iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VII **Month of publication:** July 2021

DOI: https://doi.org/10.22214/ijraset.2021.37094

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

An Emperical Investigation of Socio-Economic Parameters Associated with Crime in Karnataka Using Predictive Analysis

Dr. Monica R Mundada[1], Punith S[2], Ravva Venkata Subba Aravind[3], Rashad Mohamed Khan[4], Shashank Prakash Patil[5]
[1, 2, 3, 4, 5]*Department of Computer Science and Engineering, Ramaiah Institute of Technology*, Bangalore, India

Abstract: ***The crime rate in India has been on a rise with growing population and rapid development. Crimes are a social nuisance and brings disrepute to the society and nation at large. Data mining techniques have enabled models to predict crime. The law enforcement officers and police personnel's need to spend umpteen time to analyze crime from the crime reports published by National Crime Records Bureau and respective State police departments. Therefore, there is necessity of a mechanism which can predict crime by identifying factors responsible for increase in crime rate. This project is an attempt to address this. Socio-economic variables like poverty, urbanization, literacy, and the demographic and social composition of the population are recorded from Census data. Machine Learning algorithms namely Multiple Linear Regression, Random Forest Regressor and Generalized Linear Models are trained and deployed. These algorithms have been implemented assuming both linear and non-linear ship of data and subsequent results have been compared. Further, choropleth maps have been plotted to represent and understand data in a simple way. The results show that socio-economic factors are good indicator in explaining crime rates and good performance is observed with few models***

Keywords: crime, socio-economic, regression, analysis, Karnataka

I. INTRODUCTION

Crime is an evil that degrades the quality of life and affects everybody in a civilized society. Crime sabotages the potential of the country to promote peace and development from a wider perspective [1]. Historically solving crimes has been the privilege of legal enforcement specialists. But lately computer analysts have aided the legal enforcement in solving crimes by discovering crime patterns. High crime rates can significantly impact and negatively influence foreign and domestic investments [2]. Violence affects human wellbeing in indirect ways, as when armed conflicts undermine economic growth or the functioning of public services. Crime rate has been increasing over the period of years. There is an ardent need and necessity to understand the pattern of crime and suggest effective measures to curb crime as much as possible. There also exists the need to understand the influence of various socio- economic and regional factors which contribute to the increase in crime rate and devise tools to mitigate it. One of the important reasons why violence is an under-researched issue in development studies is the paucity of relevant data. Basic data on criminal violence in developing countries are seldom available in a convenient and reliable form [3]. The Indian government, however, publishes good deal of information on crimes with a dedicated agency called National Crime Records Bureau responsible for it [4]. This project is a preliminary attempt to analyse these data and through a thorough understanding of the various factors associated with crime, make relevant predictions which help in curtailing crime. Whilst the crime records for the entire country are available, this research's' main concern and objective is to explore the links between crime rates in the various districts in the Karnataka state socioeconomic variables as poverty, urbanization, literacy, and the demographic and social composition of the population.

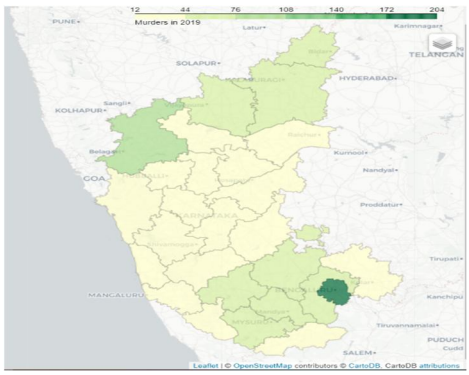


Figure 1: Choropleth map of Karnataka representing crime intensity of different district for crime Murder in 2019

