Crime Analysis in India with Interactive Visualization

Avani Vaishnav
B.E ISE
Ramaiah Institute of Technology, Bangalore

Ayana Holla P.
B.E ISE
Ramaiah Institute of Technology, Bangalore

Aishwarya Vijaykumar Sheelvant
B.E ISE
Ramaiah Institute of Technology, Bangalore

ABSTRACT
Evaluating and predicting the crime rate in any country is critical for the concerned government authorities to find ways to minimize its effect on the community and to devise methods to ultimately curb such practices. Previous study suggests a correlation between education in a country and the rate of poverty, education, and unemployment as well as unemployment and poverty in the country. This paper tries to introduce a computational model to analyze the relationship between education, poverty and unemployment rates to the crime rates in each state of India as well as the rate of crime contribution to the total crime rate in each state. Data is sourced from verified and trusted government data collection websites to keep the study authentic. The paper uses Machine Learning to recognize the influence each of the chosen socio-economic indicators has on crime as a whole. This approach consists of three components: data collection and pre-processing, employing the proposed machine learning algorithms such as simple linear regression and multiple linear regression, and lastly, visualizing data.

General Terms
Data Science, Machine Learning, Crime Prediction.

Keywords
Ordinary Least Square, Visualization, Data Cleaning, Analysis of Crime Dataset, Regression.

1. INTRODUCTION
The prediction of crime occurrences has received attention due to its prospective benefits [1][2][3]. There are myriad factors involved in the prediction of crime – physical and psychological factors, rational and irrational behavior, upscale or destitute community, etc. All of these aspects combine to make the crime rate volatile and very difficult to accurately forecast with a high degree of accuracy. Fundamental crime reduction includes the strengthening of patrols, which is financially and humanly expensive. In addition, many such records show that patrols are generally carried out based on the position of identified crime-ridden districts or police scientific expertise.

There seems to be a vicious cycle between poverty, education, unemployment, and crime. Developing counties (such as India) can be subject to extreme poverty which leads to people being denied or choosing not to educate their children and instead of finding minimum wage jobs to support their families. Over time this leads to a steep decrease in education rates in the country. Structural changes such as technological change, globalization, and the shifting economic environment along with under qualification of people lead to an increase in unemployment rate in the country. The challenge of this paper is to analyze precisely the dependence that crime is driven by. With a successful model that can predict the crime occurrence, it is possible to gain insight into the drawbacks of the pre-existing system in society, and reasons for high rates of crime contribution that would otherwise not have been noticed.

2. LITERATURE REVIEW
In recent years, several studies have been conducted on the prediction of crime occurrence and its dependency on various other social factors. The results of these predictive analyses were intended to assist in crime prevention and subsequently help in policing and geographic profiling. Previous studies were conducted using data from multiple domains such as education, demographics, and economics.

For this implementation, several journal papers have been referred and a brief description of a few has been mentioned below.

[4] This paper aims to predict crime types and crime hotspots using real-time datasets. The datasets used are for two regions: Denver and Los Angeles. The datasets have attributes such as crime type, crime occurrence location, time, month, date. Additionally, demographic datasets of both the cities were used to analyze the safety of the neighborhoods. After preprocessing of data initial data analysis is done through statistical methods. Visualization is done to compare the crime datasets of both the cities, for example, the percentage of crime occurrences over the 12 months in Denver and Los Angeles, the percentage of crime occurrences over the days of the week in Denver and Los Angeles etc. Apriori algorithm is applied on both datasets to extract frequent patterns. Multiple experiments were conducted to find the optimum support values. It is found that Denver has 62 frequent patterns and Los Angeles has 59 patterns. Crime prediction was performed using two models the Naive Bayes and the Decision Tree. Naive Bayes proved to be simpler and more accurate in the study. Many interesting relations were found by visualization such as people's age and gender distribution vary between dangerous and safe locations. The results of the study could potentially be used to raise awareness regarding criminal hotspots in a specific location and at a particular time.

Machine learning techniques were utilized to examine crime data and find the economic elements that influence crime in India in this paper [5]. The unemployment rate and the gross domestic product (GDDP) were used as independent input variables in experiments employing various machine learning techniques such as Decision Trees, Random Forest, Linear Regression, and Neural Networks where target variables are theft, burglary, and robbery. The correlation coefficient r, the coefficient of determination R2, Absolute Mean Error, and Accuracy were used to evaluate the regression model's performance. The Linear Regression algorithm outperformed the other three machine learning techniques. According to the study, there is a one-way causal relationship between

unemployment and robbery. This causality is extremely similar to the result of our paper, which found that literacy rate and crime contribution rate had unidirectional causation.

The crime rates from 2001 to 2012 were studied in this research [6], and the results were used to rank the states and union territories based on their average IPC crime rate. The proposed workflow begins with the acquisition of raw data followed by data cleansing. K-means clustering is a two-step process that begins with the development of a model and ends with the application of the model for classification. The clusters are constructed based on the similarity of the criminal attributes and are discovered using k-means clustering. For categorization, the data is subjected to the Random Forest method and neural networks. The crime hotspots are indicated on the India map using Google marker clustering for visualization. The proposed model's correctness is measured and verified using WEKA. The cases have a high accuracy of 99.93 percent, indicating that the authors are correct.

