Proceedings of the International Conference on Sustainable Computing and Data Communication Systems (ICSCDS-2023)
DVD Part Number: CFP23AZ5-DVD; ISBN: 978-1-6654-5579-4

CRIME PREDICTION AND ANALYSIS USING MACHINE LEARNING

Sameya Khatun [1]
GRIET, Hyderabad, Telangana
khansameya043@gmail.com

Kavyasree Banoth [2]
GRIET, Hyderabad, Telangana
bkavyasree2018@gmail.com

Akshara Dilli [3]
GRIET, Hyderabad, Telangana
aksharareddydilli10@gmail.com

Soundarya Kakarlapudi [4]
GRIET, Hyderabad, Telangana
kakarlapudisoundarya@gmail.com

Sri Vani Karrola [5]
GRIET, Hyderabad, Telangana
srivaniprogrammer0707@gmail.com

G.Charles Babu [6]
GRIET, Hyderabad, Telangana
charlesbabu26@gmail.com

Abstract *- **One of our society's most important problems is crime. It is the most visible part of our civilization. As a result, one of the most crucial jobs is crime prevention .Machine learning approach can better help in the prediction and analysis of the crime. The subject of machine learning crime prediction in India has been addressed through a number of prediction-based theories. Finding the dynamic character of crimes becomes a difficult challenge. The goal of crime prediction is to lower crime rates and discourage criminal activity. In order to discover the proper predictions of crime by using learning-based techniques, this study provides many machine learning algorithms, such as Naive Bayes, Support Vector Machine, Linear Regression, Decision Tree, Bagging Regression, Stacking Regression, and Random Forest Regression algorithms. Comparing the Naïve Byes algorithm to other machine learning models such as SVM, bagging, Linear Regression, Decision tree, stacking, and Random Forest, it is used to create configurations that are specific to a certain domain. On the test data, the suggested technique had a classification accuracy of 99.9%. It is discovered that the model has a stronger predictive impact than the earlier one. When compared to baseline studies that just looked at crime data sets based on violence, the model is found to have greater predictive power. The outcomes demonstrated that criminological theories are compatible with any actual evidence on crime. The suggested method was discovered to be helpful for making potential crime predictions.***

Keywords: Crime prediction, Support Vector Machine, Linear Regression, Decision Tree

1. INTRODUCTION

Numerous criminologists and researchers have recently used a variety of modeling and statistical tools to conduct extensive study and make numerous predictions about how to reduce crime. Due to the fact that crime rates are still rising, it may be necessary to do some significant research that will inform decision-makers and the relevant department about the difficulties and problems related to crime prediction and control methods. If managed manually, a human's skill set cannot keep track of criminal histories. Therefore, it is necessary to identify in a creative approach that will aid in the analysis of material related to crime. This research makes an argument for its novelty using empirical machine learning analysis and the supplementary contributions listed in this section.

2. RELATED WORK

2.1 Exploring Local Crime Patterns with Geographically Weighted Regression

AUTHORS: M. Cahill and G. Mulligan
ABSTRACT: The current study investigates a spatial distribution of violent crime and associated factors in Portland, Oregon using a structural model. The results from a global ordinary least squares model, which employs common structural measurements acquired from an opportunity frame work and is believed to be applicable to all sites within the study area, are Presented in the report. Then, geographically weighted regression (GWR), a substitute for such traditional approaches of modeling crime, is introduced. The GWR approach is used to estimate a local model and generates a set of map able parameter estimates as well as spatially varying values of significance. It is discovered a number of structural factors have associations with crime that differ greatly by place. According to the results, a mixed model that includes both fixed and spatially variable factors may produce the best realistic model of criminality. The current study demonstrates show GWR can be used to look into local factors that influence crime rates and the mis specification of an international model of urban violence.

2.2 Using criminological theory and GIS Techniques to forecast crime using risk terrain modeling

AUTHORS: J.M.Caplan, L.W.Kennedy, and J.Miller.
ABSTRACT: The two main goals drive research that is presented here. The first is to use risk terrain modeling(RTM) to foretell shooting-related crime. The risk terrain maps that were created using RTM assess the risks of up coming shootings as they are distributed over a geography using a variety of contextual data pertinent to the opportunity structure of shootings. The second goal was to evaluate the risk terrain maps' capacity for forecasting over two six month periods and contrast it with that of retroactive hot spot maps. The results show that risk terrains are significantly more accurate at predicting future shootings across a

variety of cut points than retroactive hotspot mapping. Additionally, risk landscape maps generate data that can be quickly and effectively operationalized by police administrators, such as for allocating police patrols to clustered high-risk regions.