II. LITERATURE SURVEY

Early theories of crime highlight the effect of poverty and social deprivation on crime rates [5]. Among economists, Fleisher thoroughly studied criminal behaviour. He argued that crime rates are associated with unemployment and low income levels [6]. Dubey opines that that economic, sociocultural and political factors play an important role in crime and crime control practices in India. They are of the opinion that the increasing crime rate has been influenced by the constant political tussle in India as well as due to the financial crisis [7]. Gary Becker presented a model based on the cost of crime. He explained the economics of crime in terms of the cost and benefits of a particular crime. He stated that the cost of different punishments to an offender could be made comparable by converting them into their monetary equivalent or worth [8]. Dutta and Husain investigated the impact of variables like police force and arrest rate. Socioeconomic variables like poverty and urbanization on crime in India were also studied. They concluded that these parameters are likely to have a significant impact on crime rates [9]. Gumus studied the effect of per capita income, income inequality, population, and presence of black population on the crime rate in the US and stated that these all are important determinants of the crime rate. Police expenditures and unemployment rate also have an impact on crime but not as much as other factors. Shingleton created a multi-variable regression model using predetermined environmental factors affecting crimes. This model is used to predict future violence levels using statistical analysis [10]. Uddin, & O., Osemengbe applied data mining in the context of law enforcement and intelligence analysis. The research held the promise of alleviating crime related problems. Data mining techniques as a means of detecting crimes like sex crime, theft, fraud, arson, gang drug have been discussed. [11]

In another work, a random forest regressor is used to predict crime and quantify the influence of urban indicators on homicides. Data from urban indicators of all Brazilian cities is used to train the model, and necessary conditions are studied for preventing underfitting and overfitting in the model. After training the model, the algorithm predicts the number of homicides in cities with an accuracy of up to 97% of the variance explained [12]

III. MATERIALS AND METHODS

A. *Study Area*

India is divided into 28 states and 9 union territories with Karnataka being the 6th largest state area wise. The latest data available on National Crime Records Bureau show that out of the total rape cases in India which have been reported, four out of five rape victims belong to these 10 states - Haryana, Jharkhand, Odisha, Rajasthan, Uttar Pradesh, Madhya Pradesh, Maharashtra, Kerala, Assam and Delhi. The number of total reported rape cases in these 10 states has almost doubled in the last 10 years - from 12,772 in 2009 to 23,173 in 2019. The rest 26 states have reported almost the same numbers as they did in 2009 [13]. Karnataka, however, recorded one of the lowest crime rates in rape at 1.6 per lakh of population. Only Tamil Nadu and Bihar have a lower rate [14]. According to the Census 2011, Karnataka has 30 districts, 176 sub-districts (Taluks) and 29,340 villages and its total population is 6,10,95,297.

B. *Data Collection*

The statistics of rape, murder and robbery cases in Karnataka are obtained from the official website of Karnataka State Police [15]. Annual crime reports from 2015 to 2019 are studied and taken for analysis. The crime report of 2019 mentions 37 distinct divisions comprising districts and cities jurisdiction whereas crime reports of 2015-18 mentions 36 distinct divisions. In contrast, there are 30 districts mentioned in 2011 census, hence the districts and cities mentioned in the crime report numbering 36 and 37 have been distributed accordingly. Belgavi city and Belgavi district are grouped under singular division Belgaum, Hubli-Dharwad city and Dharwad are grouped under Dharwad, D.K and Mangaluru city are grouped under Dakshina Kannada, Mysuru district and Mysuru city are grouped under Mysore, Kolar and KGF are grouped under a singular division Kolar.

The determinants of crime are rural population percentage which indicates the percentage of rural population in Karnataka state, urban population percentage which indicates the percentage of urban population in Karnataka state, literacy rate which indicates the percentage of literate population across the state, sex ratio which signifies number of females for every 1000 male, scheduled caste (SC) percentage indicates the percentage of scheduled caste across the state, scheduled tribe (ST) percentage which indicates the percentage of scheduled caste across the state, marginal workers percentage which indicates percentage of people who have worked for less than 6 months in a year, main workers percentage which indicates percentage of people who have worked for more than 6 months in a year, population density which indicates number of people per unit area police density (area-wise) and police density (population-wise). The data is collected from the Census of India data published by Directorate of Census Operations, Karnataka [16]. Choropleth thematic maps are used to visualize the crime intensity in each district. While Figure 1 represents crime intensity of Murder only for the year 2019, in order to study the distribution of crime across different districts, crime intensity over a period of years must be visualized. An average number of rape cases from 2015-2019 for each district is tabulated and the same is mapped in the following figure. The same can be performed for other crimes as well.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

