Using Machine Learning to Analyze Patterns in Serial Killers

Fernando Martinez
UTRGV
fernando.a.martinez01@utrgv.edu

Mason Garza
UTRGV
maria.garza01@utrgv.edu

***Abstract*—Through the use of machine learning, we apply random forest algorithms to create predictive models to analyze more than a thousand serial killers, and be able to predict motivation, potential number of victims, and the sex of serial killers. This is to help law enforcement with profiling serial killers that have yet to be caught, and help increase chances of capture.**

***Index Terms*—random forest, machine learning, serial killers, criminology**

I. Introduction

Criminology is a vital part of law enforcement and public safety in the modern age. Action and public policy made according to the findings of studies in this field have proven effective and popular. However, efforts to reduce and understand serial killings haven't received much of this windfall. Perhaps by using machine learning, we can close the gap.

II. Motivation and Goals

After watching ”The Ted Bundy Tapes”, one of us was infuriated by how such a deadly killer could get away with the terrible acts he did for as long as he did. The idea of using machine learning to profile serial killers and prevent extensive suffering was proposed then. Our main goal would be to create a model that could predict the actions or profile of a serial killer at large, so that law enforcement could take preemptive actions to prevent greater suffering. In addition to the practical application of profiling deadly persons to prevent murders, we also had a secondary goal of trying to gain a psychological understanding of serial killers, and perhaps see what makes them go off the deep end. In practical terms, our goal is to train a machine learning model that was able to accurately predict three specific features of a killer, their Motivation, their Number of Victims, and their Sex.

III. Background

A. Serial Killers

The FBI can broadly define a serial killing as ”The unlawful killing of two or more victims by the same offender(s), in separate events.”. Although disputes exist on the finer details, this remains a robust definition. However, potentially included in this definition are organized criminals, war criminals, and such other criminal whose actions are at the behest of an illegal organization. Although they are no less deadly, the scale of an illegal organization is much larger than the granular individual, which is what we are especially interested in. Therefore for our own purposes, we found it useful to add the following caveat, that a serial killing is also a crime that is done by those who are acting independently or alone. We allow for those who have accomplices.

B. Existing Literature

Although literature on serial killings exists, the application of machine learning, or any kind of data analysis, is rare or non-existent. The closest we found were articles detailing the use of regression algorithms to analyze criminal patterns in two North American regions, the state of Louisiana [1] and the city of Vancouver [2]. Most studies were not very recent, and the techniques used were not particularly useful to us. It was necessary to invent new methods, and focus on serial killers rather than general crime analysis.

IV. Dataset

A. Choosing Our Dataset

Part of the difficulty of conducting a data analysis on serial killers is their limited number. Public databases for serial killers are either nonexistent or too obscure for us to find. By necessity, we opted for a private resource for our data, the Radford/FGCU Serial Killer Database Project. A cross-university collaboration, it is a non-governmental private database of serial killer in the world. It contains over 3000 killers of a small variety of malevolent crimes, with about 2500 serial killers. Every killer may have a description of up to 170 features, such as their sex, birth year, method of killing, etc. Entries of the database have proof of their features and of their existence through the citation of public and private sources. This was a very robust database so we gained the use of it after an application process, and a lengthy wait time.

B. Important Limitations

The data from this database is private for a very important reason, ”All files and data should be considered private and sensitive; they may also have copyrights or laws pertaining to possession. Do NOT store any files on your personal computer, and clear your browser cache periodically.” [3] To respect this request, we modified the data we retrieved from the database to me anonymous, and condensed for our usage.

V. Preprocessing

A. Encoding

Although most features were described with a manageable and numeric value, some were described with a textual value that needed to be converted or deleted. For example, the value entered for motivation, which was a text entry from a consistent list of categories, for which there were 11 categories and a varying number of subcategories. To make the training easier, we limited the number of motivation categories to the 11 general categories, and gave an integer 1 - 11 for each category. Later on we would trim these down further to a binary feature, whether or not the motivation of the killer was Enjoyment. An example of a deleted entry was the state of the killing. Although we originally intended to one-hot encode this with 51 columns (representing all the U.S. states + D.C.). Subdivisions of other countries past and present also had to be represented. We decided to excise this feature as a result.