2.3 A more effective way to classify algorithms for predicting crime.

AUTHORS: Babakura, M. D. Sulaiman, and M. A. Yusuf
ABSTRACT: Lawen forcement agencies now have access to detailed information. Due to the increasing accessibility of information technologies, regarding a number of crimes. Finding a model (or function) that represents and distinguishes data classes or concepts is a critical part of the classification process. The intention is to forecast crime labels using the model. In this work, classification is used to analyze a crime data set and forecast the “crime category" for several states in the United States of America (USA).The socioeconomic data from the US Census of 1990 were used to build the real-world crime data set that was used in this study.

2.4 Spatial-temporal pattern analysis and prediction for urban crime.

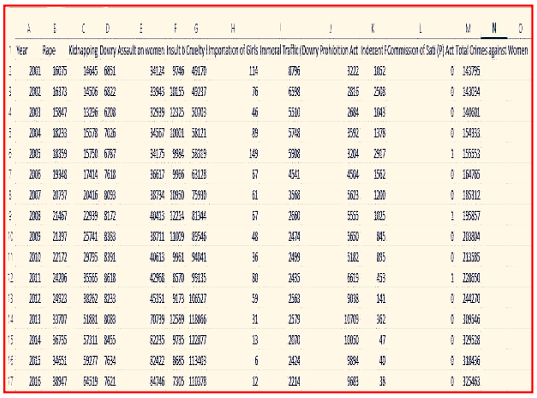
AUTHORS: Z. Li, T. Zhang, Z. Yuan, Z. Wu, and Z.Du

ABSTRACT: This study quantifies the crime data from the original case file in order to investigate the fundamental traits of urban crime in China. The essential characteristic and its rule are validated by contrasting the observed with the projected outcome of the crime circumstance. The second step is an examination of the case's internal features based on the quantity of cases, the timing, and the place of the occurrence. Thirdly, a crime prediction model based on the ARIMA is provided to forecast the crime scenario over time. The results reveal that the projected outcomes exhibit the same criminal characteristics and are consistent with the genuine values.

3. METHODOLOGY

In order to validate the predicted results, these models (for instance Naïve Bayes, SVM, Linear Regression, and Decision tree, Bagging Regression, Stacking Regression and Random Forest Regression) are developed by the suitable model parameter values and utilized to CAW dataset in this article. The following stages are used to build the suggested methodology:

Step 1: Load dataset
Load the CAW dataset, which has 13 columns and 18 rows, each column states types of crimes.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Year	Rape	Kidnapping	Dowry	Assault on women	Insult to	Cruelty by	Importation of Girls	Immoral Traffic	Dowry Prohibition Act	Indecent R	Commission of Sati (P	Act	Total Crimes against Women	
2	2001	16075	16665	6851		34124	57496	49170	124	8796	3222	1652	0	143795	
3	2002	16373	14506	6822		33945	10155	49237	76	6198	2815	2568	0	143254	
4	2003	15847	13256	6208		32939	12325	50703	46	5510	2684	1043	0	140681	
5	2004	18233	15578	7026		34567	10001	58221	89	5748	3592	1378	0	154333	
6	2005	18359	15750	6787		34175	9984	56319	149	5398	3204	2917	1	155253	
7	2006	19348	17424	7618		36617	9958	63128	67	4541	4564	1561	0	164705	
8	2007	20737	20416	8093		38734	10959	75930	61	3568	5623	1200	0	185312	
9	2008	21467	22599	8172		40403	12214	81344	97	3590	5555	1025	1	195857	
10	2009	21397	25741	8383		38711	13009	89546	48	3474	5650	845	0	206804	
11	2010	22172	26755	8391		40615	9961	94041	56	3499	5102	895	0	211385	
12	2011	24206	35565	8618		42968	8570	95535	80	3455	6025	453	1	226591	
13	2012	24923	38262	8233		45351	9175	106527	59	2563	9018	141	0	244270	
14	2013	33707	51681	8083		70739	12589	118866	31	2579	11705	162	0	309546	
15	2014	36735	57311	8455		82235	9735	122877	13	2030	12650	47	0	329528	
16	2015	34651	59377	7634		82422	8665	113403	6	1424	9894	43	0	316496	
17	2016	38947	64519	7621		84746	7205	111078	12	2214	9683	16	0	325481	

