IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114100

Prediction on the Combine Effect of Population, Education and Unemployment on Criminal Activity Using Machine Learning

Soumayadip Saha[1], Joyitree Mondal[2], Arnam Ghosh[3], Mrs. Sulekha Das[4], Dr. Avijit Kumar Chaudhuri[5]

[1] UG-Computer Science and Engineering, Techno Engineering College Banipur

[2] UG-Computer Science and Engineering, Techno Engineering College Banipur

[3] UG-Computer Science and Engineering, Techno Engineering College Banipur

[4] Assistant Professor, Techno Engineering College Banipur

[5] Assistant Professor, Techno Engineering College Banipur

Abstract: Criminology is a complex subject. There are various factors which affect rate of crime in a particular society. Few of the main aspects are population, unemployment and education. In this article we would explore the correlation of rate of violent crimes in areas where majority of the inhabitants are afro Americans with the aforesaid aspect. Following are the effects in brief of mentioned reasons influencing rate of crime. Unemployment is one of the multiple causes contributing higher rate of crimes in contemporary society. Relation of crime with unemployment is manifested well through the offensive conducted by the unemployed. At a time when there is a death of suitable jobs or opportunities with crime is undermine by the promise of expect rewards in the mind of an individual. Good education is essential for development of professional skill / abilities in individuals. Members of community with higher level of education can avail rewarding jobs resulting in higher gross capita income. Consequently, risk associated with crime outweighs lure of benefits. Apart from the above good education tunes psychological, moral and social upbringing of an individual and thereby contributes the most to make her a better person. A person with rewarding jobs and high moral value would most likely endeavor to avoid activities which is not ethical and/or legal in the eyes of law. Rapid population growth, if not properly managed can have negative impact on crime rate. This is especially evident in areas such as those inhibited by impoverished afro American population. Reason could be deteriorated standard of living as available resources become scarce. Unfavorable living and economic condition can be conducive environment for criminal activities.

Keywords : Multiple linear Regression, Cross Validation, Confusion Matrix

1. INTRODUCTION

Statistics is one of the most widely uses aspects which is use in science, medicine, crime rate, fisheries etc. There are many methods in statistics but one of the most useful methods is regression. Regression is basically six types which are simple linear regression, multiple linear regression, logistic regression, ordinal regression, multinomial regression and discriminate analysis. In this project we basically used multiple linear regressions for analysis the crime. In multiple linear regressions (MLR) there is one dependent variable where independent variable can be two or more than two. The main objective of this project to find the suitable model and the reason behind the crime and how the crime increases for some reason. First, we find the dependent variable and independent variables. After choosing the dependent and independent variables we create the formula and did training and tasting over the data. After that we did cross validation and confusion matrix to find the more correct rate of crime. Training and testing data was for 2/3 data, 1/2 data and 1/4 data and the cross validation was for 10-fold cross validation means we divide the dataset into 10 sub-dataset and then we find the crime rate for each 10-fold.

Crime – this word is just like a destruction of entire humanity and way of development in today's world. The actual definition of crime is a vast number of hardships and complexities and it's like a toxic which spoils the growth of nation. It is not universally accepted and it's socially built and altered reality. Simply crime can be defined as a criminal offence against any person or an organization with an intent to harm them directly or indirectly that is illegal and punishable under the country law. Crimes like robberies, looting, sexual harassment, killing people or try to kill anyone

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114100

are one of the major crimes that is happening in our day-to-day life. So, in this project we are going to analysis crime using multiple regression.

Based on crime dataset many works have been done already that's why we used some different methods for analysis the crime rate. Our main motive was to predict the crime value properly and accurately. In our daily life we can see every day many crimes like robbery, murder, sexual harassment etc. are happening that's why we tried to find the actual rate of crime which will help the police and government to know about crime rate and they can reduce the crime rate by reducing the crimes actual reason which we mentioned in this project.

