GROUP 159: Kevin Thomas Henely, Liutao Du, Mohammed Nasim Alvi, Songhee Yim, Cemil Selcukler

Real Time-Traveler Crime Alert

Intro/ Motivation

In a world of abundant threats, tourists stand out as easy targets for a variety of compelling reasons. They often possess valuables, which can attract individuals with ill intentions. Furthermore, tourists' vulnerabilities are exacerbated by their limited familiarity with the area. Surprisingly, tourists frequently opt not to report grievances, and civil or criminal offenses, prioritizing an uninterrupted travel experience. The limitations of existing crime analysis methods are significant, impeding their efficacy. A notable deficiency is the absence of a comprehensive dashboard capable of categorizing potential crimes in specific locations by their risk level, which would be a valuable tool for travelers seeking to prevent potential crimes.

This project embarks on the journey to enhance crime prediction through innovative approaches. We introduce a risk scoring system in the form of a heatmap, designed to assist users in identifying potential high-risk areas. Furthermore, we plan on developing prediction algorithms that utilize novel ML models to predict crime types. In addition, we employ association confidence clustering to compute the relationship between various crime types. These multifaceted strategies promise to provide valuable insights into the realm of crime analysis.

Problem Definition

Crime data analysis is a new field with many difficulties. Geospatial analysis methods are currently either computationally inefficient or lack enough real-time data to be effective. We aim to address these issues by developing an approach that uses geographic crime data to compute and rank probabilities of various crime scenarios. This real-time data aims to reduce travel risk face in unfamiliar locales with improved accuracy when compared to contemporary methods. We aim to use hotspot clustering techniques to be able to provide causal explanations of the risk factors that may lead to victimization. We can then relay our crime risk results in an informative map UI that acts as a localized alert system that notifies travelers of the crime risks in their immediate area.

Literature Review

There are various AI/ML techniques that are used today to support data mining and analysis. Classification methods such as XGBoost (Yan, 2022), and Ensemble Techniques (Ara'ujo, 2018 and Alhmuhanna 2021) are a few leading methods. However, the accuracy of these prediction models typically hovers around 40-60%. Geospatial Hotspot Clustering is another popular approach, and is computationally less costly allowing for efficient/robust training. Here are a few specific techniques that were helpful:

a) <u>Aggregate Incident Methods</u> (Giannis, 2005) use spatial association tests to generate thematic grid maps. Cluster correlations are computed based on aggregate crime counts and weighted via spatial matrix.

b) <u>Point Pattern Analysis</u> identifies discrete crime event clusters, based on representative parameters. kNN clustering, spatial/temporal analysis (STAC), k-means clustering are a few examples.

These clustering algorithms are not good at identifying causality (Estivill-Castro, 2021). These models also lack interpretability which is relevant in crime prevention since there are implications on communities.

c) <u>Kernel Density Estimation (KDE)</u> (Levine, 2008): A prospective risk map is created with interpolated data derived from known crime incident locations.

d) <u>Risk Terrain Modeling (RTM)</u> (Caplan, 2011): RTM identifies the spatial influence of multiple risk factors within a specific area to assess net risk. These risk factors are used to create separate KDE map layers that are combined to produce a composite risk terrain map.

While helpful in identifying crime patterns, these models tend to have poor accuracy. The concept also assumes that the past patterns persist into the future, which may not always be true. Since crime reporting tends to come from various sources, and some examples of crimes are poorly tracked, there is a need to integrate multiple datasets (Feng, 2019). There are also deficiencies in how this information is shared with users. Studies show that travelers are not aware of crime risk surrounding them (Liu. et al., 2023).

Proposed Method #1: Heatmap of Risk Score

A Heatmap of Risk Scores for crime occurrence would be an innovative tool to support tourist safety. It visually represents the likelihood of crimes occurring in different areas by utilizing a color-coded system. This system analyzes historical crime data, demographic information, and can incorporate real-time reports from local law enforcement and social media trends to predict potential crime hotspots. Advanced algorithms, including machine learning models, identify patterns and correlations that may be overlooked by human analyses. The proposed Heatmap Risk Score model utilizes the below risk score calculation method. The result identified the entry with the highest normalized risk score, indicating a significant concentration of risk factors. Subsequently, we utilize a Gaussian Mixture Model (GMM) to spatially group the dataset into clusters to identify distinct patterns.