[7] This paper gives a comparison of three algorithms namely SVM, Random Forest, Linear Regression. A website has also been established which can be used for prediction, observation, and viewing visualizations. The types of crime considered are - crimes against women, crimes against children, crimes under the Indian Penal Code (IPC) and crimes under Special and Local Laws (SLL). Linear Regression has been implemented along with Mean Squared error calculation. Among the four techniques used it is stated that Linear Regression is the most accurate. Random forest and SVM algorithms tend to overfit the data and the results produced were not accurate when compared to the actual trends. Visualization is also done to provide a clear representation. Various factors such as population, literacy rate corresponding to each demographic area are considered.

[8]This paper uses regression analysis based on geographic location, time and type of crime to predict crime. The authors have used a database to store crime records and other related observations. This analysis is done state-wise and also focuses on different types of crime like rape, assault, theft, dowry, etc. in each state. Based on these attributes the geographic hotspots can be found out. Few visualizations have also been done such as a pie chart to show the number of rape cases state-wise with respect to each age group.

[9]This paper focuses on finding out the best algorithm for predicting crime. The five algorithms compared are Decision Trees, Random Forest, SVM, Logistic Regression, K Nearest Neighbor. The features used for predicting are District Boundary, Police Service Area boundary, crime category code, year of observations, month, area, etc. The comparison is made based on different metrics such as recall, F1 score, precision, accuracy. It is concluded that Decision tree algorithms are the most efficient in predicting crime. Various visualizations have been done to analyze the records such as crime rate by year, city-wise crime rate, classification of crime rate by PSA.

The proposed system in this paper [10] firstly aims to make the best use of resources, identify crime hotspots and allocate vigilant resources such as police officers, police cars, and weapons, as well as reschedule patrols based on the susceptibility of a location. Secondly, to ensure a safer society by reducing crimes such as murder, rapes, thefts, drug trafficking, and other smuggling. The dataset is analyzed using supervised machine learning techniques (SMLT) to capture a variety of information, including variable identification, univariate analysis, bivariate analysis, and multivariate analysis. The goal is to classify whether the crime rate is high or low which helps in identifying the crime hotspots in the real-time world.

This paper [11] examines crime analysis and prediction utilizing a variety of data processing techniques. A key contribution made in this research is to present a method that supports data processing and categorization, which becomes more difficult when dealing with enormous amounts of data; as a result, machine learning minimizes the amount of time required to predict crime, analyze data, and produce graphs. The data comes from the Chicago data portal. Data smoothing is the next stage. The third stage is prediction, for which logistic regression was used since it performed better on the Chicago Dataset than decision tree and random forest algorithms. Data visualization is the final phase.

The next paper [12] proposes a prototype based on an actual and original dataset found on the Kaggle computing gadgets repository website. 'Crime record' is the name of this labeled dataset. The paper proposes the use of random forest and light GBM algorithms after data pre-processing. The goal of this study is to determine the type of crime that is occurring using algorithms on datasets, with a focus on time and location. Random forest algorithm achieves an accuracy of 95% and light GBM 98% because the ensemble approach integrates many tree classifiers and has considerably higher prediction results, the usage of bagging results in the highest precision.

3. PROPOSED MODEL

3.1 Predictive Model

Multiple Linear Regression (MLR) is a predictor whereby a single dependent variable is estimated using more than one independent variable. MLR is a simple and growing approach to linear regression analysis.

Multiple regression is, in essence, the extension of Ordinary Least-Squares (OLS) regression involving more than one independent variable [13]. Ordinary squares are the least-squares linear approach for estimating unknown parameters in a linear regression model. OLS selects the parameters of a linear function of a series of independent variables by the minimum square principle: to minimize the sum of the squares of the differences between the observed values (values of the variable being observed) and predicted values in the data set [14].

3.2 Functional Diagram of Proposed Model

It can be divided into divided into six steps:

a. Data collection and organization

b. Pre-processing and cleaning

c. Linear and Multiple Linear Regression Analysis

d. Evaluation and verification of results of R-square value

e. Projection and inference of data

f. Visualization

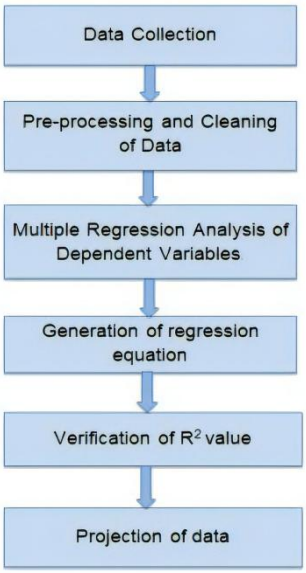


Figure 1: Flowchart of Proposed Model

3.3 Methodology

Multiple linear regression is an extension of Linear Regression. Multiple Linear Regression makes use of two or more independent variables to predict the outcome of a single continuous dependent variable. The equation for multiple linear regression is as follows:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 \ldots\ldots + b_pX_p$$

Where $\hat{Y}$ is the expected variable, $X_1 \ldots X_p$ are independent variable, b_0 is the value of $\hat{Y}$ when all the independent variables are equal to zero and $b_1 \ldots b_p$ are the estimated regression coefficients.[15]

Multiple linear regression has a number of assumptions which are often not possible to test before running the model. Therefore, model diagnostics should be run after fitting the model to test the following assumptions:

1. Data must not have multicollinearity
2. The residual errors should be approximately normally distributed
3. Homoscedasticity
4. Independency of errors

Ordinary Least Squares (OLS) model was used which corresponds to minimizing the sum of square differences between the observed and predicted values.