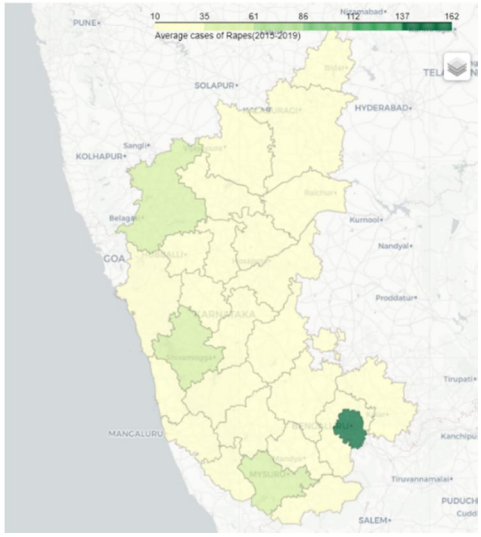


Figure 2: Choropleth map of Karnataka representing average Rape cases (2015 – 2019) in different districts

Bangalore city has the highest average of 162.4 cases followed by Mysuru at 43.2 cases per year. Yadgir district has the lowest average of 8.8 cases per year.

C. Models Used

The relationship between the dependent variables and the independent variables may be linear or non-linear. In order to capture the both linear and nonlinear relationship between the data, 3 different models namely Multi Linear Regression, Random Forest Regressor and Generalized Linear Model are employed. The Multi Linear Regression captures the linear relationship in the data and the Random Forest Regressor and Generalized Linear model captures the non-linear Relationship in the data

1) *Multi Linear Regression:* The Simple Linear Regression model is a simple model which works on one independent variable and only one dependent variable. For more than one independent variables, the model will become Multiple Linear Regression Model. Thus, Multiple Linear Regression Model is used to predict the crime status as are more than one parameters affecting crime in negative way or positive way are being considered. The Regression Model provides enough information about how independent or input variables X are affecting the output or dependent variable Y. The Linear Regression equation can be expressed as

$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + \theta_4 X_4 + \theta_5 X_5 \ldots$ eqn. 1

In the analysis, different parameters have been taken which affect the crime wherein dependent variable (y) will be the number of crimes and the independent variables (x) will be Rural Population Percentage, Urban Population Percentage, Literacy Rate, Sex Ratio, Scheduled Caste (SC) Population Percentage, Scheduled Tribe (ST) Population Percentage, Marginal Workers Percentage, Main Workers Percentage, Population Density, Police Station Density Area Wise, Police Station Density Population Wise. Here, the data is first trained with all the variables and after that the statistics of the model are taken. The statistics will include the following:

a) R squared: It gives the degree of closeness between the actual and predicted values. More The R square value the best fit it is
b) Adjusted R squared: The Adjusted R-squared takes into the account the number of independent variables used for predicting the target variable. By doing this we can determine whether adding the new variables to the model actually increases the model fit
c) P value: In statistics p value is the probability of obtaining the results at least as extreme as the observed results. The smaller the p value the stronger evidence
d) T value: The t-value measures the size of the difference relative to the variation in the sample data

The Multi linear Regression model can be constructed by the following approaches:

a) *Top-Down Approach:* In the top-down approach the model is first built and the available statistics are noted. Here the variables are recursively removed based on the above statistics criteria and the model building is stopped after there is no significant increase in the R squared and both the R-squared and the adjusted R-squared values are close.

b) *Bottom-Up Approach:* In this approach, model building is start with single variables at first and then, each variable is added and then statistics are checked to see if there is a significant difference between the R-squared and adjusted R-squared then a new variable is added. Also, after the new variable is added, the model is checked with the p-value and t-statistic.

2) *Random Forest Regression:* Random Forest regression is a supervised Learning model that uses ensemble method for regression. A random forest regression tree operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees. The model is built by the process called hyper parameter tuning, parameters which define the model architecture is defined as hyper parameter tuning. Here a set of probable values are taken for each variable and the model is trained with these parameters. The number of iterations depends on the number of hyper parameters and the range of variables. More the domain of the variables the model becomes more complex. To reduce the overfitting in the Random Forest AdaBoost (Adaptive Boosting) and Gradient Boosting can be used.

