

[Click here for more details.](#)

PREDICT CLICKED ADS CUSTOMER CLASSIFICATION BY USING MACHINE LEARNING

BY: DEA DAHLILA


About Me

“I'm a Fresh Graduate of Geophysics Engineering with a keen interest in data science and analysis. I'm passionate about extracting meaningful insights from data and using them to drive informed decision-making.”

Let's Connect

✉ ddahlila6@gmail.com

 [linkedin.com/in/dea-dahlila/](https://www.linkedin.com/in/dea-dahlila/)

 github.com/Dea-dahlila



Overview

A company in Indonesia wants to assess the effectiveness of an advertisement they have aired. This is important for the company to determine the reach of the marketed advertisement and attract customers to view it. By analyzing historical advertisement data and discovering insights and patterns, it can assist the company in determining marketing targets. The focus of this case is to create a machine learning classification model that functions to determine the right target customers.

WHAT IS THE PROBLEM?

BACKGROUND

A company in Indonesia aims to assess the effectiveness of their aired advertisements by employing a machine learning classification model to identify the right target customers and enhance marketing reach.

GOALS

Determine the right customers to increase marketing reach.

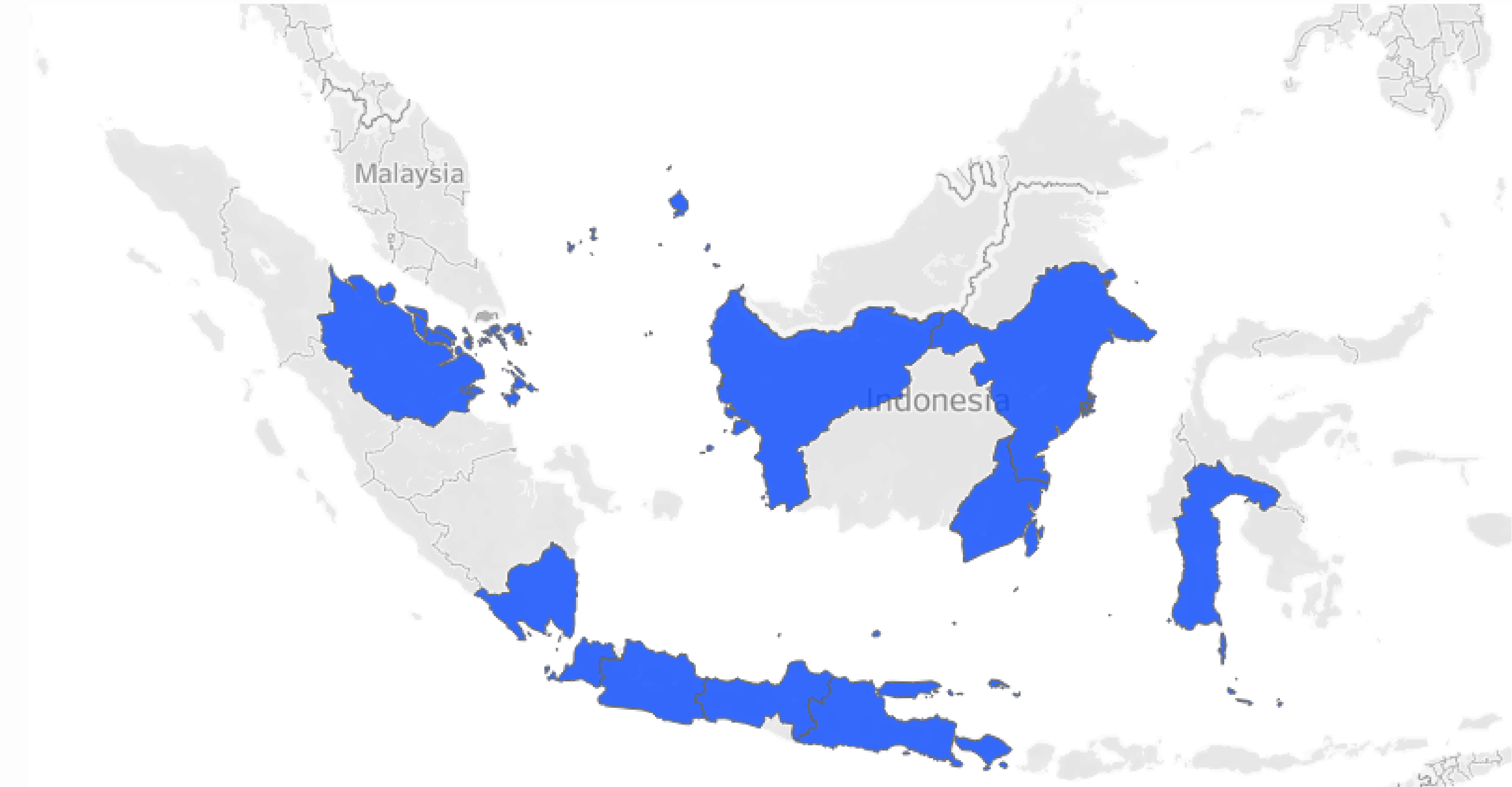
OBJECTIVES

Create a model classification to predict the right target customers effectively.

BUSINESS METRICS

- Accuracy of the Classification Model
- Conversion Rate

USER ORIGIN



DATASET

```
<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
```

| Shape

1000 data rows, 10 features.

| Missing Value

4 features that has missing value.

