JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 10, ISSUE 04, 2023

K-nearest Neighbour for Exploratory Data Analysis of Crime Data

M. Syamala Sai Sree[1], A. Srirupa[2], Ch. Kavya[2], B. Sravya [2], Ch. Varshini[2]

[1,2]Department of Information Technology

[1,2]Malla Reddy Engineering College for Women (A), Maisammaguda, Medchal, Telangana.

ABSTRACT

Crime analysis and prevention is a systematic approach for identifying and analyzing patterns and trends in crime. The system can predict regions which have a high probability for crime occurrence and can visualize crime prone areas. With the increasing advent of computerized systems, crime data analysts can help the Law enforcement officers to speed up the process of solving crimes. Using the concept of data mining, real time and location data, the system can extract unknown, useful information from unstructured data. Here we have an approach between computer science and criminal justice to develop a data mining, real time and location data procedure that can help solve crimes faster. The causes of crime occurrence like criminal background of offender, political enmity etc. we are focusing mainly on crime factors of each day. To have a better response towards criminal activity, it is very important that one should understand the patterns in crime.

Keywords: Crime analysis, Data analysis, K nearest neighbour.

1. INTRODUCTION

The crime rate increases on a daily basis. Crime as the word suggests is the violation that people do, and it is usually performed against the laws, and it is punishable. Crime cannot be predicted since it is not systematic. Also, modern technologies and hi-tech methods help criminals in achieving their goals. According to the Nigeria police, crimes like burglary, arson and it likes to have been decreased while crimes like sex filming, rape, robbery, fraud, kidnapping etc. are increasing rapidly. Though crime victims might not be easily predicted but the time and location can be predicted based on the probabilities of its occurrence. The predicted results cannot be assured of 100% accuracy, but these results show that the software program helps in reducing crime rates to a certain extent by providing security in vital areas where crimes can easily occur. Developing such a powerful crime analytics tool we must collect crime records and evaluate them. Criminal and sociology scholars are analyzing the pattern of criminal activity and its relationship with the area. Researchers have shown that many crook activities are taking place in a region. This is called a hotspot. Machine learning can be used to become aware of hotspots by way of data pushed approach.

It is only within the last few decades that technology has made spatial data mining a practical solution for wide audiences of Law enforcement officials which is affordable and available. Since the availability of criminal records is limited, the collection of crime data from various sources like newspapers, new websites, blogs, social media etc. This huge data is used as a record for creating a crime record database. So, the main challenge is developing a better, efficient crime pattern detection tool to identify crime patterns effectively.

Mining data is a technique for dealing with large data indexes in order to recognise patterns and build up a group to deal with difficulties via information analysis. The devices that have been used enable future samples to be accepted. Deep Learning is a method for analysing data from an informational collection in order to transform it into a suitable structure that can be used for further processing [1].

Crime Detection: In most nations, the police are responsible for the detection of criminal activity, while law enforcement organisations may be tasked with the finding of certain forms of criminal activity (e.g., customs departments may be charged with combating smuggling and related offenses).

Crime detection may be divided into three distinct phases: the discovery that a crime has been committed, the identification of a suspect, and the gathering of sufficient evidence to indict the suspect in front of a court of law. Many crimes are found and reported by somebody other than the police. This is a common occurrence (e.g., victims or witnesses) [2].

2. LITERATURE SURVEY

Kamoun et. al [3] reviews the defensive usage of AI/MLS in cybersecurity and then presents a survey of its offensive use. Inspired by the System-Fault-Risk (SFR) framework, we categorize AI/MLS-powered cyberattacks by their actions into seven categories. We cover a wide spectrum of attack vectors, discuss their practical implications and provide some recommendations for future research.

Fatima Dakalbab et. al [4] investigates AI strategies in crime prediction. They conduct a systematic literature review (SLR). This review evaluates the models from numerous points of view, including the crime analysis type, crimes studied, prediction technique, performance metrics and evaluations, strengths and weaknesses of the proposed method, and limitations and future directions. They review 120 research papers published between 2008 and 2021 that cover AI approaches for crime prediction. They provide 34 crime categories researched by researchers and 23 distinct crime analysis methodologies after analyzing the selected research articles. On the other hand, we identify 64 different machine learning (ML) techniques for crime prediction. In addition, we observe that the most applied approach in crime prediction is the supervised learning approach. Furthermore, they discuss the evaluation and performance metrics, as well as the tools utilized in building the models and their strengths and weaknesses. Crime prediction AI techniques are a promising field of study, and there are several ML models that researchers have applied. Consequently, based upon this review, they provide advice and guidance for researchers working in this area of study.

Shah et. al [5] described the results of certain cases where such approaches were used, and which motivated us to pursue further research in this field. The main reason for the change in crime detection and prevention lies in the before and after statistical observations of the authorities using such techniques. The sole purpose of this study is to determine how a combination of ML and computer vision can be used by law agencies or authorities to detect, prevent, and solve crimes at a much more accurate and faster rate. In summary, ML and computer vision techniques can bring about an evolution in law agencies.

