Proceedings of the International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS 2023)
IEEE Xplore Part Number: CFP22DN7-ART; ISBN: 979-8-3503-0085-7

Crime Pattern Analysis and Prediction using Regression Models

Sheshang Degadwala
Associate Professor & Head of Department, Dept. of Comp. Engineering, Sigma University, Gujarat, India
sheshang13@gmail.com

Dhairya Vyas
Research Scholar, The Maharaja Sayajirao University of Baroda, Vadodara, Gujarat, India
dhairya.vyas-cse@msubaroda.ac.in

Akash Rajeshkumar Raval
Head of Department & Assistant Professor, Department of Chemical Engineering
Sigma University,
Vadodara, Gujarat, India
akashravalchemical@gmail.com

Mukesh Soni
Department of CSE, University Centre for Research & Development
Chandigarh University, Mohali, Punjab-140413, India
mukesh.research24@gmail.com

2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS) | 979-8-3503-0085-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSSAS57918.2023.10331747

***Abstract*— This research intends to address the pressing concern of women's safety by developing predictive models for identifying regional women's crimes, utilizing a range of regression models including Support Vector Regression (SVR), Logistic Regression, Decision Tree (DT), K-Nearest Neighbor (KNN), Random Forest (RF), Linear Regression (LR), and Polynomial Regression. With a comprehensive dataset encompassing historical crime statistics, demographic attributes, and socioeconomic factors, the study aims to construct accurate predictive models by considering crucial elements such as population density, education levels, employment rates, urbanization metrics, and police coverage. Through rigorous cross-validation and comparison using metrics like mean squared error, R-squared, and accuracy, the research seeks to identify the most effective model for predicting crimes against women in different Indian regions. The outcomes of this study have far-reaching implications, offering data-driven insights for authorities, policymakers, and advocacy groups to enhance women's safety by strategically allocating resources, optimizing methods, and implementing proactive measures, thus contributing to an overall safer environment and improved well-being for women throughout the nation.**

Keywords*— Raga Identification, Indian Classical Music, FCN-Based Model, Deep Learning, Convolutional Networks, Performance Evaluation.

I. Introduction

Crime pattern analysis and prediction is an area where machine learning methods have made significant strides in recent years [1-3]. The purpose of this research is to employ machine learning techniques, primarily regression models, to predict female victimization in different parts of India.

The major goal of this study is to construct accurate and robust regression models for predicting crimes against women at the regional level by drawing on historical crime data, demographic information, and socio-economic indicators [4-8]. By delving into these connections, we may start to see trends, potential danger zones, and areas where women are more at risk. Law enforcement, lawmakers, and groups concerned about women's safety may use this information to more precisely allocate resources, create focused plans, and take preventative action.

The use of regression models to determine associations between predictor factors and the dependent variable is crucial to the success of this research. Linear Regression, Decision Tree Regression, Random Forest Regression, K-Nearest Neighbor Regression, Support Vector Regression, Logistic Regression, and Polynomial Regression are some of the selected regression models. Varying models have different strengths and methods for capturing different facets of the intricate connections between crime rates and factors including population density, literacy, employment, urbanization, and the presence of law enforcement [6-12]. The best model for predicting regional female victimization in India will be determined via this rigorous review procedure.

The results of this study have important implications for law enforcement, lawmakers, and groups promoting women's rights, all of which seek to ensure the protection of women. Accurate regional crime forecasts allow for the creation of data-driven strategies, the implementation of carefully targeted interventions, and the most effective use of available resources. A culture in which women may prosper, feel protected, and contribute to the advancement of the country can be created via the active addressing and prevention of crimes against women.

The organization of the paper is section 1 introduction, section 2 review of literature or related works, section 3 is proposed system along with flow diagram, section 4 is results and analysis and at end conclusion of the paper.

II. Related Works

This literature review summarizes studies that have used data mining and machine learning methods to analyze and predict crime, with a special emphasis on violence against women. Studies were chosen for their contributions to our knowledge of crime trends, prediction models, and the causes and consequences of violence against women.

A.A. Biswas et al. [1] conducted a study on forecasting crime trends and patterns in Bangladesh using machine learning models. Their research highlighted the significance of machine learning techniques in predicting crime patterns and facilitating effective law enforcement strategies.

