# Predict Clicked Ads Customer Classification by using Machine Learning

**Created by:**
**Muhammad Oky Hariawan**
okyhariawan@gmail.com
https://www.linkedin.com/in/okyhariawan/

"A bachelor's degree in business management with 2 year's expertise in Customer Service and Workforce Management. A data-driven & tech-savvy person who has huge interest in data analytics and currently learning data science. Skilled at analyse data, keen attention to detail, ability to effectively prioritize and execute tasks in a high-pressure environment. Proficient using python, SQL, Microsoft Power BI, Excel VBA, & other statistical tools."

**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 ”

# EXPLORATORY DATA ANALYSIS

| | Feature | Data Type | Null | Null (%) | Unique | Unique Sample |
|---|---|---|---|---|---|---|
| 0 | Unnamed: 0 | int64 | 0 | 0.0 | 1000 | [0, 1, 2, 3, 4] |
| 1 | Daily Time Spent on Site | float64 | 13 | 1.3 | 890 | [68.95, 80.23, 69.47, 74.15, 68.37] |
| 2 | Age | int64 | 0 | 0.0 | 43 | [35, 31, 26, 29, 23] |
| 3 | Area Income | float64 | 13 | 1.3 | 987 | [432837300.0, 479092950.00000006, 418501580.0,... |
| 4 | Daily Internet Usage | float64 | 11 | 1.1 | 955 | [256.09, 193.77, 236.5, 245.89, 225.58] |
| 5 | Male | object | 3 | 0.3 | 2 | [Perempuan, Laki-Laki, nan] |
| 6 | Timestamp | object | 0 | 0.0 | 997 | [3/27/2016 0:53, 4/4/2016 1:39, 3/13/2016 20:3... |
| 7 | Clicked on Ad | object | 0 | 0.0 | 2 | [No, Yes] |
| 8 | city | object | 0 | 0.0 | 30 | [Jakarta Timur, Denpasar, Surabaya, Batam, Medan] |
| 9 | province | object | 0 | 0.0 | 16 | [Daerah Khusus Ibukota Jakarta, Bali, Jawa Tim... |
| 10 | category | object | 0 | 0.0 | 10 | [Furniture, Food, Electronic, House, Finance] |

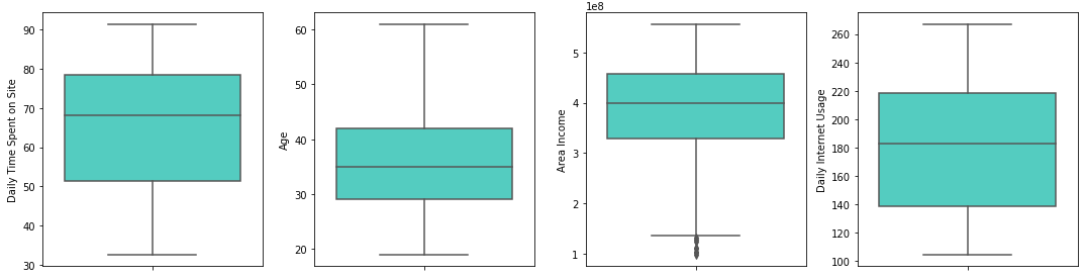
**Insight**:

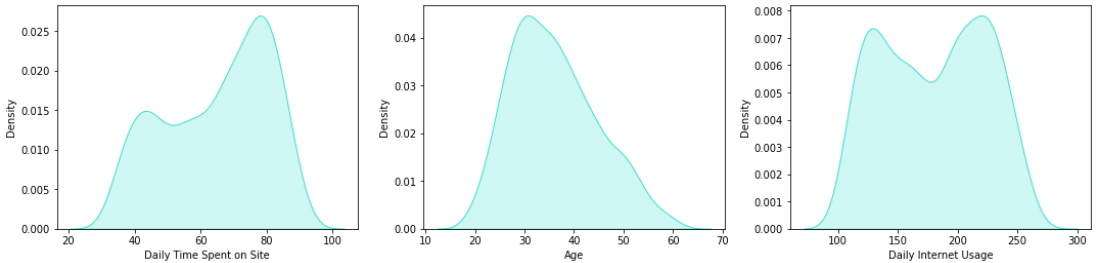
- Data ini terdiri dari 1000 baris dan 11 kolom
- Tidak terdapat data yang duplicated
- Terdapat missing value pada kolom:
  - Daily Time Spent on Site
  - Area Income
  - Daily Internet Usage
  - Male
- Clicked on Ad akan menjadi target feature
- Data ini memiliki tipe data int64 (2), float64 (3), object(6)

