Victim Relationship Prediction

Done by:
Renad Albishri
Amjad Althinyyan

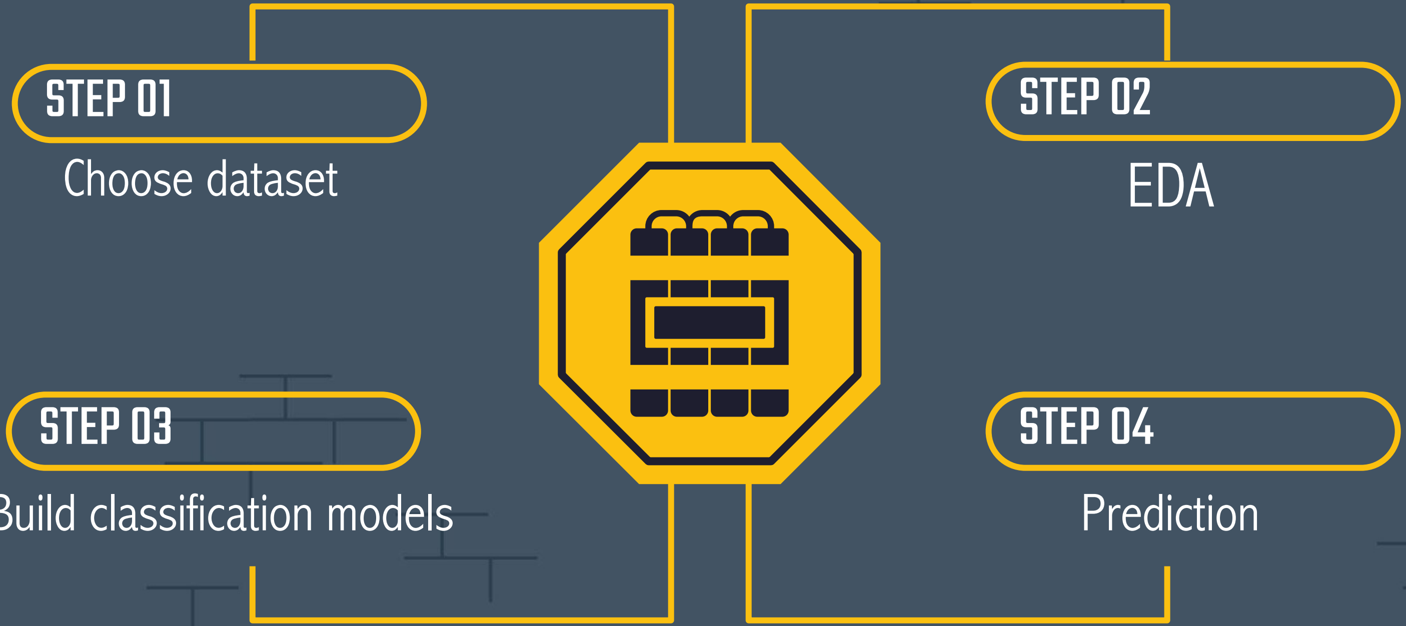
Overview

In this project, we are working on a dataset that consists of information about the crimes in USA, to identify a pattern that may be used to control and limit these crimes, by using Classification.

Goals

- Build models to predict the perpetrator's relationship with the victim.
- Choose the model that give us the best predict.

Methodology



STEP 01

Choose dataset

STEP 02

EDA

STEP 03

Build classification models

STEP 04

Prediction

Dataset

Dataset used

Homicide dataset

Which contains?

Homicide Report from 1976 to 2014

Numbers of rows & columns

638454 rows
24 columns

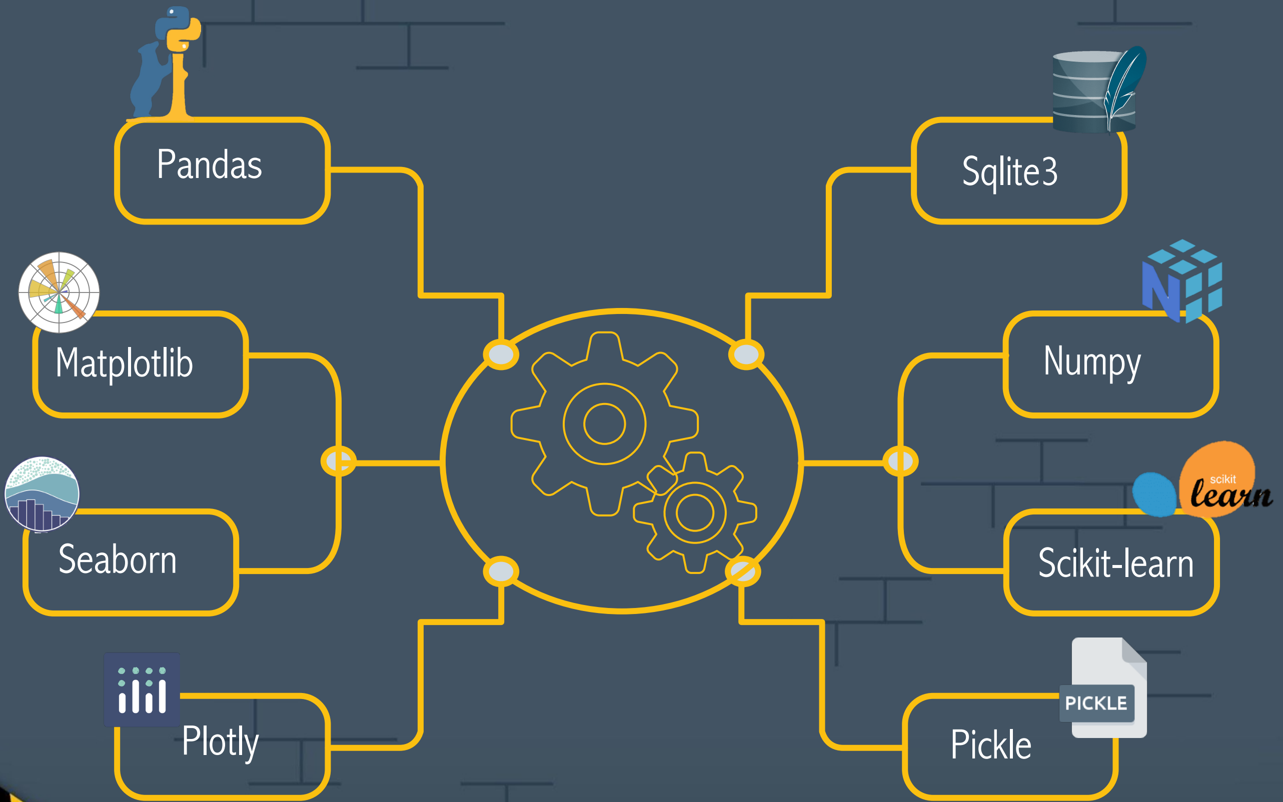
Features

Age , Race, Gender, Ethnicity of victims and perpetrators, ,Weapon used

Target

Relationship between victim and perpetrator

Tools and libraries



Pandas

Matplotlib

Seaborn

Plotly

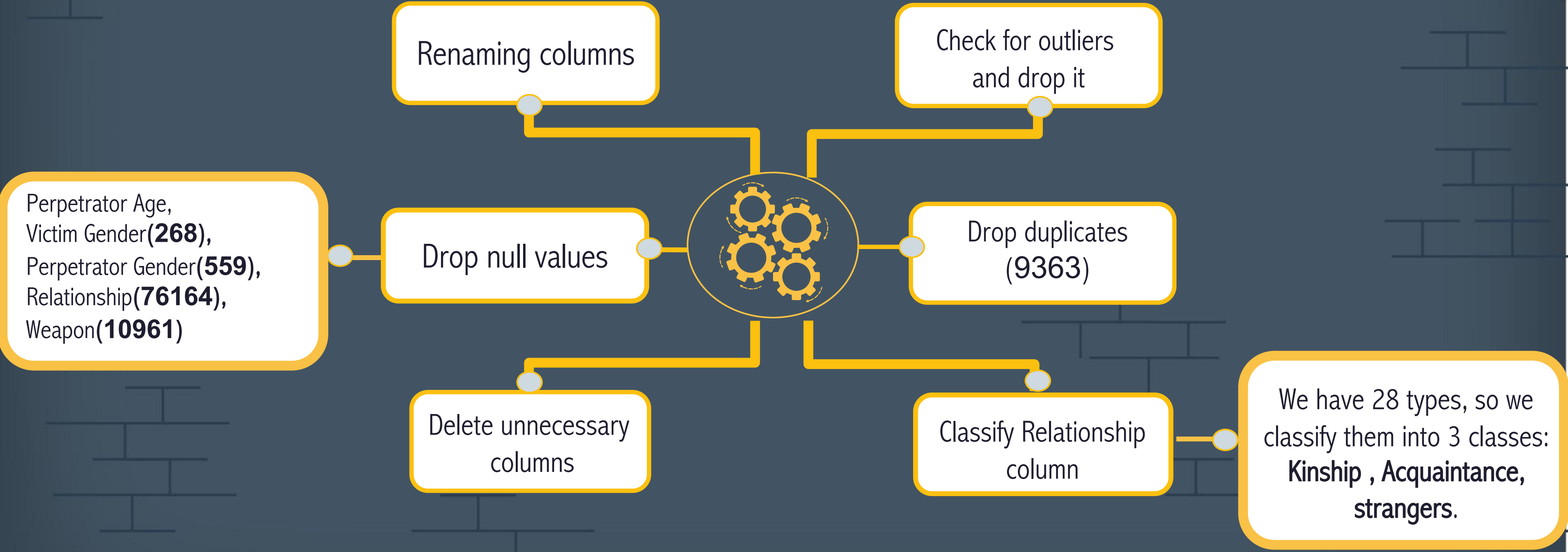
Sqlite3

Numpy

Scikit-learn

Pickle

EDA



Renaming columns

Check for outliers and drop it

Drop null values

Perpetrator Age,
Victim Gender**(268)**,
Perpetrator Gender**(559)**,
Relationship**(76164)**,
Weapon**(10961)**

Drop duplicates (9363)

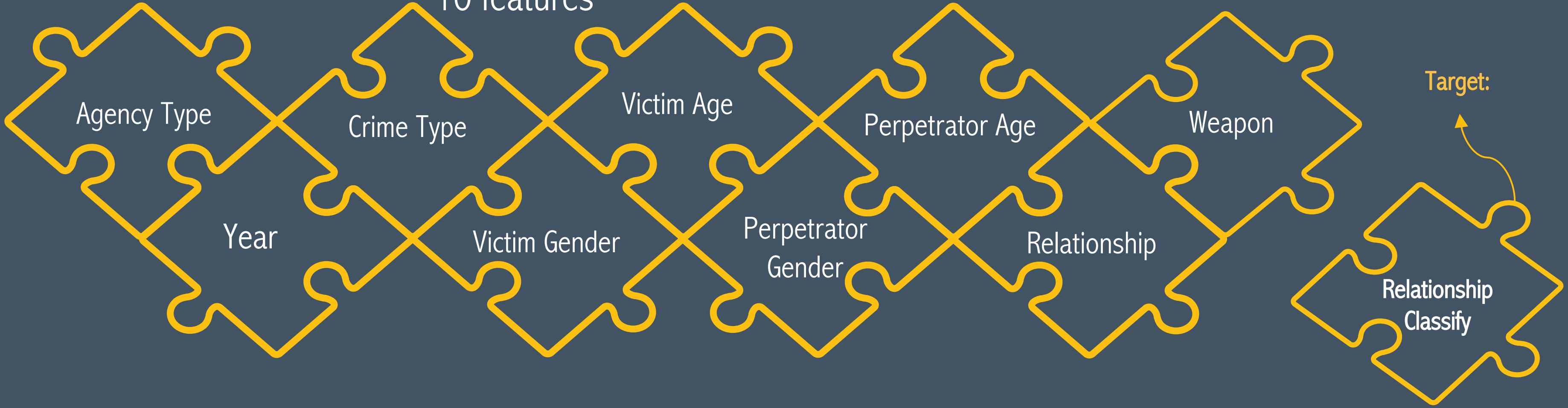
Delete unnecessary columns

Classify Relationship column

We have 28 types, so we classify them into 3 classes: **Kinship , Acquaintance, strangers**.

EDA cont...

After cleaning the data, the dataset becomes:
284,629 rows
10 features



Agency Type

Year

Crime Type

Victim Gender

Victim Age

Perpetrator Gender

Perpetrator Age

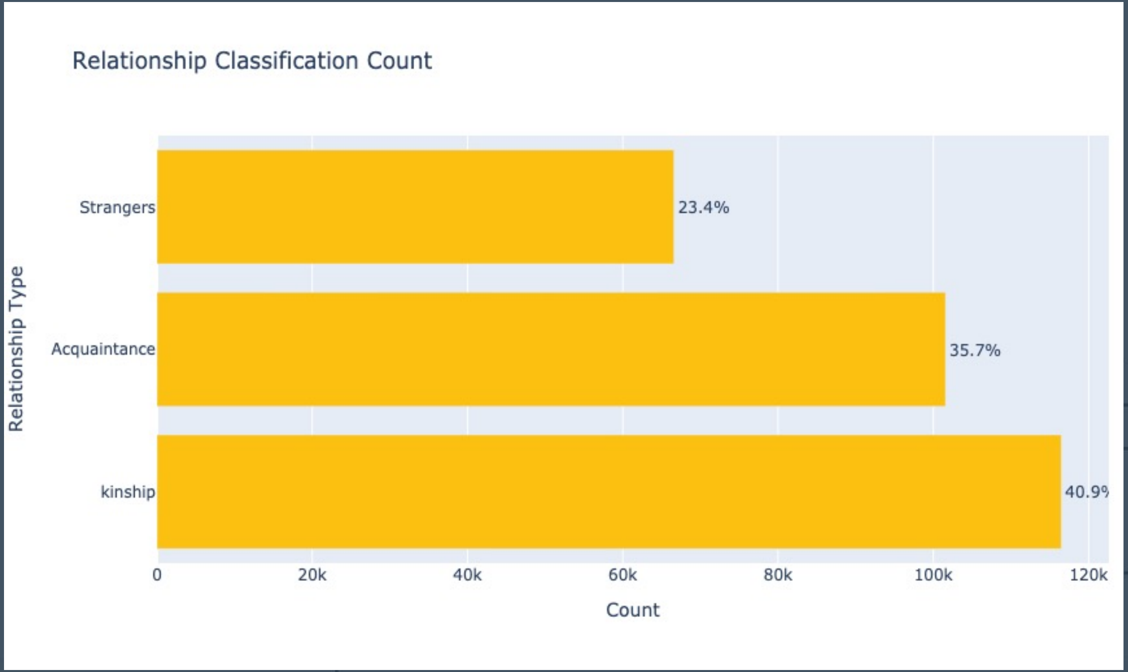
Relationship

Weapon

Target:

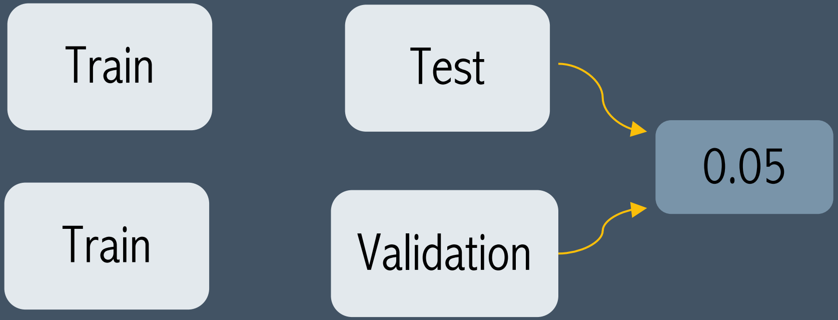
Relationship Classify

Classify the Relationship



Split Data

After converting to dummy variables, the dataset becomes:
284,629 rows
78 features



Train

Test

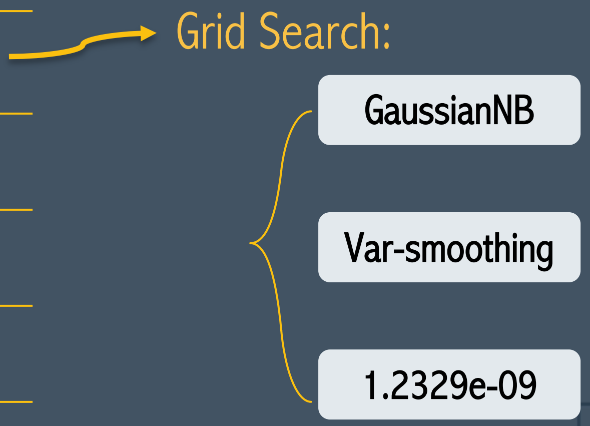
Train

Validation

0.05

Predicted models before feature engineering

Model	Training	Validation
Logistic Regression	0.9304	0.9308
GaussianNB	**0.9986**	**0.9987**
Random Forest Classifier	0.8175	0.8169
BernoulliNB	0.9594	0.9601
Decision Tree Classifier	0.9903	0.9893
Gradient Boosting Classifier	0.8673	0.8661



Grid Search:

GaussianNB

Var-smoothing

1.2329e-09

Feature Engineering

Model	Training	Validation
Logistic Regression	0.8911	0.8908
Decision Tree Classifier	0.9903	0.9893
Random Forest Classifier	0.9037	0.9050
BernoulliNB	0.9592	0.9599
GaussianNB	0.9985	0.9987
Gradient Boosting Classifier	0.8673	0.8661

First iteration:

- Difference between victim and perpetrator age column.

Feature Engineering

Model	Training	Validation
Logistic Regression	0.9359	0.9374
Decision Tree Classifier	0.9903	0.9893
Random Forest Classifier	0.8935	0.8942
BernoulliNB	0.9576	0.9589
GaussianNB	**0.9988**	**0.9989**
Gradient Boosting Classifier	0.8673	0.8661

✓ **Second iteration:**

- ➕ Difference between victim and perpetrator age column.
- 🗑 Victim ,perpetrator age columns.

Feature Engineering

Model	Training	Validation
Logistic Regression	0.9277	0.9268
Decision Tree Classifier	0.9903	0.9893
Random Forest Classifier	0.9073	0.9075
BernoulliNB	0.9576	0.9589
GaussianNB	0.8807	0.8786
Gradient Boosting Classifier	0.8673	0.8661

Third iteration:

- Number of crimes column for each state.

Feature Engineering

Model	Training	Validation
Logistic Regression	0.9339	0.9350
Decision Tree Classifier	**0.9903**	**0.9893**
Random Forest Classifier	0.8957	0.8953
BernoulliNB	0.9576	0.9589
GaussianNB	**0.9968**	**0.9971**
Gradient Boosting Classifier	0.8673	0.8661

Fourth iteration:

- ✚ Adding state weapon count column.

Selected model

Final Shape:

Same number of rows (284,629), 77 features.

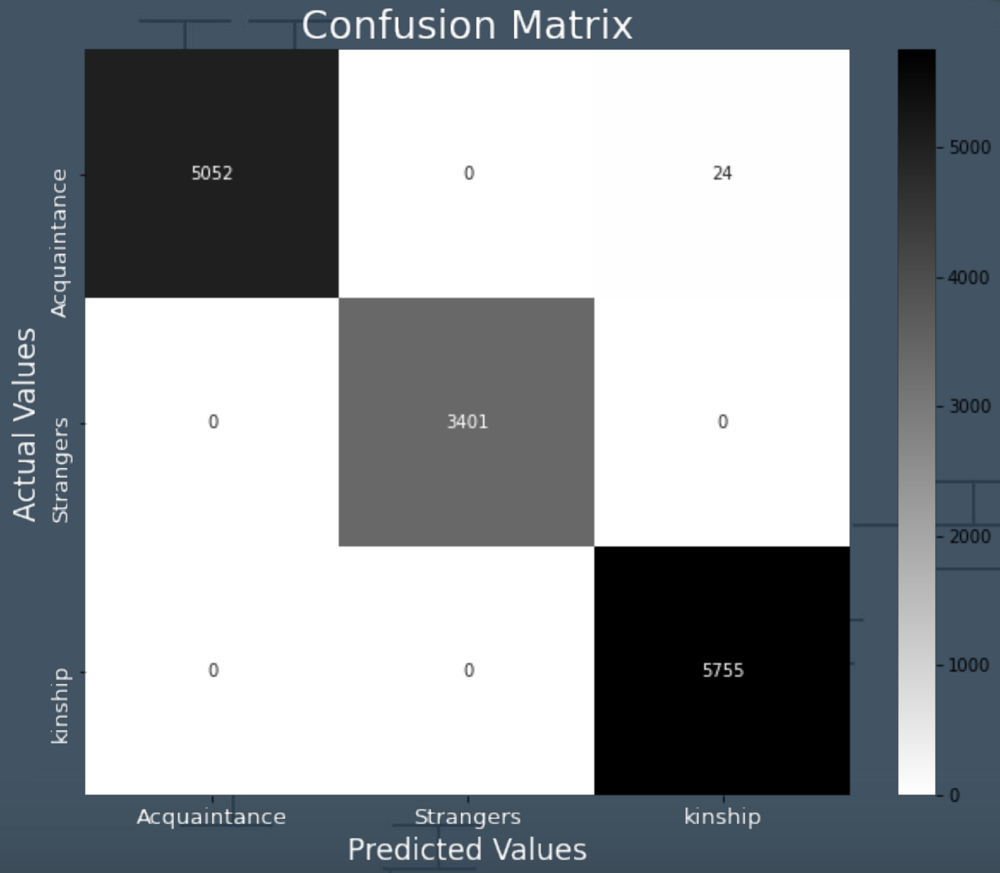
Retrain Model:

Train + Validation = Train set

Train and Test Score :

Model	Training	Testing	Error Rate
GaussianNB	0.9988	0.9983	0.002

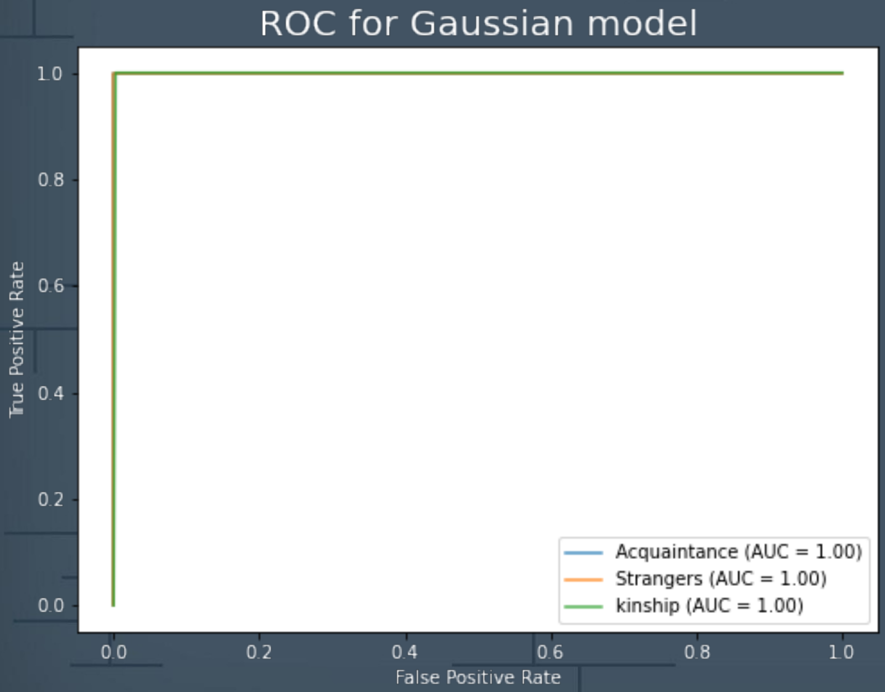
Confusion Matrix

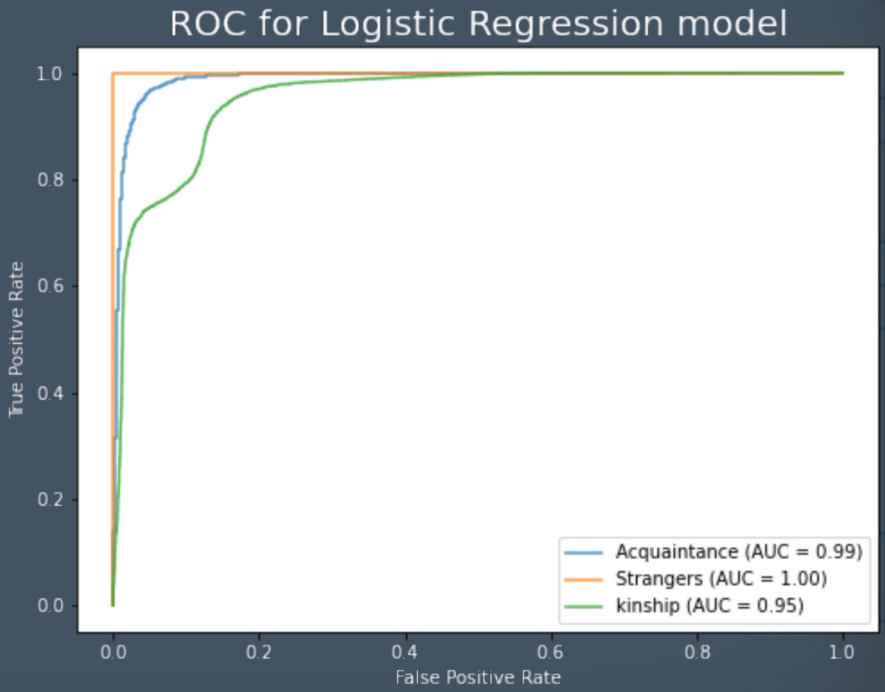


Confusion Matrix

Actual Values \ Predicted Values	Acquaintance	Strangers	kinship
Acquaintance	5052	0	24
Strangers	0	3401	0
kinship	0	0	5755

Roc


ROC for Gaussian model
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
Acquaintance (AUC = 1.00)
Strangers (AUC = 1.00)
kinship (AUC = 1.00)


ROC for Logistic Regression model
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
Acquaintance (AUC = 0.99)
Strangers (AUC = 1.00)
kinship (AUC = 0.95)

THANK YOU!