Predict Clicked Ads Customer Classification by using Machine Learning

Created by:
Yasmin Fauziah
yasminfauziah63@gmail.com
https://www.linkedin.com/in/yasmin-fauziah-85b738239/

"Bachelor of Physics from Padjadjaran University> Someone who enjoys learning new things, has good analytical and planning skill. Enjoy to solve problem related to data analysis using Excel, SQL, Phython and Looker Studio. Have a high interest in a career in the data field."

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

Overview

"A company in Indonesia wants to know the effectiveness of an advertisement that they air, this is important for companies to be able to know how much the advertisement is marketed so that it can attract customers to see advertisements.By processing historical advertisement data and finding insights and patterns that occur, it can help companies determine target marketing, the focus of this case is to create a machine learning classification model that functions to determine the right target customers "

Exploration Data Analysis (EDA)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
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   Male                      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)
memory usage: 86.1+ KB
```

- **Description**

Dataset that contains information related to personal browsing history made by Ads Company.

- **Shape**

1000 Row and 11 Feature

- **Datatypes**

Float64 (3 Feature), Int64 (2 Feature), object (6 Feature)

Untuk selengkapnya, dapat melihat jupyter notebook disini

Statistical Analysis

Data Numerical

```
df[cats].describe()
```

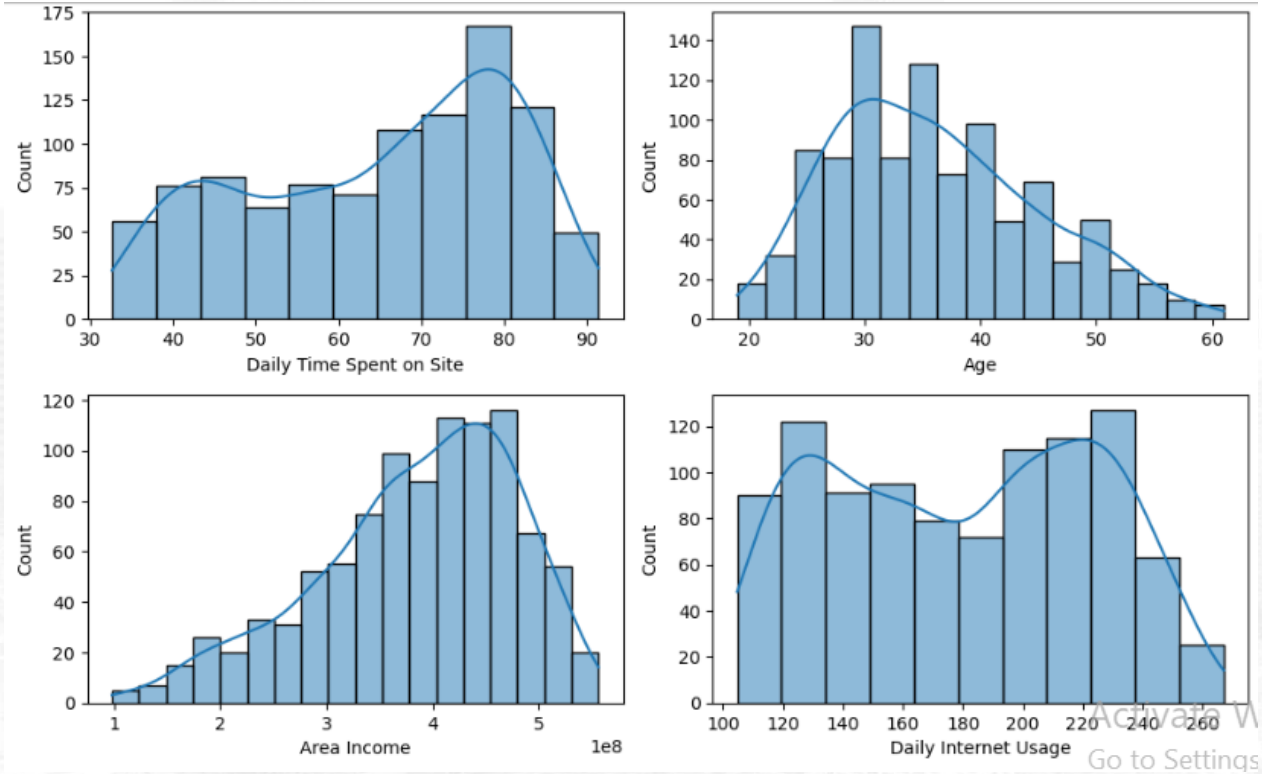
	Male	Timestamp	Clicked on Ad	city	province	category
count	997	1000	1000	1000	1000	1000
unique	2	997	2	30	16	10
top	Perempuan	5/26/2016 15:40	No	Surabaya	Daerah Khusus Ibukota Jakarta	Otomotif
freq	518	2	500	64	253	112

Data Categorcal