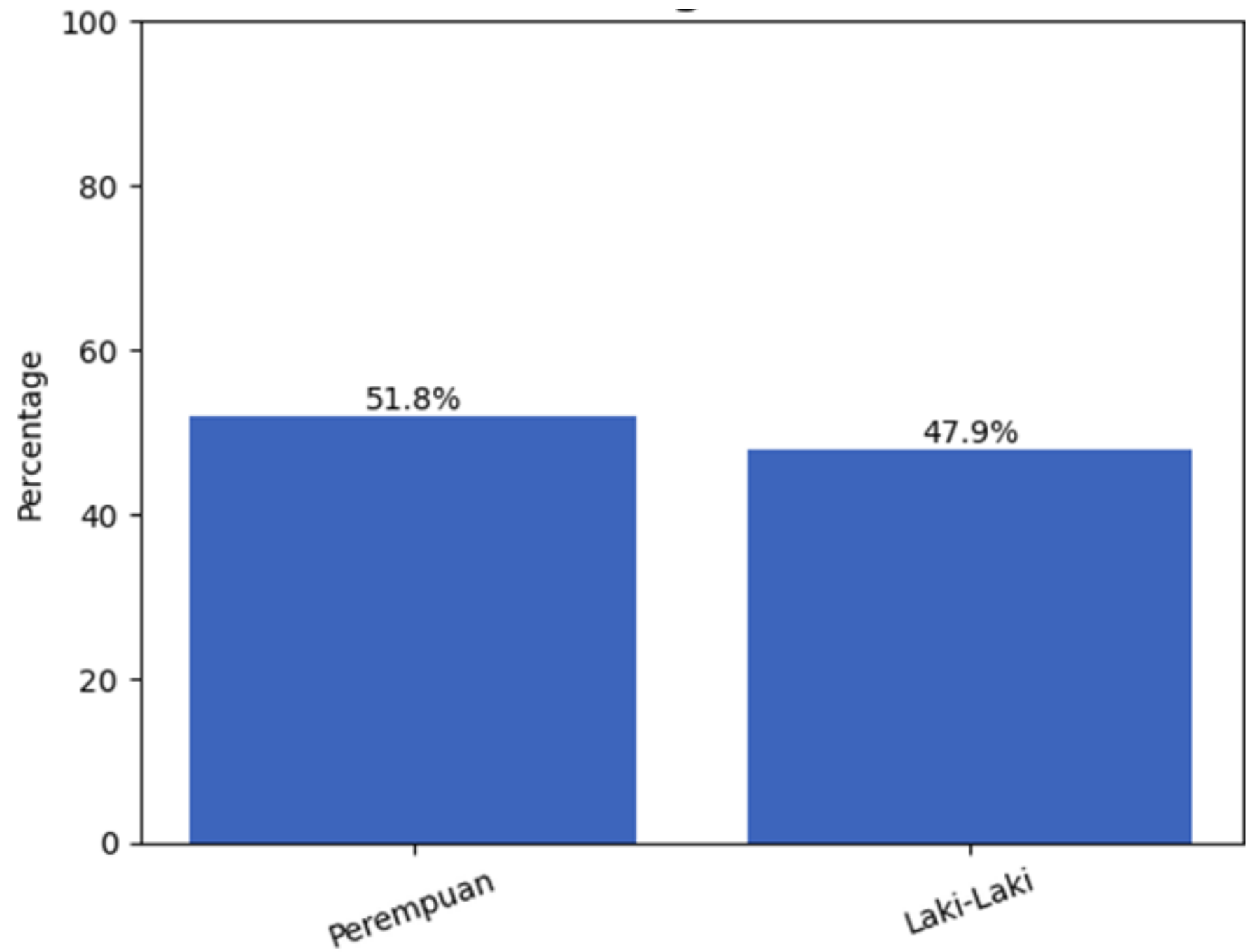
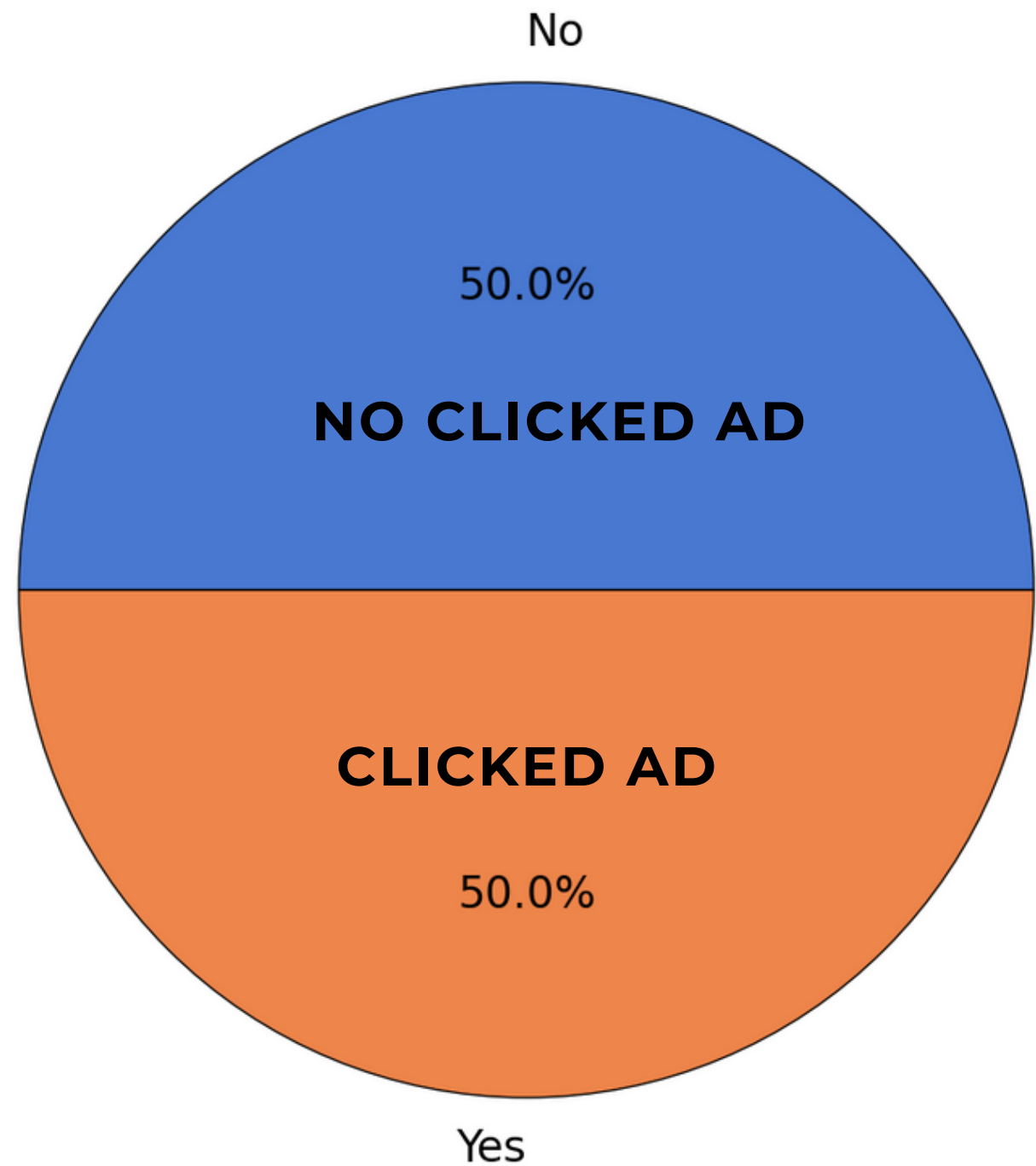
| Dtype

float64 (3 features), int64 (2 features), object (6 features).

