Machine Learning- Based Criminal Activities Analysis System

Parveen Shaheen[1], Anuradha Deolase[2]
[1]Dept of Computer Science
[2]Assistant Professor, Dept of Computer Science
[1, 2] VITM Indore M.P

Abstract- *Crimes hurt any society both socially and economically. Law enforcement bodies face numerous challenges while trying to prevent crimes. We propose a Machine learning-based criminal activities analysis system (MLBCAAS) to assist law enforcement bodies to perform descriptive, predictive, and prescriptive analysis on crime data. MLBCAAS has a modular architecture where each component is built separately from the other. MLBCAAS also supports plugins enabling future feature expansions. The platform can ingest any crime dataset with the required attributes to map the dataset to attributes required by the platform. It can then analyze them, train models, and then visualize data. MLBCAAS also combines census data with crime data to achieve a more comprehensive crime analysis and its impact on society. Moreover, with the combination of census data and crime data, MLBCAAS provides process-reengineering steps to optimize resource allocations of police forces. In this dissertation work, I am going to implement a Machine learning-based criminal activities analysis system. I will analyze two different crime datasets for any two cities and provide a comparison between the two datasets through statistical analysis. I will use MySQL database (phpMyAdmin) to store datasets and PHP as a frontend while using wamp as a server to the application. The primary purpose of the dissertation is to produce interesting frequent patterns for criminal hotspots using the Apriori algorithm. In addition, I will use a Decision Tree and a Naïve Bayesian classifier to predict potential crime types. The result of this can be used to raise people's awareness regarding the dangerous locations and to help agencies to predict future crime in a specific location within a particular time.*

Keywords- Component, formatting, style, styling, insert

I. INTRODUCTION

Crimes are a social nuisance, and it has a direct effect on society. Governments spend lots of money through law enforcement agencies to try and stop crimes from taking place. Today, many law enforcement bodies have large volumes of crime data, which need to be processed to turn into useful information.

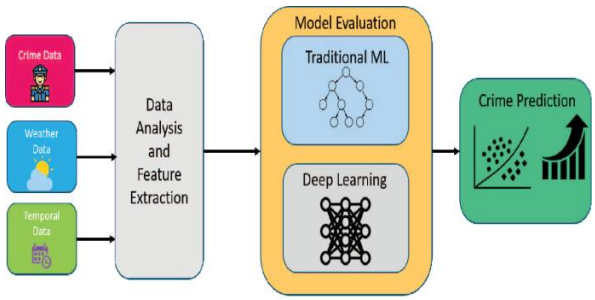

Fig 1.1 Data Flow Model

Crime data are complex because they have many dimensions and in different formats, e.g., most of them contain string and narrative records. Due to this diversity, it is not easy to mine them using statistical and machine learning data analytics tools off the shelf. It is the primary reason for the lack of a general platform for crime data mining. While there are some propitiatory platforms to predict and analyze crime data, they are focused only on specific areas of crimes, not extensible. They do not provide an API to integrate with other tools. Moreover, the same tool cannot be used for the analysis and planning, such as patrol beads and district boundaries.

Motivation:

High or increased crime levels make communities decline, as crimes reduce house prices, neighbourhood satisfaction, and the desire to move negatively. To reduce and prevent crimes, it is essential to identify the reasons behind crimes, predict crimes, and prescribe solutions. Due to large volumes of data and the number of algorithms needed to be applied to crime data, it is unrealistic to do manual analysis. Therefore, it is necessary to have a platform capable of applying any algorithm required to do a descriptive, predictive, and prescriptive analysis on a large volume of crime data. Through those three methodologies, law-enforcement authorities will be able to take suitable actions to prevent the crimes.

Moreover, by predicting the highly likely targets to be attacked during a specific time and geographical location,

police will identify better ways to deploy the limited resources and find and fix the problems leading to crimes. Several applications are already developed for crime analysis. These tools are developed to help the police identify different crime patterns and even predict criminal activities. They are complex software which needs much training before use. Designing a tool that is easy to use with minimal training would help law-enforcing bodies worldwide reduce crimes.

