87

Michigan State University

Ethnicity Prediction of Civilians Killed by Police in the USA

Zachary Harless

A57241787

Final Report

1 Introduction

Surviving in the United States of America can mean entirely different things depending on where a person lives and what color skin they happen to be born with. It is this difference that plagues the nation with unnecessary violence resulting in tragic loss of life. In recent years there has been growing discourse pertaining to the role that police play in society and how often they choose to use lethal force on civilians to preserve law and order. This project aims to develop an ethnicity-prediction model based on several machine-learning classification techniques that will attempt to predict the ethnicity of civilians that have been slain by police in the year of 2015.

2 Related Work

There have been several reports written on the topic of disproportionate police violence towards minorities in the United States over the past few years, utilizing public datasets to determine underlying patterns and trends in the data. One such report written by Sinyangwe [1] used data from the 37 largest police jurisdictions in the US to showcase the disparities between how often white and black people were arrested as well as how often they were killed per 100 residents in each jurisdiction. The results showed that black people were consistently more likely to be both arrested and killed by police by as much as 22 times. Another study conducted by Eng and Wenig [2] found similar results by using SQL analytics to normalize the number of fatalities per each demographic, which ultimately found that Native American and Black people suffered from almost three times as many fatal police shootings as white people did. A similar calculation was performed on the dataset used for this report but the results were slightly different as the dataset Eng and Wenig used has entries from 2015 to 2020 but the one used for Figure 1 on the next page only contains data from 2015. They also found that the use of body cameras during these fatal incidents have decreased since 2016 (with only 14.9% of incidents having the camera on) which is extremely concerning to the public.

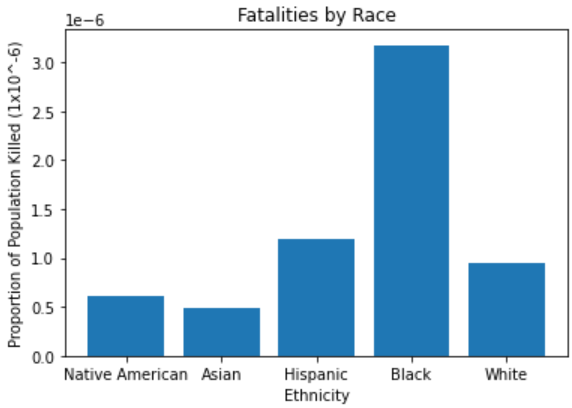

Figure 1: Victims/population proportion by ethnicity

3 Data

The data source used for this report is the public Kaggle FiveThirtyEight dataset on Police killings in the US in 2015 [3]. The dataset includes 466 entries with 34 parameters for every entry.

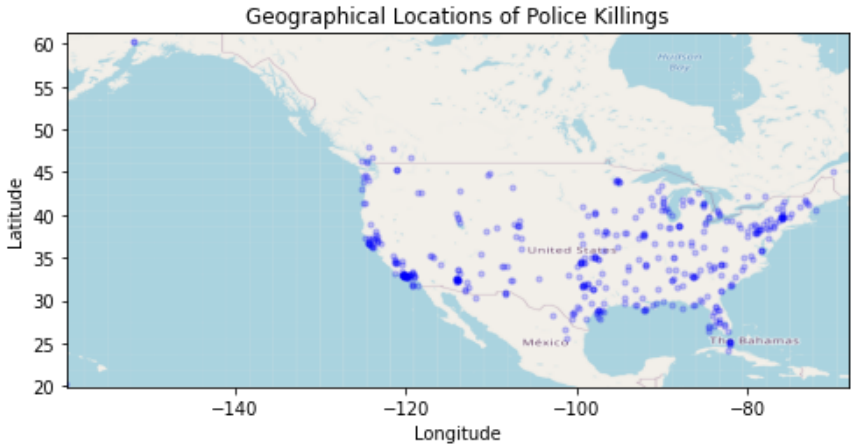

Figure 2: Locations of civilians killed in the US by the police

To initially process the data, first (i) all of the rows with missing values had to be dropped from the data frame, (ii) some parameters had to be converted into floats and integers, (iii) all rows with an "Unknown" age value were dropped from the dataframe, (iv) the gender parameter was converted into binary with "Male" as 0 and "Female" as 1, (v) the "armed" parameter was also converted into binary by establishing "Firearm" and a "Knife" values as a 0 and all other non-lethal weapons or unarmed values as a 1, and finally (vi) all of the parameters whose values were in the format of strings (such as the victim's name, the month they were killed, etc.) were removed. After the initial processing the data was cut down to 432 rows and 25 parameters as opposed to the initial 466 rows and 34 parameters.

