Prediction of Crime Occurrence using Multinomial Logistic Regression

R. Rajadevi, E. M. Roopa Devi, S. Vinoth Kumar

***Abstract:** In order to uncover hidden patterns and correlations, data analysis examines large amounts of data. Analysis of crime isa systematic approach to the identification and analysis of crime patterns and itstrends. This plays a role in the planning of problems with crime and in formulating strategies for crime prevention. Instead of focusing on causes of crime such as criminal offender background, this work focuses primarily crime factors happened on every day. This work can predict the category of crime that has a higher likelihood of occurrence in those areas and can visualize in the form of histogram and heat map by category of crime, crime by day of week and month. The study depends on a lot of variables like class, latitude, longitude, etc. For forecast, the multinomial logistic regression method is used. For weekdays, the district and the hour of the accident are used as predictors.This algorithm is used because its target variable has more than two values and no ordering in the response variable.This provides greater efficiency for handling datasets with multi class labels. This forecast can be helpful in predicting the occurrence of crime in vulnerable areas, which in turn minimizes the crime rate by providing the patrol in those areas.*

***Keywords:** Data Analytics, Prediction, Regression ,Machine Learning.*

I. INTRODUCTION

Big data analytics involves collecting data from different resources to manipulate and then finally deliver as useful products to the organization.

It is useful to integrate raw data acquired from different sources into a data item, forms the core of Big Data Analytics.There are two methodologies in Data Analytics : Exploratory Data Analytics (EDA) and Confirmatory Data Analytics (CDA). EDA is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. We can just see what the data tell us beyond the formal modeling task. In EDA data are explored which can provide information about the numbers of factors required to represent the data. CDA is a multivariate statistical procedure through which we can test how well the measured variables represent the construct. CDA is a tool which can be useful to confirm the measurement theory.

There are five Characteristics which is the building blocks of an efficient data analytics solution: Accuracy, Completeness, Consistency, Uniqueness and Timeliness. There is an another characteristic in data analytics called as Data Visualization which describes the presentation of abstract information in graphical form.

It allows users to spot patterns, trends, and correlations that otherwise might go unnoticed in traditional reports, tables, or spreadsheets. There are two basic types of Data Visualization: Exploration and Explanation. By using these categories we have many ways to make data can be visual. The most common types of data visualization are Heat map Cartogram, Choropleth, Dot Distribution Map, Connected Scatter Plot, Polar Area Diagram, Time Series, Pie Chart, Histogram, Scatter Plot, Dendrogram, Ring Chart, Tree Diagram, Alluvial Diagram, Node-Link Diagram, Matrix. A heat map is a two-dimensional representation of data in which values are represented0020by colors. It provides an immediate visual summary of information. It provides easy understanding of complex data sets.

II. RELATED WORK

. The existing system deals with large set of data and it consist of centralized database. Running Algorithm like Multinomial Logistic Regression has higher time complexity. It consists of 39 categories of crime but classification in somehow difficult. System will consists of poor accuracy and replicated values which will lead to large time consuming. As the system is centralized and it does not distribute the task or data, the retrieval and processing of the system consumes large amount of time. The analysis task is more complex and identifying error rate is difficult.

III. PROBLEM STATEMENT

The existing method is more complex for analysis and also it has the complex structure which provides complex view to the users. The designers of the System felt so difficult for giving such proper working model. It has large space complexity and time complexity. It seems to be difficult to predict the crime and to process the data from the records of crime. It neither displays the result in the pictorial form nor in the comparative manner. The algorithms such as Random Forest and Naive Bayes has higher complexity in both time as well as space The huge data makes higher complexity in prediction

Revised Manuscript Received on January 30, 2020.
* Correspondence Author

R. Rajadevi*, Department of Information Technology, Kongu Engineering College, Perundurai, India. E-mail: rajdevi@kongu.ac.in

E. M. Roopa Devi, Department of Information Technology, Kongu Engineering College, Perundurai, India. E-mail: roopadevi@kongu.ac.in

S. Vinoth Kumar, Department of Information Technology, Kongu Engineering College, Perundurai, India. E-mail: vinoths@kongu.ac.in

Retrieval Number: B7663129219/2020©BEIESP
DOI: 10.35940/ijitee.B7663.019320
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.

