Predict Ads Click

TEAM MEMBERS

Jai Prakash	**G Ashok**	**Hema Varshini**	**K Vaisakh**	**Pranavi**
CB.EN.U4CSE20027	CB.EN.U4CSE20020	CB.EN.U4CSE20026	CB.EN.U4CSE20069	CB.EN.U4CSE20063

Agenda

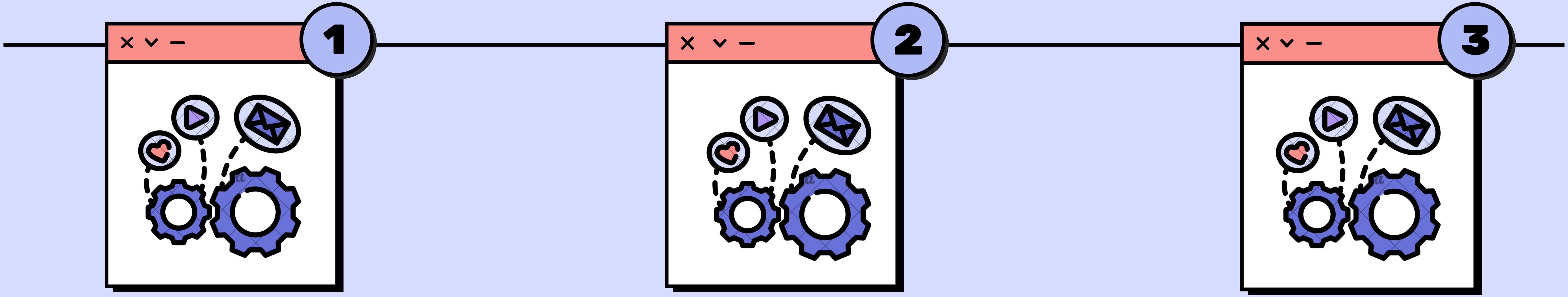
1. Objective
2. Problem Statement
3. Write your agenda point
4. Data Description
5. Data Visualization
6. Baseline Models
7. Conclusion

Objective

In this article, we will work with the advertising data of a marketing agency to develop a machine learning algorithm that predicts if a particular user will click on an advertisement.

The data consists of 10 variables: 'Daily Time Spent on Site', 'Age', 'Area Income', 'Daily Internet Usage', 'Ad Topic Line', 'City', 'Male', 'Country', Timestamp' and 'Clicked on Ad'.

Problem Statement

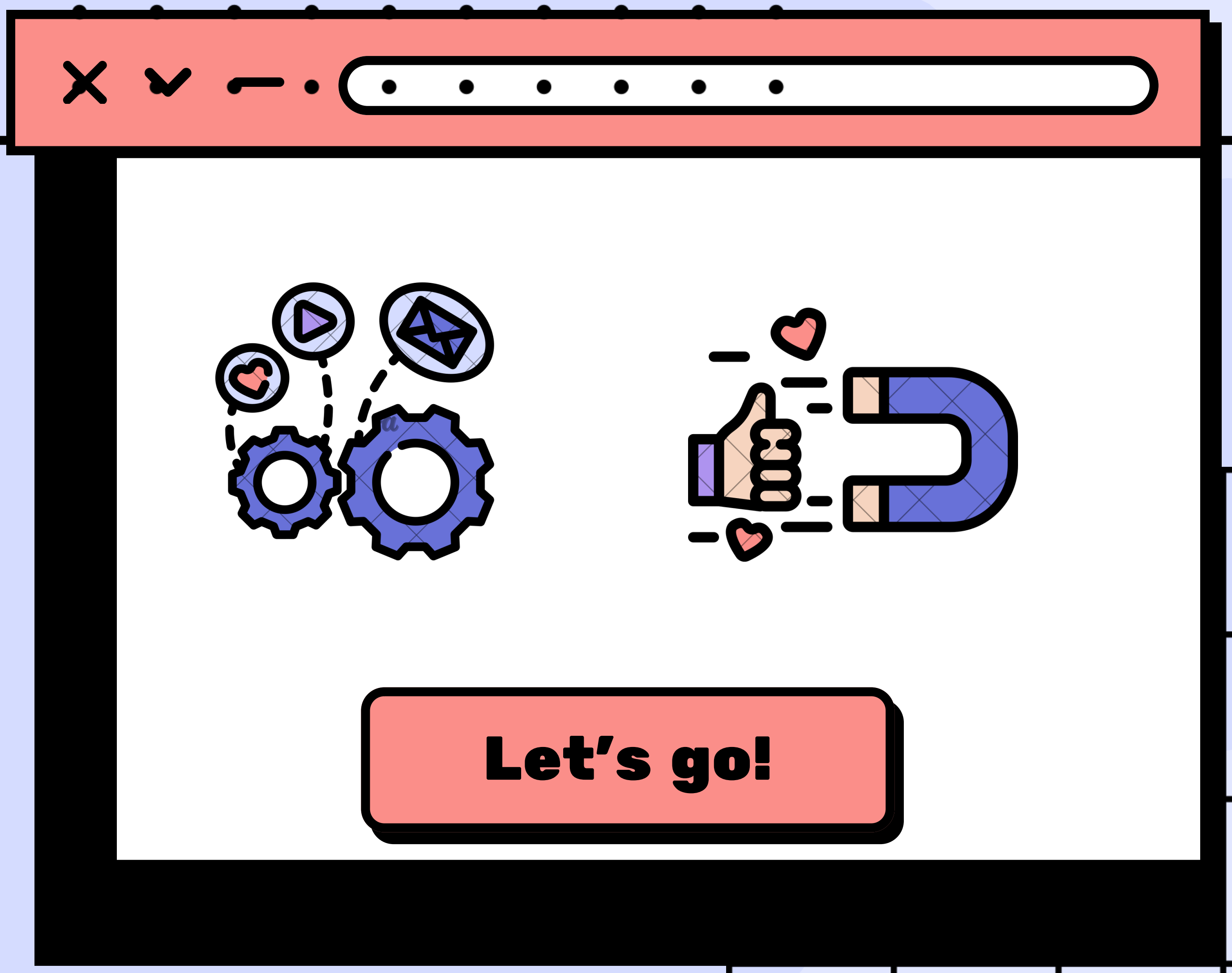


The main variable we are interested in is 'Clicked on Ad'. This variable can have two possible outcomes: 0 and 1 where 0 refers to the case where a user didn't click the advertisement, while 1 refers to the scenario where a user clicks the advertisement.