Problem Statement

How to develop a software platform to conduct descriptive, predictive, and prescriptive analysis of diverse crime data and place? We aim to find the criminal hotspots using a hypothetical dataset of the city of Indore. We will try to find the type of crime in a particular area and its chances. The algorithms we will be using for the project are A priori, Decision Tree and Naïve Bayes classifier.

Research Objective

- Develop a platform that can be used to analyze crime data using descriptive and predictive data analytics techniques.
- The proposed platform analyses the spatial and temporal (time of day, day of week, and seasons) relationships in crime data. • Suggest appropriate process reengineering steps and resource allocations based on the spatial and temporal relationships.
- Analyze the relationship between crime data and census data.

II. LITERATURE SURVEY

There have been incalculable workdone regarding crime. Extensive datasets have been reviewed, and data such as area and crime have been analyzed to enable individuals to pursue law authorizations. Existing techniques have utilized these databases to recognize crime hotspots depending on the area. Although crime areas have been recognized, there is no data accessible that incorporates the crime event date and time alongside systems that can precisely anticipate what crimes will happen later on. We analyzed some of the past work and the research papers regarding crime prediction. Below is a brief description of some of the past works.

1) Crime effects on Society

A crime can be defined as any action or omission that violates a law, which results in a punishment. Usually, what constitutes a crime depends on the government bodies and laws that are in existence in those places. To understand the nature of crimes, one has to understand its Spatio-temporal dimensions and the nature of the crime, the victim-offender relationship, the role of guardians, and the history of similar incidents [8]. Regardless of why crimes occur, they put a strain on the communities, towns, and cities. Usual monetary costs associated with them include the cost of policing crime and prosecuting those who commit crimes. Non-monetary costs consist of social costs, affecting the quality of life, mental health, and physical security of people living in those areas. Crimes are a social nuisance, and solving them faster is very important and will pay for itself.

2) Criminology Theories

According to John and David, theories of crimes can be divided into two categories: those that seek to explain the development of criminal offenders and those that seek to explain the development of criminal events.

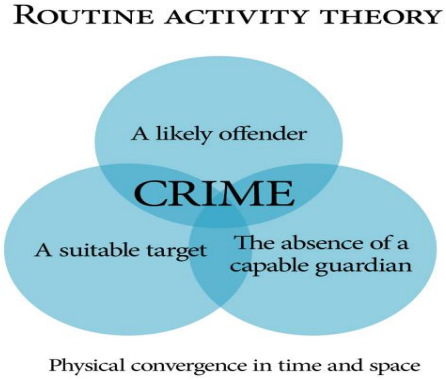

Fig 2.1 Routine activity Theory

Criminology has been mainly developed through theories and research on offenders. Only recently, it has begun to explain the crimes rather than the criminality of people involved in them. Criminology consists of many theories that explain how and why some offenders act in the way they do. Following are some theories that explain how places are associated with crimes.

1. Rational Choice suggests that offenders will select targets and define means to achieve their goals in a manner that can be explained. Further, it can be explained that human actions are based on rational decisions; probable consequences inform them of that action [9].

2. Routine Activity Theory This theory explains the occurrence of crimes as the result of several circumstances. Namely, a motivated offender, a desirable target, target and offender must be at the same place simultaneously, and lastly,

absent of other types of controller's intimate handlers, guardians, and place managers [9].

3. Crime Pattern Theory This theory combines the above two theories. It says that how targets come to offenders' attention is influenced by the distribution of crime events over time, space, and among targets. An offender will come to know of criminal opportunities while engaging in their legitimate day-to-day work. So, a given offender will only know about a subset of available targets.

III. RESEARCH & METHODOLOGY

3.1 Solution

I have completed research like collecting data, what algorithms I will use, how and what data is required to filter. My proposed work will try to solve all limitations and complexities of older systems. In my present report, I have included all points so that anyone can understand my proposed work. My work will try my best to answer and meet all the above-listed points. Now onward, the only task left is to implement it through programming languages to demonstrate my work practically. I am going to implement it in PHP with MySQL as a database. After implementation, it could be run through the internet/any private network. A training session is also required. Any user or a law enforcement body who has a crime data set can use this feature to understand the severity of the crime. As a result, could take the necessary steps to allocate resources effectively. The following sections describe the architecture and different modules implemented in the solution.