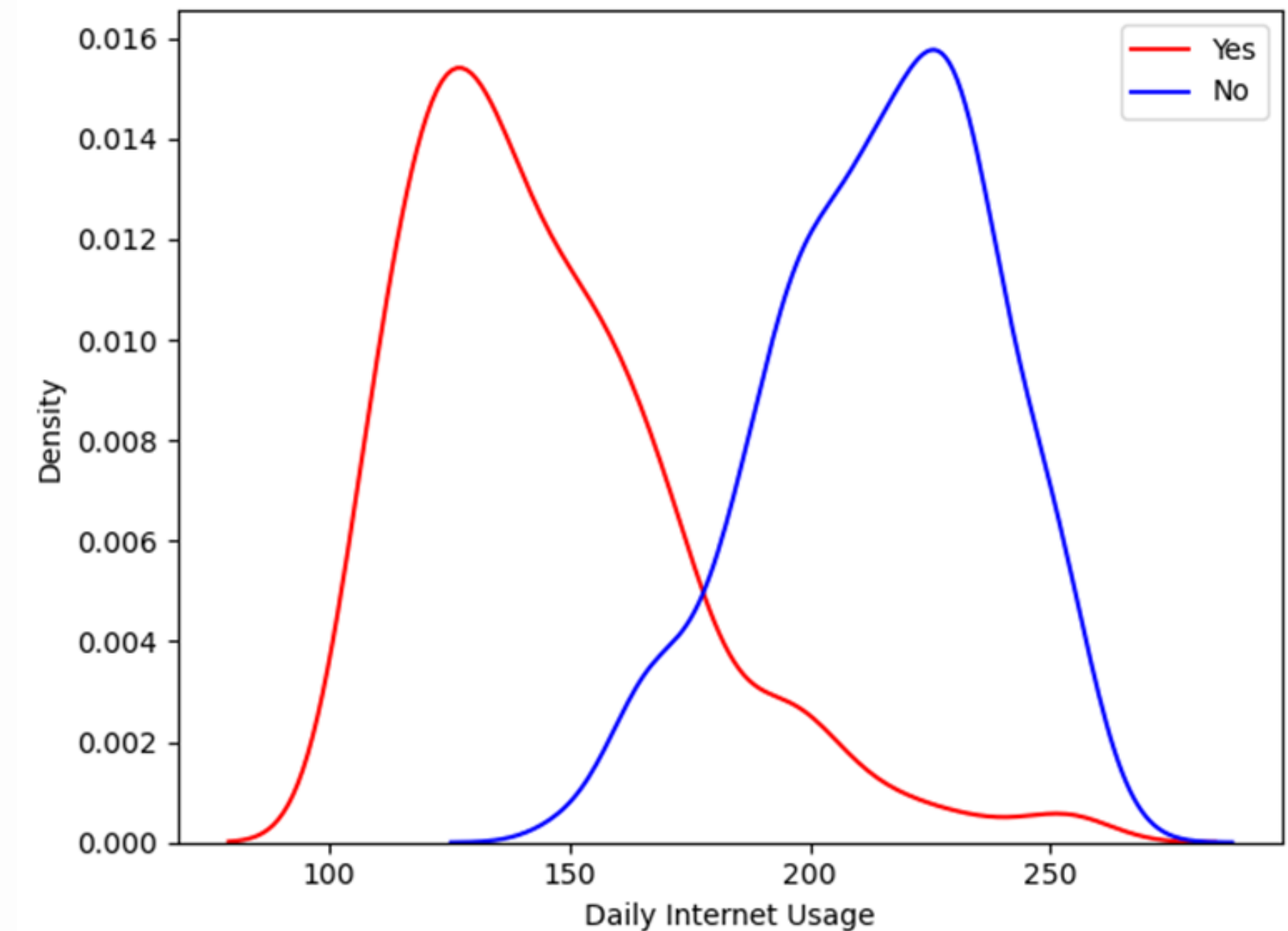
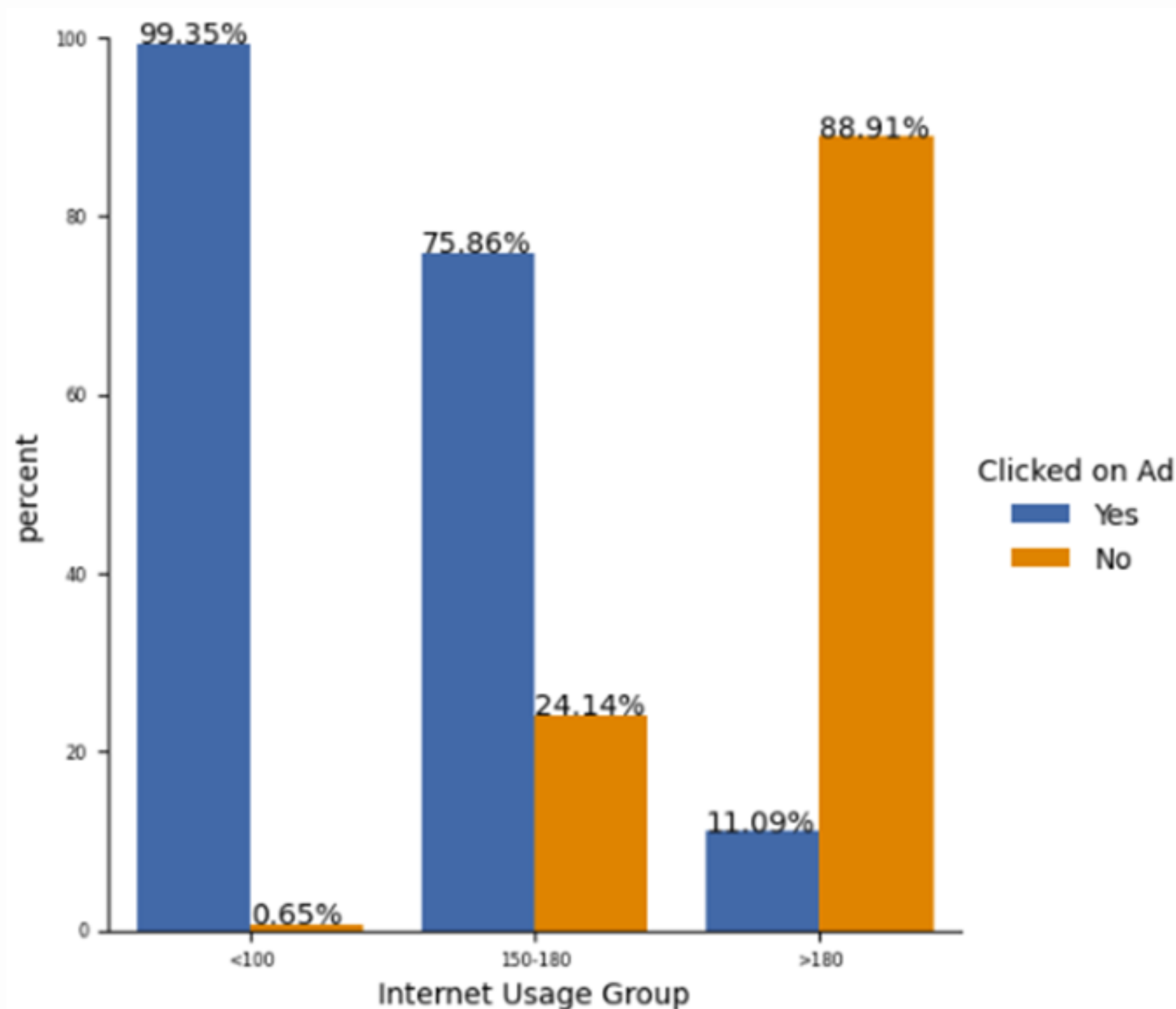
| Duplicated