4. IMPLEMENTATION

4.1 Data Collection

Data used in this project is a combination of features extracted from four individual datasets. The data for Literacy Rate as well as Population of each state in India was scraped from the 15th Official Census calculation conducted in 2011. Data for Poverty Rate in India is based on Tendulkar poverty estimation and was appropriated from the official website of data.gov.in. The data for unemployment rates of each state in India was scraped from the NSSO (National Sample Survey Office) report for unemployment released by the Ministry of Statistics and Programme Implementation. The crime dataset was sourced from the NCRB website. The final dataset used contained the features Literacy Rate, Unemployment Rate, Poverty Rate, and Crime Rate for each state in India. The description of the considered attributes are given below

1. **Literacy Rate:** Also called the "effective literacy rate"; the total percentage of the population of an area at a particular time aged seven years or above who can read and write with understanding. A person, who can only read but cannot write, is not literate.
2. **Poverty Rate:** The Tendulkar Committee stipulated a daily target of Rs.32 and Rs.47 per capita spending in rural and urban areas, respectively, and achieved a cut-off of around 30 percent of the population below the poverty line. [16]Here the model is trying to assess if poverty causes an individual to commit crimes to earn for a living, like theft etc.
3. **Unemployment Rate:** [17] National Sample Survey Office (NSSO), an organization under Ministry of Statistics and Programme Implementation (MoSPI) measures unemployment in India on following approaches:
    1. Usual Status Approach
    2. Weekly Status Approach
    3. Daily Status Approach

Figure 2 shows the final merged dataset that is used for further analysis and visualization.

	State/UT	Unemployment rate	Poverty Rate	% Crime contribution	Literacy rate
0	Andhra Pradesh	2.0	9.20	8.2	67.02
1	Arunachal Pradesh	2.2	34.67	0.1	65.38
2	Assam	4.6	31.98	2.9	72.19
3	Bihar	3.4	33.70	5.8	61.80
4	Chhattisgarh	1.4	35.93	2.5	70.28
5	Delhi	3.8	9.91	2.3	86.21
6	Goa	4.9	5.09	0.1	88.70
7	Gujarat	0.5	17.63	5.3	78.03
8	Haryana	2.9	11.16	2.6	75.55
9	Himachal Pradesh	1.3	8.00	0.6	82.80
10	Jammu & Kashmir	3.4	10.35	1.1	67.16

Figure 2: Combined Dataset

Another dataset used in this analysis for creation of India maps is in GeoJSON format. A GeoJSON is an open standard geospatial data interchange format that represents simple geographic features and their nonspatial attributes. This dataset is used in mapping and creating boundaries of states of India in the visualization. The dataset is sourced from an authentic opensource platform (GitHub). While the dataset has 10 raw attributes but only three are of value in the analysis:

1. **Name:** It is the name of the state or union territory. This attribute helps to uniquely identify and merge the geospatial dataset with the socio-economic dataset which is used for analysis.
2. **Type:** Gives information if the entry is a state or union territory.
3. **Geometry:** This is the most important feature of this dataset. It contains mapping information in the

form of polygons or multi-polygons. The set of coordinates in this attribute combine to form a closed shape of that corresponding state.

4.2 Data Pre-processing

Data pre-processing includes methods to remove any anomalies in data such as missing value, inconsistent values or duplicate values. The dataset used does not have any duplicate values.

To begin the analysis, firstly the dataset is sorted in ascending order. Any records with NaN (Not a Numeric) values or any missing data is dropped from further analysis.

4.3 Feature Selection

In this method any expendable features that do not contribute to the study are removed from further analysis. This is done to increase computation speed, reduce memory usage and remove any discrepancies that could hinder the analysis. The final attributes included in the analysis Poverty rate, Unemployment rate, % Crime contribution and Literacy rate.

4.4 Training

This method divides the dataset into train and test sets randomly. The data is split in the ratio 8:2 such that 80% of the data is to be used to train the prediction model and the subsequent 20% is used to test predictions of the model

4.5 Prediction

Results from the prediction model are analyzed and inferences drawn, results of which are shown in the next section.

5. RESULT AND TESTING

Table 1: Result of In-Sample and Out-Of-Sample Forecast

Graphics	Top
R^2 in- sample	0.2642488982254946
R^2 out- sample	-0.08472276527621658
In-sample Estimate (ISE)	2.723789078128938
Out-sample Estimate (OSE)	3.4864155499445566

Table 1 gives a summary of the In-Sample Estimate (ISE) and Out-Of-Sample Estimate (OSE) of data. ISE utilizes a subset of the complete data to train the model and forecasts values outside of the estimation period. This is done to assess the ability of the model to predict known values. Whereas an OSE trains the model on the complete set of data and performs predictions of this set. OSE values are used to assess the ability of the model to forecast unknown values.

The obtained ISE and OSE values are not very large hence, the model seems to be a good fit.