The random forest regression has the following parameters
n_estimators = number of trees in the random forest
max_feature = the maximum number of features considered to separate a node
max_depth = maximum number of levels in each decision tree
min_samples_split = min the number of data points placed in a node before the node was split
bootstrap = method of sample data points (with or without modification)
random_state = type of sample distribution with replacement or without replacement

By changing these hyper parameters and checking the accuracy, a model with good accuracy can be obtained

3) *Generalized Linear Model:* The relationship between the dependent variables and the independent variables can be non-linear. In non-linear regression there are many models which can fit the model. In the generalized linear model, the models are checked with the available non-linear functions.

In the Generalized linear model function available in Python, the following distributions can be tried:

a) Binomial
b) Gaussian
c) Poisson
d) Gamma
e) Inverse Gaussian

Here the prominent variables are

- *P-value:* n statistics p value is the probability of obtaining the results at least as extreme as the observed results. The smaller the p value the stronger evidence
- *Z-value:* It gives how many standard deviation units the point is away from the mean. The close the value the best approximate it is
- *Chi-square:* a chi square test is used to test the comparison between the two variables. Larger value means there is significant relation between variables

The Generalized linear model can be constructed by the following approaches

- *Top-Down Approach:* In the top-down approach, the model is first built and the available statistics are noted. Here the variables are recursively removed based on the above statistics criteria and the model building is stopped after there is z-value is too small
- *Bottom-Up Approach:* In this approach, it is started with single variables at the beginning and then each variable is added and checked if the z-value and p-value is small

In generalized linear model the model can be trained with different distribution and the results can be compared to obtain the best model

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

IV. RESULTS AND DISCUSSIONS

For the given crime all the three machine learning models are trained namely the Multi Linear regression, Generalized Linear model and Random Forest regression. The following are the determinants of crime:

RURPO	Rural population percentage
URBPOP	Urban population percentage
LITRATE	Literacy rate
SEXRATIO	Sex Ratio
SCPOP	Scheduled Caste (SC) population percentage
STPOP	Scheduled Tribe (ST) population percentage
MARWORK	Marginal workers percentage
MAINWORK	Main workers percentage
POPDEN	Population density
POLDENAREA	Police density area wise
POLDENPOP	Police density population wise

A. Multiple Linear Regression Model for Rape

The model is trained with the Top-Down approach after the removal of the variables which don't meet the given criteria. The following table is obtained:

Covariates	Coefficient	p-value
RURPO	1.2125	0.039
URBPOP	1.2396	0.032
STPOP	0.2545	0.056
MARWORK	-1.1415	0.057
MAINWORK	-1.0162	0.073
POPDEN	0.2527	0.014

After running four iterations with removing variables the table is obtained. From the table it is clear that by the multi linear regression: count of rural population, count of urban population, ST percentage, marginal workers in each district, number of main workers in each district and density of the population are the prominent factors

The multiple linear regression equation looks like

Y (No of Rape cases)

=c+(1.215) RURPO+(1.2396)URBPOP+(0.2545)STPOP+(-1.415)*MARWORK+(-1.0162)*MAINWORK+(0.2527)*(POPDEN)

B. Random Forest Regression for Rape

After training the random forest regressor with the following hyper parameters:

1) The n_estimators can take values from 1 to 10
2) The max_features takes value either auto or sqrt
3) The max_depth can take values from 1 to 20
4) Min_sample_split can take values from 1 to 20
5) Boot strap takes value either True or False
6) Random State can take values either 0 or 1

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

After checking with all the possible hyperparameters the following parameters got the good accuracy

Parameter	Domain
n_estimators	6
Max_features	auto
Max_depth	10
Min_samples_split	5
bootstrap	False
Random_State	0

C. Generalized Linear Model for Rape

After the model is trained with the top-down approach and with the different distributions the following parameters are obtained

Covariates	Coefficient	p-value
RURPO	0.0670	0.000
URBPOP	0.0573	0.000
LITRATE	0.0293	0.005
SCPOP	0.0260	0.031
STPOP	0.0455	0.000
MARWORK	-0.0924	0.001
MAINWORK	-0.073	0.000
POPDEN	0.0014	0.042

From the table it is clear that by the generalized linear model count of Rural Population, count of urban population, Literacy rate, SC percentage, ST percentage ,Marginal workers in each district, number of main workers in each district and Density of the population are the prominent factors.

