A Proposed Big Data Analytics Model for Crimes Predication based on Spatial and Temporal Criminal Hotspot

By

Salah El-Din Abd El-Mohaimen Ibrahim & Prof. Christina Albert Reyad

Computer and Information System Department, Sadat Academy for Management Science, Cairo, Egypt

SalahEl-Din@Outlook.Com, sams.christina.albert@gmail.com

Abstract

Big data is simply large, voluminous data collected from different sources which could be structured or unstructured. The ancient processing system may not be successful in processing such voluminous data. Big data analytics (BDA) uses extensive techniques and tools for analyzing large. This Paper introduced a solution to the crime prediction problem using machine learning algorithms. to classify this monstrous data. SVM and K-NN, Random Forest, and MLP algorithms solve this problem. In this Paper addresses, the researcher finds the best SVM algorithm, the main thing to keep in mind here is that these are to give you the most accurate answer possible as quickly as possible. The Predictive model shows that the prophet model handles outliers well and it is robust to missing data and shifts in the trend. These outcomes are going to benefit the police departments to better understand crime issues and provide insights that will enable them to track activities, predict the likelihood of incidents, effectively deploy resources, and optimize the decision-making process. After applying the proposed model to many machine learning algorithms, the paper concluded that the best crime prediction algorithm in terms of accuracy is in order as follows: SVM algorithm with an accuracy of 0.99997, time 1100 seconds, followed by K-NN algorithm with an accuracy of 0.999976, time 2450 seconds, then Random Forest algorithm with accuracy 0.999996, Time 1420 seconds, then MLP algorithm with accuracy 0.995886, time 1346 seconds. The best algorithm to achieve the best accuracy in the least time is SVM.

Keywords: Big Data, Machine Learning Algorithms, K-NN, Random Forest, Adaboost, SVM , MLP , XGBoost , Logistic Regression and Gradient boost.

1. Introduction:

Big Data is an emerging term in information and communication technologies (ICT) nowadays. The term indicates the increasing size of data in addition to the complexity which requires different processing and analytical tools. This paper aims at discovering knowledge out of Data and presenting it in a form that is easily comprehensible to humans. The exponentially increasing population leads to an increase in crime and in turn generates a huge chunk of data that could be analyzed for the government to make critical and essential decisions to maintain law and order. With the increasing concern about the crime rate, this becomes really necessary. [Nadathur A S, et al, 2018]

Criminal activity is inevitably a part of urban life. All city dwellers, therefore, have an interest in the improvement of our understanding of crime and its patterns. Police departments in particular could use this improved understanding to allocate their resources more and better serve their communities. Knowing the spatial and temporal patterns of criminal activity would allow police to deploy the right officers where and when they are most needed, and being able to predict criminal activity would allow them to anticipate and combat surges in crime. [Cherian J, and Dawson M, 2015]

Crimes are one of the major threats to society and also to civilization. The traditional crime-solving techniques are unable to live up to the requirement of existing crime scenarios. The most challenging area in these crimes is identifying the sets of crimes committed by the same individual or same group. Criminals are humans. Then they tend to do the same work in the same way. Surveys state that 50% of crimes are done by 10% of criminals. [Waduge N, 2017]

Exploring criminal behavior is a key issue in criminology. The factors that reinforce violent criminal behavior are not taken seriously. Appropriate understanding and interpretation of this motivational procedure are critical. [Saeed U et al, 2015]

Before machine learning, the first techniques simply extrapolated historical trends. This works well at very long-time scales but obviously cannot predict deviations from these trends. Slightly more advanced techniques, such as multivariate regression, did improve predictive power, but still only relied on historical crime data. There are other factors that are known to affect crime rates, such as weather and socioeconomic factors (Cohn, 1990; Carlen, 1988), and the mentioned methods do not take these into account. The use of additional datasets does cause the number of features to rise greatly, and newer machine learning methods are better suited to handle a large number of features and find the complex relationships between these factors. [Stec A and Klabjan, D. 2018], Machine learning has become a vital part of crime detection and prevention. [McClendon L and Meghanathan N, 2015]