Table 2: Summary of Linear Regression Model

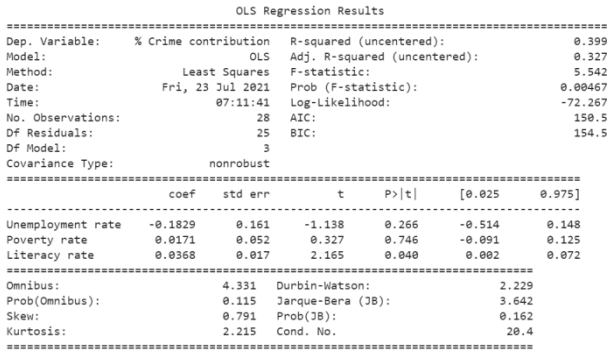
Independent Feature	Slope	Intercept	r-value	p-value	stder
Literacy	-0.9364	80.60503	-0.34786	0.034795	0.425336
Unemployment	-0.3469	4.62736	-0.29377	0.0867312	0.1965246
Poverty	-0.06516	17.975434	-0.019121	0.9131835	0.5931381

Table 2 shows the summary of the linear regression model. The significant testing of each independent variable is carried out to test whether it helps predict or explain the independent variable and this is done using the test of hypothesis. If a variable does not reject the null hypothesis, it does not have a linear relationship with the dependent variable and is excluded from the analysis.

In this analysis the p-value of only literacy rate is below the alpha value (0.05), therefore it rejects the null hypothesis. Linear regression is performed as preliminary test for multiple linear regression. Since there are three independent variables and one dependent variable multiple linear regression is employed.

The terms in linear regression are explained as follows:

1. **Slope**: slope of the regression line
2. **Intercept**: Intercept of regression line
3. **R-value:** Correlation coefficient
4. **P-Value**: The p-value for the hypothesis test whose null hypothesis is that the slope is zero
5. **Stderr**: Standard error of the estimated slope.



OLS Regression Results

Dep. Variable:	% Crime contribution	R-squared (uncentered):	0.399
Model:	OLS	Adj. R-squared (uncentered):	0.327
Method:	Least Squares	F-statistic:	5.542
Date:	Fri, 23 Jul 2021	Prob (F-statistic):	0.00467
Time:	07:11:41	Log-Likelihood:	-72.267
No. Observations:	28	AIC:	150.5
Df Residuals:	25	BIC:	154.5
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Unemployment rate	-0.1829	0.161	-1.138	0.266	-0.514	0.148
Poverty rate	0.0171	0.052	0.327	0.746	-0.091	0.125
Literacy rate	0.0368	0.017	2.165	0.040	0.002	0.072

Omnibus:	4.331	Durbin-Watson:	2.229
Prob(Omnibus):	0.115	Jarque-Bera (JB):	3.642
Skew:	0.791	Prob(JB):	0.162
Kurtosis:	2.215	Cond. No.	20.4

Figure 3: Summary of OLS Model

Figure 3 shows the summary of the OLS model. The summary provides several measures to give us an idea of the data distribution and behavior. The summary helps to analyze the data and check whether it has corrected characteristics to give us confidence in the resulting model. [18]

The first section gives a general information about the model such as the dependent variable and when the model was built. This section also includes Df Residuals and Df Model. Df here, stands for degree of freedom which indicates the number of independent values that can vary in an analysis without breaking any constraints [19]. Residuals in regression are simply the error rate which is not explained by the model. It is the distance between the data point and the regression line. Thus df (Residual) is the sample size minus the number of parameters being estimated [20]. Df Model is simply the number of X variables baring the constant variable.

The R-squared and Adj. R-squared values measure the performance of the model. [21]An adjusted R-squared value of 0.32 indicates that the model explains approximately 32 percent of the variation in the dependent variable.

Prob (F-Statistic) indicates the overall significance of the regression. It depicts the probability of the null hypothesis being true. A value close to 0 is expected. The value obtained is 0.00467.

AIC stands for Akaike's Information Criteria and is used for model selection. It is calculated as the number of parameters minus the likelihood of the overall model. A lower AIC implies a better model. Whereas, BIC stands for Bayesian information criteria and is a variant of AIC, where penalties are made more severe [22].

Table 3: Obtained AIC and BIC values

Measure	Value
AIC	150.5
BIC	154.5

The strength and type of relationship between the independent variable and dependent variable is given by the coefficient. The values obtained are:

Table 4: Obtained Coefficient values

Feature	Coefficient value
Unemployment Rate	-0.1829
Poverty Rate	0.0171
Literacy Rate	0.0368

Omnibus/Prob (Omnibus) are tests of the skewness and kurtosis respectively of the residual. To indicate normalcy a value close to 0 is required. Prob (Omnibus) performs a statistical test indicating the probability that the residuals are normally distributed [23]. A value close to 1 is expected. As shown in the above summary Omnibus is very high and the Prob (Omnibus) is relatively low. In this model the values obtained are:

Table 5: Obtained Omnibus and Prob (Omnibus) Values

Measure	Value
Omnibus	4.331
Prob (Omnibus)	0.115

Skew measures the symmetry of the data. A value close to 0 is expected which indicates the data follows a normal distribution. The value obtained in the analysis is 0.791.