Step 2: Data Preprocessing
Data cleansing and preparation for a machine learning model need data pre- processing, which also increases the model's precision and efficacy.
The following actions are involved:
To obtain the dataset Library imports Importing datasets Finding Missing Data Feature scaling

Step 3: Split data
Splitting data into train and test data

Step 4: Model Generation
We have used following 8 models
• Linear Regression
• SVM with Sequential Minimal Optimization (SMO)
• Naïve Bayes Regression with Linear Model
• Decision Tree
• Support Vector Machine
• Bagging Regression
• Stacking Regression
• Random Forest Regress

Step 5: Building the models
Similar to the Linear Regression model (Shown below in the picture) build other models like Naïve Bayes, SVM and Decision tree, Bagging Regression, Stacking Regression and Random Forest Regression.

Linear Regression

```
In [11]: from sklearn.linear_model import LinearRegression
         regressor = LinearRegression()
         regressor.fit(X_train,y_train)
         print ("Your Test Set is:- \n",X_test)
         #X_tst = np.array([2015, 0, 0,  0, 0, 0, 0, 0, 0, 0,  0, 0])
         accuracy = regressor.score(X_test,y_test)
         print ("\nYour Prediction has the accuracy of-",accuracy*100,"%")

         y_prediction = regressor.predict(X_test)
         print ("\nPredicted Total Crime for above given years is:-\n",y_prediction)

         Your Test Set is:-
          [[2.0090e+03 2.1397e+04 2.5741e+04 8.3830e+03 3.8711e+04 1.1009e+04
           8.9546e+04 4.8000e+01 2.4740e+03 5.6500e+03 8.4500e+02 0.0000e+00]
          [2.0010e+03 1.6075e+04 1.4645e+04 6.8510e+03 3.4124e+04 9.7460e+03
           4.9170e+04 1.1400e+02 8.7960e+03 3.2220e+03 1.0520e+03 0.0000e+00]
          [2.0070e+03 2.0737e+04 2.0416e+04 8.0930e+03 3.8734e+04 1.0950e+04
           7.5930e+04 6.1000e+01 3.5680e+03 5.6230e+03 1.2000e+03 0.0000e+00]
          [2.0100e+03 2.2172e+04 2.9795e+04 8.3910e+03 4.0613e+04 9.9610e+03
           9.4041e+04 3.6000e+01 2.4990e+03 5.1820e+03 8.9500e+02 0.0000e+00]]

         Your Prediction has the accuracy of- 99.99995782840291 %

         Predicted Total Crime for above given years is:-
          [203782.71602995 143814.1626291  185304.92756927 213566.69507595]
```

Step 6: Accuracy Comparisons

In contrast to other machine learning models like SVM, bagging, Linear Regression, Decision tree, stacking, and Random Forest, the Naïve Byes algorithm is used to create domain specific configurations and gives us99.9%accuracy.

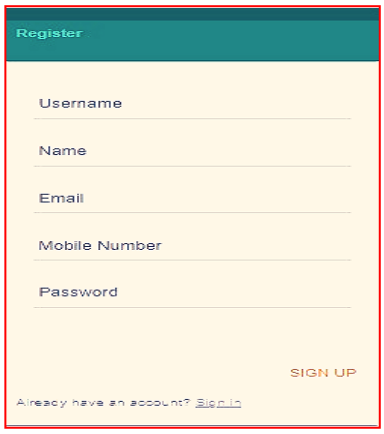
Step 7: Load Model

We use the Sci kit-learn (Sklearn) model since it is the most powerful and reliable Python machine learning package available at this time. Classification, regression, clustering, and dimensionality reduction are just some of the powerful techniques for statistical modeling and machine learning that are made available via a Python-consistent interface. NumPy, SciPy, and MatPow are the foundations upon which this library was built.

