CAPSTONE PROJECT

ML PROJECT

SAMSUNG INNOVATION CENTER

BY

S.Vithal-119cs0010

Agenda

1. Problem Statement
2. Data Given in a Dataset
3. Visualization with inference
4. Steps done
5. Choosen Model
6. Conclusion

Problem Statement

Hotel Reservation:

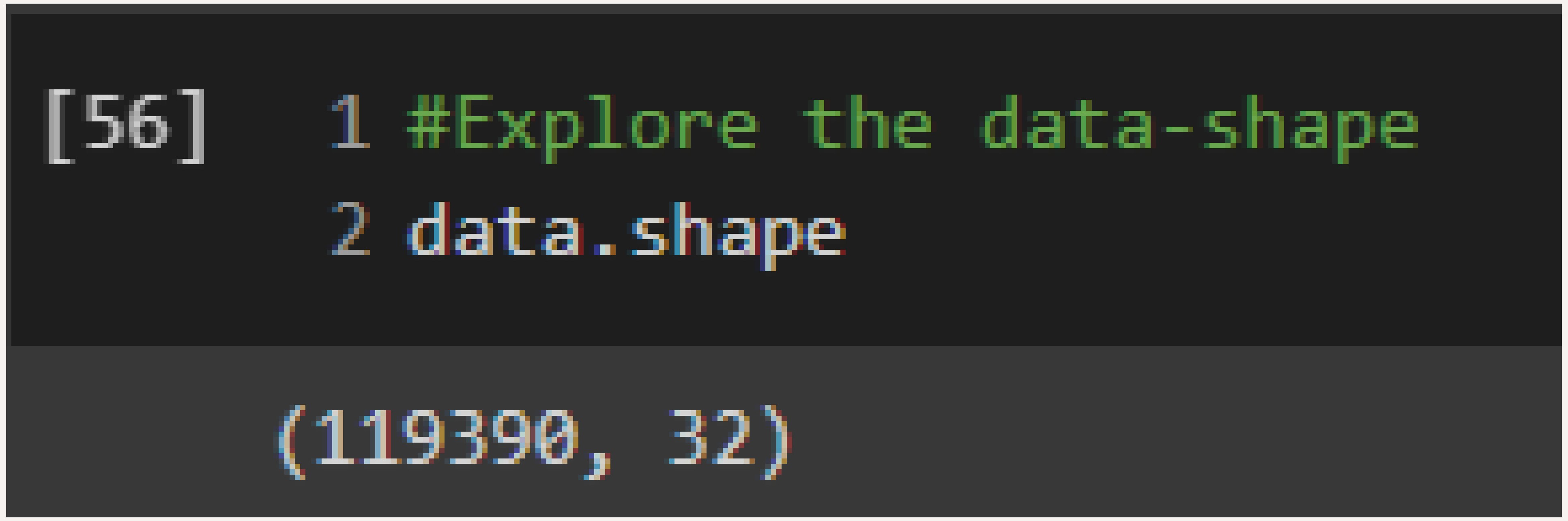
A machine learning model to predict whether the customer cancels there hotel reservation or not . By using given dataset containing data of reservations made by customers from different places in different hotels.

Data given in Dataset

Given a Dataset Hotel_Bookings.csv contains 119390 rows × 32 columns which describes all the features of the hotel and booking details including their Arrival Timings, booking stats from which country they are from, through which agent they have booked etc..

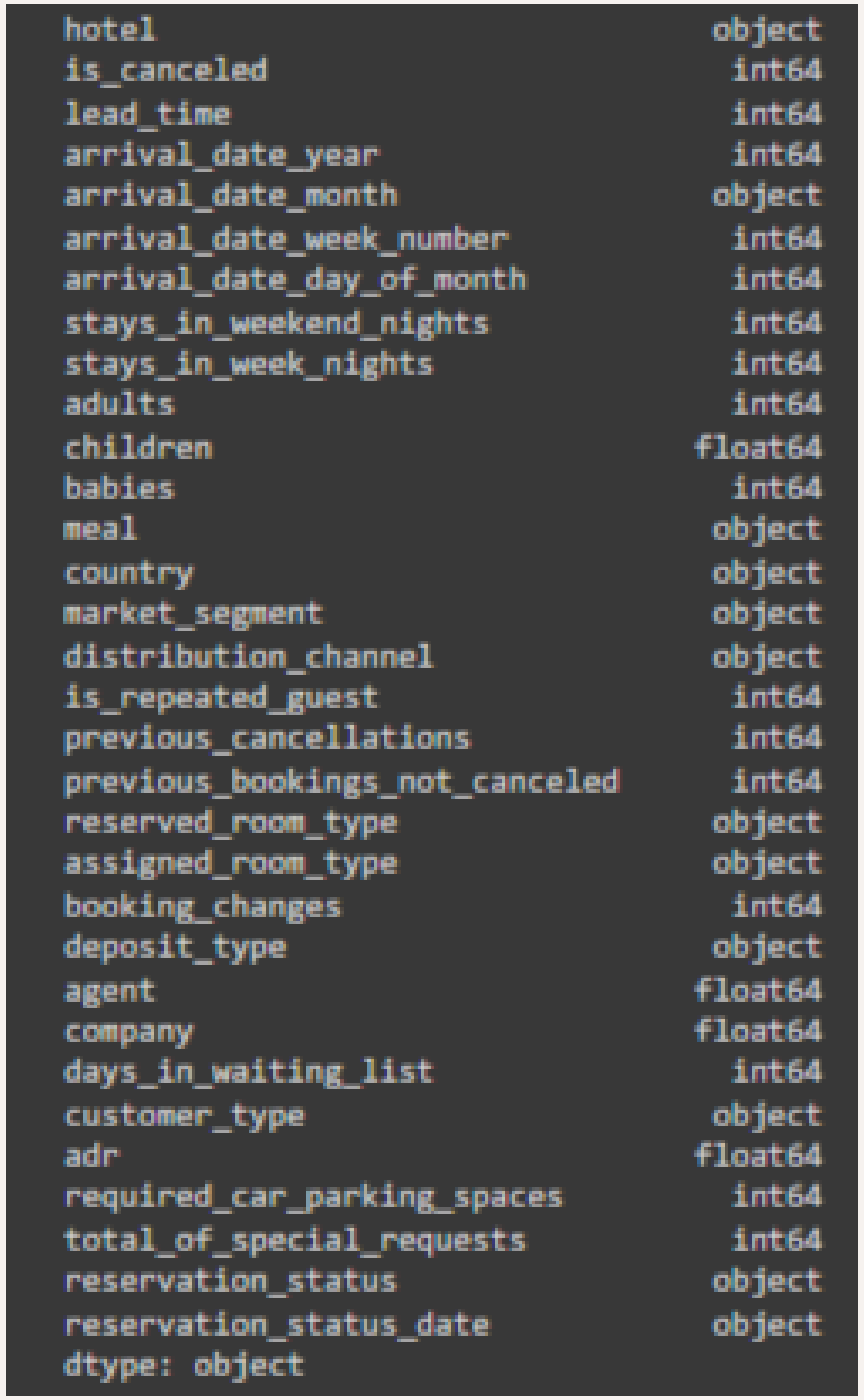
DATASET USED:
https://raw.githubusercontent.com/Premalatha-success/Datasets/main/hotel_bookings.csv

Datatypes and DataShape in Data



```
[56]  #Explore the data-shape
      data.shape

(119390, 32)
```