<u>*Risk Score Calculation:*</u>

1. *Defined weights for each factor: 'Vict_Age' (0.4), 'Premis_Cd' (0.3), and 'Weapon_Used_Cd' (0.3).*
2. *Calculated the risk score for each row using the weighted sum of factors.*
3. *Normalized the risk scores using Min-Max scaling to bring scores within a uniform range (0 to 1).*

<u>*GMM model:*</u>

1. *Used the Gaussian Mixture Model (GMM) to cluster the dataset.*
2. *Utilized BIC scores to find the best fit number of components value (n_components)*

This approach is taking the feature vector and then normalizing the data so all the features are scaled 0 to 1. Then weights are assigned to each features and an experimental optimization approach is taken to compute the weight of each feature. We measured performance indicators such as silhouette score, and bayesian-information-score (BIC) to determine the best features for crime clustering. We <u>**intuit**</u> this normalization step is better than many crime clustering approaches because it is typically difficult to incorporate different features types, some of which may be numerical (i.e. age), others discrete (victim sex). We believe with an experimental approach we can discover the best features to use for crime clustering. Once we develop the optimal clusters though trial and error, we then intend to present this data in the form of a heatmap presented to user in a webapp UI where the user can have an integrated map experience and generate a color signature for regions that have high crime density. This will allow the user to have access to easily interpretable crime data in real-time while they are traveling in their route. This a significant improvement from the state of the art which typically does not allow real-time weather style models for crimes.

Figure 1: Heatmap UI

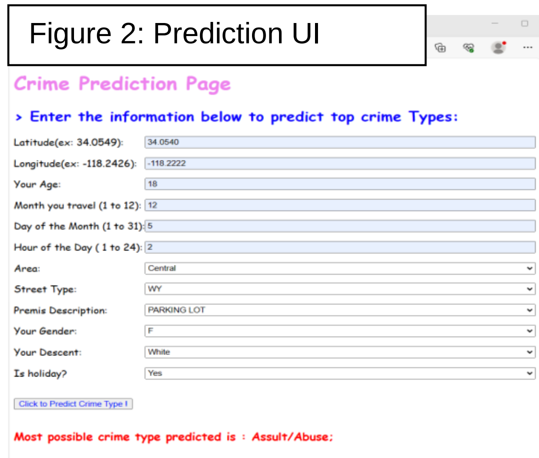
Proposed Method #2: Crime Type Prediction

With the normalized data we can now employ various machine learning (ML) Classification algorithms to predict the most probable crime type based on the user's input features, essentially a classic crime classification problem. In order to get the best performance, we run several major classifiers and compare. These algorithms include K-Nearest Neighbors (knn), Decision Tree classifier (DT), random forest (RF), and gradient boosted classifier, supportive vector machine-based classifier (SVM), a Bayesian classifier, and Extreme Gradient Boosting (XGBoost). XGBoost model is developed using the XGBoost library, all other algorithms are done with the sklearn library.

To run the ML pipeline, we first preprocess the dataset for data cleaning and scaling. Then we work on the feature engineering to select the important features and to reduce the features used int the model for better performance. After that the data is split into training and testing sets for modeling. Next, various classification algorithms are tested and evaluated for their performance. For feature selection and engineering, we use a stepwise selection approach to filter out the features that do not have a significant impact on the prediction. Our intuition for success comes from our use of the principle components analysis for feature reduction. For model tuning, we tested two approaches, grid search and random search, to turn the model's hyper parameters for its best performance. The random search method did not work because of the huge size of the training 25+ features. Grid search works better since it can reduce the number of options to search. For model comparison, we use cross validation to select the best one to use, based on metrics of accuracy, f1, precision and recall scores. This was done using sklearn's metrics methods.

In the end, the most appropriate classification model is stored on a Flask-based server as a pre-trained model with Pickle. For user prediction, the server will call the model and fit in the user's input parameter values and make the prediction (see below the interface and prediction result). We **intuit** that our approach will be successful because it exhaustively consider several supervised learning models. Moreover, we consider ensemble style learners which typically exhibit quite powerful predictive results while being very computationally efficient since one of the many aggregate models is bound to find success. We also present the user with a novel, interactive UI to conduct this data analytics on their own from a WebApp interface. This will allow data analysts to take advantage of this learning model and conduct their own classification trials.