2. LITERATURE REVIEW

Machine learning is a simple way which helps to predict the data easily. Machine Learning can be favorable for predicting the crime accuracy. In this paper a dataset has chosen which is http://archive.ics.uci.edu/ml/datasets/communities+and+crime to predict the crime accuracy in some different state based on those states- population, unemploy and education conditions. Based on crime analysis many projects have been done yet and, in those projects, some other fields and different process for finding the accuracy of crime has been used like- I - JEN), Nazlena Mohamad Ali et al. [2] describe the approach, user reviews and the approach envisaged. Their main goals were to devise crime - related events, to examine the use of crime - related events for enhancing classification and to create a customizable news recuperation system for crime. Sutapat Thiprungsri [3] analyze the probability of using auditing clustering technology. She investigates the application of the accounting cluster analysis, especially in audit divergence. The aim of the study is to examine the use of fraud filtering technology during an audit. De Bruin et al. [4] established a framework for patterns in crime that uses a new distance measure to compare and cluster all individuals based on their profiles. B. Swadi Al – Janabi [5] introduces a proposal for the analysis and detection of crime and criminal data using data classification algorithms and K Means data clustering algorithm. Manish gupta et al. Using MLR some other works has been done like- MLR used by Ofuolu *et al.* (2007) in their research to determine of adoption of improved fish production technologies among fish farmers in [6]. MLR that applied in the research by Khamis *et al.* (2003) states that the higher of R2 gives good result on model fitting [7]. Combination method of Pearson correlations, multiple and simple linear regression and ANOVA was used by Sain (2006) to see if there was change in measured habitat and fish metrics occurred in relation to increased urbanization [8]. Sain (2006) used multiple and linear regression to explain the strongest relationship between fish and habitat [9].

3. METHODOLOGY

3.1) Data: - In this paper data were taken from UCI Machine Learning Repository. The data combines socio-economics data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR, http://archive.ics.uci.edu/ml/datasets/communities+and+crime. In this dataset there are total 128 column. Out of these 5 columns are- 'State', 'Population', 'PctBSorMore', 'Unemploy' and 'Violent Crime'. "Violent Crime" tells us the crime rate per 100k population.

Basically, crime is depended on the others 4 column like "State", "Population", "PctBSorMore", and "Unemploy". In every state there are different economical condition different rate of population and based on economical condition in every state education rate, unemploy rate are different that's why in every state there are different crime rate per 100k people. Population and unemploy are also a great reason for increasing the crime. Because if population will increase everybody will not get the same education and if there is no job and all will become unemploy then the crime rate will automatically increase for surviving.

3.2 Table 1-For attributes mean and standard deviation

Attributes	Definition	Mean	Standard deviation
State	US state	28.68	16.39
Population	population for community	0.06	0.13
PctBSorMore	percentage of people 25 and over with a bachelor's degree or higher education	0.36	0.21
Unemployed	percentage of people 16 and over, in the labor force, and unemployed	0.36	0.2
ViolentCrimesPerPop	total number of violent crimes per 100K population	0.24	0.23

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114100

4. RESEARCH METHOD

Multiple linear regression are the method of statistics in regression that used to analyse the relationship between single response variable (dependent variable) with two or more controlled variables (independent variables). This method was selected for this research because there were more than controlled variables. In this research, response variable is violent crime (Y) while state (X1), Population (X2), education (X3), Unemployed (X4), are controlled variables.

4.1 Accuracy of Difference between Actual Data and Calculated Data-

We check the accuracy for 2/3, 1/2 and 1/4 data.
In this research, the hypotheses that used:
H0: b_1=b_2=b_3=b_4=0
Ha: At least one of the b_1, b_2, b_3 and b_4 does not equal to 0
which says that
H0: None of the controlled variable X_1, X_2, $X_3$and X_4 is significantly related to Y
Ha: At least one of the controlled variables X_1, X_2, $X_3$and X_4 is significantly related to Y
The model of multiple regression can be represented as:

$$Y = a+b_1X_1 + b_2X_2 + \ldots\ldots\ldots\ldots\ldots\ldots + b_nX_n$$

And the a and b formula are

$$b_i = \sum_{i=1}^{n} \frac{(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

$$a = \frac{(\sum Y \times \sum X^2)-(\sum X \times \sum XY)}{(n\times\sum X^2)-(\sum X\times\sum X)}$$

where,
y=crime
a=Constant variable
b_1=Coefficient of first control variable, X_1
b_2=Coefficient of second control variable, X_2
b_3=Coefficient of third control variable, X_3
b_4=Coefficient of forth control variable, X_4
X_1=controlled variable(state)
X_2=controlled variable(population)
X_3=controlled variable(education)
X_4=controlled variable(unemploy)