Kurtosis measures the curvature of the data. Higher peaks lead to greater kurtosis. Greater kurtosis can be interpreted as a tighter clustering of data around zero, implying a better model with few outliers. Kurtosis value obtained in the analysis is 2.215.

Durbin-Watson is a test for homoscedasticity. Expected value is between 1 and 2. Durbin-Watson value obtained is 2.229

Jarque-Bera (JB)/ Prob (JB), like the Omnibus test, this value indicates both skew and kurtosis. [24] It is supposed to agree with the Omnibus test. The values obtained are:

Table 6: Obtained JB and Prob (JB) Values

Measure	Value
JB	3.642
Prob (JB)	0.162

6. VISUALIZATION

Figure 4 shows states of India and their percent contribution to total crimes in India. The y-coordinate denotes percent contribution to crime and x-coordinate shows the corresponding state. Graph shows that Madhya Pradesh has the highest crime contribution in India.

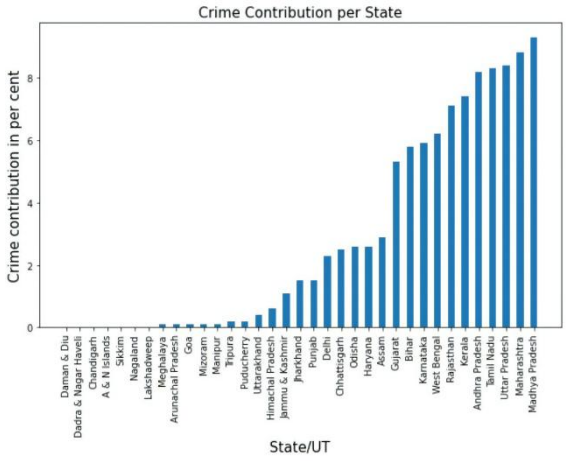


Figure 4: Crime Contribution per State

Error! Reference source not found. shows the correlation matrix of the dataset. A correlation matrix displays the correlation coefficients for the various different variables considered, viz unemployment rate, poverty rate, literacy rate and per cent crime contribution. A dark blue square in the matrix indicates a strong positive correlation between the corresponding variables and a deep red color indicates a strong negative correlation between the variables.

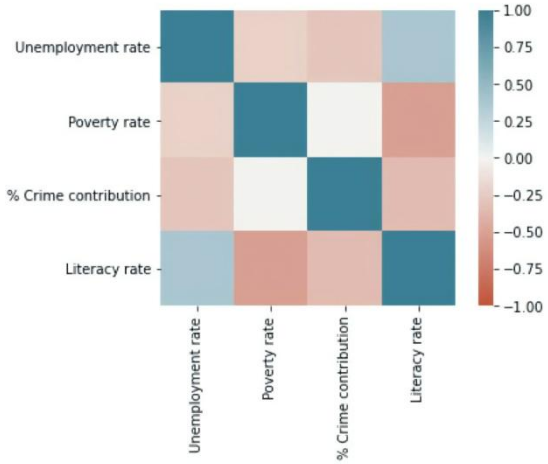


Figure 5: Confusion Matrix for Attributes

Figure 6 is an interactive box plot that graphically depicts the various groups of numeric data through their quartiles. The lines extending from the top and bottom of the box indicate variability outside the upper and lower quartile. Middle line of the box indicates 2nd quartile of median value of that group, upper line indicates 1st quartile and last line indicates 3rd quartile of the data.

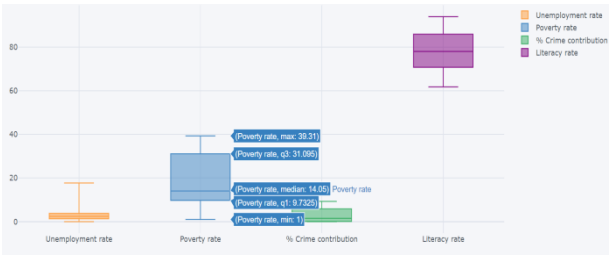


Figure 6: Descriptive Analysis of Data using Box Plot

Figure 7 shows a bar chart depicting the difference in actual and predicted values of the linear regression model.

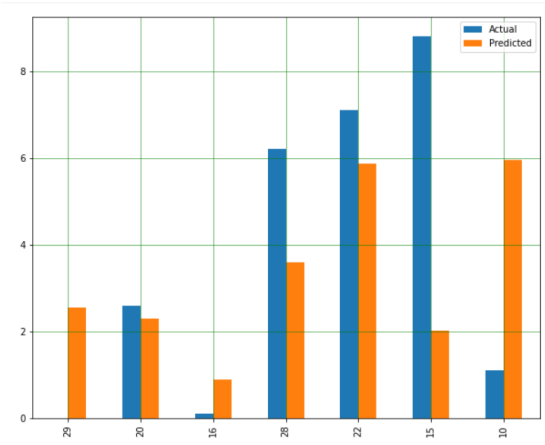


Figure 7: Comparison between Actual and Predicted Values

Figure 8 shows the comparison of literacy rate and gender for each state of India. The x-coordinate is the state and corresponding y-coordinate denotes rate of literacy.