We will see if we can use the other 9 variables to accurately predict the value 'Clicked on Ad' variable. We will also perform some exploratory data analysis to see how 'Daily Time Spent on Site' in combination with 'Ad Topic Line' affects the user's decision to click on the add.

1

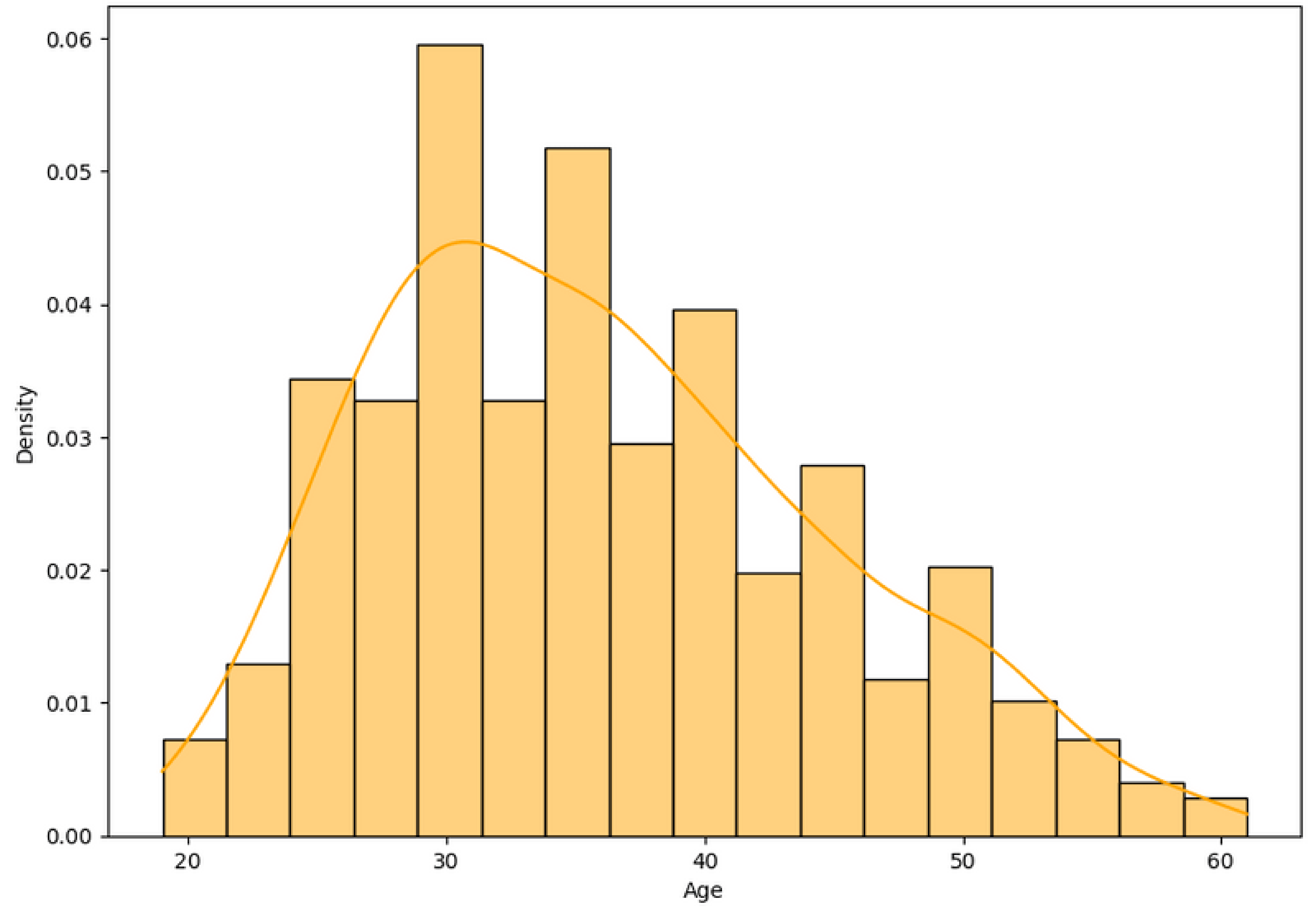
Data Description



```
Daily Time Spent on Site  1000 non-null   float64
Age                       1000 non-null   int64
Area Income               1000 non-null   float64
Daily Internet Usage      1000 non-null   float64
Ad Topic Line             1000 non-null   object
City                      1000 non-null   object
Male                      1000 non-null   int64
Country                   1000 non-null   object
Timestamp                 1000 non-null   object
Clicked on Ad             1000 non-null   int64
```