D. Multple Linear Regression Model for Rape

The model is trained with the Top-Down approach after the removal of the variables which don't meet the given criteria. The following table is obtained

Covariates	Coefficient	p-value
RURPO	2.6734	0.001
URBPOP	2.4662	0.003
SEXRATIO	-0.2356	0.002
STPOP	-1.0061	0.027
POPDEN	0.0661	0.042

From the table it is clear that by the multi linear regression count of Rural Population, count of urban population, Sex ratio,ST percentage and Density of the population. are the prominent factors.

The multiple linear regression equation for the murder is

Y(No of Murder cases)=(2.6734)*RURPO+(2.4662)*URBPOP+(-0.2356)*SEXRATIO+(-1.0061)*STPOP+(0.0661)*POPDEN

E. Random Forest Regression for Murder

After training the random forest regressor with the following hyper parameters

1) The n_estimators can take values from 1 to 10
2) The max_features takes value either auto or sqrt
3) The max_depth can take values from 1 to 20
4) Min_smaple_split can take values from 1 to 20
5) Boot strap takes value either True or False
6) Random_State can take values either 0 or 1

After checking with all the possible hyperparameters the following parameters got the good accuracy

Parameter	Domain
n_estimators	5
Max_features	quto
Max_depth	10
Min_samples_split	8
bootstrap	True

F. Generalized Linear Model for Murder

After the model is trained with the top-down approach and with the different distributions the following parameters are obtained

Covariates	coefficient	p-value
RURPO	0.0631	0.000
URBPOP	0.0629	0.000
SEXRATIO	-0.0035	0.000
SCPOP	0.0240	0.000
STPOP	-0.0327	0.000
POPDEN	0.0017	0.000

From the table it is clear that by the multi linear regression count of Rural Population, count of urban population, Sex Ratio, SC percentages percentage, Density of the population are the prominent factors

G. Multiple Linear Regression Model for Robbery

The model is trained with the Top-Down approach after the removal of the variables which doesn't met the given criteria the following table is obtained

covariates	Coefficient	p-value
RURPO	1.466	0.013
URBPOP	2.249	0.037
SCPOP	1.365	0.012
STPOP	-1.571	0.059
MAINWORK	-2.027	0.052
POPDEN	35.995	0.029

From the table it is clear that by the multi linear regression count of Rural Population, count of urban population, SC percentage ST percentage, number of main workers in each district and Density of the population are the prominent factors

The multiple Regression equation for the robbery look like
Y(No of Robbery Cases)=(1.466)*RURPO+(2.249)*URBPOP+(1.365)*SCPOP+(-1.571)*STPOP+(-2.037)*(MAINWORK)+35.995(POPDEN)

H. Random Forest Regression for Robbery

After training the random forest regressor with the following hyper parameters

1) The n_estimators can take values from 1 to 10
2) The max_features takes value either auto or sqrt
3) The max_depth can take values from 1 to 20
4) Min_smaple_split can take values from 1 to 20
5) Boot strap takes value either True or False
6) Random_State can take values either 0 or 1

After checking with all the possible hyperparameters the following parameters got the good accuracy

Parameter	Domain
n_estimators	10
Max_features	sqrt
Max_depth	6
Min_samples_split	8
bootstrap	False
Random_state	0

I. Generalized Linear Model for Robbery

After the model is trained with the top-down approach and with the different distributions the following parameters are obtained

covariates	Coefficient	p-value
RURPO	0.0350	0.000
URBPOP	0.0547	0.000
LITRATE	0.0224	0.001
SCPOP	0.0497	0.000
STPOP	-0.0436	0.000
MAINWORK	-0.0479	0.000
POPDEN	0.0024	0.000
POLDENPOP	0.1170	0.000

From the table it is clear that by the multi linear regression count of Rural Population, count of urban population, Literacy Rate, SC percentage, ST percentage, Main workers in each district, Marginal workers in each district, number of main workers in each district, Density of the population. Police station density per population are the prominent factors