0 data rows.

CLICKED ON AD & GENDER

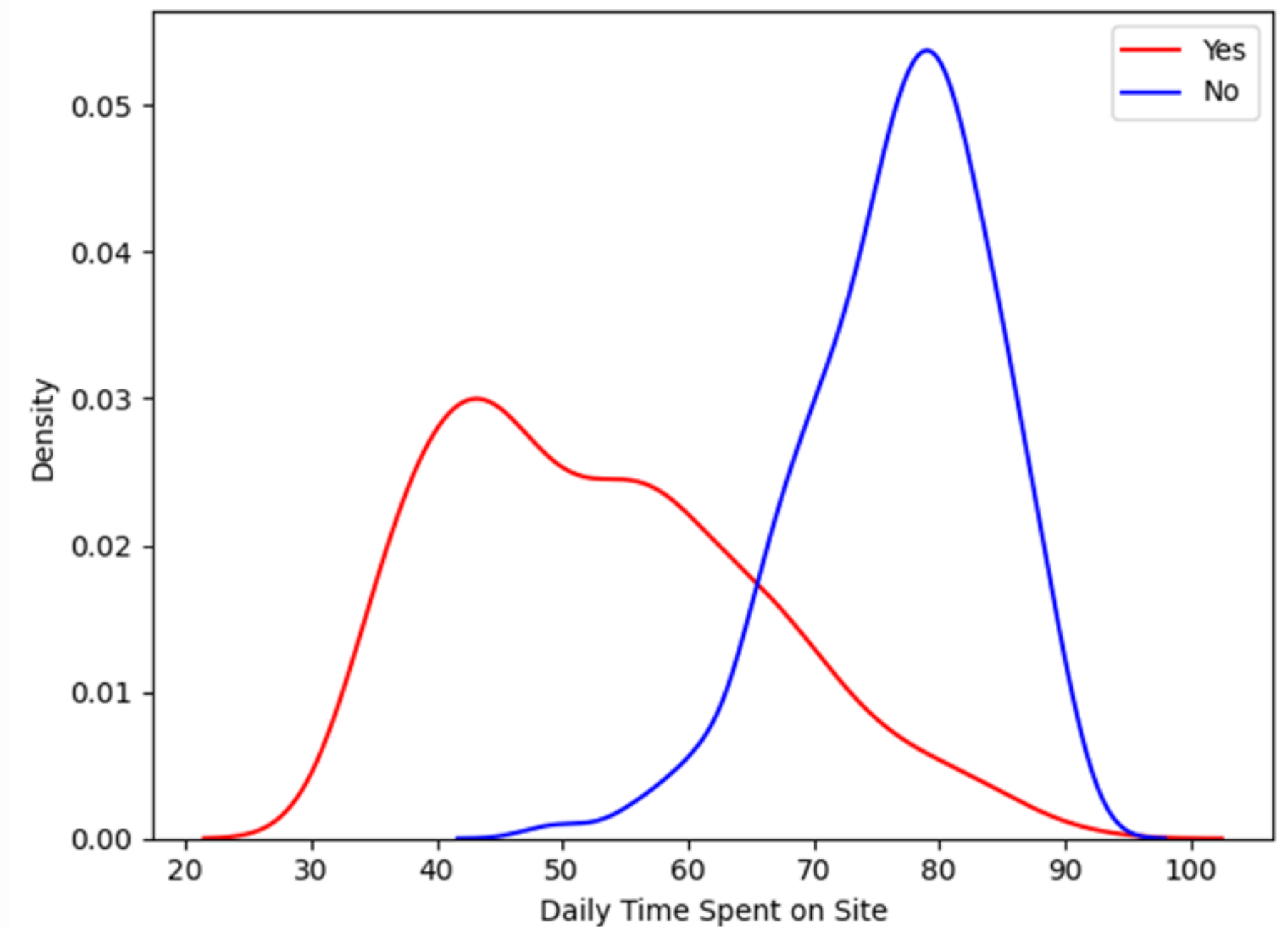
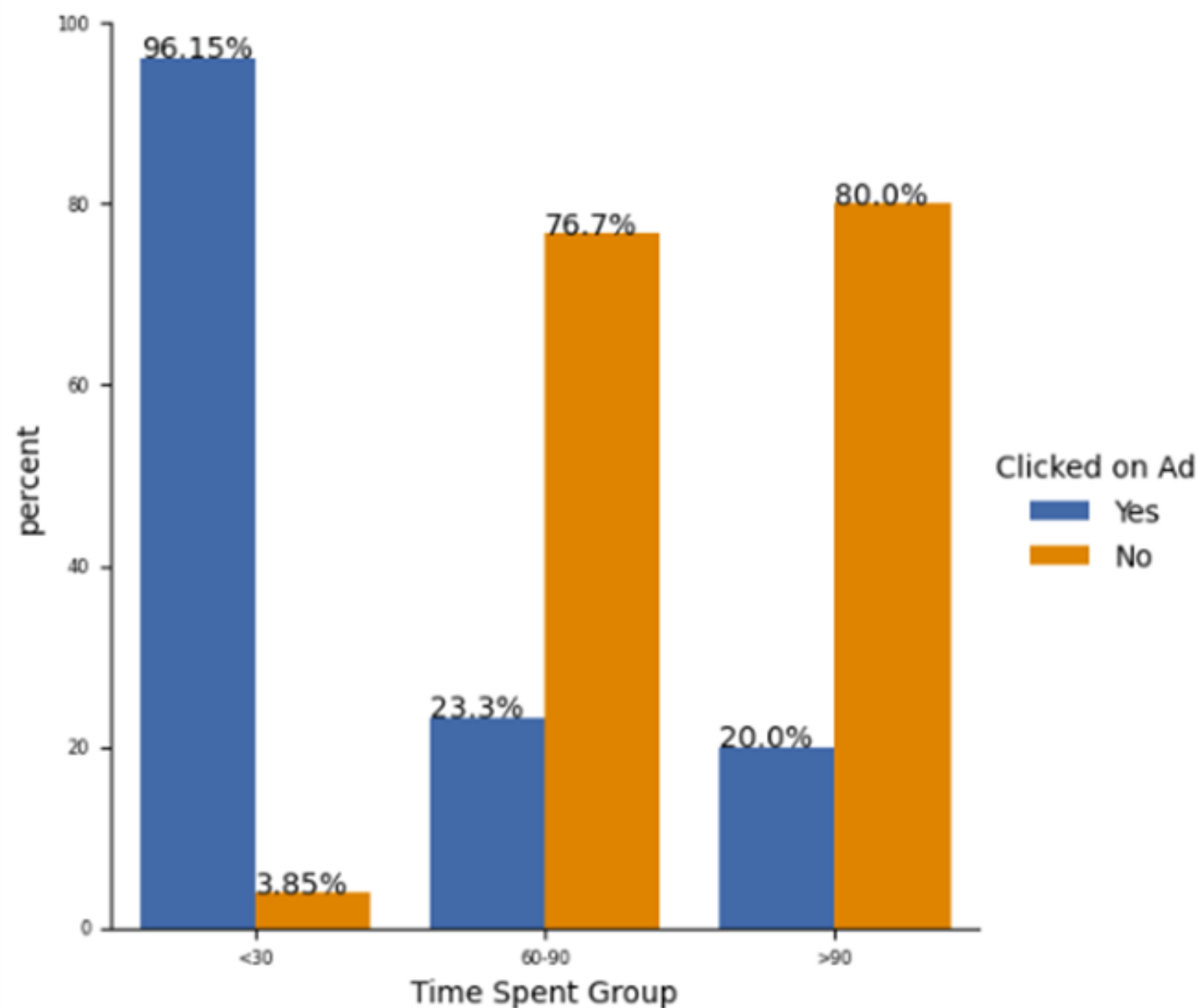


