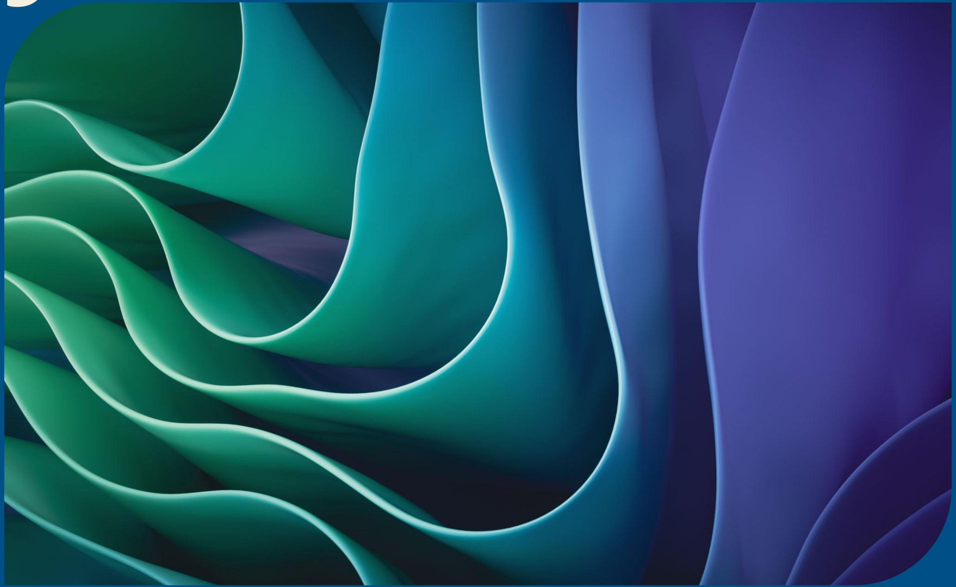
NYPD Traffic Stop Data Analysis

Samuel Preston & Dean Filippone



Traffic Stop Information

- From NYC open Data for NYPD Vehicle Stop Reports
- New York City is the most populous city in the US
- Traffic Stop statistics can be used by law enforcement to make better informed decisions
- The number of traffic stops has increased over the years

Key Question

Can we use circumstantial data of traffic stops with machine learning to predict what stops will / should result in an arrest?

Exploratory Data Analysis

The Data

Arrest Rate

Key metric we are trying to predict

Demographics

Age, Gender, Race

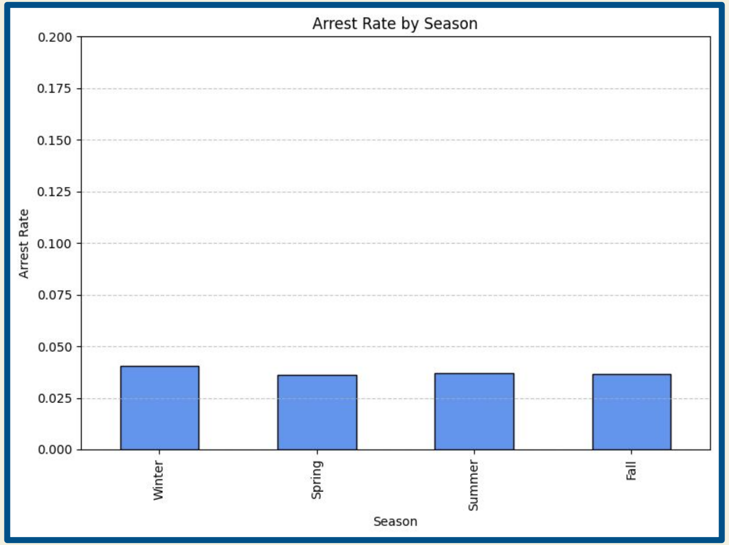
Resulting Data

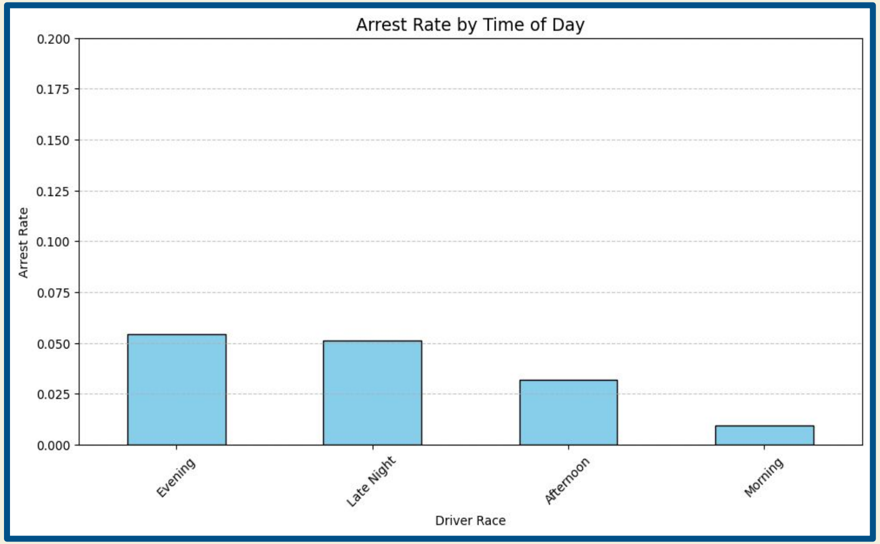
Summons_issued, vehicle_searched, vehicle_seized, force_used

Circumstantial

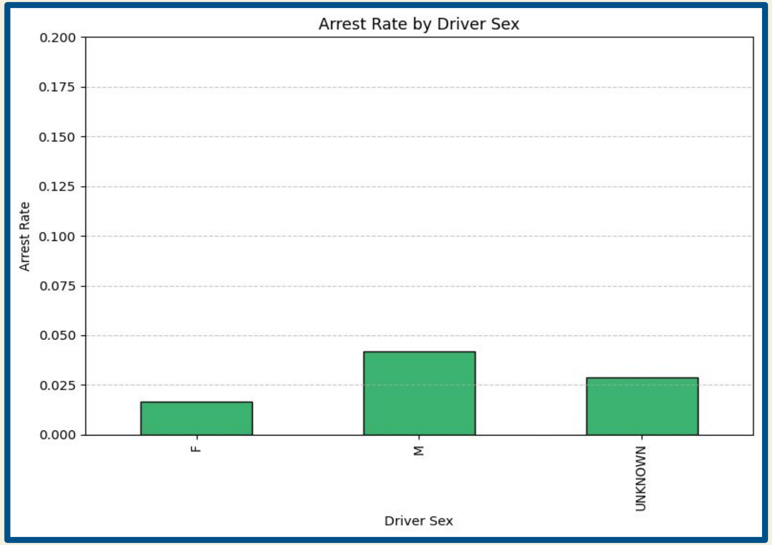
Time of day, how, month, season, checkpoint_stop, vehicle_type, location

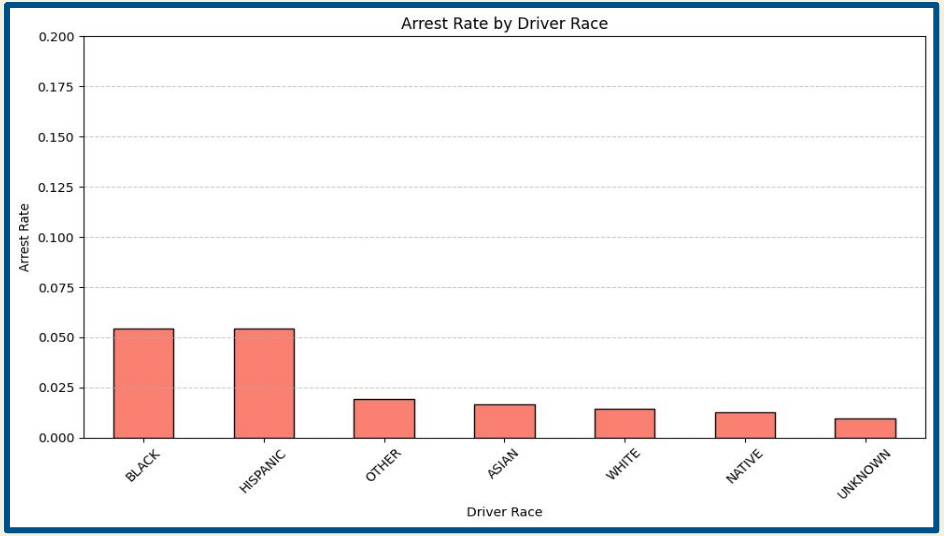
EDA


Arrest Rate by Season
Arrest Rate
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Winter
Spring
Summer
Fall
Season

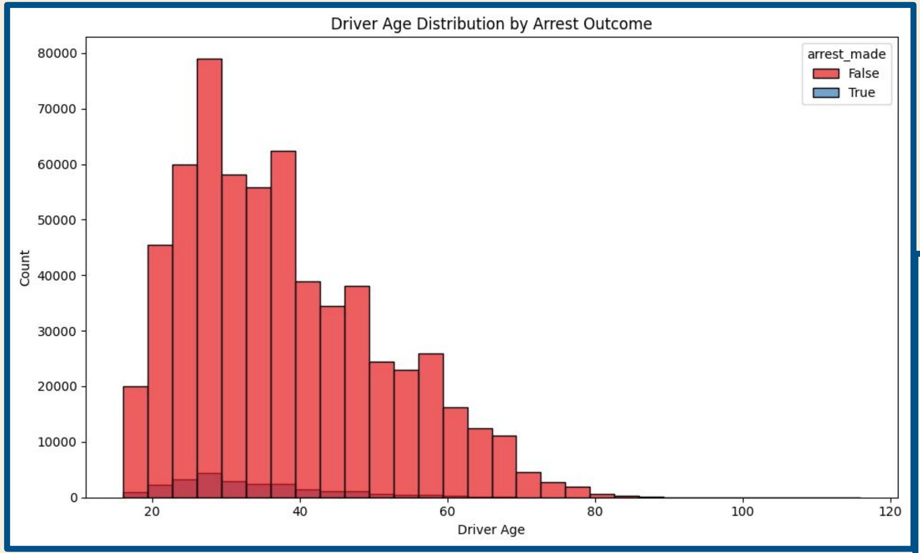

Arrest Rate by Time of Day
Arrest Rate
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Evening
Late Night
Afternoon
Morning
Driver Race

EDA

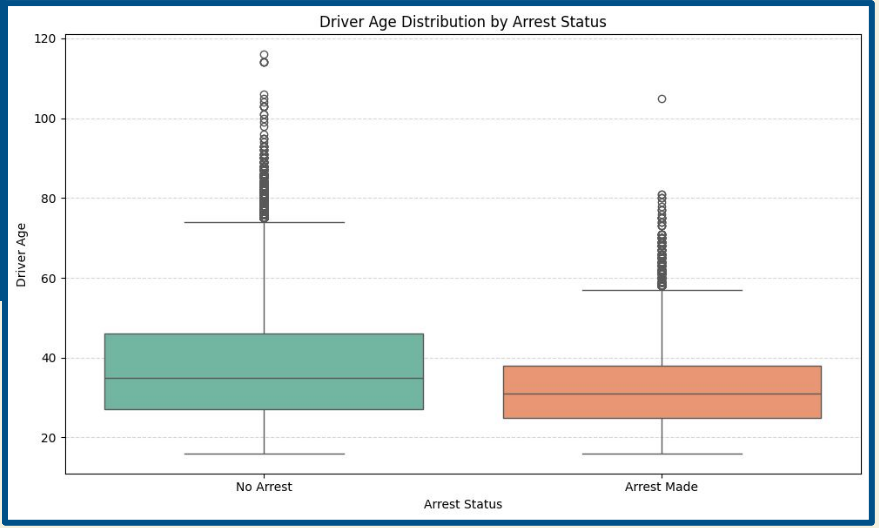

Arrest Rate by Driver Sex
Arrest Rate
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
F
M
UNKNOWN
Driver Sex