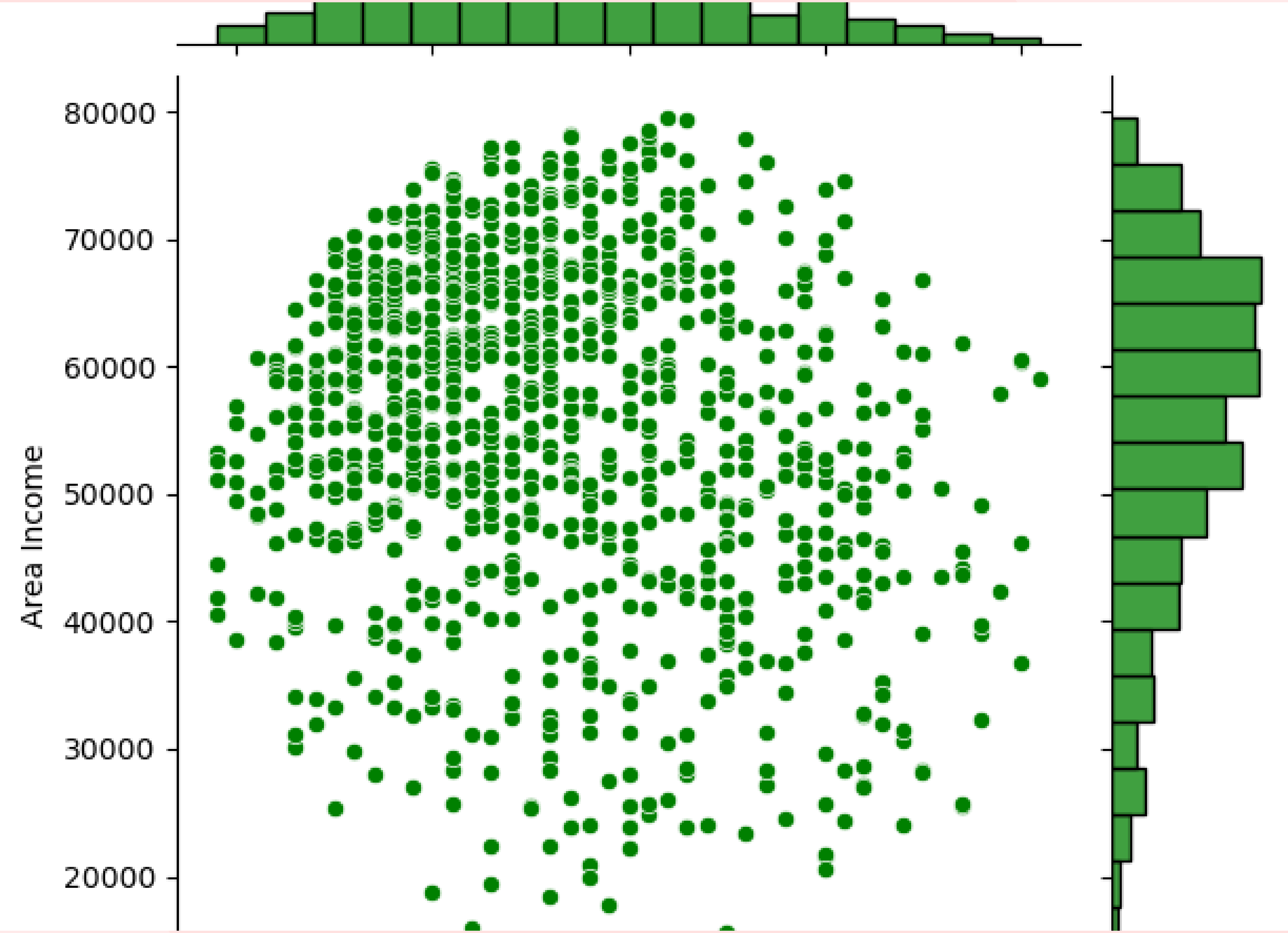
Data Visualization

Distplot



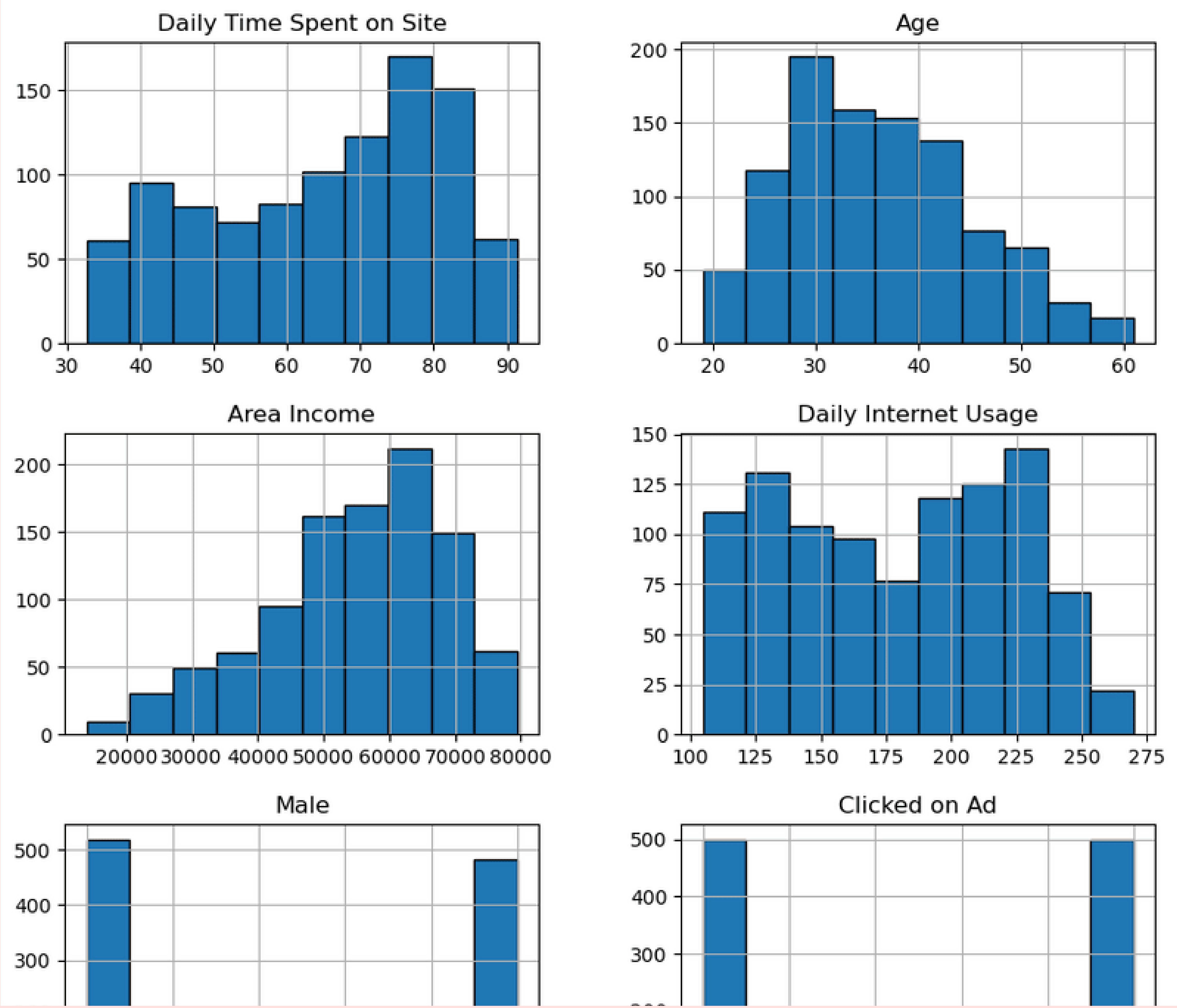
Data Visualization

Jointplot



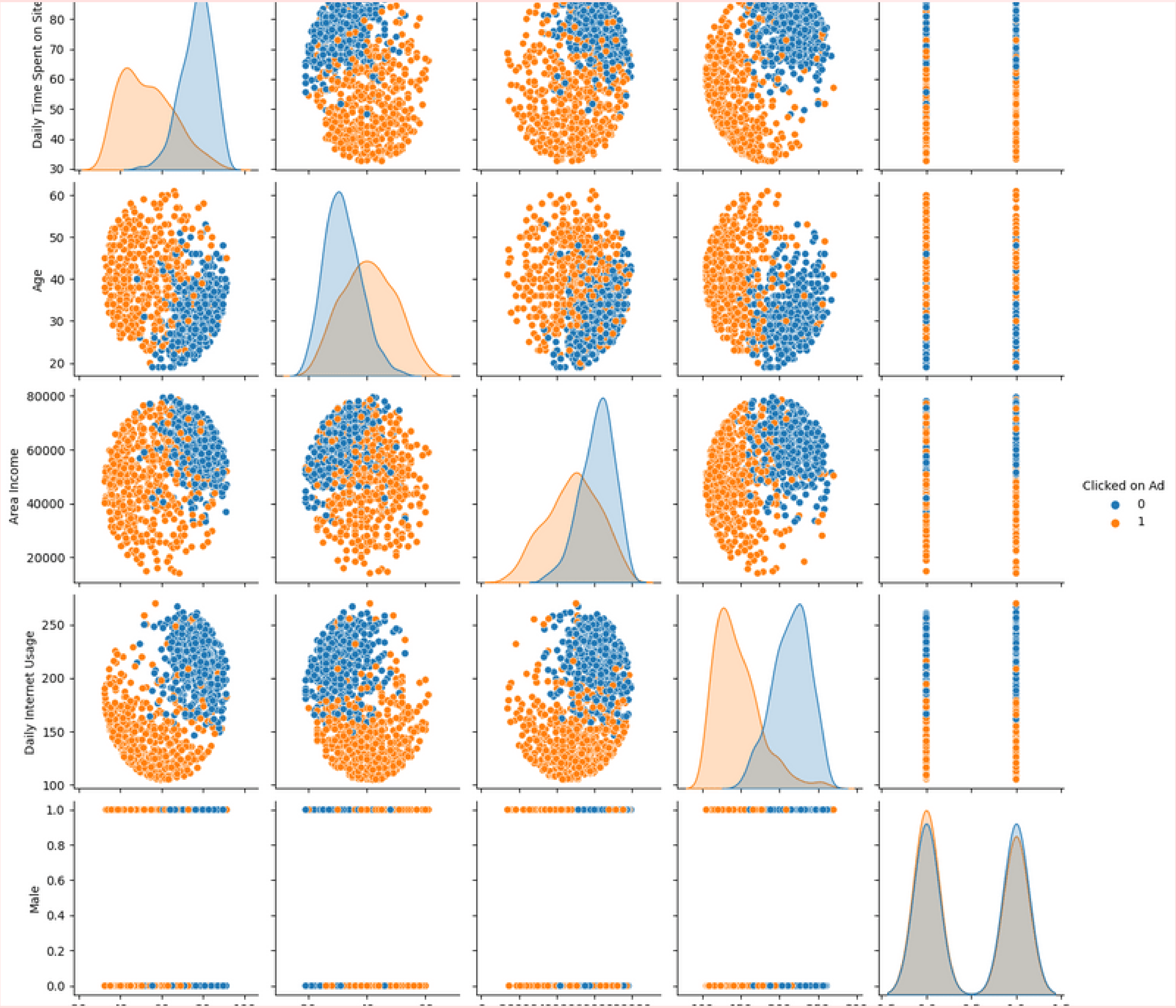
