IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Crime Rate Analysis and Prediction System

[1]Pranav KM, [2]P Avinash, [3]Prashanth KG, [4]Manojkumar CK, [5]Asha MS

[1,2,3,4]UG - Student, Department of Computer Science Engineering, DSATM, Bangalore

[5]Asst. Prof, Department of Computer Science Engineering, DSATM, Bangalore

Abstract: Crime analysis and prevention is a methodical methodology for locating and examining criminal behaviour trends. Our algorithm can identify places with a high likelihood of crime and may depict crime-prone zones, which are regions with a high level of crime. As technology in computerised systems advances, crime data analysts may be able to assist law enforcement officials in hastening the investigation of crimes. The main goals of this research are to categorise different types of crime using clustering approaches that are based on frequency and regulations. The crime data in this project is analysed using a clustering approach, and the K-Means algorithm is employed to cluster the recorded data. Following categorization and clustering, we may forecast a crime using the previous incidents' historical data. The suggested technique can identify places with a greater likelihood of crime and can discriminate between areas with a higher crime rate.

Index Terms **- Crime Analysis, data mining, clustering, prediction.**

I. INTRODUCTION

Categorizing and resolving the increasing volume of criminal cases has become a formidable challenge. Halting criminal activity requires an awareness of its patterns within a particular region. Crime-solving entities can enhance their effectiveness by comprehending the trends in criminal activity within a given area. Utilizing diverse algorithms, machine learning can detect patterns in criminal activity in a specific location. This study leverages crime data collection to predict the types of crimes that are likely to occur in a particular area, expediting the classification of criminal cases and the ensuing actions required.

The topic of crime is crucial to numerous public safety and protection issues, and gaining a better understanding of it can yield many benefits. For instance, it can encourage law enforcement to employ targeted and sensitive tactics to decrease crime, and foster stronger collaboration among individuals and government agencies to create safer communities. The study of crime patterns using data has become a burgeoning and ongoing research field, thanks to the Big Data era and the availability of fast and efficient algorithms for data analysis. Crime causes significant harm to communities and imposes significant costs. The ability to extract essential data and reveal previously undiscovered criminal patterns from a vast crime database is known as "crime data mining."

Researching crime models and making forecasts can help stop the impact of crime in different areas, for which various tools for data generation and acquisition are available. This study offers a comprehensive examination of all criminal data mining methodologies, and proposes several classification methods based on necessary criteria for crime data prediction to address the issue. Each method yields various accuracy and predicted results, and one approach may provide higher accuracy numbers than other strategies proposed for the same problem. As criminals have been a societal problem worldwide for a long time, action must be taken to prevent crime and ensure public safety. Historical crime data analysis can help identify criminal patterns and anticipate crimes, surpassing the current police tactics of catching criminals after the crime is committed due to technological advancements.

To forecast crime-prone locations, clustering methods are used to group relevant data into required groups using various clustering algorithms. The vast volume of crime datasets and the complexities of data interactions make criminology an excellent field for implementing data mining techniques, which focus on the scientific study of crime, criminal behaviour, and law enforcement to determine the characteristics of crime. This field is one of the most critical areas where data mining techniques can yield significant outcomes. The first step in further analysis is identifying crime characteristics. Data mining techniques provide extremely helpful information to aid and support police forces. Clustering algorithms transform datasets into clusters, which are then investigated to identify crime hotspots. These clusters graphically depict a collection of crimes superimposed on a map of the police jurisdiction, with high-density clusters becoming crime hotspots, and those with fewer people being disregarded. Preventive measures are put in place based on the type of crime in crime-prone locations.

K-means clustering is the most widely used clustering technique in scientific and commercial applications, as it is ideal for clustering huge datasets with lower computational complexity. It has been successfully utilised in various fields such as marketing strategies, data analysis, spatial analysis, astronomy, and agriculture. It is frequently employed as a pre-processing step for several other algorithms. Clustering is preferred over other supervised learning methods such as classification, as crimes vary greatly in type, and crime records are frequently filled with unresolved cases. As a result, classification models relying on previously solved crimes may not provide excellent prediction quality for future criminal activities.

II. LITERATURE SURVEY

Paper [1]: This study introduced a theory that proposes a society without crime, and explored the importance of data mining technology. The researchers also developed an application that aims to identify criminal activity through the use of an improved Decision Tree Algorithm. Specifically, the study implemented a better ID3 Algorithm that takes into account data entropy created by different classes of training data. Additionally, the study proposed a novel method to enhance the Advanced ID3 classification algorithm, which combines a more effective feature selection method with it. The main goal of these improvements was to enhance the overall efficiency and effectiveness of the algorithm in identifying suspicious emails related to criminal activity.