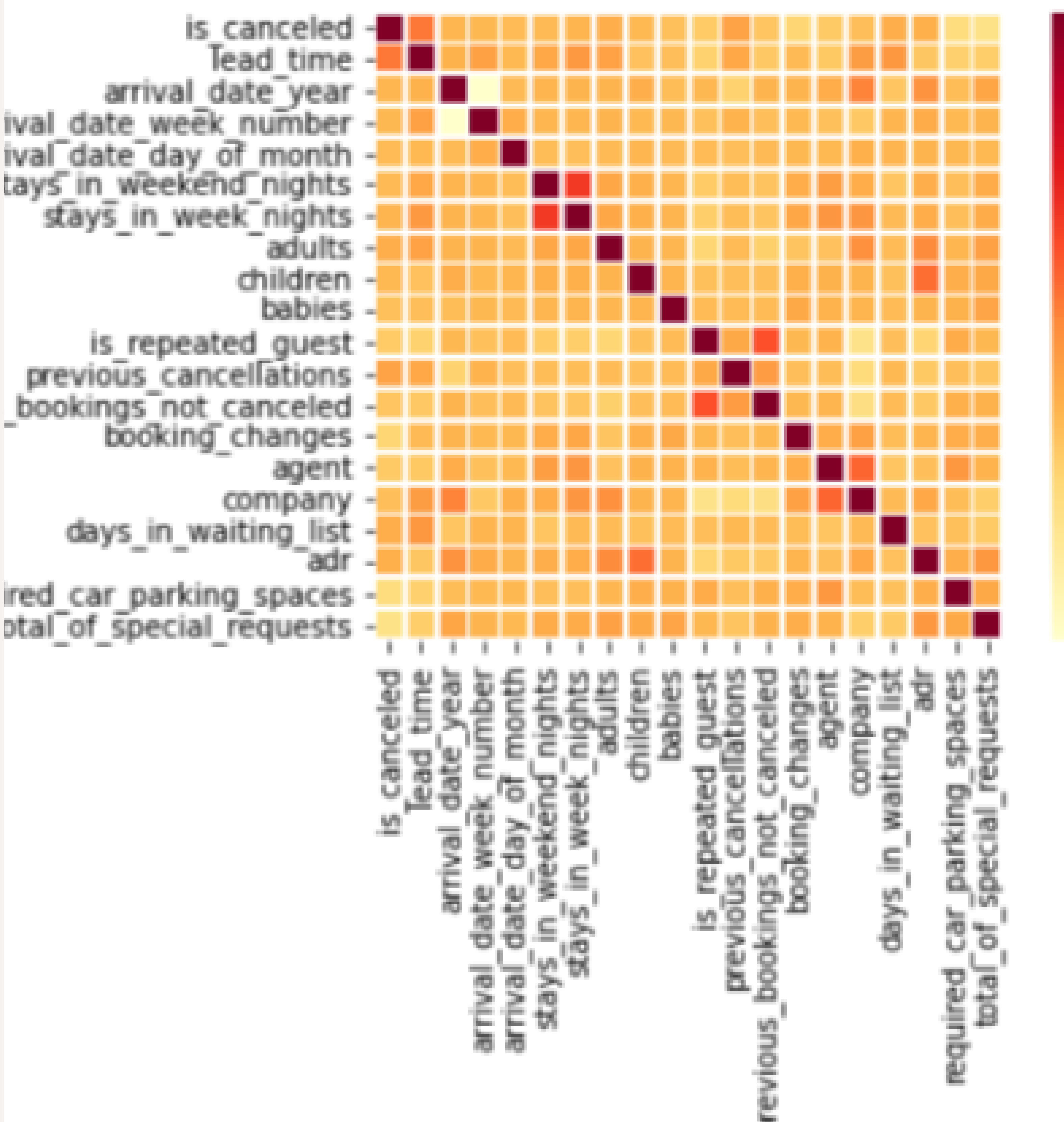


```
hotel                              object
is_canceled                         int64
lead_time                           int64
arrival_date_year                   int64
arrival_date_month                 object
arrival_date_week_number            int64
arrival_date_day_of_month           int64
stays_in_weekend_nights             int64
stays_in_week_nights                int64
adults                              int64
children                          float64
babies                              int64
meal                               object
country                            object
market_segment                     object
distribution_channel               object
is_repeated_guest                   int64
previous_cancellations              int64
previous_bookings_not_canceled      int64
reserved_room_type                 object
assigned_room_type                 object
booking_changes                     int64
deposit_type                       object
agent                             float64
company                           float64
days_in_waiting_list                int64
customer_type                      object
adr                               float64
required_car_parking_spaces         int64
total_of_special_requests           int64
reservation_status                 object
reservation_status_date            object
dtype: object
```

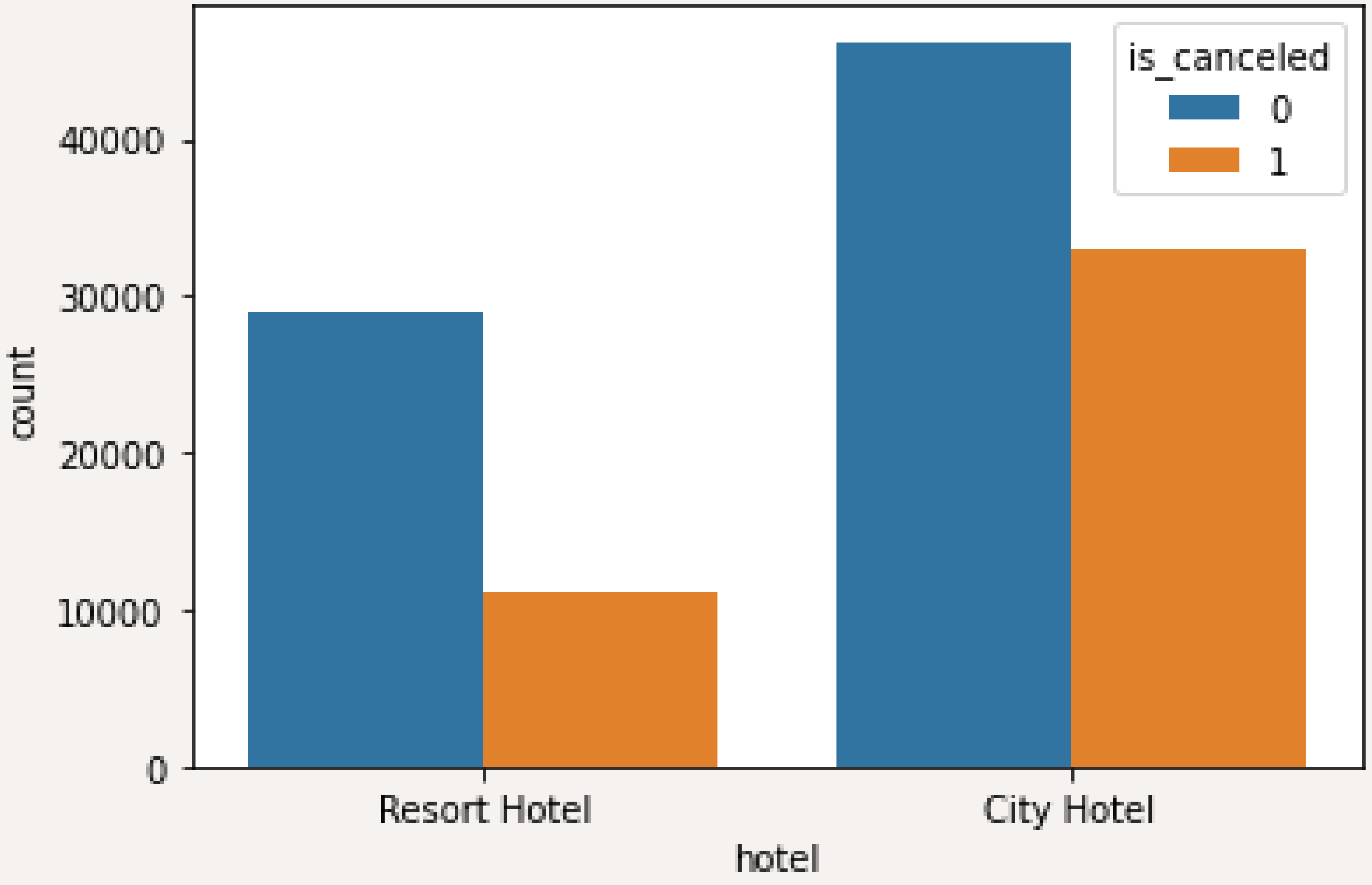
Visualization

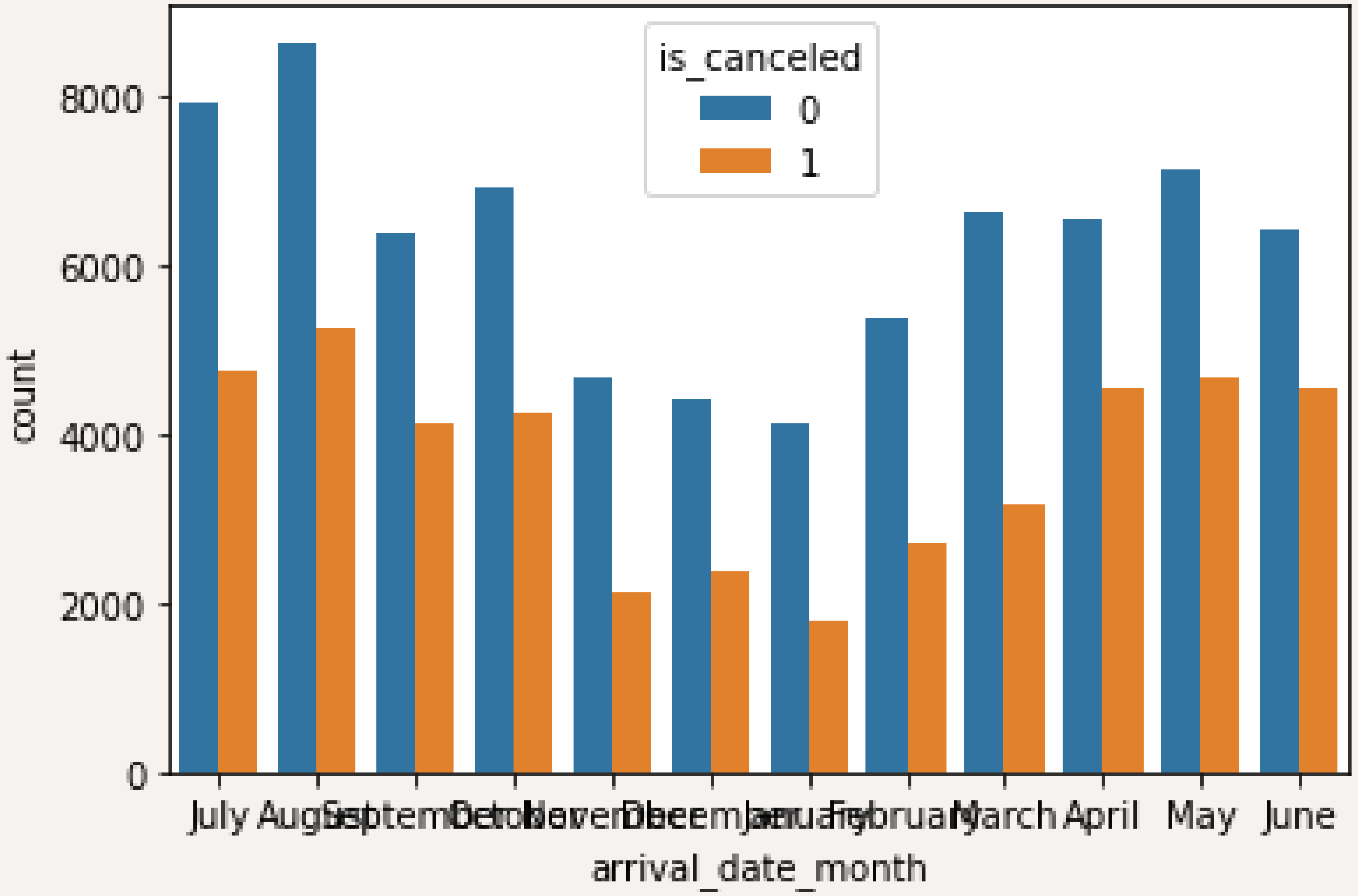
Correlation Visualization

Correlation visualization summarizes the association between two variables. I will be ranging from -1 to +1