```
df[nums].describe()
```

	Unnamed: 0	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

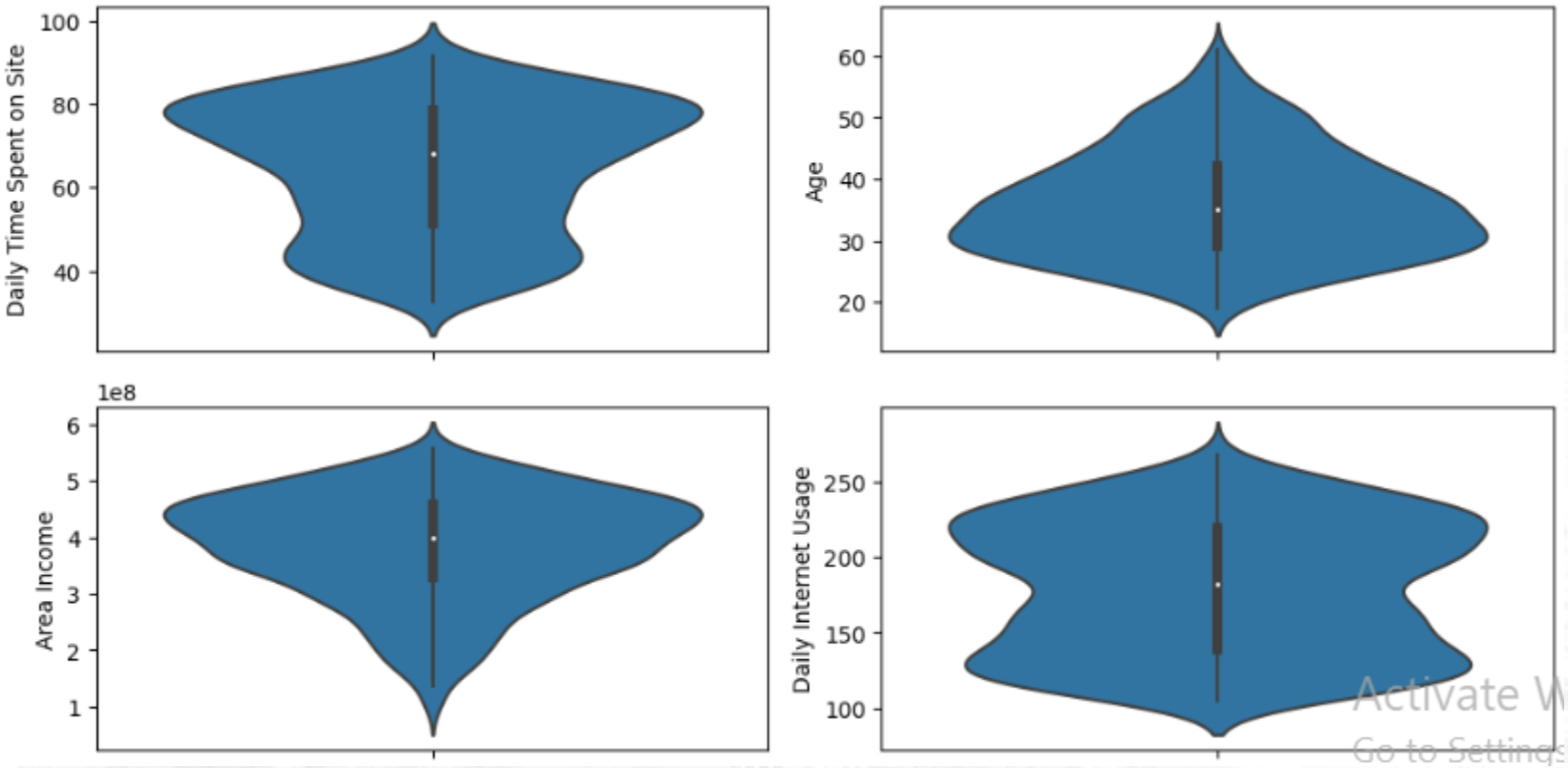
Univariate Analysis



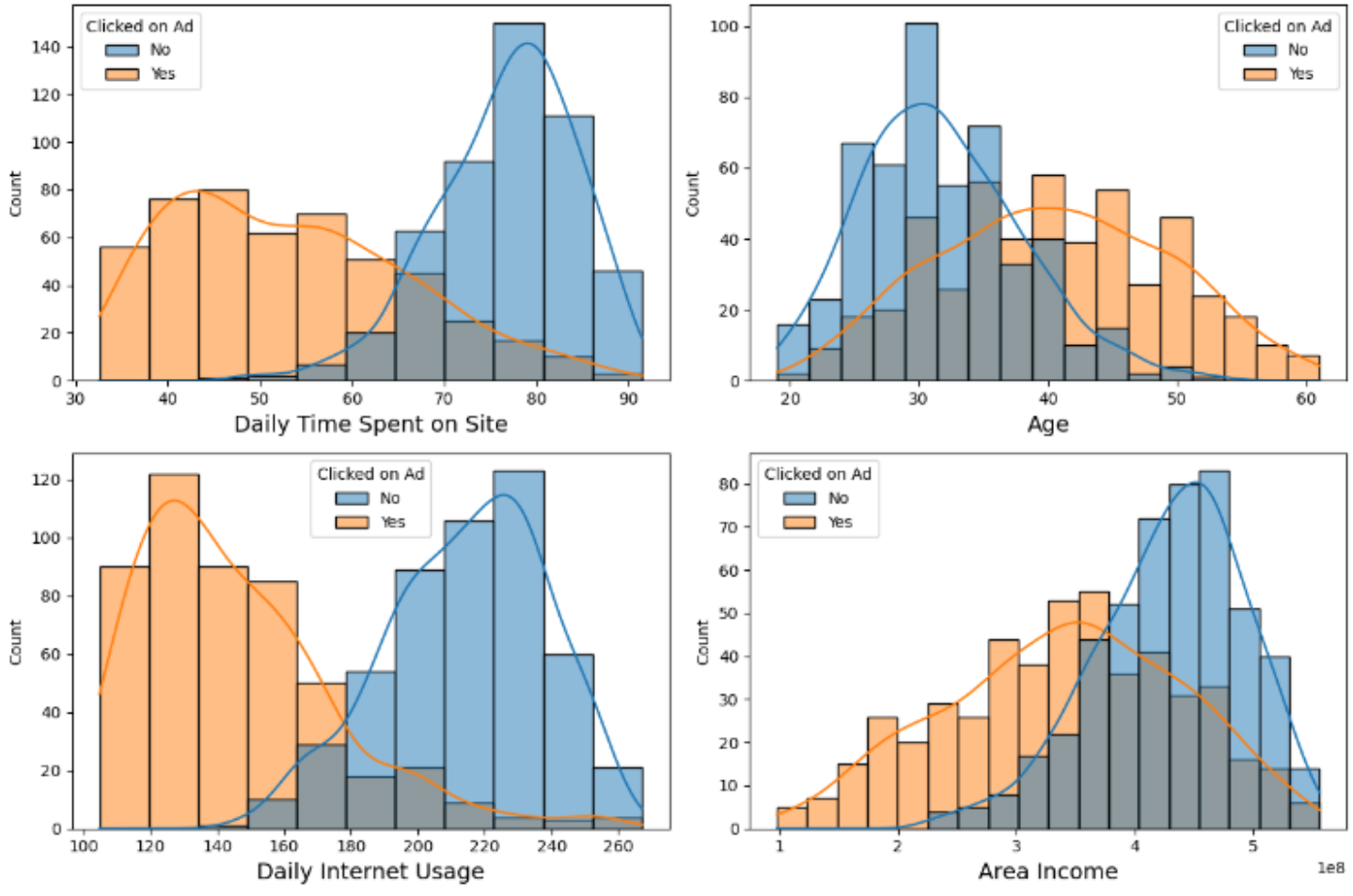
Interpretation:

- The distributions of **"Daily Time Spent on Site"** and **"Daily Internet Usage"** features have a bimodal distribution.
- The distribution of **"Age"** feature has a negatively skewed distribution
- The distribution of **"Area Income"** feature has a positively skewed distribution

Univariate Analysis



Bivariate Analysis



Bivariate Analysis

Interpretation:

Daily Time Spent on Site

- Customers with Daily Time Spent on Site 35 - 45 are more clicked on ad

- Customers with Daily Time Spent on Site 70 - 80 are more didn't clicked on ad

Age

- Customers with Age 35 - 45 are more clicked on ad

- Customers with Age 25 - 35 are more didn't clicked on ad

