JOURNAL OF
EMERGING INVESTIGATORS

Article

Implementing machine learning algorithms on criminal databases to develop a criminal activity index

Krishiv Aggarwal[1,2], Akash Iyer[1,3], Gautam Nair[1,4], Risab Sankar[1,5], Ananya Balaji[1,6], Bharat Kathi[1,7], Samart Boranna[1,4], Vineet Burugu[1,8], Larry N. McMahan[1]

[1] Aspiring Scholars Directed Research Program, Fremont, California
[2] Irvington High School, Fremont, California
[3] Mission San Jose High School, Fremont, California
[4] Branham High School, San Jose, California
[5] Dougherty Valley High School, San Ramon, California
[6] West Windsor-Plainsboro High School North, Plainsboro, New Jersey
[7] Valley Christian High School, San Jose, California
[8] Westwood High School, Westwood, California

SUMMARY

Criminal activity is a major concern in today's society. Local police agencies collect and analyze vast amounts of data to prevent future crimes from taking place and protect vulnerable populations. However, despite the availability of publicly accessible data, such as the Open Justice California website, there remains a lack of efficient methods for relaying this information to the public in a digestible format. Our research aims to bridge this gap by utilizing machine learning techniques to correlate crime data with a range of explanatory factors. We first employed a clustering algorithm to normalize the data based on population. Then we tested five different predictive algorithms to determine the most effective machine learning model. Our results indicated that a neural network approach was more accurate based on our training in predicting crime rates. Additionally, we hypothesized that higher median income, lower population density, lower unemployment duration, and lower median age would be associated with lower crime rates and that these associations would be statistically significant. Our results show that median income, population, and unemployment duration all have a significant correlation with crime rates in California while median age does not.

INTRODUCTION

Tracking criminal activity and crime rates is an issue that is pervasive in the United States, whether it be in the most urban or rural parts of the country. As a result, the value of determining a solution to mitigate and control criminal activity is of utmost importance. When analyzing crime data, looking into demographic, geographic, and other conditions can factor into an area's crime rate (1,2). Other research papers have already used machine learning models to predict crime rates in specific counties or states (1,2). One concluded that a Light gradient boosting machine (LGBM - a form of deep neural networks) approach is best while the other concluded a long short-term memory (LSTM - a form of recurrent neural network) approach is best (1,2).

Machine learning is a type of artificial intelligence that involves training a computer system on a large dataset such that it can make predictions or decisions without being explicitly programmed to do so – for example, auto-correct, search suggestions, and what ads you see. Machine learning can be used to predict crime by analyzing patterns and trends in historical crime data (3). This can help law enforcement agencies identify areas where crime is more likely to occur and deploy resources accordingly (3). However, there are several potential pitfalls to using machine learning for crime prediction. One is that the predictions may be based on biased data, which can lead to discrimination against certain groups of people. These biases can be deeply rooted, originating from social stereotypes. These stereotypes can be both positive and negative, but they tend to be more negative and often lead to stigmatization, discrimination, and unequal treatment of individuals or groups (4, 5). Additionally, the predictions may not take into account important contextual or social factors that influence crime, leading to inaccurate or unfair results (6).

Criminal activity functions as a nondeterministic polynomial time hardness problem (NP-hard problem) due to several factors. NP-hard problems are complex computational problems that require a significant amount of time and computational resources to solve (7). There are four main reasons for this: complexity, nonlinearity, uncertainty, and dynamic nature (7). Criminal activity is complex and involves a wide range of factors such as individual motivations, social, and economic factors, cultural and demographic factors, and legal and law enforcement factors (7). Next, criminal activity is a nonlinear problem, meaning that small changes in one factor can have significant impacts on the overall behavior of the system (7). In addition, criminal activity is inherently uncertain, making it difficult to accurately predict and understand (7). It is influenced by a range of internal and external factors that are often difficult to quantify and measure. Lastly, criminal activity is dynamic and constantly evolving. Criminals adapt to changing conditions and law enforcement strategies, making it challenging to predict and prevent criminal behavior (7).