Harris et. al [6] established the state of the field in application of AI to policing of financial markets and take an interdisciplinary look at opportunities to enhance the use of AI in policing more broadly. The chapter begins with an explanation of the laws designed to combat crime on financial markets—in particular the offences of insider trading and market manipulation. This is followed by an analysis of the current state of the field and discussion of the application of AI to detect and deter financial crime, as well as the use of AI in other areas of policing. The chapter concludes with lessons and opportunities from the application of AI to policing financial crime, noting the risks and limitations of an AI approach and the challenges and opportunities for expanding the application of AI to policing more generally.

Kim et. al [7] examines related cases around the world, and categorizes them according to their specific modi operandi, as well as the initial responses of national authorities to this emerging crime. By analyzing the dynamics of this new crime trend, this study aims to propose preemptive and preventative measures to address this new threat. Research has identified three main steps in SIM Swapping crime: 1) personal data theft, 2) fraudulent copying of SIM card, and 3) exploitation of falsely-obtained mobile service for perpetration. Research has also found that the subscriber authentication procedure involved in replacing a SIM card is vulnerable to identity theft, especially in

jurisdictions which have implemented eSIM. Therefore, it will be upon governments to enforce a stronger user authentication and information security regime for mobile carriers, introduce an online payment system devised with a data-sharing mechanism connecting mobile carriers and financial services, and raise public awareness on SIM Swapping and information security in general.

Nicholls et. al [8] presented the compound the necessity for use of detection techniques based on Graph/Group based Anomaly Detection to combat financial crime. The authors accept the challenge particularly in obtaining labeled datasets, and the expertise required in labeling ground truths where one is not already available. With the advancement of cryptocurrency and its deepening entrenchment into the financial ether, it is not surprising that anti-money laundering in cryptocurrency research has been initiated. The authors suspect a closer examination of cryptocurrency and its integration into the public domain by the respective Revenue Commissioners and law enforcement authorities of the varying countries worldwide, resulting in increased output of research, particularly in the Group/Graph based Anomaly Detection domain.

RAZA et. al [9] focuses on the spam classification approached using machine learning algorithms. Furthermore, this study provides a comprehensive analysis and review of research done on different machine learning techniques and email features used in different Machine Learning approaches. Also provides future research directions and the challenges in the spam classification field that can be useful for future researchers.

Bada et. al [10] investigates the afore-mentioned domain to answer the question: what is the state-of-the-art in the academic field of understanding, characterising and profiling cybercriminals. Through the application of the PRISMA systematic literature review technique, we identify 39 works from the last 14 years (2006-2020). Our findings demonstrate that overall, there is lack of a common definition of profiling for cyber-offenders. The review found that one of the primary types of cybercriminals that studies have focused on is hackers and the majority of papers used the deductive approach as a preferred one. This article produces an up-to-date characterisation of the field and also defines open issues deserving of further attention such as the role of security professionals and law enforcement in supporting such research, as well as factors including personality traits which must be further researched whilst exploring online criminal behaviour. By understanding online offenders and their pathways towards malevolent behaviours, they can better identify steps that need to be taken to prevent such criminal activities.

3. PROPOSED SYSTEM

The K-nearest neighbors (KNN) classifier is a popular algorithm used in machine learning for classification tasks. It is particularly useful for exploratory data analytics of crime data due to its simplicity and ability to make predictions based on similar instances. The KNN classifier is a non-parametric algorithm that makes predictions based on the similarity of instances in a dataset. It operates by assigning a class label to a new data point based on the majority class label of its k nearest neighbors in the feature space. The distance metric, such as Euclidean distance, is used to measure the similarity between instances.

The need for KNN in the context of crime data analysis arises from the desire to identify patterns, trends, and relationships within the data. By utilizing KNN, analysts can classify new instances based on their proximity to existing labeled instances. In the case of crime data, this means that KNN can be used to classify new criminal incidents by comparing them to historical data.

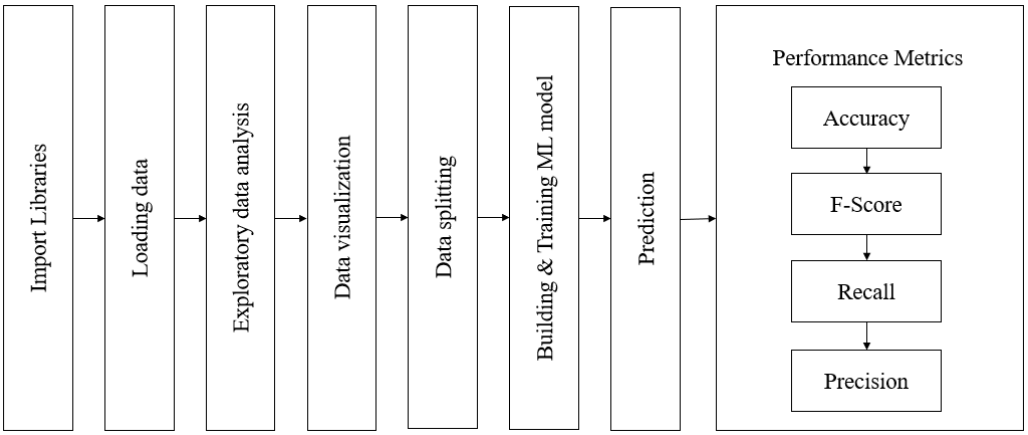


Fig. 1: Block diagram of proposed system.