Figure 2: Prediction UI



Proposed Method #3: Clustering Discovery and Association Confidence Mining

This innovation and experiment serves to compute the cause-effect relationship between crime types. We aim to develop and test knowledge discovery techniques that can compute the cause and effect relationship between variables. While typical statistical computations exist to compute feature relationships they can be computationally expensive in the data rich geo-spatial domain. As such, any algorithm that may allow users to navigate and find pattern between feature maps can be highly useful in understanding the bias and constraints on data. This sort of association rule mining can be used in tandem with many other predictive models to improve performance if we can demonstrate it is successful. We consider particularly our interest of being able to identify murder risk with high confidence using other crime occurrences. Our approach is better than the state of the art because typically when analysts try to create predictions on murder trends they find it difficult as murder occurrence is quite sparse compared to other crime data. By relying on predictive confidence between murder and other more common crime types we might be able to improve the ability to predict murder.

The baseline algorithm built here is called the knowledge discovery and spatial data mining or KDSDM (Estivill-Castro, 2001). It is conceptually simple to understand, but very powerful. The first step in the algorithm is to identify area clusters of crimes, or hotspots. Many algorithms may be used such as DBSCAN, kNN, or kMeans. For the purpose of this project we proceed with kMeans as it demonstrated the clearest separation between clusters and is computationally efficient. Once a set of crimes is identified as a cluster, a convex-hull (or bounding area) is drawn around each cluster. We repeat this for various crime types computing the convex-hull area for each case. The algorithm then applies Bayes Rule to compute the probability of one crime to predict the other, but instead of probability of individual, this approach uses the convex-hull area for the computation:

$$\text{Confidence} = \Pr[X \cap Y] / \Pr[X] = \text{Area}[X \cap Y] / \text{Area}[X]$$

In this formula we compute the confidence (or the probability of Y given X) in terms of area association. Thus Area[X ∩ Y] is the intersectional area of the cluster areas of crime X and crime Y. and Area[X] is simply the cluster area of crime X. We **intuit** this will be quite successful as bayesian mathematics are typically difficult to translate into a geospatial domain. But, our approach is attempting to do just that.

Experiment/Evaluation #1: Heatmap of Risk Score

<u>Questions:</u> Having outlined the innovative methods we'll employ to improve the geospatial assessment of crime-risk, we seek next to develop controlled experiments and evaluations to determine the effectiveness of our new methods. In each case our experimental approach will have an independent and dependent variable, with various trails where all other variables are controlled. In the first experiment set out to answer which gender and which geolocations have the highest crime prevalence.

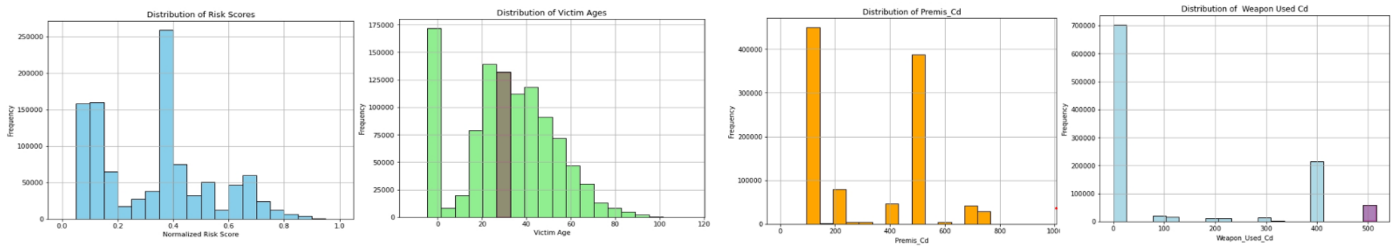
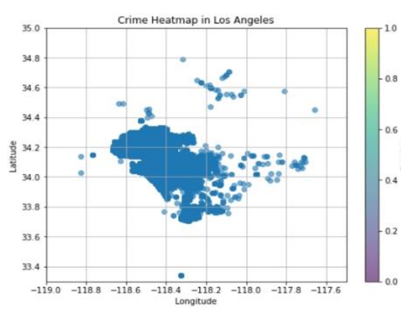


Figure 4: Distribution of various features (Left to Right: Risk Score, Age, Premise, Weapons)

<u>Observations:</u> In the dataset (Los Angeles Crime data), the highest normalized risk score (0.999999999) highlighted key features of the highest-risk victims. Notably, 'victim age' at 32, 'premise/area' recorded as 900 (MTA Redline to Union Station), and 'weapon used' was (unknown weapon/other) marked at 500.0. In fact, we observed several of the highest risk score data points to have similar values. This underscored a convergence of multiple risk factors within this region (*Figure 1*). This convergence emphasizes the importance of pattern identification. If we can identify a pattern of victim and perpetrator feature values we might be able take actions or even inform the police department to effectively address the combined risk factors and prevent similar incidents from occurring.

