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 3, March 2022

DOI: 10.17148/IJARCCE.2022.11389

Prediction on requirement of police recruitment on the basis of community population and rate of violent crimes where Colored (Afro) Americans live

Anand Kumar Jha[1], Prithwish Raymahapatra[2], Nandini Ghosh[3], Sayantika Bose[4], Sulekha Das[5]

UG- Computer Science and Engineering , Techno Engineering College Banipur , Habra ,Kolkata[1-4]

Assistant Professor, Computer Science Engineering, Techno Engineering College Banipur, Habra, WestBengal[5]

Abstract : There is a general tendency that effective police force has a significant protective effect on violent crimes. The effect of police on crime operates through both a deterrence and an incapacitation way.The deterrence theory is based on the fact that an increase in capacity of police force enhances a typical offender's chance of being caught and thereby decreases crime particularly those which are violent in nature.The police departments actively focus their resources on the incapacitation of individuals posing the greatest risk to society, which make the incapacitation channel also an important factor. In this perspective adequate human resources (deployed police personnel) play a vital role. This article goes on to explore the relationship between population of a community of coloured (Afro-American) people, number police staff employed and violent crime occurrences in particular. The findings from this paper suggest that numbers of police recruited/ employed for maintaining law and order of a community/ area with coloured (Afro-American) inhabitation depend on population of that community/area and number of prevalent violent crimes. Our findings suggest that higher numbers of police not only reduce crime rates but also increase the share of crime, and in particular violent crime. This data will be analyzed using multiple linear regression in machine learning method using python programming language.

Keywords: Forecast; data Science, data analysis, Regression analyses; Stepwise multiple regression

1. INTRODUCTION:-

Statistics analysis is widely used in all aspects such as in business, science, medicine, fisheries, data science, real time data analysis and also in social sciences (Sarker et al., 2006). There are many methods in statistics through which we can perform this analysis and one of them is called regression. Using these methods in statistics, we can plan the production to check what the customer likes and wants, and you can also check the quality of the products far more efficiently with statistical methods. In fact, many business activities can be completed with statistics including deciding a new location, marketing the product, and estimating what will be the profit on the new product. Statistics basically have six branch such as linear regression, multiple linear regression, logistic regression, ordinal regression, multinominal regression and discriminant analysis.it is used to preliminary edit, used to detect glaring omissions and inaccuracies(often involved respondent follow up) and completeness, Legibility. One method that is categorized in the stepwise-type procedures is stepwise regression also used in this paper. The main objective of this paper is to predict the number of police required to control the rate of crime. Exploratory Data Analysis is an important aspect of any data science project. It forms the initial steps before moving into the Machine learning aspects. The proposed model is tested on the "Communities and Crime Data Set" from the UCI Machine Learning Repository.

In today's world crime rates are increasing dramatically and due to this reason the police requirement have also increased. In every country , there are some factors on which this police requirement are dependent on , those some factors can be population of that country and crime rate of that country. In this paper , we have used some statistical methods to predict the police requirement in different countries. There are many factors on which this police requirement depends on but in this paper we have used only two fields i.e. crime rate and population of that country.

Statistical methods like linear regression and multiple regression are widely used by researchers for the analysis of this database to predict the police requirement. And the researchers have also used Machine Learning procedure to do the same and have got the precise results on it.

We have chosen some factors like crime rates and population of a country because we thought that it can be the main factors on which the police requirement depends on and using the MLR and ML procedure we have seen that these factors relate directly to it and have also got very accurate and precise results.

We have searched some research papers and found that there are no such papers on these particular fields which we are working on. But we have found some papers using MLR and ML procedures to predict the crime rate. For example -

1. **Comparison of Machine Learning Algorithms for Predicting Crime Hotspots**

https://ieeexplore.ieee.org/abstract/document/9211482 ,This paper has used machine learning algorithms to predict the crime hotspots in different zones in China and has used some historical data for their work.

2. **Crime Analysis Using Machine Learning**

https://ieeexplore.ieee.org/abstract/document/8614828 , This paper investigates machine learning based crime prediction this paper has used the vancouver data of last 15 years for its analysis and has also used two different data processing methods like K-nearestneighbors and boosted decision tree to obtain a accuracy of 39% - 44% .

LITERATURE REVIEW:-

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use data as input to predict new output values.

Intan Martina Md Ghania and Sabri Ahmadb[5] used Multiple Regression Method to Forecast Fish Landing in their research paper. Using ML method, they got two models and between those models, model 2 is better. They also got the finest result.

H. C. Hamaker[6] described a brief description about the multiple regression. How multiple linear regression works and its formulas and also where it can be used.

Julie Barber and Simon Thompson[7] depicted the usefulness of generalised linear models (GLMs) for regression analyses of cost data and they use the ML method also.