Arrest Rate by Driver Race
Arrest Rate
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
BLACK
HISPANIC
OTHER
ASIAN
WHITE
NATIVE
UNKNOWN
Driver Race

EDA

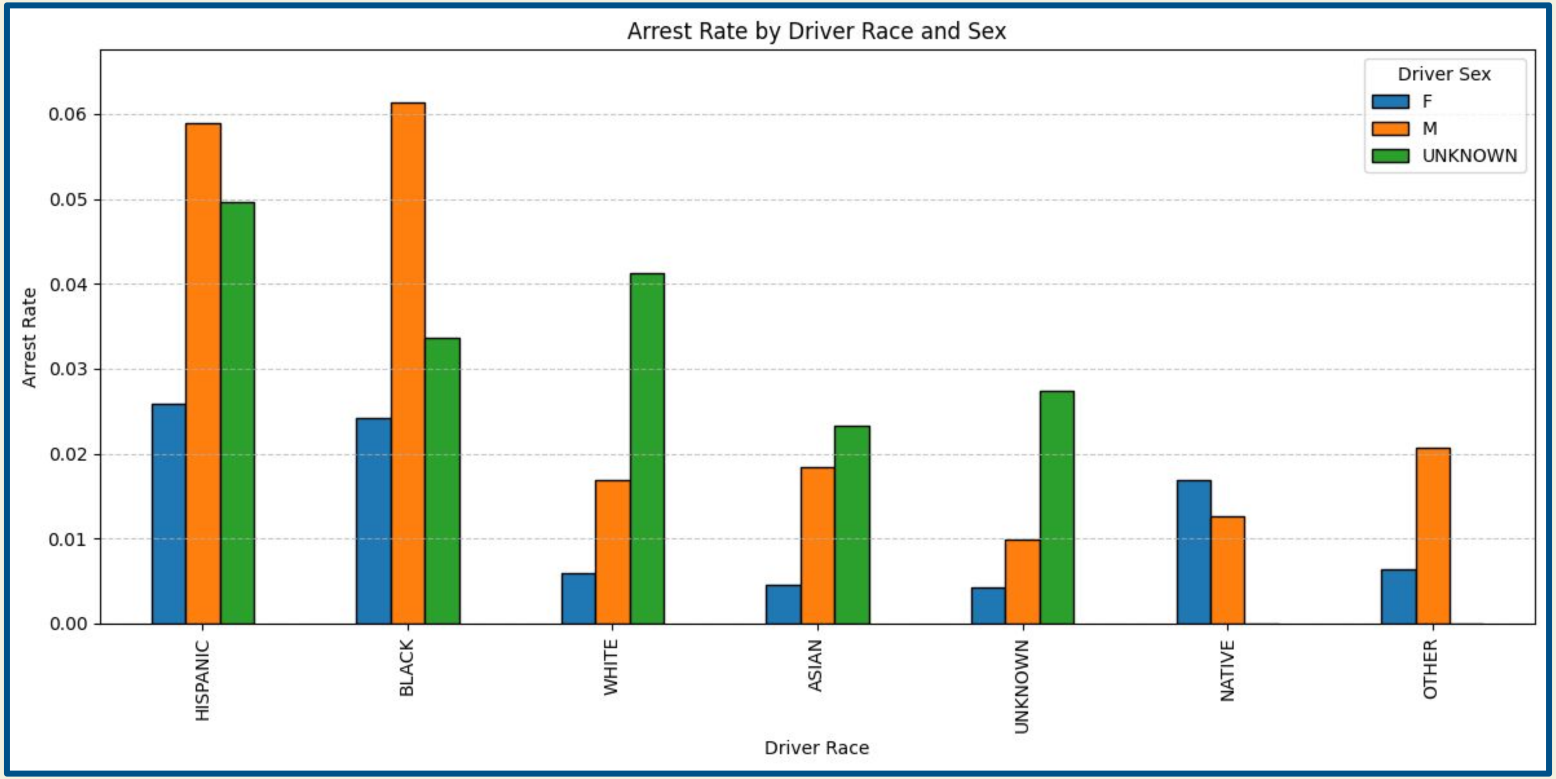


Driver Age Distribution by Arrest Outcome
Count
0
10000
20000
30000
40000
50000
60000
70000
80000
20
40
60
80
100
120
Driver Age
arrest_made
🟥 False
🟦 True

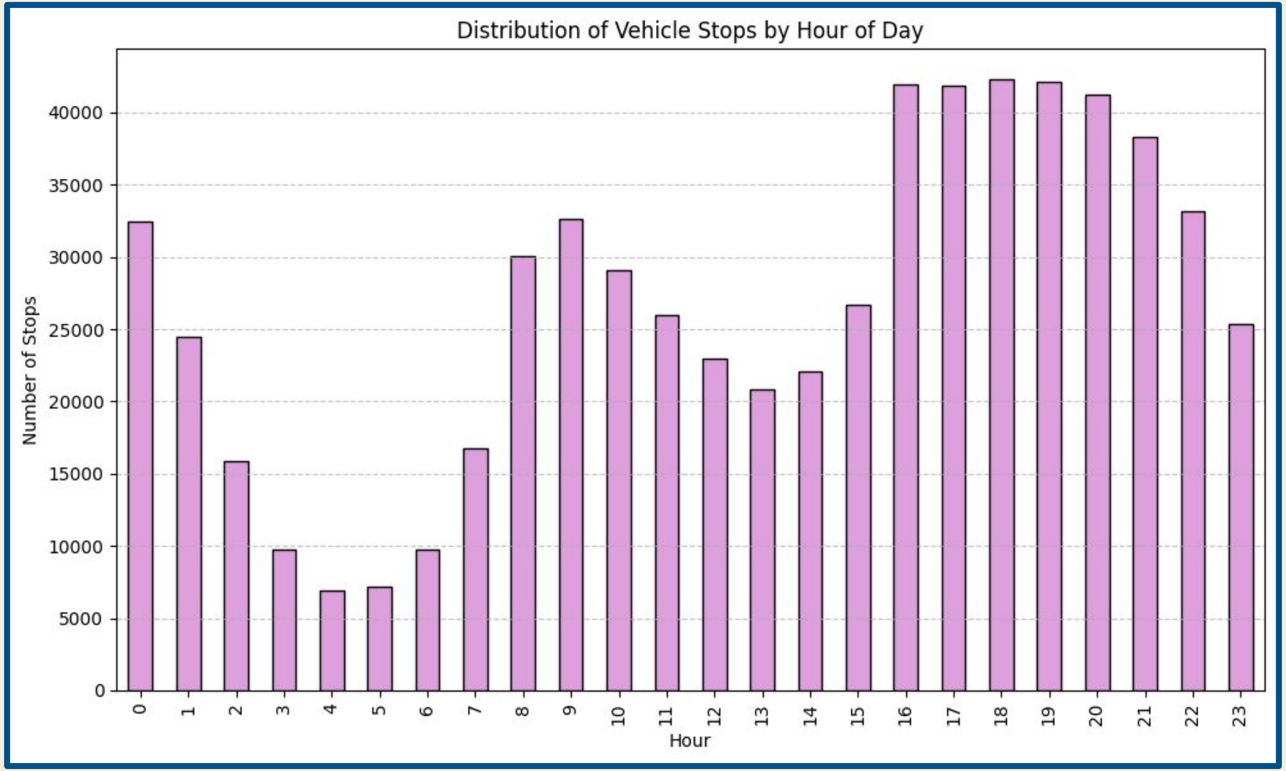


Driver Age Distribution by Arrest Status
Driver Age
20
40
60
80
100
120
No Arrest
Arrest Made
Arrest Status

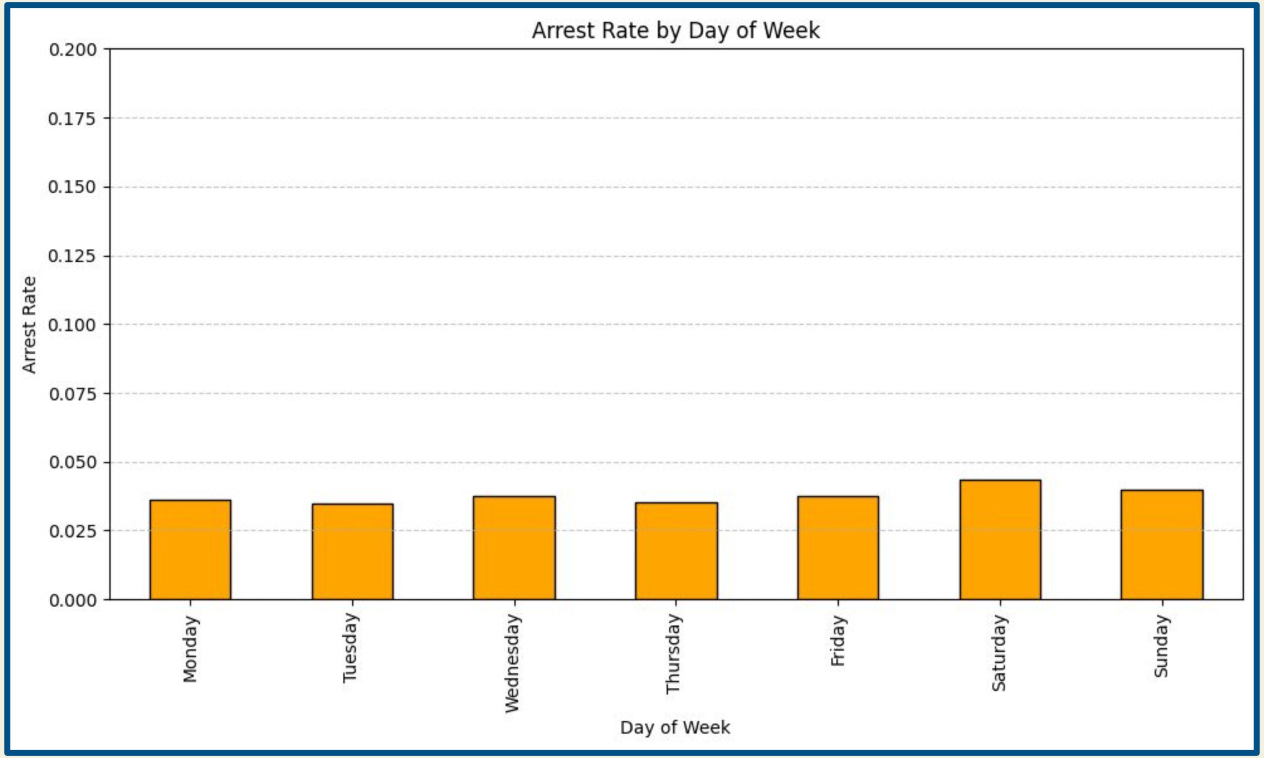
EDA


Arrest Rate by Driver Race and Sex
Arrest Rate
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Driver Sex
F
M
UNKNOWN
HISPANIC
BLACK
WHITE
ASIAN
UNKNOWN
NATIVE
OTHER
Driver Race