Daily Internet Usage

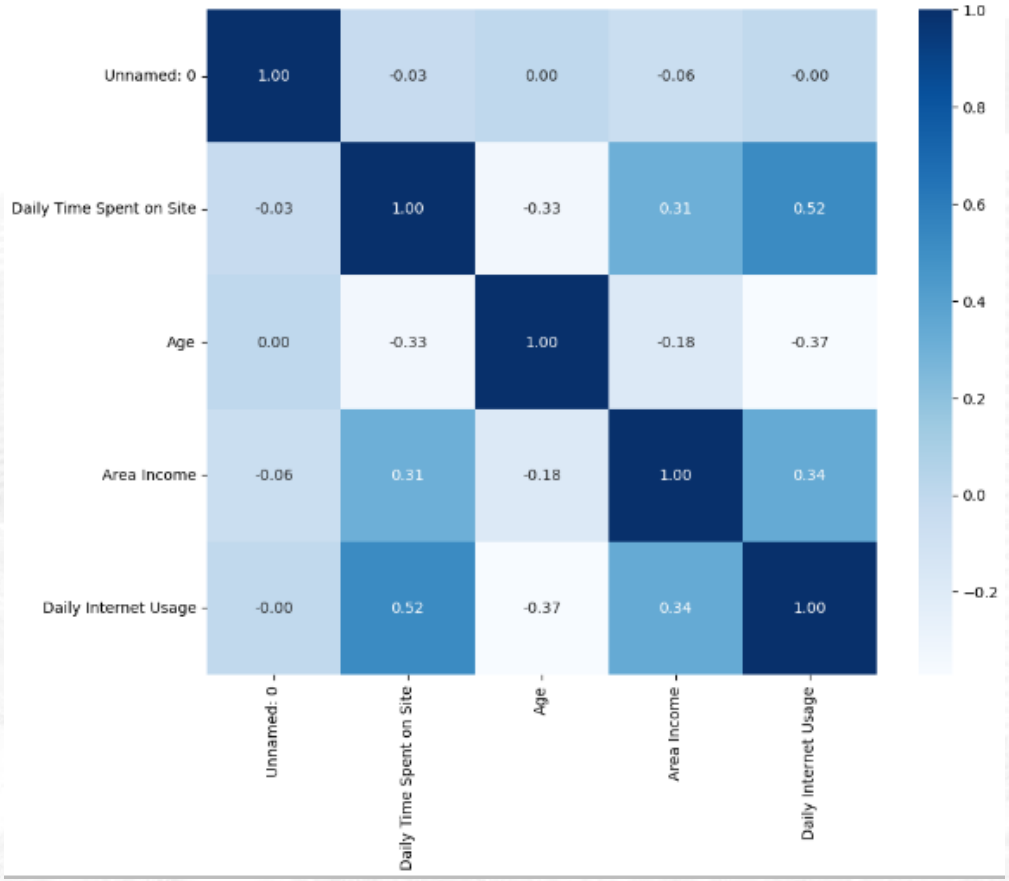
- Customers with Daily Internet Usage of 100 - 150 are more clicked on ad

- Customers with Daily Internet Usage of 175 - 225 are more didn't clicked on ad

Area Income

- Customers with an income range of around 380 - 460 million didn't clicked on ad more than those with other income ranges

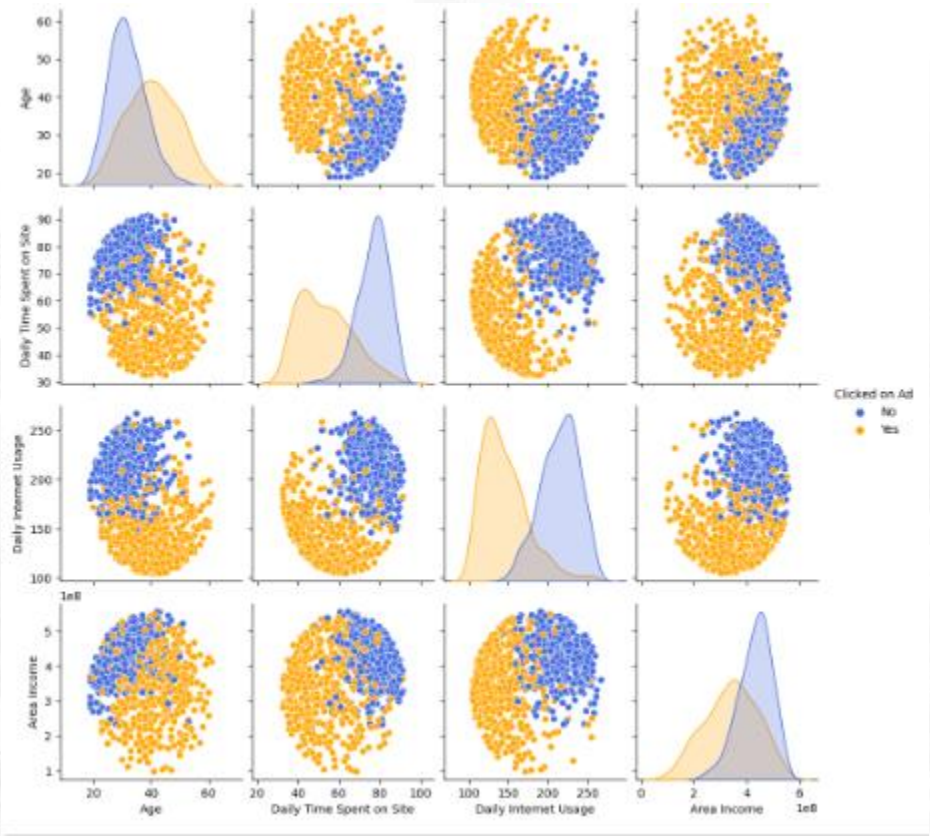
Multivariate Analysis



Interpretation:

- The **“Daily Internet Usage”** feature has a positive correlation with **“Daily Time Spent on Site”** feature equal to 0.52 and 0.34 with **“Area Income”** feature.
- The **“Daily Internet Usage”** and **“Age”** features have a negative correlation equal to -0.37
- The **“Daily Time Spent on Site”** feature has a negative correlation with **“Age”** feature equal to -0.33

Multivariate Analysis



Data Cleaning & Preprocessing

Handling Missing Value

