Prediction of Solved Homicides Using a Classification Method

Lamija Zukic*, Samed Jukic*

*International Burch University
lamija.zukic@stu.edu.ibu.ba
samed.jukic@ibu.edu.ba

Original research

Abstract: *Homicide rates are still high in the world and they are the worst crime in human existence. Despite all the technological advances and usage of information by various agencies, the number of homicides is not decreasing. Homicide prediction in certain countries should notably be the number one priority, which can help the government to easily identify the kind of profile they are looking for, or even help them prevent those cases. This paper compares different Machine Learning Techniques classifications of homicide prediction. Random Forest (RF), Random Tree, J48, Naive Bayes and k-Nearest-Neighbor (KNN) were tested to determine which method provides the best results in homicide prediction classification. The results of sample accuracy for all algorithms were around 99%, which clearly shows that all algorithms give great results. However, J48 is the best technique applied on the dataset, as it classified all instances correctly.*

Keywords: Classification, data analysis, homicide, machine learning, prediction.

1. Introduction

Homicide continues to prevail in most news. Unfortunately, every evening news ends up broadcasting at least one homicide within the country's region. This has been an issue and still is in most countries, but especially in the United States of America (USA). It is widely known that USA government agencies have Crime Departments that are advanced in the fields of data gathering and data processing for various needs. However, it seems that the homicide rate in the USA is still high, and it is a big problem as homicide is the most violent form of crime. Term homicide clearance rate refers to the percent of homicides that lead to arrest or charge, meaning that the case can be declared as solved.

On the other hand, uncleared homicide means that the suspect is not found due to the lack of data gathered by the police. This can lead to a serious problem, as the perpetrator is freely moving, and can result in another homicide as revenge. Homicide clearance rates currently range up to 65% in the USA [1], 95% in Japan [1] and 75% in Canada [2]. With evolving technologies, nowadays many tools are used for analytics and predictions in crimes. Having a huge amount of raw and meaningful data creates many opportunities for getting the desired output that can help the government make accurate decisions. In the case of homicide, that should be the number one priority, crime agencies should use that data to make predictions to prevent homicide crime. Various machines are applied for different reasons and different outputs.

This research paper uses a USA homicide dataset to predict homicide clearance, in terms of whether the crime is solved or not. Before ML models are applied, data is preprocessed and cleaned. Different machine learning methods are applied to the data. Random Forest (RF), Random Tree, C4.5, Naive Bayes and k-Nearest-Neighbor (KNN) were compared based on different performance evaluation criteria.

The organization of the paper is as follows. Section 2 presents literature background work on the prediction of homicide, whereas Section 3 describes the Homicide dataset and ML techniques applied. Section 4 presents the experimental results. Finally, Section 5 concludes the paper.

2. Literature Review

Few research papers analysis of homicide clearance data in certain countries and the factors that lead to the decision whether the crime is solved or not. A paper *The Value of Life in Deaths provide* used multiple regression to determine there were extralegal factors, particularly racial and gender, that affected clearance rates. The results showed that homicide clearance varied by several extralegal factors. For instance, cases involving non-white victims or older victims were less likely to be solved, thus white victim homicides had a 42 percent greater chance of being solved than non-white cases. Also, cases that involved younger children were more likely to be solved than cases with older victims. Multiple regression used in this paper resulted in binary outcomes that predicted the chance of homicide clearance, and it determined which homicides were most likely to be cleared. The model proved to be very successful [3].

Another paper used a sophisticated statistical approach multilevel latent class analysis [MLCA], which concluded that more stranger homicides were not solved, which might have been due to missing data on the victim/perpetrator relationship. Also, robberies or similar felonies that resulted in homicides may be mistakenly characterized as stranger killings [4]. McClendon and Meghanathan implemented three algorithms, Linear Regression, Additive Regression, and Decision Stump (DS) using the same finite set of features on crime unnormalized dataset to identify violent crime patterns from two datasets [5]. Regression proved to be the best method for predicting the crime data based on the training set input. The algorithm that had the greatest

correlation coefficient value and generated the lowest error values is the linear regression algorithm. The DS algorithm proved to be the least accurate one.

