Final Report
Survey of Machine Learning Models in Crime Prediction

Na Le, Jinyi Ouyang, Jude (Ken Yoong) Lim
Virginia Tech School of Computer Science
n4l2@vt.edu,jinyi@vt.edu,lkyoong428@vt.edu

Abstract

With millions of crimes reported each year in the United States, safety and security have always been top concerns of citizens and lawmakers alike. To assess the threat and improve overall safety of communities, many attempts have been made to apply machine learning algorithms in clustering and predicting crime in different areas. These predictions allow for more effective utilization of law enforcement while helping identify and address social issues associated with high crime rates. This paper reviews 14 published works to identify popular machine learning algorithms used in the domain of crime prediction. We then conduct tests with a variety of experiments and techniques with these algorithms to assess these algorithms' performance and their shortcomings.

1 Introduction

It is estimated that more than 8 millions of crime offenses have been committed each year in the United States since 1960. Regardless of our position in a community, be it a resident, student, or just a visitor, safety and security will be one of our top priorities when in an area. Advancements in data mining and machine learning presents an exciting opportunity to apply these techniques to ensure the safety of our communities.

However, utilization of machine learning techniques in this space is not without its issues. The stakes are extremely high in this space, and any unaddressed mistakes from an algorithm has the potential to severely disrupt many lives with issues such as over-policing and wrongful incercerations. As such, it is important that models being applied to this domain are as precise as possible.

Though the papers we studied provided us many meaningful and thoughtful insights of crime data mining in crime investigation, these models were not fine-tuned with different imputation techniques and were not compared thoroughly. Therefore, we wanted to fill this gap by answering two main research problems: which machine learning algorithms perform best in crime data mining and what could affect the performance of these models.

We studied a large number of existing published works in order to accurately determine popular machine learning algorithms in the domain of crime prediction in section 2. We then use this knowledge to inform our algorithm selection as well as our experiment methodology. This process is detailed in section 3. We report the results and observations from our experiments in section 4. Finally, we conclude by discussing the implications of our results as well as potential ways forward in the domain.

2 Literature Review

We reviewed 14 papers [1][2][3][4][5][6][7][8][9][10][11][12][13][14] related to the utilization of Machine Learning Models in Crime Prediction to lay the foundation for the different algorithms. From these papers we were able to gain a general overview of a large variety of machine learning algorithms used in crime prediction.

Broadly, most applications [1] - [14] of machine learning in the space of crime prediction scan through recorded information from public databases to formulate the patterns of how, where and when a crime happens and then predict the location and the type of offense of a potential suspect using predictive analytics. While these applications can potentially support community confidence in criminal justice, they are also questionable in terms of efficiency and effectiveness. It is important that the models are bias-free so the polices do not get a wrong suspect and detain the real culprit in time.

There are also other projects that look into machine learning in the context of crime investigation. Papers [15][16] provide application examples of machine learning in real-world cases and addresses the implementation and interpretation problems in the previous works that lead to models' bias and wrong accusation. Paper [17] proposes potential prevention for the discrimination issue in crime data mining.

Finally, other surveys such as [18][19][20] revisited the advantages and disadvantages in using data mining methods to find the relationship between demographic factors and crime rate. They also investigated the crime patterns from the correlation between space and time and addressed the challenges from previous studies. These surveys provided us more context of how the criminal phenomena interacted with the outside world.

3 Methodology

3.1 Algorithm and Metric Selection

We found that papers [1], [3] - [8], [10] - [13] were more relevant to our current project since they discusses the machine learning algorithms in more details with deeper discussion on the pros and cons of these algorithms. In Lin et al.'s, Safat's and Kim et al.'s studies, different models were compared and contrasted, including K-nearest neighbors, Decision Tree, Random Forests and Naive Bayes Algorithms. We noticed that Risk Terrain Models and Kernel Density Estimation have been introduced in Wheeler and Steenbeek's study, where they discussed and compared them with Random Forests.