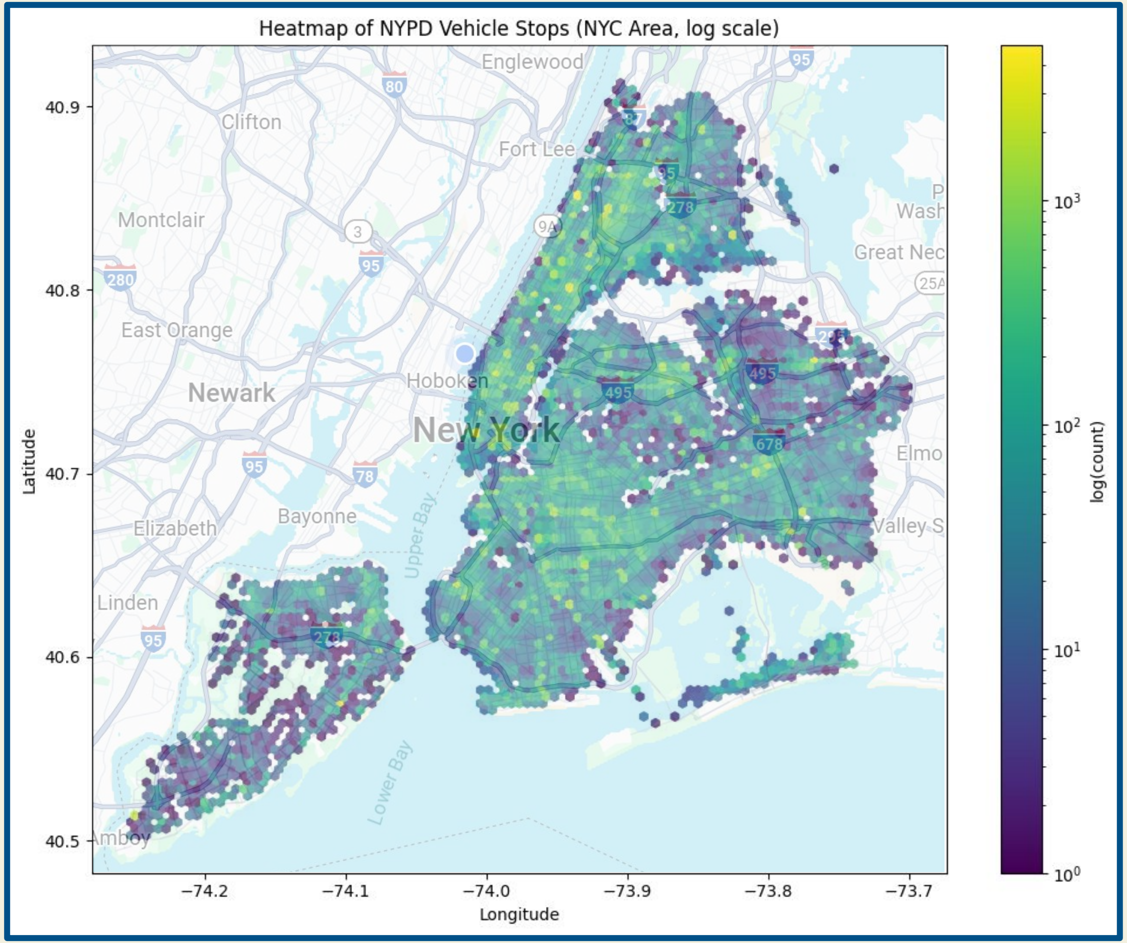
EDA


Distribution of Vehicle Stops by Hour of Day
Number of Stops
0
5000
10000
15000
20000
25000
30000
35000
40000
Hour
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

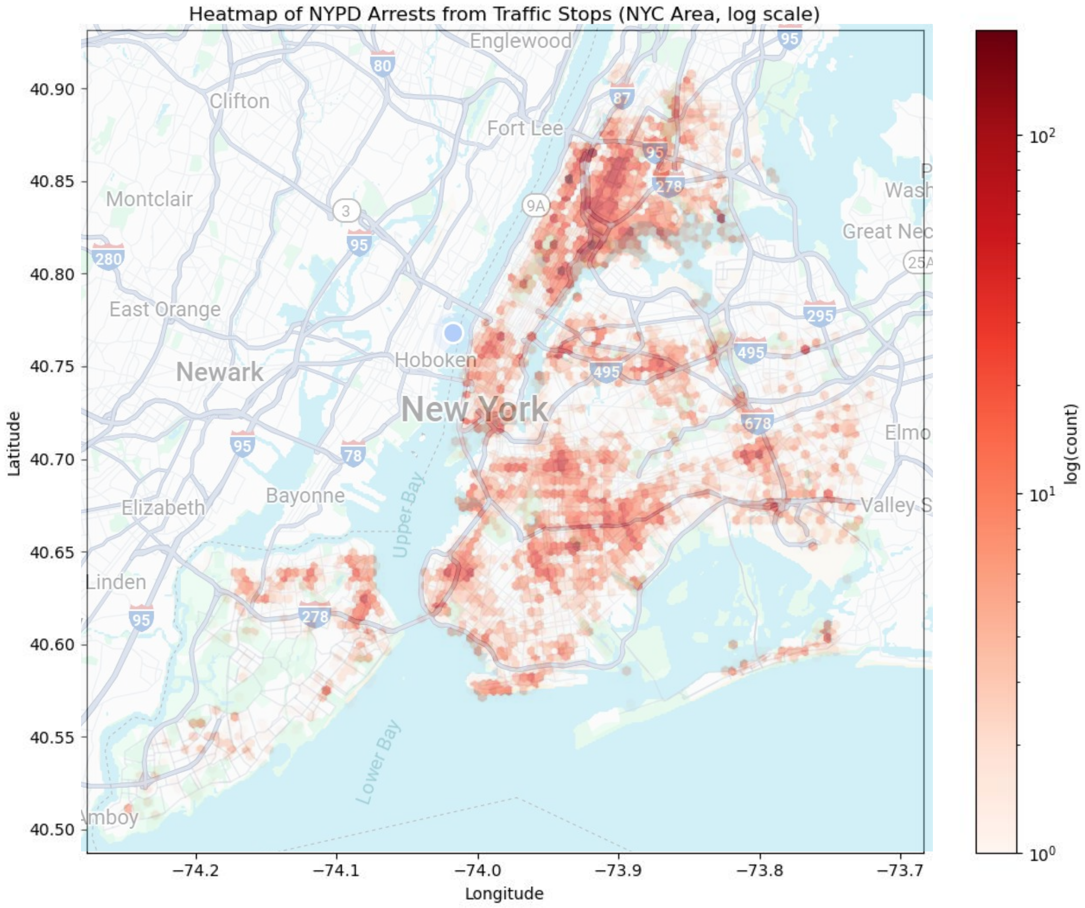
EDA


Arrest Rate by Day of Week
Arrest Rate
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of Week

EDA


Heatmap of NYPD Vehicle Stops (NYC Area, log scale)
Latitude
Longitude
40.9
40.8
40.7
40.6
40.5
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
log(count)
10^3
10^2
10^1
10^0
Englewood
Clifton
Fort Lee
Montclair
East Orange
Newark
Hoboken
New York
Elizabeth
Bayonne
Linden
Amboy
Great Nec
Elmo
Valley S
Upper Bay
Lower Bay

EDA


Heatmap of NYPD Arrests from Traffic Stops (NYC Area, log scale)
Latitude
40.90
40.85
40.80
40.75
40.70
40.65
40.60
40.55
40.50
Longitude
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
log(count)
10^2
10^1
10^0
Englewood
Clifton
Fort Lee
Montclair
East Orange
Newark
Hoboken
New York
Bayonne
Elizabeth
Linden
Upper Bay
Lower Bay
Great Nec
Elmo
Valley S
Wash
mboy
80
87
95
278
9A
3
280
295
495
678
78
25A

The Models

Some Metrics

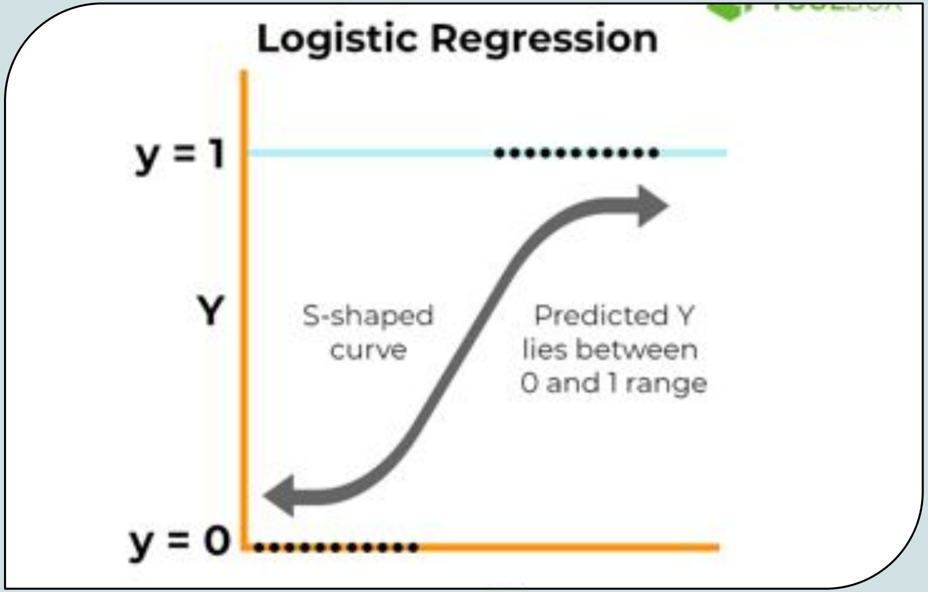
$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

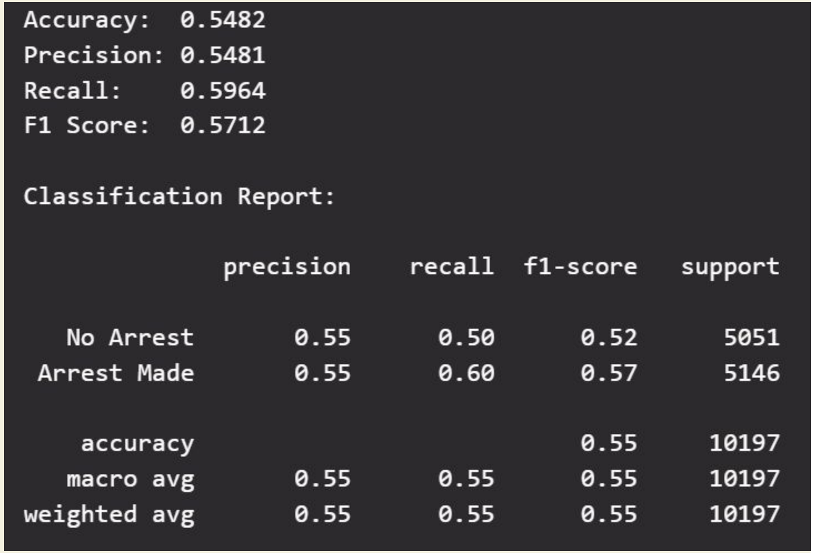
$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Logistic Regression

