Mini Project

Stanford Open Policing Project Dataset

Team Kaggle

TABLE OF CONTENTS

INTRODUCTION

Stanford Open Policing Project Dataset - San Francisco

https://openpolicing.stanford.edu/data/

Objective

- ➔ To determine the reasons of individuals being stopped by police while driving
- ➔ To determine if there are any biases for a police to stop someone based on their gender, race, age etc

PROBLEM STATEMENT

If you get stopped by the police in San Francisco, are you being stopped for a valid reason?

DATA PREPARATION

Predictors: **date**, **district**, **subject_age**, **subject_race**, **subject_sex**, **contraband_found**, **search_vehicle**, **reason_for_stop**
Response: **outcome**

Columns Removed:

- Raw Row No./Time/Location/Lat/Lng/Type **(Not needed)**
- Arrest made/citation issued/warning issued **(stated in outcome)**
- Search conducted/Raw result of contact description **(duplication)**
- Search basis **(not useful due to significant portion of other)**
- Raw search vehicle description **(significant portion of categorical data & similar to outcome)**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 905070 entries, 0 to 905069
Data columns (total 22 columns):
 #   Column                              Non-Null Count   Dtype  
---  ------                              --------------   -----  
 0   raw_row_number                      905070 non-null  object 
 1   date                                905070 non-null  object 
 2   time                                905035 non-null  object 
 3   location                            905027 non-null  object 
 4   lat                                 903373 non-null  float64
 5   lng                                 903373 non-null  float64
 6   district                            852883 non-null  object 
 7   subject_age                         846182 non-null  float64
 8   subject_race                        905070 non-null  object 
 9   subject_sex                         905070 non-null  object 
 10  type                                905070 non-null  object 
 11  arrest_made                         905070 non-null  bool   
 12  citation_issued                     905070 non-null  bool   
 13  warning_issued                      905070 non-null  bool   
 14  outcome                             889389 non-null  object 
 15  contraband_found                    53381 non-null   object 
 16  search_conducted                    905070 non-null  bool   
 17  search_vehicle                      905070 non-null  bool   
 18  search_basis                        53381 non-null   object 
 19  reason_for_stop                     902858 non-null  object 
 20  raw_search_vehicle_description      905070 non-null  object 
 21  raw_result_of_contact_description   905070 non-null  object 
dtypes: bool(5), float64(3), object(14)
memory usage: 121.7+ MB
```

DATA PREPARATION

NULL VALUES

- Not all NaN means no data
- Age/District/Search Vehicle/Reason for Stop have missing data
- Contraband found just means there is no search conducted
- Outcome just mean they are stopped without valid reason

	date	district	subject_age	subject_race	subject_sex	outcome	contraband_found	search_vehicle	reason_for_stop
46	2014-08-01	NaN	NaN	hispanic	male	citation	NaN	False	Mechanical or Non-Moving Violation (V.C.)