B. Empty Cells

The preliminary data set we had about 2500 killers with 173 features. However the data was extremely ”porous”, as described by one associate. Not a single killer had all the 173 features described, every single one had at least one empty cell. This is big problem, we needed a full data set to begin training our model.

A method was devised to cut out empty cells. Any features which were described in less than 80% of the serial killers were cut out of the working data set. After this culling process, every killer with an empty cell was removed from the data set. This gave us a working data set of the following dimensions: 1100 killers with 60 features each. This process was done with a python script

VI. Methods

A. Environment

For this project we chosen to use MATLAB for the training and testing of our models. It has mature machine learning libraries and tools that can be used with great ease and effectiveness. We never ran into a situation where MATLAB lacked a functionality that we needed, and the actual run time of the code was relatively short.

B. Choosing an Algorithm

We needed to choose two types of algorithms, one that could do category predictions (for Motive and Sex), and one that can do regression predictions(for Number of Victims). We tried as many as possible, including but not limited to:

- Linear and Logistic Regression
- Support Vector Machines
- Decision Trees
- Naive Bayes
- Feed Forward Neural Networks (FFNN)
- Random Forests

We normalized the data when necessary. In all methods we used 10-fold cross validation, Except for FFNN, in which we used a 15% holdout.

In our preliminary testing, two methods were giving the most promising results, FFNN, and Random Forests. The accuracy of both were comparable. The accuracy of FFNNs for classifying sex was 91.4%, while the accuracy of Random Forest was 92.1%. We ultimately decided to use Random Forests for our study. This is because the interpretability of Random Forests model is more feasible than the that of a Neural Network, which is something we prioritized from the start.

C. Random Forests

Random Forests are what some may call an *Ensemble Method*. The general idea of ensemble methods is to train many ”weak learners” in order create a ”strong learner”. In the case of Random Forest, the weak learners are decision trees.

Random Forests also employ *Bagging*, Bootstrap Aggregation. When a new tree is trained, it is trained using a uniformly random sample of features from the original dataset with replacement. Sampling with replacement allows for a sample of training features to be repeated. This is ”Bootstrapping”. Then, once all the trees are trained, they can be fed the testing data. Every tree will give a different result for the test data, but the average (for regression) or majority (for classification) result will be the one given by the forest. This is ”Aggregation”.

In addition, when trees are trained in a Random Forest, the features chosen to train the tree are also chosen in a uniformly random way with replacement.

Random Forests have many attractive qualities. Because Random Forests are essentially a collection of Decision Trees, they don’t need to be trained with normalized data, and can handle both continuous and categorical data simultaneously. Much like decision trees, they have white box qualities, as you can look at the trees that produce the results. One can also calculate the importance of variables by averaging the error difference as the values of a variable are permuted across all the trees. Bagging also makes Random Forests resilient against variance and overfitting. It’s likely that those are the reasons why it performed so well with our data.

VII. Training Data, Models, and Testing

Two Classification Random Forests were trained for Motivation. One is an 11 Class model where tried to classify for all 11 categories: Financial Gain, Attention, Enjoyment, Anger, Mental Illness, Cult, Avoid Arrest, Organized Crime, Convenience, Wildwest Outlaws, and Mutiple Motives. Of these original 11, Enjoyment is the most common, followed by Financial Gain. In the second model we opted for a Binary Motive classifier, by collapsing all the categories into just two: Enjoyment and Other.

One Regression Random Forest was trained for predicting the number of victims. We excised a number of features from the data set which we considered trivially correlated to the number of victims, such as the number of suspected victims, the number of male victims, etc. The average number of victims in our training set was 5, with a standard deviation of 5.1.

Finally, a Classification Random Forest was trained for predicting the sex of the killer. Much like the victim model, certain features were removed from the the working data set, such as the White Male feature. The vast majority of the killers were male.

All of these models were 10-fold cross-validated. The error measure for the classification forests was the misclassification rate, and the error measure for the regression forest was the mean squared error.

VIII. Results

The models for motivation had fair accuracies. The 11 Class model could correctly classify a 64.6% of the test data, while the Binary Motive classifier had an accuracy of 81.6% (Fig. 1). We produced confusion matrices of both results (Fig. 2, 3).

The model for victims ended with a root mean squared error of 4.8 victims (Fig. 4).

The model for predicting sex was 92.1% accurate. A confusion matrix was also produced for this (Fig. 5)

We found the accuracies for the models that predicted Binary Motive, number of victims, and the sex of the victims appreciable and worthy of further examination.