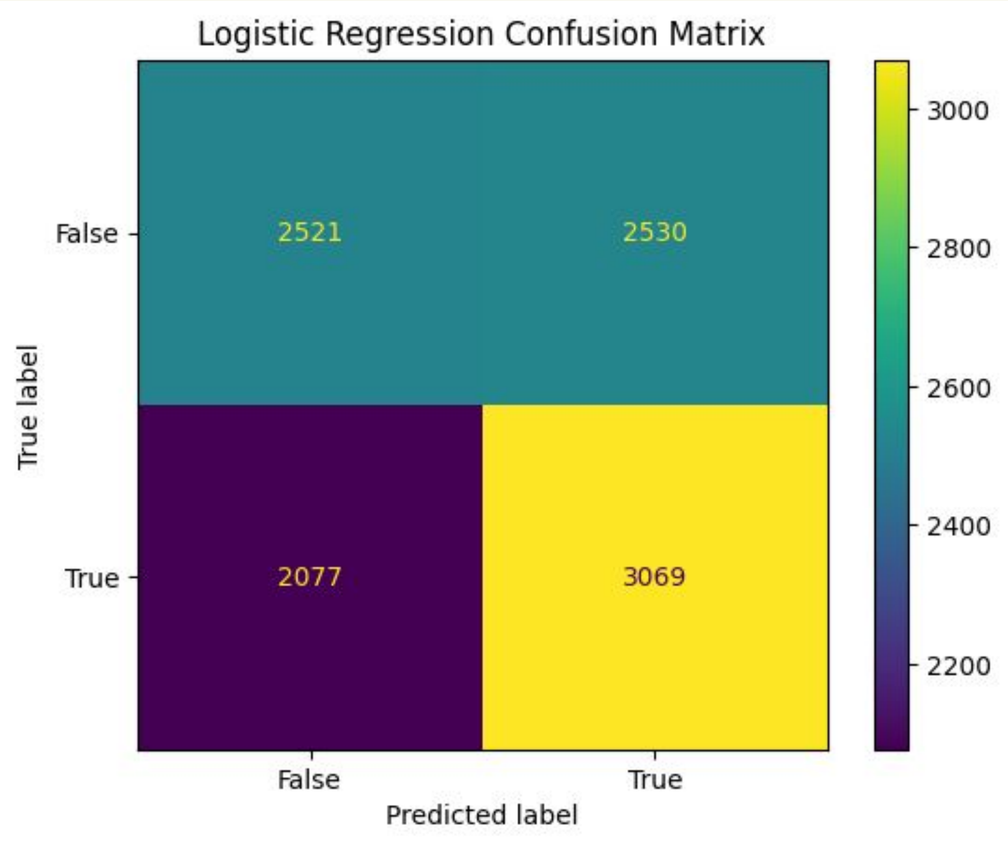
It models the probability of the occurrence of a binary event using a logistic function. Despite its name, it is used for classification, not regression



Logistic Regression

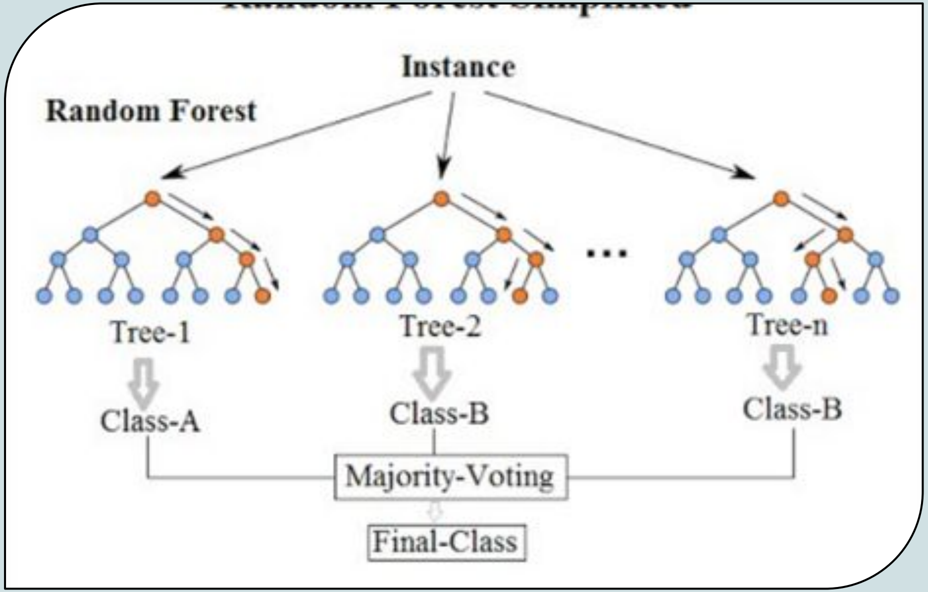


```
Accuracy:  0.5482
Precision: 0.5481
Recall:    0.5964
F1 Score:  0.5712

Classification Report:

              precision    recall  f1-score   support

   No Arrest       0.55      0.50      0.52      5051
 Arrest Made       0.55      0.60      0.57      5146

    accuracy                           0.55     10197
   macro avg       0.55      0.55      0.55     10197
weighted avg       0.55      0.55      0.55     10197
```

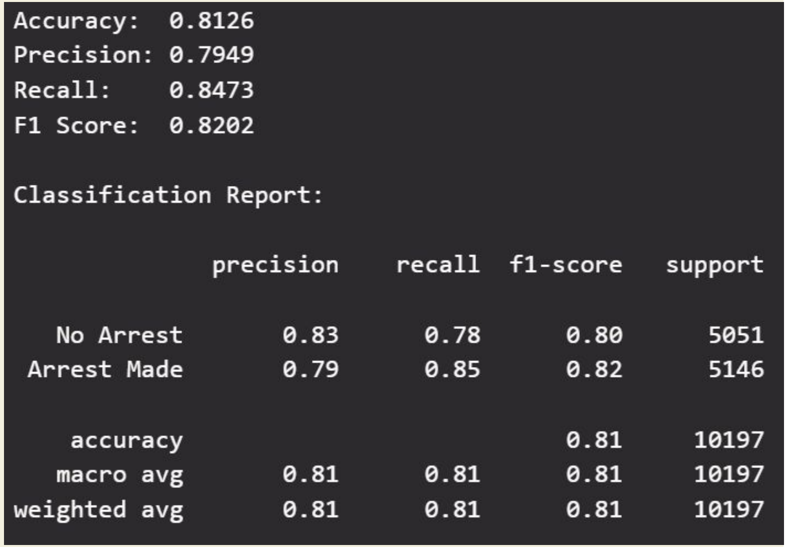


Random Forest Classifier

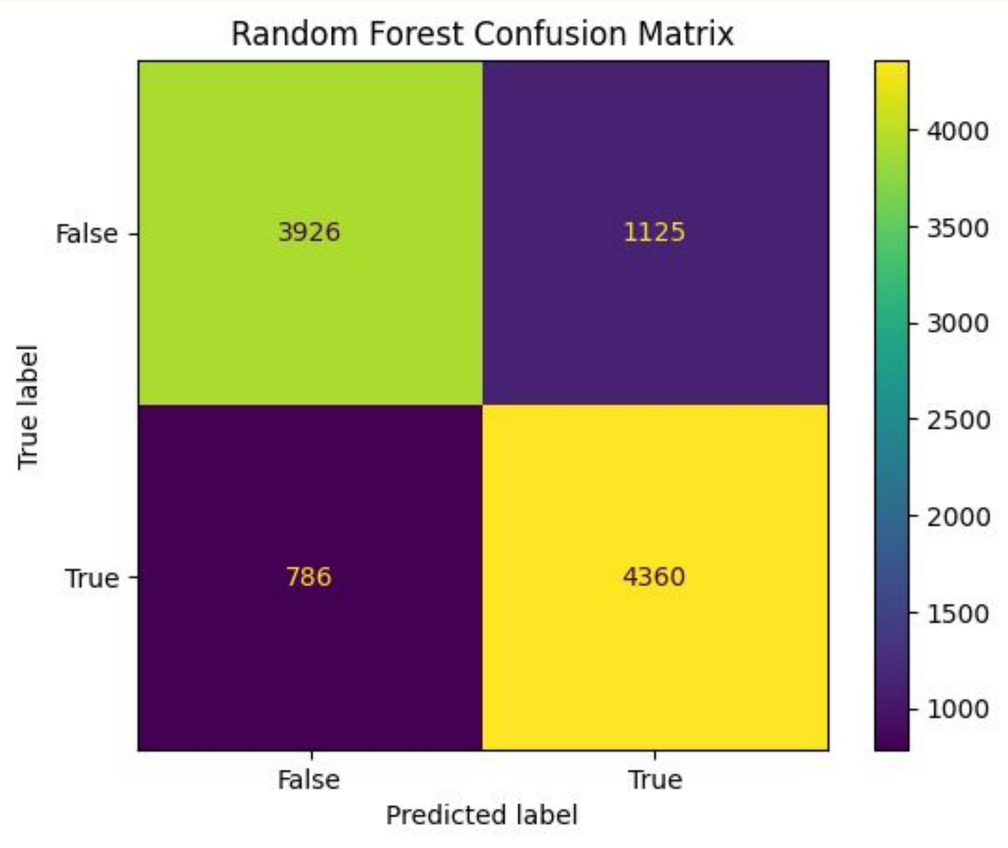
It builds multiple decision trees during training and merges their predictions. This ensemble approach improves generalization and reduces overfitting.



Random Forest Classifier

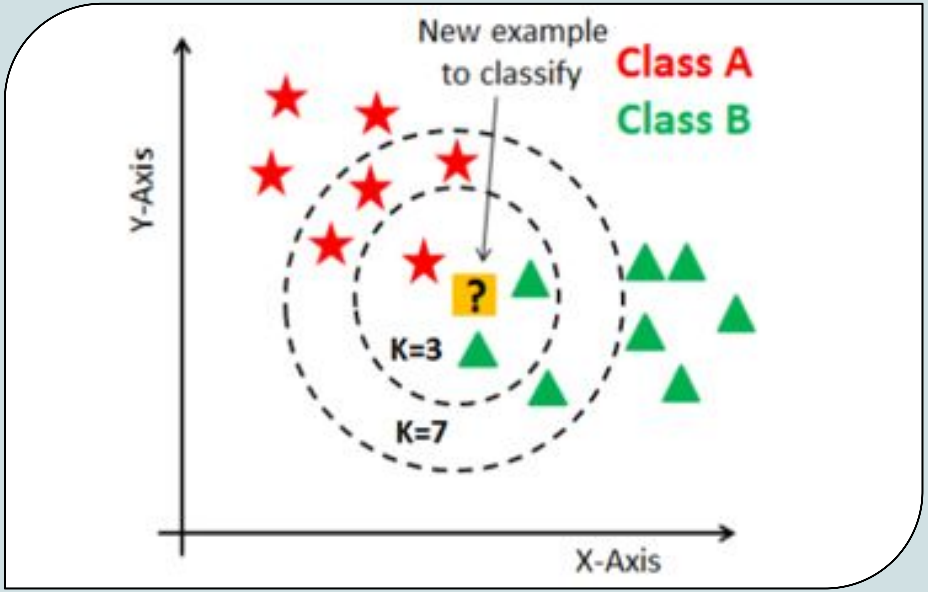


```
Accuracy:  0.8126
Precision: 0.7949
Recall:    0.8473
F1 Score:  0.8202

Classification Report:

              precision    recall  f1-score   support

   No Arrest       0.83      0.78      0.80      5051
 Arrest Made       0.79      0.85      0.82      5146

    accuracy                           0.81     10197
   macro avg       0.81      0.81      0.81     10197
weighted avg       0.81      0.81      0.81     10197
```