47	2014-08-01	NaN	NaN	other	female	citation	NaN	False	Moving Violation
48	2014-08-01	NaN	NaN	asian/pacific islander	male	citation	NaN	False	Mechanical or Non-Moving Violation (V.C.)
49	2014-08-01	NaN	NaN	other	male	citation	NaN	False	Moving Violation
50	2014-08-01	NaN	NaN	hispanic	male	NaN	NaN	False	Mechanical or Non-Moving Violation (V.C.)

	date	district	subject_age	subject_race	subject_sex	outcome	contraband_found	search_vehicle	reason_for_stop
33211	2007-01-01	C	48.0	black	male	citation	False	True	Mechanical or Non-Moving Violation (V.C.)
33212	2007-01-01	H	16.0	asian/pacific islander	male	citation	search not conducted	False	Moving Violation
33213	2007-01-01	H	23.0	asian/pacific islander	male	citation	search not conducted	False	Mechanical or Non-Moving Violation (V.C.)
33214	2007-01-01	G	20.0	black	male	warning	search not conducted	False	Moving Violation
33215	2007-01-01	A	35.0	white	male	without valid reason	search not conducted	False	Moving Violation

CHARACTERISTICS

	date	district	subject_age	subject_race	subject_sex	outcome	contraband_found	search_vehicle	reason_for_stop
0	2007-01-01	C	56.0	white	male	citation	search not conducted	False	Mechanical or Non-Moving Violation (V.C.)
1	2007-01-01	B	32.0	white	male	citation	search not conducted	False	Moving Violation
2	2007-01-01	I	57.0	asian/pacific islander	female	citation	search not conducted	False	Moving Violation
3	2007-01-01	A	31.0	hispanic	male	warning	search not conducted	False	Moving Violation
4	2007-01-01	J	37.0	hispanic	female	citation	search not conducted	False	Moving Violation

	date	district	subject_age	subject_race	subject_sex	outcome	contraband_found	search_vehicle	reason_for_stop
count	844223	844223	844223.000000	844223	844223	844223	844223	844223	844223
unique	3254	13	NaN	5	2	4	3	2	7
top	2009-03-20	H	NaN	white	male	citation	search not conducted	False	Moving Violation
freq	627	116470	NaN	350317	595578	589829	794624	794624	526427
mean	NaN	NaN	37.830546	NaN	NaN	NaN	NaN	NaN	NaN
std	NaN	NaN	13.634125	NaN	NaN	NaN	NaN	NaN	NaN
min	NaN	NaN	10.000000	NaN	NaN	NaN	NaN	NaN	NaN
25%	NaN	NaN	27.000000	NaN	NaN	NaN	NaN	NaN	NaN
50%	NaN	NaN	35.000000	NaN	NaN	NaN	NaN	NaN	NaN
75%	NaN	NaN	47.000000	NaN	NaN	NaN	NaN	NaN	NaN
max	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	NaN	NaN

01 CATEGORICAL

Most of the data are categorical

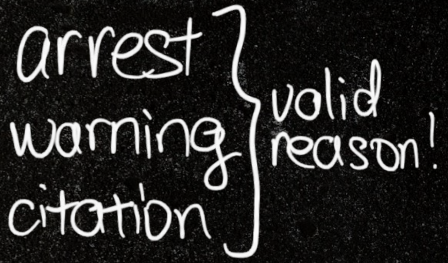
02 BINARY OUTCOME

Outcome is simplified to with or without valid reason

03 MULTIVARIATE

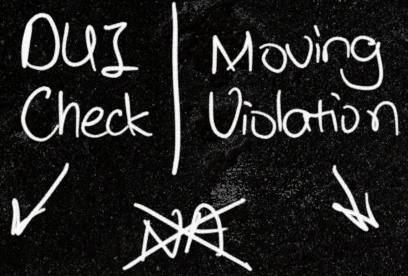
Multiple variables will affect the outcome

FEATURE ENGINEERING



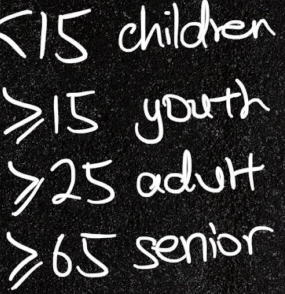
BINARY

Arrest, Warning, Citation grouped as valid reason, and NA as no valid reason



SPLITTING

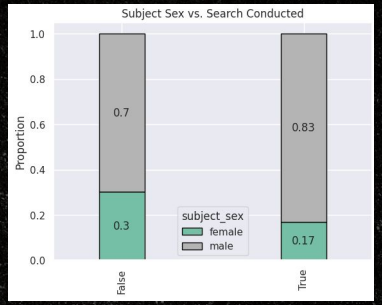
Reason_for_stop sometimes has multiple entries, therefore data is split



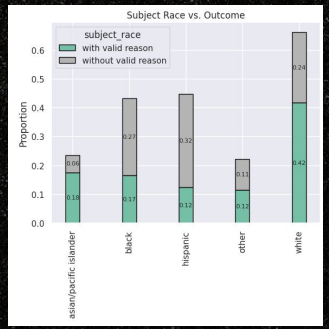
ENCODING

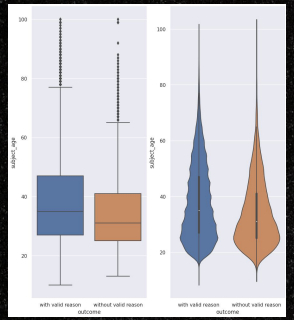
Age, an ordinal data, converted into categorical data

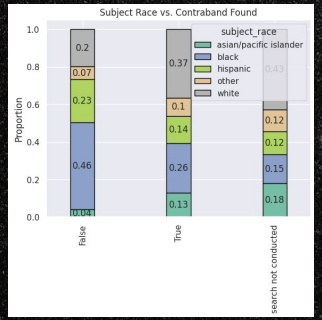
EXPLORATORY ANALYSIS



Police stops are generally fairly objective, but vehicular search is biased. Age, gender, and particularly race does affect if you will be stopped for a valid reason or not.







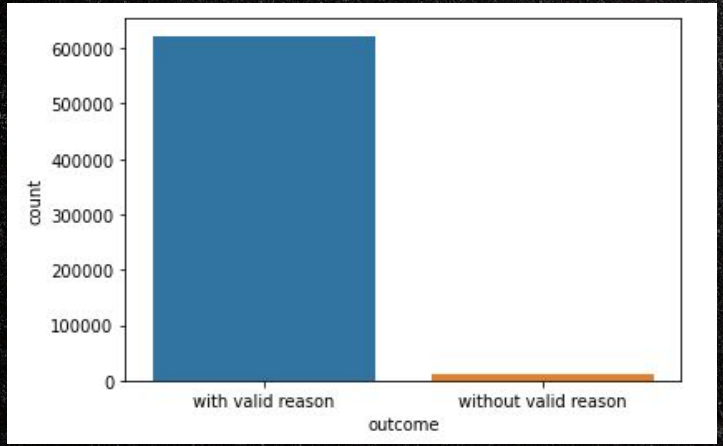
MODEL BUILDING

Train & Test Data

Split the dataset into 75% train and 25% test