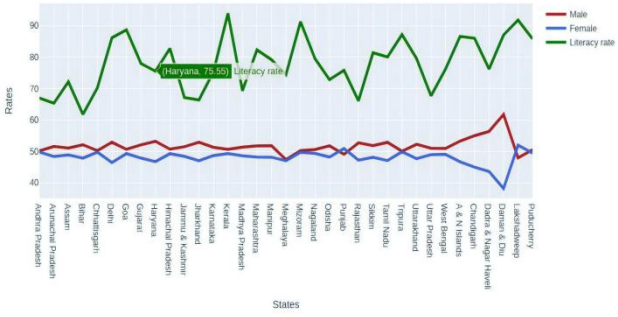


Figure 8: Comparison between Literacy Rate Vs. Gender

Figure 9, Figure 10, Figure 11 and Figure 12 show crime contribution, literacy rate, unemployment rate and poverty rates per state respectively. These visualizations were created using Python packages Geopandas and Matplotlib.

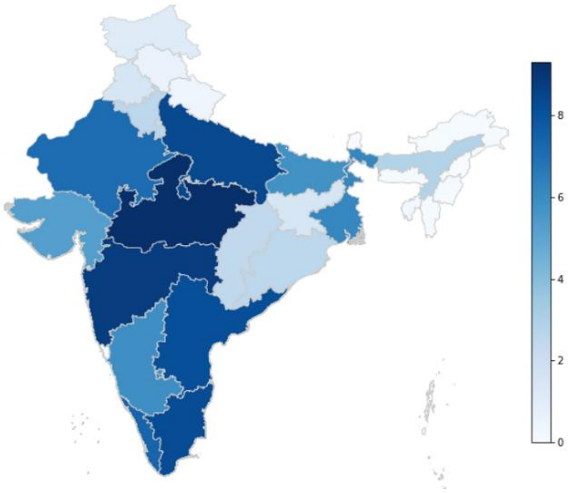


Figure 9: Crime Contribution per State

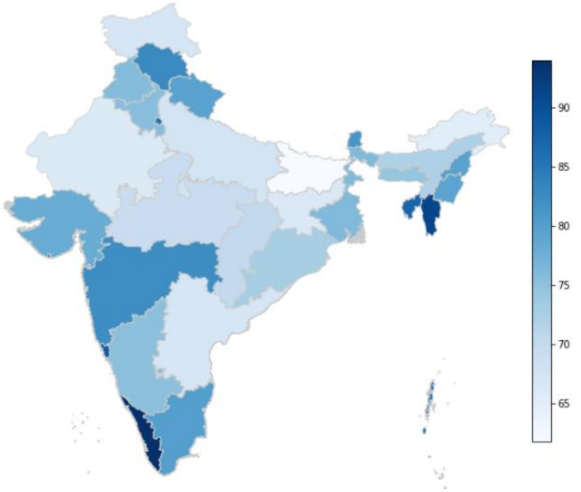


Figure 10: Literacy Rate per State

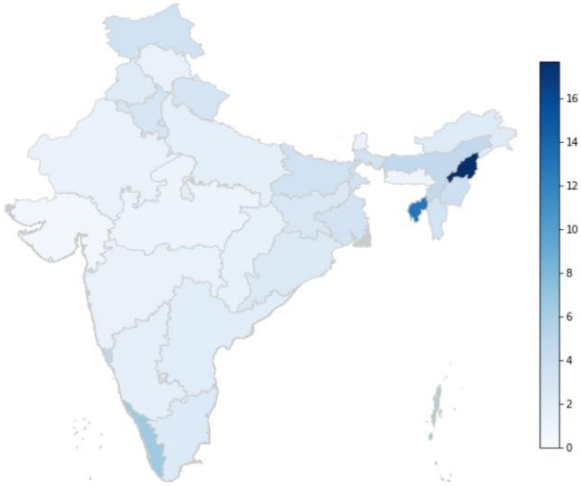


Figure 11: Unemployment Rate per State

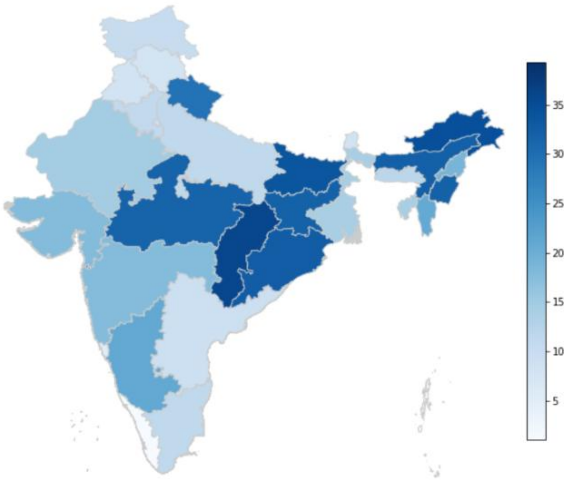


Figure 12: Poverty Rate per State

Figure 13, Figure 14, Figure 15 and Figure 16 show percentage crime contribution, literacy rate, unemployment rate and poverty rate in an interactive map format. This implementation made use of Python libraries Geopandas and Folium. These visualizations are interactive because hovering the mouse over them provides more information about the attributes. For example, by placing mouse over the state of Jammu and Kashmir we can get added information that the rate of crime is 1.1.