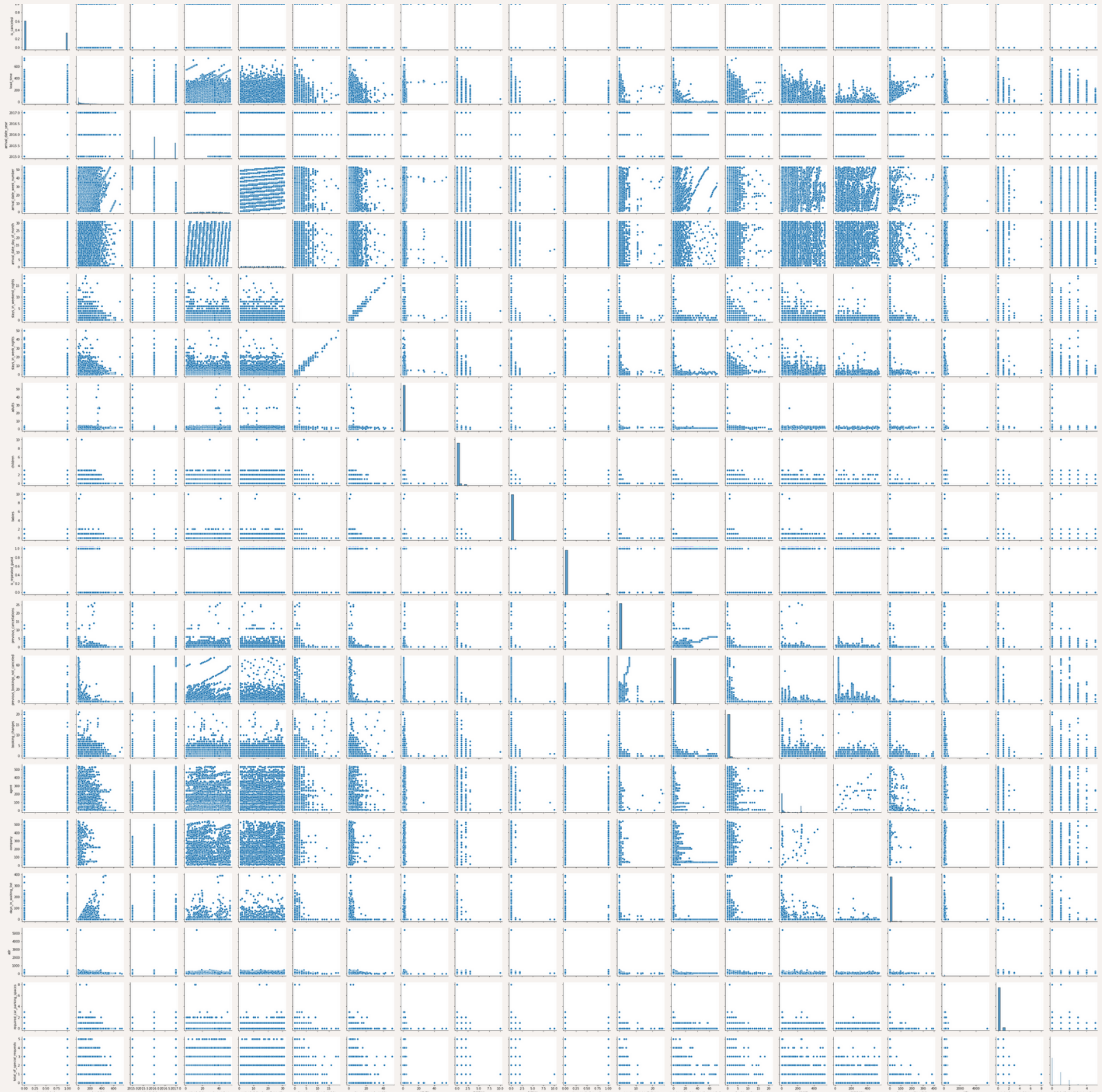


Visualizing from feature to feature

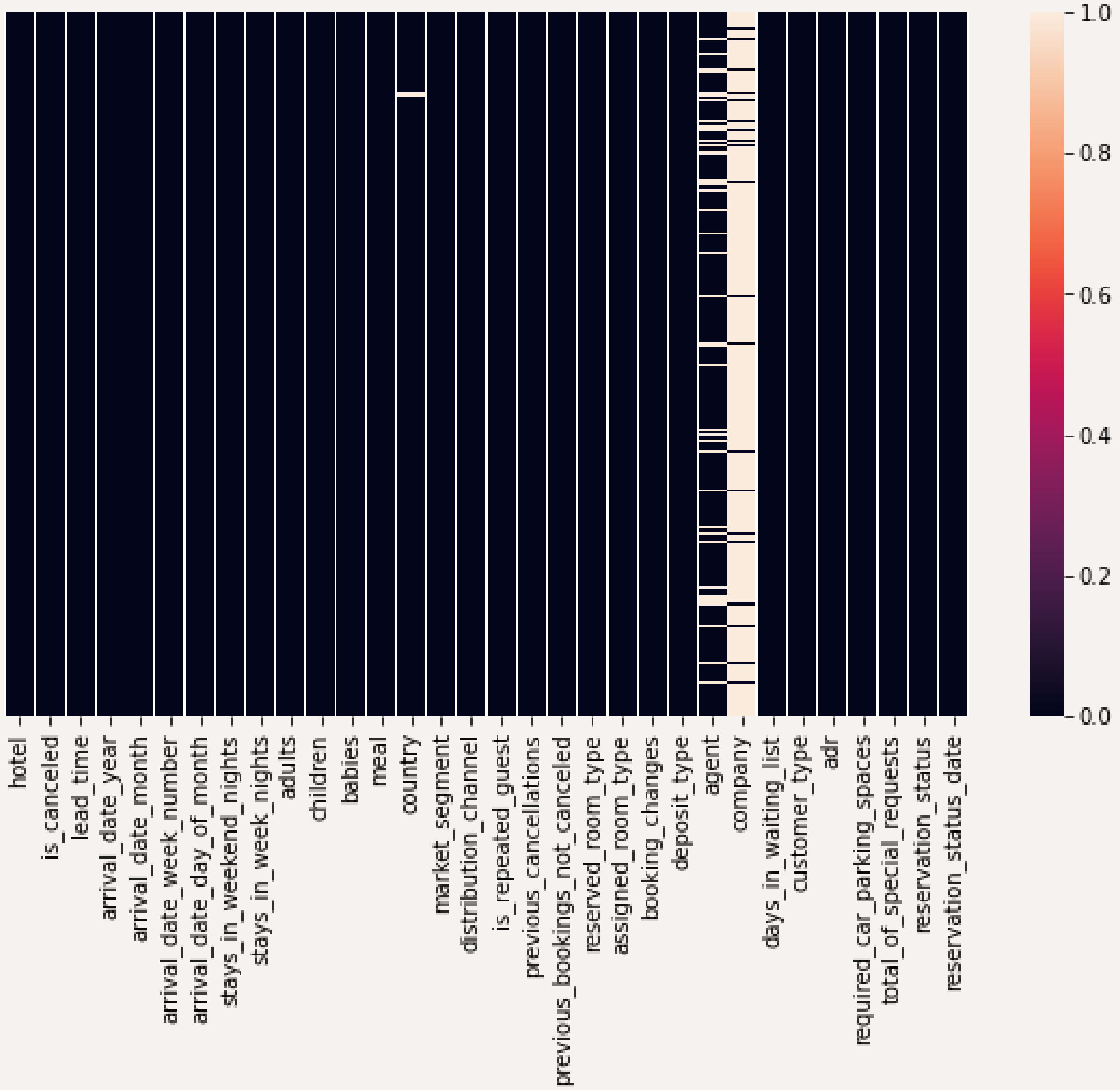


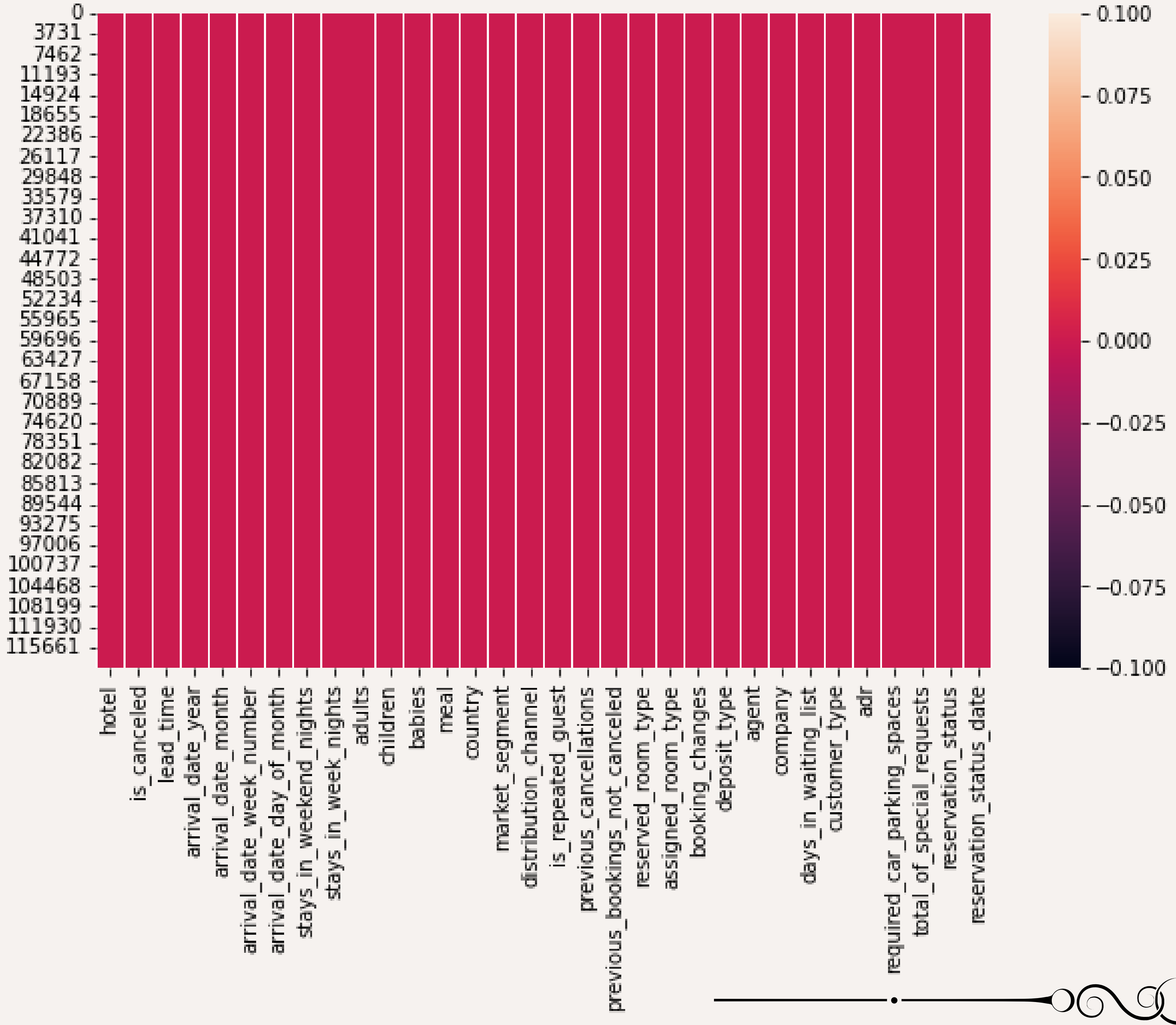


Visualizing pairplot



Visualizing null values before and after





Inference on the data

- Correlated all the features of the dataset in order to get the relationship

between the features.

- Then we have checked the null-values of all features and we had observed that the

feature named ‘company’ has high null values. So, we can drop that company while training the data from the dataset.

- For the remaining features which consists of null values, they will be replacing with

median/mode/mean of that remaining non–null data in that feature.

Steps done

1)imported required libraries

- Basic and most important libraries
- Model evaluation tools
- Data processing functions

2) Importing dataset

- Exploring the data shape
- Exploring the Datatypes

3)Correlation between all the Features

4)Describe Statistical Summary

5)Check for Null Values

- Replace Null values with median for int and float and mode for object type
- verfying the null values