S. Lavanyaa et al. [2] explored crime against women in Tamil Nadu, India, using data mining techniques. Their study emphasized the analysis and prediction of crime against women, providing insights for improving the efficiency of the Tamil Nadu Police Department.

B. Sivanagaleela and S. Rajesh [3] focused on crime analysis and prediction using the Fuzzy C-Means algorithm. Their research aimed to identify crime patterns and predict future crime occurrences, demonstrating the effectiveness of fuzzy clustering techniques in crime analysis.

Khushabu A. Bokde et al. [4] proposed a crime detection technique using data mining and K-means clustering. Their study highlighted the application of clustering algorithms in identifying crime patterns and assisting law enforcement agencies in crime detection and prevention.

Priyanka Das et al. [5] conducted a behavioral analysis of crime against women using a graph-based clustering approach. Their research demonstrated the usefulness of graph-based clustering in identifying behavioral patterns and detecting anomalous activities related to crimes against women.

G. Vicente et al. [6] focused on the spatial patterns and temporal trends of dowry deaths, a specific type of crime against women, in the districts of Uttar Pradesh, India. Their study revealed insights into the spatial distribution and temporal variations of dowry deaths, facilitating targeted interventions and policy planning.

P. Tamilarasi et al. [7] conducted research on the diagnosis of crime rates against women using machine learning algorithms. Their study employed k-fold cross-validation and various machine learning techniques to accurately diagnose crime rates against women.

S. Lavanyaa et al. [8] explores the application of data mining techniques for analyzing and predicting crime against women in Tamil Nadu. This section presents a review of related work in the field of crime analysis and prediction, emphasizing the use of data mining techniques.

Shiju Sathyadevan et al. [9] explored crime analysis and prediction using data mining techniques. Their research highlighted the application of data mining techniques in crime analysis and prediction, enabling proactive measures for crime prevention.

Priya Gandhi and Shayog et al. Sharma [10] employed predictive modeling techniques to address the problem of crime against women. Their research focused on developing predictive models to identify potential crime hotspots and improve women safety.

Hyeon-Woo Kang et al. [11] proposed a prediction model for crime occurrence using multimodal data and deep learning techniques. Their study emphasized the integration of diverse data sources and the application of deep learning algorithms for crime prediction.

Ritvik Chauhan et al. [12] conducted a spatial-temporal analysis of crime against women in India. Their research utilized geospatial techniques to map and analyze the spatial and temporal patterns of crimes against women, providing insights for targeted interventions.

Bhajneet Kaur et al. [13] explored the factors affecting crime against women using regression and K-means clustering techniques. Their study identified key factors influencing crimes against women and developed regression and clustering models for predicting crime occurrences.

Kaur, Ahuja, et al. [14] present a study on crime against women, focusing on analysis and prediction using data mining techniques. The authors investigate patterns and trends related to crimes against women by employing data mining methods. They analyze various factors contributing to these crimes and develop prediction models. This research contributes to understanding the dynamics of crimes against women and highlights the potential of data mining for predicting such incidents.

Patel et al. [15] contribute to the field of crime against women analysis and prediction in India by utilizing supervised regression techniques. The authors focus on understanding and forecasting crimes against women through the application of supervised regression methods. They develop models that consider relevant features to predict the occurrence of such incidents. This work underscores the significance of using regression-based approaches for addressing the issue of crimes against women in India.

Gupta et al. [16] describe an analysis of criminal spatial events in India using exploratory data analysis and regression techniques. The authors delve into the spatial distribution of criminal incidents and employ exploratory data analysis to uncover patterns and trends within the data. They also utilize regression methods to model the relationships between various factors and criminal events. The study contributes to the understanding of the spatial dynamics of crime in India and highlights the potential of data analysis and regression in addressing criminal issues.

Alves, Ribeiro, and Rodrigues [17] focus on crime prediction using urban metrics and statistical learning methods. The authors explore the use of urban metrics, which are measures related to the urban environment, to predict crime occurrences. They combine these metrics with statistical learning techniques to develop models for crime prediction. By doing so, they offer insights into the potential of incorporating urban features and advanced statistical methods for anticipating criminal activities.