J. Comparison Between Models

The following table drawn from taking the 5 samples randomly and testing each algorithm with same data the metric is the square root of the distance

	Multi linear regression	Random Forest Regression	Generalized linear model
Rape	13.03	4.675	5.77
Murder	23	5.23	23.7
Robbery	17.2	3.43	7.23

For the rape crime RFR and GLM has shown a good performance. Next for the murder crime RFR has shown a good performance and the MLR and GLM has similar error

For the robbery the RFR did best and the GLM has shown moderate performance

V. CONCLUSION

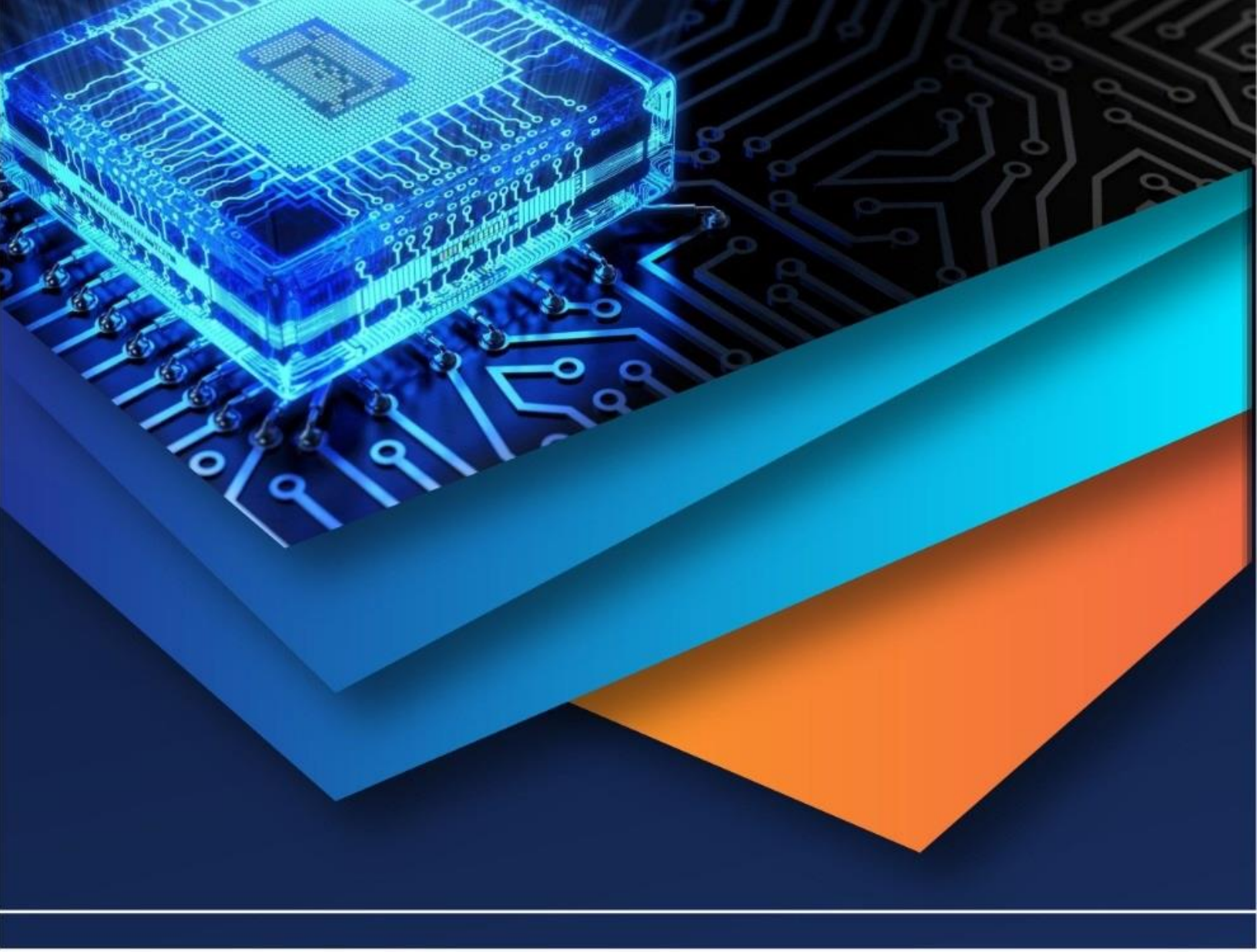
The research work aims to have a significant impact and benefit the safety and well-being of every citizen of Karnataka, and will provide for a pathway to adapt similar research for the other states in the country and prove as an effective measure to curb crime in the respective states and the country at large. This research provides a mechanism for the maintenance of law and order by helping the concerned Government departments in identification of various districts in the state which have a high rate of crime using the visualizations. It also provides for the police department to predict crime that could take place by using the predictive analysis technique which would enable the cops to plan and prepare well in advance by taking reasonable steps to prevent the crimes such as increasing patrols in the districts, installation of CCTV cameras at location where the occurrence of crime is high. Since the majority of crime takes place due to lack of education where the literacy rate is low, the police department can take steps to create awareness among people where the literacy rate is low.