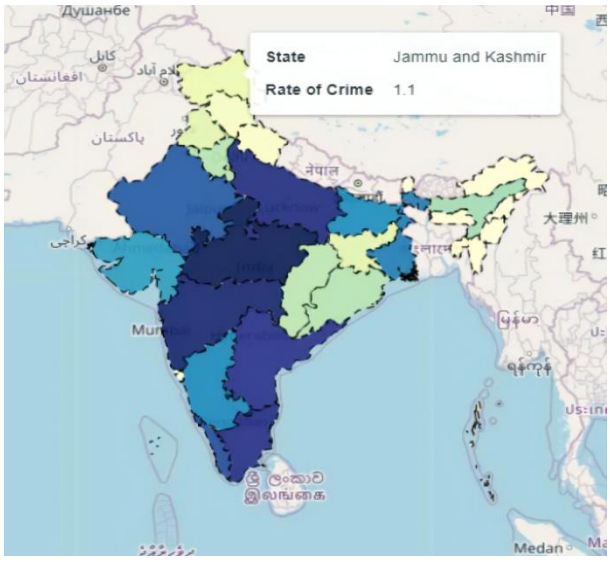


Figure 13: Interactive Map of Crime

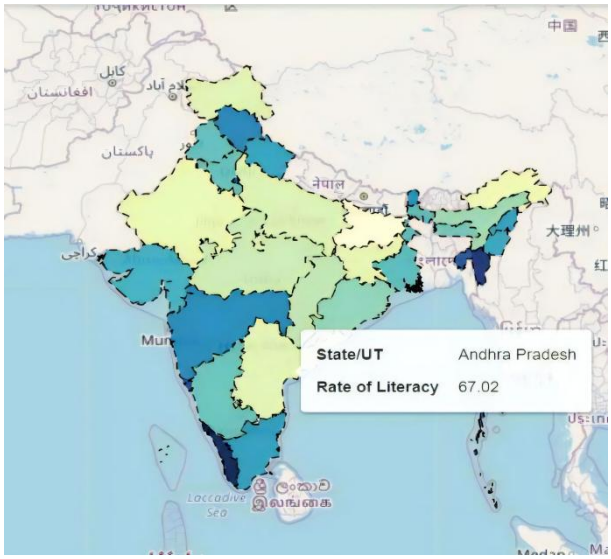


Figure 14: Interactive Map of Literacy Rates

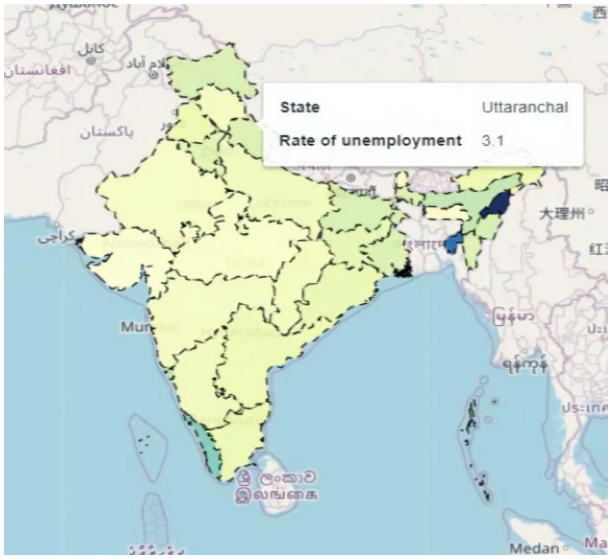


Figure 15: Interactive Map of Unemployment Rates

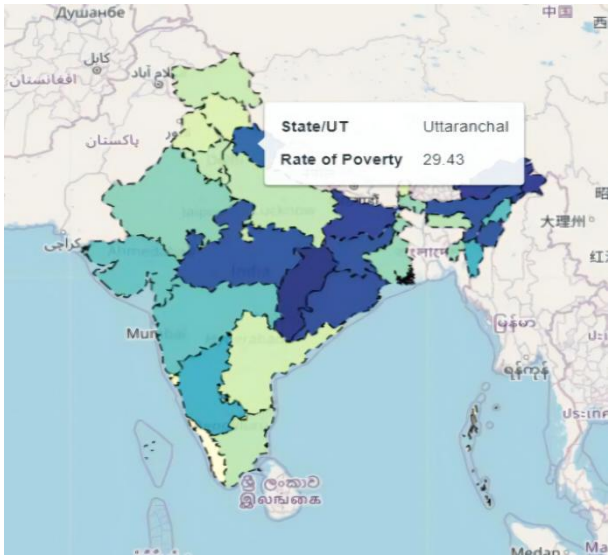


Figure 16: Interactive Map of Poverty Rates

7. CONCLUSION

This model was built with an aim to help Government agencies and law enforcement bodies inculcate some changes in the present system. From results of the model obtained it can be concluded that the predicted results are not 100% accurate but this model can be used to reduce crime rate to a certain extent by providing security in crime sensitive areas. The data visualizations also help us analyze which are the sensitive areas with respect to poverty, unemployment, literacy and crime. Since the model has been built on real time data the results were not as expected. The project's possible future work will include further refinement of the model and also different methods of analysis for more accurate results. There is no user interaction provided in this project. A UI can be provided where users can get past information. Possible improvement in the collection of data and as well as creation of a database consisting of data over the years will be key focus of the future work.