Paper [2] : The paper's authors emphasized the importance of utilizing data mining techniques, clustering, and classification to effectively investigate crimes and identify criminals. They developed the Intelligent Crime Investigation System (ICSIS) which could identify a criminal based on the evidence gathered from the crime scene. Clustering was used to identify patterns of criminal behavior since each crime has a specific pattern. A supervised learning method, Naive Bayes, was used to train the database to predict possible suspects based on criminal histories. This research work proposes a multi-agent system for pattern recognition in criminal activity, where agents are responsible for identifying the location, time, role, trademark, and substance of the criminals. The entities are then returned to the bean for property exposure and converted into objects. To identify the most likely suspects from crime data, the Nave Bayes classifier is used to classify the killers and suspects, and the criminals are grouped according to the model to help identify typical criminal behavior.

Paper [3]: The researchers utilized different data mining techniques with the aid of the quick miner tool to study and forecast crime rates. Their analysis on criminal activity involved the use of the K-Means Clustering method, with an emphasis on identifying crime patterns, predicting criminal behavior based on the spatial distribution of available data, and detecting crimes. Additionally, the study tracked the changes in homicide crime rates between consecutive years.

Paper [4]: This study proposes a novel approach to clustering and predicting crimes based on real-world data. Previous approaches to crime prediction did not address important parameters such as the impact of outliers in data mining preprocessing, the quality of train-test data, and feature significance. To address these gaps, the study utilized the Generation Algorithm to improve outlier detection in the preprocessing phase and defined the fitness function using accuracy and classification error parameters. Overall, this research reflects a unique method for clustering and predicting crimes based on actual data, with a focus on improving outlier detection during the preprocessing phase through the use of accuracy and classification error parameters to define the fitness function.

Paper [5]: This study presents a method for identifying high-risk crime areas that can be used to create maps of crime-prone zones. Rather than focusing solely on crimes, the authors examine the daily factors that contribute to criminal activity. They use the Naive Bayes, Logistic Regression, and Support Vector Machine classifiers to classify daily crime trends and criminal factors. The authors also employ an Apriori algorithm in their crime pattern recognition section to detect tendencies and patterns in criminal activity. To predict potential locations for crime, they use a Decision Tree algorithm.

Paper [6]: This study aimed to provide an alternative approach for predicting criminal activity by utilizing social network interactions at a regional level. The researchers collected individual data from these interactions to gain insights into regional behavior. They developed a method for classifying several features from Foursquare and Gowalla applications, which are commonly used in the San Francisco Bay Area. By analyzing the geographic data obtained from maps, the researchers were able to track crime patterns and occurrences and identify metropolitan areas with significant criminal activity. The study focused on the use of regional information to investigate crime in urban areas, and the Haversine formula was employed to estimate and display the distance between the crime scene and the location.

Paper [7]: This study proposed a method of grouping criminals based on their criminal history. The criminal profile for each offense and year is obtained from the database, and a distance matrix is calculated. Next, a profile distance matrix is created for the following year, which includes the frequency value. Clusters are then generated using a naive clustering technique. By establishing a criminal profile that depicts a year's worth of criminal activity, a large group of offenders can be easily analyzed, and future suspicious behavior can be anticipated. This can aid in obtaining a clear understanding of the various types of criminal careers that are presently available. To extract the elements required for determining a person's criminal career, the program was tested on the actual Dutch National Criminal Record Database.

Paper [8]: This study presents a technique that utilizes social media usage, call records, and location markers to aid criminal analysis and make quicker and more precise conclusions. The authors also introduce a novel approach that utilizes human social interaction to detect suspicious behavior based on social network feeds. A sequence of inference rules is employed to identify the entity's suspicious movements. Their model is capable of predicting and characterizing human behavior using real-world data sources.

Paper [9]: The author of this article has developed a forensic tool that can locate the primary members of a criminal organization, including its top leaders, with the aim of weakening the organization's power by removing these individuals. This forms the basis of their approach. Their new project, called SIIMCO, creates a graph-based network representation of the criminal organization using either mobile communication data or criminal records. These networks serve as the foundation of the system, representing the criminal organization or crime incident reports. The connections or lines of contact between two criminal vertex points indicate their interpersonal relationships. The author has used formulas to measure the influence and importance of each vertex relative to other vertices, or criminals in the graph, in order to determine each vertex's overall impact.