)identifying and removing duplicate values

- Identify total number of duplicated values
- Dropping all duplicate values

Steps done

- Verifying duplicates are any

7)Encoding: Used label encoder

8)Evaluating a classification model

- Dividing data into Input X variables and Target Y variable.
- Y with only ‘is canceled’ feature and X with the remaining features.

9)Applied following Algorithms to find best model

- logistic regression
- KNN
- SVM(Linear kernel)
- Naive Bayes
- Dession Tree
- Bagging Classifier

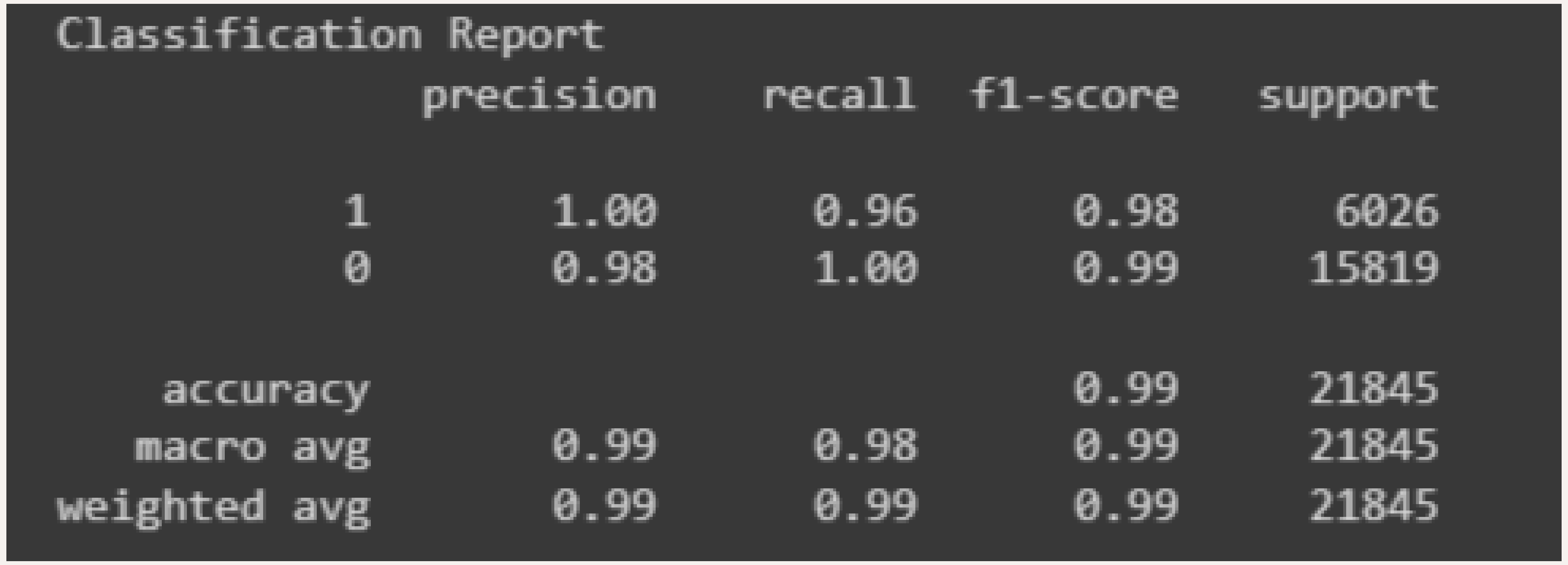
Steps done

10)Classification report and cofusion matrix for every model

Evaluating Models

Logistic Regression

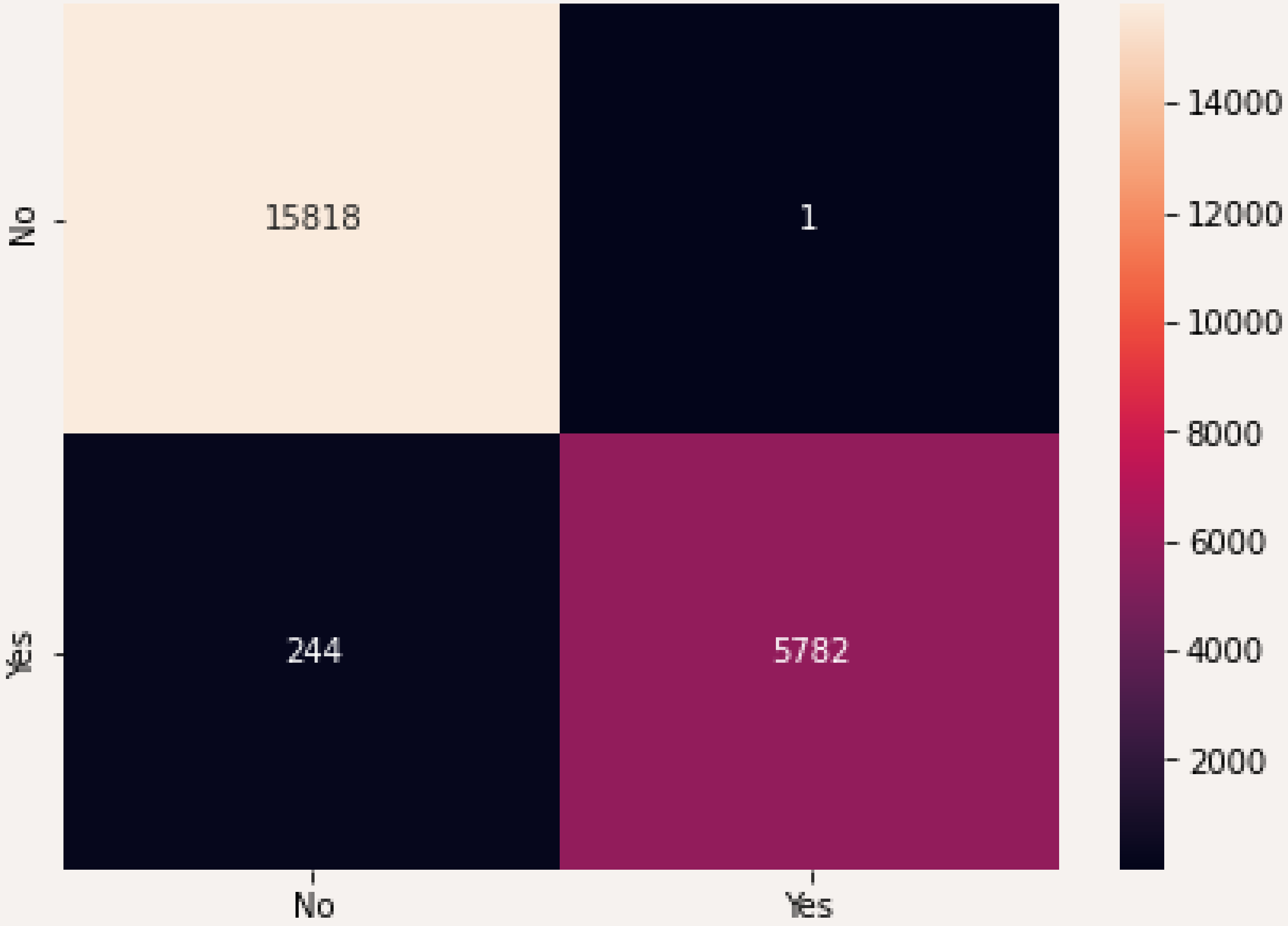
Classification Report after training:



```
Classification Report
              precision    recall  f1-score   support

           1       1.00      0.96      0.98      6026
           0       0.98      1.00      0.99     15819

    accuracy                           0.99     21845
   macro avg       0.99      0.98      0.99     21845
weighted avg       0.99      0.99      0.99     21845
```