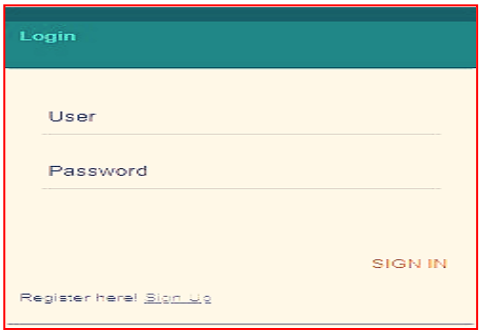
Naive Bayes Regression with Linear Model

```
from sklearn.linear_model import BayesianRidge
clf = BayesianRidge(compute_score=True)
clf.fit(X_train,y_train)
```

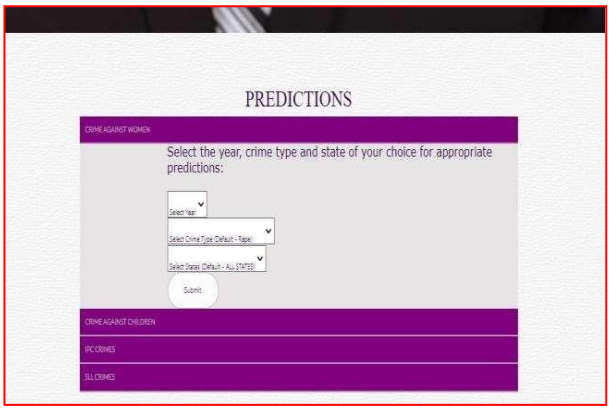
Step 8: User Register



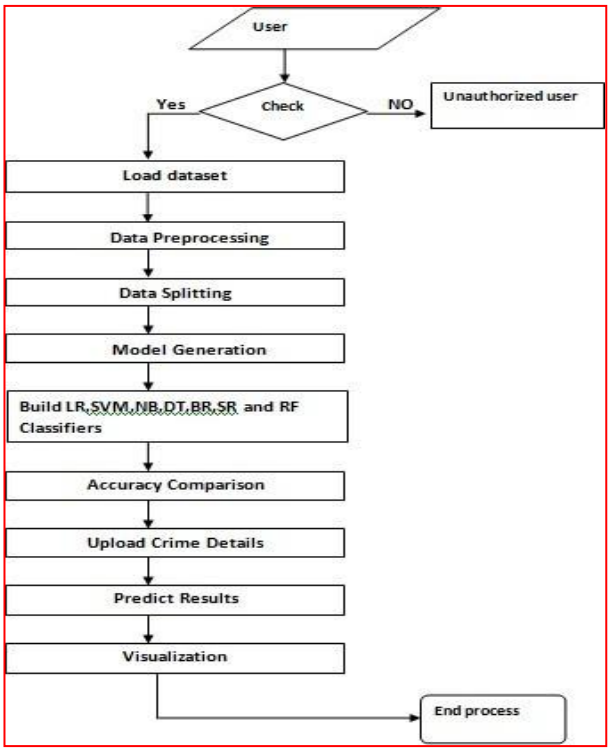
Step 9: User Login



Step 10: Upload Crime Details



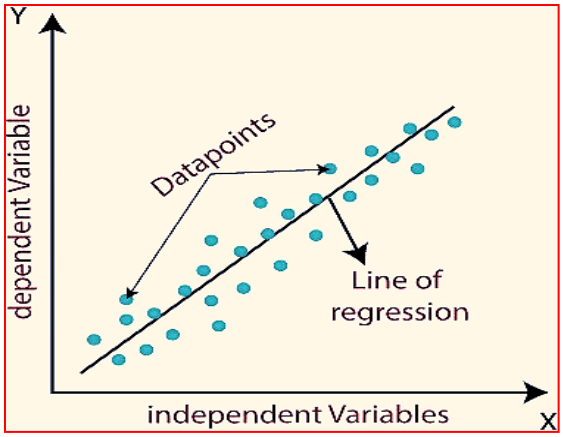
Block Diagram



4. MODULE DESCRIPTION

4.1 Linear Regression

Linear regression provides forecasts for continuous /real /numerical variables such as sales, salary, age, and product price. In a linear model, the connection between independent(x) and dependent (y) variables is shown using the linear regression procedure (y). A slanted straight line, representing the connection between the variables, is the output of the linear regression model. Have a look at the example below:



4.2 Naïve Bayes Regression with Linear Model

On the real m of classification techniques, the Naïve Bayes algorithm is a supervised learning strategy grounded in the Bayes theorem. Its primary use is in text classification problems when a large training set is available. Among the many classification algorithms available, the Naive Bayes Classifier stands out as one of the simplest and most accurate. It helps in developing quick models for machine learning that can provide reliable predictions. This model makes predictions based on the probability of events occurring. Among the many applications for Naïve Bayes algorithms are spam filtering, sentiment analysis and article classification.