IV. PROPOSED SYSTEM

The proposed system is to identify and visualize the occurrence of the crime with higher probability of critical areas. These results are used to predict crime rates in sensitive areas .. The analysis depends on several factors such as latitude and longitude etc.., Data are collected, classified and visualized using graphs. Multinomial logistic regression algorithm is used for prediction. Day of week, District and hour of the incident are used as predictors.

This reason behind in selecting the algorithm is because there are more than two values for target variable and response variables are not ordered. Then the main objective is to predict the area with higher crime rate and to provide security so the crime rate will be reduced. The System has various modules involving Data Visualization and Classification. The data source is the static dataset which is going to be used as a source for the data analysis process. R Studio is used to analyze and visualize the data. The dataset is stored as a csv file for processing which is the first step that has to be done for data processing.

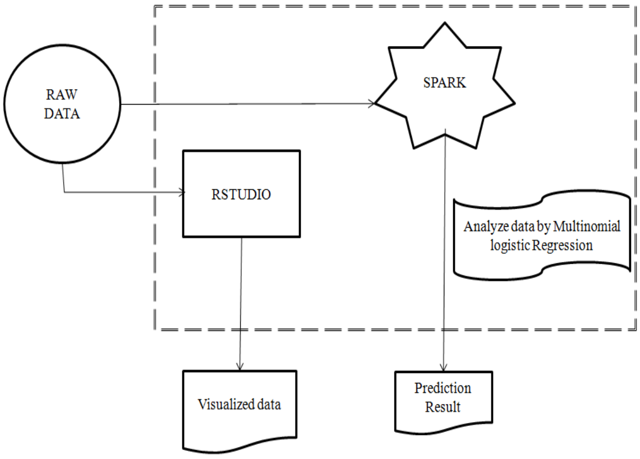


Fig.1.1 Framework to perform classification

Import the dataset in the program for training. The category of crime will be removed and the test dataset will be loaded as a csv file. The dataset will be trained in the desired model by multinomial logistic regression. After running the model for several iterations the output file will be generated with the probabilistic values for various crime categories. This system design consists of all the modules which are going to process the datasets by analytics. The raw data which consists of test data and training data. Working directory is set and the data sets are loaded. The data set is given to spark environment and the day of week, district of the crime and the hour of the incident are used as predictors from the dataset. The data is pre-processed by using the methods that deal with the missing values to maximize the crime rate prediction.

The multinomial regression model learns from the parameters weekday, when the crime occurred regularly, city, hour.

Multinomial Logistic Regression

Relationship of the variable in the dataset is identified by using regression techniques. The categorical variables based on the independent variables can be categorized using Logistic regression .For example to predict whether user will buy a product or not (YES/NO) .Binomial logistic regression model is further modified to categorize the dependent variable which contains nominal values with more than two level using multinomial logistic regression. The dataset contains 39 categorical values like bribery, gambling, runaway, fraud, vehicle theft etc. A multinomial Logistic regression technique is applied to predict the crime based on the input variables.

$$Y1 = Ln(\frac{p2}{p1}) \qquad (1)$$

$$Y2 = Ln(\frac{p3}{p1}) \qquad (2)$$

$$\mathrm{P1} = 1/(1 + \mathrm{e^\wedge y1} + \cdots \mathrm{e^\wedge y24}) \qquad (3)$$

$$\mathrm{P2} = \mathrm{e^\wedge y1}/(1 + \mathrm{e^\wedge y1} + \cdots \mathrm{e^\wedge y24}) \qquad (4)$$

$$Pn = \mathrm{e^\wedge y(n-1)}\ /(1 + \mathrm{e^\wedge y1} + \cdots \mathrm{e^\wedge y2} \qquad (5)$$