Vivek et al. [18] present research on spatio-temporal crime analysis and forecasting using Twitter data and machine learning algorithms. The authors leverage Twitter data to analyze the spatial and temporal patterns of crime incidents. They employ various machine learning algorithms to develop models that can forecast crime trends. This work showcases the integration of social media data and machine learning for enhancing crime analysis and prediction capabilities.

Collectively, this research highlights the value of data mining and machine learning strategies for examining and forecasting violence against women. They help with evidence-based policy making, specific interventions, and boosting women's safety by shedding light on crime trends, risk factors, and prediction models.

III. Proposed Methodology

Predicting where crimes against women would occur in India's regions using regression models is possible with some careful planning. These procedures guarantee accurate data processing, model selection, assessment, and interpretation, which in turn yields useful insights for enhancing measures to reduce criminal activity.

A. Data Collection:

Collect historical crime statistics on crimes against women in India on a regional level. Data on population density, literacy, employment, urbanization, and law enforcement presence are all important indicators of potential crime rates and should be collected. It is compiled from various sources, including the National Crime Records Bureau (NCRB) of India, which is responsible for collecting and analyzing crime data in the country. The dataset contains detailed information on various aspects of crime, including the type of offense, the location where the crime occurred, the gender and age group of the victims and offenders, and other relevant attributes. The data is organized in a structured format, making it suitable for analysis and exploration. Here are some key features and attributes present in the dataset:

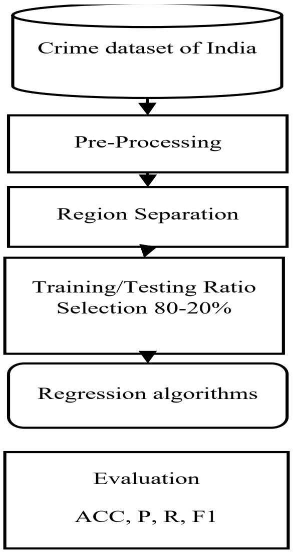


Fig. 1. Proposed System Flow

Crime Type: The dataset covers a wide range of criminal activities, including murder, kidnapping, robbery, burglary, theft, assault, sexual offenses, cybercrime, and more. Each crime is categorized under specific types.

Location: The dataset provides information about the state and city where the crime took place. This allows for analyzing crime patterns and comparing crime rates across different regions in India.

Victims and Offenders: The dataset includes details about the victims and offenders involved in each crime. It provides information about the gender and age group of the victims and offenders, enabling demographic analysis.

Crime Rates: The dataset provides statistical information such as the number of reported crimes, crime rates per lakh of population, and crime rates per 100,000 populations for each crime type and location. These figures can be used to assess the prevalence of different crimes and their distribution across India.

Time Period: The dataset specifically covers crime data for the year 2019. It allows researchers and analysts to study crime trends and patterns for that particular year.

B. Data Preprocessing:

Using pandas removes any inaccurate or unnecessary information from the data set. To guarantee that all variables are measured on the same scale, normalize, or standardize the data. Separate the data into a training set and a test set so that you can compare the results. Using

C. Region Separation

Using Pandas find the areas in India where crime statistics and forecasts are most wanted. This may be done based on political subdivisions like states, counties, or municipalities.

D. Train/Test Split

Typically, eighty percent of the data is used for training and twenty percent is used for testing. The precise division, however, might change depending on the amount and complexity of the collection.

E. Regression algorithms:

Select appropriate regression models for foretelling regional female victimization in India. In this setting, several different models are used, including Linear Regression, Decision Tree, Random Forest, K-Nearest Neighbor, Support Vector Regression, Logistic Regression, and Polynomial Regression.