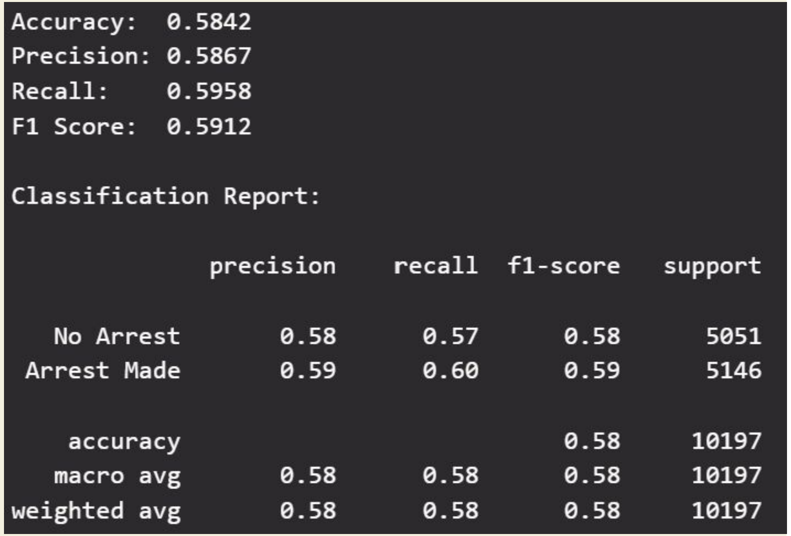


K-Nearest Neighbor

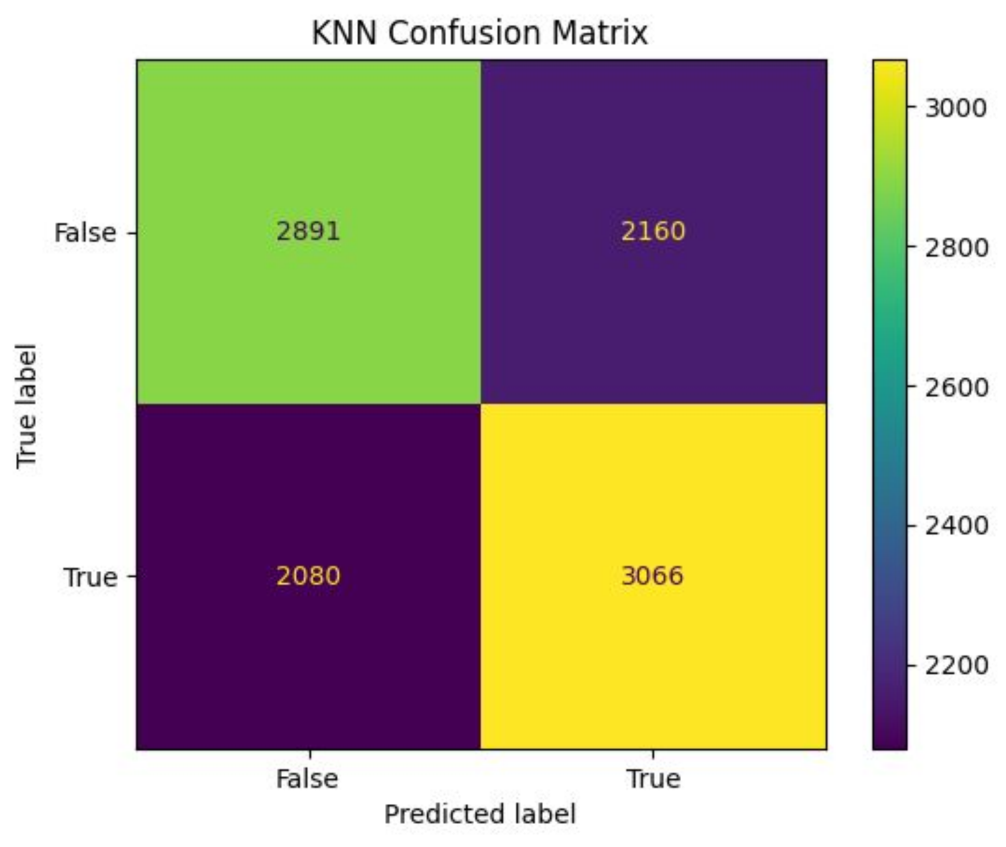
It classifies a data point based on the majority class of its k-nearest neighbors. The choice of k and the distance metric (e.g.Euclidean Distance) are important parameters



K-Nearest Neighbor

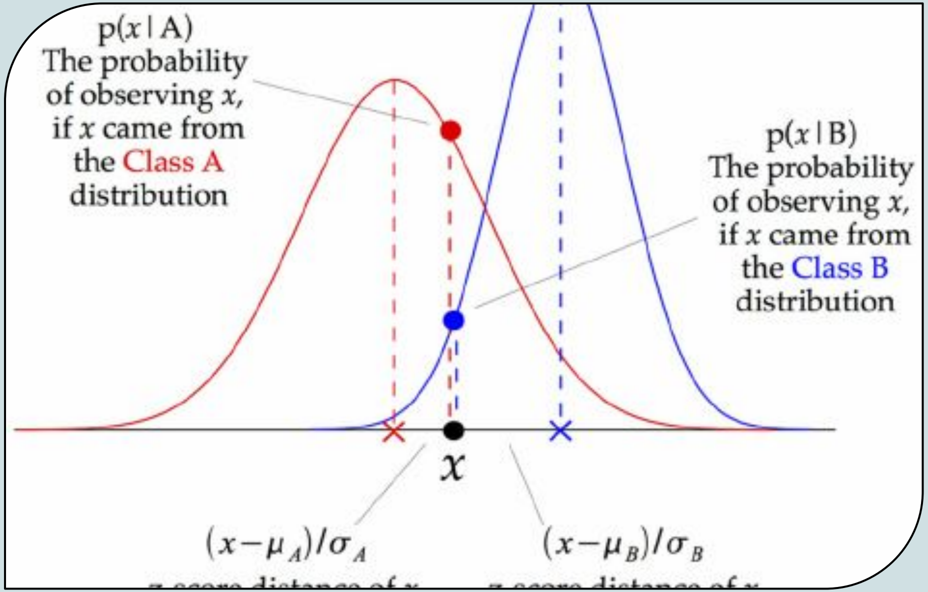


```
Accuracy:  0.5842
Precision: 0.5867
Recall:    0.5958
F1 Score:  0.5912

Classification Report:

              precision    recall  f1-score   support

   No Arrest       0.58      0.57      0.58      5051
 Arrest Made       0.59      0.60      0.59      5146

    accuracy                           0.58     10197
   macro avg       0.58      0.58      0.58     10197
weighted avg       0.58      0.58      0.58     10197
```

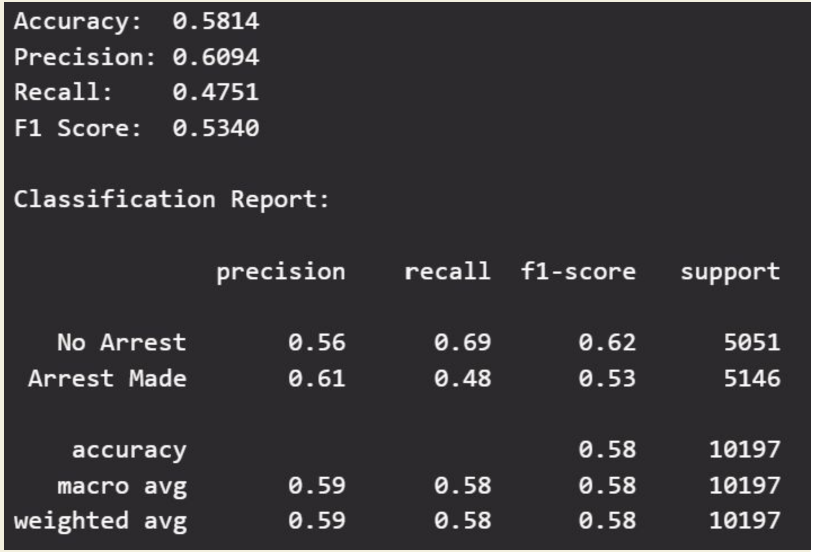


Gaussian Naive Bayes

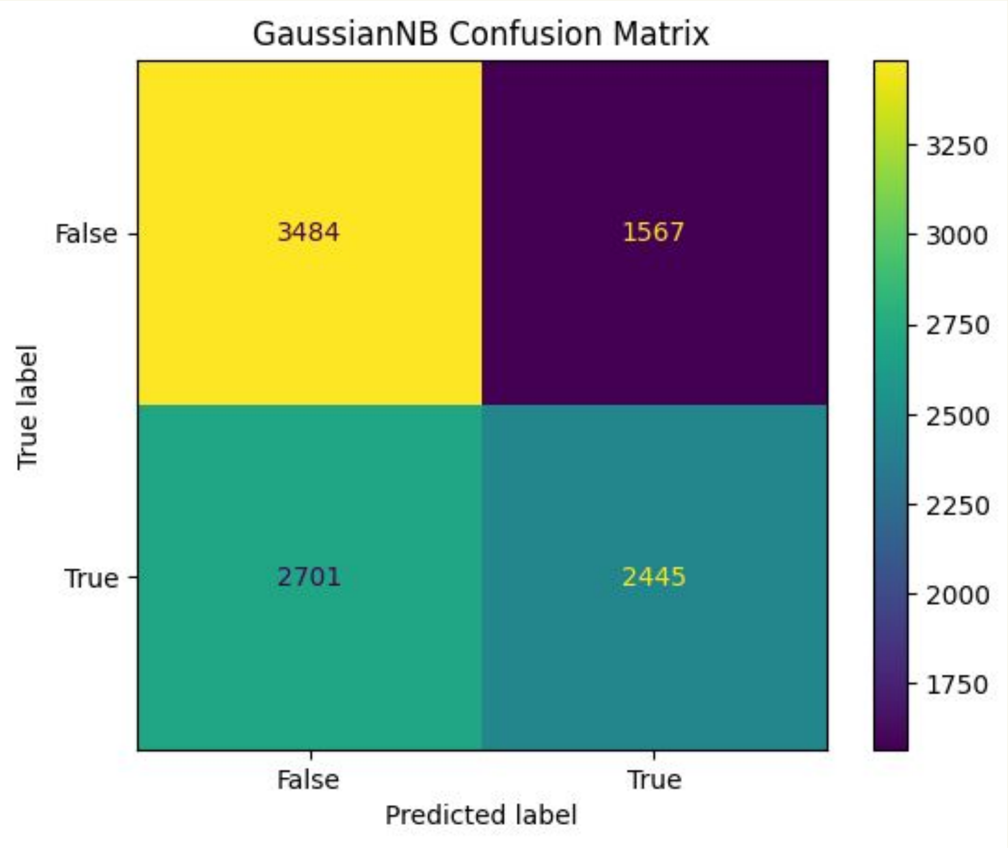
It is based on Bayes' Theorem and assumes that the features used to describe an observation are conditionally independent give the class label


p(x | A)
The probability of observing x, if x came from the Class A distribution
p(x | B)
The probability of observing x, if x came from the Class B distribution
x
$(x-\mu_A)/\sigma_A$
z-score distance of x
$(x-\mu_B)/\sigma_B$
z-score distance of x