The above equation No. 3 to 5 is used to find the probabilities The features in the dataset are given as input to model. The values in the features are converted to numerical values. The linear regression model make use of all the features x1,x2,x3… xn and W is initial weight for all the features W=[w1,w2,w3….wn]. The model generated using the features and weights like w1*x1, w2*x2,w3*x3 …… wn*xn. During the training phase the initial weight is updated. Softmax function is used to find the probability of each feature and it depends on the updated weights. The one hot encoding is applied for conversion of categorical values to binary format. A matrix is constructed with number of input features and class variable. Distance is calculated using entropy function using Softmax and matrix. During training the expected weights will be obtained .Prediction is done using the weights obtained during training phase. The above procedure is repeated for each record in the training data set until loss function value is minimal. After completing the all the iteration weights obtained it utilized for predicting the expected class. This regression model handles larger data sets with higher efficiency. Evaluation metrics are applied to determine the efficiency of the algorithm.

V. RESULT AND DISCUSSION

The San Francisco crime data is collected from the Kaggle website. The algorithm chosen for this solution, is a variation of multinomial logistic regression, a classification model based on regression where the dependent variable (what is used as predictor) is categorical (opposite of continuous), implemented in Spark environment. It provides two dataset: a train data set and a test dataset. The train dataset is made of 878049 records and the test dataset, of 884262 records. Both of them contains incidents from January 1, 2003 to May 13, 2015. After pre-processing the data, the next step is to create and train the model.

The model would forecast the crime class using the weekday when the incident took place, the district where it happened and the time it occurred, as predictors. The following are features found in the crime dataset which is used for prediction of crime .

These are the attributes of the dataset:

ATTRIBUTE	DESCRIPTION
Dates	Indicates the Timestamp of the incident.
Category	Category of the incident.
Descript	A short description of the incident.
DayOfWeek	Incident occurred on this Day of the week
PdDistrict	Under which police department the crime occurred
Resolution	Outcome of the incident.
Address	Address of the incident.
X	Longitude
Y	Latitude

The following fig 5.1 indicates the number of crimes happened during the respective years

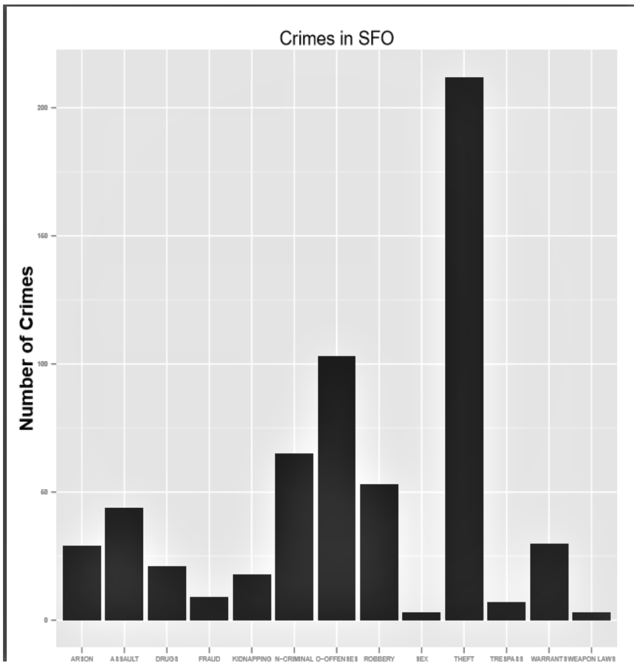
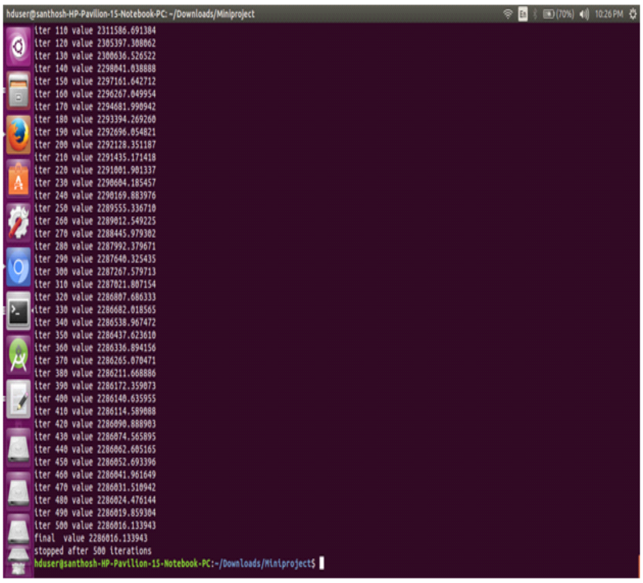
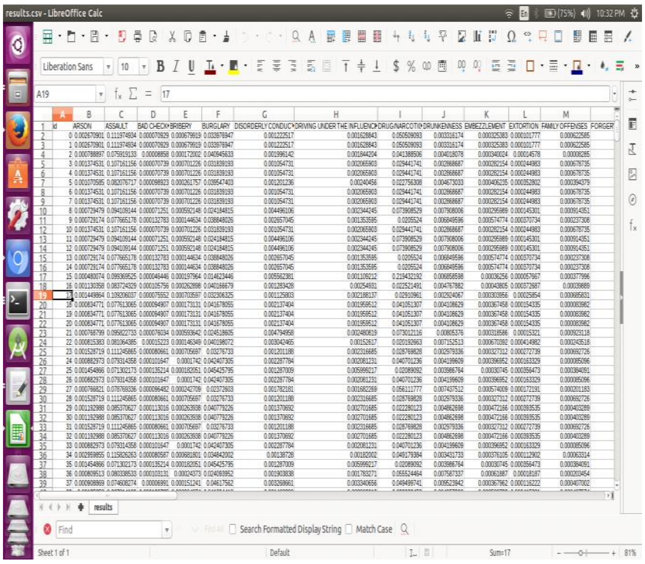


Fig :5.1 Classification of Crimes



```
hduser@santhosh-HP-Pavilion-15-Notebook-PC: ~/Downloads/Miniproject
iter 110 value 2311586.691384
iter 120 value 2305397.308062
iter 130 value 2300636.526522
iter 140 value 2298041.038888
iter 150 value 2297161.642712
iter 160 value 2296267.049954
iter 170 value 2294681.990942
iter 180 value 2293394.269260
iter 190 value 2292696.054821
iter 200 value 2292128.351187
iter 210 value 2291435.171418
iter 220 value 2291001.901337
iter 230 value 2290604.185457
iter 240 value 2290169.883976
iter 250 value 2289555.336710
iter 260 value 2289012.549225
iter 270 value 2288445.979302
iter 280 value 2287992.379671
iter 290 value 2287640.325435
iter 300 value 2287267.579713
iter 310 value 2287021.807154
iter 320 value 2286807.686333
iter 330 value 2286682.018565
iter 340 value 2286538.967472
iter 350 value 2286437.623610
iter 360 value 2286336.894156
iter 370 value 2286265.070471
iter 380 value 2286211.668886
iter 390 value 2286172.359073
iter 400 value 2286140.635955
iter 410 value 2286114.589088
iter 420 value 2286090.888903
iter 430 value 2286074.565895
iter 440 value 2286062.605165
iter 450 value 2286052.693396
iter 460 value 2286041.961649
iter 470 value 2286031.510942
iter 480 value 2286024.476144
iter 490 value 2286019.859304
iter 500 value 2286016.133943
final  value 2286016.133943
stopped after 500 iterations
hduser@santhosh-HP-Pavilion-15-Notebook-PC:~/Downloads/Miniproject$ 
```

Fig: 5.2 Multinomial Regression Values



	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	id	ARSON	ASSAULT	BAD CHECK	BRIBERY	BURGLARY	DISORDERLY CONDUC	DRIVING UNDER THE INFLUENC	DRUG/NARCOTIC	DRUNKENNESS	EMBEZZLEMENT	EXTORTION	FAMILY OFFENSES	FORGER
2	0	0.002670901	0.111974934	0.000070929	0.000679919	0.033976947	0.001222517	0.001628843	0.050509093	0.003316174	0.000325383	0.000101777	0.000622585	
3	1	0.002670901	0.111974934	0.000070929	0.000679919	0.033976947	0.001222517	0.001628843	0.050509093	0.003316174	0.000325383	0.000101777	0.000622585	
4	2	0.000788897	0.075919133	0.00008858	0.000172002	0.040945633	0.001996142	0.001844204	0.041388506	0.004018078	0.000340024	0.00014578	0.00008285	
5	3	0.001374531	0.107161156	0.000070739	0.000701226	0.031839193	0.001054731	0.002065903	0.029441741	0.002868687	0.000282154	0.000244983	0.000678735	
6	4	0.001374531	0.107161156	0.000070739	0.000701226	0.031839193	0.001054731	0.002065903	0.029441741	0.002868687	0.000282154	0.000244983	0.000678735	
7	5	0.001070585	0.082076717	0.000098923	0.000261757	0.039547403	0.001201236	0.00240456	0.022756308	0.004673033	0.000406235	0.000352802	0.000394379	
8	6	0.001374531	0.107161156	0.000070739	0.000701226	0.031839193	0.001054731	0.002065903	0.029441741	0.002868687	0.000282154	0.000244983	0.000678735	
9	7	0.001374531	0.107161156	0.000070739	0.000701226	0.031839193	0.001054731	0.002065903	0.029441741	0.002868687	0.000282154	0.000244983	0.000678735	
10	8	0.000729479	0.094109144	0.000071251	0.000592148	0.024184815	0.004496106	0.002344245	0.073908529	0.007908006	0.000295989	0.000145301	0.000914351	
11	9	0.000729174	0.077665178	0.000132783	0.000144634	0.038848026	0.002657045	0.001353595	0.0205524	0.006849596	0.000574774	0.000370734	0.000237308	
12	10	0.001374531	0.107161156	0.000070739	0.000701226	0.031839193	0.001054731	0.002065903	0.029441741	0.002868687	0.000282154	0.000244983	0.000678735	
13	11	0.000729479	0.094109144	0.000071251	0.000592148	0.024184815	0.004496106	0.002344245	0.073908529	0.007908006	0.000295989	0.000145301	0.000914351	
14	12	0.000729479	0.094109144	0.000071251	0.000592148	0.024184815	0.004496106	0.002344245	0.073908529	0.007908006	0.000295989	0.000145301	0.000914351	
15	13	0.000729174	0.077665178	0.000132783	0.000144634	0.038848026	0.002657045	0.001353595	0.0205524	0.006849596	0.000574774	0.000370734	0.000237308	
16	14	0.000729174	0.077665178	0.000132783	0.000144634	0.038848026	0.002657045	0.001353595	0.0205524	0.006849596	0.000574774	0.000370734	0.000237308	
17	15	0.000480074	0.099369525	0.000046446	0.000197964	0.014623446	0.005562381	0.001109212	0.219432192	0.006858598	0.00036256	0.000057667	0.000377996	
18	16	0.001130358	0.083724329	0.000105756	0.000262898	0.040166679	0.001283428	0.00254931	0.022521491	0.004767882	0.00043805	0.000372687	0.00039889	
19	17	0.001449864	0.109206037	0.000075552	0.000703597	0.032306325	0.001125803	0.002188137	0.02910961	0.002924067	0.000303956	0.00025854	0.000685831	
