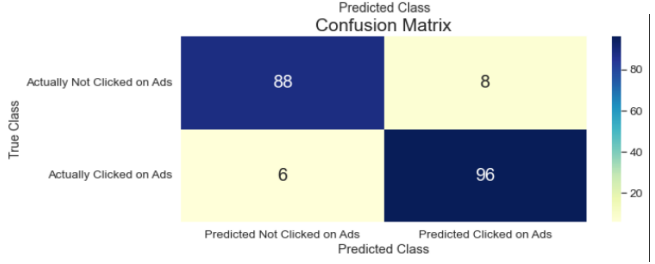
Logistic Regression



KNN



DT



Hasil Modelling II (Dengan Normalisasi)

Intepretasi:

1. Setelah dilakukan normalisasi data, nilai evaluasi pada confusion matrix cenderung membaik/lebih tinggi
2. Terdapat peningkatan nilai accuracy pada model Decision Tree namun nilai accuracy pada model KNN mengalami penurunan.
3. Dengan normalisasi data sekalipun nilai accuracy terbaik tetap pada model yang menggunakan algoritma Decision Tree
4. Pencarian Feature Importance akan menggunakan model Decision Tree

	Log Res	KNN	Decision Tree
Accuracy	0.484848	0.59596	0.929293
Recall	0.000000	0.50000	0.941176
Precision	0.000000	0.63750	0.923077
F1 Score	0.000000	0.56044	0.932039

	Train	Test
Log Res	0.508217	0.484848
KNN	0.768647	0.595960
Decision Tree	1.000000	0.929293

Business Recommendation & Simulation

- Tampilkan ***Feature Important*** dari hasil model machine learning

- Tulislah **rekomendasi bisnis** berdasarkan EDA dan Feature Important

- Tulislah sebuah simulasi perusahaan dalam marketing yang menunjukan ***cost, revenue, dan profit*** **sebelum dan setelah menggunakan model machine learning**. Tunjukan perbedaan dari kedua simulasi tersebut.

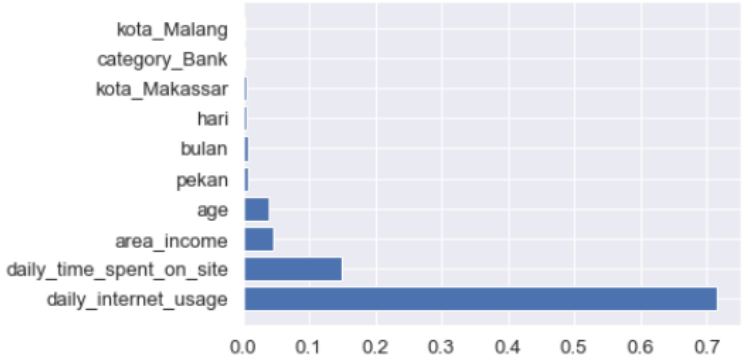
- Tulislah pula **simpulan** yang didapat dari proses tersebut

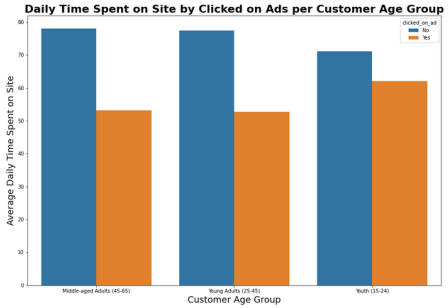
Untuk selengkapnya, dapat melihat jupyter notebook disini
https://colab.research.google.com/drive/15SHMTPKUUGJg7FojdHdC1VWMcxFzkSyG?usp=sharing

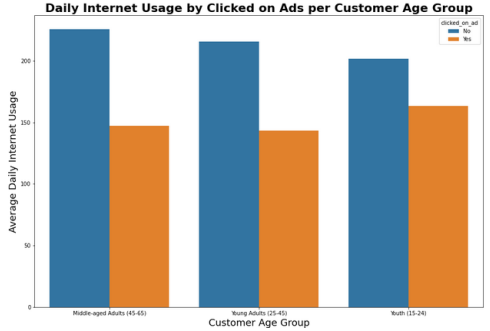
Feature Importance & Business Recommendation

Dengan model Decision Tree didapati 10 feature paling berpengaruh dapat dilihat pada visualisasi disamping. Tiga feature teratas yaitu daily_internet_usage, daily_time_spent_on_website, serta area_income.

