UNIVERSITY SCHOOL OF INFORMATION COMMUNICATION AND TECHNOLOGY



**TERM PAPER REPORT (IT-669)**

**on**

**Crime Prediction Using Data Mining**

**Mentor:**

Dr. Reena Gupta

**Submitted By:**

Hitesh Sharma

MCA SE (2022-2024)

Enrollment Number: 01416404522

**ACKNOWLEDGEMENT**

With candour and pleasure, I take the opportunity to express my sincere thanks and obligation to my esteemed guide Dr.Reena Gupta. It is because of her able and mature guidance and co-operation, without which it would not have been possible for me to complete my project. It is my pleasant duty to thank all the staff members of the computer Centre who never hesitated to help me from time to time during the project.

**Introduction**

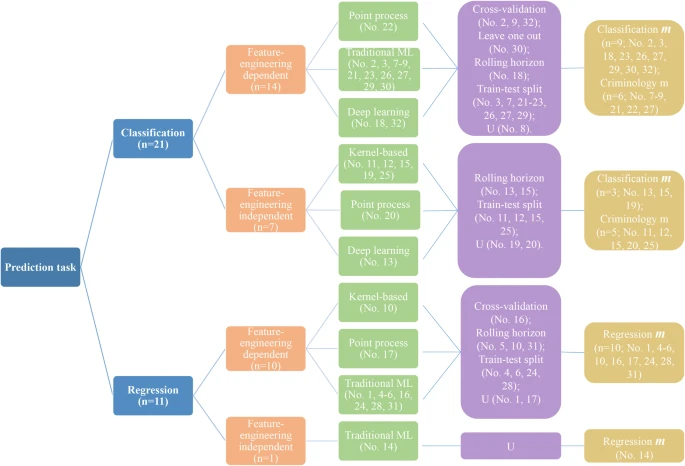
* The safeguarding of citizens and the security of the state is one of the most crucial jobs of the government. With the recent, quite promising developments in the domain of Data Analysis it is now possible to finally use the mountains of data collected by the authorities to find hotspots and then develop relevant strategies. The absolute aim of my term paper is to identify and exploit the underlying patterns and similarities between crimes. Patterns when fed into an algorithm can help in predictions of future crime. This approach greatly saves the manhours put in by authorities, so they can direct their resources elsewhere.
* Applications of Big Data, AI/ML, and data mining are showing great promise when it comes to data analyzing. Taking advantage of these we can devise methods to generalize the data, group the similar type together, and cluster the ones with similar attributes. Coupled with the sophisticated hardware we can use these techniques to improve the surveillance and prepare for forthcoming events in advance. Approaches like entity extraction, clustering, association rule mining, decision trees, support vector machines, naive Bayes rule, neural networks and social network analysis amongst others will put to use.

**Name of the Paper - A systematic review on spatial crime forecasting**

The hotspots (also known as binary classification) approach is the most common kind of predicting inference found. The majority of techniques utilized in this study were traditional machine learning techniques, along with kernel density estimation-based approaches and, less frequently, point process and deep learning techniques. Prediction Accuracy, Prediction Accuracy Index, and F1-Score rank as the best indicators of evaluating performance.

We distinguished five different categories of inference: (1) hotspots (which accounted for the majority of studies), (2) number of crimes, (3) crime rate, (4) crime category, (5) percentage of crimes in clusters, and (6) cluster characteristics. The authors suggested a mix of conventional machine learning techniques, kernel density estimation-based approaches, and less frequently point process and deep learning techniques for forecasting. There are several measures that can be used to evaluate performance, however, the Prediction Accuracy Index, and F1-Score are the top ones. Last but not least, the train-test split was the most used validation strategy.

The integration of spatial or spatiotemporal information into conventional prediction algorithms is one of the most promising areas of current research. Surveillance plots (Fig. Below) give a more thorough picture of the precision of the projected data with regard to the performance evaluation. Police can use them as a decision-making tool to determine the threshold that balances prediction accuracy with the size of patrol zones because they include the area coverage on the x-axis.



**Shortcomings**

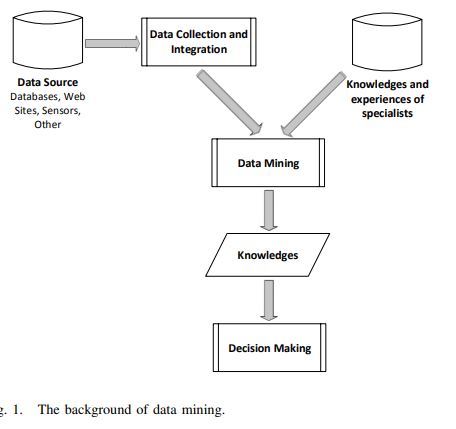
Overall, key elements of study experiments are not consistently disclosed, and the spatial unit of analysis and sample size are frequently undescribed things. Similarly, the crafting procedures for methods relying on feature engineering are not sufficiently documented. The aforementioned factors make it challenging to replicate a study or compare its findings to those of a potential future study.