3.2 Features provided by the platform

The MLBCAAS provides the following features.

1. Redraw efficient police jurisdiction boundaries.
2. Query data in the crime dataset.
3. Predict crime categories for a given crime scene.
4. Preprocess the uploaded crime dataset

3.3 Architecture

Figure 8 depicts the high-level architecture of the proposed Inquisitors Machine learning-based criminal activities analysis system (MLBCAAS). Data Receiver is used to provide data to the platform, and the data persistence unit is used to store data used by the system and trained models. The preprocessor is used to preprocess raw data received by the Data Receiver, which feeds into the next layer. It consists of two core modules, Statistical Analyzer and Machine Learner. Using those two modules, the Descriptive analyzer, Prescriptive analyzer, and Predictive analyzer have been implemented. Also, a few other components provide a vital contribution to the MLBCAAS. The functionalities provided by prescriptive, descriptive and predictive analyzers are exposed to the user through an API. The results provided by MLBCAAS according to the user's requests can be visualized using the Visualizer component.

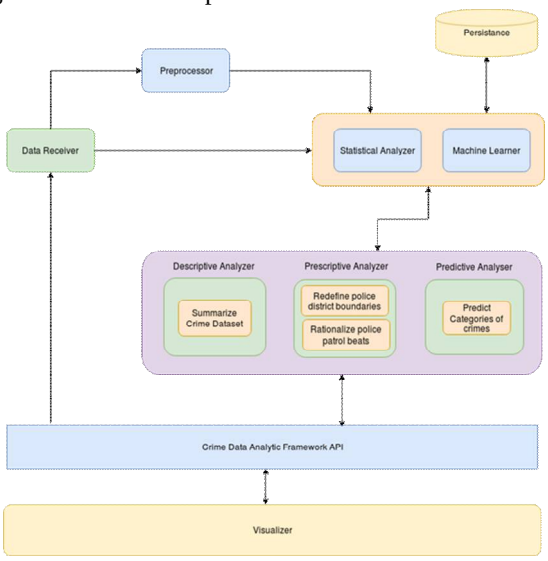

Fig.3.1. High-level architecture of the proposed platform.

3.4 Design Documents

1)Use Case View

The crime data Analysis platform has a design to handle specific use cases. Figure 9 shows use cases of the MLBCAAS.

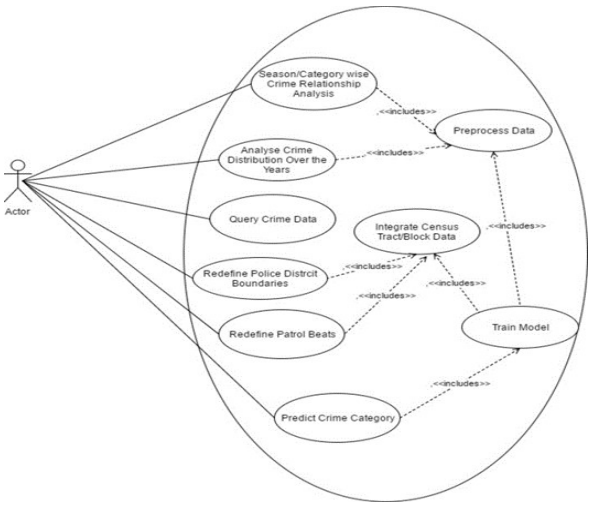

Fig. 3.2. A use case diagram for MLBCAAS

2)Development view

Figure 10 explain the whole system from the developer's perspective. Six main components are loosely coupled with each other.