Further, to narrow down and identify important covariates and factors associated with increasing crime, the following can be undertaken in future:

1) *Disintegration of Research:* Contract the research area into smaller sub-divisions and regions and identify local factors associated in that region. This will prove to be an effective strategy as analysing a smaller area will also introduce new local factors which have influence on the increasing crime rate. Analysing the demography of an area will also help in understanding neighbouring regions' properties
2) *Analysis Of Low Crime Incidence Regions:* Study the data and factors associated with crime rate in areas and regions which have a considerably low crime rate. Analyse these regions and incorporate results in other districts with high crime rate.

REFERENCES

[1] Arthur, J.A. (1991) 'Socio-economic predictors of crime in rural Georgia'. Criminal Justice Review, 16: 29-41.
[2] Becker, G.S., (1968) 'Crime and Punishment: An Economic Approach', Journal of Political Economy, 76 (2): 169-217
[3] Drèze, Jean, and Reetika Khera. "Crime, Gender, and Society in India: Insights from Homicide Data." Population and Development Review, vol. 26, no. 2, 2000, pp. 335–352. JSTOR, www.jstor.org/stable/172520. Accessed 22 June 2021.
[4] https://ncrb.gov.in/
[5] Shaw, Clifford R. and McKay, Henry D. (1942) Juvenile Delinquency and Urban Areas. The University of Chicago Press,
[6] Fleisher, B., (1963) 'The Effect of Unemployment on Juvenile Delinquency'. Journal of Political Economy, 71(6): 543-555
[7] Dubey and Aggarwal (2015) Crime, Crime Rates and Control Techniques: A Statistical Analysis, lawoctopus.com
[8] Becker, G.S., (1968) 'Crime and Punishment: An Economic Approach', Journal of Political Economy Bourguignon, Ehess And Delta (1999), Crime, Violence and Inequitable Development, Annual World Bank Conference on Development Economics, Washington, D.C., April 28-30, 1999.
[9] Dutta, M. And Husain, Z. (2009). Determinants of crime rates: Crime Deterrence and Growth in post liberalized India.
[10] Shingleton et al., "Crime Trend Prediction Using Regression Models for Salinas, California", Dissertations and Theses , Naval Postgraduate School, Monterey, California, 2012.
[11] Uddin, & O., Osemengbe & Uddin, Osemengbe. (2014). Data Mining: An Active Solution for Crime Investigation.
[12] Alves, Luiz & Valentin Ribeiro, Haroldo & Rodrigues, Francisco. (2017). Crime prediction through urban metrics and statistical learning.
[13] Rai, Dipu. "India's 10 most dangerous states for women". India Today, 10th October, 2020
[14] Bharadwaj, K.V Aditya. "Karnataka registers the highest number of sedition cases in the country". The Hindu. October 1, 2020.
[15] Karnataka State Police www.ksp.gov.in.
[16] Census Of India 2011 Primary Census Abstract Data Highlights Karnataka Series 30. Data Product Code : 29-007-2011-PCA Data Highlights
[17] Hyper Parameter Tuning the random forest regression in python www.towardsdatascience.com
[18] What are T-value and P-value in Statistics? blog.minitab.com
[19] Generalized Linear Models introduction to advanced statistical models www.towardsdatascience.com



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY