JOURNAL OF CRITICAL REVIEWS

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

CRIME DATA ANALYSIS USING MACHINE LEARNING

C. Gazala Akhtar[1], B. Srija[2], D. Sneha[2], D. Deekshitha [2], T. Bhavana[2]

[1]Assistant Professor, [2]UG Scholar, [1,2]Department of CSE-Cyber Security

[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 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 an 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. Instead of focusing on 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: FBI crime data, machine learning, crime analysis.

1. INTRODUCTION

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, the modern technologies and hi-tech methods help criminals in achieving their goals. According to the Nigeria police, crimes like burglary, arson and it likes 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 prediction based on the probabilities of its occurrence. The predicted results cannot be assured of 100% accuracy but these results shows that the software program helps in reducing crime rate to a certain extent by providing security in vital areas where crimes can easily occur. Developing such a powerful crime analytics tool we have to collect crime records and evaluate them. Criminal, and sociology scholars are analysing 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 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 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 of 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 for 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

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

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.

3. PROPOSED SYSTEM

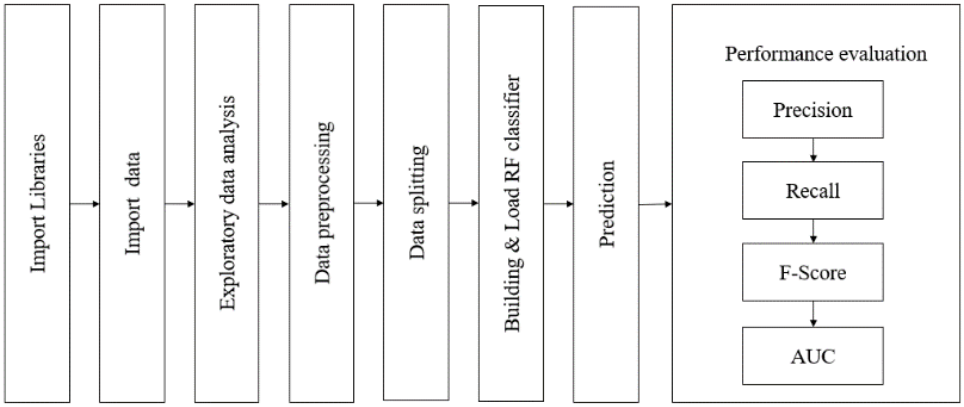


Fig. 1: Block diagram of proposed system.

Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Step 1: In Random Forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

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

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

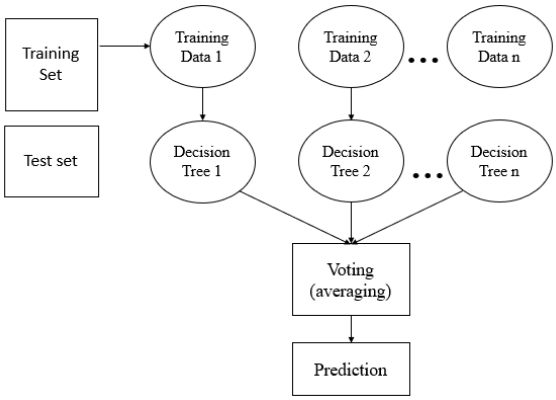


Fig. 2: Random Forest algorithm.

Important Features of Random Forest

- **Diversity**- Not all attributes/variables/features are considered while making an individual tree, each tree is different.
- **Immune to the curse of dimensionality**- Since each tree does not consider all the features, the feature space is reduced.
- **Parallelization**-Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.
- **Train-Test split**- In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.
- **Stability**- Stability arises because the result is based on majority voting/ averaging.

Assumptions for Random Forest

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random Forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

Below are some points that explain why we should use the Random Forest algorithm

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

Types of Ensembles

Before understanding the working of the random forest, we must look into the ensemble technique. Ensemble simply means combining multiple models. Thus, a collection of models is used to make predictions rather than an individual model. Ensemble uses two types of methods:

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

Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest. Bagging, also known as Bootstrap Aggregation is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation.

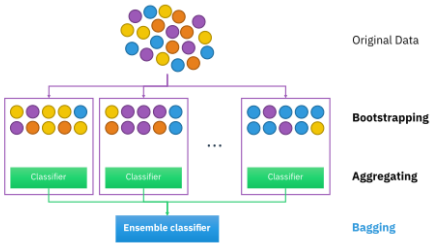


Fig. 3: RF Classifier analysis.

Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

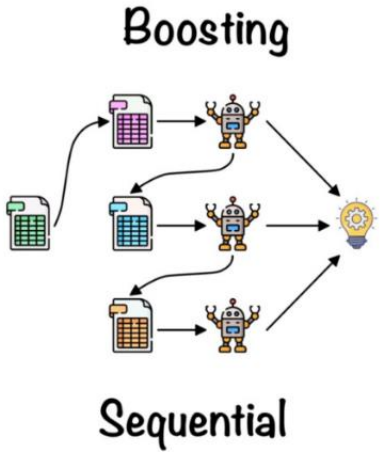


Fig. 4: Boosting RF Classifier.

Advantages of proposed system

- It can be used in classification and regression problems.
- It solves the problem of overfitting as output is based on majority voting or averaging.
- It performs well even if the data contains null/missing values.
- Each decision tree created is independent of the other thus it shows the property of parallelization.
- It is highly stable as the average answers given by a large number of trees are taken.
- It maintains diversity as all the attributes are not considered while making each decision tree though it is not true in all cases.
- It is immune to the curse of dimensionality. Since each tree does not consider all the attributes, feature space is reduced.

4. RESULTS AND DISCUSSION

Sample dataset

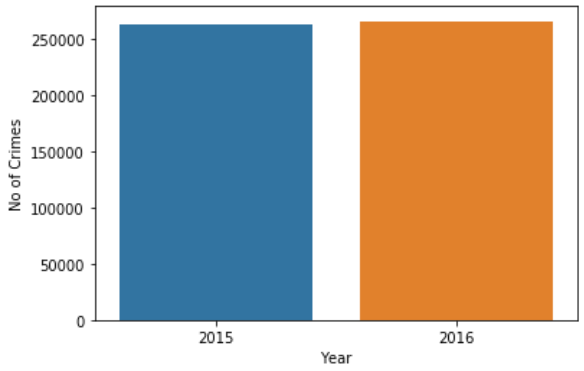
	Unnamed: 0	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	...	Ward	Community Area	FBI Code	X Coordinate
0	3	10508693	HZ250496	05/03/2016 11:40:00 PM	013XX S SAWYER AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE	APARTMENT	True	...	24.0	29.0	08B	1154907.0
1	89	10508695	HZ250409	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

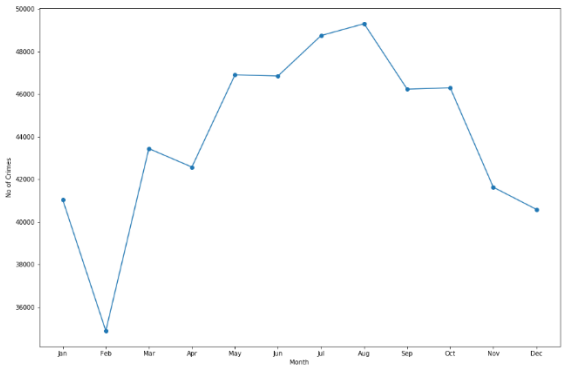
Y Coordinate	Year	Updated On	Latitude	Longitude	Location
1893681.0	2016	05/10/2016 03:56:50 PM	41.864073	-87.706819	(41.864073157, -87.706818608)
1864330.0	2016	05/10/2016 03:56:50 PM	41.782922	-87.604363	(41.782921527, -87.60436317)
1904819.0	2016	05/10/2016 03:56:50 PM	41.894908	-87.758372	(41.894908283, -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)