CLICKED ON AD & INTERNET USAGE



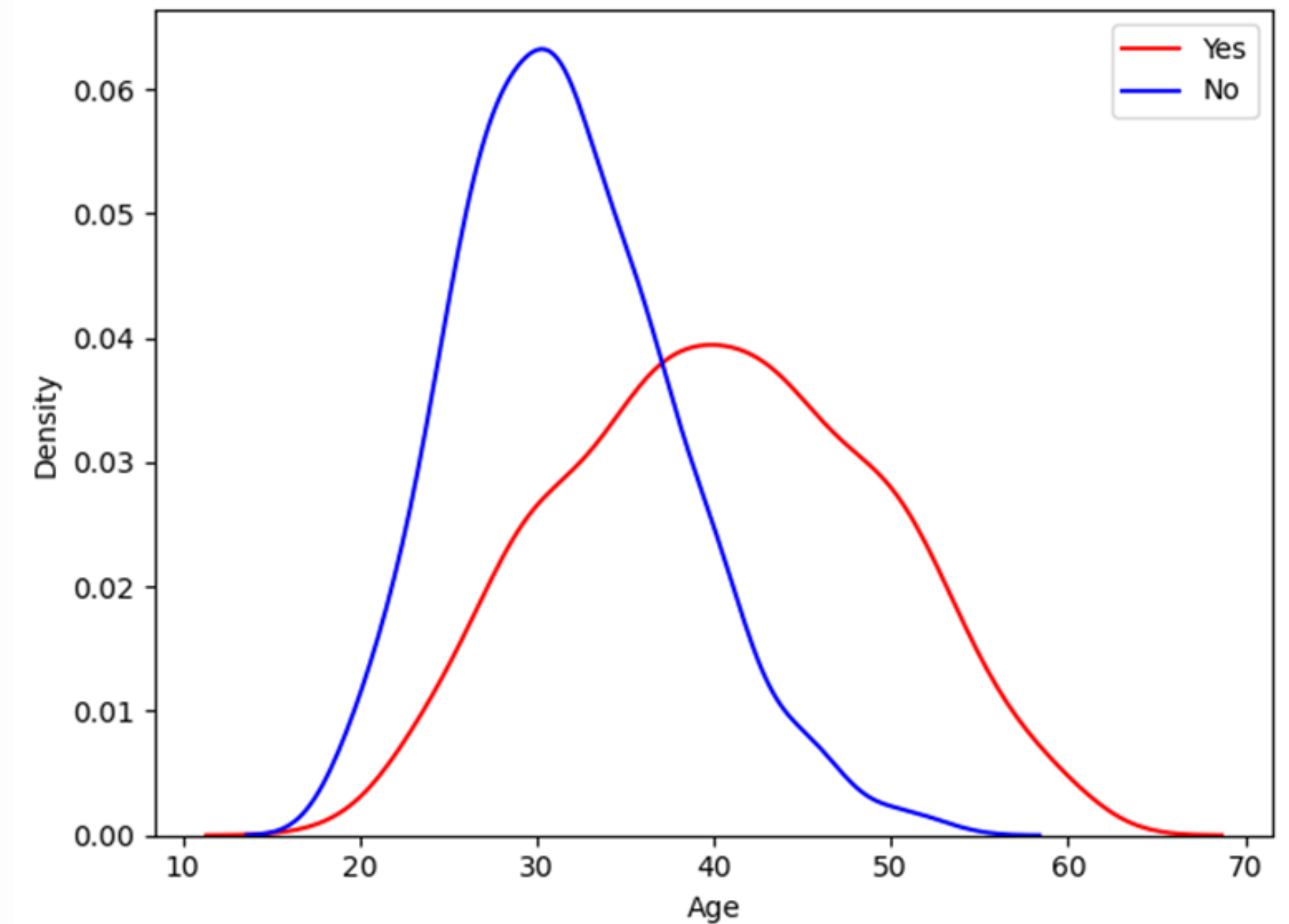
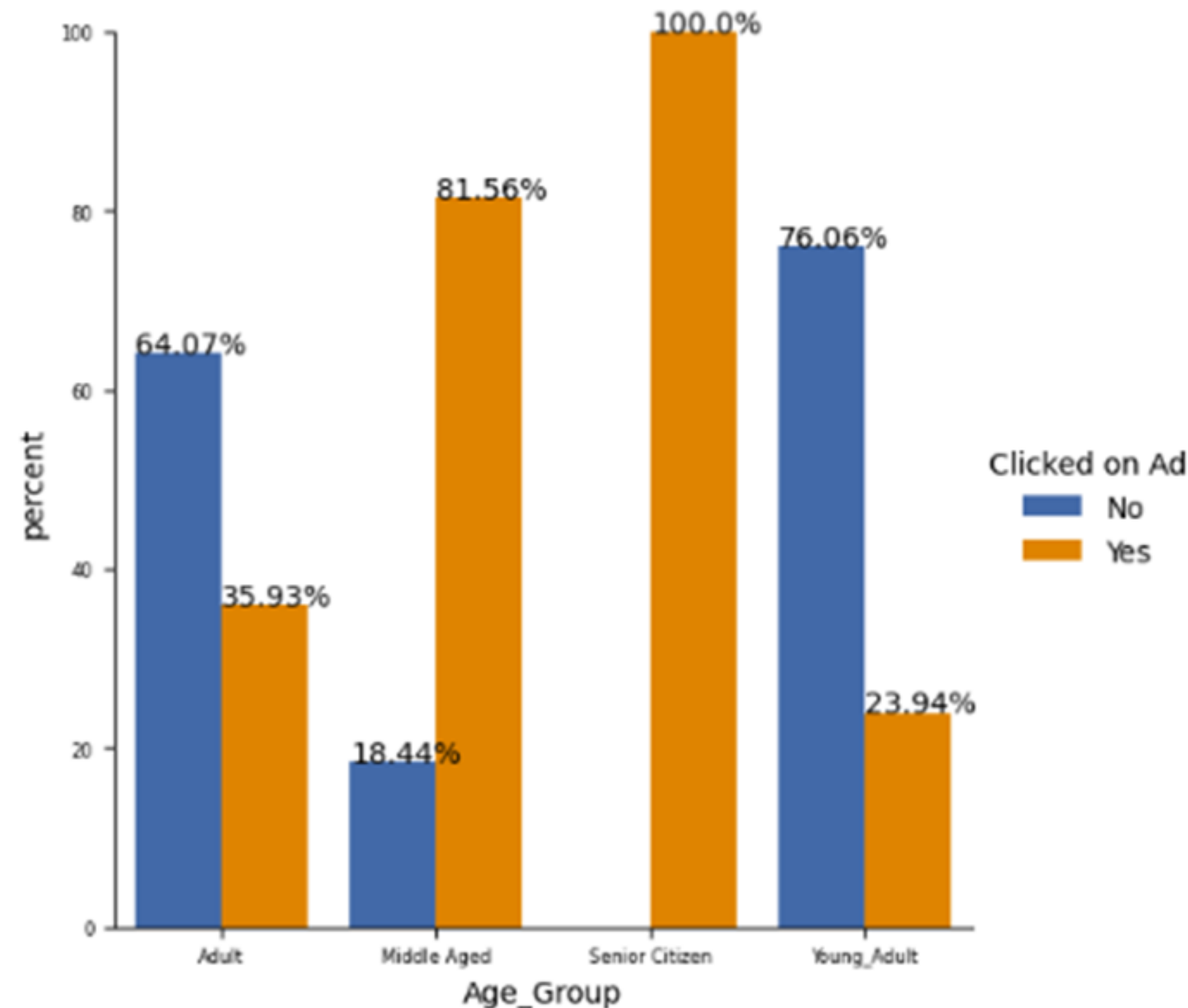
The internet usage and the duration of user visits on a website have a similar distribution. That is, potential users can be found when they visit a website for only a short duration.

CLICKED ON AD & TIME SPENT



Based on the visualization results, it can be observed that users who have infrequent internet usage have a higher potential for clicking on ads compared to users who have frequent internet usage. This suggests that users with low internet usage tend to pay more attention to the presence of ads on a website.

CLICKED ON AD & AGE



From the visualization, it turns out that the potential market is actually among older people. This may be because young people are more careful and selective when browsing the internet. And young people are very quick to notice when there are advertisements on a website.