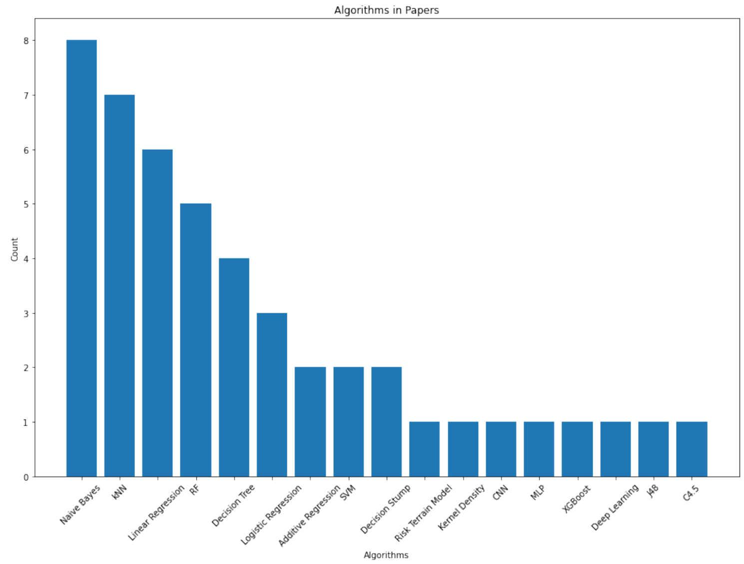
Figure 1 shows the distribution of algorithms mentioned in the papers we reviewed. We chose some of the more common models observed for testing. We also chose a few less mentioned, but well-performing algorithms to test in our project. We eventually arrived at a list of 8 algorithms to test, namely: Decision Tree (DT), Naive Bayes (NB), K-Nearest Neighbors (KNN), K-Means Clustering (KM), Multilayer Perceptron (MLP), Support Vector Machine (SVM), Random Forest (RF) and Logistic Regression (LR).

Our literature review initially drove us to assess the algorithms using runtime, precision, recall, and F-measure [9][13]. We also planned on including accuracy when assessing the algorithms as it was observed to be quite a popular metric [2][3][6][9][13]. However, after some deliberation, we decided to focus on precision as we decided that in the real world, it would be more important that people do not get falsely accused of crimes.

3.2 Experimental Setup

We utilized 5-fold cross validation with our prepared data to assess the effectiveness of each algorithm. The performance across all folds is averaged and recorded. This process is repeated for each different missing value filling and dimensionality reduction technique we tried.

Figure 1: Algorithm occurrence frequency in reviewed papers.



3.3 Data

We utilized the Communities and Crime from the UCI Machine Learning repository with 1994 entries and 124 features. This dataset integrated the socio-economic data from the 1990 US Census, crime data from the 1995 FBI UCR and law enforcement data from the US LEMAS survey, so it included the fundamental elements that could affect the crime rate in an area such as population, household size, race, education, salary, and age.

Our first step in processing the data is to remove non-predictive columns such as location data. As we only selected classifiers to experiment with, we discretized the continuous dependent variable "ViolentCrimesPerPop" to better tailor the dataset to our needs.

3.4 Filling techniques

As with all datasets, many entries of this data contain missing values. Prior research has proven that the method at which these values are filled can heavily influence the performance of a given algorithm. As such, we investigated a variety of filling techniques to identify the best method for filling missing values in data for this particular domain.

We investigated filling missing values with: mean, median, k-nearest neighbour, forward-fill, and interpolation.

3.5 Tuning Hyperparameters

We experimented with different hyperparameter values for all 8 algorithms and arrived at the following values. The values in Table 1 produced the best performance for any given missing value filling technique and were used for all subsequent experimentation.

3.6 Dimensionality Reduction

Dimensionality reduction transforms any given dataset into a lower dimension while retaining the meaning of the original data. Historically, these techniques are commonly utilized to reduce the computational requirements of an application when dealing with large datasets containing many features. Dimensionality reduction can also increase the reliability of the results from various algorithms by making sure only relevant features affect the final model.

Table 1: Hyperparameter value of classifiers.

	Parameter	Value
DT	splitters	random
	max_feature	auto
NB	var_smoothing	1.00E-02
KNN	n_neighbors	7
	weights	distance
KM	max_iters	600
	n_clusters	2

	Parameter	Value
MLP	activation	relu
	solver	sgd
	max_iter	1000
SVC	kernel	linear
	gamma type	scale
RF	criterion	entropy
	max_depth	15
LR	solver	liblinear

Table 2: Performance comparison of machine learning algorithms in crime prediction.

	Filling	**DT**	**NB**	**KNN**	**KM**	**MLP**	**SVM**	**RF**	**LR**
Precision	Median	0.62	0.5962	0.6127	0.5559	0.6574	0.6426	0.6477	0.6199
	Mean	0.617	0.6015	0.6158	0.5917	0.6488	0.6454	0.6474	0.62
	kNN	1	0.7709	0.6784	0.598	0.962	0.8099	0.9456	0.743
	ffill	0.6216	0.6415	0.6173	0.2883	0.639	0.6601	0.6622	0.6186
	interpolation	0.6009	0.6298	0.6092	0.3377	0.6404	0.6483	0.6437	0.6162
Runtime	Median	0.1919	0.0219	0.0699	0.4473	3.704	0.3161	0.8355	0.2759
	Mean	0.1926	0.0214	0.0684	0.373	3.7305	0.3229	0.8363	0.1898
	kNN	0.0603	0.0218	0.0694	0.4245	2.3662	0.2856	0.731	0.186
	ffill	0.2033	0.0216	0.0662	0.4681	4.7035	0.5064	0.896	0.187
	interpolation	0.2364	0.0219	0.069	0.4097	3.7787	0.326	0.9898	0.1871

We picked Principal Component Analysis (PCA) as it represents one of the most common dimensionality reduction techniques used in both research and industry.