**In the following pages we have made a rough analysis of the data of the respective research paper. We tried to describe and understand the data through various visualization analysis and different plotting schemes.**

**Name of the Paper - A Survey of Data Mining Techniques for Analysing Crime Patterns**

This paper provides an overview of studies on the many applications of data mining as well as recommendations for employing these approaches to solve crimes. Additionally, it addresses the issues and inadequacies in the field of crime data mining.

Data mining is the process of analysing historical data to discover patterns, trends, and information. The first crucial elements for investigation in order to uncover the concealed knowledges are as follows: 1) The data must be accurate and sufficient to be used for analysis. 2) The expertise and knowledge of specialists. In order to help with decision-making and problem-solving, data mining techniques produce knowledge results.



A. Association Rule Mining

This type of unsupervised learning is used to uncover concealed insights in unlabeled data. Additionally, association rules can find relevant object co-occurrences in huge data sets. In the fundamental association rule, there are two elements to the rule. 1) The antecedent, also known as the left side or hand side (LHS). 2) The ensuing, often known as the right side or right hand side (RHS). LHS RHS is a type of generic association rule, where LHS and RHS are separate item-sets. The RHS item-set is likely to occur if the LHS item-set does. The crucial statistical parameters of support and confidence should be combined for the effective finding of association rules. Such metrics have a value between 0 and 1. A rule on associations is likely to be uninteresting if it has very little support. In order to evaluate the dependability of association rules, one uses the confidence measure. The higher the confidence level for a particular rule A B in a transaction set T, the more likely it is that B will be present in T that contains A. Support and confidence results should exceed user-specified thresholds in order to find co-occurrences between two data sets.

B: Clustering

A method of unsupervised data analysis is clustering. Using this method, similar data are divided into the same group while different data are divided into the other group. There are numerous algorithms, including Expectation-Minimization, Hierarchical Clustering, and K-means.

K-means Algorithm: The user first enters the k centroids number. The desired clusters are indicated by the letter K. Every cluster needs a centroid, which is the average value for the cluster. The closest centroid is then given to each data record. The centroid of each cluster is updated by computing the mean cluster once all input data records have been assigned. The assignment and improving the centroids will be repeated until the most recent centroids do not change.



C: Classification

This approach of supervised learning is used to classify items into one of numerous pre-established categories. Several issues with numerous applications have seen extensive use of the categorization techniques.

Decision Tree: The root node, internal nodes, and leaf or terminal nodes make up the majority of the structure. The root node, which serves as the top of the tree and is utilised to split records in the data set, is selected from an important property. The qualities that are test conditions are contained in the internal nodes. Every internal node has unique qualities. The class label or prediction outcome is represented by the leaf or terminal nodes. In the decision-making process, the test condition is compared before the root node is taken into account. The test's results specify the subsequent appropriate branch to take into account the following internal nodes. Up until a leaf node is reached, the internal nodes are slowly taken into account. The record is given the class label or the predicted outcome.

From one of the references of the above paper, we tracked down a dataset and ran analysis over it. To understand the underlying data and the rough pattern before we dive deeper into classification, clustering or association.

**References**

[1] J. Wang, J. Hu, S. Shen, J. Zhuang and S. Ni, "Crime risk analysis through big data algorithm with urban metrics", Phys. A Stat. Mech. Appl., vol. 545, May 2020.

[2] F. Yi, Z. Yu, F. Zhuang, X. Zhang and H. Xiong, "An integrated model for crime prediction using temporal and spatial factors", Proc. IEEE Int. Conf. Data Mining (ICDM), pp. 1386-1391, Nov. 2018.

[3] R. K. Wortley and L. A. Mazerolle, Environmental Criminology and Crime Analysis, vol. 6, 2016.

[4] A. Belesiotis, G. Papadakis and D. Skoutas, "Analyzing and predicting spatial crime distribution using crowdsourced and open data", ACM Trans. Spatial Algorithms Syst., vol. 3, no. 4, pp. 12, 2018.

[5] N. H. M. Shamsuddin, N. A. Ali and R. Alwee, "An overview on crime prediction methods", Proc. 6th ICT Int. Student Project Conf. (ICT-ISPC), pp. 1-5, May 2017.

[6] H. B. F. David and A. Suruliandi, "Survey on crime analysis and prediction using data mining techniques", ICTACT J. Soft Comput., vol. 7, no. 3, pp. 1459-1466, Apr. 2017.

[7] C. Chauhan and S. Sehgal, "A review: Crime analysis using data mining techniques and algorithms", Proc. Int. Conf. Comput. Commun. Autom. (ICCCA), pp. 21-25, May 2017

[8] S. Prabakaran and S. Mitra, "Survey of analysis of crime detection techniques using data mining and machine learning", J. Phys. Conf. Ser., vol. 1000, no. 1, Apr. 2018.

[9] H. Hassani, X. Huang, E. S. Silva and M. Ghodsi, "A review of data mining applications in crime", Stat. Anal. Data Mining ASA Data Sci. J., vol. 9, no. 3, pp. 139-154, Jun. 2016.

[10] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering", 2007.

[11] F. Weidt and R. Silva, "Systematic literature review in computer science-a practical guide", vol. 1, 2016.

[12] U. Thongsatapornwatana, "A survey of data mining techniques for analyzing crime patterns", Proc. 2nd Asian Conf. Defence Technol. (ACDT), pp. 123-128, Jan. 2016.