3.1 K-Nearest Neighbor (KNN) Algorithm

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.
- It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Steps

1. Get labeled data: The labeled data consists of features and labels. Features are the characteristics or the property of the object whereas labels are the class of the object with those features.
2. Convert labeled data to encoded data: Usually computations are based on numerical form so we convert the data to numeric form by encoding them.
3. Create feature set: Creating a set of features by packing the features.
4. Split the data for train and test: The data are split training and testing. Usually, 80% for training and 20% for testing but can select based on need.
5. Train the classifier: Training the classifier with the training data by specifying the value of k. Use k =3 for binary classification, i.e., two labels classification. If used k =1 then it is simply a nearest neighbor classifier.
6. Test the classifier: Testing the classifier with the testing data.
7. Evaluate: Evaluating the classifier using confusion matrix and its evaluation metrics i.e., accuracy, precision, recall, et cetera.

Advantages

KNN is advantageous for exploratory data analytics of crime data due to the following reasons:

- Non-parametric Approach: KNN is a non-parametric algorithm, meaning it doesn't make any assumptions about the underlying data distribution. This makes it suitable for analyzing crime data, which can often have complex and diverse patterns.
- Flexibility: KNN allows for easy adaptation to changing data patterns. As new crime data is collected, the KNN model can be updated and retrained to incorporate the latest information.
- Interpretable Results: KNN provides intuitive and interpretable results. By identifying the K nearest neighbors to a given instance, analysts can gain insights into the characteristics and attributes that contribute to the classification decision.

Comparison of KNN and Random Forest with FBI Crime Data

KNN and Random Forest are both machine learning algorithms, but they have different characteristics and strengths. When it comes to analyzing FBI crime data, the choice between these algorithms depends on the specific requirements and objectives of the analysis.

KNN is better than Random Forest for exploratory data analytics of crime data in certain scenarios:

- Proximity-based Classification: KNN excels at proximity-based classification. If the goal is to identify similar crime incidents based on their attributes and classify new incidents accordingly, KNN is well-suited for this task. It considers the local patterns and relationships within the data.
- Interpretability: KNN provides transparent and interpretable results. It directly shows which instances in the training data are closest to a given test instance, making it easier to understand the reasoning behind the classification decision.

4. RESULTS AND DISCUSSION

Sample dataset

	Unnamed: 0	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	...	Ward	Community Area	FBI Code	X Coordinate
0	3	10508693	HZ250496	05/03/2016 11:40:00 PM	013XX S SAWYER AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE	APARTMENT	True	...	24.0	29.0	08B	1154907.0
1	89	10508695	HZ250409	05/03/2016 09:40:00 PM	061XX S DREXEL AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE	RESIDENCE	False	...	20.0	42.0	08B	1183066.0
2	197	10508697	HZ250503	05/03/2016 11:31:00 PM	053XX W CHICAGO AVE	0470	PUBLIC PEACE VIOLATION	RECKLESS CONDUCT	STREET	False	...	37.0	25.0	24	1140789.0
3	673	10508698	HZ250424	05/03/2016 10:10:00 PM	049XX W FULTON ST	0460	BATTERY	SIMPLE	SIDEWALK	False	...	28.0	25.0	08B	1143223.0
4	911	10508699	HZ250455	05/03/2016 10:00:00 PM	003XX N LOTUS AVE	0820	THEFT	$500 AND UNDER	RESIDENCE	False	...	28.0	25.0	06	1139890.0