4.2 Confusion-Matrix

After finding the accuracy of difference between actual data and calculated data we did the Confusion Matrix. In this Confusion Matrix process first, we find the **TP** – which stands for "**TRUE POSITIVE**" means accuracy of classified positive data, **TN** – which stands for "**TRUE NEGATIVE**" means actual value is positive but predicted data is negative, **FP** – which stands for "**FALSE POSITIVE**", means which remark that actual value is negative but predicted data is positive, **FN** – which stands for "**FALSE NEGATIVE**" means which remark that actual data and the predicted data both are negative. After that we find the accuracy, sensitivity and specificity,
Where-
Accuracy: -
It's the ratio of the correctly labeled subjects to the whole pool of subjects.
Accuracy is the most intuitive one [10].

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100$$

Sensitivity: -
Sensitivity is the basically how sure we are that we didn't miss any positive data means the total % of positive value.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \times 100$$

Specificity: -
Specificity is basically the all true negative means we don't want any false positives.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114100

Specificity = $\frac{TN}{TN+FP} \times 100$

After finding the accuracy, sensitivity and specificity we find the and standard deviation, R^2, F1 score, kappa test, and the formula of and standard deviation, R^2, F1 score, kappa test is in the below-

Standard deviation: -

Standard deviation is a number that describes how spread out the values is. A low standard deviation means that most of the numbers are close to the mean (average) value. A high standard deviation means that the values are spread out over a wider rang [13].

Formula of Standard Deviation is: -

$$\text{RMSE}=\sqrt{\text{MSE}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

R_Square: -

The R2 score is a very important metric that is used to evaluate the performance of a regression-based machine learning model. It is pronounced as R squared and is also known as the coefficient of determination. It works by measuring the amount of variance in the predictions explained by the dataset [14].

$$R^2=1-\frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2}$$

Where,

$\hat{y}$ = Predicted value of Y

$\overline{Y}$ = mean Value of Y

F1 score: -

The F1 score is defined as the harmonic mean of precision and recall. As a short reminder, the harmonic mean is an alternative metric for the more common arithmetic mean. It is often useful when computing an average rate. In the F1 score, we compute the average of precision and recall [15]. For, finding the F1 score we have to find first precession and recall

$$\text{Precession} = \frac{\text{TP}}{\text{TP+FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

$$\text{F1 Score} = \frac{2*Recall*precession}{Recall+Precession}$$

Kappa Test: -

In essence, the kappa statistic is a measure of how closely the instances classified by the machine learning classifier matched the data labeled as ground truth, controlling for the accuracy of a random classifier as measured by the expected accuracy [16]. For finding kappa test first we have to find the Observed Agreement, Expected Agreement

$$\text{Observed Agreement} = \%\ (\text{Overall Accuracy})$$

$$\text{Expected Agreement} = \frac{(TP+FP)*(TP+FN)*(FN+TN)*(FP+TN)}{100}$$

$$\text{Kappa Test} = \frac{Observed\ Agreement - Expected\ Agreement}{100 - Expected\ Agreement}$$

4.2 Cross-Validation

After finding the accuracy of difference between actual data and calculated data and Confusion Matrix we did cross validation. In this cross-validation process first we divide the whole dataset into 10 sub-dataset and then we find the Confusion Matrix, Accuracy, Sensitivity, Specificity, Standard Deviation, R_Square, F1 Score, Kappa test of 10 sub-dataset elements randomly and differently.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