```
dfp = df2.isna().sum()*100/len(df2)
print(round(dfp, 2).sort_values(ascending=False))
```

```
Daily Time Spent on Site    1.3
Area Income                 1.3
Daily Internet Usage        1.1
Gender                      0.3
Age                         0.0
Timestamp                   0.0
Clicked on Ad               0.0
city                        0.0
province                    0.0
category                    0.0
dtype: float64
```

There are 4 features that have null values including **“Daily Time Spent on Site”, “Area Income”, “Daily Internet Usage” dan “Gender”**. The four features are filled with median and mode values.

Handling Data Duplicate

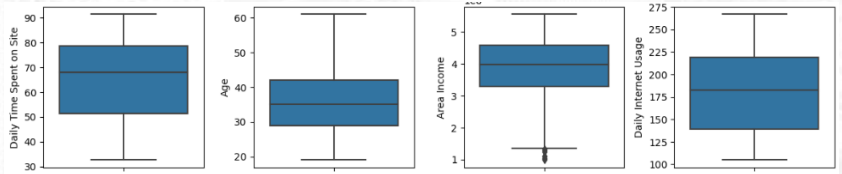
```
df2.duplicated().any()
```

```
False
```

No duplicate data

Handling Outlier

```
Total Rows BEFORE Outlier Handling Z-Score = 1000
Total Rows AFTER Outlier Handling Z-Score = 997
```



Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Cleaning & Preprocessing

Rakamin Academy

Feature Extraction

```
df2['year'] = df2.Timestamp.dt.year
df2['month'] = df2.Timestamp.dt.month
df2['week'] = df2.Timestamp.dt.isocalendar().week
df2['day'] = df2.Timestamp.dt.day
```

Extracted new columns from **“Timestamp”**

Feature Encoding

```
# Gender
mapping_gender = {
    'Perempuan' : 0,
    'Laki-Laki' : 1
}
dfe['Gender'] = dfe['Gender'].map(mapping_gender)
```

```
# Clicked on Ad
mapping_ads = {
    'No' : 0,
    'Yes' : 1
}
dfe['Clicked on Ad'] = dfe['Clicked on Ad'].map(mapping_ads)
```

```
for i in ['city','province','category']:
    onehots = pd.get_dummies(dfe[i], prefix=i)
    dfe = dfe.join(onehots)
```

Feature Selection

```
dfe.drop(columns=['Timestamp', 'city', 'province', 'category'], inplace=True)
```

Split Data

```
X = dfs.drop(columns=['Clicked on Ad'])
y = dfs['Clicked on Ad']
```

```
print(X.shape)
print(y.shape)
```

```
(997, 65)
(997,)
```

```
y.value_counts()
```

```
0    500
1    497
Name: Clicked on Ad, dtype: int64
```

```
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size = 0.3, random_state = 42)
```

```
print('X_train size : ', X_train.shape)
print('X_test size  : ', X_test.shape)
print('y_train size : ', y_train.shape)
print('y_test size  : ', y_test.shape)
```