IX. Model Analysis

It was previously mentioned that a method exists that one can use to measure the importance of a variable in Random Forest model. We employed this method on the three models we felt were worth examining and were able to get graphs showing the importance of every feature to the models. We also looked at a small subset of the trees in each model to try and gain further understanding of the results.

A. Binary Motive

In the Binary Motive model (Fig 6), one variable stands out in importance from the rest, Var 24. Var 24 corresponds with the feature that states whether or not the killer had raped. We found this result to be shocking, yet intuitive, as sexual element is often present in the Enjoyment motive. Also significant, and possibly related to the aforementioned rape, is the number of male victims and the number of female victims.

B. Victims

In the victim model (Fig 7), the most important variables are related to time. The year of the first kill, and the killer's birth year. This is potentially related to the disparity in kill counts between certain periods of time, which the model picked up on. Also important is whether or not the killer took what is called "a possession trophy". A possession trophy is a personal possession of the victim (not a body part) that is taken by the killer as a commemoration of their kill. We have no theory for why this may be significant.

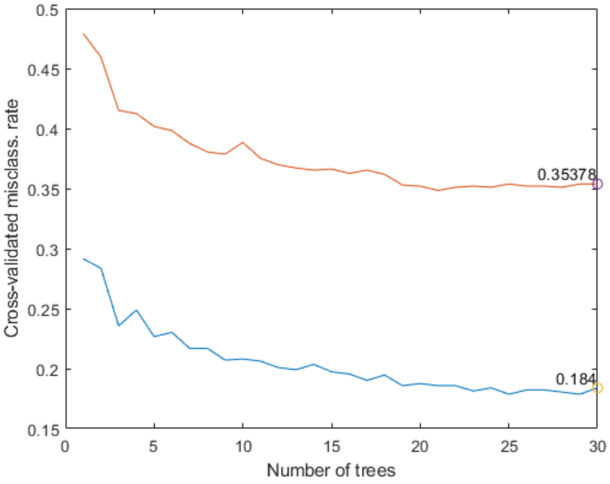


Fig. 1. Error plot of the motive models, red is all 11, blue is Binary

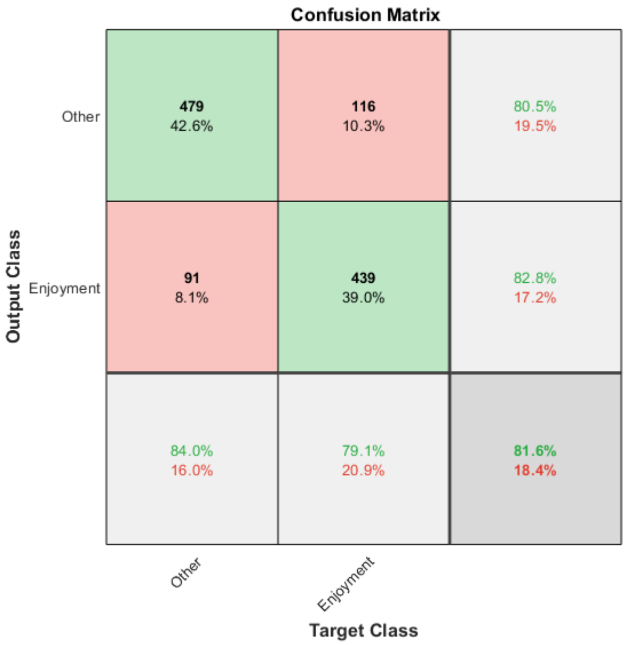
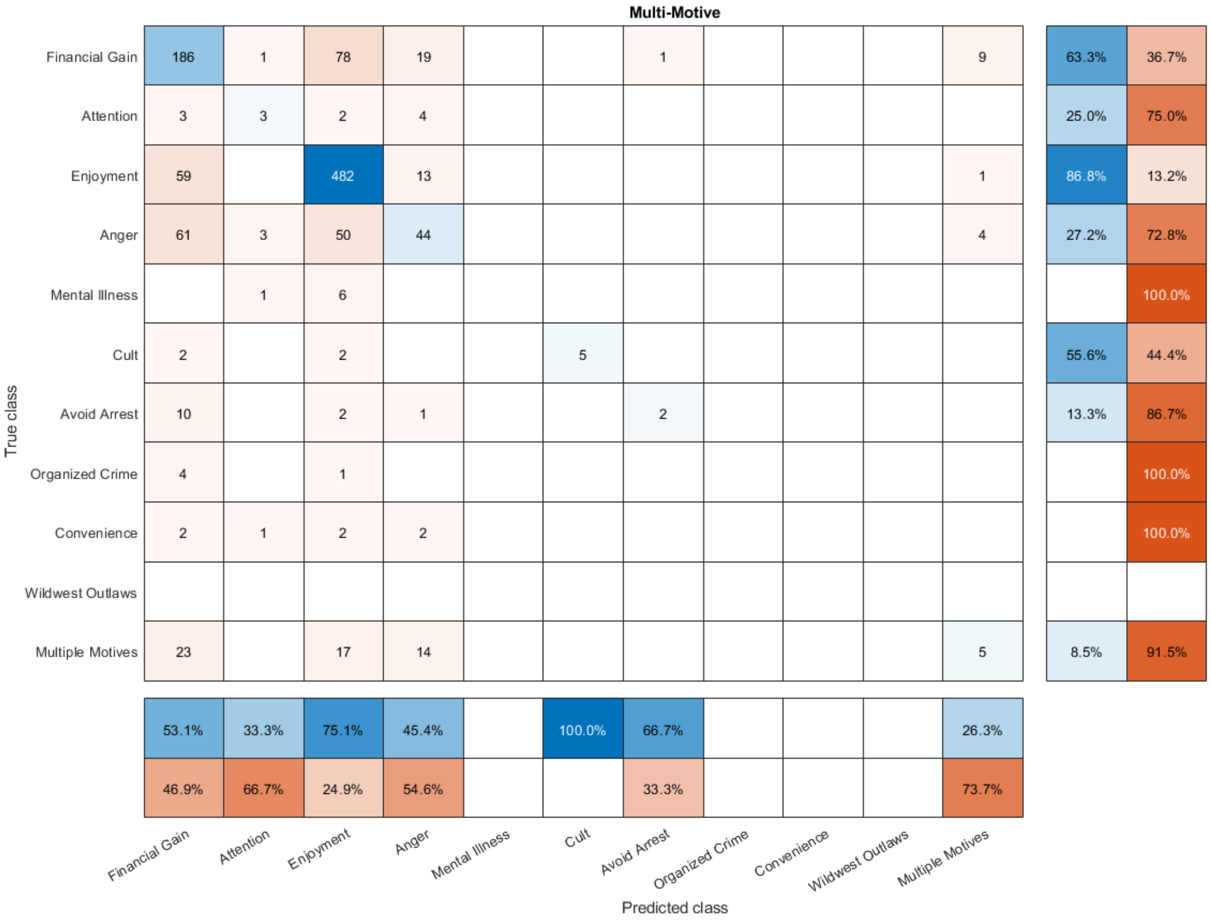


Fig. 2. Binary motive confusion matrix

C. Sex

In the sex model (Fig 8), the most important variable is also related to time, the birth year of the killer. However, another notable pattern is that many variables share a large significance in this model, such as the killer's motivation and whether or not the victims were of adult age.



	Financial Gain	Attention	Enjoyment	Anger	Mental Illness	Cult	Avoid Arrest	Organized Crime	Convenience	Wildwest Outlaws	Multiple Motives		
Financial Gain	186	1	78	19			1				9	63.3%	36.7%
Attention	3	3	2	4								25.0%	75.0%
Enjoyment	59		482	13							1	86.8%	13.2%
Anger	61	3	50	44							4	27.2%	72.8%
Mental Illness		1	6										100.0%
Cult	2		2			5						55.6%	44.4%
Avoid Arrest	10		2	1			2					13.3%	86.7%
Organized Crime	4		1										100.0%
Convenience	2	1	2	2									100.0%
Wildwest Outlaws													
Multiple Motives	23		17	14							5	8.5%	91.5%
	53.1%	33.3%	75.1%	45.4%		100.0%	66.7%				26.3%		
	46.9%	66.7%	24.9%	54.6%			33.3%				73.7%		

Fig. 3. Confusion matrix for all motives.