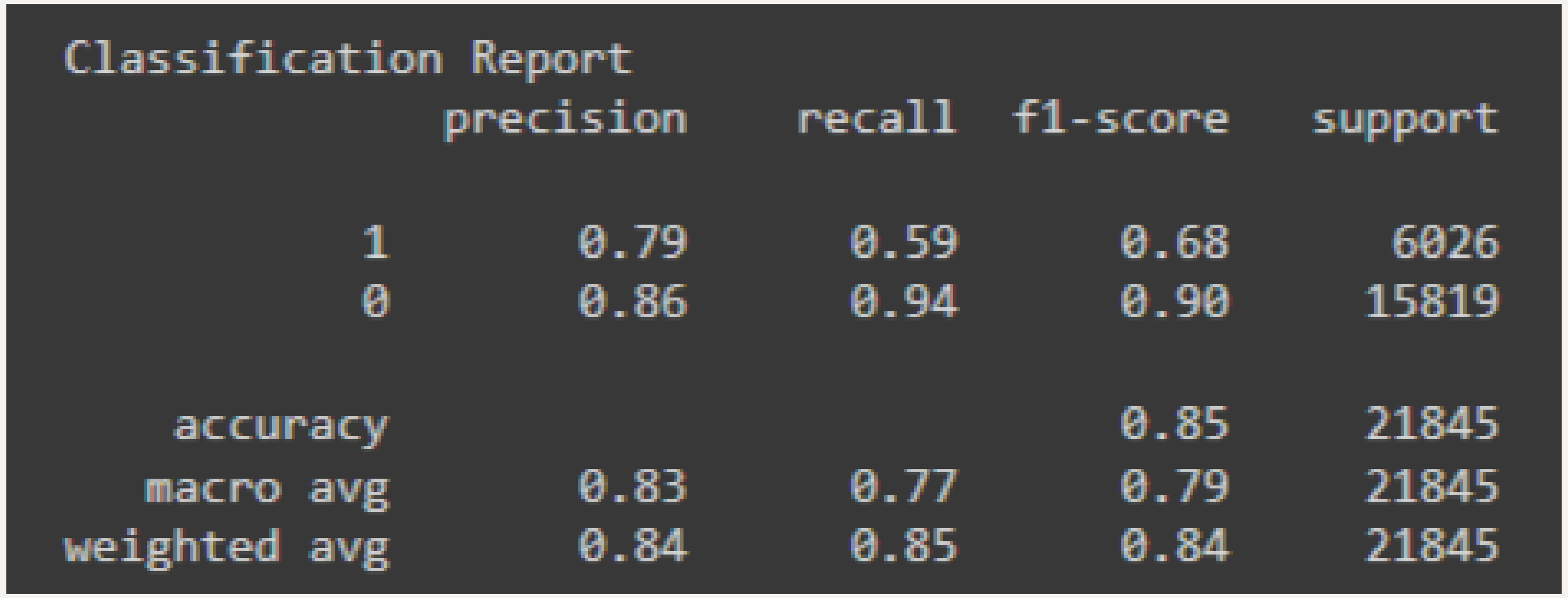


Confusion matrix
for
Logistic Regression

Evaluating Models

KNN

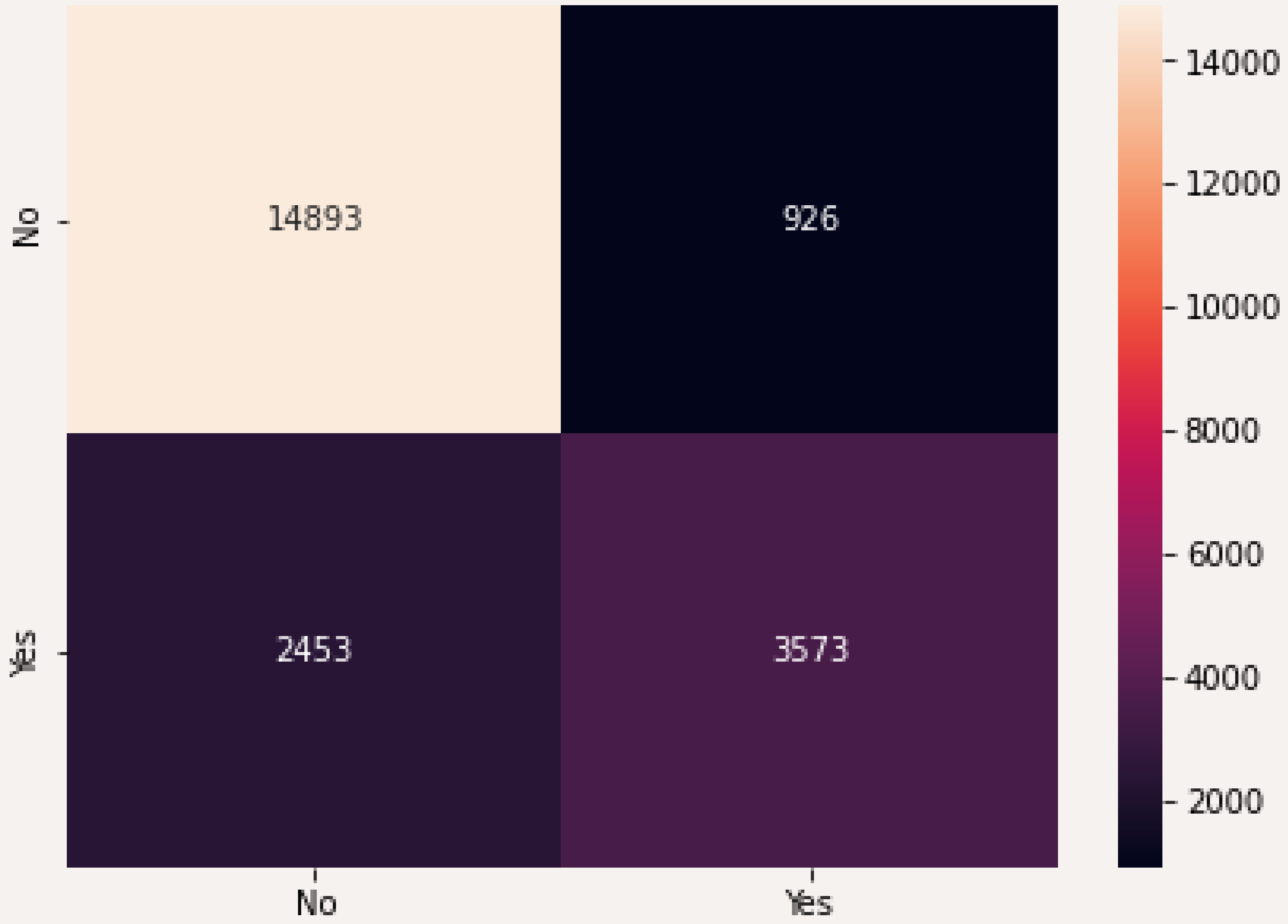
Classification Report after training:



```
Classification Report
              precision    recall  f1-score   support

           1       0.79      0.59      0.68      6026
           0       0.86      0.94      0.90     15819

    accuracy                           0.85     21845
   macro avg       0.83      0.77      0.79     21845
weighted avg       0.84      0.85      0.84     21845
```