```
X_train size :  (697, 65)
X_test size  :  (300, 65)
y_train size :  (697,)
y_test size  :  (300,)
```

Modelling Without Normalization & Standarization

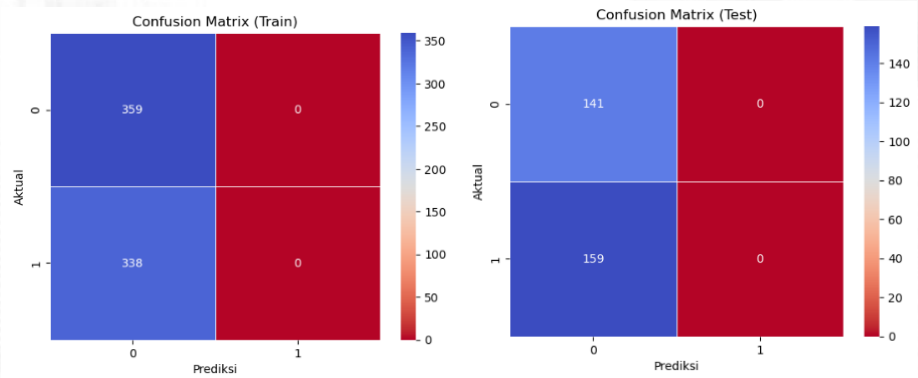
Model	Accuracy	Recall	Precision	F1-Score
Logistic Regression	0.470	0.00	0.00	0.00
Decision Tree	0.940	0.943	0.943	0.943
RandomForest	0.957	0.962	0.956	0.959
K-Nearest Neighbor	0.630	0.688	0.553	0.613
Gradient Boosting	0.967	0.974	0.962	0.968

Modelling With Normalization & Standarization

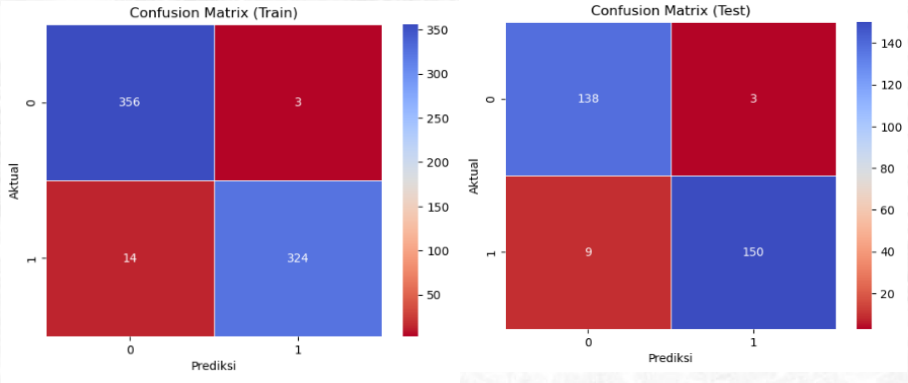
Model	Accuracy	Recall	Precision	F1-Score
Logistic Regression	0.960	0.980	0.943	0.962
Decision Tree	0.937	0.938	0.943	0.940
RandomForest	0.950	0.962	0.943	0.952
K-Nearest Neighbor	0.757	0.816	0.698	0.752
Gradient Boosting	0.967	0.974	0.962	0.968

Untuk selengkapnya, dapat melihat jupyter notebook disini

Logistic Regressioon Without Normalization & Standarization

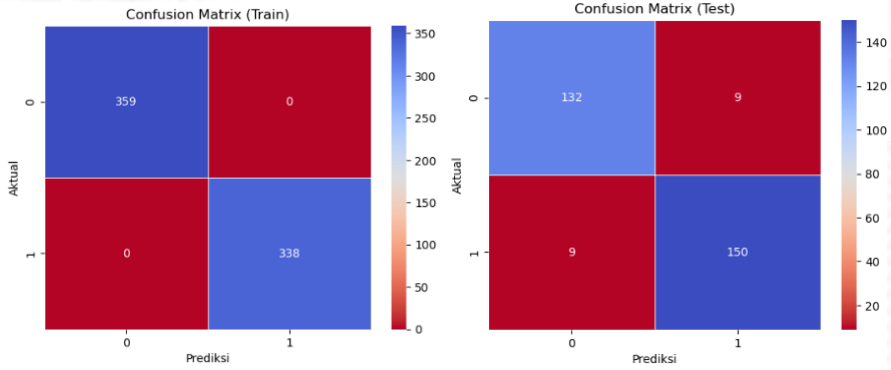


Logistic Regressioon With Normalization & Standarization

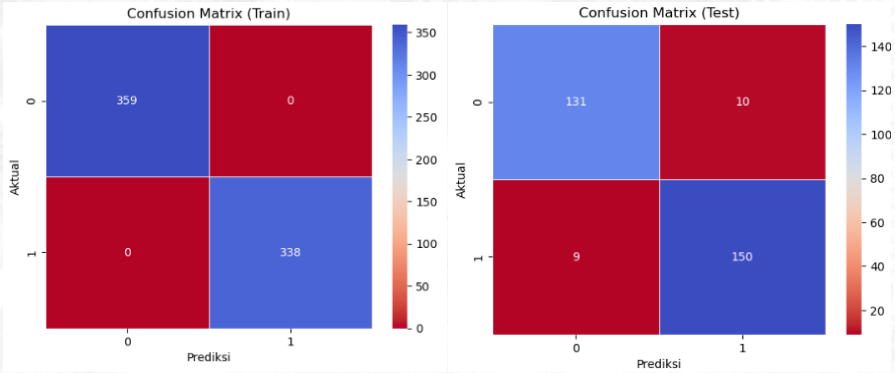


Data Modeling

Decision Tree Without Normalization & Standarization

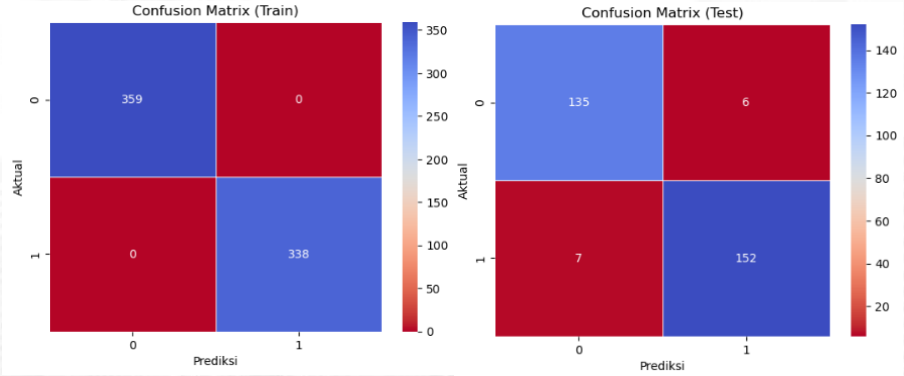


Decision Tree With Normalization & Standarization

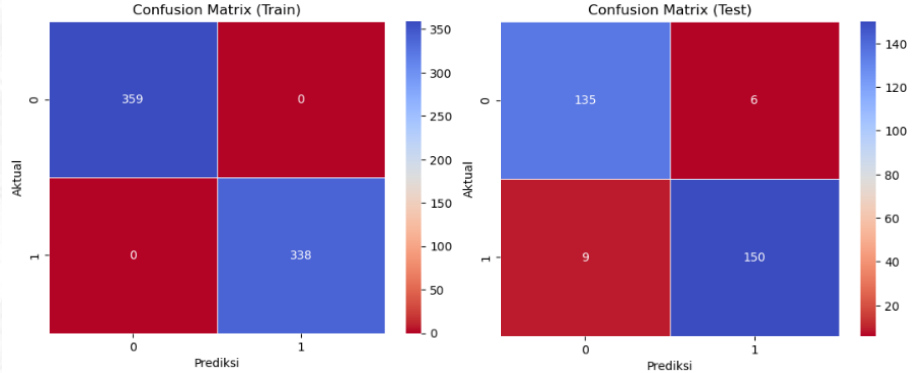


Data Modeling

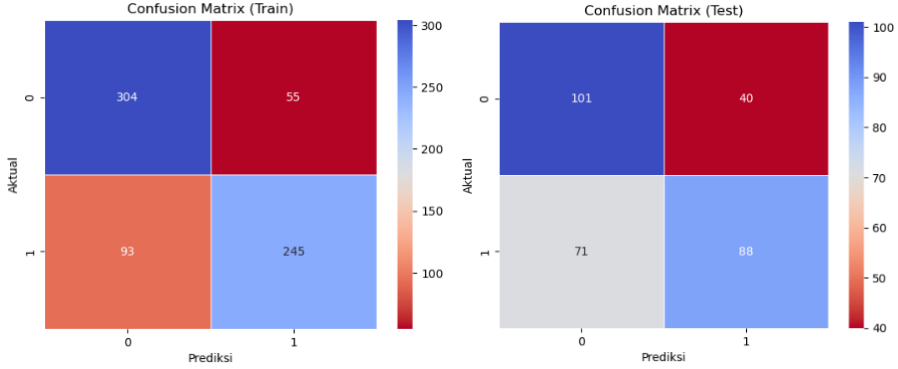
Random Forest Without Normalization & Standarization



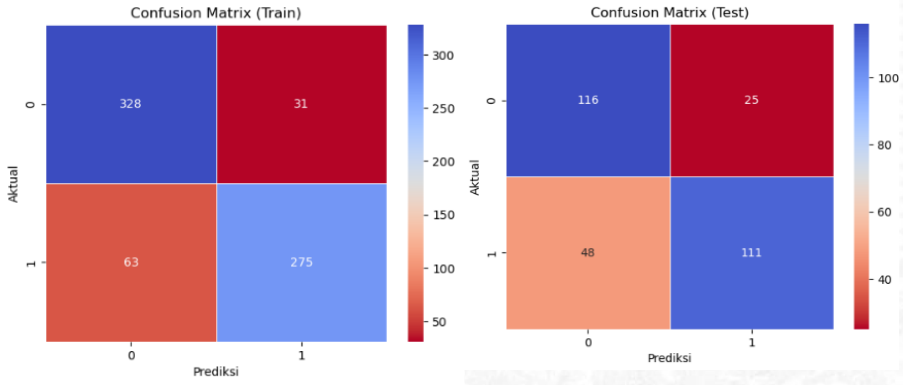
Random Forest With Normalization & Standarization



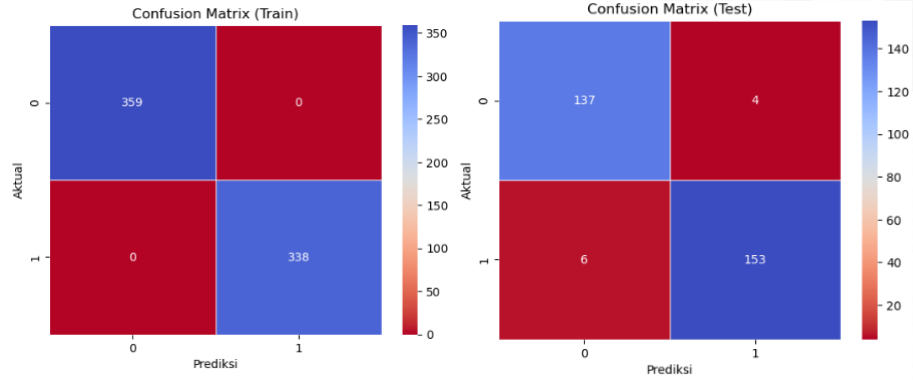
K-Nearest Neighbor Without Normalization & Standarization