Pseudo Code

```
1.Start
2. Read Number of Data (n)
For i=1 to n:
Read Xi and Yi
Next i
3. Initialize:
$sumX = 0, sumX2 = 0, sumY = 0, sumXY = 0$
4. Calculate Required Sum:
For i=1 to n:
$sumX = sumX + Xi$
$sumX2 = sumX2 + Xi * Xi$
$sumY = sumY + Yi$
$sumXY = sumXY + Xi * Yi$
Next i
5. Calculate Required Constant a and b of $y = a + bx$:
$b = (n * sumXY - sumX * sumY) / (n * sumX2 - sumX * sumX)$
$a = (sumY - b * sumX) / n$
6. Display value of a and b
7. Stop
```

F. Model Evaluation:

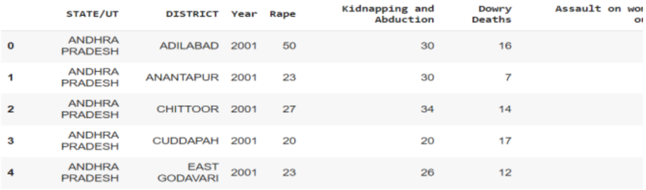
Evaluate the efficacy of each regression model by calculating its F1 score, accuracy, precision, recall, and mean squared error (MSE). To ensure the accuracy of the models and prevent overfitting, cross-validation should be used.

Examine the outcomes and explanations offered by the selected regression model. Study the factors that significantly affect crime against women. Learn how to make better policies and interventions by analyzing the connections between poverty and criminal activity.

IV. Result and Analysis

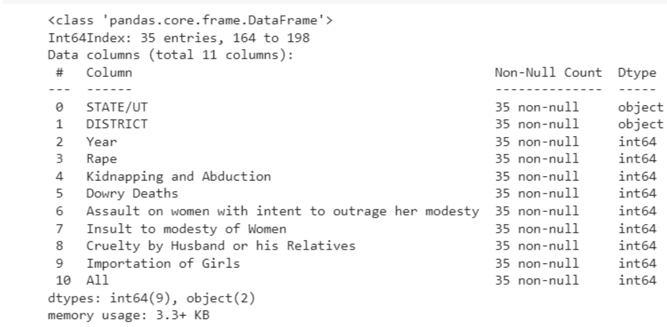
Metrics like R2Score, EVS, MSE, MAE, and RMSE are often employed in predictive modelling and data analysis. When attempting to predict sea level rise using climate data, they are crucial in assessing the performance of models.

Now showing the results different models from reading fig.2 and 3 data reading which has columns 10 columns.



	STATE/UT	DISTRICT	Year	Rape	Kidnapping and Abduction	Dowry Deaths	Assault on wo o
0	ANDHRA PRADESH	ADILABAD	2001	50	30	16	
1	ANDHRA PRADESH	ANANTAPUR	2001	23	30	7	
2	ANDHRA PRADESH	CHITTOOR	2001	27	34	14	
3	ANDHRA PRADESH	CUDDAPAH	2001	20	20	17	
4	ANDHRA PRADESH	EAST GODAVARI	2001	23	26	12	