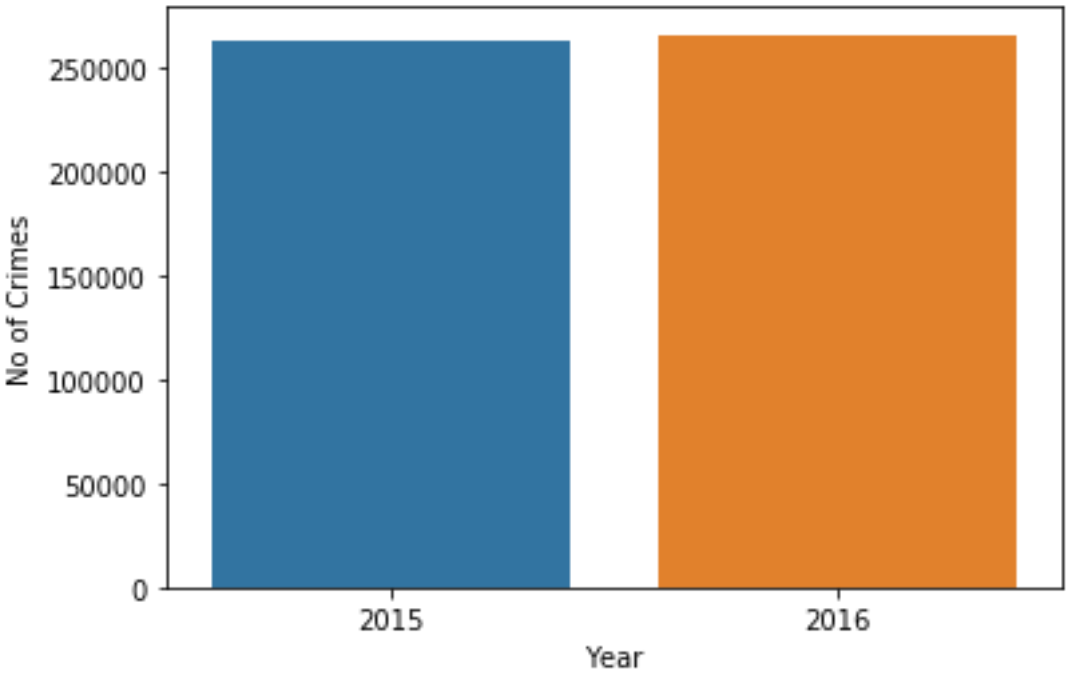
5 rows × 23 columns

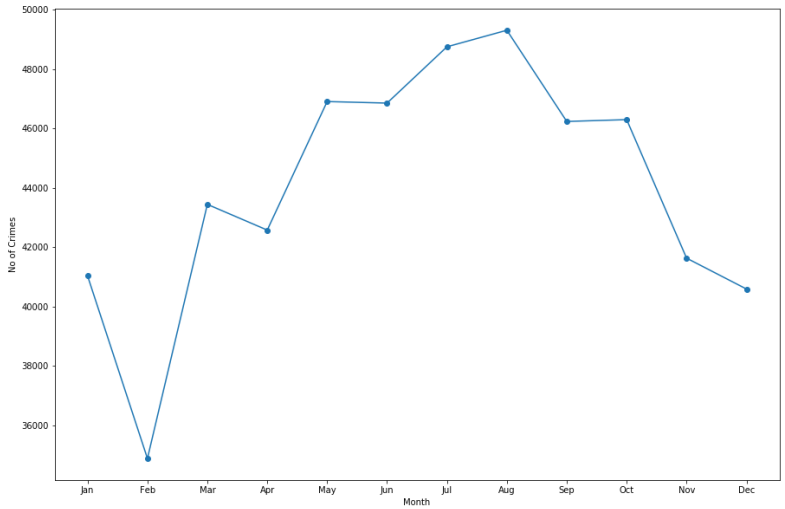
Y Coordinate	Year	Updated On	Latitude	Longitude	Location
1893681.0	2016	05/10/2016 03:56:50 PM	41.864073	-87.706819	(41.864073157, -87.706818608)
1864330.0	2016	05/10/2016 03:56:50 PM	41.782922	-87.604363	(41.782921527, -87.60436317)
1904819.0	2016	05/10/2016 03:56:50 PM	41.894908	-87.758372	(41.894908283, -87.758371958)
1901475.0	2016	05/10/2016 03:56:50 PM	41.885687	-87.749516	(41.885686845, -87.749515983)
1901675.0	2016	05/10/2016 03:56:50 PM	41.886297	-87.761751	(41.886297242, -87.761750709)

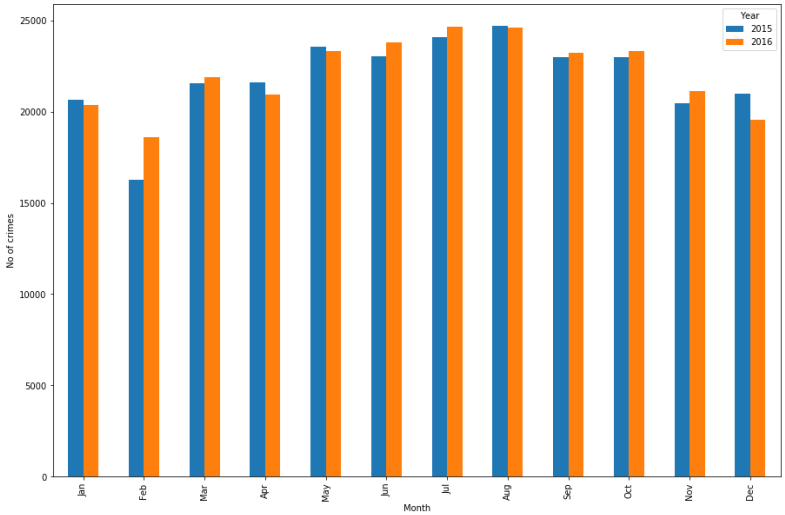
List of Attributes

```
['Unnamed: 0',
 'ID',
 'Case Number',
 'Date',
 'Block',
 'IUCR',
 'Primary Type',
 'Description',
 'Location Description'
 'Arrest',
 'Domestic',
 'Beat',
 'District',
 'Ward',
 'Community Area',
 'FBI Code',
 'X Coordinate',
 'Y Coordinate',
 'Year',
 'Updated On',
 'Latitude',
 'Longitude',
 'Location']
```