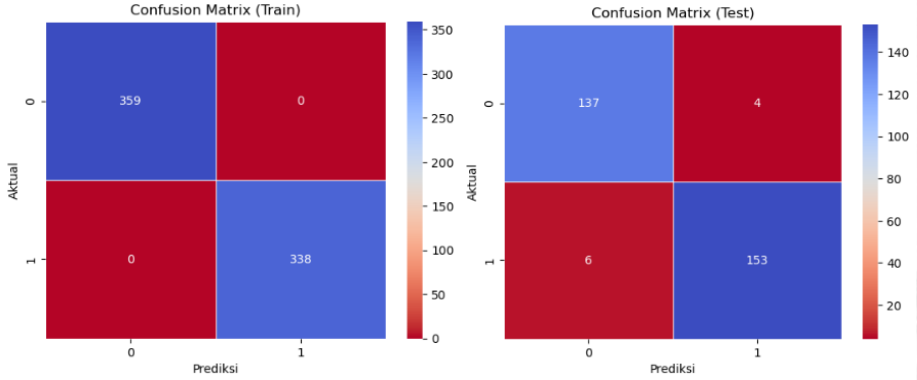
K-Nearest Neighbor With Normalization & Standarization



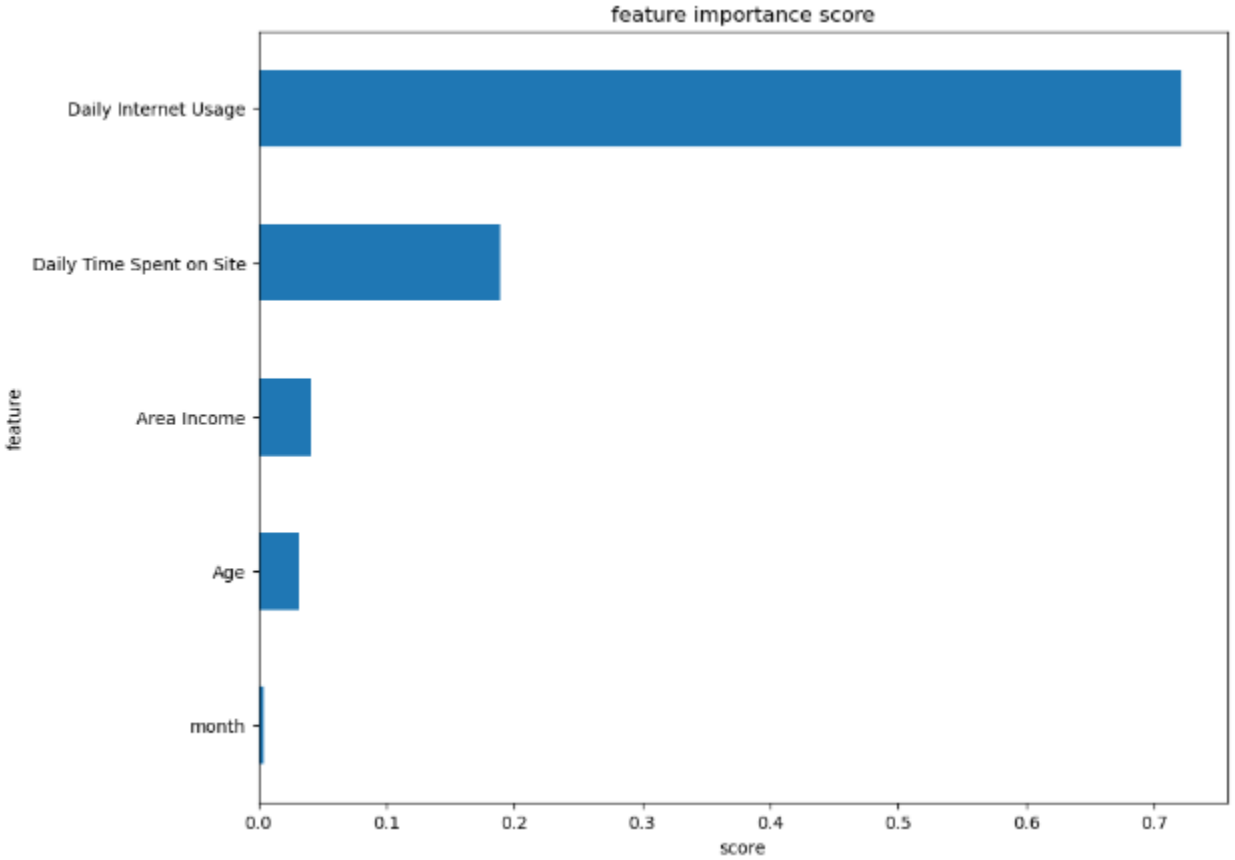
Gradient Boosting Without Normalization & Standarization



Gradient Boosting With Normalization & Standarization

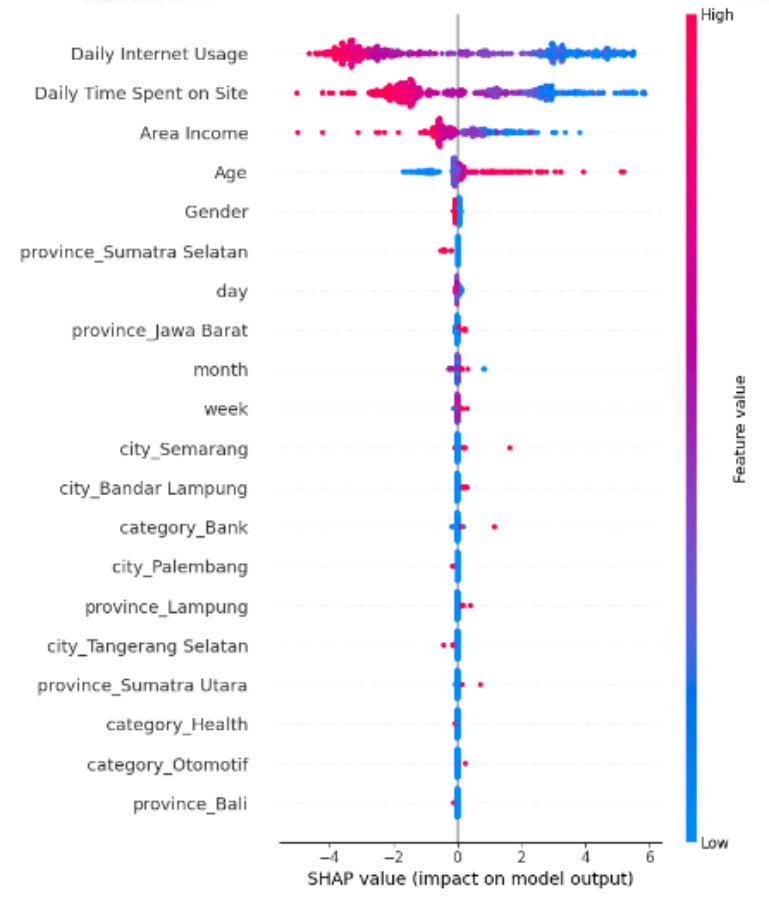


Data Modeling



Top four features to predict advertisements clicked by customers are **Daily Internet Usage, Daily Time Spent on Site, Area Income,** and **Age**. Daily Internet Usage has the highest importance score highest, followed by Daily Time Spent on Site. This shows that the higher the daily internet usage and daily time spent on site, the higher the likelihood of customers to click on ads.In addition, customers with higher area incomesand younger age are more likely to click on ads as well.

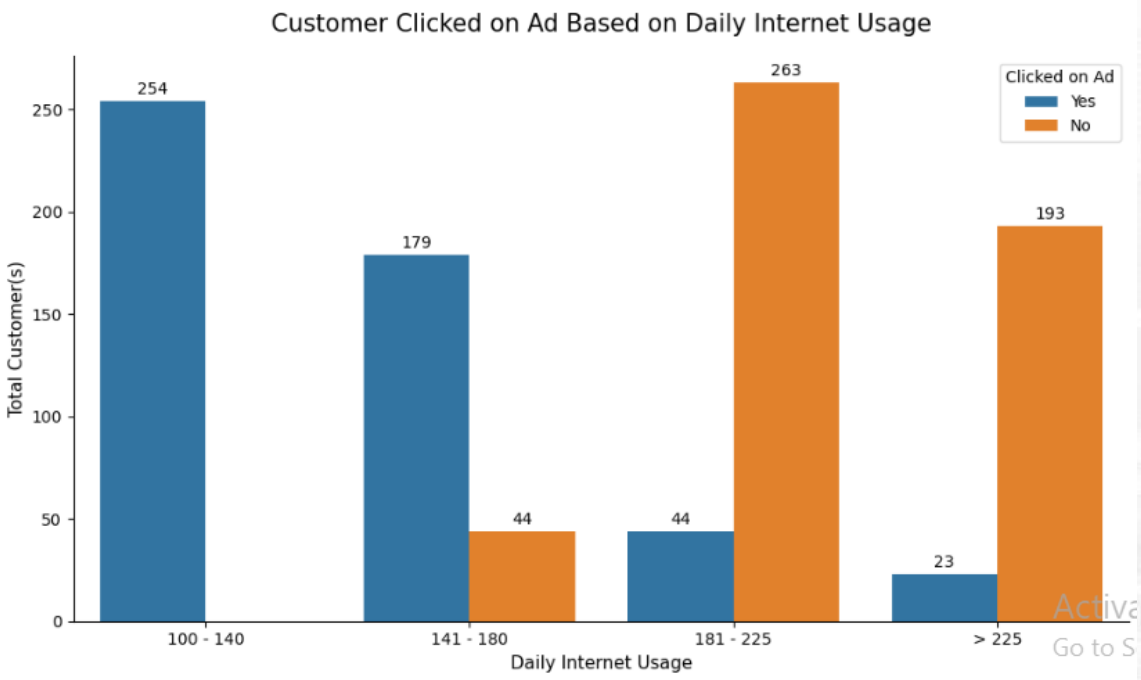
Business Recommendation & Simulation



Top four features to predict advertisements clicked by customers are **Daily Internet Usage, Daily Time Spent on Site, Area Income,** and **Age**. Daily Internet Usage has the highest importance score highest, followed by Daily Time Spent on Site. This shows that the higher the daily internet usage and daily time spent on site, the higher the likelihood of customers to click on ads.In addition, customers with higher area incomesand younger age are more likely to click on ads as well.

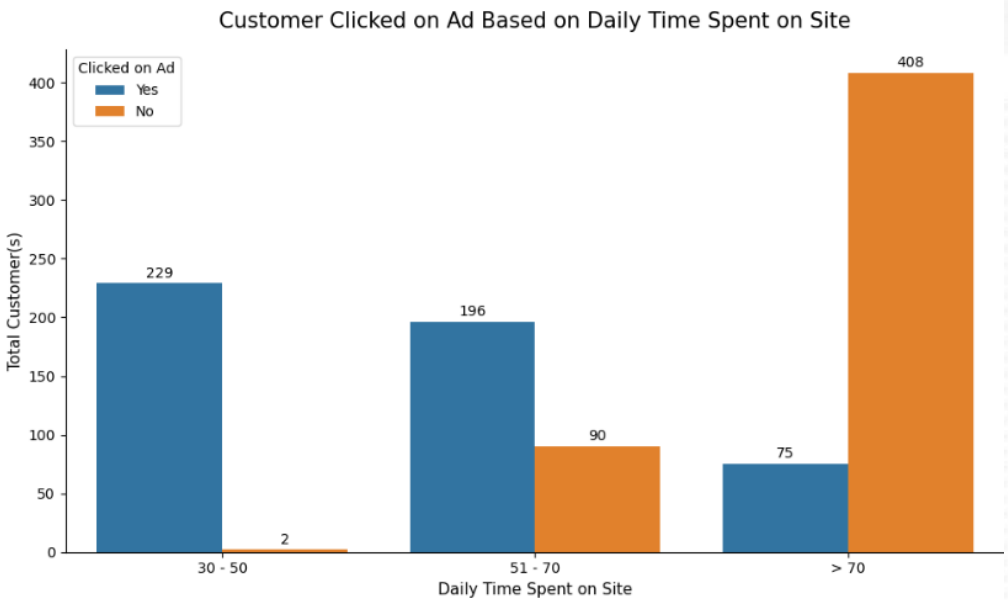
Untuk selengkapnya, dapat melihat jupyter notebook disini

Business Recommendation & Simulation



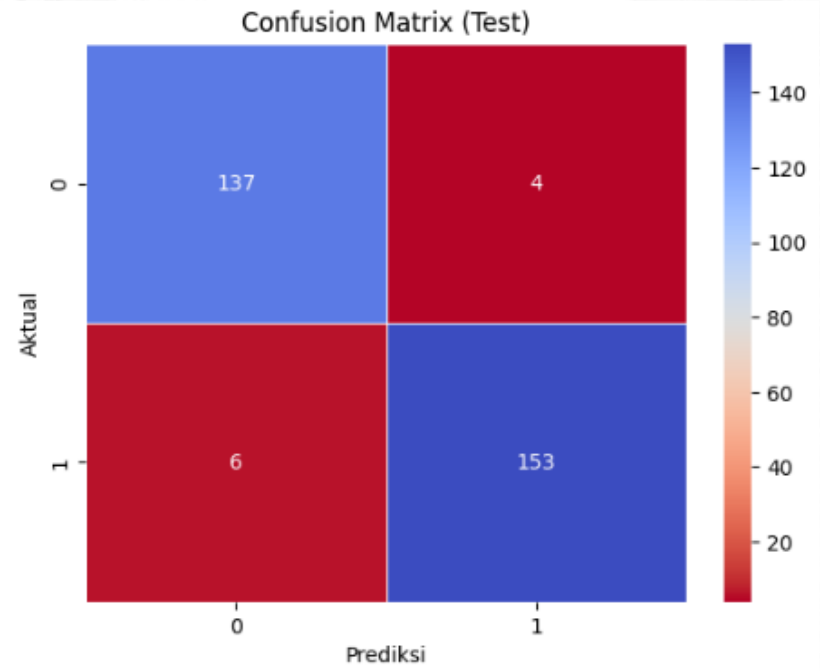