Untuk selengkapnya, dapat melihat jupyter notebook disini

# EXPLORATORY DATA ANALYSIS

## UNIVARIATE ANALYSIS



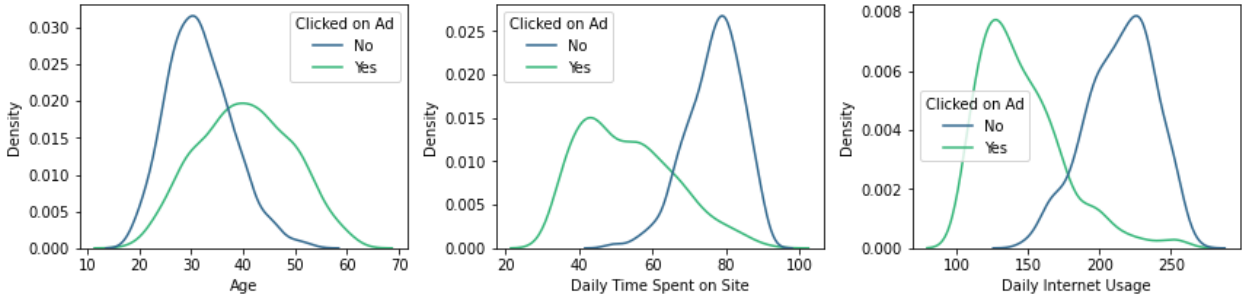




**Insight**:

- Terdapat adanya outlier pada feature Area Income
- Feature Daily Time Spent on Site & Daily Internet Usage memiliki distribusi bimodal dengan 2 modus
- Feature Age mendekati distribusi normal dengan umur rata-rata 36 tahun