Another paper performed ML algorithms in crime prediction in Canada. Kim and Joshi used classification methods to identify patterns and create predictions. K-nearest neighbor (KNN) and Decision Tree algorithms were implemented to analyze the crime dataset. Their dataset consisted of more than 500,000 records with an accuracy between 39% and 44%. The accuracy turned out to be below as a prediction model, however, the accuracy can be increased by tuning the algorithms and crime data for specific applications [6]. Rolf Loeber & Lia Ahonen examined the Pittsburgh Youth Study, which began in 1987. The study was conducted among boys from public schools who were randomly selected from 1, 4, and 7 grades, respectively. To identify the number of high-risk males, they used a so-called screening assessment, which consists of collecting information from participants, their parents, and teachers. For the next assessment, 30% of the most antisocial boys were chosen, with 30% of boys randomly selected from the remaining 70%. In the following assessment, selected boys had carried out face-to-face conversations at half-yearly intervals throughout approximately, the next 10 years [10]. According to Rolf Loeber & Lia Ahonen, the study is uniquely created to investigate individuals' delinquency and substance use, as well as their continuation and renouncement from such behavior. The information collected up to 2009, 37 males were convicted for homicide between the ages 15 and 29, while in total 39 males were victims of homicide. The average age of perpetual is 19.7, while the average age of the victim is 22.7. An interesting data gathered from the study is that 32 out of 37 convicted homicide offenders were African American, and 37 out of 39 are homicide victims. Another important piece of data pointed out in this study is the usage of guns as the main weapons. To sum up, these homicides were mostly carried out by African American males that involved guns, drugs, and gangs. There were few anomalies, but such were excluded from the result analysis. In this study, the authors explain how three different types of predictors need to be considered: factors in family and neighborhood environment, early behavioral factors such as conduct problems, and early offenses such as self-report and arrest. The study also identifies the prediction of homicide victims, using regression analysis. Same characteristics and predictors are shared between victims and offenders [10].

This study worked on a detailed and thorough analysis of a small and selected number of participants, which involved yearly face-to-face conversations throughout the years, where people were able to follow their life and their behavior on a certain basis. Predictions in terms of predictors were identified, as to what factors in one's life can lead up to homicide and victim.

3. Dataset and Machine Learning Techniques

3.1. Dataset

The homicide database in the United States, compiled and gathered by the Murder Accountability Project is made available to the world. The dataset includes murders from the FBI's Supplementary Homicide Report from 1980 to 2014, rounding up to almost 35 years, on more than 600,000 murders. This research paper uses homicide records from 2010 to 2014 as those are the most recent data available to the public. The dataset contains 20 relevant features, such as information about homicide time and place, victim and perpetrator details, the relationship between the two, and the weapon used. Dataset has been preprocessed to clear noisy data to provide more meaningful results. Python language has been used for data cleaning. Crime in the US has been a big topic for decades between the crime agencies and politics. As each country is fighting to drop its crime rate, so is the US. Having an accurate database of such information can be helpful in the future once applying a proper ML model thus analyzing the results.

3.2. Data Preprocessing

Python language continues to be a very popular and well-maintained open-source programming language in Data Science. Before ML models were performed, the dataset went through data preprocessing in Python. A new data frame has been created and later used for machine learning. Some features were not relevant, so they were excluded from the dataset. Due to the very large dataset only records for the years 2010 through 2014 were taken into account, as they are the most recent public data. Records where the feature Relationship is Unknown and where MurderSolved has value Yes, were removed from the dataset as well. These records had most of the features classified as unknown, and such data would not contribute to the research. However, this significantly reduced time and CPU memory for the model to build. Also, certain features, such as victim age and perpeterur age had few values identified as outliers and they made no sense. This problem would impact the analysis, so methods such as aggregation and mean replacement were used to overcome this problem.