4.3 Decision Tree

Decision trees, a kind of supervised learning, may be used to both classification and regression issues, but they are often employed to address the former. It's a classifier organized like a tree, with the nodes representing characteristics in the data set, the branches representing rules for making that classification, and the leaf nodes representing the actual classification. Decision Node and Leaf Node are the two types of tree nodes.

A Decision

Node is a choice making tool and so contains numerous branches, where as a Leaf Node is the end result of a decision and thus has no additional branches. To make a decision, one uses a node called a" Decision, "which contains multiple "branches," and "Leaf" nodes, which are the consequence of the decision and have no "branches." It is a graphical depiction of all feasible out comes to a problem/decision depending on specific criteria.

4.4 Bagging Regression

Before aggregating each forecast (either by voting or by averaging) to get a final prediction, an ensemble meta estimate or known as a bagging regression or fits base regression or distinct random subsets of the original data set.

4.5 Random Forest Regression

Classifier known as Random Forest employs many decision trees applied to various subsets of a dataset and then averages the results to increase the predicted accuracy of the dataset as a whole. There is less of a chance of over fitting and better accuracy if there are more trees in the forest.

4.6 Support Vector Machine

One of the most common supervised learning techniques used for classification and regression is the Support Vector Machine (SVM). In order to efficiently classify new data points in the future, the SVM approach looks for the best line or decision boundary that can split n-dimensional space into classes. A hyper plane represents the ideal boundary for making a choice. Selective feature selection is used to choose which extreme vectors and points will be utilized to construct the hyper plane. The SVM method employs support vectors as a means of symbolizing such severe conditions. In the diagram below, you can see how the decision boundary or hyper plane is utilized to determine two distinct categories.

5. DATASET DESCRIPTION

The CAW data set, which has 13 columns and

18 rows each column, denotes types of crimes.

	Year	Rape	Kidnapping and Abduction of Women & Girls	Dowry Deaths	Assault on women with intent to outrage her modesty	Insult to the modesty of Women	Cruelty by Husband or his relatives	Importation of Girls from Foreign Country	Immoral Traffic (P) Act	Dowry Prohibition Act	Indecent Representation of Women (P) Act	Commission of Sati (P) Act	Total Crimes against Women
0	2001	16075.0	14645.0	6851.0	34124.0	9746.0	49170.0	114	8796.0	3222.0	1052.0	0.0	143795.0
1	2002	16373.0	14506.0	6822.0	33943.0	10155.0	45237.0	76	6568.0	2816.0	2508.0	0.0	143034.0
2	2003	15847.0	13296.0	6208.0	32939.0	12325.0	50703.0	46	5510.0	2894.0	1043.0	0.0	146891.0
3	2004	18233.0	15578.0	7026.0	34567.0	10001.0	58121.0	89	5748.0	3592.0	1378.0	0.0	154333.0
4	2005	18359.0	15750.0	6787.0	34175.0	9984.0	58319.0	149	5908.0	3204.0	2917.0	1.0	155553.0

The following are the crimes in our data set.

1. Rape
2. Abduction and Kidnapping of Women & Girls
3. Deaths by dowry
4. Assaulting women with the intention of offending their modesty
5. Insulting women's modesty
6. Husband's or his relatives cruelty
7. Indecent Representation of Women(P)Act
8. Dowry Prohibition Act
9. Immoral Traffic(P) Act
10. Commission of Sati(P) Act Total Crimes against women
11. Commission of Sati(P) Act
12. Total Crimes against Women

6. RESULTS AND DISCUSSION

Crime Predictions:

Select the respective year ranging from (2017-2020)

Check the Appropriate Box for the Offense You Committed (for instance like Rape, Kidnapping and Abduction, Dowry Deaths, Assault on women, Insult to the modesty of Women, Cruelty by Husband or his relatives and Total Crimes against Women) Finally, click and submit to view forecasts for the states you chose (West Bengal, Uttar Pradesh, Tripura, Telangana, Tamil Nadu, Sikkim, Rajasthan, Punjab, Odisha, Nagaland, Mizoram, Meghalaya, Manipur, Maharashtra, Madhya Pradesh, Kerala, Karnataka, Jharkhand, Jammu and Kashmir, Himachal Pradesh, Haryana, Gujarat, Goa, Chhattisgarh, Bihar.