# EXPLORATORY DATA ANALYSIS

## BIVARIATE ANALYSIS

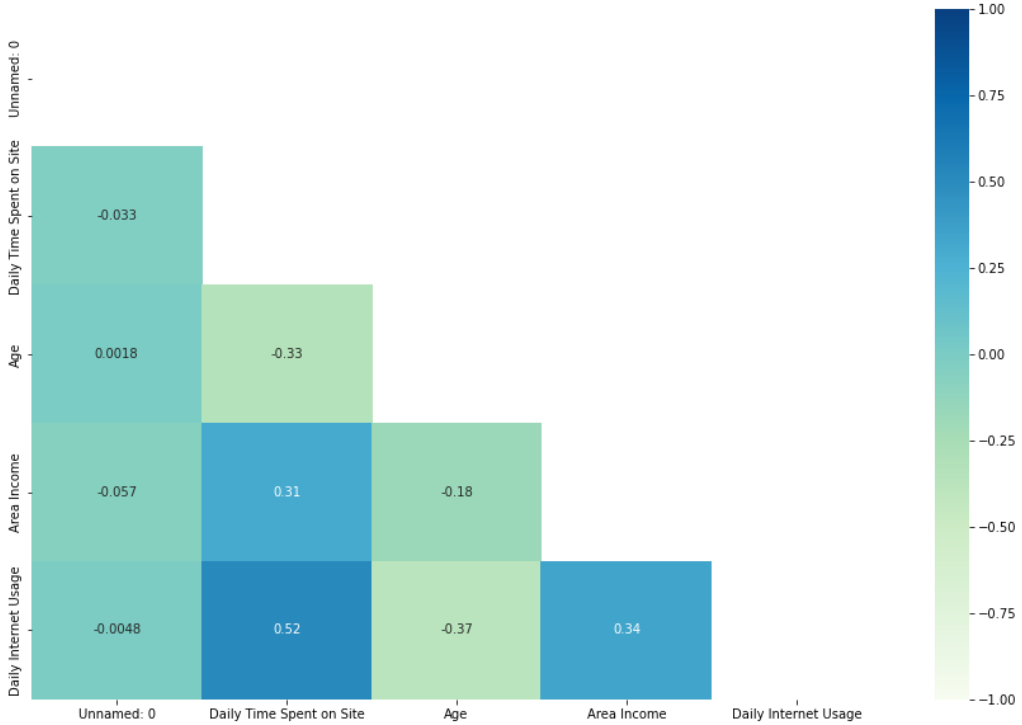



**Insight**:

- Customer yang mengklik iklan memiliki usia yang lebih tua dibandingkan dengan customer yang tidak mengklik iklan dan Customer lebih muda juga terlihat lebih menghindari klik iklan.
- Customer yang menghabiskan waktu lebih sedikit setiap harinya cenderung untuk mengklik iklan.
- Customer yang lebih sedikit menggunakan internet tampak cenderung mengklik iklan dan customer dengan penggunaan internet yang lebih banyak cenderung untuk tidak mengklik iklan

# EXPLORATORY DATA ANALYSIS

## MULTIVARIATE ANALYSIS




**Insight**:

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

# DATA PREPROCESSING

## Handle Null Value

- Untuk numerical feature (Daily Time Spent on Site, Daily Internet Usage, Area Income) akan di imputasi meggunakan Median
- Untuk categorical feature (Male) akan di imputasi menggunakan Modus

| | feature | missing_value | percentage |
|---|---|---|---|
| 0 | Daily Time Spent on Site | 13 | 1.3 |
| 1 | Area Income | 13 | 1.3 |
| 2 | Daily Internet Usage | 11 | 1.1 |
| 3 | Male | 3 | 0.3 |

## Handle Duplicated Value

- Tidak ditemukan adanya data duplicated dalam dataset ini

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

```
0
```

## Feature Engineering

- Convert feature Timestamp menjadi beberapa kolom baru (year, month, week, day)
- Mengubah nama feature Male menjadi Gender demi mempermudah interpretasi

```
df_clean['Timestamp'] = pd.to_datetime(df_clean['Timestamp'])
# Create a new feature called Year
df_clean['Year'] = df_clean.Timestamp.dt.year
# Creates a new column called Month
df_clean['Month'] = df_clean.Timestamp.dt.month
# Creates a new column called Week
df_clean['Week'] =df_clean.Timestamp.dt.dayofweek
# Creates a new column called Day
df_clean['Day'] = df_clean.Timestamp.dt.day

# rename feature Male to Gender
df_clean.rename(columns = {"Male" : "Gender"}, inplace = True)
```

# DATA PREPROCESSING

## Feature Encoding

- Label encoding untuk Gender & Clicked on Ad
- One hot encoding untuk city, province, & category

```
# label encoder
mapping_male = {
    'Laki-Laki' : 0,
    'Perempuan' : 1}

mapping_clicked = {
    'No' : 0,
    'Yes' : 1}

df_clean['gender_mapped'] = df_clean['Gender'].map(mapping_male)
df_clean['adclicked_mapped'] = df_clean['Clicked on Ad'].map(mapping_clicked)
```

```
# handle dengan one hot encoding
for cat in ['city', 'province', 'category']:
  onehots = pd.get_dummies(df_clean[cat], prefix=cat)
  df_clean = df_clean.join(onehots)
```

## Feature Selection

- Menghapus feature yang tidak relevant & redundant yaitu:
- Unnamed: 0, Timestamp, Clicked on Ad, city, province, category, Gender

## Split Target & Feature

- Feature mapping_clicked akan menjadi target feature sedangnkan lainnya akan menjadi features

```
X = df_model.drop(labels=['adclicked_mapped'],axis=1)
y = df_model[['adclicked_mapped']]
```

# DATA MODELING

## MODELING TANPA NORMALIZATION

| Model Name | Accuracy | Recall | Precision | Duration |
|---|---|---|---|---|
| K-Nearest Neighbors | 0.673333 | 0.640000 | 0.685714 | 0.009152 |
| XgBoost | 0.960000 | 0.966667 | 0.953947 | 0.418589 |
| Random Forest | 0.963333 | 0.960000 | 0.966443 | 0.473676 |
| Gradient Boosting | 0.960000 | 0.960000 | 0.960000 | 0.646277 |
| LightGBM | 0.973333 | 0.966667 | 0.979730 | 0.219908 |