```
outcome
with valid reason        622237
without valid reason      10930
dtype: int64
```

Majority: With valid reason
Minority: Without valid reason



MODEL BUILDING

Random Oversampling

Examples in the minority class are selected randomly and duplicated.

Pros:

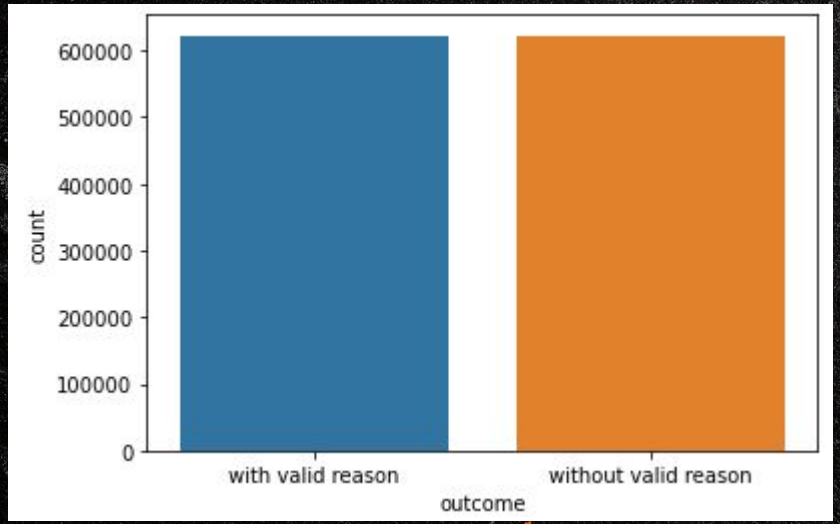
Does not lose any informations

Cons:

It can cause overfitting and poor generalisation to the test set

Majority
Minority