Paper [10]: The focus of this study is on the care of mentally ill inmates during their incarceration, and the authors have put forth a hypothesis regarding their treatment. They propose that the Social Security Number (SSN) of mentally ill offenders, along with their personal and professional criminal records, can be used to identify them. Based on their mental health status, these offenders are classified into three levels of risk potential: "high," "medium," and "low." The aim is to differentiate between misdemeanors and felonies that may or may not be referred, depending on the mental state of the offenders. Continuous data collection and monitoring of their behaviors help distinguish mentally ill offenders from other dangerous criminals who pose a threat to other inmates. In addition, the study divides the mental health of inmates into "referred" and "not-referred" groups.

III. PROBLEM STATEMENT

The main challenge is that with the increasing population, crime rates are also expected to rise in different areas, making it difficult for officials to accurately predict crime rates. Despite the officials' efforts to address various issues, predicting future crimes may still be challenging, and reducing crime rates may not always be feasible. Previous studies have analyzed large datasets to extract information such as location and types of crimes to aid law enforcement. Using these databases, existing methods have identified crime hotspots based on location, and there are various mapping applications that display crime locations and types in cities. However, while crime locations have been identified, information on the date and time of the crime occurrence is often lacking, and accurate techniques for predicting future crimes are not yet available.

IV. IMPLEMENTATION

The project was implemented using Python language, specifically using the Anaconda distribution for machine learning purposes. Anaconda is a popular Python distribution, formerly known as Continuum Analytics, that offers more than 100 new packages for scientific computing, data science, statistical analysis, and machine learning. The team found Anaconda to be an easier tool to work with for several reasons, including installing Python on multiple platforms, creating separate environments, dealing with insufficient privileges, and quickly getting started with specific packages and libraries.

The data for this project was scraped from the publicly available data on the Bangalore police website, which was compiled by different police stations in the city. To limit the area for prediction and reduce complexity, the team started the implementation in Bangalore. The data was sorted and converted into a new format with time-stamp, longitude, and latitude, which served as input for the machine learning model to predict crime rates in specific locations or cities. Entries were made to help the machine learn the algorithms and the expected output. The accuracy of different algorithms was measured, and the one with the highest accuracy, Random Forest, was used for the prediction kernel.

4.1 KNN (K-Nearest Neighbours)

K nearest neighbors (KNN) is a popular non-parametric lazy learning algorithm used in pattern recognition and data mining. It employs a similarity measure, such as a distance function, to classify new cases based on stored cases. This approach is also known as instance-based learning. KNN stores all available cases and assigns new instances to the most common class amongst its K nearest neighbors, where K is a user-defined parameter. The algorithm is widely used due to its simplicity and effectiveness in classification tasks.

4.2 RANDOM FOREST

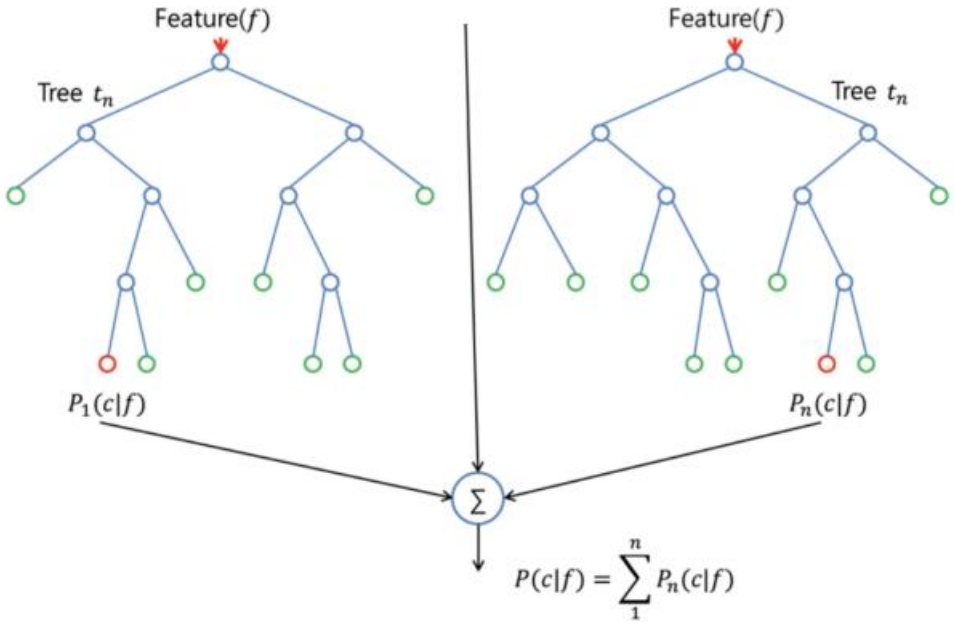
Random Forests is an ensemble learning method that has gained popularity due to its ability to combine multiple classifiers and produce accurate predictions on test data. The algorithm is designed to minimize variance and prevent overfitting on the training data by using randomness during split decision-making. Random Forests uses a family of classifiers, denoted as $h(x|\theta 1)$, $h(x|\theta 2)$,..., $h(x|\theta k)$, where each member of the family is a classification tree, and k is the number of trees selected from a model random vector. The parameter vectors, θk, are chosen randomly, and each classification tree in the ensemble is built using a different subset, $D\theta k(x,y) \subset D(x,y)$, of the training dataset, where D(x,y) represents the entire training dataset.