A previous study identified crime patterns with K-means

clustering to predict the time of a crime (8). K-means clustering is unique in its ability to "produce" optimal solutions. In this study, they determined significant attributes for the clusters, later learning that to get the quality of input required, they need skilled professionals to map the data (8). Although this method is difficult, it was required as certain crimes have factors that affect them more (8). On the other hand, some papers have focused on the probability of past crimes repeating themselves. Lum et al. attempted to improve the efficiency of risk assessment instruments but their fairness has been questioned due to a lack of data (9). They found that there was a significant amount of uncertainty surrounding an individual despite how much data they had, and the probability of the individual committing a crime varies significantly (9). Additionally, in another paper, the fairness of risk assessment instruments in recidivism, which is the tendency of a convicted criminal to reoffend, was tested. Instead of testing how fair decisions should be made, they surveyed users on how they perceive and reason about fairness in algorithmic decision-making (10). They also identify possible pathways that they can take to address these concerns of unfairness in recidivism risk assessment instruments (10).

Our objective was to develop a criminal activity index using machine learning techniques that can be used for various purposes, such as identifying safe places to purchase property, travel routes, areas that require an increase or decrease in police resources, and the safest time of day to engage in activities. To develop our index, we considered various environmental factors, including median income, population, unemployment duration, median age, county, and year, as these factors are known to be correlated with crime rates (5,6). We used the California OpenJustice Arrest Dataset to draw conclusions and predictions about criminal rates through our trained models in counties throughout California, creating a tabular index that will benefit the public in making informed decisions.

We carefully considered the potential ethical issues surrounding crime prediction algorithms, particularly their susceptibility to racial bias, and sought to reduce bias as much as possible by for example taking all of California into account, though previous research highlights it is impossible to have an efficient machine learning model without any bias (11). It's important to note that the potential pitfalls of using predictive policing algorithms have been well documented, as they may lead to unfair targeting and harassment of certain communities. Additionally, we recognize the historical issue of redlining and its impact on property ownership and the allocation of resources. Therefore, we will ensure that our index does not perpetuate these issues and that our work considers these historical injustices.

RESULTS

We used data from Open Justice California website and Orange data mining to run our tests. Our results from two sets of five different machine learning models – the first set with all four factors (Neural Network, Random Forest, SVM, Naive Bayes, kNN) and another set without median age – predict crime rate. These models show that neural networks perform better by all metrics in both experiments on the test data. The first metric, area under the curve (AUC), is calculated based on average predicted vs actual data points where the closer the number is to one the more accurate it is, with scores of 0.920 and 0.918 respective to first set and second set. The second metric is, Classification Accuracy (CA), which is totally correct over several points, and the closer the CA is to 1 the better, scores of 0.739 and 0.740, F1-score, is a harmonic mean of precision and recall, of 0.739 and 0.739, and Precision/Recall scores of 0.739 and 0.740 (**Table 1**). Random forest and support vector machine (SVM) performed just below the neural network. Random forest scores respectively: 0.897, 0.709, 0.705, 0.704, 0.709 and SVM scored respectively: 0.9, 0.699, 0.698, 0.703, 0.699 (**Table 1**). While Naive Bayes and k nearest neighbors (kNN) are relatively far behind. Naive Bayes scored respectively: 0.867, 0.632, 0.628, 0.629, 0.632 (**Table 1**). kNN scored respectively: 0.841, 0.605, 0.598, 0.598, 0.605 (**Table 1**).

The goal of our model was to determine their relation/correlation with crime rate based on our factors. In an effort to comprehend the underlying patterns within the dataset, we conducted a bivariate analysis of a number of variables. One of the initial observations was the inverse relationship between median income and crime rate per 100,000 individuals, where areas with low median income demonstrated significantly higher crime rates compared to those with high median income (R-value = -0.6779, **Figure 1**). Among the four factors, median income had the highest F-value of 131.072, indicating the strongest association with crime rates. This suggests that areas with lower median incomes may have higher crime rates. We also examined the impact of unemployment duration on the crime rate and found that low unemployment durations tend to be correlated with low crime rates, but a scattered pattern emerged as the unemployment duration increased (R-value = 0.3626, **Figure 2**). Notably, we found a relatively scarce number of counties that had high crime rates

Factors Used	**Median Income, Population, Unemployment Duration, Median Age**					**Median Income, Population, Unemployment Duration**				
Model	Neural Network	Random Forest	SVM	Naive Bayes	kNN	Neural Network	Random Forest	SVM	Naive Bayes	kNN
AUC	0.92	0.897	0.9	0.867	0.841	0.918	0.901	0.902	0.872	0.841
CA	0.739	0.709	0.699	0.632	0.605	0.74	0.719	0.699	0.631	0.605
F1	0.739	0.705	0.698	0.628	0.598	0.739	0.716	0.697	0.62	0.598
Precision	0.739	0.704	0.703	0.629	0.598	0.74	0.715	0.703	0.621	0.598
Recall	0.739	0.709	0.699	0.632	0.605	0.74	0.719	0.699	0.631	0.605

Table 1: The results of different models we trained using two different sets of data. One data set used all factors and the other removed median age. The models are ordered from best to worst precision going from left to right.

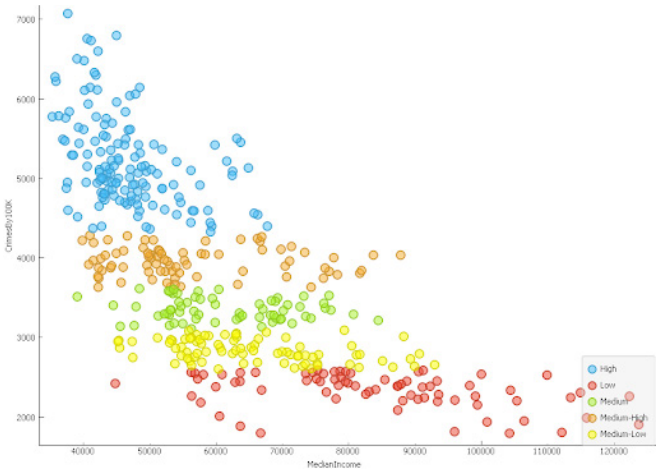

Figure 1: Median Income's correlation with Crime Rate. Each circle represents one of California's counties in a specific year (2011-2019) and the color indicates the number of crimes in 100rds of thousands. Crime rate is shown as the average number of crimes per 100k people. This data is from the Open Justice California website and is presented in a bivariate model.

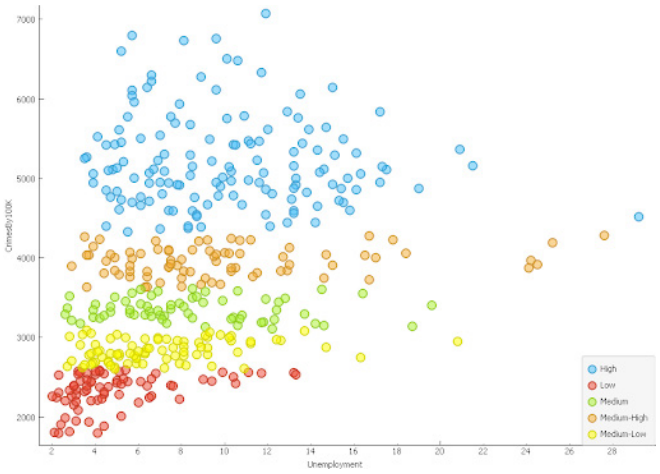

Figure 2: Average Unemployment duration's correlation with Crimes Per 100k. Each circle represents one of California's counties in a specific year (2011-2019) and the color indicates the number of crimes. Crime rate is shown as the average number of crimes per 100k people.

with low unemployment and vice versa. We also evaluated the population in relation to the crime rate, which indicated that counties with low populations exhibited high crime rates (R-value = -0.326, **Figure 3**). Population and unemployment duration also had a strong association with crime rates, with F-values of 35.191 and 23.893, respectively. This suggests that areas with higher population densities and higher unemployment duration may also have higher crime rates. An analysis of median age against crime rate did not reveal any significant correlations (R-value = 0.0938, **Figure 4**). Finally, the median age had a lower but still statistically significant F-value of 7.122. This suggests that areas with higher median ages may have lower crime rates, although the correlation is smaller than that of the other factors. However, normalizing the data to a normal distribution assumption does not explain this phenomenon. Further analysis of population versus median income sheds light on this discrepancy, as the results indicated that a low population is often associated with low median income, thereby correlated with higher crime rates (**Figure 5**).

DISCUSSION

The results of our study provide important insights into the relationship between various socioeconomic factors and crime rates. Our bivariate analysis revealed an inverse relationship between median income and crime rate, as areas with low median income demonstrated significantly higher crime rates compared to those with high median income. Furthermore, our analysis of unemployment duration and crime rate revealed that low unemployment durations tend to be correlated with low crime rates, but a scattered pattern emerged: as the unemployment duration increased a weaker association occurred over time. This suggests that while unemployment may be a contributing factor to crime rates, it is not the sole determining factor, and other variables may also be at play.

Our analysis of median age against crime rate did not reveal any significant correlations, indicating that median age was a slightly correlated factor in this context. However, it is worth noting that the effect of median age was statistically significant but smaller than that of the other factors. This suggests that while median age may not be a strong predictor of crime rates, it may still have some impact. The population size was also a significant predictor of crime rates, with low-population areas exhibiting higher crime rates. This relationship may be partially explained by the fact that low-population areas are often associated with low median incomes, which in turn can correlate with higher crime rates.

Neural networks performed the best in our study due to their inherent ability to model complex, non-linear relationships, which are often present in real-world data. In the context of crime prediction, factors such as median income, population, unemployment duration, and median age are likely to interact with each other in complex ways to influence crime rates. Neural networks are capable of capturing these interactions through their hidden layers, which can learn and represent high-level features in the data. However, while neural networks performed the best in our study, it's important to note that they are not always the best choice for every problem. One of the main considerations when choosing a machine learning model is the interpretability of the model. Neural networks, due to their complexity, are often referred to as "black box" models, meaning that it can be difficult to understand exactly how they are making their predictions. This can be a significant drawback in fields where interpretability is important, such as in the criminal justice system, where decisions can have serious consequences for individuals' lives. Other models like Random Forest and Support Vector Machines (SVM) might not perform as well as neural networks in terms of prediction accuracy, but they offer more interpretability. For instance, Random Forests provide feature importance scores that can help in understanding which factors are most influential in predicting crime rates. SVMs, on the other hand, are based on the concept of finding a hyperplane that best separates the classes in the data, which can be easier to visualize and understand in certain cases.

This highlights the potential of machine learning in

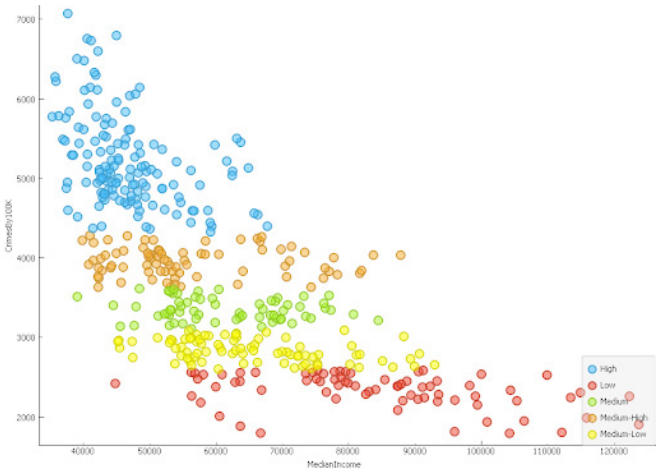

Figure 3: Population's correlation with Crimes Per 100k. Each circle represents one of California's counties in a specific year (2011-2019) and the color indicates the number of crimes. Crime rate is shown as the average number of crimes per 100k people.

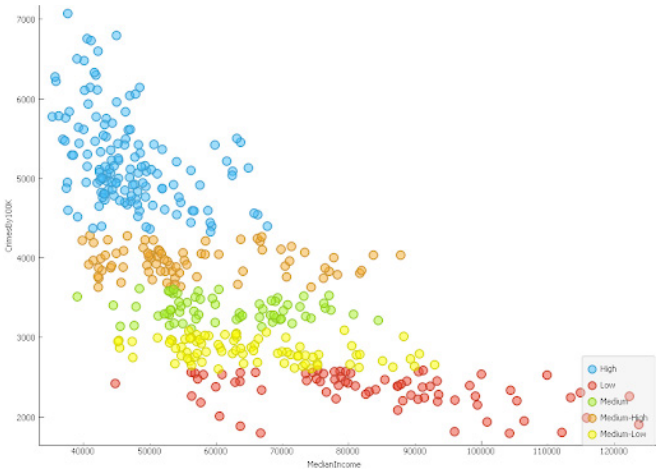

Figure 4: Median Age's correlation with Crimes Per 100k. Each circle represents one of California's counties in a specific year (2011-2019) and the color indicates the number of crimes. Crime rate is shown as the average number of crimes per 100k people.

predicting crime rates and identifying areas that are at risk of high crime rates. Overall, our study contributes to our understanding of the complex relationship between socioeconomic factors and crime rates. The findings suggest that addressing poverty and economic disadvantage may be key to reducing crime rates in high-risk areas. However, other factors such as unemployment, population density, and age demographics also need to be taken into account when formulating crime prevention strategies.

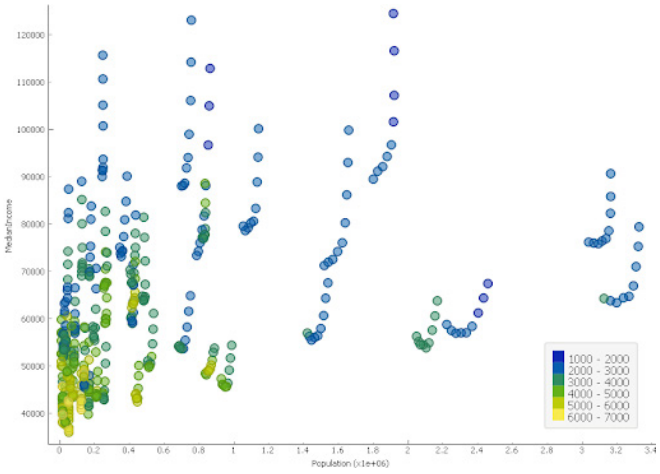

Figure 5: Population and Median Income Correlation with Crimes Per 100k. Each circle represents one of California's counties in a specific year (2011-2019), and the color indicates the average number of crimes per 100k people.

MATERIALS AND METHODS

For the analysis, crime data were obtained from the State of California DOJ Arrests dataset (12). This dataset included a summary of each California county with counts of several types of crime. The factors studied that affected crime data were obtained using the DataCommons API (13). From this API information, median income, unemployment duration, median age, and population were collected for each county across the years 2011-2019. Each county had nine sets of data points, and due to there being 58 counties in California, a total of 522 data point sets were available. To consider outliers, we restricted the data to only data point sets with a population between 20,000-5,000,000 (7.5-92.5%), leaving us with 441 data points. The datasets were then merged and correlated based on County Name and Year. All data was then stored in a SQL database to be exported as a flat file for data mining using Orange Data Mining (14).

Before modeling, all crime data were normalized to the population, providing the variable Crimes/100K population to remove data skews due to population in the bivariate analysis. The Crimes/100K variable was then clustered using Louvain Clustering, to split the variable into five discrete categories, "Low", "Medium-Low", "Medium", "Medium-High", and "High" Crime counties, creating the final prediction variable.

We conducted a one-way ANOVA test using Orange Data Mining to identify which of the four factors impacted the crime rate the most in our experiment.