Gaussian Naive Bayes

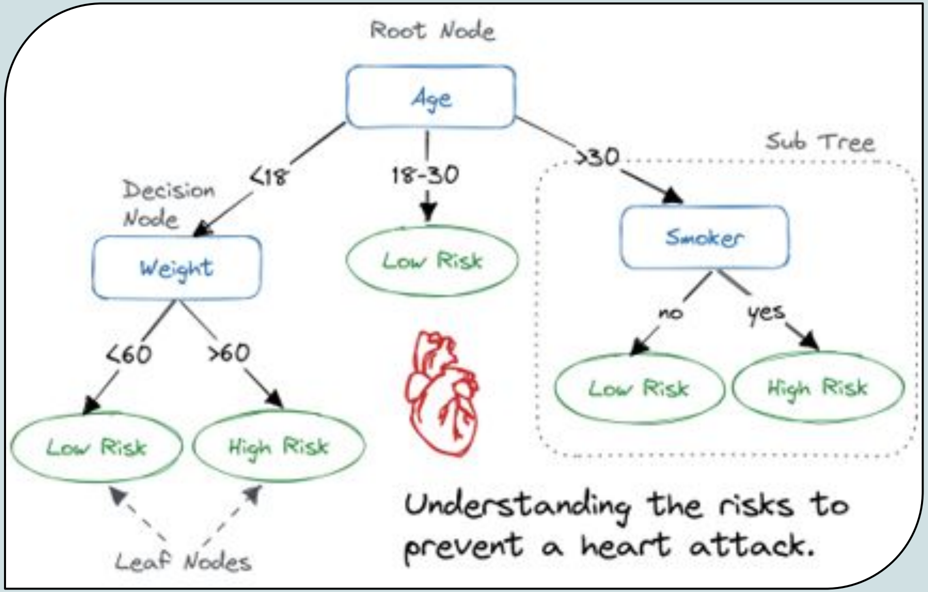


```
Accuracy:  0.5814
Precision: 0.6094
Recall:    0.4751
F1 Score:  0.5340

Classification Report:

              precision    recall  f1-score   support

   No Arrest       0.56      0.69      0.62      5051
 Arrest Made       0.61      0.48      0.53      5146

    accuracy                           0.58     10197
   macro avg       0.59      0.58      0.58     10197
weighted avg       0.59      0.58      0.58     10197
```



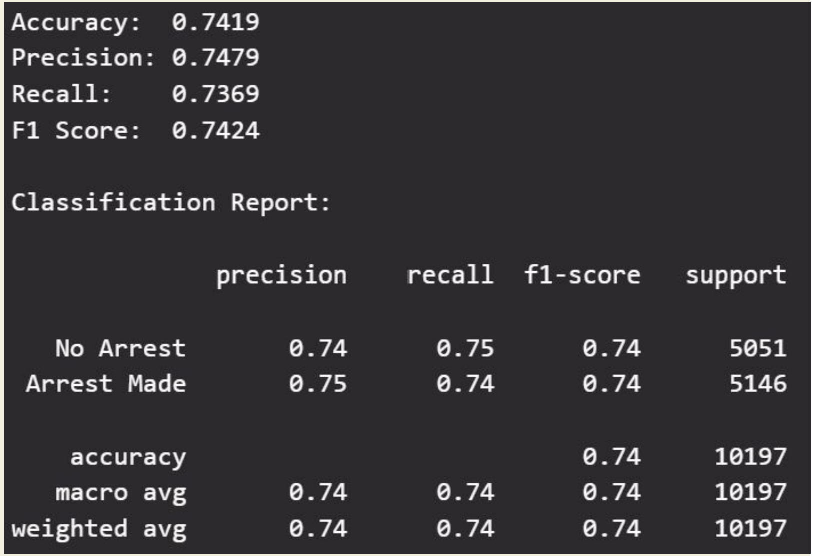
Decision Tree Classifier

Decision trees recursively split datasets into subsets based on the most significant feature at each node, forming a tree structure to facilitate decision making, making them useful for both classification and regression tasks

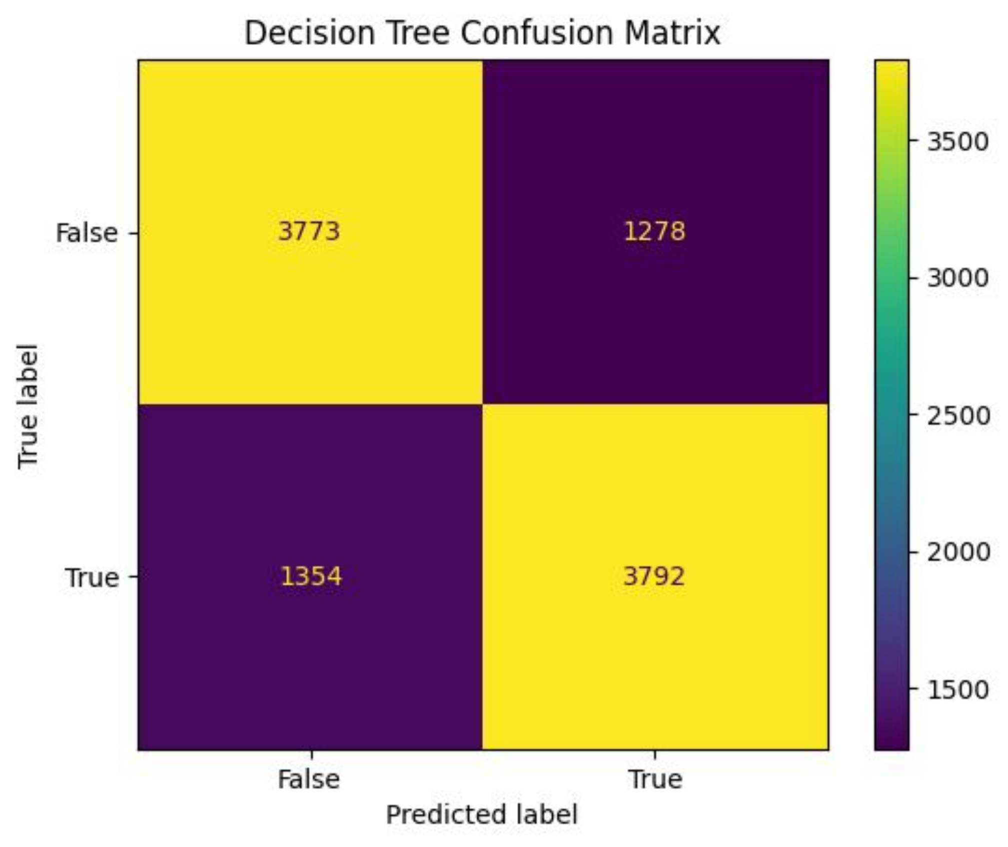


Understanding the risks to prevent a heart attack.

Decision Tree Classifier

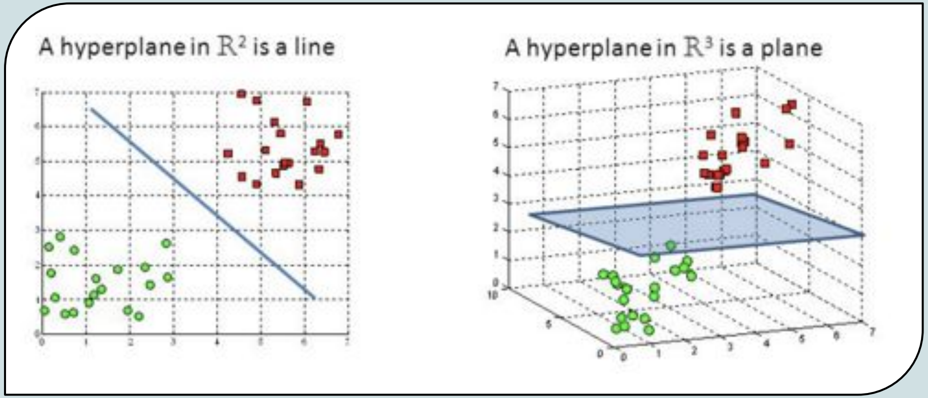


```
Accuracy:  0.7419
Precision: 0.7479
Recall:    0.7369
F1 Score:  0.7424

Classification Report:

              precision    recall  f1-score   support

   No Arrest       0.74      0.75      0.74      5051
 Arrest Made       0.75      0.74      0.74      5146

    accuracy                           0.74     10197
   macro avg       0.74      0.74      0.74     10197
weighted avg       0.74      0.74      0.74     10197
```



Support Vector Machine

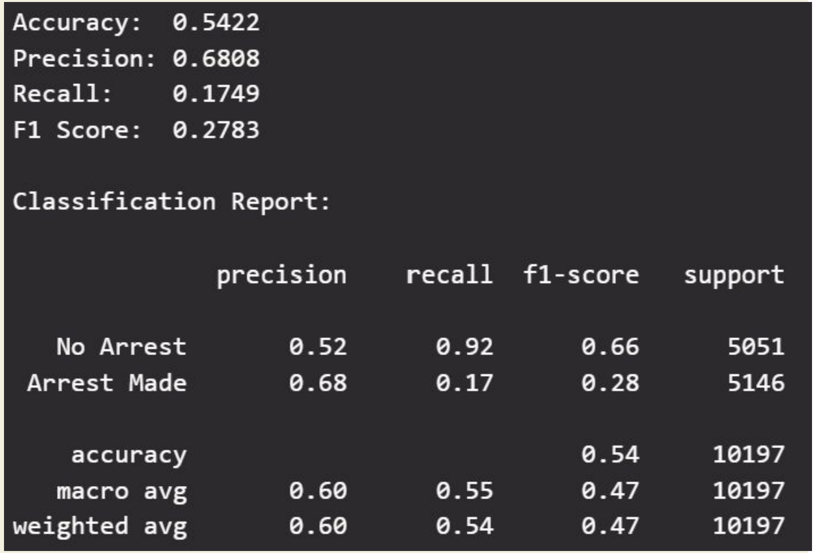
Finds the hyperplane that best separates data points of different classes in a high-dimensional space. Kernel functions enable SVMs to handle non-linear decision boundaries



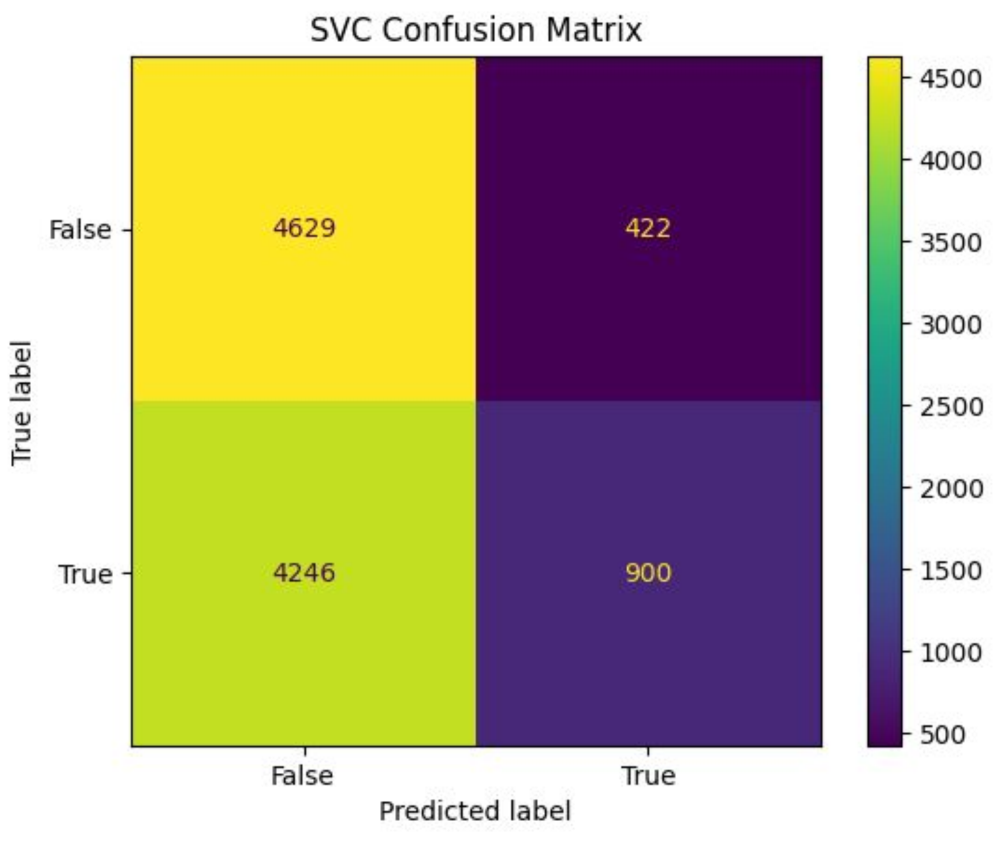
A hyperplane in $\mathbb{R}^2$ is a line

A hyperplane in $\mathbb{R}^3$ is a plane

Support Vector Machine

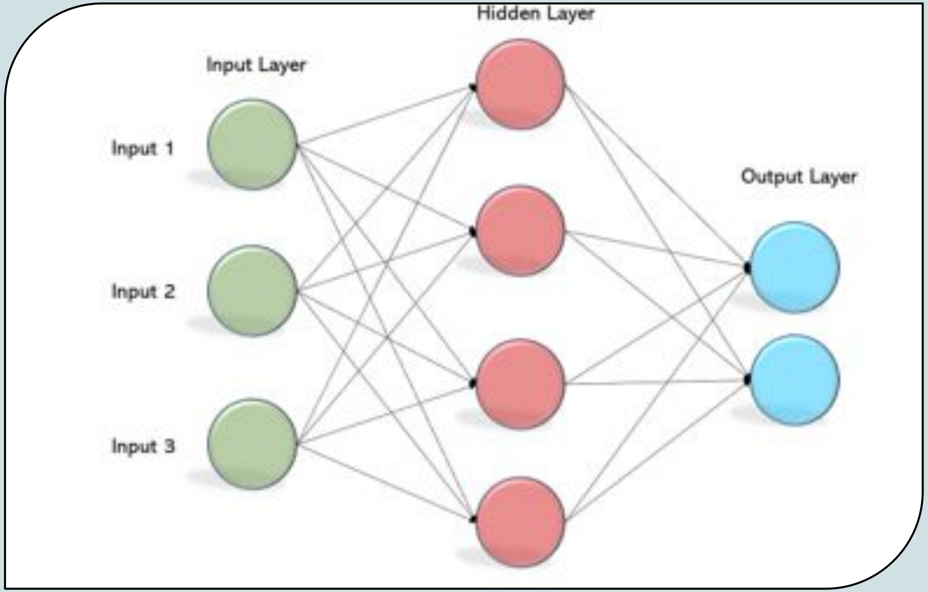


```
Accuracy:  0.5422
Precision: 0.6808
Recall:    0.1749
F1 Score:  0.2783

Classification Report:

              precision    recall  f1-score   support

   No Arrest       0.52      0.92      0.66      5051
 Arrest Made       0.68      0.17      0.28      5146

    accuracy                           0.54     10197
   macro avg       0.60      0.55      0.47     10197
weighted avg       0.60      0.54      0.47     10197
```

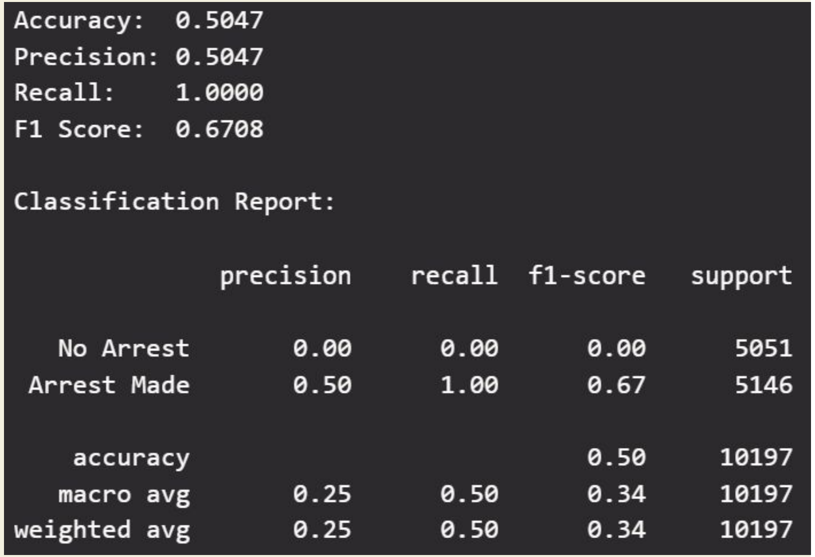


Multi-Layer Perceptron

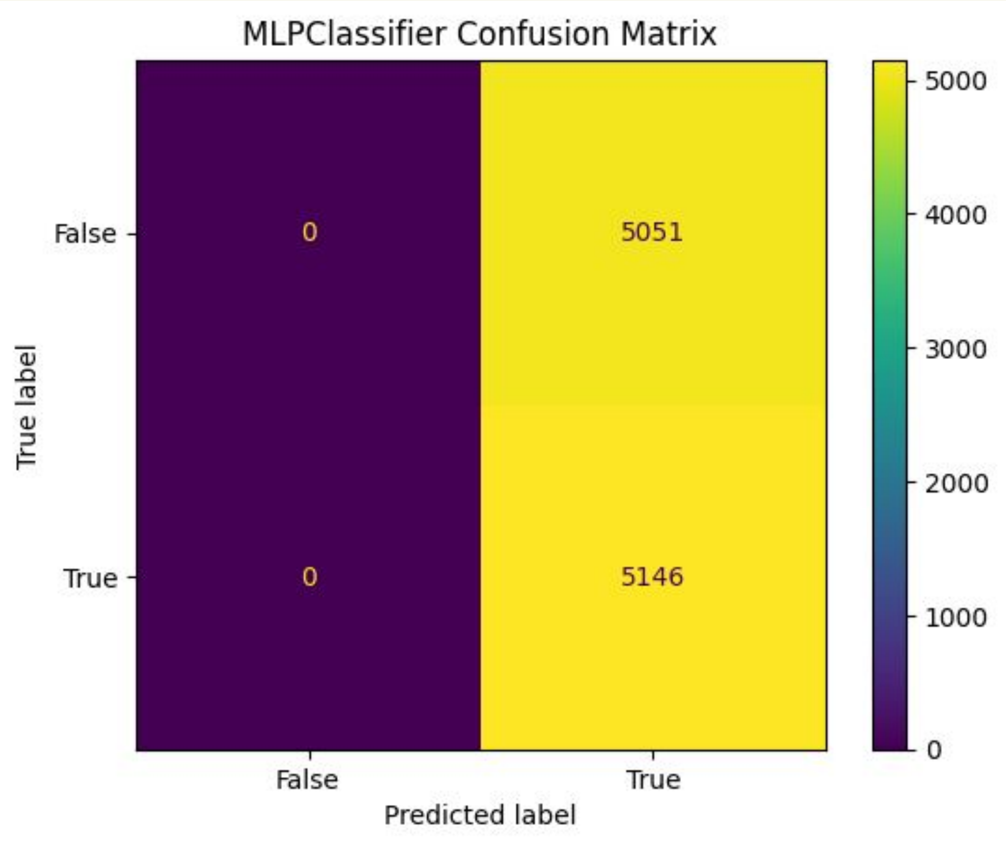
MLP is a type of ANN that can be used for classification tasks. The term 'Perceptron' refers to the individual nodes in the network and 'multilayer' indicates there are multiple layers of these nodes



Multi-Layer Perceptron

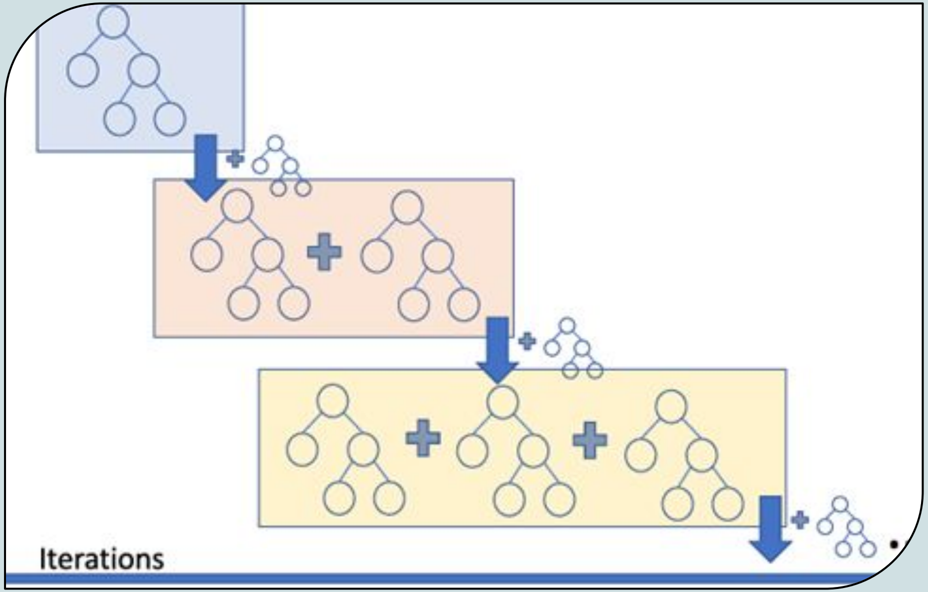


```
Accuracy:  0.5047
Precision: 0.5047
Recall:    1.0000
F1 Score:  0.6708

Classification Report:

              precision    recall  f1-score   support

   No Arrest       0.00      0.00      0.00      5051
 Arrest Made       0.50      1.00      0.67      5146

    accuracy                           0.50     10197
   macro avg       0.25      0.50      0.34     10197
weighted avg       0.25      0.50      0.34     10197
```

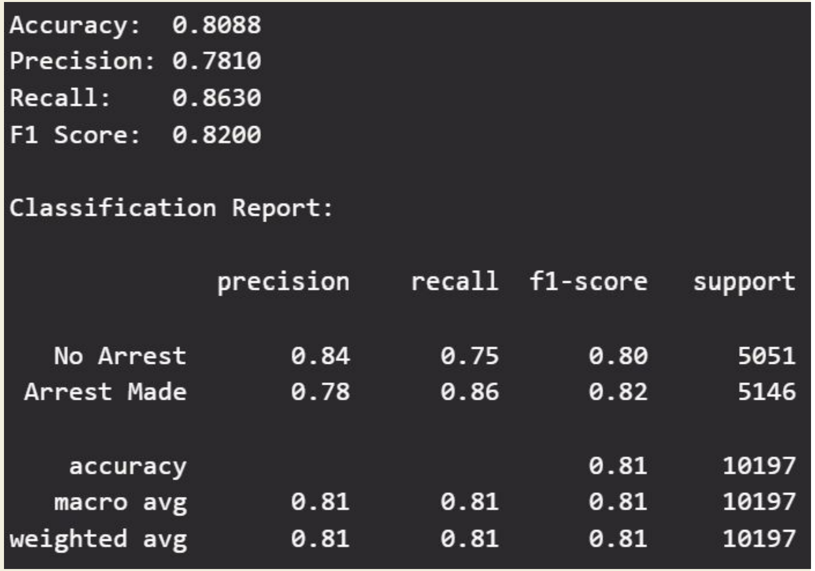


Gradient Boosting Classifier

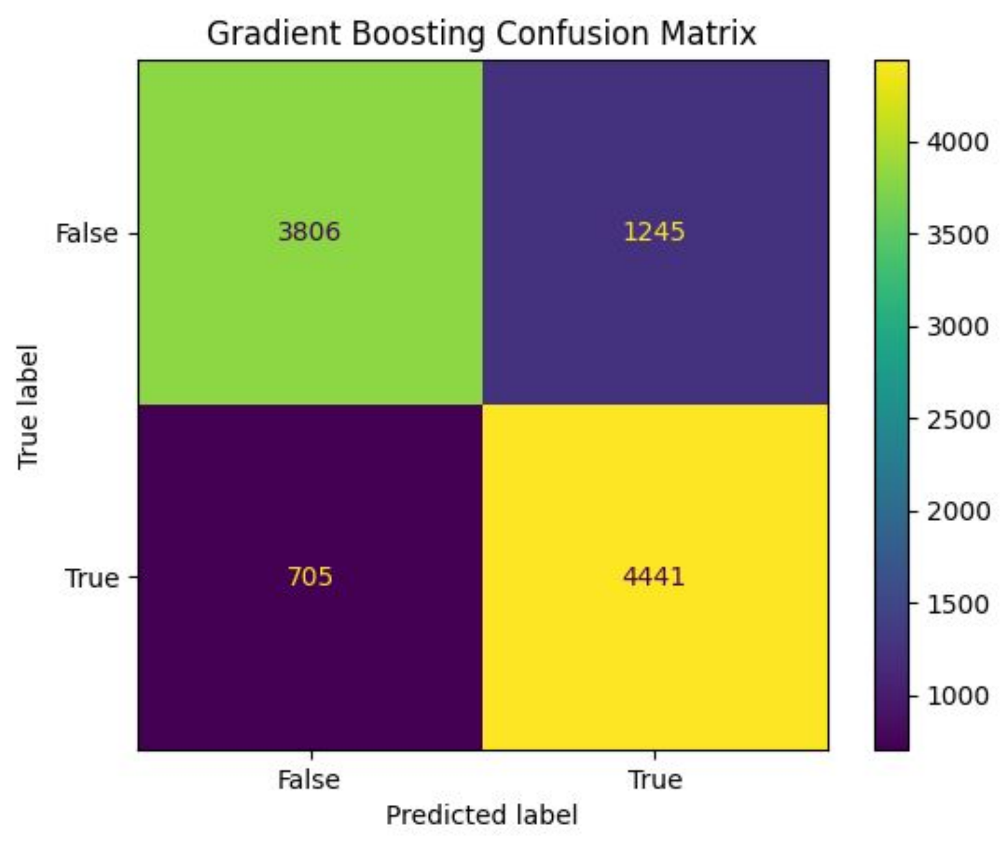
Gradient Boosting Classifier is an ensemble learning technique that builds a series of weak learners, usually decision trees, sequentially, each correcting the errors of its predecessor, ultimately creating a strong predictive model



Gradient Boosting Classifier

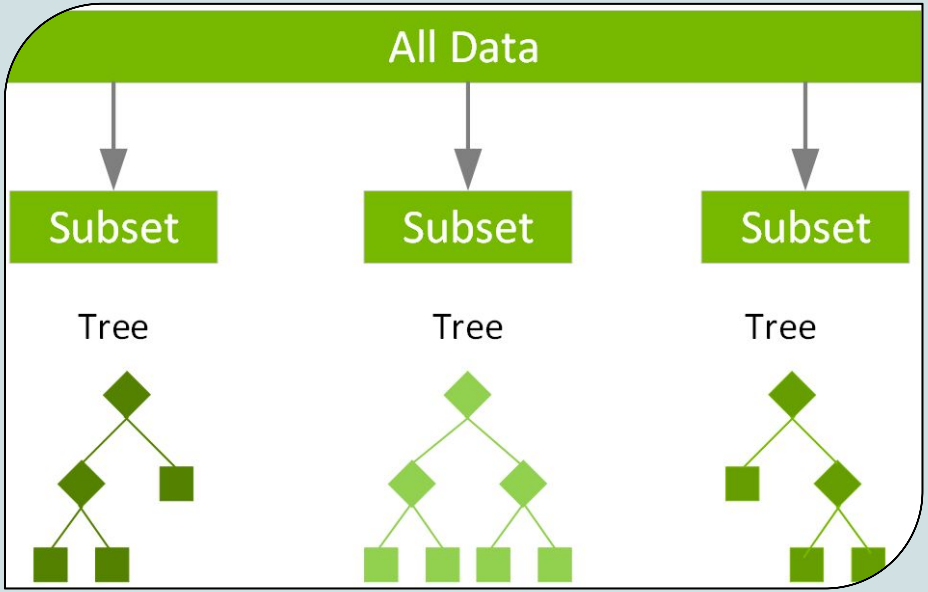


```
Accuracy:  0.8088
Precision: 0.7810
Recall:    0.8630
F1 Score:  0.8200

Classification Report:

              precision    recall  f1-score   support

   No Arrest       0.84      0.75      0.80      5051
 Arrest Made       0.78      0.86      0.82      5146

    accuracy                           0.81     10197
   macro avg       0.81      0.81      0.81     10197
weighted avg       0.81      0.81      0.81     10197
```

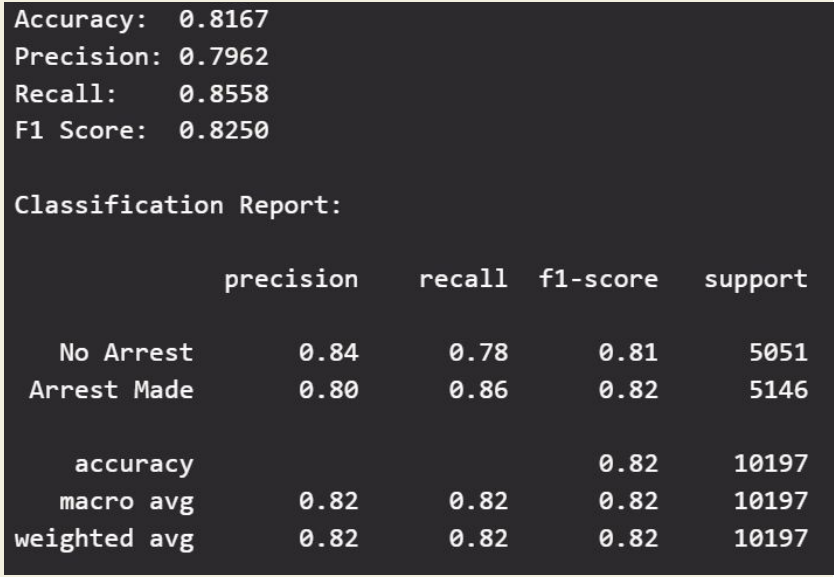


XGBoost Classifier

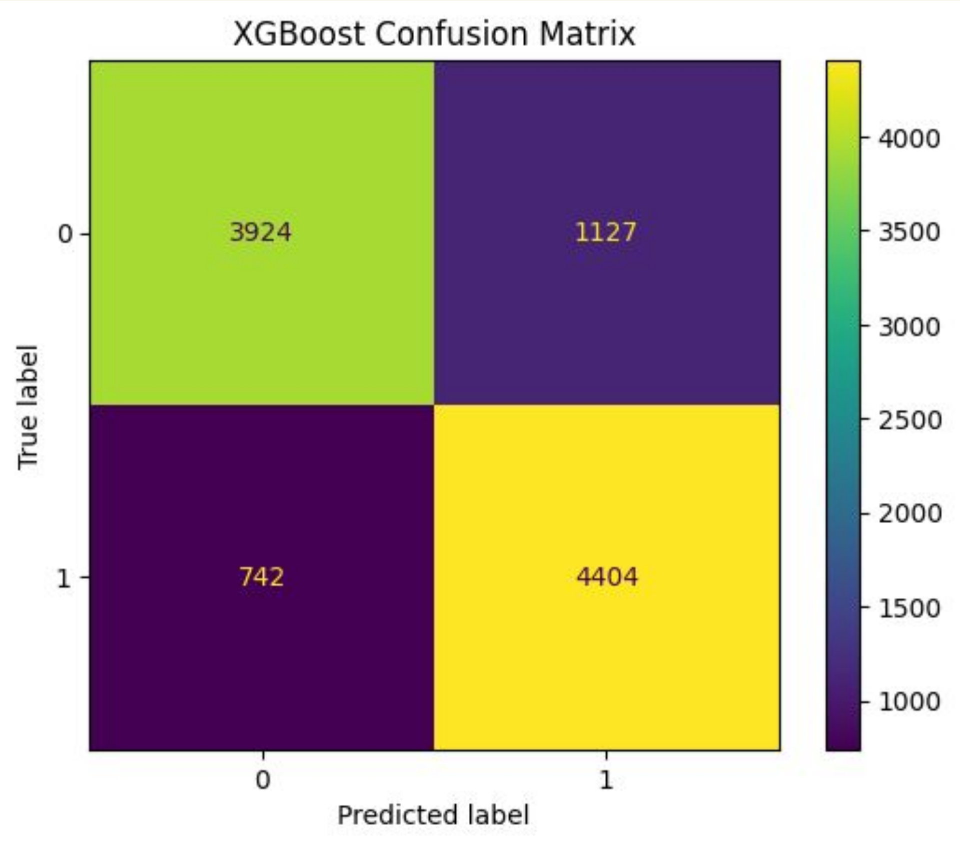
XGBoost (Extreme Gradient Boosting) is an optimized and scalable gradient boosting algorithm that enhances decision trees by employing a regularized objective function, parallel tree construction, and additional features, providing high predictive accuracy and efficiency, making it a popular choice for various machine learning tasks.



XGBoost Classifier

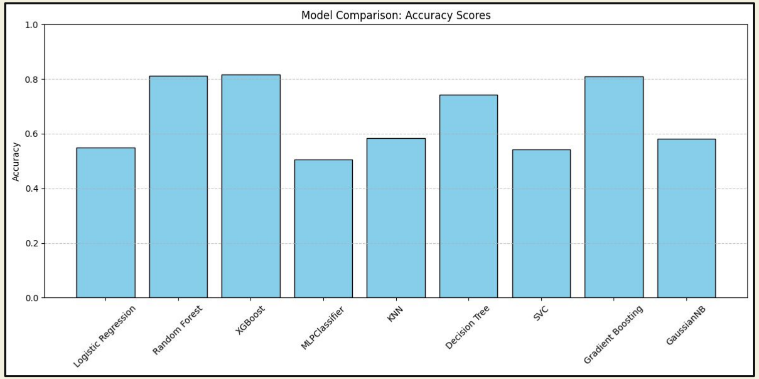


```
Accuracy:  0.8167
Precision: 0.7962
Recall:    0.8558
F1 Score:  0.8250

Classification Report:

              precision    recall  f1-score   support

   No Arrest       0.84      0.78      0.81      5051
  Arrest Made       0.80      0.86      0.82      5146

    accuracy                           0.82     10197
   macro avg       0.82      0.82      0.82     10197
weighted avg       0.82      0.82      0.82     10197
```



Model Results

The results overall were questionable as inherently this type of analysis itself is flawed. It is near impossible to predict whether an arrest was made based on circumstantial information, without any of the information related to the crime or reason they were pulled over.



Model	Accuracy
XGBoost	81.67%
RF	81.26%
GBoost	80.88%
DTree	74.19%
KNN	58.42%
NBayes	58.14%
Logistic	54.82%
SVC	54.22%
MLP	50.47%

Best Model: XGBoost (81.67%)

- Handles complex, nonlinear feature interactions (dataset has many non-linear feature interactions)
- Penalizes overfitting via built-in regularization (Generalizes better thanks to L1/L2)
- Learns from prior mistakes through boosting

Future Improvements

- Bring in more core information, such as what each command code means
    - This would give far more contextual supporting data to each stop to enable the analysis to take into account WHY the stop was made
- Run a similar experiment / analysis with purely the circumstantial data
    - I.e. driver race, sex, location of stop etc.
    - This could act as a preliminary warning for an officer on how an interaction may go based on past history

Thank You!