Machine Learning can be used to identify the patterns of crimes. The data to feed this Machine Learning approach can be taken from past crime records, social media sentiment analysis, weather data, etc. There are five steps in crime prediction using Machine Learning. Those are data collection, data classification, pattern identification, prediction, and visualization. The resources of law enforcement authorities can be used effectively by using crime prediction methods [Waduge N, 2017]

2. Objective of Big Data Analytics Model for Predication based on Spatial and Temporal Criminal Hotspot

The main objective of this paper is to design and implement a new model that will solve these problems by applying machine learning algorithms to a dataset of criminal activity to predict attributes and event outcomes. also, will make a comparison between different machine learning algorithms techniques and achieve a fair judgment between these techniques.

The core of these objectives include:

- Improve crime prediction accuracy.
- Using a machine learning algorithm for giving higher accuracy in predicting crimes.
- Find locative and temporal criminal hotspots using real-world datasets of crimes.
- Identify the location of the crime and when they occur frequently.
- Paper of criminal activity in a geographic area helps to understand the underlying pattern of the crime the area suffers from.
- Predict what type`of crime might occur next in a specific location within a particular time.

This paper is to improve the quality of prediction processes in the datasets. This comparative paper seeks out a fair judgment and knows which algorithm is best to fit in with the introduced dataset. The paper moved on to certain phases described in the model of the experimentation as shown in Fig.1. The model includes many phases: data collection, data preprocessing and selection, transformation phase, selection of Big Data mining tool, selection of programming language, and selection of data mining algorithms depending on the dataset binary classification or multi-classification.

While processing the four phases, this paper starts with data collection, data preprocessing, and filtering to clean, integrate, and transform the data. Then, the required fields for data mining are selected by the programming language. The data was transformed into a certain file format, which is acceptable by Big Data mining tools. The Big Data mining tools were different data mining algorithms based on the dataset Multi classification which are tested by Python.

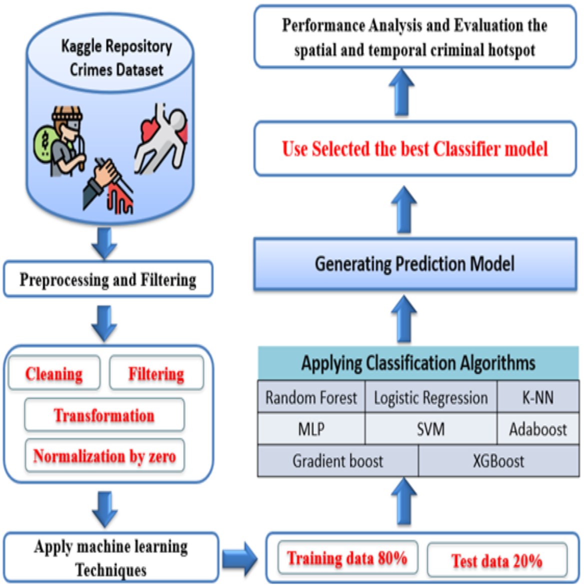
Kaggle Repository
Crimes Dataset
Preprocessing and Filtering
Cleaning
Filtering
Transformation
Normalization by zero
Apply machine learning Techniques
Training data 80%
Test data 20%
Applying Classification Algorithms
Random Forest
Logistic Regression
K-NN
MLP
SVM
Adaboost
Gradient boost
XGBoost
Generating Prediction Model
Use Selected the best Classifier model
Performance Analysis and Evaluation the spatial and temporal criminal hotspot

Fig. 1 The conceptual model of the comparative paper with different classifiers.

2.1 Data Collection and Selection Phase

In this experiment, the dataset is a big dataset which is obtained from Kaggle www.kaggle.com. The dataset contains more than 6,000,000 records/ rows of data and cannot be viewed in full in Microsoft Excel., and 23 features, three classes. the datasets, feature extraction, and feature selection should be done properly. as shown in table 1.

Table 1 The description of attribute of the dataset.

ID - Unique identifier for the record.
Case Number - The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.
Date - Date when the incident occurred. this is sometimes a best estimate.
Block - The partially redacted address where the incident occurred, placing it on the same block as the actual address.
IUCR - The Illinois Uniform Crime Reporting code. This is directly linked to the Primary Type and Description. See the list of IUCR codes at https://data.cityofchicago.org/d/c٧ck٤٣٨-e.
Primary Type - The primary description of the IUCR code.
Description - The secondary description of the IUCR code, a subcategory of the primary description.
Location Description - Description of the location where the incident occurred.
Arrest - Indicates whether an arrest was made.
Domestic - Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.
Beat - Indicates the beat where the incident occurred. A beat is the smallest police geographic area - each beat has dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has ٢٢ police districts. See the beats at https://data.cityofchicago.org/d/aerh-rz٧٤.
District - Indicates the police district where the incident occurred. See the districts at https://data.cityofchicago.org/d/fthy-xz٣r.
Ward - The ward (City Council district) where the incident occurred. See the wards at https://data.cityofchicago.org/d/sp٦-٣٤z٧٦.
Community Area - Indicates the community area where the incident occurred. Chicago has ٧٧ community areas. See the community areas at https://data.cityofchicago.org/d/cauq٨-yn٦.
FBI Code - Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS). See the Chicago Police Department listing of these classifications at http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html.
X Coordinate - The x coordinate of the location where the incident occurred in State Plane Illinois East NAD ١٩٨٣ projection. This location is shifted from the actual location for partial redaction but falls on the same block.
Y Coordinate - The y coordinate of the location where the incident occurred in State Plane Illinois East NAD ١٩٨٣ projection. This location is shifted from the actual location for partial redaction but falls on the same block.
Year - Year the incident occurred.
Updated On - Date and time the record was last updated.
Latitude - The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
Longitude - The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
Location - The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.

2.2 Selection of Big Data Mining Tool

Analyzing Big Data can be very cumbersome and challenging. There is no particular software that can be used for the analysis. Different enterprises use different tools for Big Data analysis. However, the tool to use depends on the type of data one needs to analyze. The choice of tools can also affect the quality of your data which can have a significant impact on your analysis. In this paper, some tools can be used to analyze both structured and unstructured data (Big Data).

Python is an easy to learn, powerful programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python's elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms (Python Documentation: https://docs.python.org/2/tutorial/). A language excels at string processing - that is the manipulation of string lists a few languages with good string processing capabilities and compares them in terms of the degree to which they are still being actively developed by a community of developers testing whether they are object-oriented.

3. Machine Learning Algorithms

Machine learning is basically an artificial intelligence (AI) wherein a machine can learn on its own code ability without providing the explicit programmer. The main motive is that when a problem is countered it doesn't write the program again, but it changes its own code according to the new scenario that was discovered. Itself learned what must be learned from provided data scenario, past expressions, and learning past experiences it comes up with a new situation. [Kumar R and Nagpal B, 2019]

K Nearest Neighbor Classifier is a supervised machine learning algorithm useful for classification problems. It works by finding the distances between a query and all the examples in the data, selecting the specified examples that are closest to the query, and then voting for the most frequent label.[Kumar A et al, 2020]

Random forests are ensembles of decision trees. Random forests are one of the most successful machine learning models for classification. They combine many decision trees to reduce the risk of over-fitting.[kumar LS et al, 2022]

Support Vector Machine is one of the most popular a supervised machine learning algorithm, that can be used for both classification and regression problems. But it is preferred for classification problems.[Cervantes J et al, 2020]

Logistic regression is a statistical model that is used to model the probability of a certain class or event existing.it is used for predicting categorical dependent variable using another set of independent variables. [Louise A et al, 2021]

3.1 Evaluation

The comparison of the different machine learning algorithms mentioned in Section 3 is based on the following measured parameters [[Andrew M and Brynjolfsson E, 2012]

Training Time: As an accomplished machine learning algorithm are measured. Time taken to build the model is called training time. This varies on the implementations of the algorithms.

Accuracy: It is the percentage of the correctly classified instances which are the total number of predictions that were correct.

Precision: It is the element of the identified items that are correct and used to measure how well the proposed algorithm matches with the ground truth. It's likewise known as Positive Predictive Value (PPV).

Recall: It is another measure used to compute how the proposed algorithm matches the ground truth. Recall, or sensitivity or consistently True Positive

Rate (TPR), which is a measure of the number of true positives relative to the sum of the true positives and the false negatives. It is the element of items that were correctly detected among all the items that should have been detected.

F-measure: It is sensitivity, an overall measure of how well we have been able to classify the ground truth foregrounds and backgrounds.

From the previously mentioned measured parameters, we can compare the accuracies provided by all the algorithms on a dataset. Here, the focus is mainly on comparing major parameters like accuracy and training time in order to decide which machine learning algorithm is better suited for a selected type of data.

3.2 Experimental Result and Analysis

According to the previous techniques which have been mentioned in the previous section, these classification techniques will be applied to many different data sets with different sizes so that a comparison paper could be done. Evaluation will depend on many standards, but the main important ones are accuracy and time to build the model.

Certain comparative studies conducted earlier have shown that a particular algorithm has performed better on its data set and its conclusions; however, it differs from each other. So, to apply what has been mentioned Python Language will be used to get the required results of the comparative paper.

The analysis and studies either have used a very minimal set of classifiers or have used data sets that are diverse resulting in an advantage or bias for a particular algorithm. Keeping that in mind, we have included a good number of classifiers in our analysis and used data sets that are diverse. The following sections describe the results obtained in our analysis. Finally, this comparison paper was made in order to make a fair judgment and know which algorithm best is to fit with the introduced dataset.

3.3 Classification Results Using Random Forest Algorithm

In this paper results show a Random Forest classification algorithms result, First, in order to measure the model performance, The dataset was split into two parts, a training set which was 80% used to train the model of the actual dataset and 20% was used as a testing set to train the model. In order to evaluate the performance of the algorithms by several multiclass classification evaluations and regression metrics, in order to measure its accuracy, precision, recall, F1-Score, false positives rate (FPR), and true positives rate (TPR), which are defined by equations by means of a confusion matrix and finally, Time is taken to build model per second.

Table 2 Comparison of several different Classes random forest.

Classes	Precision	Recall	F-١Score	Accuracy
Primary type	0.9999227851504006	0.9999227846205683	0.9999227846205683	0.9999227846205683

In the fig.2. presents more than of classes in the primary type of the result of the technique can be show the following:

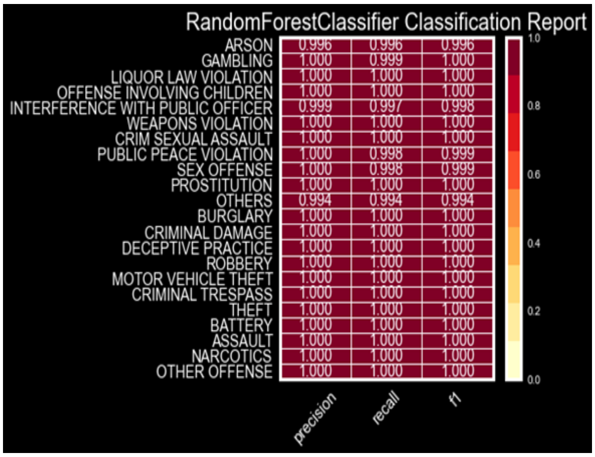

Fig. 2. Random Forest Classifier Classification Report

3.4 Classification Results Using Logistic Regression

In this paper results show a Logistic Regression classification algorithms result, First, in order to measure the model performance, The dataset was split into two parts, a training set which was 80% used to train the model of the actual dataset and 20% was used as a testing set to train the model. In order to evaluate the performance of the algorithms by several multiclass classification evaluations and regression metrics, in order to measure its accuracy, precision, recall, F-Measure, false positives rate (FPR), and true positives rate (TPR), which are defined by equations by means of a confusion matrix and finally, Time is taken to build model per second.

Table 3 Comparison of several different Classes Logistic Regression

Classes	Precision	Recall	F-۱Score	Accuracy
Primary type	0.637953909124772	0.6847190764104676	0.6847190764104676	0.6847190764104676

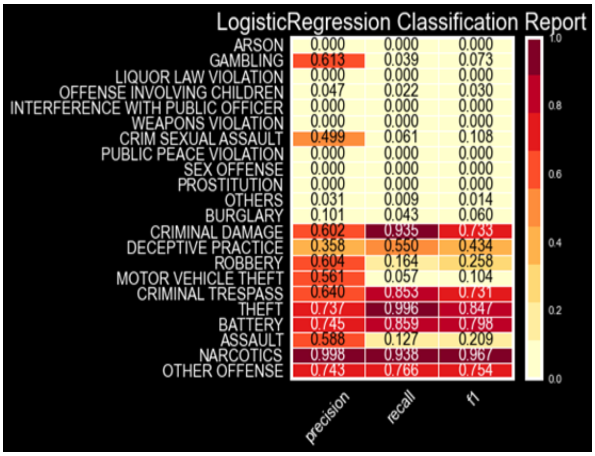

Fig. 3. Logistic Regression Classifier Classification Report and Visualizer

3.5 Classification Results Using K-NN

In this paper results show a K-NN classification algorithms result, First, in order to measure the model performance, The dataset was split into two parts, a training set which was 80% used to train the model of the actual dataset and 20% was used as a testing set to train the model. In order to evaluate the performance of the algorithms by several multiclass classification evaluations and regression metrics, in order to measure its accuracy, precision, recall, F-Measure, false positives rate (FPR), and true positives rate (TPR), which are defined by equations by means of a confusion matrix and finally, Time is taken to build model per second.

Table 4 Comparison of several different Classes KNN

Classes	Precision	Recall	F-۱Score	Accuracy
Primary type	0.9999613920428628	0.9999613923102841	0.9999613923102841	0.9999613923102841

In the fig. 4 presents more than of classes in the primary type of the result of the technique can show that in fig.4.