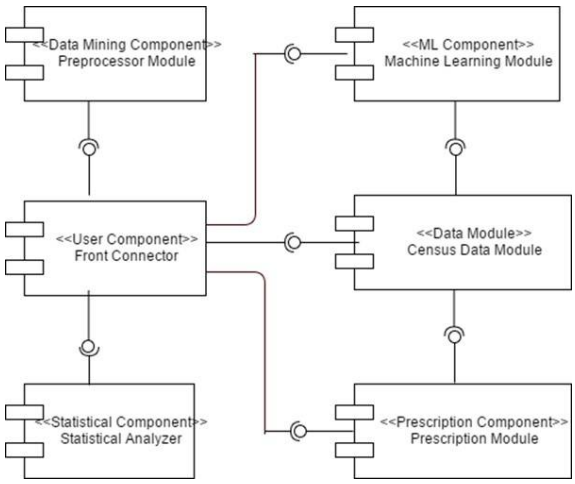

Fig.3.3. A use case diagram for MLBCAAS

3)Preprocessor module

Preprocessor Module consists of the features and functionalities needed to preprocess the data before feeding them into other modules such as Machine Learner and Statistical Analyzer. Within this module, functionalities needed to preprocess the input data have been implemented.

4)Front connector module

The front connector provides API to the user by hiding the complexity of the Crime Data Analytic Framework. Users can feed data into the framework and get results back using the front connector module.

5)Statistical Analyzer module

The statistical analyzer provides basic statistics of the data set the feed to the framework. It provides API to get Simple statistics like mean, variance median, column statistics to complex statistics like F.P. growth Algorithm based frequentitemsets. The Statistical Analyzer module provides an API to perform fundamental statistical analysis on the crime dataset provided by the use.

6)Machine learning module

The Machine Learning component is built on top of Apache Spark MlLib and created to hide the complexity of spark Machine learning algorithms.

7)Census Data module

This module keeps various kinds of census data like census block data, census tract data race data.

3.5 Data Descriptions

This platform is based on the Apache Spark engine. Therefore, within the framework inside the platform, data will be kept and processed as data structures used by Apache Spark such as Data Frame, Dataset. However, since the user should not know how to use spark, those spark data structures have been hidden within some new custom data structures, enabling a user to use the platform easily without knowledge about Apache Spark. Basic data flow within the platform is shown in Figure 11.

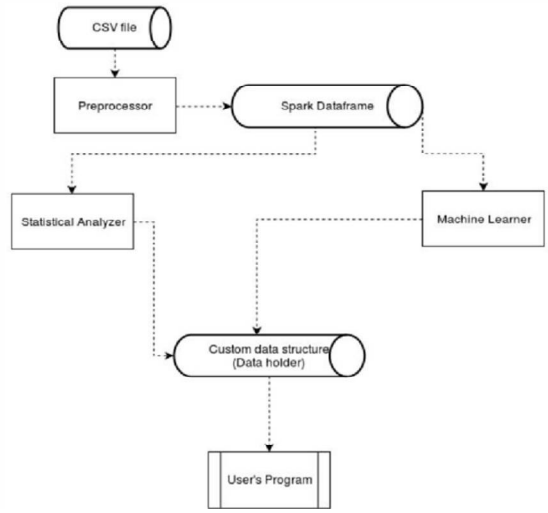

Fig.3.4. Data flow within MLBCAAS

3.6 Implementation

We implemented a web application to visualize the project output to the user. This web application provides a user with access to all the implemented features of the platform, including descriptive, predictive, and prescriptive analytics.

1) Web Application

Machine learning-based criminal activities analysis system performs descriptive, prescriptive, and predictive analysis of crime data and provides user-friendly and effective ways to analyze crime data. A web application has been developed to interact with the user to make crime data analysis feasible and practical. This has been developed using PHP along with MySQL MVC.

Crime Data Analytics Platform provides visualization using a web application built using MySQL framework and PHP. Reasons for selecting MySQL framework for our platform development are as follows:

- MySQL provides an apparent contrast between controllers, JavaBean models, and views.
- MySQL is very flexible. It is entirely based on interfaces.
- MySQL is view-agnostic. It does not push a developer to use D.B. and can use alternative view technologies.

Since we are using Apache spark java, we had to use a framework compatible with java platforms. Therefore, we used MySQL for the implementation of the web application. Reasons to use PHP along with MySQL framework.