```
outcome
with valid reason        622237
without valid reason     622237
dtype: int64
```



MODEL BUILDING

Synthetic Minority Over-Sampling Technique for Nominal (SMOTEN)

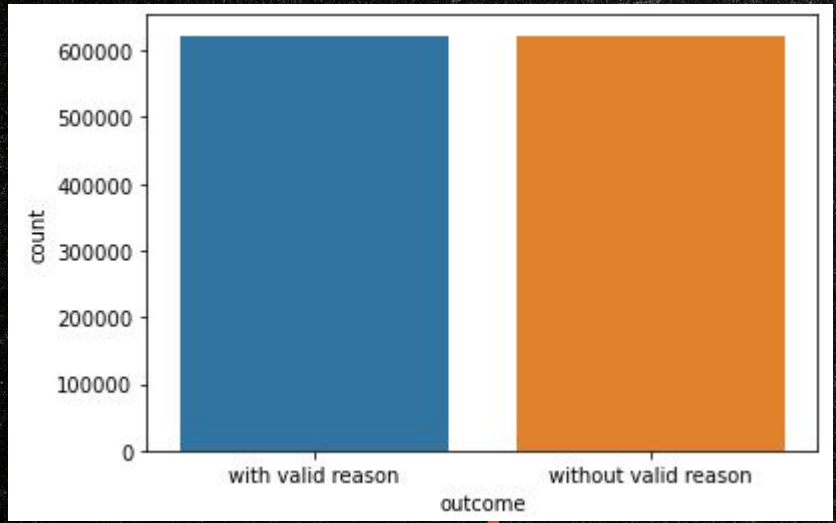
Uses a k-nearest neighbour algorithm to create new synthetic samples to balance the dataset.

Cons:

The synthetic samples are created without considering the majority class.

Majority
Minority

```
outcome
with valid reason        622237
without valid reason     622237
dtype: int64
```



MODEL BUILDING

Combination

Random Undersampling and Oversampling

- Undersample majority class (with valid reason) to 400 000.
- Oversample minority class (without valid reason) to 400 000.

Random Undersampling and SMOTEN

- Undersample “with valid reason” to 400 000.
- SMOTEN “without valid reason” to 400 000.

Majority
Minority