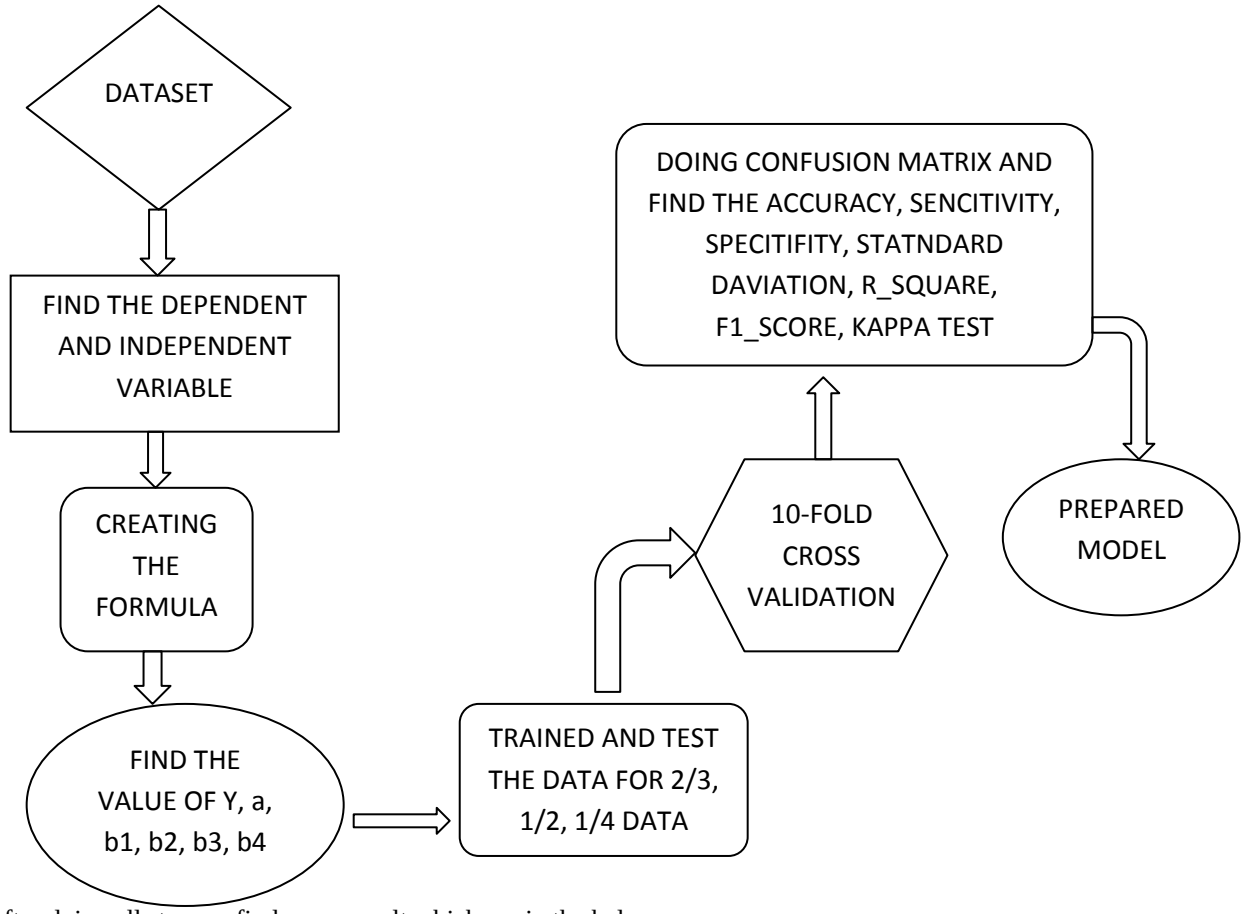
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114100

5. RESULT

In this paper whatever we used based on that we made a flowchart of that which is given in the below: -



After doing all step we find some result which are in the below

5.1 Accuracy of difference between actual data and calculated data

In this method we took training data and test data for 2/3, 1/2 and 1/4 value.

1)	Accuracy of 2/3 data	83.13
2)	Accuracy of 1/2 data	95.98
3)	Accuracy of 1/4 data	98.39

5.2 Confusion-matrix: -

For 2/3 Data-

Confusion Matrix	552 73 28 11
Accuracy	94.13
Sensitivity	98.05
Specificity	72.28
Standard Deviation	0.15
R_SQUARE	0.26
Precession	0.95
Recall	0.98
F1 Score	0.97
Kappa Test	1.0

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114100

For 1/2 data –

Confusion Matrix	956 28 11 1
Accuracy	98.8
Sensitivity	99.9
Specificity	71.79
Standard Deviation	0.21
R_SQUARE	0.08
Precession	0.99
Recall	1.0
F1 Score	0.99
Kappa Test	1.0