4 Results

Of all filling techniques as shown in Table 2, it was noted that kNN imputation resulted in the best performance. Intuitively this makes sense as areas with similar demographics and properties are more likely to share other properties such as crime rates. The performance of other fill methods were quite similar to each other for most algorithms, with no method performing notably better than others. The most precise algorithms included Decision Tree, Multi-layered Perceptrons, and Random Forests.

Table 3 summarizes the output of precision and runtime of each model when we applied PCA to the data with number of components (C) from 1 to 9.

We noted that most models' performances peaked at 6 components and the weakest performing model, K-means was significantly improved by the application of PCA. However, all results with PCA were poorer than when PCA was not used. The best performing algorithm with PCA was multi-layered perceptron while the lowest performing was K-Means at 9 components. It should also be noted that PCA reduced runtime across all algorithms.

5 Discussion and Conclusion

Some limitations were identified in our study. The dataset used was quite dated, and might not be the best indicator as to how these algorithms would perform with modern data. While we thought the results observed were valid on a general level, the procurement and usage of a more modern dataset would give our results more relevance to modern implementations of machine learning in crime prediction.

Dimensionality reduction, while successful in reducing runtime, ultimately worked against the needs of the domain. While some might argue that speed was important when trying to prevent crime, we ultimately decided that the ability to accurately identify crime factors was much more important. Furthermore, utilizing dimensionality reduction on the data obfuscated the descriptive nature of the

Table 3: Performance of 8 models with different PCA components

C		**DT**	**NB**	**KNN**	**KM**	**MLP**	**SVM**	**RF**	**LR**
1	Precision	0.5737	0.6058	0.5794	0.3850	0.5929	0.5675	0.5737	0.5428
	Runtime	0.0276	0.0100	0.0392	0.2392	5.3438	0.2560	0.5814	0.0313
2	Precision	0.5998	0.5969	0.6207	0.3631	0.6212	0.5885	0.6235	0.5553
	Runtime	0.0248	0.0059	0.0168	0.0564	2.0747	0.1235	0.3298	0.0186
3	Precision	0.6235	0.5671	0.6173	0.3574	0.6310	0.6182	0.6492	0.5895
	Runtime	0.0275	0.0059	0.0167	0.0526	1.9936	0.1010	0.3407	0.0189
4	Precision	0.6369	0.5709	0.6384	0.3503	0.6733	0.6481	0.6551	0.6244
	Runtime	0.0323	0.0058	0.0200	0.0488	2.0439	0.1096	0.4264	0.0203
5	Precision	0.6384	0.5886	0.6439	0.3693	0.6739	0.6402	0.6548	0.6224
	Runtime	0.0274	0.0059	0.0180	0.0486	2.0754	0.1177	0.4337	0.0235
6	Precision	0.6592	0.5899	0.6461	0.5450	0.6941	0.6457	0.6639	0.6280
	Runtime	0.0376	0.0060	0.0205	0.1969	2.2606	0.1210	0.4311	0.0270
7	Precision	0.6554	0.6171	0.6713	0.3649	0.7256	0.6902	0.6962	0.6100
	Runtime	0.0371	0.0063	0.0211	0.3048	2.2084	0.1279	0.4351	0.0294
8	Precision	0.6561	0.6211	0.6784	0.3440	0.7137	0.7046	0.7160	0.6408
	Runtime	0.0371	0.0058	0.0200	0.1557	2.0240	0.1217	0.4433	0.0309
9	Precision	0.6641	0.6119	0.6907	0.3265	0.7411	0.7117	0.7132	0.6453
	Runtime	0.0424	0.0064	0.0228	0.2033	2.1331	0.1342	0.5411	0.0254

features, making the model completely uninterpretable and reducing user trust, another important factor in this domain.

One of the most notable findings in our experiments was that with the correct combination of algorithm and filling method, near-perfect performance can be achieved. However, we theorized that this might be due to overfitting. Furthermore, many other uncertainties exist when applying machine learning in the real world such as social changes over time and adversarial attacks on the system, casting doubt on these algorithms' performance in real-world applications.

Our results matched the results from the existing works quite well. Our best performing algorithms, namely: decision trees, multi-layered perceptrons, and random forests were observed to be commonly high-performing algorithms in the papers we reviewed. This result showed that the research in the domain was going in a good direction and producing increasingly precise techniques.

While our results looked promising, it was important that checks and balances exist in systems that utilize machine learning for crime prediction. Implementations of explainable AI systems, as well as having humans-in-the-loop to catch potential mistakes was the key to ensuring the success of these systems while algorithmic performance progresses to the levels required for true automation.