3.1 Parameter Selection

After separating the data into training and testing sets, sklearn's [4] SelectKBest was utilized to calculate the best *k* features to be fed into the models. After several tests of multiple different *k* values, *k* = 5 appeared to result in models with the highest prediction accuracy

4 Model Design/Creation

The model was designed around predicting the binary value of a new custom parameter "iswhite" that was derived from converting the "raceethnicity" parameter (which had values of White, Black, Asian, Hispanic, and Native American) into the newly created 'iswhite' variable where "White" would equal a value of 0 and all other ethnicities would receive a value of 1. This parameter will become the data labels and the rest of the reformatted parameters will be the predictors to try and correctly predict whether the victim was white or not for of any given row in the dataset. Due to the binary nature of the parameter, classification methods were used to train models and predict that "iswhite" values more so than regression methods.

5 Methods

The parameters were tested using several different machine learning classification techniques including KNN, Random Forests, Naïve Bayes, and Support Vector Machines. All of these models were implemented using the scikit-learn library.

5.1 KNN

K-Nearest Neighbors is a clustering technique that works by assigning a value based on the average of the *n* neighbor values around it. In order to find the value *n* that resulted in the highest prediction accuracy, a loop was created to iterate through several different values of *n* and ultimately *n* = 2 was the most accurate.

5.2 Random Forest

Random Forests is a classification method that consists of many individual Decision Trees whose results are then averaged or taken the mode of. In order to form the most accurate Random Forest model the sklearn GridSearchSV function was utilized to determine the optimal parameters. A random forest model was then created using the newly found optimal hyperparamters max_feautures = 0.25, n_estimators = 250, and min_samples_split = 4

5.3 Naïve Bayes

The Bernoulli Naïve Bayes (also known as the condition probability) is more suitable for binary and Boolean features which is reflective of some of the key predictors in this dataset so it seemed to be a good model to implement. To make the model as accurate as possible the sklearn GridSearchSV function was used again and found the optimal parameters of: alpha= 1, binarize=0.0, and fit_prior=True

5.4 SVM

Support Vector Machines are a supervised learning method used for both classification and regression. Once again the sklearn GridSearchSV function was used to identify the best possible parameters for the most accurate model estimation and the parameters found were as follows: C=1, degree = 3, kernel = 'rbf', and gamma = 0.001.

6 Results/Conclusion

This project attempts to predict the ethnicity of a civilian killed by the police in the US in 2015 by utilizing features such as age, whether the victim was armed, and the racial makeup of the local area where the killings occurred. Four separate Classification models were used to help make these predictions after careful tuning of the hyperparameters and multiple trials. These models include K Nearest-Neighbor, Random Forest, Naïve Bayes (Bernoulli), and Support Vector Machines. After the models were trained and label predictions had been made, the sklearn accuracy scores were calculated and are listed in the table below. Random Forest ended up with the highest accuracy score of about 0.643 while Naïve Bayes performed the worst with an accuracy score of about 0.369. Although 0.643 isn't a very convincing level of accuracy, when considering the limited scope of the dataset and the lack of abundant data entries the value becomes more impressive.

The future work on this project could include (i) testing more classification models to see if any of them out-perform Random Forest, (ii) turning hyperparameters even further to result in higher accuracies, and (iii) implementing an additional dataset in order to train and test the models better.

	KNN	Random Forest	Naïve Bayes	SVM
Accuracy Score:	0.5	0.643	0.369	0.393

Figure 3: Table containing accuracy scores for each classification model

7 References

[1] Sinyangwe, Samuel. "The Police Departments With The Biggest Racial Disparities In Arrests And Killings." *FiveThirtyEight*, FiveThirtyEight, 4 Feb. 2021, fivethirtyeight.com/features/the-biden-administration-wants-to-address-racial-bias-in-policing-what-cities-should-it-investigate/.

[2] Eng, Chengyin, and Brooke Wenig. "Analysis of Police Fatal Shootings in the U.S." *Databricks*, Databricks, 16 Nov. 2020, databricks.com/blog/2020/11/16/fatal-force-exploring-police-shootings-with-sql-analytics.html.

[3] https://www.kaggle.com/fivethirtyeight/fivethirtyeight-police-killings-dataset, *public data source*

[4] https://scikit-learn.org/stable/, *python package used for machine learning models*