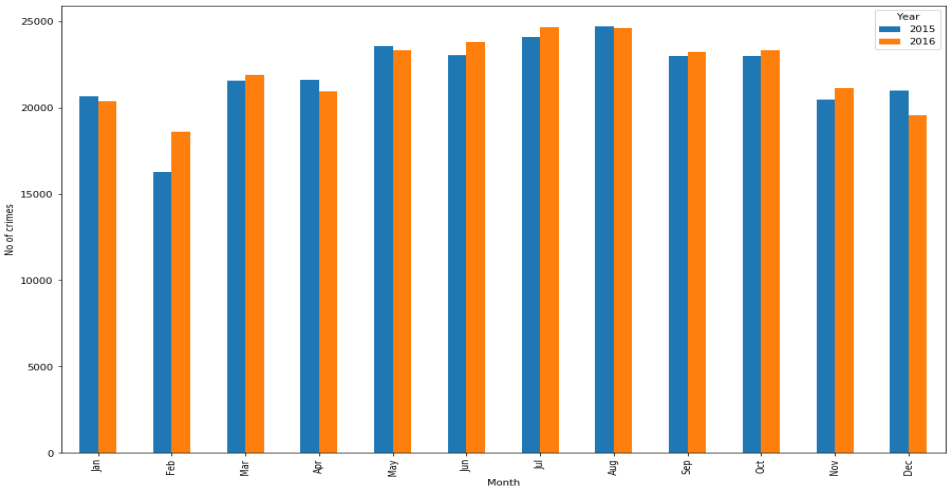
List of Attributes

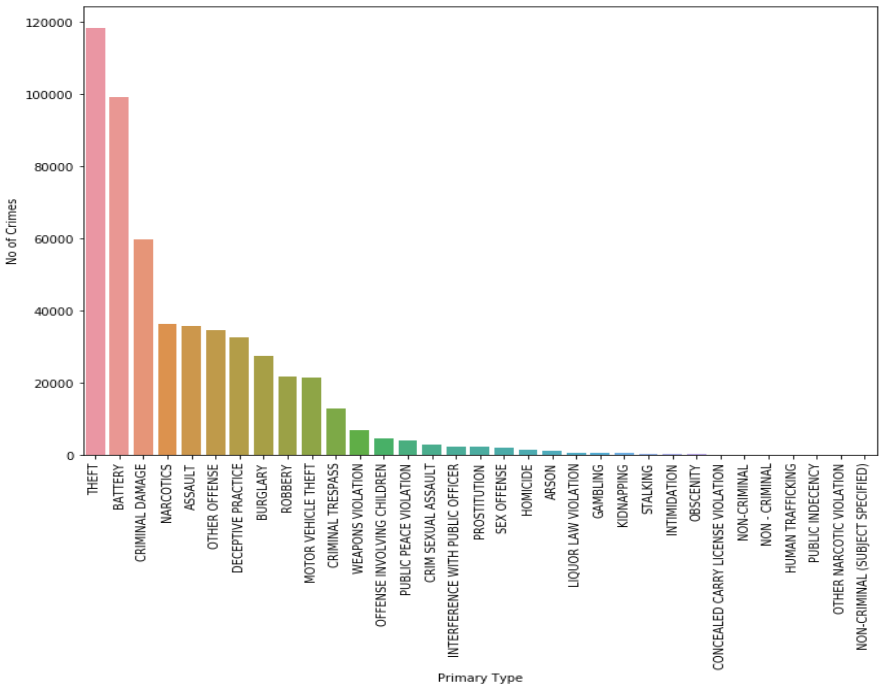
```
['Unnamed: 0',
 'ID',
 'Case Number',
 'Date',
 'Block',
 'IUCR',
 'Primary Type',
 'Description',
 'Location Description'
 'Arrest',
 'Domestic',
 'Beat',
 'District',
 'Ward',
 'Community Area',
 'FBI Code',
 'X Coordinate',
 'Y Coordinate',
 'Year',
 'Updated On',
 'Latitude',
 'Longitude',
 'Location']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