8. ACKNOWLEDGMENT

The progress of this paper would not have been possible without the unconditional support of many people, each of whom have helped us in carrying out this paper. Lastly, our deep and sincere gratitude to our families and friends for their continuous and unparalleled love, help and support.

9. REFERENCES

[1] Hajela, G., Chawla, M., & Rasool, A. 2020. A clustering based hotspot identification approach for crime prediction. Procedia Computer Science, 167, 1462-1470.

[2] Hossain S., Abtahee A., Kashem I., Hoque M.M., Sarker I.H. 2020. Crime Prediction Using Spatio-Temporal Data. In: Chaubey N., Parikh S., Amin K. (eds) Computing Science, Communication and Security. COMS2 2020. Communications in Computer and Information Science, vol 1235. Springer, Singapore.

[3] Bappee, F. K., Petry, L. M., Soares, A., & Matwin, S. 2020. Analyzing the impact of foursquare and streetlight data with human demographics on future crime prediction. arXiv preprint arXiv:2006.07516.

[4] Almanie, T., Mirza, R., & Lor, E. 2015. Crime prediction based on crime types and using spatial and temporal criminal hotspots. arXiv preprint arXiv:1508.02050.

[5] Mittal, M., Goyal, L. M., Sethi, J. K., & Hemanth, D. J. 2019. Monitoring the impact of economic crisis on crime in India using machine learning. Computational Economics, 53(4), 1467-1485.

[6] Deepika, K., & Vinod, S. 2018. Crime analysis in India using data mining techniques. International Journal of Engineering & Technology, 7(2.6), 253-258.

[7] Sathyadevan, S., Devan, M. S., & Gangadharan, S. S. 2014. Crime analysis and prediction using data mining. In 2014 First International Conference on Networks & Soft Computing (ICNSC2014*)* (pp. 406-412). IEEE.

[8] Bodare, S., Kurkute, S., Akash, M., & Pawar, R. 2019. Crime Analysis using Data Mining and Data Analytics.

[9] V, K., A, K. P., M, L., & D, L. S. 2019. Prediction of Crime Rate Analysis Using Supervised Classification. International Research Journal of Engineering and Technology (IRJET), 06(03), 6771-6775.

[10] Keerthi.R, Kirthika.B, Pavithraa.S, & V.Gowri, D. 2020. Prediction of Crime Rate Analysis Using Machine Learning Approach. International Research Journal of Engineering and Technology (IRJET), 07(09), 1617-1621.

[11] Prasad, P., Singh, R., Singh, R., & Kothari, P. R. 2020. Crime Analysis and Prediction Using Data Mining. International Research Journal of Engineering and Technology (IRJET), 07(04), 5732-5735.

[12] Pranav, D., Yamini, C., Pranathi, A., Sasidhar, A., & Ch, S. C. 2021. Machine Learning Analysis on Crime Prediction System. Machine Learning, 8(04).

[13] Chowdary, D. H. 2020. Multiple Linear Regression Explained - Analytics Vidhya. Retrieved September 14, 2021, from https://medium.com/analytics-vidhya/multiple-linear-regression-explained-215f2683cd5a

[14] Wikipedia contributors. 2021. Ordinary least squares. Retrieved September 14, 2021, from https://en.wikipedia.org/wiki/Ordinary_least_squares

[15] Multiple Linear Regression Analysis. 2013, January 17. Retrieved from Boston University School of Public Health: https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable7.html

[16] B2B. 2017. Poverty Lines in India: Estimations and Committees. Retrieved from Civilsdaily: https://www.civilsdaily.com/poverty-lines-in-india-estimations-and-committees/

[17] Team, U. T. 2021. Types of unemployment in India, causes and solutions. Retrieved from Utkal Today: https://www.utkaltoday.com/types-of-unemployment-in-india/

[18] Dismuke, C., & Lindrooth, R. 2006. Ordinary least squares. Methods and Designs for Outcomes Research, 93, 93-104.

[19] Frost, J. 2021. Degrees of Freedom in Statistics. Retrieved September 14, 2021, from https://statisticsbyjim.com/hypothesis-testing/degrees-freedom-statistics/

[20] Multiple Regression. (n.d.). Retrieved September 14, 2021, from https://people.richland.edu/james/ictcm/2004/multiple.html

[21] Lewis-Beck, M. S., & Skalaban, A. 1990. The R-squared: Some straight talk. Political Analysis, 2, 153-171.

[22] Yadav, J. 2019. Statistics: How Should I interpret results of OLS? Retrieved September 14, 2021, from https://www.linkedin.com/pulse/statistics-how-should-i-interpret-results-ols-jyoti-yadav/

[23] McCarty, K. 2018. Interpreting Results from Linear Regression – Is the data appropriate? Retrieved September 14, 2021, from https://www.accelebrate.com/blog/interpreting-results-from-linear-regression-is-the-data-appropriate

[24] Koizumi, K., Okamoto, N., & Seo, T. 2009. On Jarque-Bera tests for assessing multivariate normality. Journal of Statistics: Advances in Theory and Applications, 1(2), 207-220.