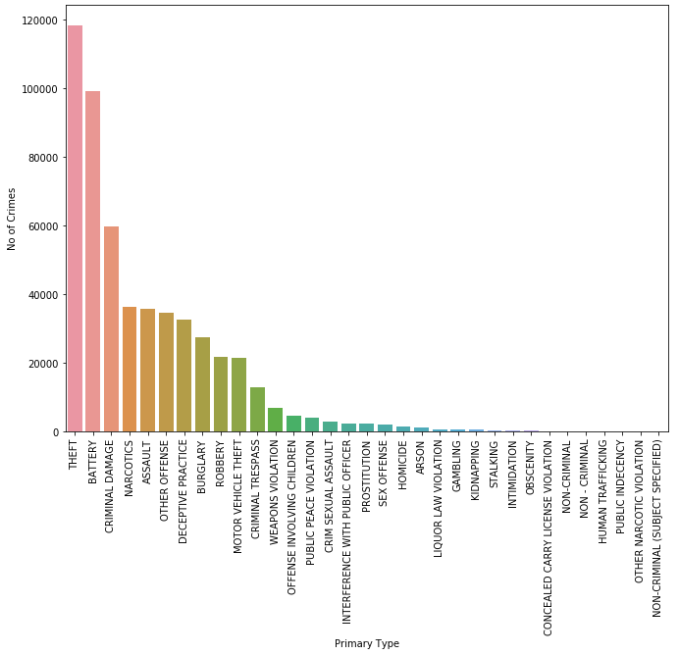
ISSN- 2394-5125 VOL 10, ISSUE 04, 2023

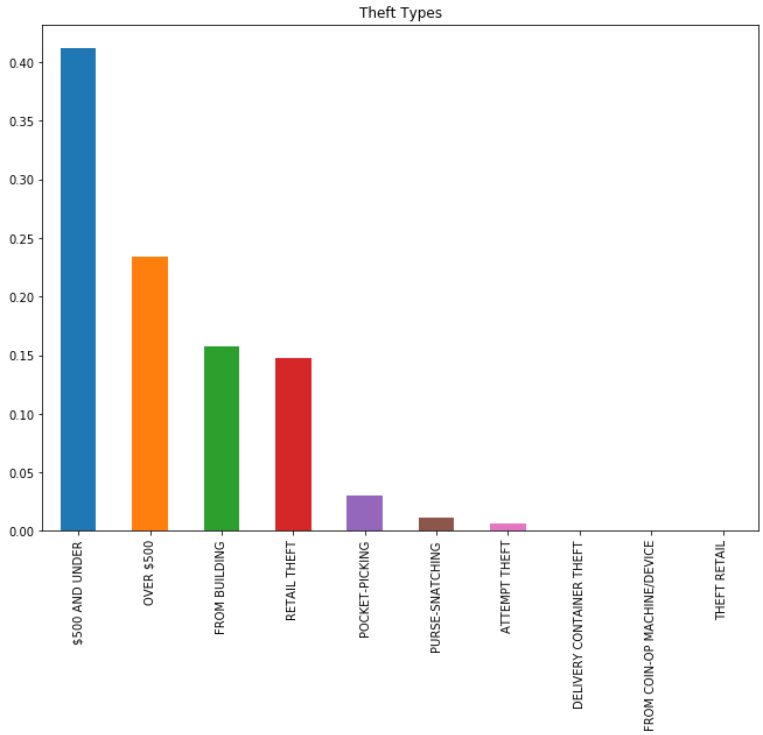


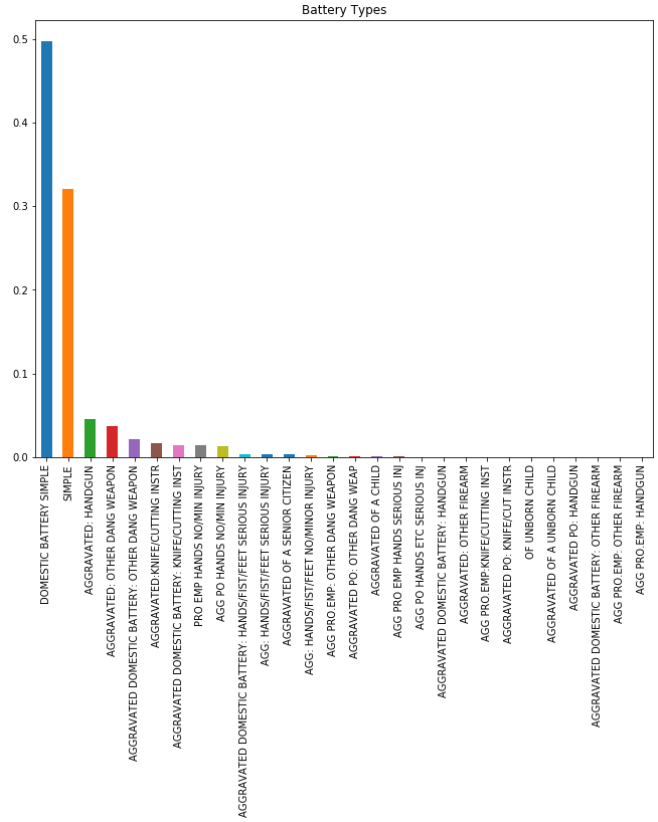


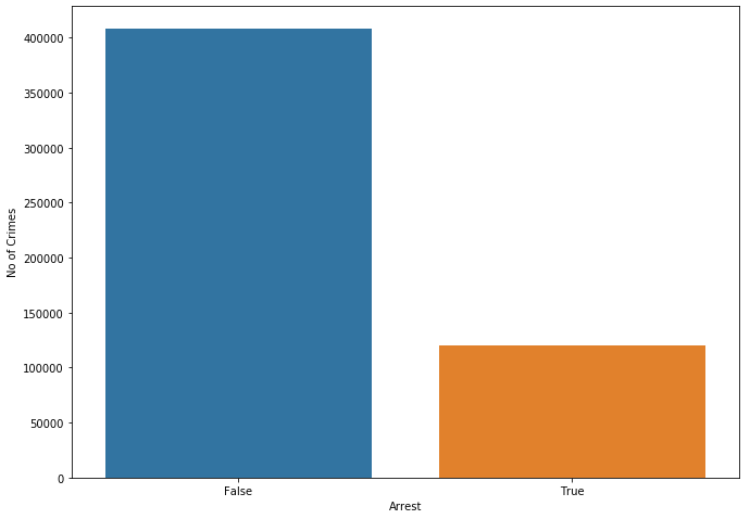


ISSN- 2394-5125 VOL 10, ISSUE 04, 2023

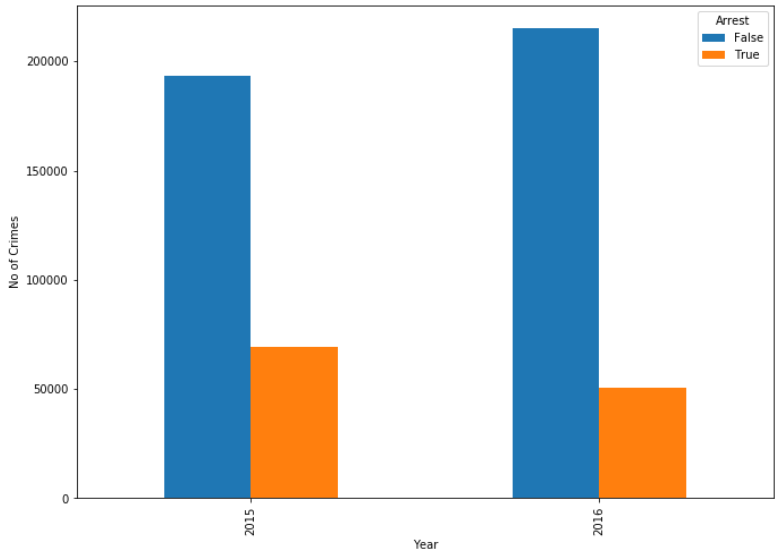


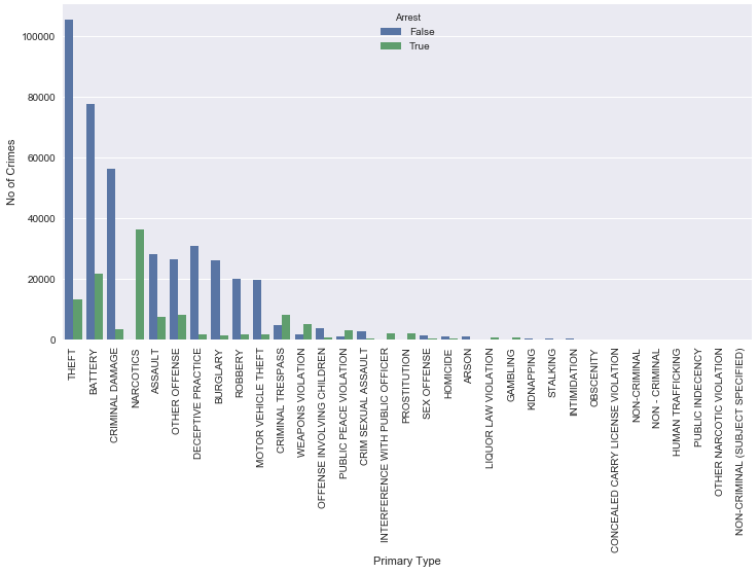


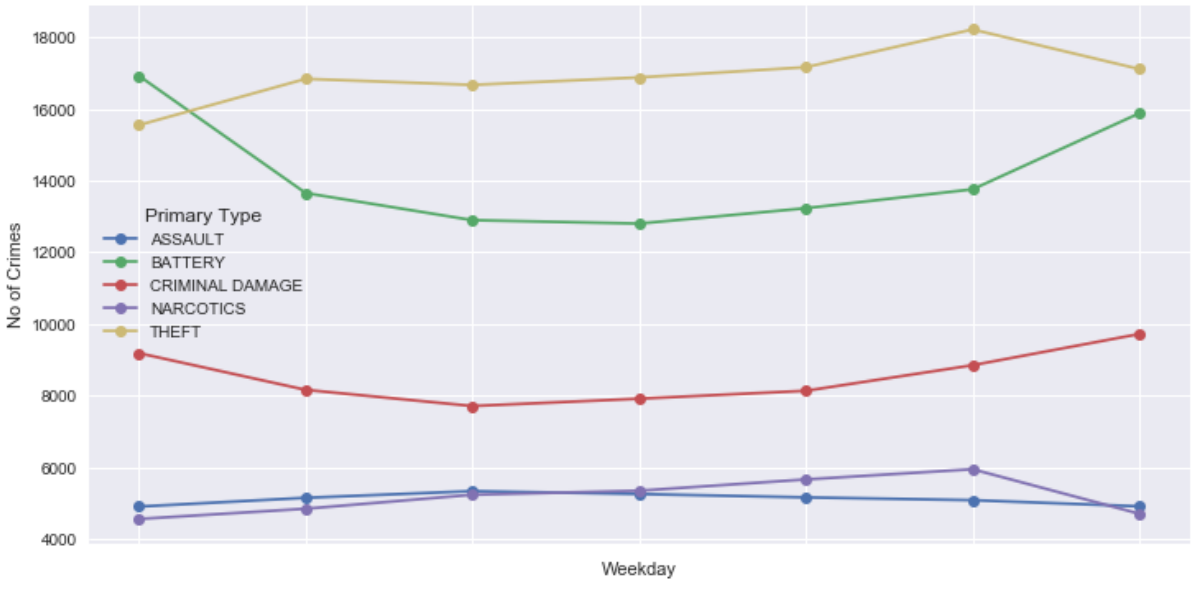

Battery Types
0.5
0.4
0.3
0.2
0.1
0.0