```
outcome
with valid reason        400000
without valid reason     400000
dtype: int64
```

MODEL BUILDING

Ordinal Variables

- Age band
    1. Children
    2. Youth
    3. Adult
    4. Senior
- Outcome
    1. With valid reason
    2. Without valid reason

Nominal Variables

- Race
- Gender
- Contraband Found
- Vehicle Searched
- Reason for Stop

MODEL BUILDING

Label Encoding

- Each label is assigned a unique integer based on alphabetical ordering
- Used on ordinal variables

0	Adult
1	Children
2	Senior
3	Youth

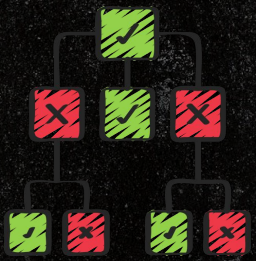
One Hot Encoding

- Each unique value in the category will be added as a feature
- Used on nominal variables

	subject_race_asian/pacific islander	subject_race_black	subject_race_hispanic	subject_race_other	subject_race_white
0	0.0	0.0	0.0	0.0	1.0
1	0.0	0.0	0.0	0.0	1.0
2	0.0	0.0	0.0	0.0	1.0
3	0.0	0.0	0.0	0.0	1.0
4	0.0	0.0	0.0	1.0	0.0

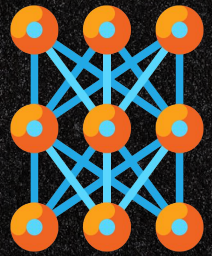
PREDICTORS & RESPONSE

MODEL USED



Machine Learning

Decision Tree



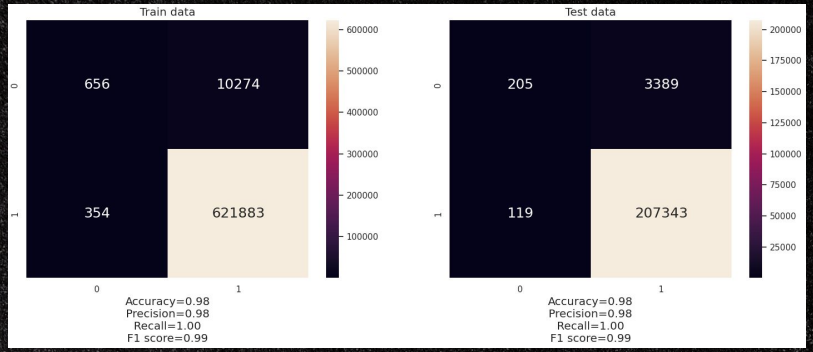
Deep Learning