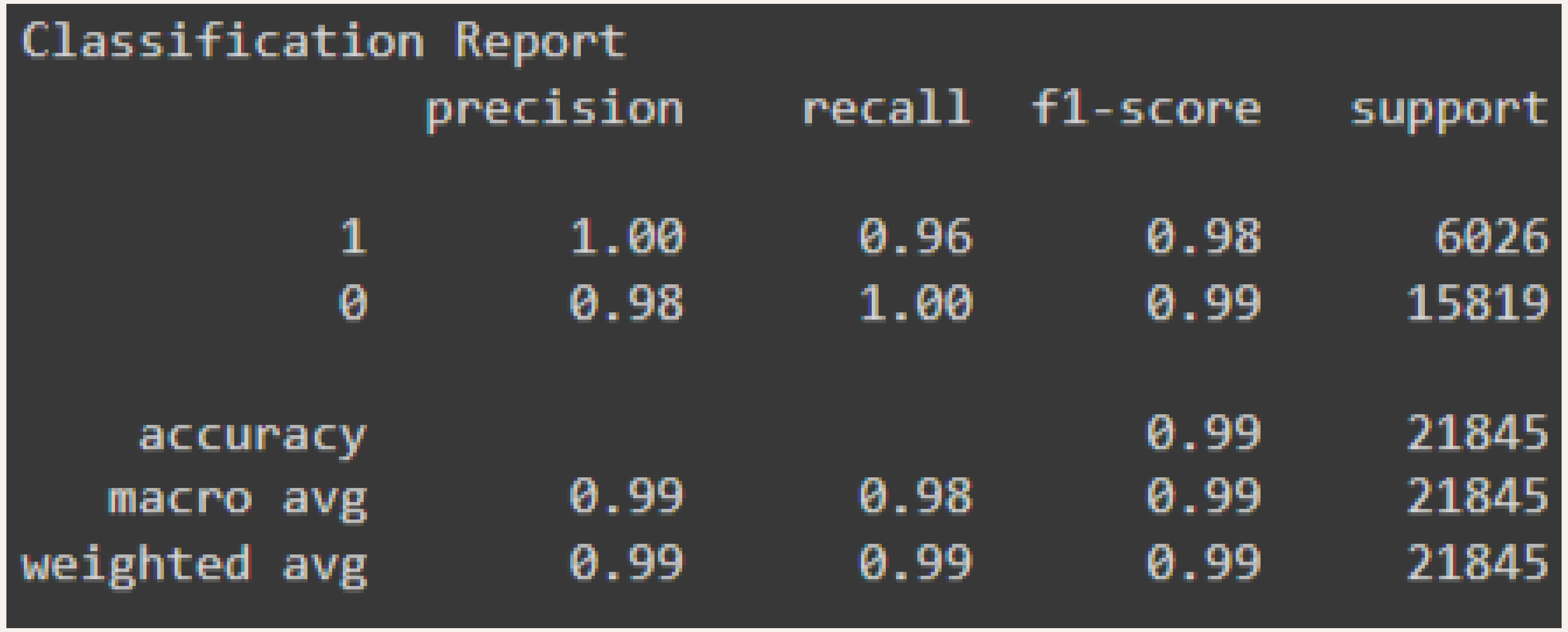


Confusion matrix
for
KNN

Evaluating Models

SVM-Linear Kernel

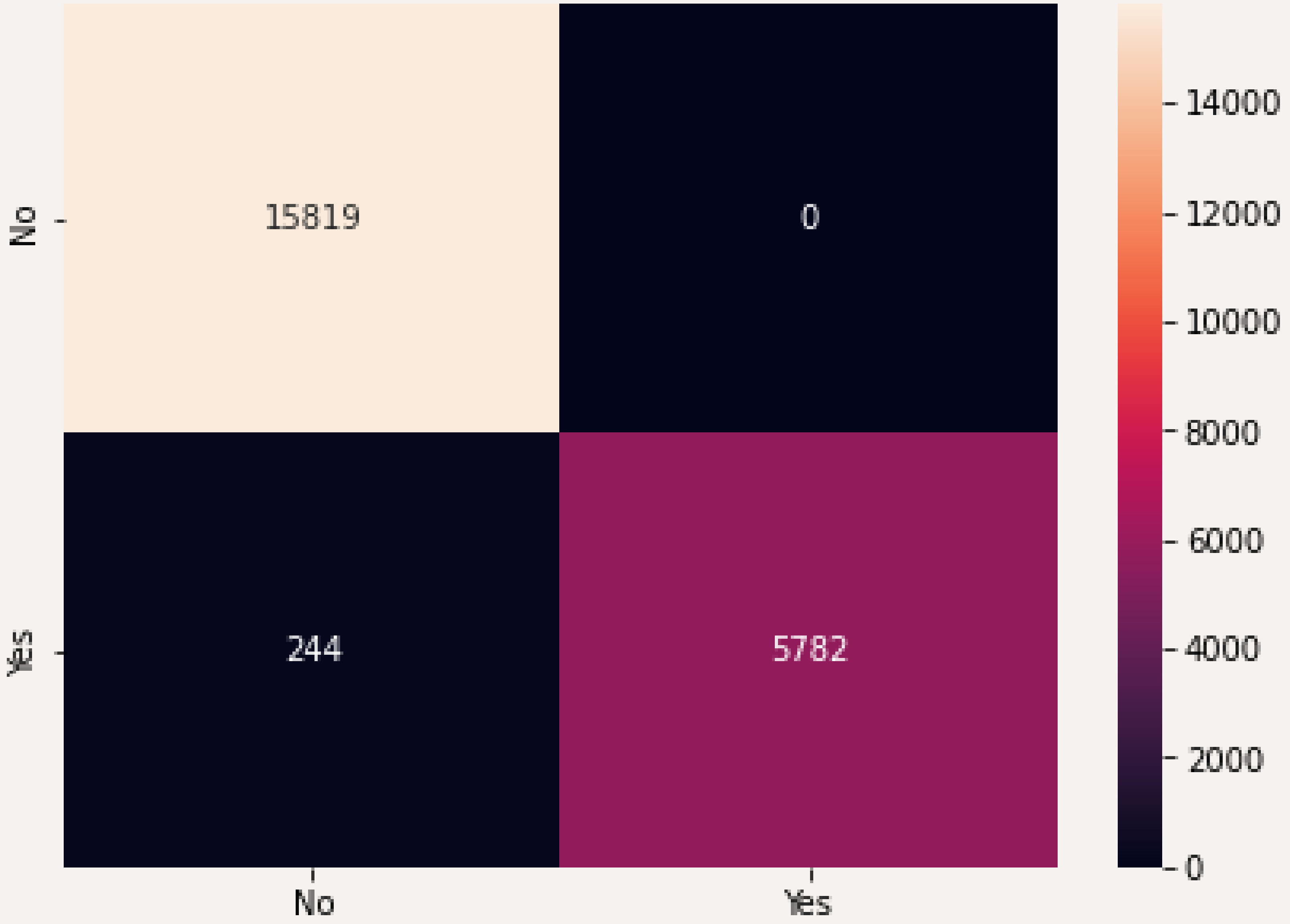
Classification Report after training:



```
Classification Report
              precision    recall  f1-score   support

           1       1.00      0.96      0.98      6026
           0       0.98      1.00      0.99     15819

    accuracy                           0.99     21845
   macro avg       0.99      0.98      0.99     21845
weighted avg       0.99      0.99      0.99     21845
```