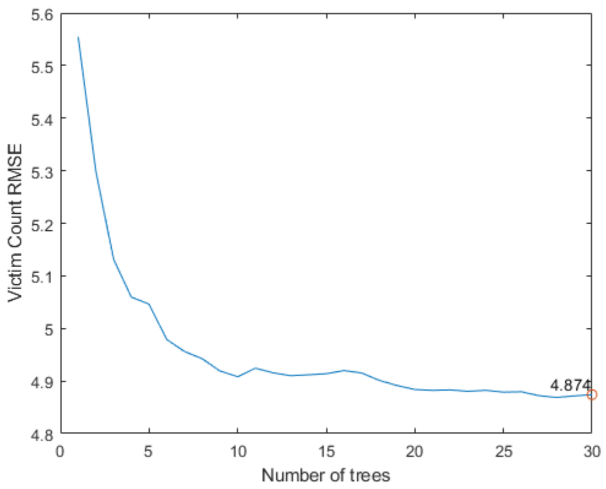
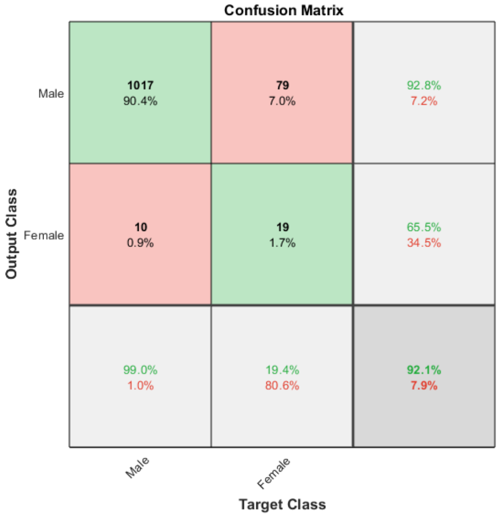


Fig. 4. Error plot for victims



	Male	Female	
Male	1017 90.4%	79 7.0%	92.8% 7.2%
Female	10 0.9%	19 1.7%	65.5% 34.5%
	99.0% 1.0%	19.4% 80.6%	92.1% 7.9%

Fig. 5. Confusion Matrix for the Sex model.

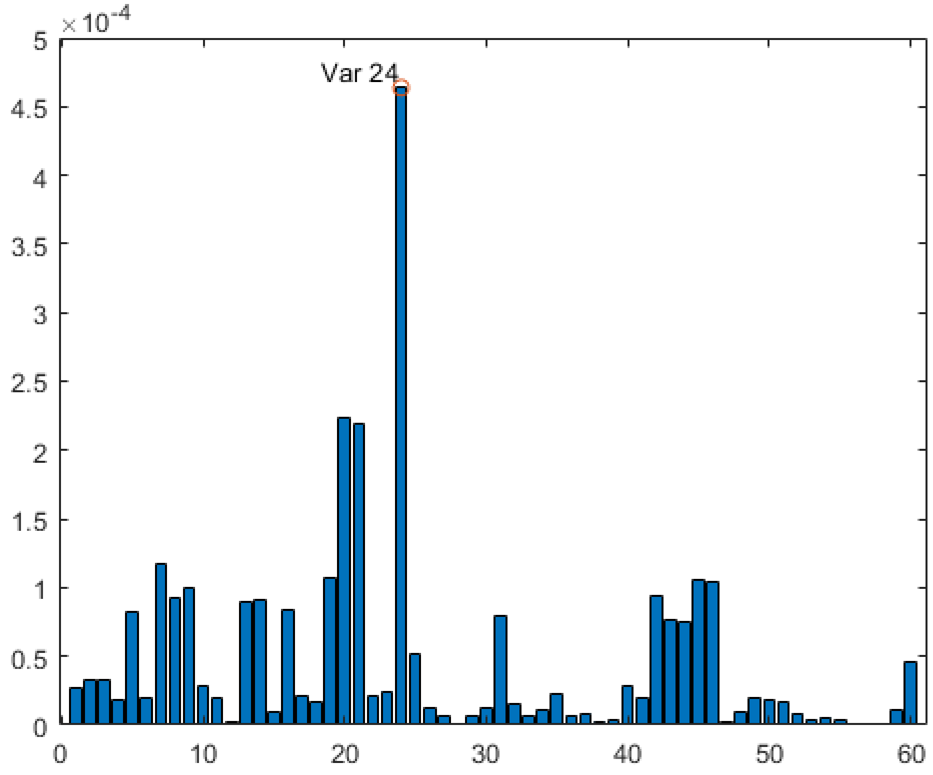


Fig. 6. Predictor Importance for the binary motive model.