3.3. Machine Learning Techniques

Classification is one of the most common applications in Machine Learning. It identifies and discovers patterns, and later on sort the data into groups based on its similarities. Classification methods build a model that predicts future outcomes using predefined classes which are based on certain criteria [12].

3.4 J48

J48 is an open-source Java implementation of the C4.5 decision tree algorithm. J48 as a decision tree classifier has additional features for missing data, continuous attribute value ranges, pruning of decision trees, rule derivation, etc. J48 uses a predictive ML model that calculates the final value from the dataset. Its structure consists of root nodes, intermediate nodes, and leaf nodes, where each node consists of a decision, leading to the final result. To calculate which attribute is the best option for splitting the tree, we use the splitting criterion [7].

3.5 Random Forest (RF)

Random Forest is a method that operates by creating multiple decision trees during its first phase or training phase. The decision of the majority of the trees is chosen by the RF as its final result. The first benefit of RF is the reduction of overfitting, as we use multiple trees, meaning that the data is fit so close. RF runs efficiently on large data, thus it produces highly accurate predictions. Also, it estimates missing data, thus maintaining accuracy when a large proportion of data is missing [8].

3.6 Naive Bayes Classifier

Naive Bayes is another classification algorithm that assumes the input values are nominal, even though numerical inputs are also supported once applied. Naive Bayes uses an implementation of the Bayes Theorem which is based on the principles of conditional probability, as each class is obtained from the training set and is adopted as independent variables. It later predicts the class with the highest probability. This algorithm has proved to be very effective (fast and easy to calculate), despite the impractical assumption where the variables are expected to be independent [7].

3.7 K-Nearest-Neighbor (KNN)

KNN is also called instance-based learning. KNN algorithm supports both classification and regression methods. It is a simple algorithm that locates or classifies a new instance closest or the most similar to the training patterns. To make a prediction, it takes the mode (most

common class) of the training pattern to find the k, most similar instance. KNN method produces a linear decision boundary [7].

4. Results and Discussion

This section presents all of the results from the implementations of the J48, Random Forest, Naive Bayes, and KNN algorithms. The algorithms were run to predict the feature CrimeSolved. The algorithm that gives the lowest error values for prediction and its accuracy is highlighted in the results presented in Tables 1 and 2.

TABLE 1. Algorithm Results.

	Sample Accuracy	**ROC Area**	**F-Measure**
J48	100%	1	1
RF	99.9%	1	1
Naive Bayes	99.9%	1	1
KNN	99.8%	0.99	0.91

TABLE 2. Confusion Matrix.

Algorithm	**Confusion Matrix**	
	Correct	Incorrect
J48	34914	0
Random Forest	34904	10
Naive Bayes	34904	10
KNN	34878	36

For all these algorithms, cross-validation of 10 folds was employed as a parameter before the model building. Cross-validation is a standard evaluation technique, which systematically runs a repeated percentage split. As a result, it gives 10 evaluation results, which are later averaged. The sole purpose of this is to give better results since the dataset is relatively large. The total number of instances was 34914. The percentage of correctly classified instances is usually called sample accuracy. All four classifiers have been implemented and tested in software package WEKA using default parameters.

Looking at the table, the overall accuracy of the algorithms provides very similar and r. The algorithm that had the greatest sample accuracy was J48 (100%) among other three algorithms, with a decision tree as an output. KNN algorithm was the least accurate model (99.8%) with the

most incorrectly classified instances (36). The model with zero incorrectly classified instances was J48.
Another preferred measure is F-Measure. F-Measure is a combined measure for precision and recall calculated as follows:

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

The ROC area for all four algorithms is high, meaning that if we were given an item from both classes, this would be a percentage of randomly putting them correctly. Given the classification task for this particular dataset and the features that have been provided, J48 and Random Forest algorithm are the most accurate of the four.
Thomas Hargrove, a founder of a nonprofit organization for crimes, developed an algorithm for detecting serial killers and their activities. The algorithm is based on clustering methods where murders of women within a close area and similar weapons are used. The algorithm organizes homicide into clusters based on victims' data, such as gender, geographic location, and weapon used, and outputs murder groups with low homicide clearance rates [9]. According to the source, the algorithm proved to be successful in catching serial killers, one in particular.

Y. Rayhan and T. Hashem developed a prediction model using Spatial-temporal systems, which describes a phenomenon in a certain location and time. The model is called AIST a.k.a. Attention Based Interpretable Spatio Temporal Network for crime prediction, and it uses historical data (past crimes), external features (e.g., traffic, point of interest), and recurring trends of crime. The model proved to be very accurate and reliable using real data. The paper compared this method to other methods such as decision tree and RNN, where the AIST method outperformed most algorithms, looking at the evaluation criteria [11].

5. Conclusion

We observe all four algorithms to be very effective and accurate in predicting the homicide clearance data based on the training set input. Looking at the confusion matrix, we can have a deeper insight into the accuracy of models. The J48 algorithm classified correctly all instances in the dataset. The reason for such accuracy might certainly be due to the data preprocessing. Certain data was transformed to bring it to a state where the algorithm can easily parse it, meaning that the data can be easily interpreted by the algorithm. Data preprocessing is a crucial part, as it directly impacts the accuracy rate.

After an analysis of all the homicides that happened between 2010 and 2014, it is concluded that most homicides are solved while a small percentage of homicides are unsolved. As the idea is to find patterns between homicides that are unsolved, the results showed interesting information. Algorithms performed provided high accuracy in classifying unsolved homicides, and the following text will explain what attributes contribute to these results. The homicide rate is slightly decreasing. This is due to evolving technologies in the world, where the USA is investing huge amounts of money, especially in government and security facilities. Tools such as CCTVs and systems can help agencies solve crimes in much easier and more convenient ways.
Predicting homicide clearance can be a crucial contribution when solving a case. Knowing a potential outcome, whether a case is solved or not makes solving a crime case easier and is a helpful tool for the government, police, and investigators.

6. References

[1] Roberts, A., (2008). *Explaining Differences in Homicide Clearance Rates Between Japan and the United States*.

[2] Mahony, TH, Turner, J (2012) *Police reported clearance rates in Canada, 2010*. Juristat article. Cat no 85-002-X. Ottawa: Statistics Canada.

[3] Lee, C. (2005). *The value of life in death: Multiple regression and event history analyses of homicide clearance in Los Angeles County*. Journal of Criminal Justice 33.

[4] Flewelling, R. L., & Williams, K. R. (1999). *Categorizing homicides: The use of disaggregated data in homicide research*. In M. D. Smith & M. A. Zahn (Eds.), Homicide: A sourcebook of social research (pp. 96 – 106).

[5] McClendon, L. and Meghanathan, N., (2015). *Using Machine Learning Algorithms to Analyze Crime Data*. ResearchGate OI: 10.5121/mlaij.2015.2101

[6] Kim, S., Joshi, P., Kalsi, P. and Taheri, P., (2018). *Crime Analysis Through Machine Learning*. 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCONalog).

[7] Brownlee, J., (2016). *How To Use Classification Machine Learning Algorithms in Weka*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/use-classification-machine-learning-algorithms-weka/>

[8] Simplilearn. 2018. Random Forest Algorithm - Random Forest Explained | Random Forest in Machine Learning | Simplilearn. [online] Available at: <https://www.youtube.com/watch?v=eM4uJ6XGnSM>

[9] Kolker, R., (2017). Bloomberg - *Serial killers should fear this algorithm*.[online] Bloomberg.com. Available at: https://www.bloomberg.com/news/features/2017-02-08/serial-killers-should-fear-this-algorithm

[10] Rolf Loeber & Lia Ahonen (2013). Journal of Youth and Adolescence - *Invited Address: Street Killings: Prediction of Homicide Offenders and Their Victims*.

[11] Y. Rayhan and T. Hashem (2020). arXiv.org - AIST: An Interpretable Attention-based Deep learning Model for Crime Prediction

[12] J. Brownlee (2020). https://machinelearningmastery.com/ - 4 Types of Classification Tasks in Machine Learning