20	18	0.000834771	0.077613065	0.000094907	0.000173131	0.041678055	0.002137404	0.001959512	0.041051307	0.004108629	0.000367458	0.000154335	0.000083962	
21	19	0.000834771	0.077613065	0.000094907	0.000173131	0.041678055	0.002137404	0.001959512	0.041051307	0.004108629	0.000367458	0.000154335	0.000083962	
22	20	0.000834771	0.077613065	0.000094907	0.000173131	0.041678055	0.002137404	0.001959512	0.041051307	0.004108629	0.000367458	0.000154335	0.000083962	
23	21	0.000768799	0.095822733	0.000076034	0.000593642	0.024518605	0.004794958	0.002480619	0.073012116	0.00805376	0.000318586	0.00015321	0.000923118	
24	22	0.000815383	0.081064385	0.00015223	0.000146349	0.040198072	0.003042465	0.00152617	0.020192663	0.007152513	0.000670392	0.000414982	0.000243518	
25	23	0.001528719	0.111245865	0.000080661	0.000705697	0.03276733	0.001201188	0.002316685	0.028769828	0.002979336	0.000327312	0.000272739	0.000692726	
26	24	0.000882973	0.079314358	0.000101647	0.0001742	0.042407305	0.002287784	0.002081231	0.040701236	0.004199609	0.000396952	0.000163329	0.000085096	
27	25	0.001454866	0.071302173	0.000135214	0.000182051	0.045425795	0.001287009	0.005999217	0.02089092	0.003986764	0.00030745	0.000356473	0.000384091	
28	26	0.000882973	0.079314358	0.000101647	0.0001742	0.042407305	0.002287784	0.002081231	0.040701236	0.004199609	0.000396952	0.000163329	0.000085096	
29	27	0.000766821	0.078769336	0.000096482	0.000242709	0.02372603	0.001782181	0.001682269	0.056111777	0.007437512	0.000574009	0.000172191	0.000201183	
30	28	0.001528719	0.111245865	0.000080661	0.000705697	0.03276733	0.001201188	0.002316685	0.028769828	0.002979336	0.000327312	0.000272739	0.000692726	
31	29	0.001192988	0.085370627	0.000113016	0.000263938	0.040779226	0.001370692	0.002701685	0.022280123	0.004862698	0.000472166	0.000393535	0.000403289	
32	30	0.001192988	0.085370627	0.000113016	0.000263938	0.040779226	0.001370692	0.002701685	0.022280123	0.004862698	0.000472166	0.000393535	0.000403289	
33	31	0.001528719	0.111245865	0.000080661	0.000705697	0.03276733	0.001201188	0.002316685	0.028769828	0.002979336	0.000327312	0.000272739	0.000692726	
34	32	0.001192988	0.085370627	0.000113016	0.000263938	0.040779226	0.001370692	0.002701685	0.022280123	0.004862698	0.000472166	0.000393535	0.000403289	
35	33	0.000882973	0.079314358	0.000101647	0.0001742	0.042407305	0.002287784	0.002081231	0.040701236	0.004199609	0.000396952	0.000163329	0.000085096	
36	34	0.002959855	0.115826263	0.000080587	0.000681801	0.034842002	0.00138728	0.00182002	0.049179384	0.003431733	0.000376105	0.000112902	0.00063314	
37	35	0.001454866	0.071302173	0.000135214	0.000182051	0.045425795	0.001287009	0.005999217	0.02089092	0.003986764	0.00030745	0.000356473	0.000384091	
38	36	0.000809513	0.080338533	0.000103131	0.00024373	0.024093952	0.001903838	0.001783271	0.055524464	0.007587337	0.00061887	0.00018187	0.000203454	
39	37	0.000908869	0.074608274	0.00006991	0.000151241	0.04617562	0.003268661	0.003340656	0.049499741	0.009523942	0.000367962	0.000116222	0.000407002	