- Customers with daily internet usage of 100-140 and 141-180 had the highest percentage who clicked on the ads. This shows that they do not have the time and are more likely to be interested in the advertised service. Companies can give special promos so that customers will not only click on the ad but will buy products.
- Improve and optimize advertising promotions to target customers who have a high percentage of not clicking ads with daily internet usage between 181-225+. The company needs to further analyze such as adjusting ad content and targeting to better match interests.

Business Recommendation & Simulation



- Customers with daily time spent on the site 30-70 have the highest percentage of customers who clicked on the ads. This indicates that they are interested in the product orservice being advertised. Adjust your ad targeting strategy and improve user experience. This may include improving website design and usability of the website, providing clear and concise ad content, and ensuring that ad placements are visible and attractive to potential customers to potential customers.
- Customers with time spent on the site each day >70 have the highest percentage of not clicking on ads. it is important to optimize the advertising campaign by adjusting the advertising content and targeting to better suit their interests and preferences. This can be achieved by analyzing their behavior on the website, conducting surveys to understand their needs or interests, and providing appropriate promotional offers or products to encourage them to click through.or appropriate products to encourage them to click and buy.

Business Recommendation & Simulation



- In overall out of 300 sample customers, machine learning will predict 143 customers not click the ad and 157 customers click the ad or 48% and 52%.
- Based on confusion matrix table beside machine learning created successfully correctly predict customers who did not clicked on the ad by 137 customers out of a total 300 sample customers or 46%.
- From 157 customers who were predicted to click ads by machine Learning 153 customers were predicted correctly or 97%. While the rest about 3% are errors

Business Recommendation & Simulation

- Assumed the company runs a Digital Marketing Business with a Cost of Rp 5000 for each customer
- Company will benefit based on the number of customers who convert with cost of Rp 10 000 per customer
- Here we will focus on the losses caused by the cost that the company has spentto serve ads but do not generate revenue for the company customers who did not click the ad

Schema Without Machine Learning

Data by dataset available 1000 customers

500 Customers did not click the ad (50%)

500 Customers clicked the Ad (50%)

Revenue = 500 x 10,000 = 5,000,000

Marketing Cost = 1000 x 5,000 = 5,000,000

PROFIT Rp 0 (+0%)

Schema with Machine Learning

Data by dataset available1000 customers

480 Customers did not click the ad (48%)

520 Customers clicked the ad (52%)

504 Customers clicked the ad (97% of 520)

Revenue = 504 x 10,000 = 5,040,000

Marketing Cost = 520 x 5,000 = 2,600,000

PROFIT Rp 2,440,000 (+93%)