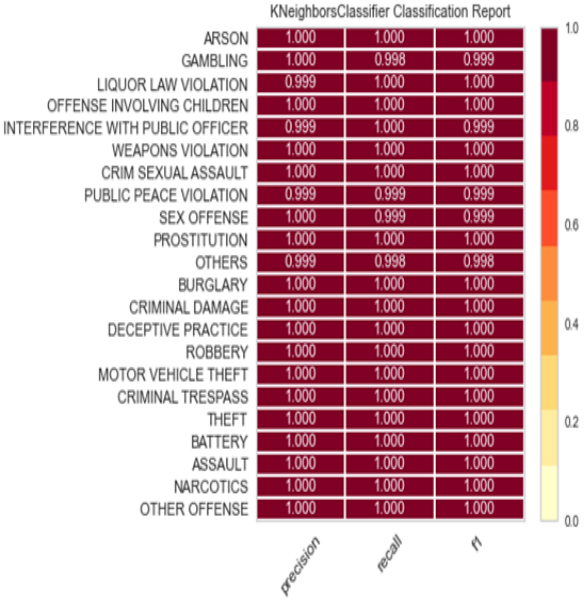

Fig. 4. KNN Classifier Classification Report and Visualizer

3.6 Classification Results Using MLP Classifier

In this paper results show an MLP classification algorithms result, First, to measure the model performance, The dataset was split into two parts, a training set which was 80% used to train the model of the actual dataset, and 20% was used as a testing set to train the model. To evaluate the performance of the algorithms by several multiclass classification evaluations and regression metrics, in order to measure its accuracy, precision, recall, F-Measure, false positives rate (FPR), and true positives rate (TPR), which are defined by equations by means of a confusion matrix and finally, Time is taken to build model per second.

Table 5 Comparison of several different Classes MLP

Classes	Precision	Recall	F-١Score	Accuracy
Primary type	0.995886918669546	0.995250084234959‎4	0.9952500842349594	0.9952500842349594

In the figure 5 presents more than of classes in the primary type of the result of the technique can show that in fig.5.

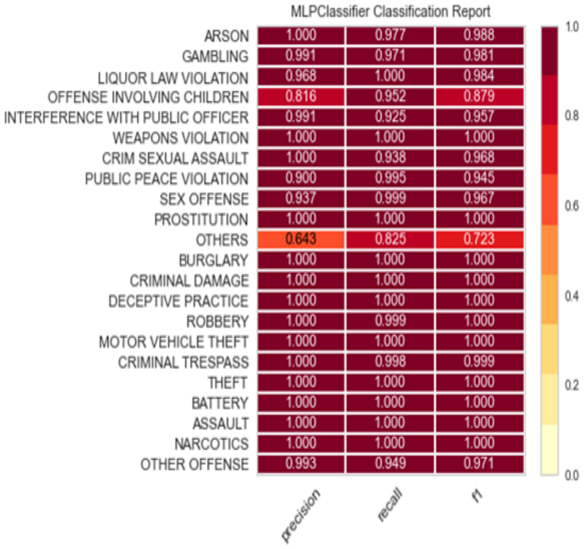

Fig. 5 MLP Classifier Classification Report and Visualizer

3.7 Classification Results Using SVC Classifier

In this paper Results, show an SVC classification algorithms result, First, to measure the model performance, The dataset was split into two parts, a training set which was 80% used to train the model of the actual dataset and 20% was used as a testing set to train the model. To evaluate the performance of the algorithms by several multiclass classification evaluation and regression metrics, in order to measure its accuracy, precision, recall, F-Measure, false positives rate (FPR) and true positives rate (TPR), which are defined by equations by means of a confusion matrix and finally, Time taken to build model per second.

Table 6 Comparison of several different Classes SVC

Classes	Precision	Recall	F-١Score	Accuracy
Primary type	0.9999719233094674	0.9999719216802067	0.9999719216802067	0.9999719216802067

In the fig.6 presents more than of classes in the primary type of the result of the technique can show that in fig.6.

Fig. 6 SVC Classifier Classification Report and Visualizer

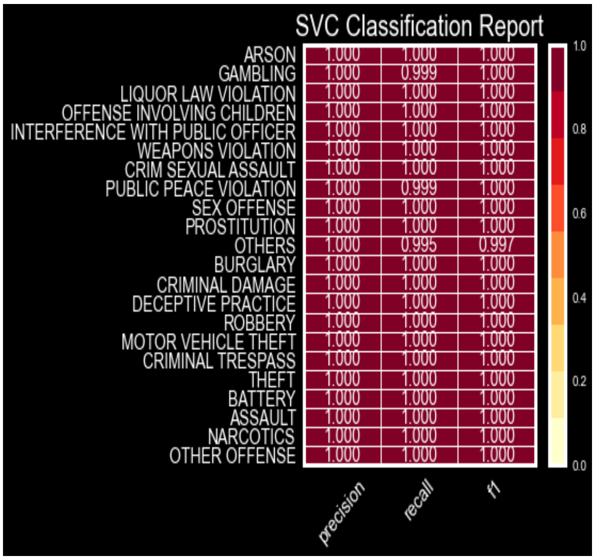

Fig. 6 SVC Classifier Classification Report and Visualizer

4 Conclusion

In this paper, Big Data analytics has outlined the challenges and the methods used for performing different operations, and it compared the most popular algorithms in Big Data analytics.

This paper offers a way to foresee and predict crimes and fraud within a city. It focuses on having a crime prediction tool that can be helpful to law enforcement and reduce crimes for the security authorities. As compared to the previous work, this work was successful in achieving the highest accuracy in prediction. Along the way, many patterns of criminal activities in various areas which will be helpful for criminal investigation were known. This pattern has much greater importance than we realize. The SVM system helps law implementing agencies with improved and exact crime analysis because achieves Hight accuracy (.099997) and less time. By traversing through the crime dataset, we must find out different reasons that lead to crime. Since in this paper is bearing in mind only some limited factors, full accuracy cannot be accomplished. For getting more accurate results in prediction we must find out more crime attributes of places instead of setting only certain attributes. Thus far this system was trained using certain attributes, but we can take into account more factors to improve accuracy. In the future, this work can be stretched to have developed classification algorithms to detect criminals more efficiently. The crime rates that are increasing non-stop may go down in the future due to such prediction techniques.

5 References

1. Nadathur A S, Narayanan G, Ravichandran I, Srividhya.S and Kayalvizhi.J, (2018), "Crime Analysis and Prediction Using Big Data", International Journal of Pure and Applied Mathematics, Volume 119 No. 12, 207211-

2. Cherian J, and Dawson M. (2015), RoboCop: Crime Classification and Prediction in San Francisco, leland stanford junior university.

3. Waduge N, (2017), "Machine Learning Approaches for Detect Crime Patterns", University of Moratuwa, DOI:10.13140/RG.2.2.11794.66246

4. Saeed U, Sarim M, Usmani A, Mukhtar A, Shaikh A B and Raffat Sh K, (2015), "Application of Machine learning Algorithms in Crime Classification and Classification Rule Mining", Research Journal of Recent Sciences, Vol. 4(3), 106114-

5. Stec A and Klabjan, D. (2018). Forecasting crime with deep learning. arXiv preprint arXiv:1806.01486.

6. McClendon L and Meghanathan N, (2015), "Using Machine Learning Algorithms to Analyze Crime Data", Machine Learning and Applications: An International Journal (MLAIJ) Vol.2, No.1

7. Kumar R and Nagpal B, (2019), "Analysis and prediction of crime patterns using big data" Institute of Computer Applications and Management, Int. j. inf. tecnol. 11:799-805

8. Kumar A, Verma A, Shinde G, Sukhdeve Y and N. Lal, (2020), «Crime Prediction Using K-Nearest Neighboring Algorithm,» International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020, pp. 14-, doi: 10.1109/ic ETITE47903.2020.155.

9. kumar LS, Pandimurugan V, Usha D, Guptha M N, Hema M.S, (2022)," Random Forest tree classification algorithm for predicating loan, Materials Today: Proceedings, Volume 57, Part 5, Pages 22162222-, ISSN 22147853-

10. Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A, (2020) A comprehensive survey on support vector machine classification: Applications, challenges and trends, Neurocomputing, Volume 408, Pages 189215-, ISSN 09252312-,

11. Louise A. Brown Nicholls, Allyson J. Gallant, Nicola Cogan, Susan Rasmussen, David Young, Lynn Williams, (2021), "Older adults› vaccine hesitancy: Psychosocial factors associated with influenza, pneumococcal, and shingles vaccine uptake", Vaccine, Volume 39, Issue 26, Pages 35203527-, ISSN 0264410-X,

12. Andrew M and Brynjolfsson E, (2012), Big Data: The Management Revolution, Harvard Business Review