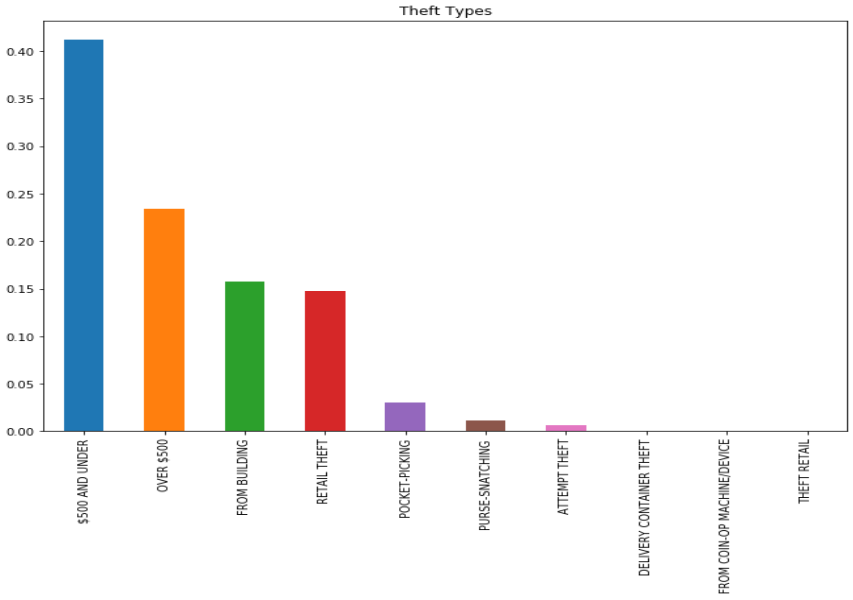
ISSN- 2394-5125 VOL 10, ISSUE 04, 2023

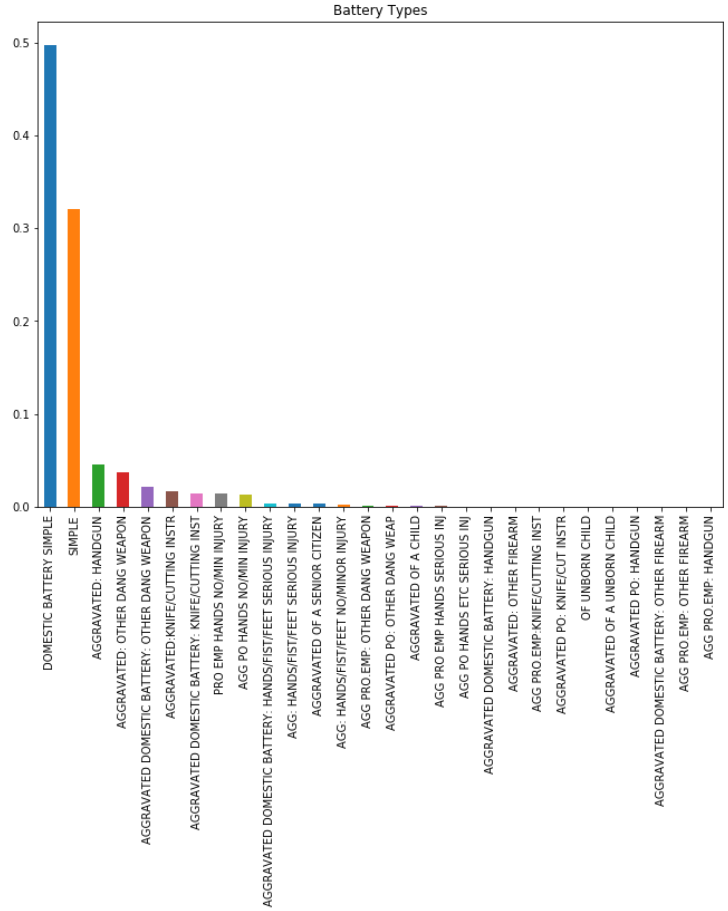




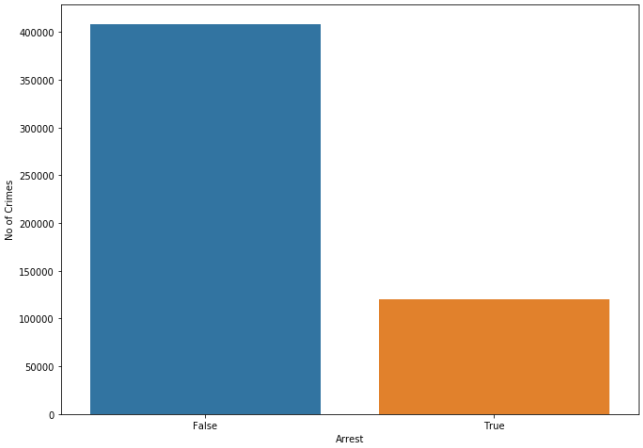


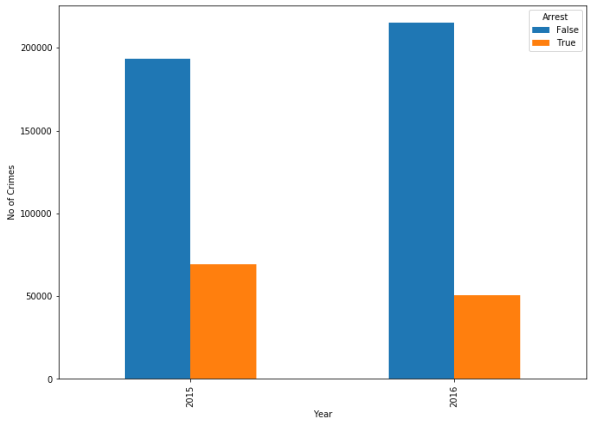
ISSN- 2394-5125 VOL 10, ISSUE 04, 2023

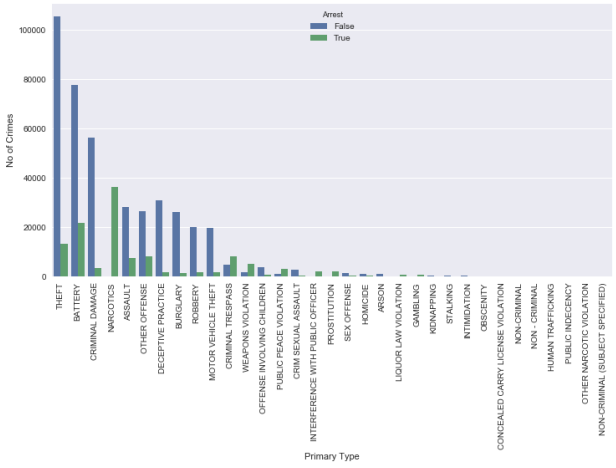


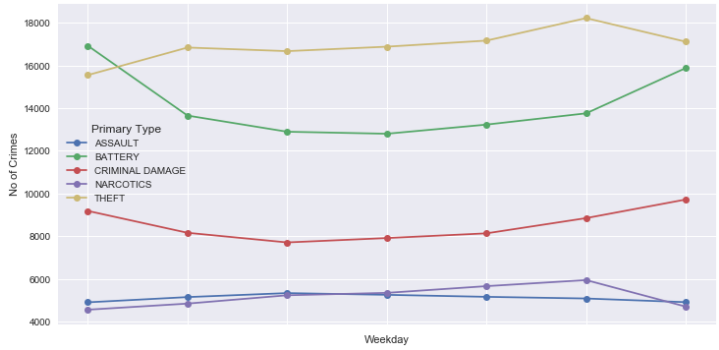


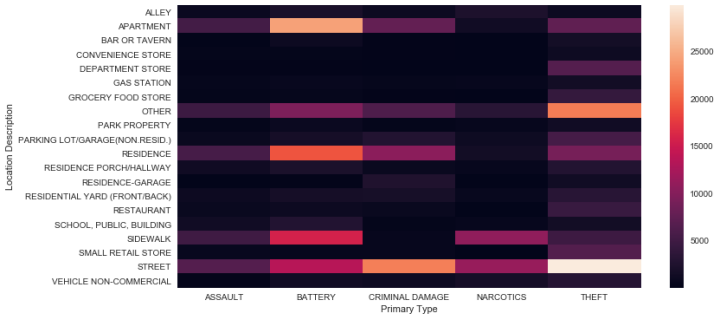
ISSN- 2394-5125 VOL 10, ISSUE 04, 2023











Model splitting, and building