Data Visualization

Histogram



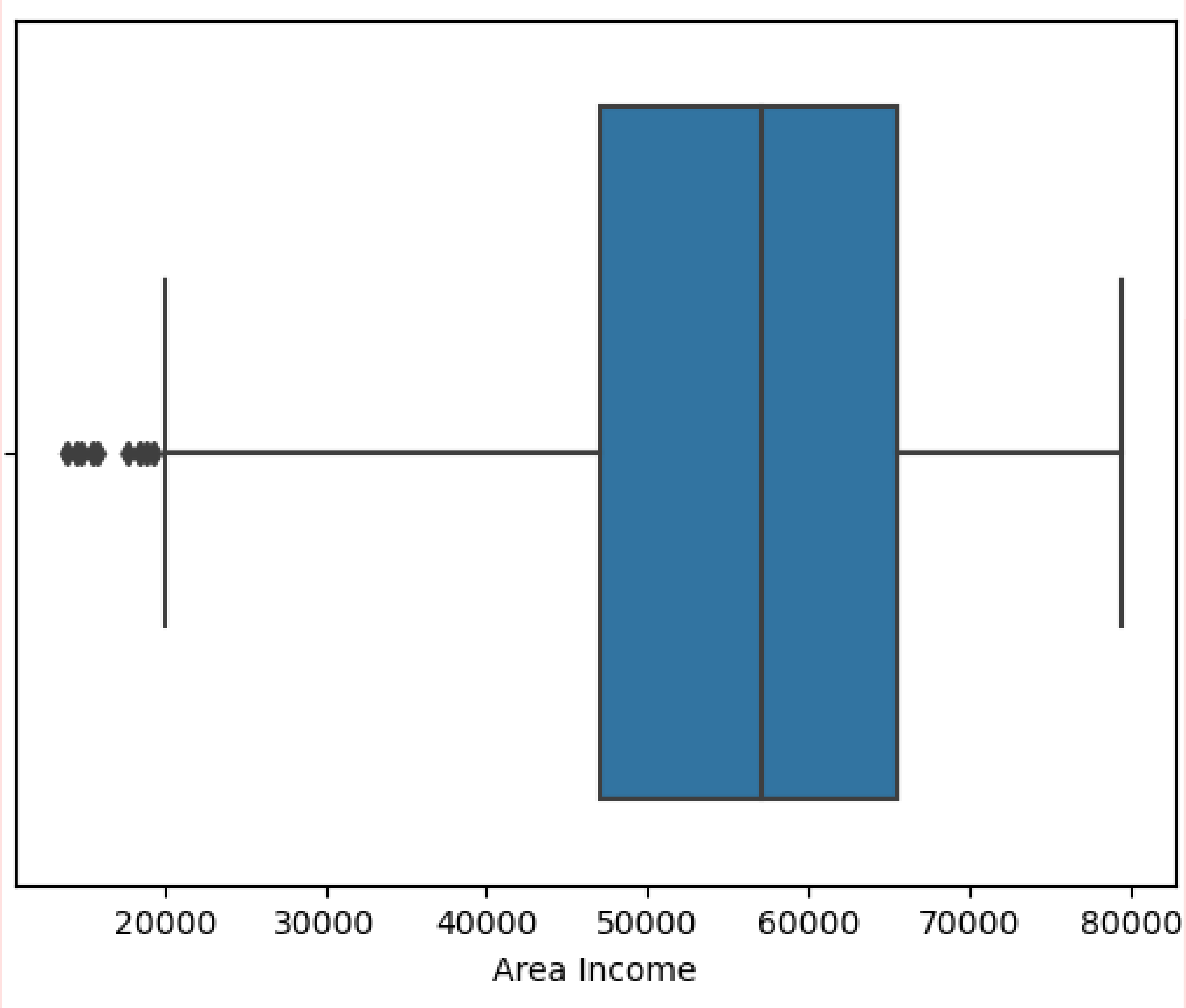
Data Visualization

Pairplot



Data Visualization

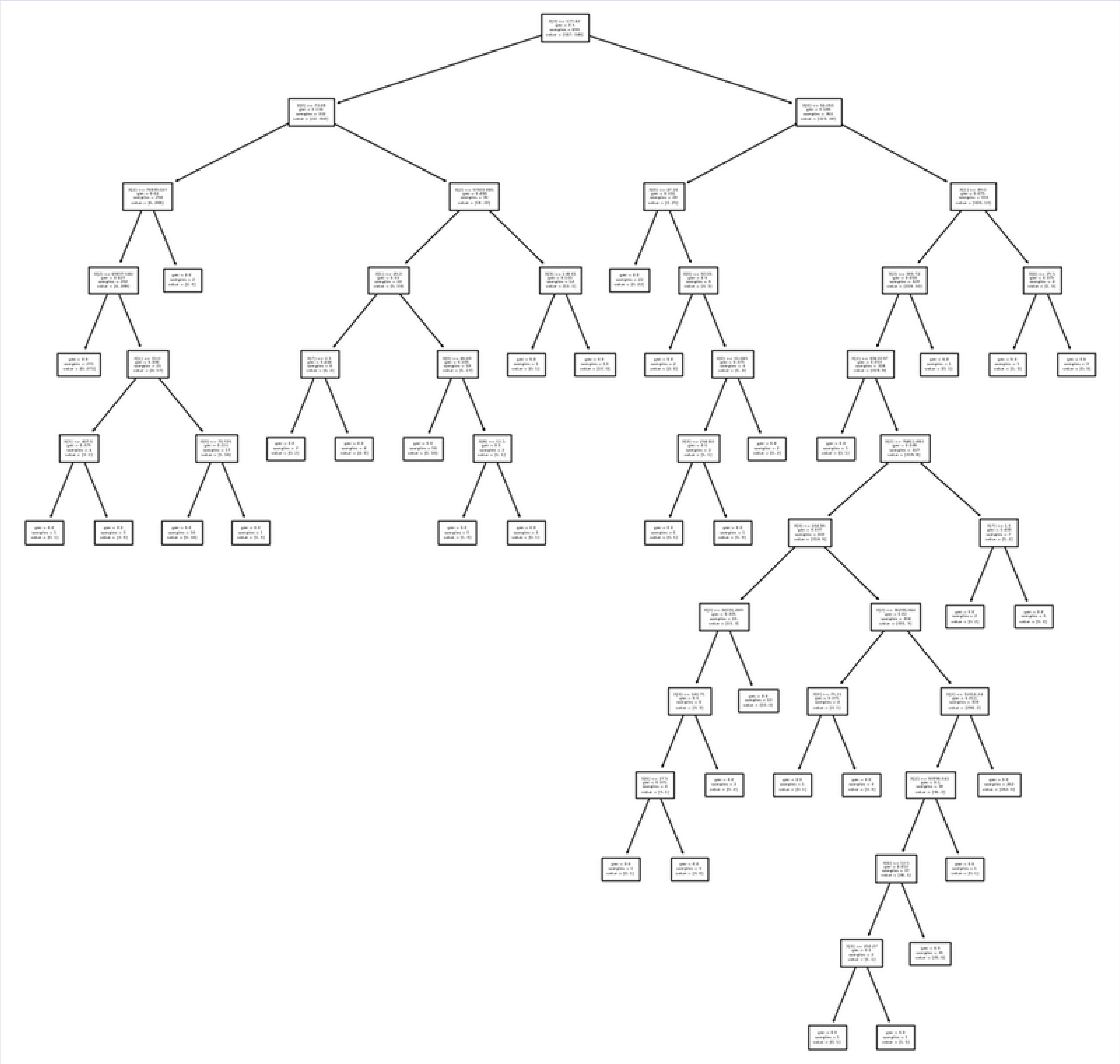
Boxplot



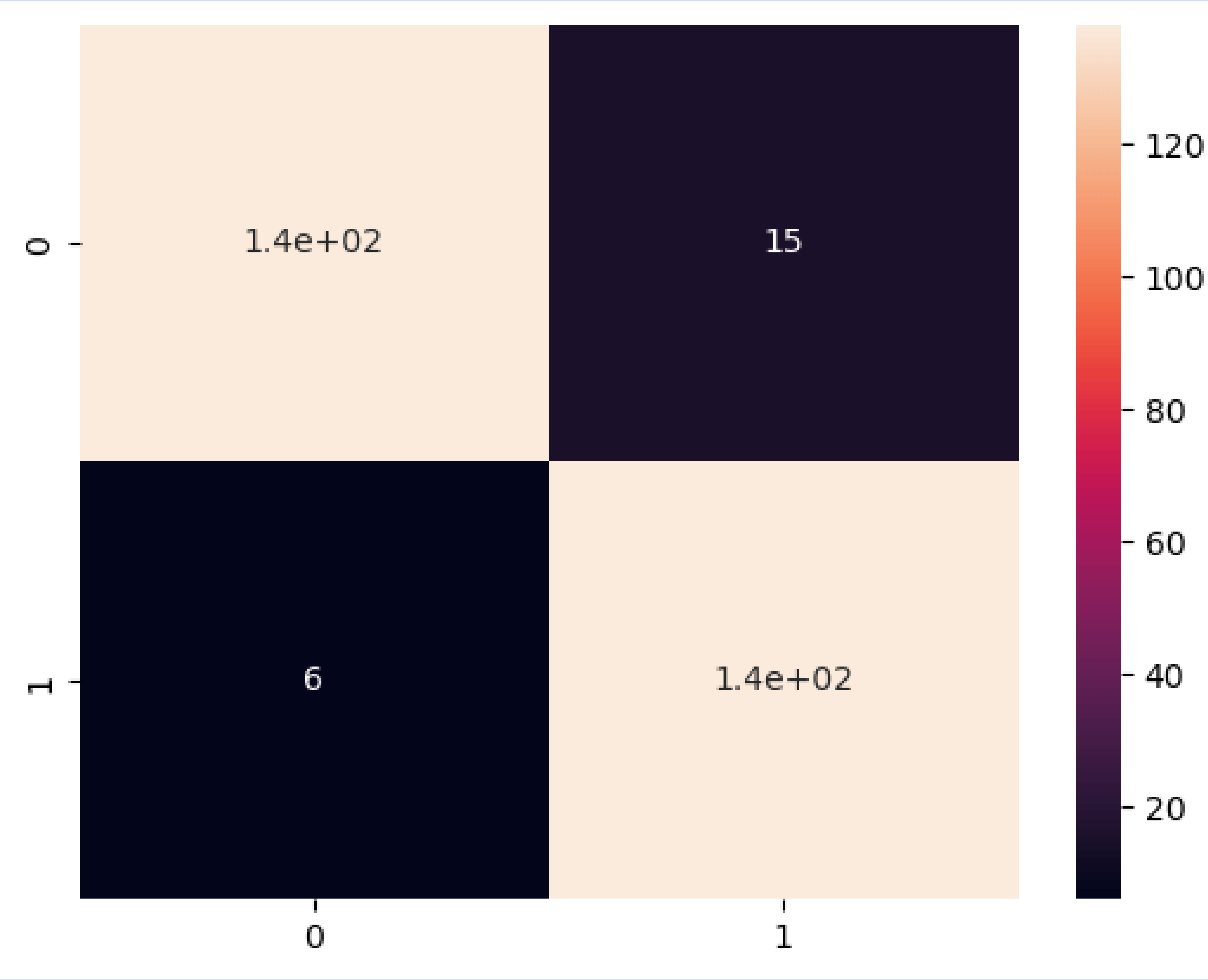