Thus, $h(x|\theta k)$ is the kth classification tree which uses a subset of features $x\theta k \subset x$ to build a classification model. Each tree then works like regular decision trees: it partitions the data based on the value of a particular feature (which is selected randomly from the subset), until the data is fully partitioned, or the maximum allowed depth is reached. The final output y is obtained by aggregating the results thus

$$y = \text{argmax}_{p \in \{h(x_1)..h(x_k)\}} \{ \sum_{j=1}^{\kappa} (I(h(x|\theta_j) = p)) \}$$

where I denotes the indicator function.

Below is the figure of Random forest Example.



4.3 TECH STACK

Anaconda Distribution (v5.1)
Python (3.6.5)
Packages Used:
- Flask (0.12.2)
- Pandas (0.22.1)
- Numpy (1.14.2)
- Sklearn (0.19.1)
- Geopy (1.13.0)

HTML 5
CSS 3
Bootstrap 4
Java Script 1.8

4.3.1 Anaconda Distribution (v5.1)

Anaconda Distribution is a software package that provides an easy way to install and manage a variety of data science and scientific computing packages, including Python, R, and Jupyter notebooks. Anaconda Distribution includes a large collection of open-source packages, including popular data science libraries such as NumPy, pandas, and scikit-learn, and is available for Windows, macOS, and Linux.

4.3.2 Python (3.6.5)

Python 3.6.5 is a version of Python that was released on March 28, 2018. It is an incremental improvement over the previous version, Python 3.6.4, and includes several bug fixes and minor enhancements. Some of the new features in Python 3.6.5 include, improved handling of Unicode characters, faster string formatting, improved error messages and debugging information, improved handling of asynchronous programming using the asyncio library.

4.3.3 Packages Used:

- Flask (0.12.2) : Flask is a lightweight web framework for Python that is designed to make it easy to build web applications. Flask is known for its simplicity and flexibility, and it is particularly well-suited for building small to medium-sized web applications and APIs.
- Pandas (0.22.1) : Pandas is a popular open-source data manipulation and analysis library for Python. Pandas provides data structures for efficiently storing and manipulating large datasets, as well as a wide range of functions for performing data analysis tasks, such as filtering, aggregation, and transformation.
- Numpy (1.14.2) : NumPy is a popular open-source library for numerical computing in Python. NumPy provides support for large, multi-dimensional arrays and matrices, as well as a wide range of mathematical functions for working with these arrays.
- Sklearn (0.19.1) : NumPy is a popular open-source library for numerical computing in Python. NumPy provides support for large, multi-dimensional arrays and matrices, as well as a wide range of mathematical functions for working with these arrays.
- Geopy (1.13.0) : Geopy is a Python library for geocoding and geolocation that provides an easy-to-use interface to several popular geocoding APIs. Geopy supports several geocoding services such as Google Geocoding API, Bing Maps API, OpenStreetMap Nominatim, and many more.

4.3.4 HTML 5

HTML5 is the latest version of the standard markup language for creating web pages and applications, released in 2014 by the W3C. It introduces several new features and enhancements over its predecessor, such as native support for multimedia, the canvas element for drawing graphics and animations, offline support, improved semantics, and better support for forms. HTML5

provides a more robust and flexible platform for building web applications, allowing developers to create richer, more interactive web experiences that can work seamlessly across different devices and platforms.