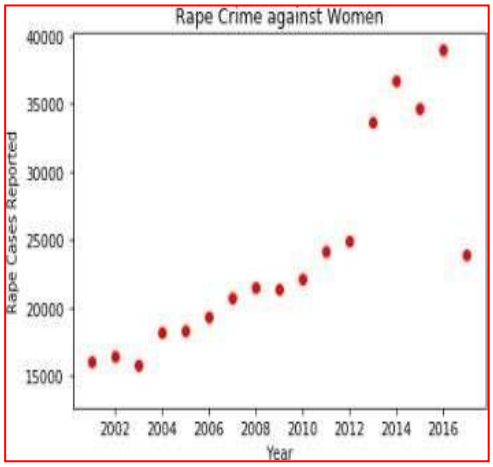


Figure 6.1: Rape crime against women

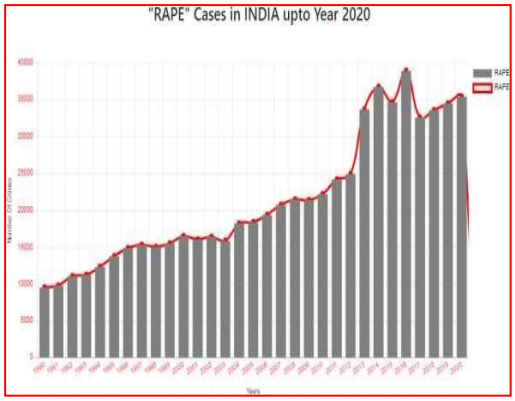


Figure 6.2: Rape

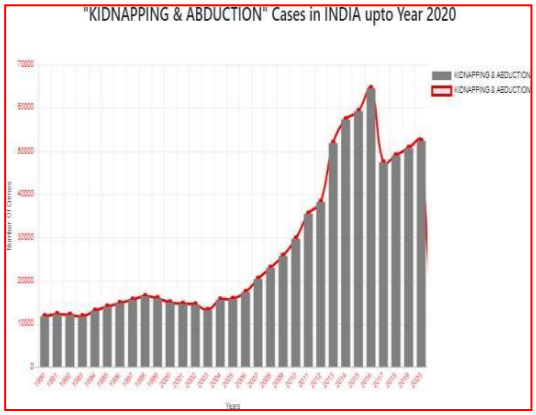


Figure 6.3: Abduction and Kidnapping of Women & Girls

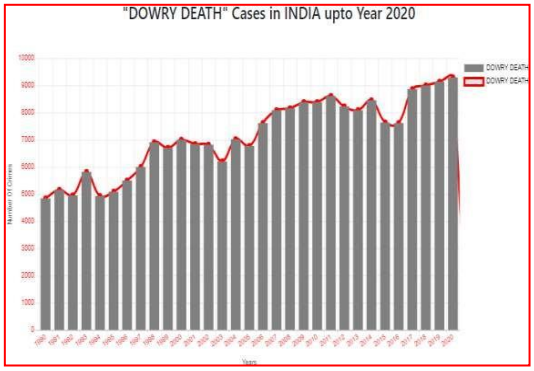


Figure 6.4: Dowry Deaths

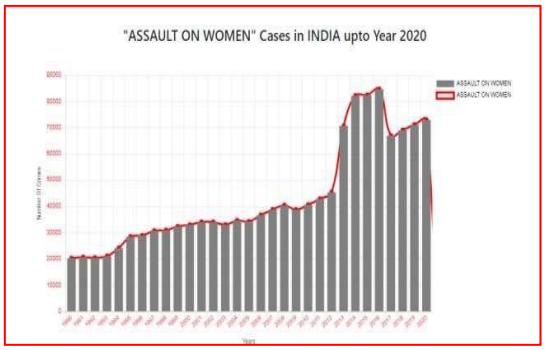


Figure 6.5: Assault on women

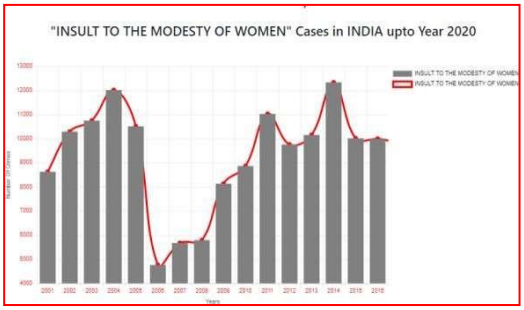


Figure 6.6: Insulting women's modesty

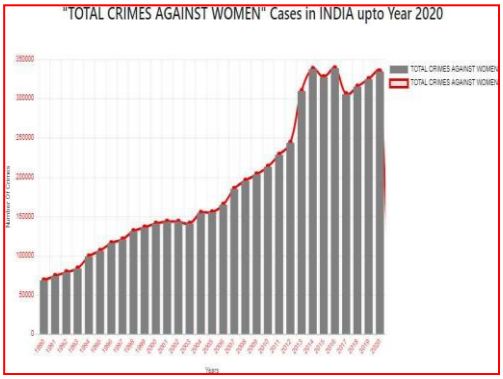


Figure 6.7: Total Crimes against Women

7. CONCLUSION

Machine learning models, such as Naive Bayes, Linear Regression, Decision Tree, SVM, Bagging Regression, Stacking Regression and Random Forest Regression algorithms, were employed in the current study to determine the most suitable crime predictions. In contrast to other machine learning models like SVM, bagging, Linear Regression, Decision tree, stacking, and Random Forest, the Naïve Byes algorithm is used to create domain-specific configurations. The conclusion suggests that a performer model does not typically function properly. The outcomes of the experiments show how effective the suggested paradigm is. During the training phase, the model's core working time grows at a rate of 99.5%. Predictability in measuring is a result of the effectiveness of the technique used to determine the appropriate course of action in criminal situations. On the testing data, the suggested technique had a classification accuracy of 99.9%.

REFERENCES

[1] Leni Marlina Muslim Muslim, A Siahaan and P Utama ,"Data Mining Classification Comparison ", IJOETT in Computer Science,2016.

[2] BHssina, MERBOUHA Abdelkarim, Hanane Ezzikouri and Mohammed. Erritali, Special Issue on Advances in Vehicular AdHoc Networking and Applications, 2014.