DATA CLEANSING & PREPROCESSING

1

Missing Value

handled missing values in columns Daily Time Spent on Site, Area Income, Daily Internet Usage, Male (Gender).

2

Duplicated Data

There is no duplicated data

3

Outliers

Handled outliers with IQR to make sure all features are stable for modelling.

4

Encoding

Handled with one-hot encoding

5

Split Data

Data Train 80%
Data Test 20%

6

Scaling

Use StandardScaler

MODELING

Model Metrics	Logistic Regression	Random Forest	Desicion Tree
Accuracy	0.97	0.97	0.95
Precision	0.99	1.00	0.97
Recall	0.96	0.94	0.93
ROC-AUC	1.00	1.00	0.95
F1	0.97	0.7	0.95

USE RECALL

Using Recall to measure how accurately the model can identify users who actually clicked on ads. Recall measures the number of users who actually clicked on ads and were correctly identified by the model.

The formula to calculate recall is:

$$\text{Recall} = \text{True Positive} / (\text{True Positive} + \text{False Negative})$$

True Positive (TP) is the number of users who actually clicked on ads and were correctly identified by the model as ad clickers. False Negative (FN) is the number of users who actually clicked on ads but were incorrectly identified by the model as non-ad clickers.

By using recall, companies can understand how well the model can recognize users who are genuinely interested in clicking on ads. If the recall is high, it means the model successfully identifies a large portion of users who are actually interested in clicking on ads. However, if the recall is low, the company may miss out on a number of potential users who are actually interested in the ads.

By increasing recall, companies can maximize the effectiveness of their ad campaigns by targeting more relevant users and increasing the likelihood of converting users into customers

HYPARAMETER TUNING

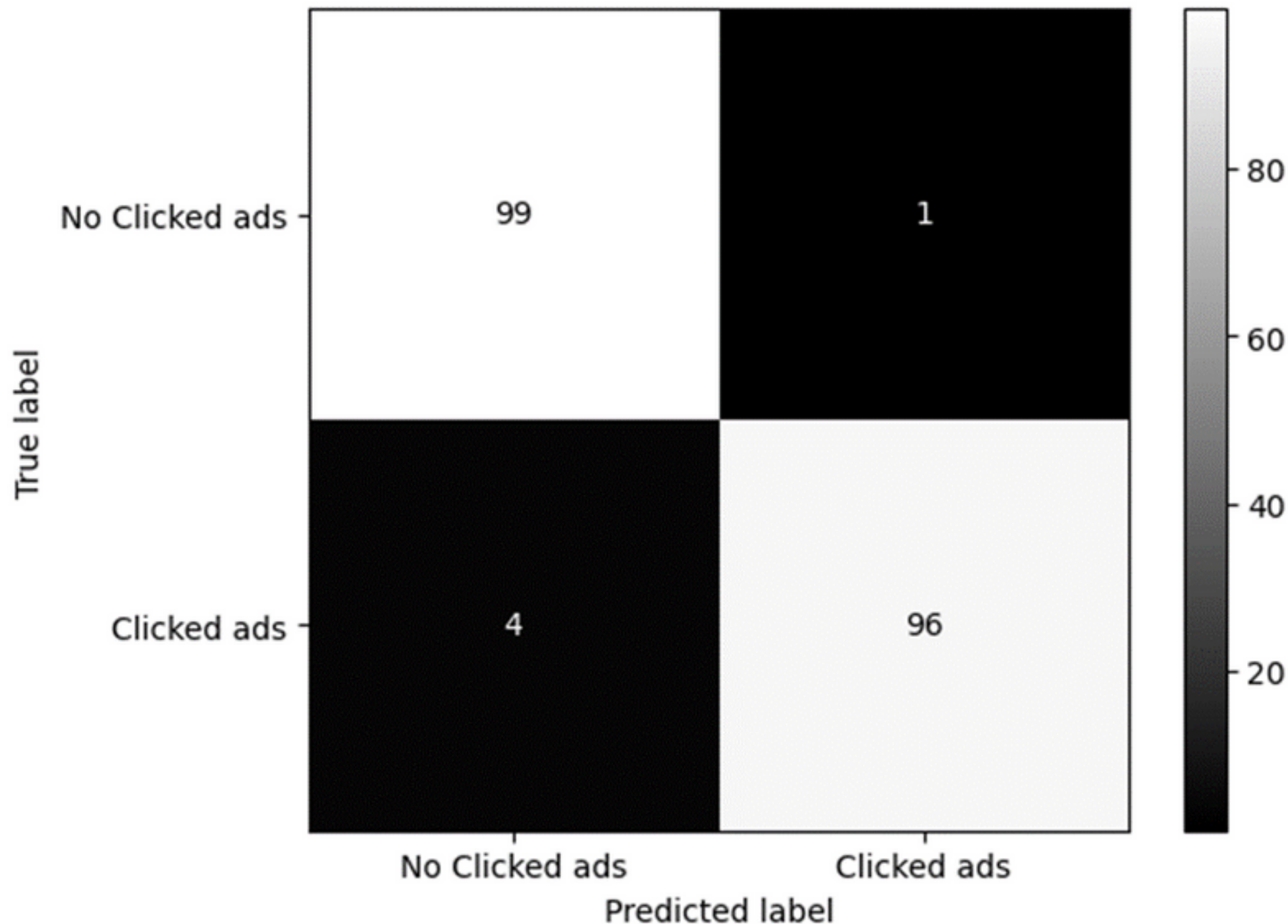
MODEL	RECALL	
	TRAIN	TEST
Logistic Regresion	0.98	0.96

MODEL	RECALL	
	TRAIN	TEST
Logistic Regresion	0.99	0.96

TUNING

IN THIS CASE, BOTH BEFORE AND AFTER HYPERPARAMETER TUNING, THE TRAIN AND TEST SCORES HAVE A VERY SMALL DIFFERENCE. IN THIS SITUATION, THERE IS NO SIGNIFICANT DIFFERENCE IN THE MODEL'S PERFORMANCE BEFORE AND AFTER HYPERPARAMETER TUNING.

CONFUSION MATRIX



Before implementing machine learning, approximately 100 out of a total of 200 customers did not click on the advertisement but were predicted as customers who would click on it. This situation caused the company to incur a 50% loss because these customers were not actually interested in the advertisement but were still treated as potential customers.