```
from sklearn.model_selection import train_test_split
```

```
test_size = 0.25
```

```
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=test_size)
print('Partitioning Done!')
```

```
Partitioning Done!
```

```
from sklearn.ensemble import RandomForestClassifier
RF= RandomForestClassifier(max_depth=5,n_estimators=100)
```

```
RF.fit(x_train,y_train)
```

```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=5, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=100,
                       n_jobs=None, oob_score=False, random_state=None,
                       verbose=0, warm_start=False)
```

Accuracy performance

```
RF.score(x,y)
```

```
0.8591857619438064
```

```
RF.score(x_train,y_train)
```

```
0.8590464983820391
```

```
RF.score(x_test,y_test)
```

```
0.8596035526291083
```

Classification report

	precision	recall	f1-score	support
0	0.86	0.98	0.92	127356
1	0.85	0.38	0.52	32413
accuracy			0.86	159769
macro avg	0.85	0.68	0.72	159769
weighted avg	0.86	0.86	0.84	159769

5. CONCLUSION

This paper used to predict the type of crime that might occur based on time and location. The algorithm involving trees showed that the predicted results is very much closer to the actual results. Thus, the dataset used, provides the maximum correct result with higher accuracy when implemented with different tree classifiers. The stated results in this paper show that Random Forest method works best and AdaBoost works least well for predicting crimes using time and location. The results in this paper provides similar results when implemented with tree-based algorithms.

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.