[3] C Zhang and S Zhang, Association Rule MM and Algorithms, S-VBerlinHeidelberg, 2002.

[4] D T. Larose, Data MM and models, Hoboken, New Jersey :Published by J Wiley & Sons, Inc2006

[5] X. Zhang, L. Liu, L. Xiao and J. Ji, "Comparison of Machine Learning Algorithms for Predicting Crime Hotspots," in IEEE Access, vol. 8, pp. 181302-181310, 202

[6] W. Safat, S. Asghar and S. A. Gillani, "Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques," in IEEE Access, vol. 9, pp. 70080-70094, 2021

.[7] Choo, Kim-Kwang Raymond. "Prediction of Crime Occurrence from Multi-Modal Data Using Deep Learning." PloS one. 12, no. 4 (2017).

[8] S. Kim, P. Joshi, P. S. Kalsi and P. Taheri, "Crime Analysis Through Machine Learning," 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 2018, pp. 415-420.

[9] 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).

[10] S. S. Kshatri, D. Singh, B. Narain, S. Bhatia, M. T. Quasim and G. R. Sinha, "An Empirical Analysis of Machine Learning Algorithms for Crime

Prediction Using Stacked Generalization: An Ensemble Approach," in IEEE Access, vol. 9, pp. 67488-67500, 2021, doi: 10.1109/ACCESS.2021.3075140.

[11] Z. M. Wawrzyniak et al., "Data-driven models in machine learning for crime prediction," 2018 26th International Conference on Systems Engineering (ICSEng), Sydney, NSW, Australia, 2018, pp. 1-8, doi: 10.1109/ICSENG.2018.8638230.

[12] H. Adel, M. Salheen and R. Mahmoud, "Crime in relation to urban design. Case study: the greater Cairo region", Ain Shams Eng. J., vol. 7, no. 3, pp. 925-938, 2016.

[13] Y. L. Lin, L. C. Yu and T. Y. Chen, "Using machine learning to assist crime prevention" in IEEE 6 th Intl. Congr. on Advanced Appl. Inform. (IIAI-AAI) , Hamamatsu, Japan, Jul. 2017.