However, after implementing machine learning, the company successfully reduced the loss significantly to only 0.5%. This indicates that the machine learning model can accurately predict customers who are not interested in the advertisement, allowing the company to optimize resources and marketing strategies more effectively. Therefore, the use of machine learning brings significant benefits to the company by reducing losses resulting from inaccurate targeting in the past.

MODEL BUSINESS SIMULATION

Assumptions:

- Cost of ads = Rp. 10,000
- Number of users = 200
- Conversion value per customer = Rp. 40,000

Without Machine Learning:

- Cost = cost of ads * number of users
- Cost = Rp. 10,000 * 200
- Cost = Rp. 2,000,000

Only 50 customers convert, so $50 * \text{Rp. } 40,000 = \text{Rp. } 2,000,000$

- Revenue = Rp. 2,000,000
- Profit = Rp. 2,000,000 - Rp. 1,000,000 = Rp. 1,000,000
- Loss percentage = $(\text{profit} / \text{cost}) * 100$
- Loss percentage = $(\text{Rp. } 1,000,000 / \text{Rp. } 2,000,000) * 100 = 50\%$

Without using machine learning, the company incurs a 50% loss.

With Machine Learning:

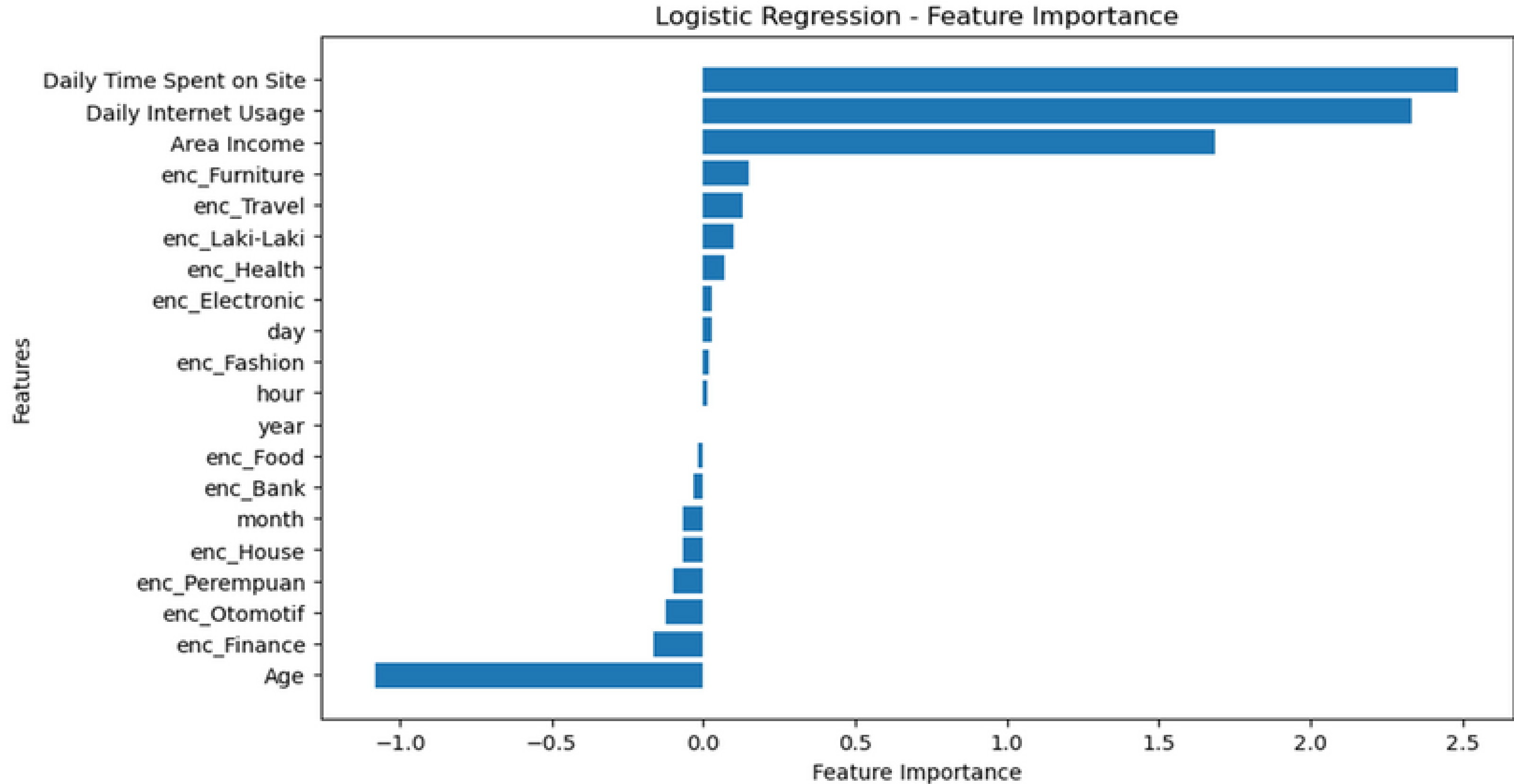
- Cost = cost of ads * number of users
- Cost = Rp. 10,000 * 200
- Cost = Rp. 2,000,000

The conversion rate increases to 96% (192 users)

- Revenue = $\text{Rp. } 40,000 * 192 = \text{Rp. } 7,680,000$
- Profit = Revenue - Cost
- Profit = $\text{Rp. } 7,680,000 - \text{Rp. } 2,000,000$
- Profit = Rp. 5,680,000

Using machine learning significantly increases the revenue

FEATURE IMPORTANCE



BUSINESS RECOMENDATION

Based on EDA and feature importance, we can conclude that:

- The data we obtained has 2 user segments, namely the upper class and lower class user segments.
 - The upper class users have the characteristics of frequent internet usage, frequent visits to product websites, relatively young age, and high income.
 - The lower class users have the opposite characteristics.
- Users from the lower economic class tend to be more easily interested in clicking on digital ads.
- Users who heavily use the internet may be more difficult to target with ads because they might already be accustomed to digital ads.
- The older generation is a potential market for digital marketing.