“

Base Models

Decision Tree



Confusion Matrix



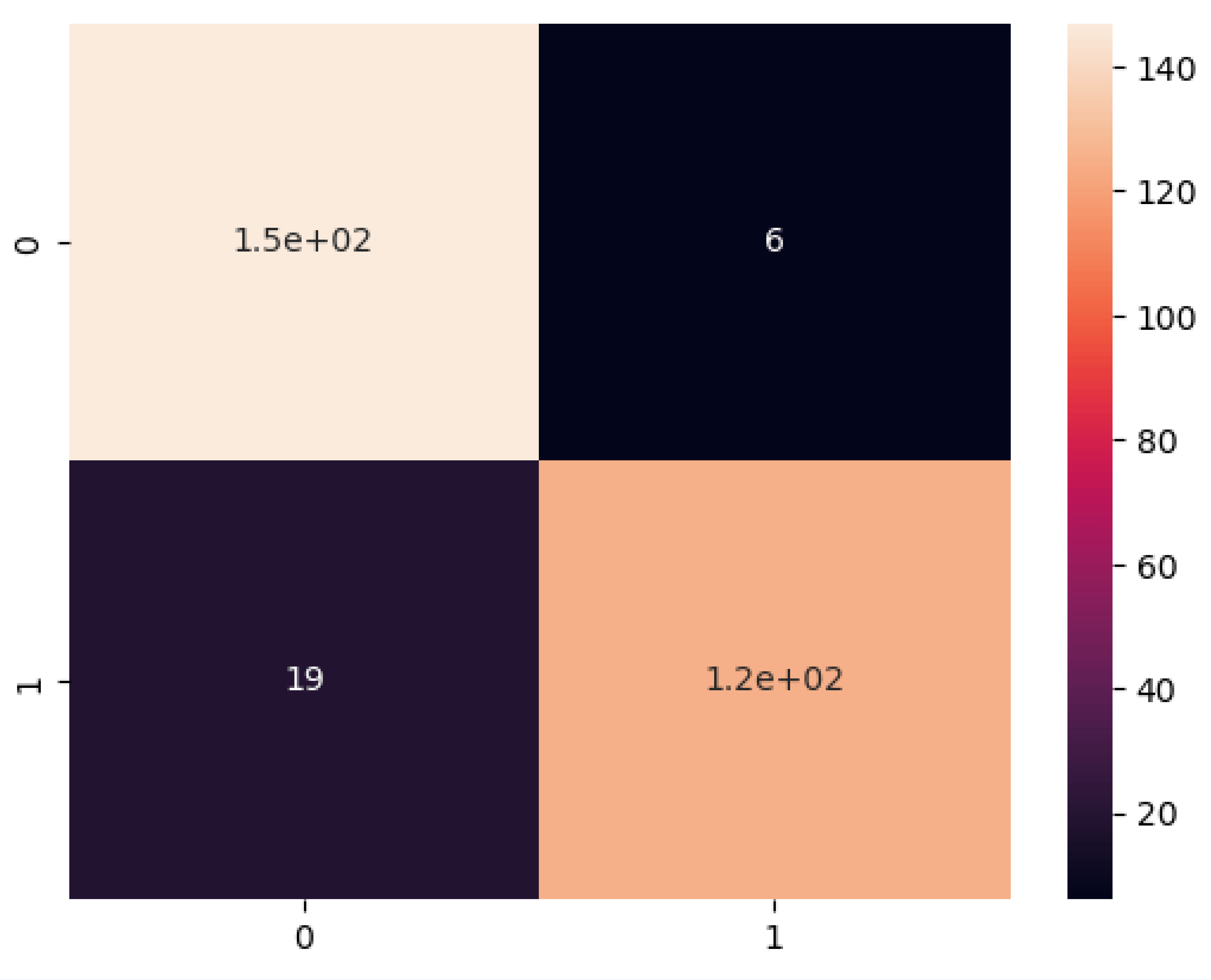
Base Models

Logistic Regression

Since the output variable is Ad Clicked whose values are either 0 or 1 so it is a binary classification problem.So we can use logistic regression Logistic regression is a type of regression we can use when the response variable is binary.