Figure 5 shows the heatmap of reported crimes by employing latitude and longitude coordinates. Various shades on the heatmap represent distinct crime densities across the areas, offering a visual insight into crime hotspots.



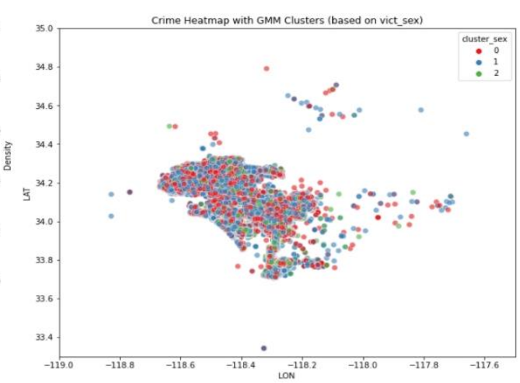


Figure 5: (Left) Heatmap

Figure 6: (Right) GMM cluster - based on victim sex

One of the most intriguing findings from the GMM experiments involved clustering by victims' sex/gender (*Figure 6*). Cluster 0 notably demonstrates a mean and median 'Vict_Sex' value of 2.0, indicating a female victim within this cluster. Moreover, the geographic coordinates (Lon_mean: 34.046, Lat_mean: -118.256 / Lat_median: 34.0599, Lon_median: -118.3274) denote the central crime locations in this cluster. On the other hand, Cluster 2 stands out with a notably lower average 'Vict_Age' (Mean: normalized to 0.160988) compared to other clusters which implies the involvement of younger individuals as victims. Understanding such variations in victim profiles among clusters assists law enforcement analysts allocate resources, or implement targeted interventions to the patterns observed within each cluster.

Experiment/Evaluation # 2: Crime Prediction

__Question:__ The question we set out to answer was whether feature (principle component decomposition) would help us increase the predictive performance of our models. Particularly if we emphasized features discovered from our other experiments to be have high risk scores and confidence values.

__Observations:__ The major challenges we faced were the large dataset for model training, the large numbers of features to filter out, and the significant amount of work on model tuning and comparison. To handle the feature selection, we employed stepwise feature selection and have found that the most important features to focus are related to location, time and the victim's person feature including age, gender and descent. We then did feature engineering to introduce new features that are related to time and location. These include the new time-related features, month, date in month, hour of the day and the hour zone in a day, and also if the day_is_holiday, to better capture the information. We also created new geographic features to divide LA into blocks with size of 5x5km, based on longitude/latitude using the Geohash library.

Even after carefully tuning the model , the prediction performance was still low even with very sophisticated models like random forest, with accuracy of about 0.3. This is not very abnormal based on the literature. In our case, this was mainly due to the large variations of classification labels in the original dataset, of 137 crime types to predict. Therefore, we eventually concentrated the crime type labels down to 46 labels, by combining the original crimes type into a bigger class. For example, all theft types (petty, grand, vehicle, etc.) are grouped into one theft category, all assault and batteries are grouped in one battery category. We were able to dramatically improve prediction performance from previous ~0.3 to ~0.5 to 0.6 after the label concentration and feature engineering by focusing on features we know to be useful from Experiment 1 and 3. While our results were mixed at 0.5, we were delighted to find that relying on features and variables with high information score increased our predictive performance.

Experiment/Evaluation #3: Clustering Discovery and Association Confidence Mining

__Question:__ This experiments seeks to answer which crime type is the best predictor of violent crimes. If we can answer this question we can use more common crime types to predict violent crimes (murder, assault, sex crimes) which usually have more sparse data.