DOMESTIC BATTERY SIMPLE
SIMPLE
AGGRAVATED: HANDGUN
AGGRAVATED: OTHER DANG WEAPON
AGGRAVATED DOMESTIC BATTERY: OTHER DANG WEAPON
AGGRAVATED:KNIFE/CUTTING INSTR
AGGRAVATED DOMESTIC BATTERY: KNIFE/CUTTING INST
PRO EMP HANDS NO/MIN INJURY
AGG PO HANDS NO/MIN INJURY
AGGRAVATED DOMESTIC BATTERY: HANDS/FIST/FEET SERIOUS INJURY
AGG: HANDS/FIST/FEET SERIOUS INJURY
AGGRAVATED OF A SENIOR CITIZEN
AGG: HANDS/FIST/FEET NO/MINOR INJURY
AGG PRO EMP: OTHER DANG WEAPON
AGGRAVATED PO: OTHER DANG WEAP
AGGRAVATED OF A CHILD
AGG PRO EMP HANDS SERIOUS INJ
AGG PO HANDS ETC SERIOUS INJ
AGGRAVATED DOMESTIC BATTERY: HANDGUN
AGGRAVATED: OTHER FIREARM
AGG PRO EMP:KNIFE/CUTTING INST
AGGRAVATED PO: KNIFE/CUT INSTR
OF UNBORN CHILD
AGGRAVATED OF A UNBORN CHILD
AGGRAVATED PO: HANDGUN
AGGRAVATED DOMESTIC BATTERY: OTHER FIREARM
AGG PRO EMP: OTHER FIREARM
AGG PRO EMP: HANDGUN