Artificial Neural Network (ANN)

DECISION TREE
(depth=6 & K-Fold cross validation)

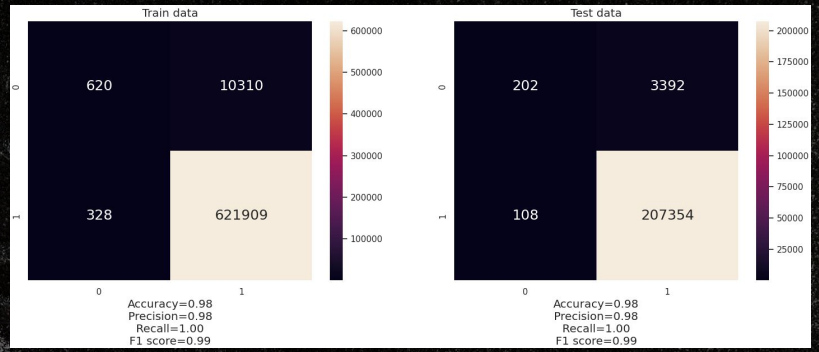
Attempt 1

depth=6 with imbalanced data



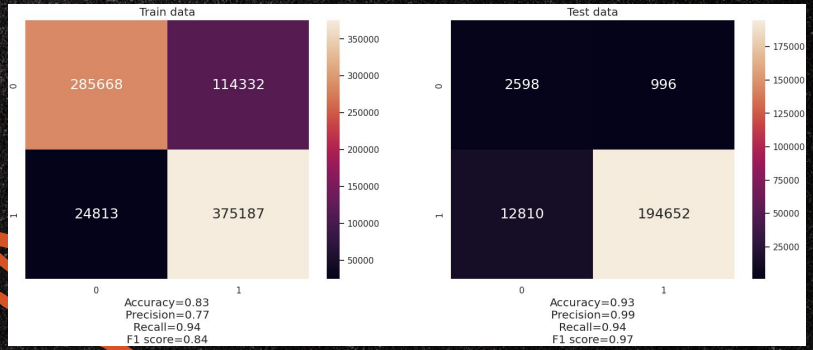
Attempt 2

optimal with imbalanced data



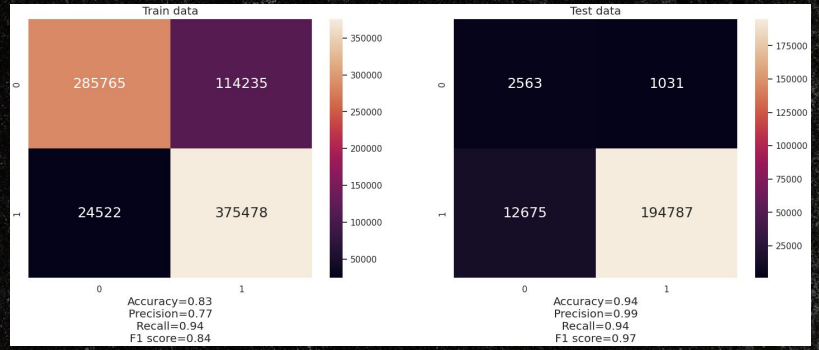
Attempt 3

optimal with random oversampling data



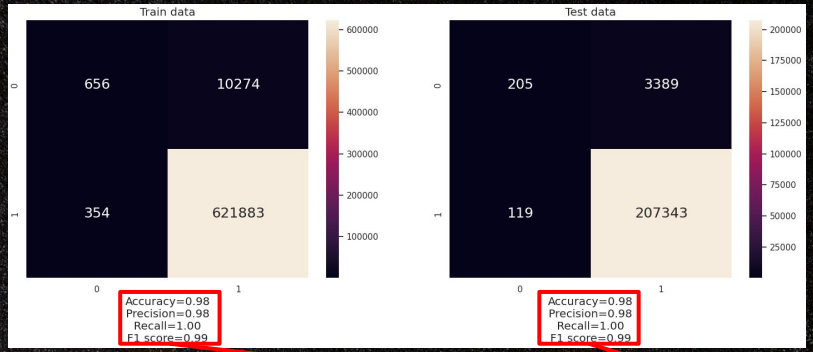
Attempt 4

optimal with SMOTEN data



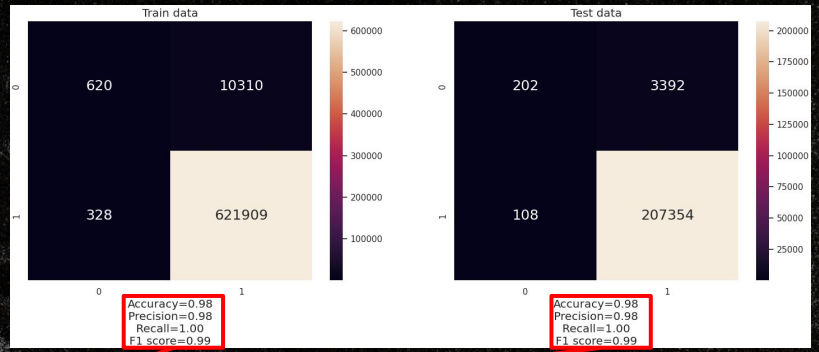
Attempt 1

depth=6 with imbalanced data



Attempt 2

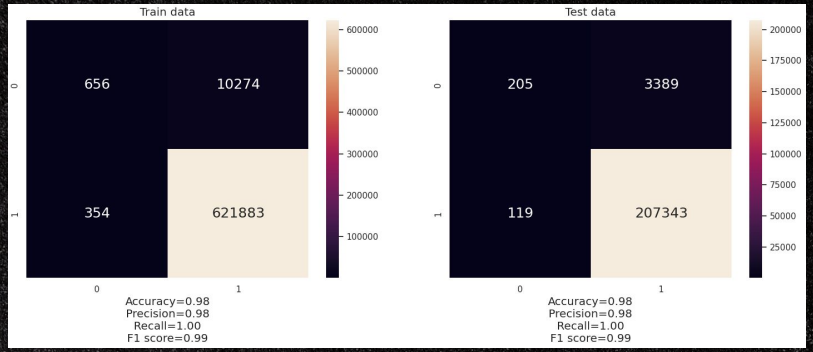
optimal with imbalanced data



Metric	Train data	Test data
Accuracy	0.98	0.98
Precision	0.98	0.98
Recall	1.00	1.00
F1 score	0.99	0.98

Attempt 1

depth=6 with imbalanced data



Train data

98% Accuracy → 83% Accuracy

100% Recall → 94% Recall

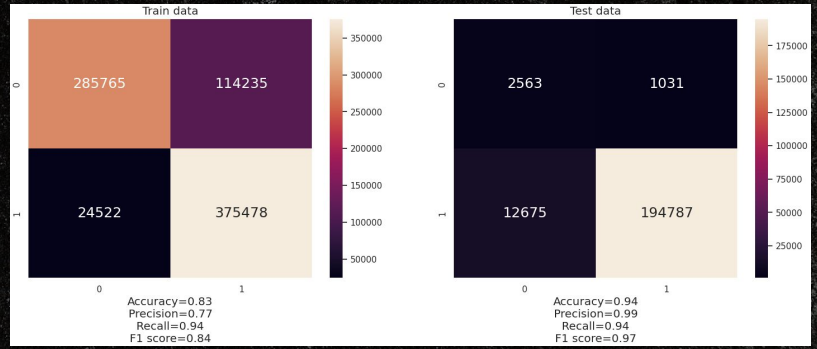
Test data

98% Accuracy → 94% Accuracy

100% Recall → 94% Recall

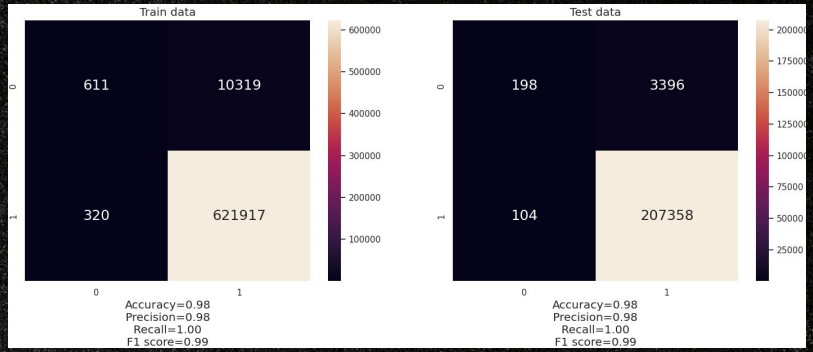
Attempt 4

optimal with SMOTEN data

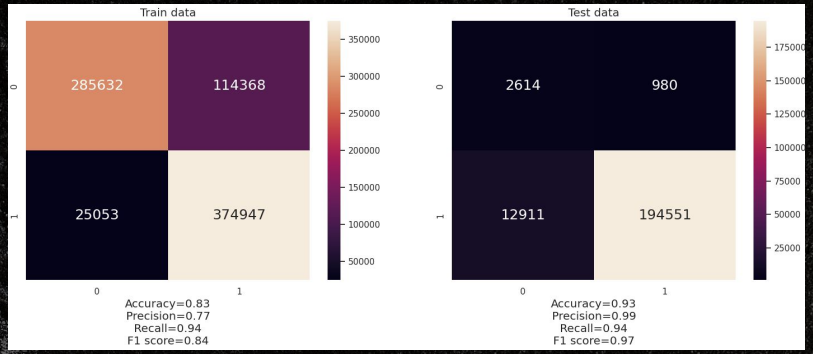


ARTIFICIAL NEURAL NETWORK (ANN)

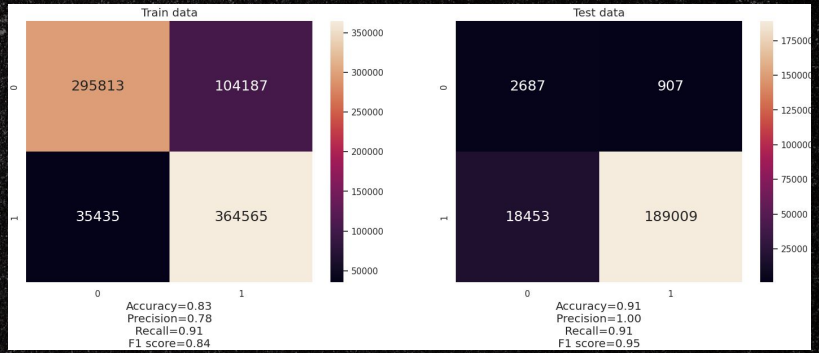