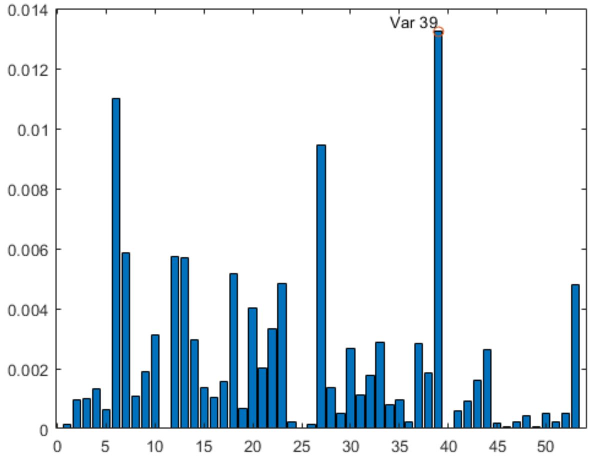


Fig. 7. Predictor Importance for the victim model

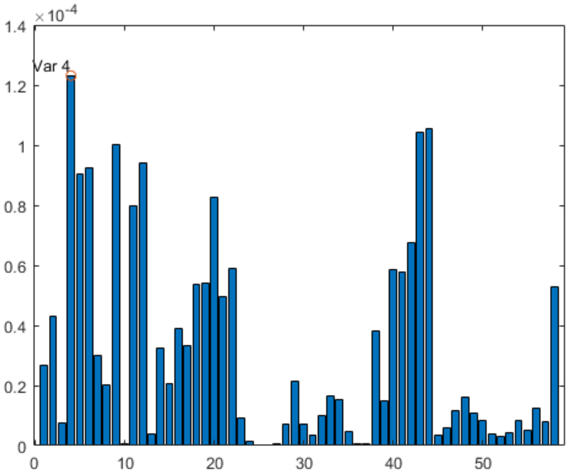


Fig. 8. Predictor Importance for the sex model

X. Difficulties and Setbacks

In trying to create these models, we made a few mistakes which will be addressed now.

Originally, our models had a much higher accuracy than they do now. Not until we began our predictor importance testing, did we realize how they became as accurate as they were. In the victim model, we originally purported that we had an LSBoost model with a root mean squared error of about 1.7. When the model which produced that accuracy was analyzed, we found that one variable was extremely dominant, the total suspected number of victims over the killer's lifetime. This was not a fair model, and so we built a new model from scratch without that feature, which used bagging instead. Something similar happened with our Sex model, we originally had a model that was a simple decision tree and managed an accuracy of about 98%. When the tree's structure was observed, we found that it was relying heavily on the White Male feature to classify male killers. This was also observed to be unfair, and so a new model was chosen from scratch. Such a thing was never observed in the motive model fortunately.

XI. Future Work

We consider these to be maiden steps. If we were to pursue this study even further, there are a handful of paths we can take. We want to explore methods of data imputation so that the data set we can work with can be larger. There are other features we can train for, such as the method of kill, the presence of mental illness, the type of victim the killer prefers etc. One suggestion was to incorporate the physical appearance of the killers into the data, since pictures are available in the database. More detailed analysis of the models can still be conducted by observing the totality of the trees, instead of a small subset of them as we did.

XII. Conclusion

Data analysis on a small data set is an inherent challenge. In spite of that, decent results were attained in this project. The models for motivation and sex reported results of 81% accuracy and 92% accuracy respectively. The Number of Victims model ended with an RSME of 4.8. The power of Random Forests were at display in this project, probably due to factors involving their flexibility and resilience in the face of data sets with high variance. Hopefully this shows that machine learning is possible and practical to use on a serial killer data set for the purpose of profiling, criminology and psychology.

References

[1] McClendon, Lawrence, and Natarajan Meghanathan. ”Using machine learning algorithms to analyze crime data.” Machine Learning and Applications: An International Journal (MLAIJ) 2.1 (2015): 1-12.

[2] Kim, Suhong, et al. ”Crime Analysis Through Machine Learning.” 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). IEEE, 2018.

[3] Radford/FGCU Database