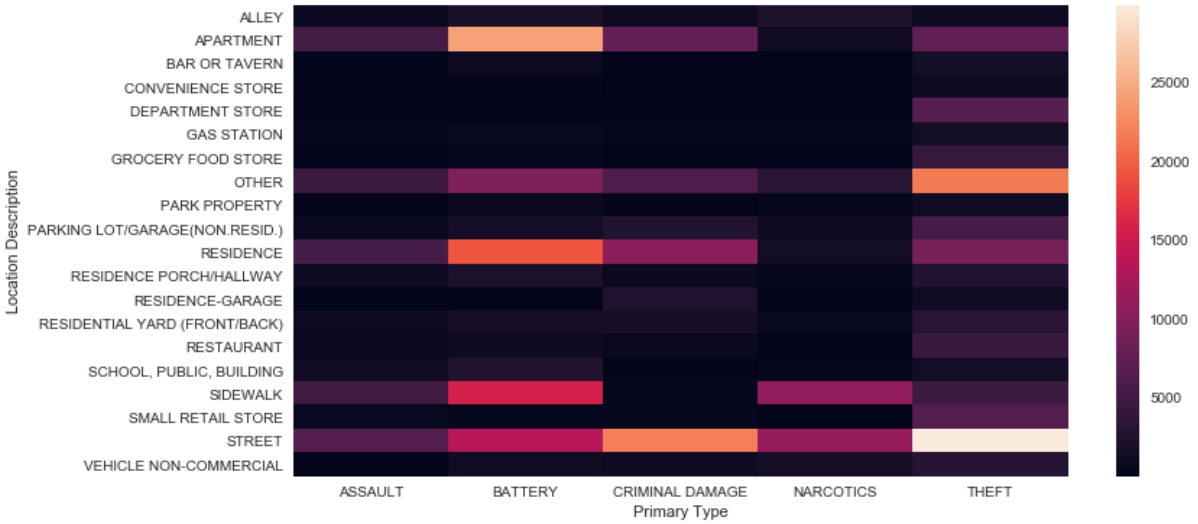

No of Crimes
400000
350000
300000
250000
200000
150000
100000
50000
0
False
True
Arrest

ISSN- 2394-5125 VOL 10, ISSUE 04, 2023









5. CONCLUSION

Crime analysis and prevention is a systematic approach aimed at identifying and analyzing patterns and trends in crime. By utilizing computerized systems and data analysis techniques, this approach enables the efficient prediction of crime occurrence and visualization of crime-prone areas. The abstract emphasizes the significance of computer science and criminal justice working together to develop data mining, real-time data, and location-based procedures that can help law enforcement officers solve crimes more rapidly. From the obtained results, the KNN Classifier outperforms the Random Forest Classifier in terms of accuracy for exploratory data analytics of crime data. The future scope involves refining the KNN Classifier through parameter tuning, feature engineering, integration with other techniques, and applying it to real-time crime prediction and prevention systems.

REFERENCES

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

[2] Kaur, B., Ahuja, L., & Kumar, V. (2019, February). Crime against women: Analysis and prediction using data mining techniques. In 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon) (pp. 194-196). IEEE.

[3] F. Kamoun, F. Iqbal, M. A. Esseghir and T. Baker, "AI and machine learning: A mixed blessing for cybersecurity," 2020 International Symposium on Networks, Computers and Communications (ISNCC), 2020, pp. 1-7, doi: 10.1109/ISNCC49221.2020.9297323.

[4] Fatima Dakalbab, Manar Abu Talib, Omnia Abu Waraga, Ali Bou Nassif, Sohail Abbas, Qassim Nasir, Artificial intelligence & crime prediction: A systematic literature review, Social Sciences & Humanities Open, Volume 6, Issue 1, 2022, 100342, ISSN 2590-2911, https://doi.org/10.1016/j.ssaho.2022.100342.

[5] Shah, N., Bhagat, N. & Shah, M. Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. Vis. Comput. Ind. Biomed. Art 4, 9 (2021). https://doi.org/10.1186/s42492-021-00075-z

[6] Harris, H. (2022). Artificial Intelligence and Policing of Financial Crime: A Legal Analysis of the State of the Field. In: Goldbarsht, D., de Koker, L. (eds) Financial Technology and the Law. Law, Governance and Technology Series, vol 47. Springer, Cham. https://doi.org/10.1007/978-3-030-88036-1_12

[7] M. Kim, J. Suh and H. Kwon, "A Study of the Emerging Trends in SIM Swapping Crime and Effective Countermeasures," 2022 IEEE/ACIS 7th International Conference on Big Data, Cloud Computing, and Data Science (BCD), 2022, pp. 240-245, doi: 10.1109/BCD54882.2022.9900510.

[8] J. Nicholls, A. Kuppa and N. -A. Le-Khac, "Financial Cybercrime: A Comprehensive Survey of Deep Learning Approaches to Tackle the Evolving Financial Crime Landscape," in IEEE Access, vol. 9, pp. 163965-163986, 2021, doi: 10.1109/ACCESS.2021.3134076.

[9] M. RAZA, N. D. Jayasinghe and M. M. A. Muslam, "A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms," 2021 International Conference on Information Networking (ICOIN), 2021, pp. 327-332, doi: 10.1109/ICOIN50884.2021.9334020.

[10] M. Bada and J. R. C. Nurse, "Profiling the Cybercriminal: A Systematic Review of Research," 2021 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), 2021, pp. 1-8, doi: 10.1109/CyberSA52016.2021.9478246.