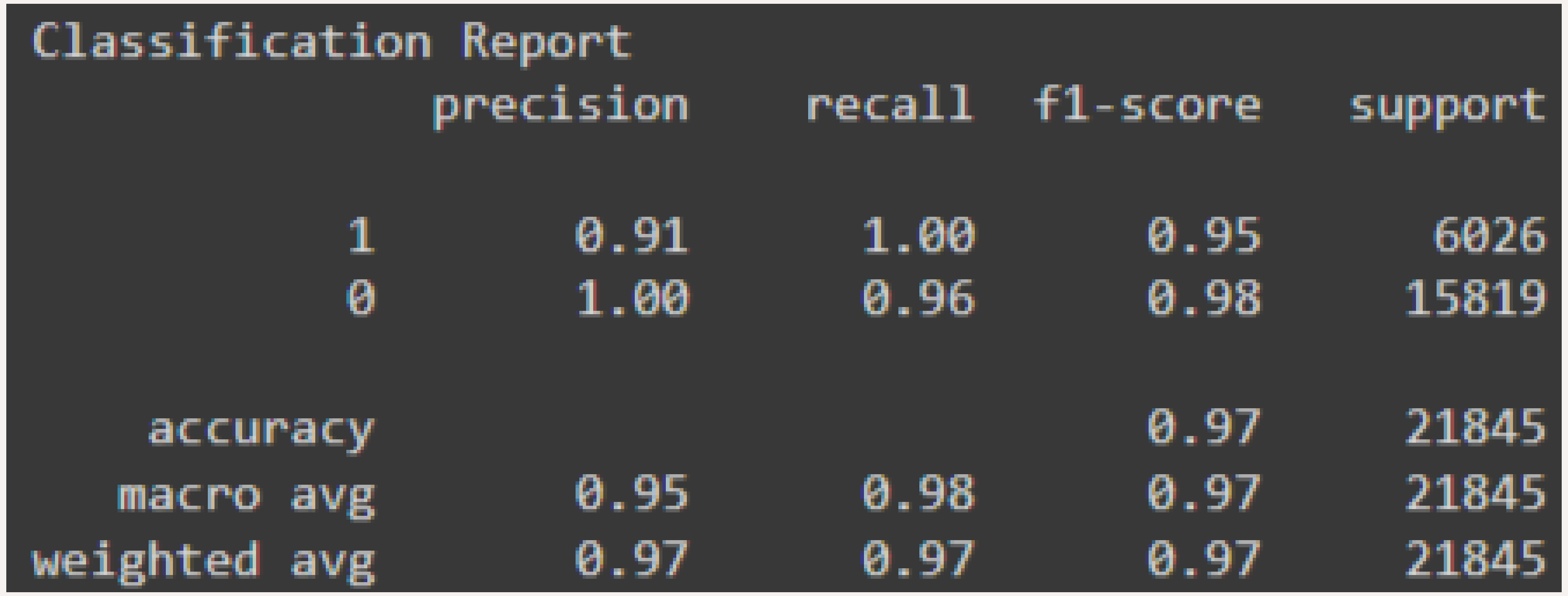


Confusion matrix
for
SVM

Evaluating Models

Naive Bayes

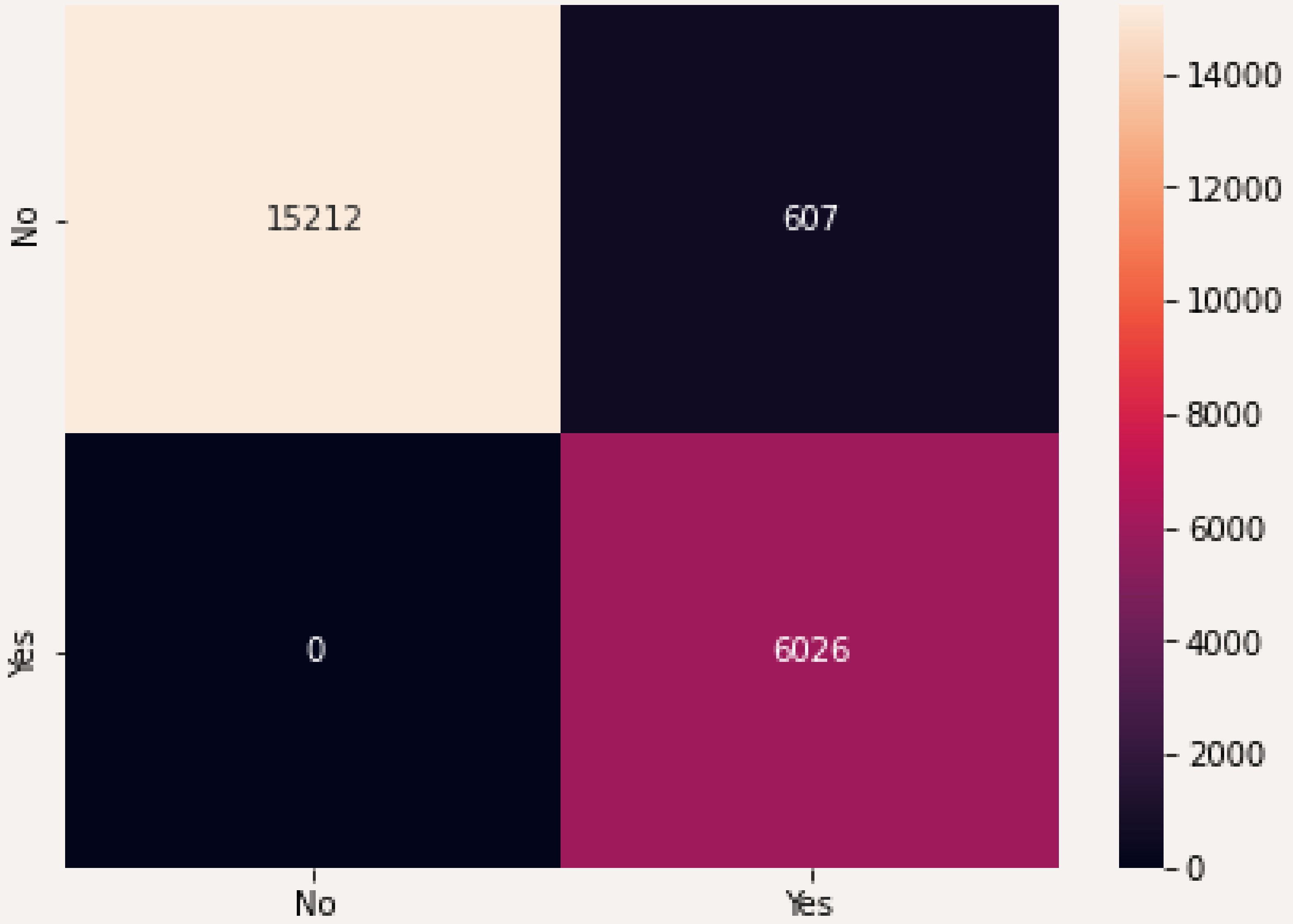
Classification Report after training:



```
Classification Report
              precision    recall  f1-score   support

           1       0.91      1.00      0.95      6026
           0       1.00      0.96      0.98     15819

    accuracy                           0.97     21845
   macro avg       0.95      0.98      0.97     21845
weighted avg       0.97      0.97      0.97     21845
```

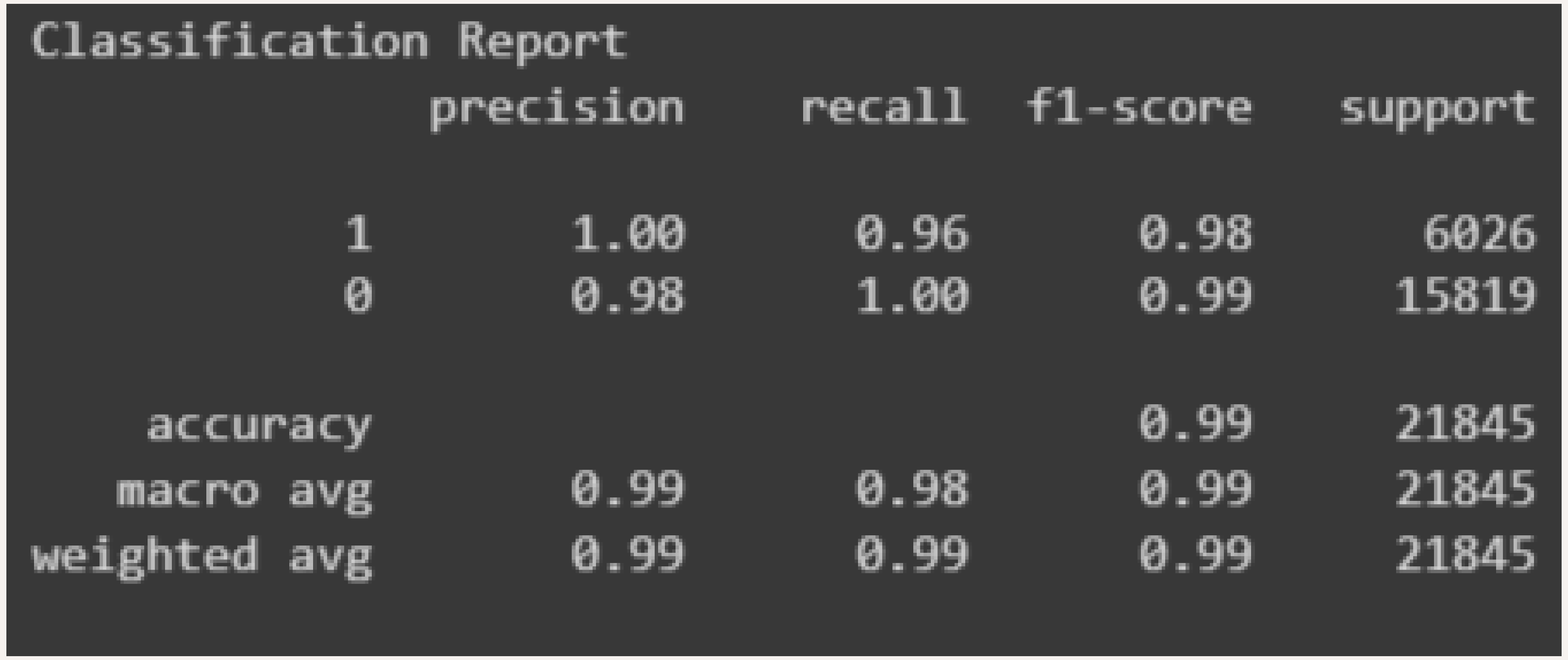


Confusion matrix
for
Naive Bayes

Evaluating Models

Decision Tree with

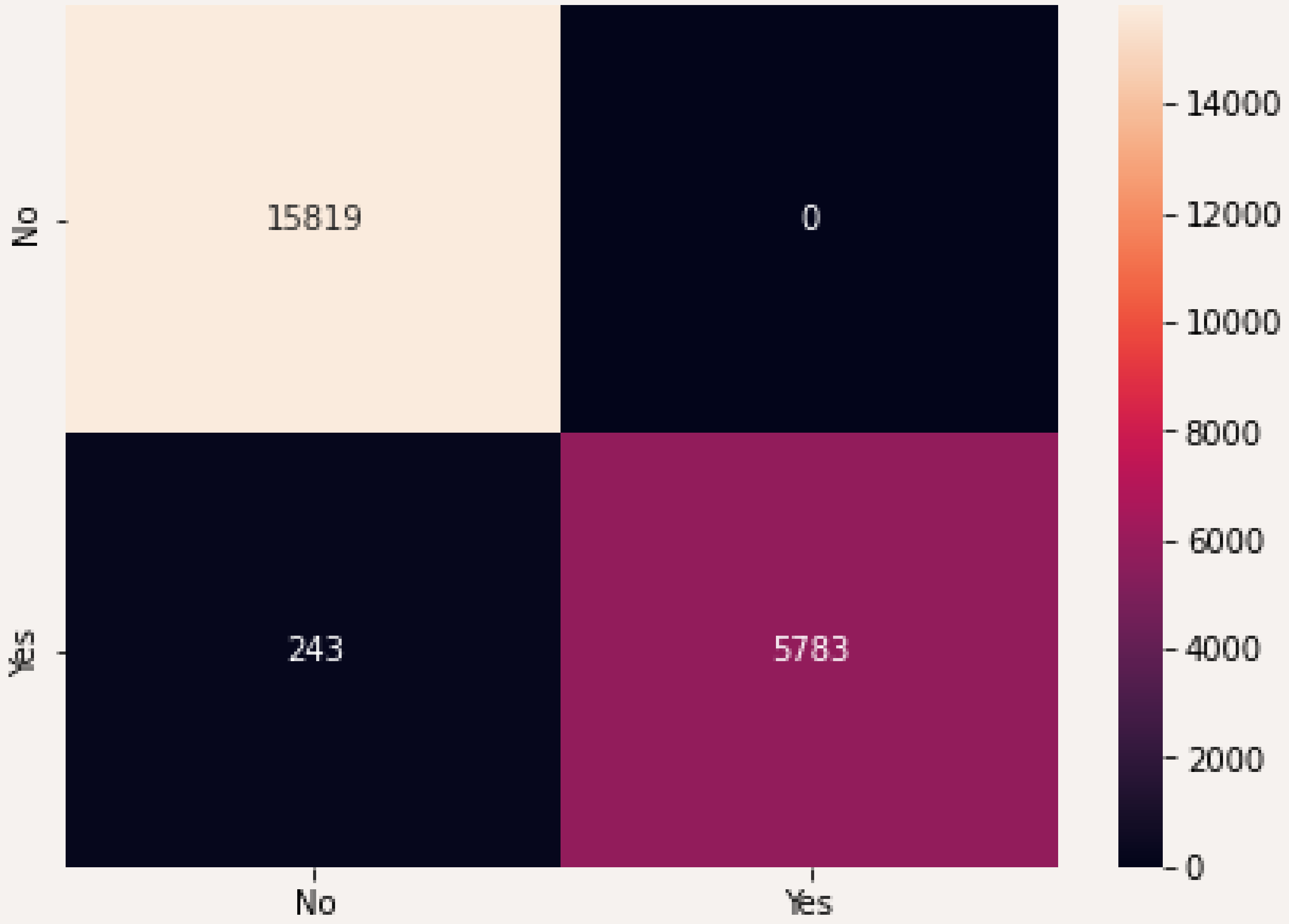
criterion = 'gini', max_depth = 1
Classification Report after training:



```
Classification Report
              precision    recall  f1-score   support

           1       1.00      0.96      0.98      6026
           0       0.98      1.00      0.99     15819

    accuracy                           0.99     21845
   macro avg       0.99      0.98      0.99     21845
weighted avg       0.99      0.99      0.99     21845
```