For 1/4 Data-

Confusion Matrix	1471 21 3 0
Accuracy	99.8
Sensitivity	100.0
Specificity	87.5
Standard Deviation	0.43
R_SQUARE	0.29
Precession	1.0
Recall	1.0
F1 Score	1.0
Kappa Test	1.0

5.3 For 10-fold cross-validation

Test Cases	Accuracy Rate	Sensitivity	Specificity	Standard Deviation	R_SQUARE	Precession	F1 Score	Kappa Test
01	98.99	100.0	84.62	0.16	0.22	0.99	0.99	1.0
02	97.49	99.47	66.67	0.16	0.23	0.98	0.99	1.0
03	98.49	100.0	72.73	0.16	0.24	0.98	0.99	1.0
04	97.99	99.46	76.92	0.16	0.23	0.98	0.99	1.0
05	96.98	99.46	64.29	0.16	0.22	0.97	0.98	1.0
06	97.99	98.92	84.62	0.16	0.24	0.99	0.99	1.0
07	96.48	99.46	53.85	0.16	0.22	0.97	0.98	1.0
08	98.49	100.0	80.0	0.15	0.24	0.98	0.99	1.0
09	96.98	100.0	62.5	0.16	0.23	0.97	0.98	1.0
10	98.99	100.0	84.62	0.16	0.22	0.99	0.99	1.0

CONCLUSIONS:

This paper using multiple regressions (MLR) to predict the crime level. We have collected the data from UCI Machine Learning Repository based on that we made a relationship between the dependent variable and the independent variable after that we perform cross validation for more accuracy. After checking the cross validation, we move to the Confusion matrix where we compare the actual target values with those predicted by the machine learning model. Using these models, we predict the accuracy as well as sensitivity and specificity for 2/3th, 1/2th, 1/4th set of data.

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114100

As per researchers we can see that how the crime is varied for whole world and humanity. So, in this project we defined some actual reason for increasing crime. We basically did this project because day-by-day this crime rate is increasing. It will help the any countries government to analyze their countries crime rate and they can also reduce the reasons behind increasing crime.

REFERENCES

Before doing that research paper some research is necessary so after analysis some paper some different concept is create and then a project can do so whatever papers are analyzed before doing this project those papers link are given in below

1. https://www.researchgate.net/profile/Natarajan-Meghanathan/publication/275220711_Using_Machine_Learning_Algorithms_to_Analyze_Crime_Data/links/571dc8ae08ae408367be5de8/Using-Machine-Learning-Algorithms-to-Analyze-Crime-Data.pdf
2. https://ieeexplore.ieee.org/abstract/document/8614828/
3. https://iopscience.iop.org/article/10.1088/1742-6596/1000/1/012046/meta
4. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3656953
5. Banfield, R. E., Hall, L. O., Bowyer, K. W., Bhadoria, D., Kegelmeyer, W. P., & Eschrich, S. (2004, June). A comparison of ensemble creation techniques. In International Workshop on Multiple Classifier Systems (pp. 223-232). Springer, Berlin, Heidelberg.
6. Ben-Shakhar, G., Lieblich, I., & Bar-Hillel, M. (1982). An evaluation of polygraphers' judgments: A review from a decision theoretic perspective. Journal of Applied Psychology, 67(6), 701.
7. Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.
8. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
9. Darmon, M., Vincent, F., Dellamonica, J., Schortgen, F., Gonzalez, F., Das, V., ... & Schlemmer, B. (2011). Diagnostic performance of fractional
10. excretion of urea in the evaluation of critically ill patients with acute kidney injury: a multicenter cohort study. Critical care, 15(4), 1-8.
11. Daubin, C., Quentin, C., Allouche, S., Etard, O., Gaillard, C., Seguin, A., ... & Du Cheyron, D. (2011). Serum neuron-specific enolase as predictor
12. of outcome in comatose cardiac-arrest survivors: a prospective cohort study. BMC cardiovascular disorders, 11(1), 1-13.
13. Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian journal of internal
14. medicine, 4(2), 627.
15. Hanley, J. A. (1989). Receiver operating characteristic (ROC) methodology: the state of the art. Crit Rev Diagn Imaging, 29(3), 307-335.
16. Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1),