4.3.5 CSS 3

CSS3 is the latest version of the style sheet language used for describing the presentation of web pages and applications. Released in 2011 by the W3C, it introduces several new features and enhancements over CSS2.1, such as flexible box layout, grid layout, transitions and animations, and media queries. These features provide developers with a more powerful and flexible platform for styling web pages and applications, allowing them to create dynamic and engaging user interfaces without relying on JavaScript or other scripting languages.

4.3.6 Bootstrap 4

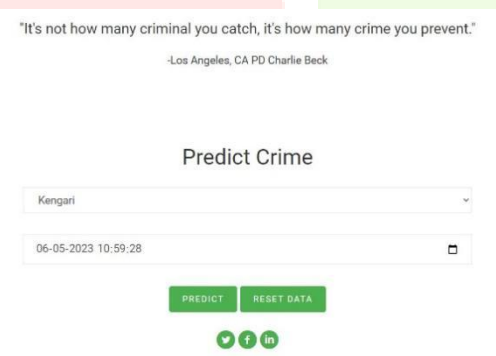
Bootstrap 4 is a widely-used front-end development framework designed to create responsive, mobile-first web applications. Created by Twitter and released in 2018, it builds upon the foundation of HTML, CSS, and JavaScript, offering new features and enhancements over its predecessor, Bootstrap 3. These include native Flexbox support, an updated grid system for improved responsiveness, a redesigned form layout and validation system, and a new card component to streamline page design. Bootstrap 4 makes it easier for developers to create modern and responsive websites that work seamlessly across different devices, making it a popular choice for web development projects.

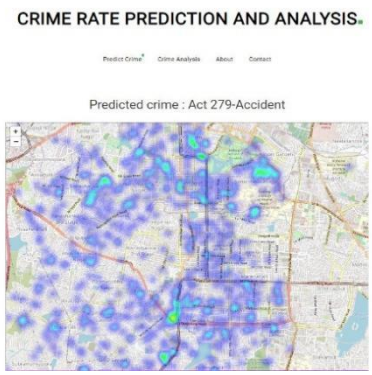
4.3.7 Java Script 1.8

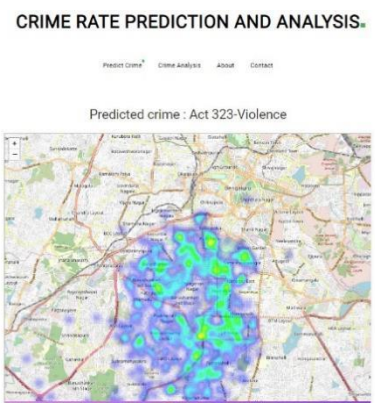
JavaScript 1.8 is a version of the popular scripting language that was released in 2008. It introduced several new features and improvements over its predecessor, JavaScript 1.7, such as generator functions, expression closures, trailing commas in object literals, and improved regular expressions. These features made JavaScript more efficient, concise, and flexible, providing developers with a more powerful tool for creating dynamic and interactive web pages and applications. JavaScript 1.8 is still widely used today, and its features have become fundamental building blocks for modern JavaScript frameworks and libraries.

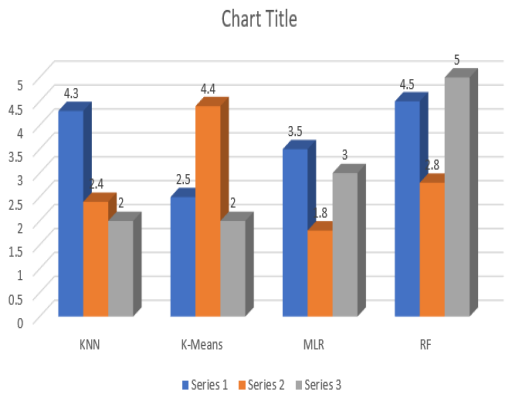
V. RESULTS

The crime rate analysis and prediction application utilizing KNN, K-means, Random Forest, and location data with a map has yielded promising results. By collecting historical crime data, including location, date, and time, we were able to pre-process and analyse the data to uncover valuable insights. Using feature selection techniques, we determined the most influential factors for crime prediction. These features were then used to train machine learning models, including KNN, K-means, and Random Forest. The models were trained on the historical crime data, enabling them to learn patterns and make accurate predictions. With the integration of location data and the map, the application provides a user-friendly interface for users to input specific dates and times to predict the likelihood of crimes occurring at particular locations. The map visualization displays the predicted crime hotspots, helping users make informed decisions and take necessary precautions.