Attempt 5
imbalanced data



Attempt 6
random oversampling data

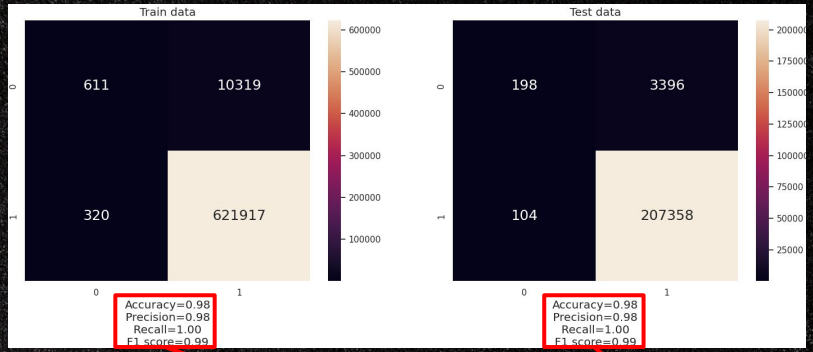


Attempt 7
SMOTEN data



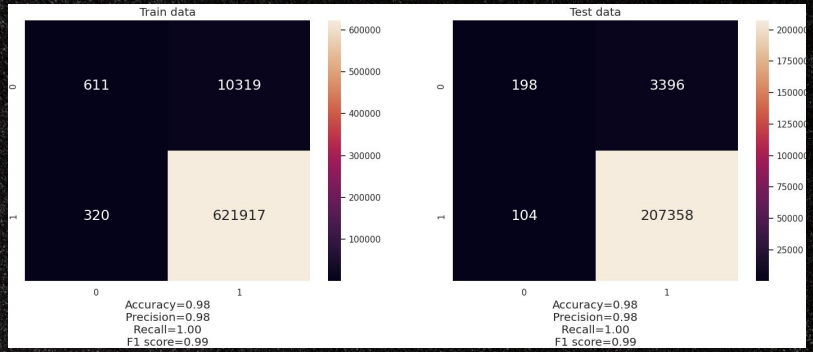
Attempt 5

imbalanced data



Metric	Train data	Test data
Accuracy	0.98	0.98
Precision	0.98	0.98
Recall	1.00	1.00
F1 score	0.99	0.99

Attempt 5
imbalanced data

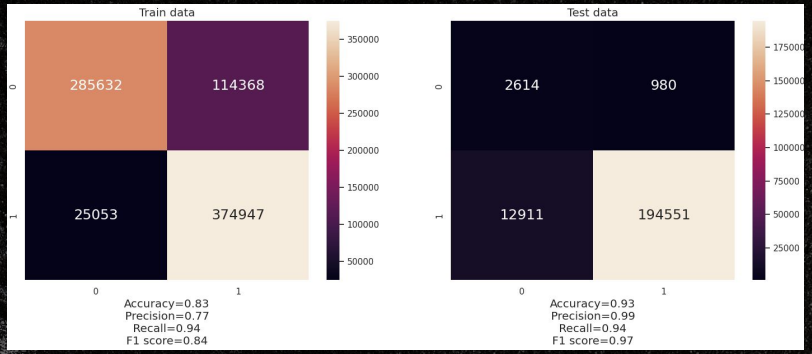


Train data

98% Accuracy → 83% Accuracy

100% Recall → 94% Recall

Attempt 6
random oversampling data



Test data

98% Accuracy → 93% Accuracy

100% Recall → 94% Recall

TEST DATA PERFORMANCE

Model	Accuracy		
	Without Downsampling & Oversampling	Downsampling & Oversampling	
		Random Undersampling & Random Oversampling	Random Undersampling & SMOTEN Oversampling
Decision Tree (depth=10)	0.98		
Decision Tree (optimal)	0.98	0.93	0.94
Artificial Neural Network (ANN)	0.98	0.93	0.91

THINGS WE APPLIED

	Encoding	Resampling	Cross-Validation	Machine Learning
New Techniques	❑ Label encoding ❑ One-hot encoding	❑ Random undersampling ❑ Random oversampling ❑ SMOTEN oversampling	❑ K-Fold cross validation	❑ Artificial neural network (ANN) (Classification)

	Data Preparation	Exploratory Data Analysis	Machine Learning
Things We Learnt Before	❑ Data Cleaning	❑ Correlation	❑ Decision Tree (Classification)

MEMBERS' CONTRIBUTION

Pipeline	Member
Data Preparation	
i) Data Cleaning	Jackson, Jordan, Jun Yong
Exploratory Analysis	
i) Feature Engineering	Jackson, Jun Yong
ii) Exploratory Analysis	Jackson, Jordan, Jun Yong

Pipeline	Member
Model Building	
i) Resampling	Peiqi, Jun Yong
ii) Label Encoding	Jordan, Jun Yong
iii) One-hot Encoding	Peiqi, Jun Yong
Machine Learning	
i) Decision Tree	Jun Yong
ii) Artificial Neural Network	Jun Yong

CONCLUSION

Outcome

Our machine learning model is able to predict if a person is stopped by a police with valid reason or without valid reason, based on the predictors:

- Age
- Gender
- Race
- Reasons for stop

Meeting Our Initial Objective

- To determine the reasons of individuals being stopped by police while driving
- To determine if there are any biases for a police to stop someone based on their gender, race, age etc

THANK YOU