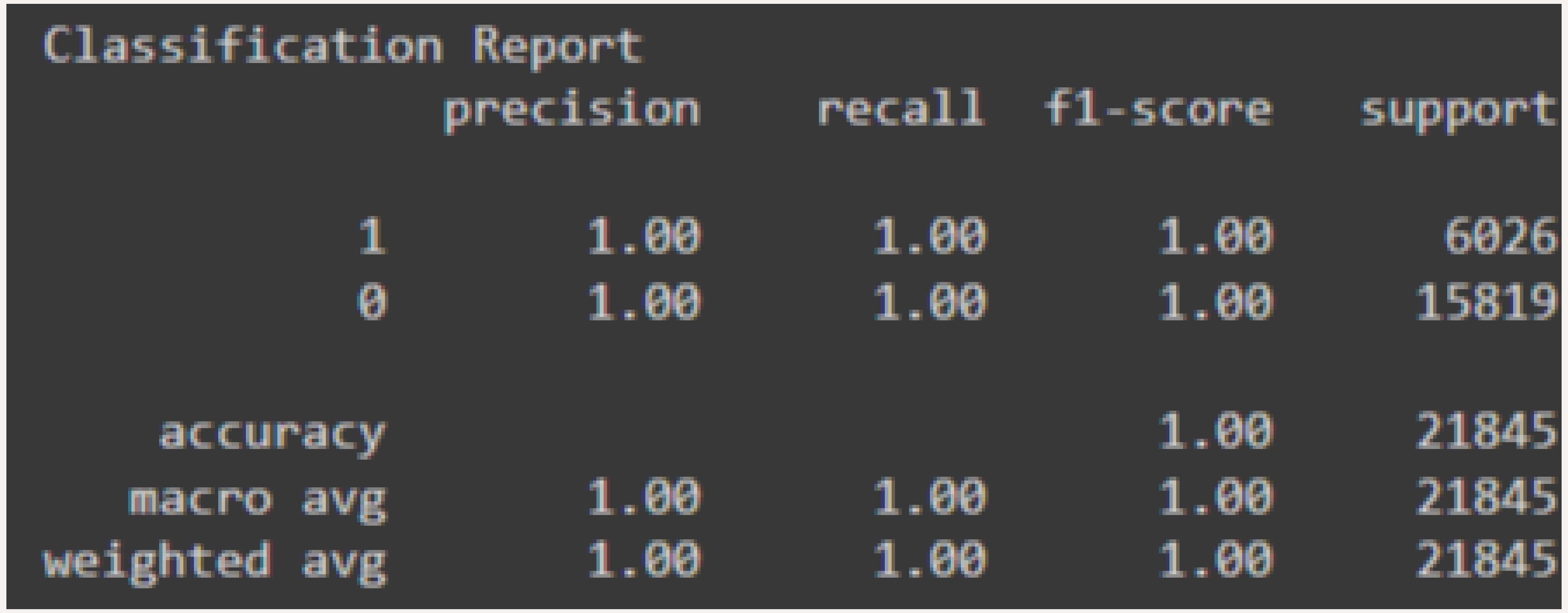


Confusion matrix
for
Decision Tree

Choosen Model

Bagging Classifier

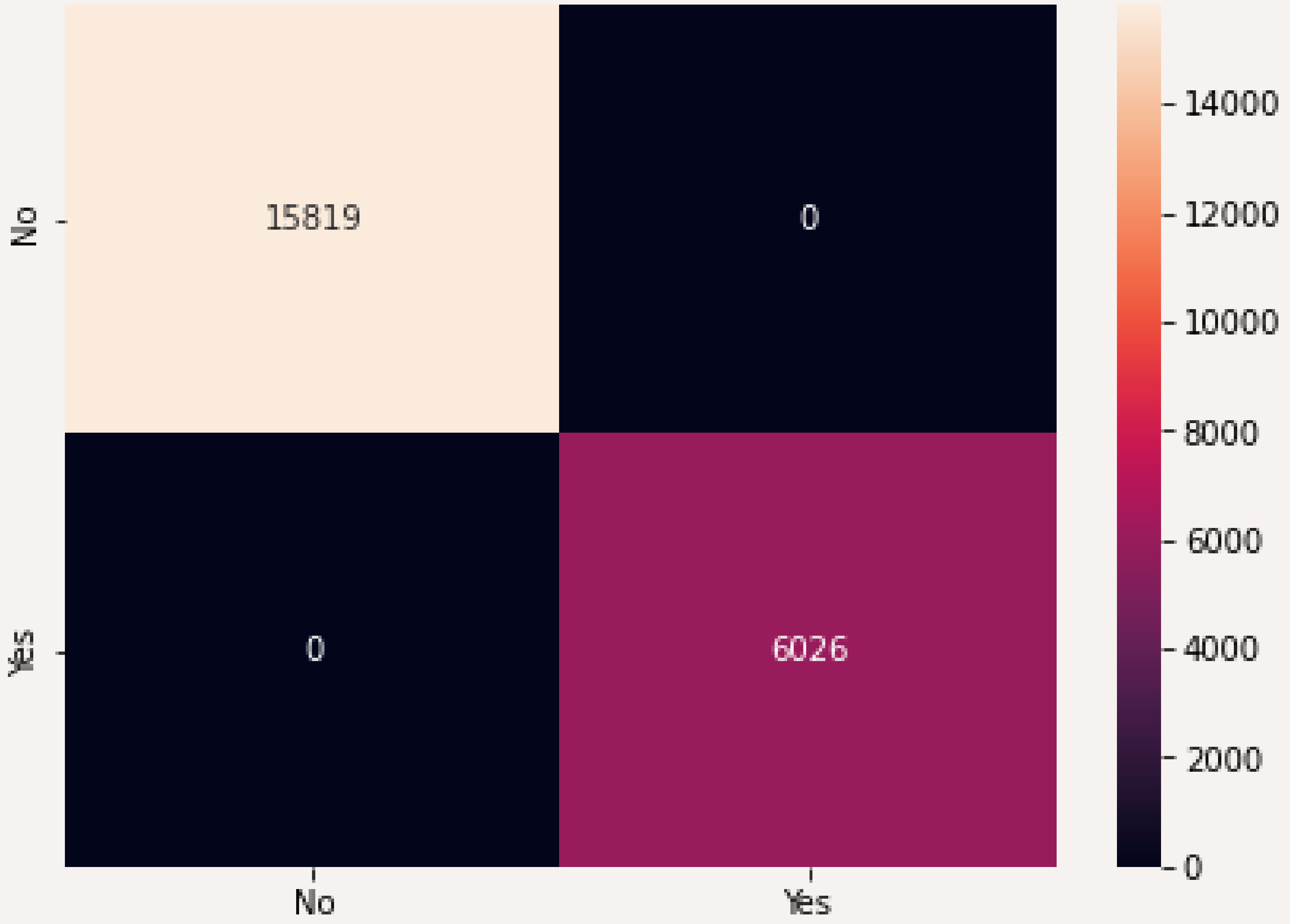
Classification Report after training:



```
Classification Report
             precision    recall  f1-score   support

           1       1.00      1.00      1.00      6026
           0       1.00      1.00      1.00     15819

    accuracy                           1.00     21845
   macro avg       1.00      1.00      1.00     21845
weighted avg       1.00      1.00      1.00     21845
```



Confusion matrix
for
Bagging Classifier

Conclusion

- In Bagging, each individual trees are independent of each other because they consider different subset of features and samples to predict a model.
- In this model n_estimators as 150 with default base_estimator and random_state=0 are used.
- From, all the evaluation with accuracy score and confusion matrix we conclude with bagging classifier we got the best model which gives 100% accuracy.

Thank You