VI CONCLUSION

Crime has a wide-ranging impact on individuals, communities, and geographical regions across the world. Accurately predicting and analyzing crime data is a challenging yet critical task. Early identification of crime trends can help prevent criminal activities. In this study, we have reviewed several established techniques for data-driven crime analysis and forecasting. We have compared papers that have a background in crime prediction and criminal identification with a theoretical study. Each approach has its advantages and limitations, and each paper employs a different methodology to identify crimes and predict criminal behavior.

By identifying crime patterns and trends, we can predict the types of crimes that may occur in a particular district during a given season. However, to enhance the performance of a single classifier that predicts crime in a specific district, it is necessary to integrate different models and consider the seasonal and time component at which crimes are more likely to occur. Crime forecasting is important in maintaining personal safety by advising individuals to avoid certain areas during specific hours, months, and seasons. Additionally, this information can help individuals make informed decisions about where to reside and vacation.

REFERENCES

[1] Mugdha Sharma, "Z-Crime: A Data Mining Tool for the Detection of Suspicious Criminal Activities based on the Decision Tree", International Conference on Data Mining and Intelligent Computing, pp. 1-6, 2014.

[2] Kaumalee Bogahawatte and Shalinda Adikari, "Intelligent Criminal Identification System", Proceedings of 8th IEEE International Conference on Computer Science and Education, pp. 633-638, 2013.

[3] Jyoti Agarwal, Renuka Nagpal and Rajni Sehgal, "Crime Analysis using K-Means Clustering", International Journal of Computer Applications, Vol. 83, No. 4, pp. 1-4, 2013.

[4] Rasoul Kiani, Siamak Mahdavi and Amin Keshavarzi, "Analysis and Prediction of Crimes by Clustering and Classification", International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No. 8, pp. 11-17, 2015.

[5] Shiju Sathya Devan, M.S. Devan, and S. Surya Gangadharan, "Crime Analysis and Prediction using Data Mining", Proceedings of IEEE 1st International Conference on Networks and Soft Computing, pp. 406-412, 2014.

[6] Yu-Yueh Huang, Cheng-Te Li and Shyh-Kang Jeng, "Mining Location-based Social Networks for Criminal Activity Prediction", Proceedings of 24th IEEE International Conference on Wireless and Optical Communication, pp. 185-190, 2015.

[7] Jeroen S. De Bruin, Tim K. Cocx, Walter A. Kosters, Jeroen F. J. Laros and Joost N. Kok, "Data Mining Approaches to Criminal Career Analysis", Proceedings of 6th IEEE International Conference on Data Mining, pp. 1-7, 2006.

[8] Ehab Hamdy, Ammar Adl, Aboul Ella Hassanien, Osman Hegazy, and Tai-Hoon Kim, "Criminal Act Detection and Identification Model", Proceedings of 7th International Conference on Advanced Communication and Networking, pp. 79-83, 2015.

[9] Kamal Taha and Paul D. Yoo, "SIIMCO: A Forensic Investigation Tool for Identifying the Influential Members of a Criminal Organization", IEEE Transactions on Information Forensics and Security, Vol. 11, No. 4, pp. 811- 822, 2016.

[10] Kevin Sheehy et al., "Evidence-based Analysis of Mentally 111 Individuals in the Criminal Justice System", Proceedings of IEEE Systems and Information Engineering Design Symposium, pp. 250-254, 2016.

[11] David J. Hand, Heikki Mannila and Padhraic Smyth, "Principles of Data Mining", MIT Press, 2001.

[12] Donald E. Brown, "The Regional Crime Analysis Program (RECAP): A Framework for Mining Data to Catch Criminals", Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, pp. 2848-2853, 1998.

[13] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "The KDD Process for Extracting Useful Knowledge from Volumes of Data", Communications of the ACM, Vol. 39, No. 11, pp. 27-34, 1996.

[14] Illhoi Yoo, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang and Lei Hua, "Data Mining in Healthcare and Biomedicine: A Survey of the Literature", Journal of Medical Systems, Vol. 36, No. 4, pp. 2431-2448, 2011.

[15] Shyam Varan Nath, "Crime Pattern Detection using Data Mining", Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, pp. 1-4, 2006.