Hal ini menunjukkan bahwa perusahaan dapat fokus mengambil Langkah bisnis kedepannya dengan memperhatikan feature feature tersebut.





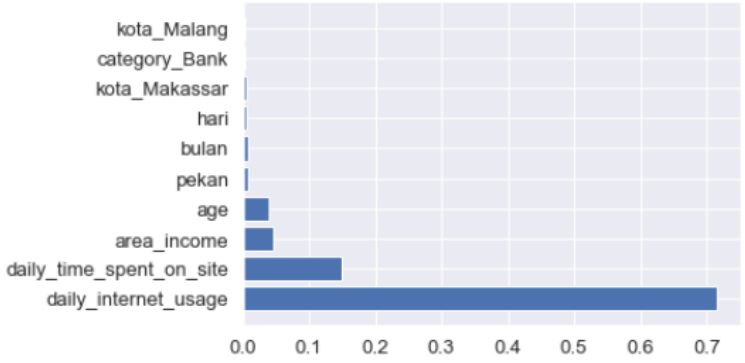


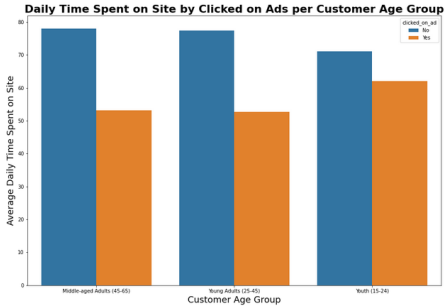
Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

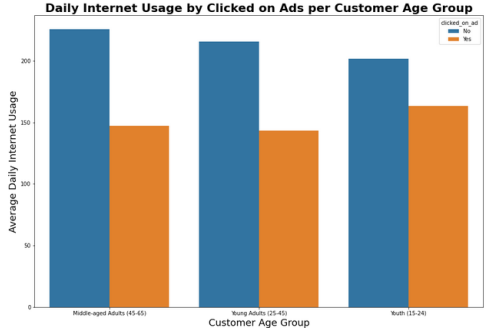
Feature Importance & Business Recommendation

Dengan Feature Importance dan EDA yang sudah dilakukan pada tahap sebelumnya maka rekomendasi bisnis yang dapat dilakukan pada perusahaan yaitu:

1. Mengurangi waktu yang dibutuhkan oleh Customer untuk berselancar pada website dengan memperbaiki UI/UX agar dapat mengurangi kemungkinan Customer tidak mengklik iklan
2. Meningkatkan kualitas iklan agar mendekati *personal behaviour* customer
3. Menyesuaikan Iklan kepada tingkat pendapatan Customer (Seperti promo, discount, loyalty club)







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

Bussiness Impact Simulation

Dengan asumsi bahwa perusahaan akan mendapatkan *revenue* sebesar 1,000 per customer yang mengklik iklan dengan biaya operasional 10% revenue, apabila terdapat 100 customer yang diprediksi mengklik iklan maka apabila tanpa model:
Maka hanya ada 52 orang yang diterima (df['clicked_on_ad'].sample(100)= 52% Yes) berarti:
52 x 1,000 = 52,000
52,000 - (52,000 x 10/100)= 46,800 total keuntungan untuk perusahaan
Namun dengan model:
Perusahaan akan mendapatkan revenue dari 92 orang (Accuracy 92%) :
92x1,000 = 92,000
92,000 – (92,000 x 10/100)= 82,800 total keuntungan untuk perusahaan

Conclusion

Perusahaan dapat menerapkan langkah bisnis yang disarankan pada bisnis recommendation supaya dapat meningkatkan kemungkinan customer dalam mengklik ad serta menurunkan kemungkinan customer untuk tidak mengklik ad.

Apabila perusahaan tetap belum mendapatkan profit maka setidaknya perusahaan sudah dapat memprediksi besaran expenditure dengan model

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