To determine the best predictive model, five machine learning models (neural network, SVM, naive Bayes, kNN, and random forest) were picked and tested on the dataset. All of these models were run through Orange Data Mining using the default model while changing some of the parameters as outlined below. For the neural network, we used ReLu activation with 100 by 100 neurons per layer, and to reduce cost, we limited the maximum number of iterations to 200. For our SVM model, we used an RBF kernel and made the iteration limit at 200. For naive Bayes and kNN models, we used the Orange default parameters. Finally, for our random forest model, we set the number of trees to ten and limited the smallest subset to five. Overall, we found that these combinations of parameters provided a good balance between accuracy and interpretability for our dataset.

Using Median Income, Median Age, Unemployment Duration, and Population as our inputs for one set and Median Income, Population, and Unemployment Duration for another, each model was trained on 80% of the data to predict crime rate, and the rest of the 20% was used for testing. Models were

ranked and chosen based on AUC, area under the receiver operating characteristics (ROC) curve, and CA all produced through Orange Data Mining.

ACKNOWLEDGEMENTS

We would like to thank Olive Children's Foundation for corporate sponsorship to fund Aspiring Scholars Directed Research Program such that we can have this opportunity to conduct this research.

Received: September 7, 2022
Accepted: April 25, 2023
Published: August 29, 2023

REFERENCES

1. Moeinizade, Saba, and Guiping Hu. "Predicting Metropolitan Crime Rates Using Machine Learning Techniques." *INFORMS International Conference on Service Science*, 26 Nov. 2019, https://doi.org/10.1007/978-3-030-30967-1_8.
2. Zhang, Xu, et al. "Comparison of Machine Learning Algorithms for Predicting Crime Hotspots." *IEEE Access*, vol. 8, 2020, pp. 181302-181310, https://doi.org/10.1109/ACCESS.2020.3028420.
3. Catel, Cagatay., et al. "Machine Learning in Crime Prediction." *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, 2023, pp. 2887-2913, https://doi.org/10.1007/s12652-023-04530-y.
4. Hadjimatheou, Katerina, and Christopher Nathan, 'The Ethics of Predictive Policing ', in Carissa Véliz, The Oxford Handbook of Digital Ethics, Accessed 2 Apr. 2023, https://doi.org/10.1093/oxfordhb/9780198857815.013.22.
5. Mi, Joyce. "Variable Selection Methods With Applications To Crime Prediction." California State Polytechnic University, Pomona, Summer 2016, pp. iv-v. scholarworks.calstate.edu/downloads/z029p6976.
6. Quednau, Joseph, "How are violent crime rates in U.S. cities affected by poverty?" The Park Place Economist: Vol. 28 2001, digitalcommons.iwu.edu/parkplace/vol28/iss1/8.
7. Johnson, Shane D. "The Complexity of Crime and Security." The Routledge Handbook of Security Studies, edited by Myriam Dunn Cavelty and Thierry Balzacq, Routledge, 2010, pp. 94-105.
8. Nath, Shyam Varan. *Crime Pattern Detection Using Data Mining - Brown University*. 2006, cs.brown.edu/courses/csci2950-t/crime.pdf.
9. Lum, Kristian, et al. "Closer than They Appear: A Bayesian Perspective on Individual-Level Heterogeneity in Risk Assessment." *ArXiv.org*, 1 Feb. 2021, https://doi.org/10.48550/arxiv.2102.01135.
10. Grgić-Hlača, Nina, et al. "Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction." *ArXiv.org*, 26 Feb. 2018, https://doi.org/10.48550/arxiv.1802.09548.
11. Berk, Richard, et al. "Fairness in Criminal Justice Risk Assessments: The State of the Art." ArXiv, 28 May 2017, https://doi.org/10.48550/arxiv.1703.09207.
12. "State of California Department of Justice." *OpenJustice*, openjustice.doj.ca.gov/data.
13. *Home - Data Commons*, datacommons.org/.
14. Bioinformatics Laboratory, University of Ljubljana. "Data Mining." *Orange Data Mining - Data Mining*, orangedatamining.com/.