- Directives in PHP bring additional functionalities to HTML.
- PHP Material provides already developed U.I. components to be used.
- Provides two-way data binding.
- Provides client-side MVC framework.
- Dependency injection

2)Features provided by the platform

- Upload a crime dataset.
- Preprocess the uploaded crime dataset.
- Discover spatial and temporal patterns of crime data using visualization features.
- Query data in the crime dataset.
- Redraw efficient police jurisdiction boundaries
- Draw efficient police patrol beats based on the crime distribution.
- Predict crime categories for a given crime scene.

IV. RESULTS & EVALUATION

While comparing all the algorithms, we learned from this project that they are not entirely reliable while the algorithms used are practical. First, we applied the apriori algorithm to find the frequent crime patterns meaning we analyzed the database to get the possibilities of a crime happening. However, they are not 100% reliable since one example shows that the chance of vehicle theft in waghodiya is a hundred per cent which is impossible to happen. So, we cannot wholly rely on the result, which gives 100% possibility. Next, we use a Decision tree as a searching algorithm that helps us find names from a large-scale database.

Finally, we use the naïve Bayes classifier to find the type of crime that can happen in a particular area at a particular time. The naïve Bayes classifier only gives the type of crime and not the probability of it happening, so we can only use it to find the crime type. Another disadvantage of using a naïve Bayes classifier in this project is that if a location has registered a crime of the same number, the result will only show the most recent crime as output. By comparing the output of all the algorithms, we found that the most efficient and accurate algorithm used is the decision tree, followed by naïve Bayes and Apriori. The efficiency of each algorithm depends on the type of application it is used in.

Accuracy: Overall correctness of the model and is calculated as the sum of correct classifications divided by the total number of classifications.

Precision: Given all the predicted labels (as given class X), how many instances were correctly predicted.

Recall (Sensitivity): For all instances that should have been labelled X, how many of these were correctly captured.

Table 4.1 shows the confusion matrix Of Naive Bayesian classification. The confusion matrix for Random Forest with 500 trees is shown in Table 4.2, while the confusion matrix for Multilayer Perceptron neural network with 500 perceptrons is shown in Table.

TABLE 4.1.CONFUSION MATRIX OF NAIVE BAYESIAN CLASSIFICATION

	Other offences (Actual)	Larceny/ Theft (Actual)	Vehicle Thefts (Actual)	Non criminal (Actual)	Drug narcotic (Actual)
Other offences (Predicted)	56232.0	24415.0	0	0	361.0
Larceny/ Theft (Predicted)	12769.0	39019.0	0	0	483.0
Vehicle Thefts (Predicted)	11093.0	16418.0	0	0	299.0
Non-criminal	12388.0	10571.0	0	0	95.0
(Predicted)					
Drug narcotic (Predicted)	14053.0	1998.0	0	0	9.0

TABLE 4.2. CONFUSION MATRIX OF RANDOM FOREST CLASSIFICATION

	Other offences (Actual)	Larceny/ Theft (Actual)	Vehicle Thefts (Actual)	Non criminal (Actual)	Drug narcotic (Actual)
Other offences (Predicted)	38423	13637	280	1840	19
Larceny/ Theft (Predicted)	8138	26624	75	54	8
Vehicle Thefts (Predicted)	4913	10686	25554	89	46
Non Criminal (Predicted)	8581	6247	124	362	3
Drug narcotic (Predicted)	7270	714	12	2671	0

Therefore, if the Naive Bayesian algorithm is selected only by looking at accuracy, the possibility of predicting important class labels such as DRUG/NARCOTIC and LARCENY/THEFT would be 0. That would be not suitable for our application. Because whenever our application predicts a theft, it would become a wrong prediction. Therefore, we have chosen the Random Forest algorithm despite the higher accuracy considering individual labels' higher precision and recall values. If we consider the Random Forest algorithm and Neural Network algorithm, the accuracy of Random Forest (0.40) is higher than the accuracy of Neural Network algorithm (0.32), and precision and recall of individual labels of Random Forest are also higher than Neural Network. We have chosen the Random Forest algorithm considering individual labels' higher accuracy, precision, and recall values.

V. CONCLUSION

In today's time where crime is increasing day by day, a significant challenge faced by law enforcement is predicting crime to protect the citizens. As data science and technology progress, tools of data mining and A.I. are now accessible to the law enforcement community. Computers can process a great many directions in seconds, sparing valuable time. Computers are likewise less prone to blunders than human investigators.

The project focuses on developing a crime prediction analysis tool for local Society using data science and technology. This project enables the law department to characterize and analyze the crime data to identify crime patterns and predict possible future crimes. The system can be a small prototype, but it can still be implemented in the real-world database.

For the future extension of the project, we can apply more classification algorithms to increase the crime prediction accuracy and enhance the system's overall performance.

VI. FUTURE WORK

As the data size and the covering geographical area increase, the solution provided by the Machine learning-based criminal activities analysis system needs more computation power. So, it needs various parallelized and distributed system techniques. The system currently uses crime data, census block data, census tract data, population data and race data to do data mining, prediction, police district boundary generation and patrol beats generation. However, the system can be extended to integrate other relevant data like offender residence, serial killer’s data.

Also, the proposed algorithm can be further improved to consider the prioritization of calls for services. The proposed algorithm can also be improved to provide police patrol routes rather than an optimal location for a police car. This can be done by considering different periods, seasons, and special occasions like New Year Eve. Furthermore, the GIS data can be integrated with the crime data set using a proper GIS plan. In this way, we can significantly improve the precision and the recall of the prediction model trained by the MLBCAAS.

The visualizer's capability can be significantly improved by integrating more visualizing models such as various graph types. Also, critical data from the backend to the front end can be optimized further. It will improve the user experience significantly; providing an interactive guide will make anyone with or without domain knowledge use the platform.

REFERENCES

[1] Almanie, T., Mirza, R., &Lor, E. (2015). Crime prediction is based on crime types and using spatial and temporal criminal hotspots—*arXiv preprint arXiv:1508.02050*.

[2] Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., &Pentland, A. (2014, November). Once upon a crime: towards crime prediction from demographics and mobile

data. In *Proceedings of the 16th international conference on multimodal interaction* (pp. 427-434).ACM.

[3] Iqbal, R., Murad, M. A. A., Mustapha, A., Panahy, P. H. S., &Khanahmadliravi, N. (2013). An experimental study of classification algorithms for crime prediction.*Indian Journal of Science and Technology*, *6*(3), 4219-4225.

[4] Nasridinov, A., Ihm, S. Y., & Park, Y. H. (2013). A decision tree-based classification model for crime prediction.In *Information Technology Convergence* (pp. 531-538).Springer, Dordrecht.

[5] Liao, R., Wang, X., Li, L., & Qin, Z. (2010, July).A novel serial crime prediction model based on Bayesian learning theory.In *Machine Learning and Cybernetics (ICMLC), 2010 International Conference on* (Vol. 4, pp. 1757-1762).IEEE.

[6] Yu, C. H., Ward, M. W., Morabito, M., & Ding, W. (2011, December). Crime forecasting using data mining techniques.In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on* (pp. 779-786).IEEE.

[7] Laalmanac.com, 'City of Los Angeles Planning Areas Map', 2015. [Online]. Available: http://www.laalmanac.com/LA/lamap3.htm. [Accessed: 20- May- 2015].

[8] Varunon9, naïve-bayes-classifier, (2018), GithubGist,https://github.com/varunon9/naive-bayes-classifier.

[9] Dave Smith, PHP Apriori Algorithm Data Miner,(2015),https://www.phpclasses.org/browse/file/61953.html.

[10] VTwo-Group, Apriori-Algorithm (2014, November), https://github.com/VTwo-Group/AprioriAlgorithm.

[11] Julian Finkler, PHP Decision Tree Classifier: Compose decision trees and evaluate subjects, (2017, July),https://www.phpclasses.org/package/10385- PHP-Compose-decision-trees-and-evaluate-subjects.html.