Fig : 5.3 Prediction using MLR

The fig 5.2 indicates the regression values obtained using multinominal regression techniques and fig 5.3 show the prediction of crime using MLR

VI. CONCLUSION

The analysis of large crime data set is done by using R Studio and the Multinomial logistic regression algorithm. Apache Spark is used to achieve a high data processing. The crime factors change over time. Because of limited factors full accuracy cannot be achieved. More dependent crime attributes has to be added for getting better results in prediction. Advanced implementations like prediction by image processing can be implemented. Big Data Analytics is a real leap forward and a great opportunity to make huge gains in performance, productivity, income and profitability. The Big Data Era is here, and these are truly revolutionary times for business and technology experts to work together and delivering on the pledge.

REFRENCES

1. Yerpude, P., & Gudur, V. (2017). Predictive modelling of crime dataset using data mining. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, *7*(4).
2. Sathyadevan, S., & Gangadharan, S. (2014, August). Crime analysis and prediction using data mining. In *2014 First International Conference on Networks & Soft Computing (ICNSC2014)* (pp. 406-412). IEEE.
3. Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., & Pentland, A. (2014, November). Once upon a crime: towards crime prediction from demographics and mobile data. In *Proceedings of the 16th international conference on multimodal interaction* (pp. 427-434). ACM.
4. A. Nasridinov, S. Ihm, and Y. Park. 'A Decision Tree Based Classification Model for Crime Prediction'. Information Technology Convergence, 531-538, 2013
5. David Smith., R Tops Data Mining Software Poll, Java Developers Journal, May 31, 2012.
6. Fox, John and Andersen, Robert (January2005). 'Using the R Statistical Computing Environment to Teach Social Statistics Courses'. Department of Sociology, McMaster University. Retrieved 2006-08-03.
7. SF Crime Classification Data-Kaggle. https://www.kaggle.com/sfcrime/data.
8. YangchangZhoa(2012).R Data mining Examples and case studies. http://rdataminig.com.

9. Feng, M., Zheng, J., Han, Y., Ren, J., & Liu, Q. (2018, July). Big Data Analytics and Mining for Crime Data Analysis, Visualization and Prediction. In International Conference on Brain Inspired Cognitive Systems (pp. 605-614). Springer, Cham.
10. Das, P., & Das, A. K. (2019). Application of Classification Techniques for Prediction and Analysis of Crime in India. In Computational Intelligence in Data Mining (pp. 191-201). Springer, Singapore.
11. Antolos, D., Liu, D., Ludu, A., & Vincenzi, D. (2013, July). Burglary crime analysis using logistic regression. In International Conference on Human Interface and the Management of Information (pp. 549-558). Springer, Berlin, Heidelberg.
12. Kang, H. W., & Kang, H. B. (2017). Prediction of crime occurrence from multi-modal data using deep learning. PloS one, 12(4), e0176244

AUTHORS PROFILE

R. Raja Devi, is received the B.E(Computer Science and Engineering) degree from Bharathiar University and also M.E(Computer Science and Engineering) degree from Annamalai University . She has 15 years of experience in teaching field. Her research interest includes web Technology , Data mining,optimization techniques,Machine learning , Service Oriented Architecture and Web Services Composition.Currently she is working as an Assistant Professor(Kongu Engineering College) and pursuing the Ph.D degree in Computer Science and Engineering. She has published four papers in international journals.

E. M. Roopa Devi is received the B.E(Electrical and Electronics) degree from Anna University and also M.E(Computer Science and Engineering) degree from Anna University . She has 9 years of experience in teaching field. Her research interest are in the area of Computer Networks, Network Security and Data mining.Currently she is working as an Assistant Professor(Kongu Engineering College) and pursuing the Ph.D degree in Computer Science and Engineering. She has published four papers in international journals.

S.Vinoth Kumaris, received the B.E(Electronics and Communication) degree from Anna University and also M.E(Computer and Communication) degree from Anna University .He has 6 years of experience in teaching field. His research interest are in the area of Computer Networks and Network Security.Currently ,He is working as an Assistant Professor in Kongu Engineering College.

.