Dari hasil modelling tanpa normalisasi terlihat jika LightGBM adalah model terbaik dengan nilai akurasi yang paling tinggi dan dengan durasi yang lebih singkat. Namun, hasil model diatas kemungkinan masih dapat lebih tinggi lagi setelah kita melakukan normalisasi

# DATA MODELING

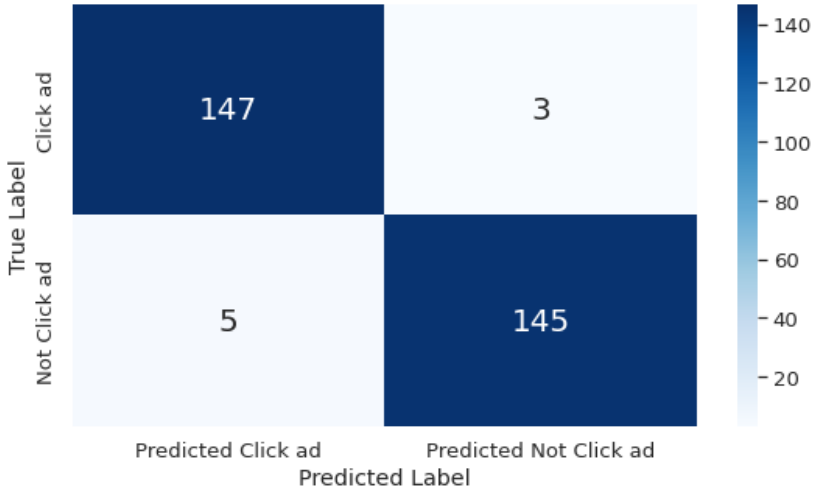
## MODELING DENGAN NORMALIZATION

| Model Name | Accuracy | Recall | Precision | Duration |
|---|---|---|---|---|
| K-Nearest Neighbors | 0.723333 | 0.686667 | 0.741007 | 0.004302 |
| XgBoost | 0.960000 | 0.966667 | 0.953947 | 0.503060 |
| Random Forest | 0.966667 | 0.966667 | 0.966667 | 0.448467 |
| Gradient Boosting | 0.960000 | 0.960000 | 0.960000 | 0.915377 |
| LightGBM | 0.973333 | 0.966667 | 0.979730 | 0.241174 |

Setelah dilakukan normalisasi terlihat apabila akurasi model K-Nearest Neighbors mengalami peningkatan. Namun, LightGBM masih menjadi model terbaik setelah dilakukan normalisasi.

# EVALUATION MODELING

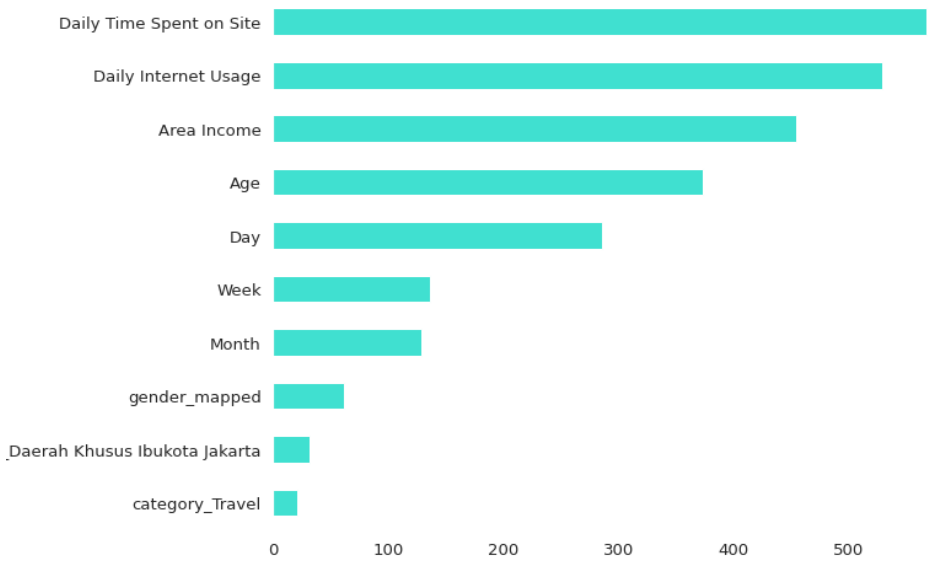
## CONFUSION MATRIX




- Tujuan dari modeling ini adalah untuk memprediksi jumlah maksimum customer potensial yang mengklik iklan.
- Oleh karena itu, kita perlu meminimalkan False Positive (5) dimana customer yang diprediksi akan mengklik iklan tetapi tidak mengklik iklan.
- Hal iniah yang kemudian dapat menimbulkan potensi kerugian