Observation: From computing confidence values we can identify the predictive power of one crime type relative to another. A single trial example can be shown below for our three violent crime prediction targets:

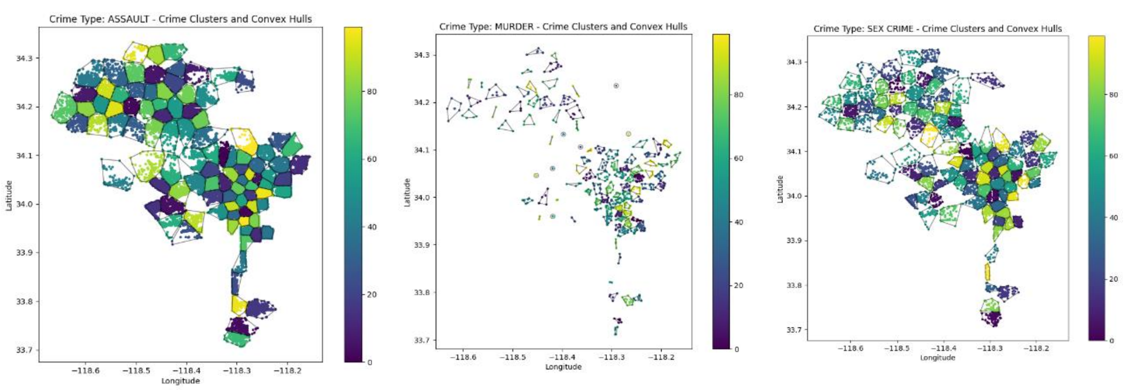


Figure 7: Cluster Regions for: (Left) Assault, (Center) Murder, (Right) Sex Crime

The confidence value can then be computed as the intersections of the target maps divided by the area of the signal variable. This is computed as the predictive confidence of the signal variable towards the target variable. The Confidence results for experiments with various crime associations can be seen below

:

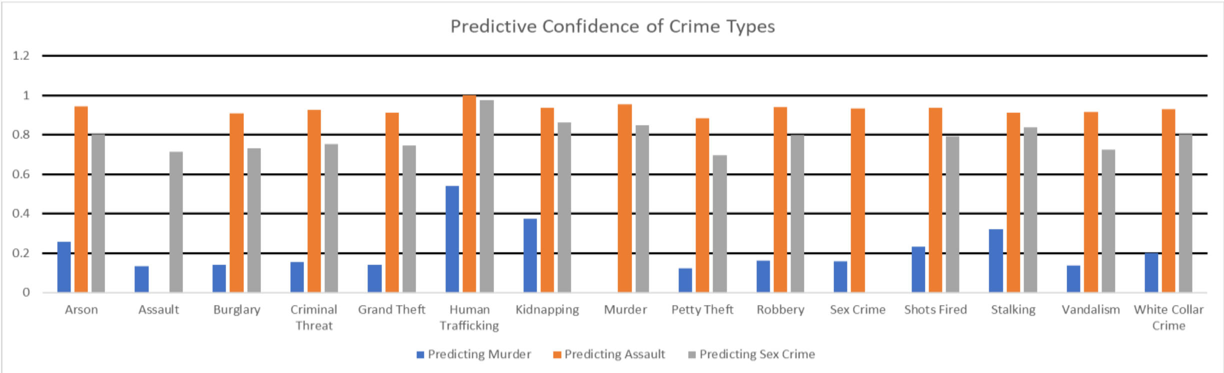


Figure 8: Results of Confidence experimentations attempting crime associations of various types

We achieved successful Confidence levels between various crime types. For example we see the confidence (predictive power) of crimes towards the identification of assault risk being very strong, all above 0.80. We see that "Murder" and "Human Trafficking" are the highest confidence values at 0.97 and 0.99 respectively. This is a great result as it indicates that any region with human trafficking or murder has a >97% risk of assault. This is a reasonable result that matches our understanding of the world. I expect regions with some violent crimes to have a high prevalence of other crimes. We can see similar associations with sex crimes. We find Kidnapping, Stalking, Human Trafficking as some of the best performing signal variables, all scoring above 0.80 in Confidence. Again, this makes rational sense as stalkers and kidnappers would correlate highly with sex crimes. This is a very useful discovery, sex crime data we found to be very sparse since they are not always reported. Finally for the prediction of murder, we see that Human Trafficking and Kidnapping are the best predictors. In general, we can see that Human Trafficking regions have the highest confidence in predicting violent crimes. Using these finding we can suggest to law enforcement analysts and other data learning models that human trafficking crime clusters may be areas with very high likelihood of violent crime, even if the analyst is lacking direct violent crime data.