To evaluate the quality of a logistic regression model is to create a confusion matrix, which is a 2×2 table that shows the predicted values from the model vs. the actual values from the test dataset

Confusion Matrix



“

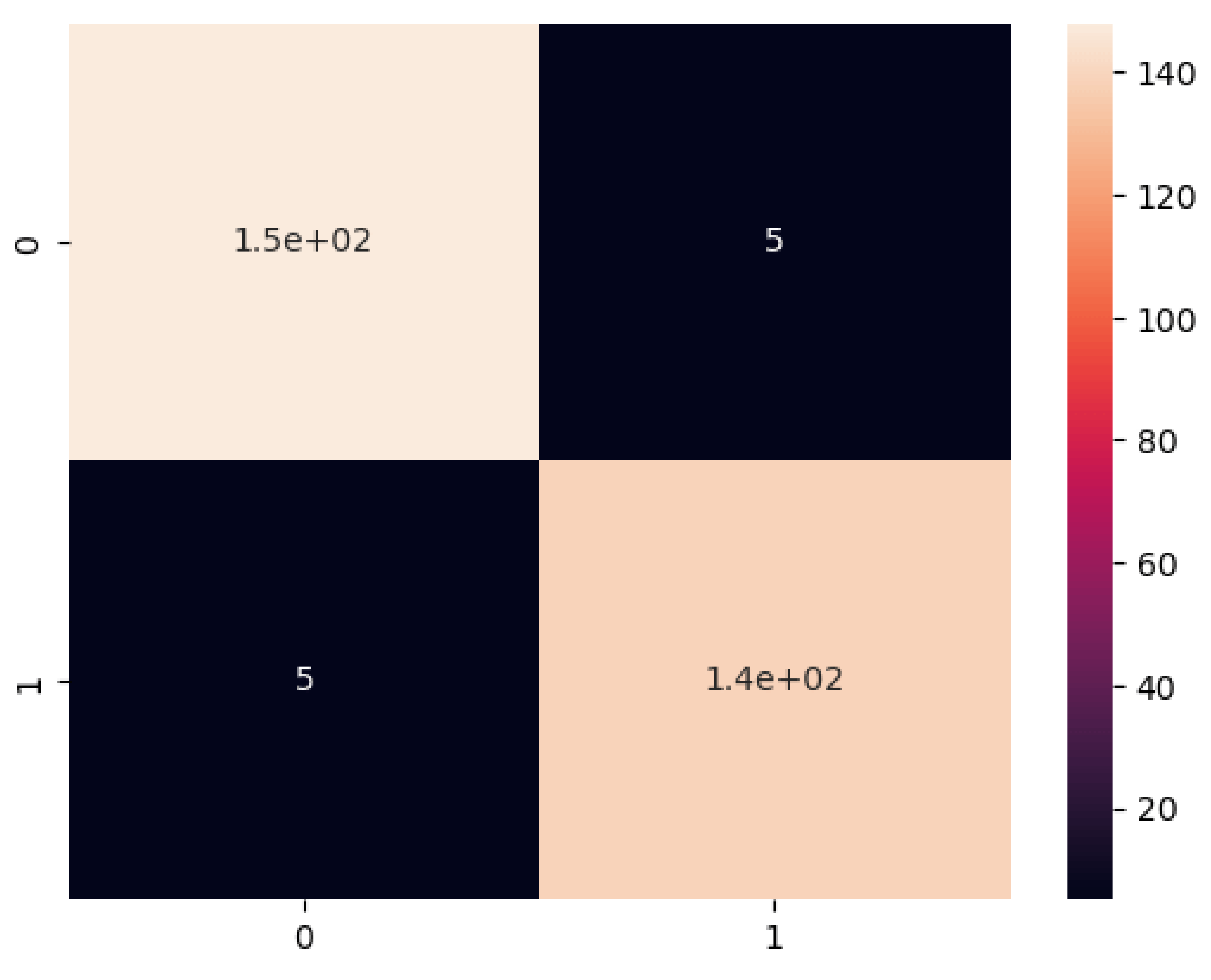
Base Models

Naive Bayes

A classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

The Naive Bayes Algorithm has trouble with the ‘zero-frequency problem’. It happens when you assign zero probability for categorical variables in the training dataset that is not available. When you use a smooth method for overcoming this problem, you can make it work the best.

Confusion Matrix



Base Models

KNN

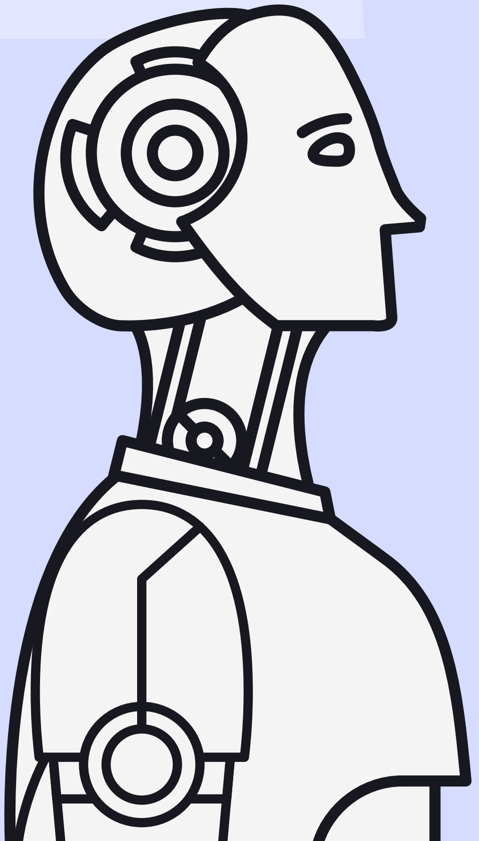
The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.

K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

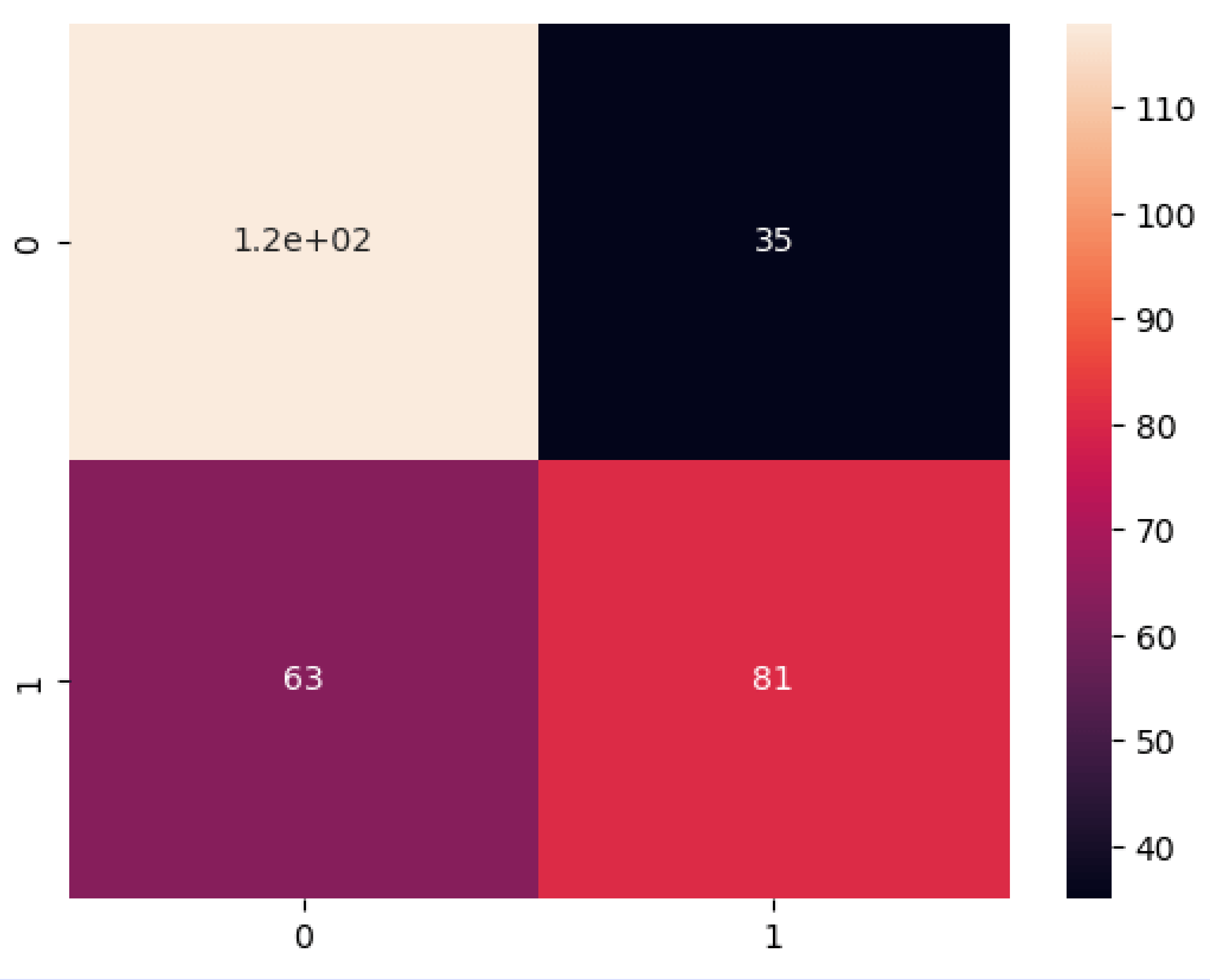
K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

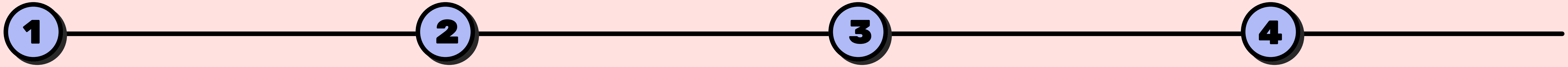
KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

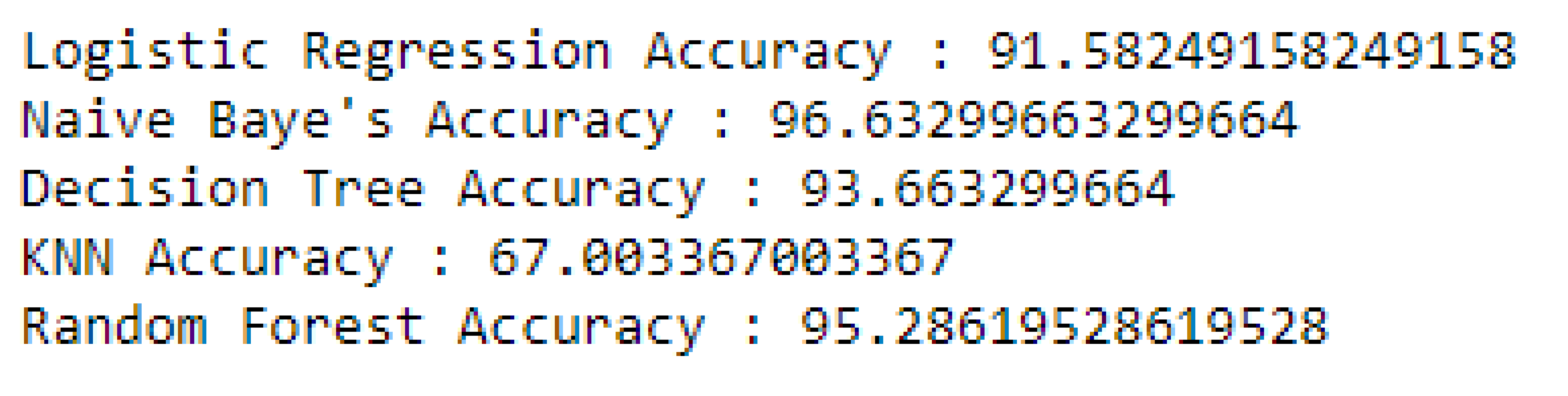


Confusion Matrix



Performance





```
Logistic Regression Accuracy : 91.58249158249158
Naive Baye's Accuracy : 96.63299663299664
Decision Tree Accuracy : 93.663299664
KNN Accuracy : 67.003367003367
Random Forest Accuracy : 95.28619528619528
```

Conclusion

- In this project, we developed an Ad prediction system that detects Ads click using realtime machine learning techniques.
- It can be concluded that the "Naive Bayes" model showed better performances in comparison to the Logistic Regression model. The confusion matrix shows us that the 140 predictions have been done correctly and that there are only 12 incorrect predictions. Additionally, Naive Bayes accuracy is better by about 3% in comparison to the first regression model.
- Data standardization, such as derivation standardization, is a useful way for improving performance,such as accuracy.
- The performance of the proposed system was verified by using different parametric measure as accuracy, sensitivity and specificity.

Thank you!