6 Contributions

Na Le - Literature review, Filling Algorithms, Tuning Hyperparameters, Drafting the Final Report

Jinyi Ouyang - Literature review, Coding Algorithms, Drafting the Final Report

Jude (Ken Yoong) Lim - Literature Review, Paper characteristic analysis, Data Cleaning, Additional Filling Algorithms, Drafting the Final Report

7 References

[1] Mahmud, S. Nuha, M. & Sattar, A. (2021) Crime Rate Prediction Using Machine Learning and Data Mining. In *Soft Computing Techniques and Applications* (pp. 59-69). Springer, Singapore.

[2] Shah, N. Bhagat, N. & Shah, M. (2021) Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. Computing for Industry, Biomedicine, and Art, 4(1), 1-14.

[3] Kouidri, M. Yasin, N. M. & Al-Garadi, M. A. (2019) Crime prediction for stop and search outcomes using machine learning.

[4] Alves, L. Ribeiro, H. V. & Rodrigues, F. A. (2017) Crime prediction through urban metrics and statistical learning. *Physica A: Statistical Mechanics and its Applications, 505,* 435-443.

[5] Paladugu, S. Yakkala, T. S. Boggarapu, N. & Modekurty, S. K. K. (2021) Crime Rate Prediction Using Machine Learning. *International Journal of Research in Engineering, Science and Management, 4* (9), 245–246.

[6] Kim, S. Joshi, P. Kalsi, P. S. & Taheri, P. (2018) Crime analysis through machine learning. In *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (pp. 415-420). IEEE.

[7] Shukla, A. Katal, A. Raghuvanshi, S. & Sharma, S. (2021) Criminal Combat: Crime Analysis and Prediction Using Machine Learning. In *2021 International Conference on Intelligent Technologies (CONIT)* (pp. 1-5). IEEE.

[8] Sun, C. C. Yao, C. Li, X. & Lee, K. (2014) Detecting Crime Types Using Classification Algorithms. *J. Digit. Inf. Manag., 12*(5), 321-327.

[9] Safat, W. Asghar, S. & Gillani, S. A. (2021) Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques. *IEEE Access, 9,* 70080-70094.

[10] Prabakaran, S. & Mitra, S. (2018) Survey of analysis of crime detection techniques using data mining and machine learning. In *Journal of Physics: Conference Series* (Vol. 1000, No. 1, p. 012046). IOP Publishing.

[11] Yerpude, P. (2020) Predictive modelling of crime data set using data mining. *International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol, 7.*

[12] McClendon, L. & Meghanathan, N. (2015) Using machine learning algorithms to analyze crime data. *Machine Learning and Applications: An International Journal (MLAIJ), 2*(1), 1-12.

[13] Lin, Y. L. Chen, T. Y. & Yu, L. C. (2017) Using machine learning to assist crime prevention. In *2017 6th IIAI international congress on advanced applied informatics (IIAI-AAI)* (pp. 1029-1030). IEEE.

[14] Wheeler, A. P. & Steenbeek, W. (2021) Mapping the risk terrain for crime using machine learning. *Journal of Quantitative Criminology, 37*(2), 445-480.

[15] Oatley, G. C. (2022). Themes in data mining, big data, and crime analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12(2)*, e1432. https://doi.org/10.1002/widm.1432

[16] Oatley G. & Brian W. (2011) Data Mining and crime analysis. *Wiley Interdisciplinary Reviews*, Vol 1. https://doi.org/10.1002/widm.6

[17] Sara H. & Joseph D.F. & Antoni M.B. Discrimination prevention in data mining for intrusion and crime detection. *2011 IEEE Symposium on Computational Intelligence in Cyber Security (CICS)*. https://doi.org/10.1109/CICYBS.2011.5949405

[18] Hamdi, A. & Shaban, K. & Erradi, A. et al. (2022) Spatiotemporal data mining: a survey on challenges and open problems. *Artif Intell Rev 55*, 1441–1488. https://arxiv.org/abs/2103.17128

[19] Li, X. & Joutsijoki, H. & Laurikkala, J. & Juhola, Martti (2015). Crime vs. demographic factors revisited: Application of data mining methods. *Webology, 12(1), Article 132*. http://www.webology.org/2015/v12n1/a132.pdf

[20] Karimi A. & Abbasabadei S. & Torkestani JA. & Zarafshan F. Process Modeling and Extraction of Patterns of Computer Crimes Using Data Mining. *Computer Science Journal of Moldova. 2020;28(1)*:45-58. Accessed May 4, 2022. https://search-ebscohost-com.ezproxy.lib.vt.edu/login.aspx?direct=truedb=a9hAN=142939123scope=site