Conclusion

We introduce a cutting-edge risk scoring system, utilizing a heatmap to help users identify high-risk areas. We identified key victim demographics (women) and geolocations in LA that has the highest risk scores. We also tested various predictive models using standard models and compared the results to models that used our PCA decomposed features (particularly those we found that had higher information scores and confidence scores from our other experiments) and we were able to show an improvement of 20% in our kNN learning models using these techniques. Finally, we computed crime clusters and computed the association confidence between various crime types. We found 'Human Trafficking' and 'Kidnapping' to be the most important predictors of violent crime. This can be very useful to law enforcement and other analysts who might lack violent crime data and can now rely on other common crime types. We were unable to show a higher than 50% prediction score, however a next step may be to utilize additional more data cleaning such as ICA and LDA in order to reduce feature complexity which may improve results further.

We made all these methods and their results accessible to users everywhere with an intuitive and interactive webapp UI. Now, the 3.9 millions citizens of LA can access crime trending and forecasting in their regions in real time and understand the crime risk they might be under when they are traveling in their city. These innovative approaches promise to enhance the field of data-driven crime analysis, fostering a safer and more informed environment for all stakeholders.

Distribution of Team Member Effort

Workload has been evenly distributed among all team members, promoting active involvement and equal contribution to the project's progress. This balanced approach has fostered collaboration, allowing diverse insights and expertise to enhance the project's overall quality and success.

Appendix A

References

[1] Salah, M. M., & Xia, K. (2022). Big Crime Data Analytics and Visualization. *ACM Digital Library*. https://doi.org/10.1145/3523089.3523094
[2] Feng, M. (2019). Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data. *IEEE*. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8768367
[3] Mandalapu, V., Elluri, L., Vyas, P., & Roy, N. (2023). Crime Prediction using Machine Learning and Deep Learning: A Systematic review and Future Directions. IEEE Access, 1. https://doi.org/10.1109/access.2023.3286344
[4] S. Kim, P. Joshi, P. S. Kalsi and P. Taheri, "Crime Analysis Through Machine Learning," 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 2018, pp. 415-420, doi: 10.1109/IEMCON.2018.8614828. https://ieeexplore.ieee.org/document/8614828
[5] Agarwal, J. (2013). Crime Analysis using K-Means Clustering. International Journal of Computer Applications, 0975–8887. https://www.academia.edu/33456150/Crime_Analysis_using_K_Means_Clustering
[6] Estivill-Castro, V. (2001). Data Mining Techniques for Autonomous Exploration of Large Volumes of Geo-referenced Crime Data. ResearchGate. https://www.researchgate.net/publication/228748459_Data_mining_techniques_for_autonomous_exploration_of_large_volumes_of_geo-referenced_crime_data
[7] L. Elluri, V. Mandalapu and N. Roy, "Developing Machine Learning Based Predictive Models for Smart Policing," 2019 IEEE International Conference on Smart Computing (SMARTCOMP), Washington, DC, USA, 2019, pp. 198-204, doi: 10.1109/SMARTCOMP.2019.00053. https://ieeexplore.ieee.org/document/8784006
[8] Das, A. K., & Das, P. (2022). Graph based ensemble classification for crime report prediction. *Applied Soft Computing*, *125*, 109215. https://doi.org/10.1016/j.asoc.2022.109215
[9] McClendon, L. (2015). Using Machine Learning Algorithms to Analyze Crime Data. *Machine Learning and Applications: An International Journal (MLAIJ)*, *2*(1). https://www.researchgate.net/profile/Natarajan-Meghanathan/publication/275220711_Using_Machine_Learning_Algorithms_to_Analyze_Crime_Data/links/571dc8ae08ae408367be5de8/Using-Machine-Learning-Algorithms-to-Analyze-Crime-Data.pdf
[10] Ara´ujo, A., Cacho, N., Bezerra, L., Vieira, C. & Borges, J. Towards a crime hotspot detection framework for patrol planning, 1256–1263 (IEEE, 2018). https://ieeexplore.ieee.org/document/8622949
[11] Almuhanna, A. A., Alrehili, M. M., Alsubhi, S. H. & Syed, L. Prediction of crime in neighbourhoods of new york city using spatial data analysis, 23–30 (IEEE, 2021). https://ieeexplore.ieee.org/document/9425120
[12] Yan, Z., Chen, H., Dong, X., Zhou, K. & Xu, Z. Research on prediction of multi-class theft crimes by an optimized decomposition and fusion method based on xgboost. Expert Systems with Applications 207, 117943 (2022). https://dl.acm.org/doi/abs/10.1016/j.eswa.2022.117943
[13] Han, X., Hu, X., Wu, H., Shen, B. & Wu, J. Risk prediction of theft crimes in urban communities: an integrated model of lstm and st-gcn. IEEE Access 8, 217222–217230 (2020). https://ieeexplore.ieee.org/document/9276416
[14] Yao, S. et al. Prediction of crime hotspots based on spatial factors of random forest, 811–815 (IEEE, 2020). https://ieeexplore.ieee.org/document/9201899
[15] Hossain, S., Abtahee, A., Kashem, I., Hoque, M. M. & Sarker, I. H. Crime prediction using spatio-temporal data, 277–289 (Springer, 2020). https://arxiv.org/abs/2003.09322
[16] Ratcliffe, J. H., & Rengert, G. F. (2008). Near-repeat patterns in Philadelphia shootings. Security