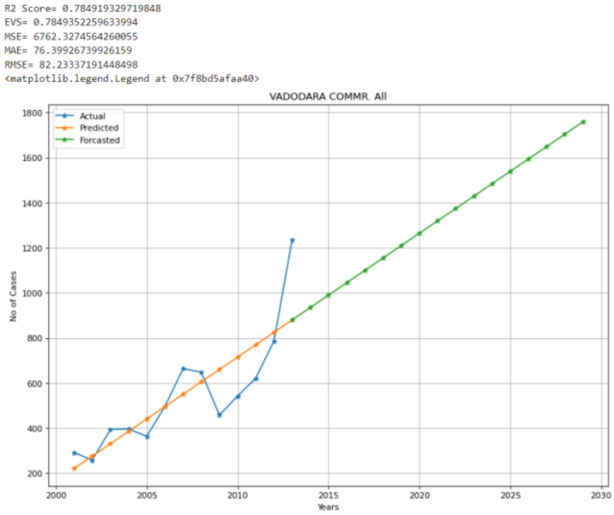
Fig. 2. Reding Dataset



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 35 entries, 164 to 198
Data columns (total 11 columns):
 #   Column                                       Non-Null Count  Dtype
---  ------                                       --------------  -----
 0   STATE/UT                                     35 non-null     object
 1   DISTRICT                                     35 non-null     object
 2   Year                                         35 non-null     int64
 3   Rape                                         35 non-null     int64
 4   Kidnapping and Abduction                     35 non-null     int64
 5   Dowry Deaths                                 35 non-null     int64
 6   Assault on women with intent to outrage her modesty  35 non-null     int64
 7   Insult to modesty of Women                   35 non-null     int64
 8   Cruelty by Husband or his Relatives          35 non-null     int64
 9   Importation of Girls                         35 non-null     int64
 10  All                                          35 non-null     int64
dtypes: int64(9), object(2)
memory usage: 3.3+ KB
```

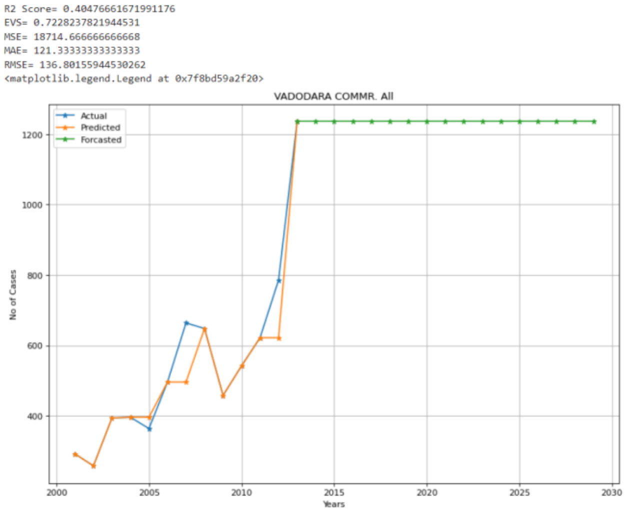
Fig. 3. Extract Gujarat Data

Data separate fig. 4 shows liner regression model forecasting which has r2 score of 0.78 and we can see that the blue line is actual data, orange is predicted line and green is forecasting line.



```
R2 Score= 0.784919329719848
EVS= 0.7849352259633994
MSE= 6762.327456426055
MAE= 76.39926739926159
RMSE= 82.23337191448498
<matplotlib.legend.Legend at 0x7f8bd5afaa40>
```

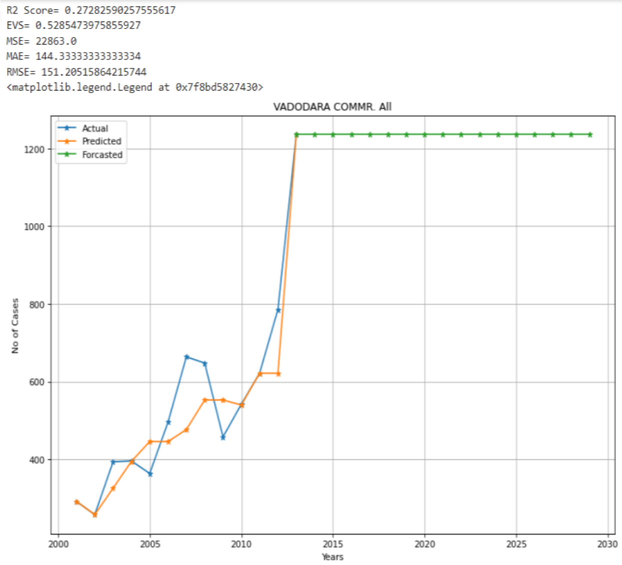
Fig. 4. Linear Regression



```
R2 Score= 0.404766661671991176
EVS= 0.7228237821944531
MSE= 18714.666666666668
MAE= 121.33333333333333
RMSE= 136.80155944530262
<matplotlib.legend.Legend at 0x7f8bd59a2f20>
```

Fig. 5. Decision Tree Model

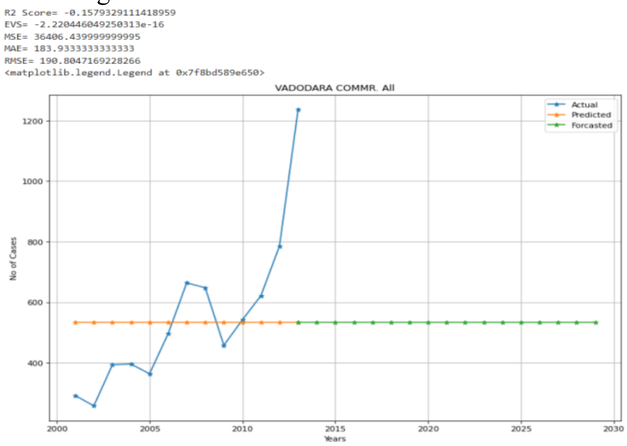
Data separate fig. 5 shows Decision tree regression model forecasting which has r2 score of 0.40 and we can see that the blue line is actual data, orange is predicted line and green is forecasting line.



```
R2 Score= 0.27282590257555617
EVS= 0.5285473975855927
MSE= 22863.0
MAE= 144.33333333333334
RMSE= 151.20515864215744
<matplotlib.legend.Legend at 0x7f8bd5827430>
```

Fig. 6. Random Forest Model

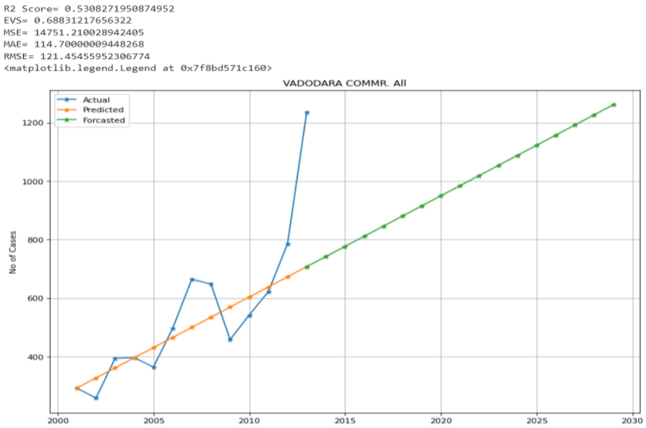
Data separate fig. 6 shows Decision tree regression model forecasting which has r2 score of 0.27 and we can see that the blue line is actual data, orange is predicted line and green is forecasting line.



```
R2 Score= -0.1579329111418959
EVS= -2.220446049250313e-16
MSE= 36406.439999999995
MAE= 183.9333333333333
RMSE= 190.8047169228266
<matplotlib.legend.Legend at 0x7f8bd589e650>
```

Fig. 7. KNN Model

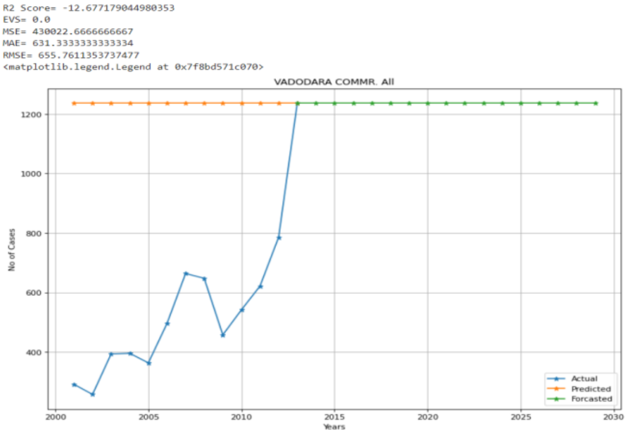
Data separate fig. 7 shows KNN regression model forecasting which has r2 score of 0.15 and we can see that the blue line is actual data, orange is predicted line and green is forecasting line.



```
R2 Score= 0.5308271950874952
EVS= 0.68831217656322
MSE= 14751.210028942405
MAE= 114.70000009448268
RMSE= 121.45455952306774
<matplotlib.legend.Legend at 0x7f8bd571c160>
```

Fig. 8. Support Vector Regression Model

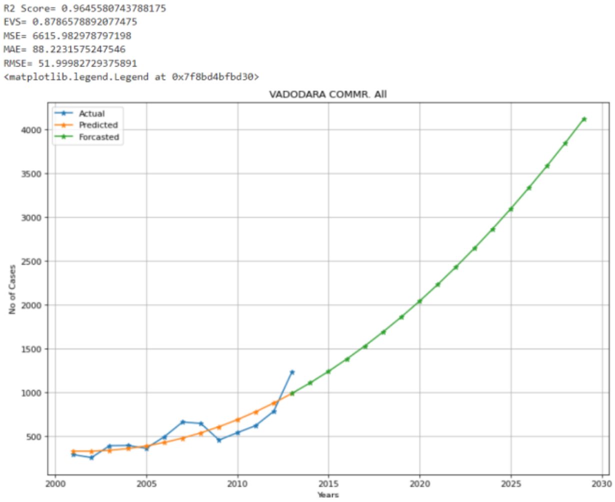
Data separate fig. 8 shows SVM regression model forecasting which has r2 score of 0.53 and we can see that the blue line is actual data, orange is predicted line and green is forecasting line.



```
R2 Score= -12.677179044980353
EVS= 0.0
MSE= 430022.6666666667
MAE= 631.3333333333334
RMSE= 655.7611353737477
<matplotlib.legend.Legend at 0x7f8bd571c070>
```

Fig. 9. Logistic Regression Model

Data separate fig. 9 shows LR regression model forecasting which has r2 score of 0.12 and we can see that the blue line is actual data, orange is predicted line and green is forecasting line.



```
R2 Score= 0.9645580743788175
EVS= 0.8786578892077475
MSE= 6615.982978797198
MAE= 88.2231575247546
RMSE= 51.99982729375891
<matplotlib.legend.Legend at 0x7f8bd4bfbd30>
```

Fig. 10. Polynomial Regression Model

Data separate fig. 10 shows Polynomial regression model forecasting which has r2 score of 0.96 and we can see that the blue line is actual data, orange is predicted line and green is forecasting line.

TABLE I. REGRESSION MODEL ANALYSIS

Model	R2-SCORE	EVS	MSE	MAE	RSME
Linear Regression	0.78	0.78	6761.32	76.39	82.23
Decision Tree	0.40	0.72	18714.66	121.33	136.80
Random Forest	0.27	0.52	22863	144.33	151.20
K-Nearest Neighbor	0.15	0.22	356406.43	183.93	190.80
Support Vector Regression	0.5308	0.68	14751.21	114.70	121.45
Logistic Regression	0.1267	0	430022.66	631.33	655.76
Polynomial Regression	**0.96**	**0.87**	**6615.98**	**88.22**	**51.99**

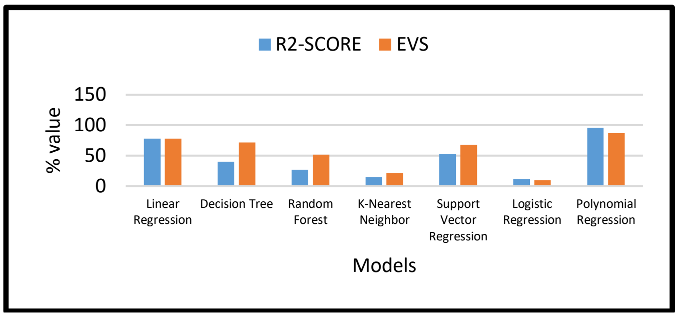


Fig. 11. Graphical Comparision of Machine Learning Models

Data separate fig. 11 shows graphical comparison pf regression models in which LR and Polynomial Regression gives best performance among all.

CONCLUSION

Ultimately, after testing several regression models for making predictions about regional female victimization in India, Polynomial Regression was found to be the most effective and accurate. After carefully comparing many different regression methods, such as K-Nearest Neighbor, Linear Regression, Random Forest, Decision Tree, Logistic Regression, Support Vector Regression, and Polynomial Regression, it was found that Polynomial Regression provided the most accurate and trustworthy predictions. The evaluation of various regression models for forecasting crimes against women reveals interesting insights. Among the models tested, linear regression exhibited the highest level of accuracy and precision, boasting an impressive R2-score of 0.78 and EVS of 0.78. This implies that the model can explain approximately 78% of the variance in the data. The model's performance is further supported by a relatively low MSE of 6761.32, MAE of 76.39, and RSME of 82.23, indicating minimal errors. In contrast, the decision tree, random forest, K-nearest neighbor, support vector regression, and logistic regression models demonstrated lower R2-scores and exhibited higher errors, suggesting decreased predictive capabilities. Notably, the polynomial regression model emerged as a standout performer, achieving an exceptional R2-score of 0.96, surpassing even linear regression in terms of accuracy.

It's worth noting, nevertheless, that the appropriate model to use in each crime study may vary depending on the dataset in question, the kind of predictors available, and the overall context of the investigation. The selection of Polynomial Regression as the optimum model for projecting regional women crimes in India should, therefore, be validated and refined via more study and analysis.

REFERENCES

[1] A.A.Biswas and S.Basak "Forecasting the Trends and Patterns of Crime in Bangladesh using Machine Learning Model," pp. 114-118, doi:10.1109/ICCT46177.2019.8969031

[2] S. Lavanyaa and D. Akila "Crime against Women (CAW) Analysis and Prediction in Tamilnadu Police Using Data Mining Techniques", pp 261-265, doi:10.31838/jcr.07.03.98

[3] B. Sivanagaleela and S. Rajesh "Crime Analysis and Prediction Using Fuzzy C-Means Algorithm", pp 595-599, doi:10.1109/ICOEI.2019.8862691

[4] Khushabu A. Bokde, Tiksha P. Kakade, Dnyaneshwari S. Tumsare , Chetan G. Wadhai and Deepa Bhattacharya "Crime Detection Technique Using Data Mining and K-Means", pp 223-226, doi:10.17577/IJERTV7IS020110

[5] Priyanka Das and Asit Kumar Das "CBehavioural analysis of crime against women using a graph-based clustering approach", doi:10.1109/ICCCI.2017.8117714

[6] G. Vicente, T. Goicoa and P. Fernandez-Rasines, M. D. Ugarte "Crime against women in India: unveiling spatial patterns and temporal trends of dowry deaths in the districts of Uttar Pradesh.", pp 655-679, doi: https://doi.org/10.1111/rssa.12545

[7] P. Tamilarasi, Dr.R.Uma Rani "Diagnosis of Crime Rate against Women using k-fold Cross Validation through Machine Learning Algorithms", pp 1034-1038, doi:10.1109/ICCMC48092.2020.ICCMC-000193

[8] S. Lavanyaa, D. Akila "Crime against Women (CAW) Analysis and Prediction in Tamil Nadu Police Using Data Mining Techniques", pp 595-599, doi:10.1109/ICOEI.2019.8862691

[9] Shiju Sathyadevan, Devan M.S, Surya Gangadharan. S. "Crime Analysis and Prediction Using Data Mining ", pp 406-412, doi:10.1109/CNSC.2014.6906719

[10] Priya Gandhi, Shayog Sharma "Approach of Predictive Modeling on Crime Against Women Problem", pp 284-288, https://www.ijrra.net/Vol5issue1/IJRRA-05-01-65.pdf

[11] Hyeon-Woo Kang, Hang-Bong Kang "Prediction of crime occurrence from multimodal data using deep learning", pp 1-19, doi: https://doi.org/10.1371/journal.pone.0176244

[12] Ritvik Chauhan, Vijay Kumar Baraik "Mapping Crime against Women in India: Spatio-Temporal Analysis, 2001-2012 ", pp 2243-2254, doi: doi.org/10.5281/zenodo.1127980

[13] Bhajneet Kaur, Laxmi Ahuja and Vinay Kumar "Factors Affecting Crime Against Women Using Regression and K-Means Clustering Techniques", pp 149-162, doi: https://doi.org/10.1007/978-981-10-3953-9_15

[14] B. Kaur, L. Ahuja, and V. Kumar, “Crime Against Women: Analysis and Prediction Using Data Mining Techniques,” in 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019, pp. 194–196. doi: 10.1109/COMITCon.2019.8862195.

[15] B. Patel and M. C. Zala, “Crime Against Women Analysis & Prediction in India Using Supervised Regression,” in 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), 2022, pp. 1–5. doi: 10.1109/ICEEICT53079.2022.9768533.

[16] U. Gupta and R. Sharma, “Analysis of criminal spatial events in india using exploratory data analysis and regression,” Computers and Electrical Engineering, vol. 109, p. 108761, 2023, doi: https://doi.org/10.1016/j.compeleceng.2023.108761.

[17] L. G. A. Alves, H. V Ribeiro, and F. A. Rodrigues, “Crime prediction through urban metrics and statistical learning,” Physica A: Statistical Mechanics and its Applications, vol. 505, pp. 435–443, 2018, doi: https://doi.org/10.1016/j.physa.2018.03.084.

[18] M. Vivek and B. R. Prathap, “Spatio-temporal Crime Analysis and Forecasting on Twitter Data Using Machine Learning Algorithms,” SN Computer.