# EVALUATION MODELING

## FEATURE IMPORTANCE




Terdapat 4 feature yang memiliki pengaruh tinggi terhadap model yaitu:

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

# BUSINESS RECOMENDATION

## CUSTOMER CLASSIFICATION

### High Class Customer

- Customer yang memiliki income yang tinggi dan usia yang muda.
- Tipe ini cenderung lebih sering menggunakan internet dan dalam waktu yang lama.
- Customer tipe ini cenderung untuk tidak mengklik iklan yang diberikan

### High Class Customer

- Customer yang memiliki income yang rendah dan usia tua
- Tipe ini tidak sering menggunakan internet dan dalam dengan waktu yang singkat
- Customer tipe ini cenderung untuk mengklik iklan

# BUSINESS RECOMENDATION

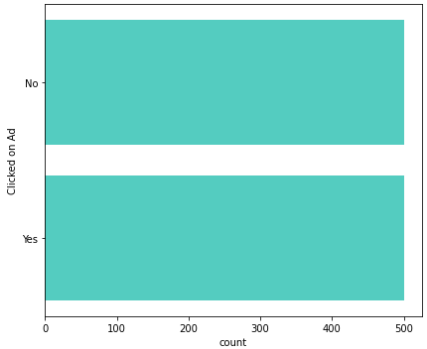
# SIMULATION

**Asumsi**
- Marketing cost per customer = Rp 10.000
- Keuntungan dari customer yang mengklik iklan = Rp 15.000
- Simulasi dilakukan dengan target populasi sebanyak 400 customer

**Tanpa Machine Learning**

Diketahui dari dataset ini distribusi Clicked on Ad adalah 50:50

Sehingga kita dapat membagi distribusi:
- Customer yang mengklik iklan (200)
- Customer yang tidak mengklik iklan (200)




**Marketing cost**
400 customers x Rp 10.000 = Rp 4.000.000

**Revenue**
200 customers x Rp 15.000 = Rp 3.000.000

**Profit**
Revenue - Cost = - (Rp 1.000.000)

Dari simulasi diatas tampak nilai potensi kerugian sebesar **Rp 1.000.000** dengan persentase kerugian **25%.**

# SIMULATION

**Asumsi**
- Marketing cost per customer = Rp 10.000
- Keuntungan dari customer yang mengklik iklan = Rp 15.000
- Simulasi dilakukan dengan target populasi sebanyak 400 customer

**Dengan Machine Learning**

Karena tujuannya menurunkan False Positive dan akurasi model sebesar 97%

Sehingga kita dapat membagi distribusi:
- Customer yang mengklik iklan (206)
- Customer yang tidak mengklik iklan (194)

**Marketing cost**
206 customers x Rp 10.000 = Rp 2,060,000 ▼

**Revenue**
206 customers x Rp 15.000 = Rp 3,090,000 ▲

**Profit**
Revenue - Cost = Rp 1,030,000 ▲

Dari simulasi diatas tampak nilai potensi keuntungan sebesar **Rp 1,030,000** dengan persentase keuntungan **50%.**

# CONCLUSION

**Dari Hasil Simulasi**

- Jumlah marketing cost yang digunakan lebih tinggi apabila tanpa machine learning dikarenakan target market yang kurang jelas. Sehingga menimbulkan potensi kerugian dari pemasaran yang kurang efisien dan dapat menyebabkan kerugian.
- Sedangkan apabila kita mengimplementasi machine learning, kita dapat lebih menspesifikasi target market sehingga dapat meminimalkan potensi kerugian dan dapat meraih keuntungan yang lebih potensial.
- Dengan mengimplementasikan machine learning kita dapat meraih potensi keuntungan sebesar 50%