Journal, 21, 58–76. https://www.researchgate.net/publication/247478132_Near-Repeat_Patterns_in_Philadelphia_Shooting
[17] Liu, W., Xu, C., Peng, Y., & Xu, X. (2023, June 16). *Evolution of tourism risk communication: A bibliometric analysis and meta-analysis of the antecedents of communicating risk to tourists*. MDPI. https://www.mdpi.com/2071-1050/15/12/9693
[18] Pina-Sanchez, J., Brunton-Smith, I., Buil-Gil, D., & Cernat, A. (2023, July 25). *Exploring the impact of measurement error in police recorded crime rates through sensitivity analysis - crime science*. BioMed Central. https://crimesciencejournal.biomedcentral.com/articles/10.1186/s40163-023-00192-5
[19] Buil-Gil, D., & I. Mawby, R. (2022, July 30). Do tourists report crime to the police? An exploratory analysis in Barcelona. https://www.tandfonline.com/doi/full/10.1080/13683500.2022.2105198
[20] Comer, Benjamin P., Cody Jorgensen, and David Carter. "Reported crime frequencies: a statistical comparison of state crime reports and the UCR." American journal of criminal justice 48.1 (2023): 151-175. https://link.springer.com/article/10.1007/s12103-021-09623-y
[21] Hart, Timothy C. "Hot spots of crime: Methods and predictive analytics." *Geographies of Behavioural Health, Crime, and Disorder: The Intersection of Social Problems and Place* (2020): 87-103. https://link.springer.com/chapter/10.1007/978-3-030-33467-3_5#:~:text=Crime%20hot%20spot%20mapping%20allows,and%20management%20of%20police%20operations.
[22] Oatley, Giles C. "Themes in data mining, big data, and crime analytics." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 12.2 (2022): e1432. https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1432
[23] Zhiwei Li, Wang, Mingjing Li, Wei-Ying. A Probabilistic Model for Retrospective News Event Detection. Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval; 2005. p.106-113. https://dl.acm.org/doi/10.1145/1076034.1076055
[24] Gionis, A. (2005). Clustering Aggregation. *IEEE*. https://i11www.iti.kit.edu/_media/teaching/winter2006/graphclustering/gmt-ca-05.pdf
[25] Townsley, M., Homel, R., & Chaseling, J. (2003). Infectious burglaries: A test of the near repeat hypothesis. The British Journal of Criminology, 43(1), 615–633. https://www.jstor.org/stable/23639045?typeAccessWorkflow=login
[26] Levine, N. (2008). The "hottest" part of a hotspot: Comments on "The utility of hotspot mapping for predicting spatial patters of crime". Security Journal, 21, 295–302. https://link.springer.com/article/10.1057/sj.2008.5
[27] Caplan, J. M., Kennedy, L. W., & Miller, J. (2011). Risk terrain modeling: Brokering criminological theory and GIS methods for crime forecasting. Justice Quarterly, 28(2), 360–381. https://www.tandfonline.com/doi/abs/10.1080/07418825.2010.486037