Mr. M. S. BARTLETT [8] gave a detailed information about multiple linear regression in his paper named 'FURTHER ASPECTS OF THE THEORY OF MULTIPLE REGRESSION'.

J.T. Lin, D. Bhattacharyya and V. Kecman [9] used multiple linear regression in the field of composites machining. They are very much successful in that case.

After studying some more papers about the multiple linear regression, researchers have got a clear idea about the MLR method. That's why in this paper MLR method is used and full research paper is depending on this method.

METHODOLOGY:-

Data Collection:-

In this paper, data were taken from the "Communities and Crime Data Set" from the UCI
Machine Learning Repository: which is available at https://archive.ics.uci.edu/ml/datasets/communities+and+crime created by Michael
Redmond(redmond '@' lasalle.edu),Computer Science,La Salle University, Philadelphia, PA, 19141, USA -- culled from 1990 US Census, 1995 US FBI Uniform Crime Report, 1990 US Law Enforcement Management and Administrative Statistics Survey, available from ICPSR at U of Michigan.
-- Donor: Michael Redmond (redmond '@' lasalle.edu); Computer Science; La Salle University; Philadelphia, PA, 19141, USA -- Date: July 2009
The dataset which has been taken from this website contains a very huge amount of data, to be more precise a total of 128 columns are there. Out of those 128 columns we have taken only 3 columns for your research work. Those 3 columns are - "LemasTotReqPerPop" i.e. total requests for police per 100K population (numeric - decimal), "population" i.e. population for community: (numeric - decimal), and "Violent Crimes PerPop" i.e. total number of violent crimes per 100K population (numeric - decimal).
"LemasTotReqPerPop " depicts the total police requirement of a country and it is so the dependent variable of our MLR and ML process and this is a very important field because if the crime rate of a country increases then the police requirement will also increase, "population" this field depicts the total population of a country this can be said as the main

DOI: 10.17148/IJARCCE.2022.11389

field because almost everything of a country depends on population of a country if the population increase then along with it everything increases, "ViolentCrimesPerPop" this field depicts the violent crime rate of a country, this field is also a main factor for the increment in police requirement and recruitment process.

Table-1 : Attributes, Mean and Standard Deviation

ATTRIBUTE	DEFINITION	MEAN	STANDARD DEVIATION
LemasTotReqPerPop	total requests for police per 100K population.	0.03	0.09
population	population for community	0.06	0.12
ViolentCrimesPerPop	total number of violent crimes per 100K population	0.24	0.23

Research method:-

Multiple linear regressions (MLR) are the method of statistics in regression that is used to analyze the relationship between a single response variable (dependent variable) with two or more controlled variables (independent variables).
Our study comprises 3 steps: data collection, data processing, and model training. The goal of our study was to develop and evaluate a machine learning approach that has the best performance for predicting how many police are required to control the number of crimes in a specific population. We selected this method for this research because there were more than controlled variables. In this research 'police require' variable is (Y) while population variable is(X1), crime variable is(X2). Here 'police require' is a dependent variable and 'population'& 'crime' are independent variables.

Step 1: Checking assumptions:-

● Primary works with fields: The first step is to build a forecasting model by checking assumptions of data. Then we have chosen the dependent and independent fields, after that we have predicted the number of policies required to control the crime rate.

Then, MLR should have a linear relationship between dependent variable and independent variables which are "police required in the area(y)" as the dependent variable and "crime rate of that area(x2)" and "population of that area(x1)" as independent variables.

Thereafter we have taken ½ ,⅘, ⅔ amount of data from the full dataset as test data and the rest are as training data to check the accuracy.

● Cross Validation: Cross validation is a technique for evaluating ML models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data. Use cross-validation to detect overfitting, ie, failing to generalize a pattern. In this particular paper we have divided the original dataset into 10 subsets and each subset's data are being checked with the given dataset(we are taking 200 datas from one subset at a time and comparing them with the rest one thousand eight hundred datas) then we have got accuracy.

Confusion Matrix: A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing. The output "TN" stands for True Negative which shows the number of negative examples classified accurately. Similarly, "TP" stands for True Positive which indicates the number of positive examples classified accurately. The term "FP" shows False Positive value, i.e., the number of actual negative examples classified as positive; and "FN" means a False Negative value which is the number of actual positive examples classified as negative. One of the most commonly used matrix while performing classification is accuracy. The accuracy, sensitivity, specificity, precision, recall, and f1-score of a model (through a confusion matrix) is calculated using the given formula below.

DOI: 10.17148/IJARCCE.2022.11389

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

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

$$\text{Specificity} = \frac{TN}{TN+FP}$$

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

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

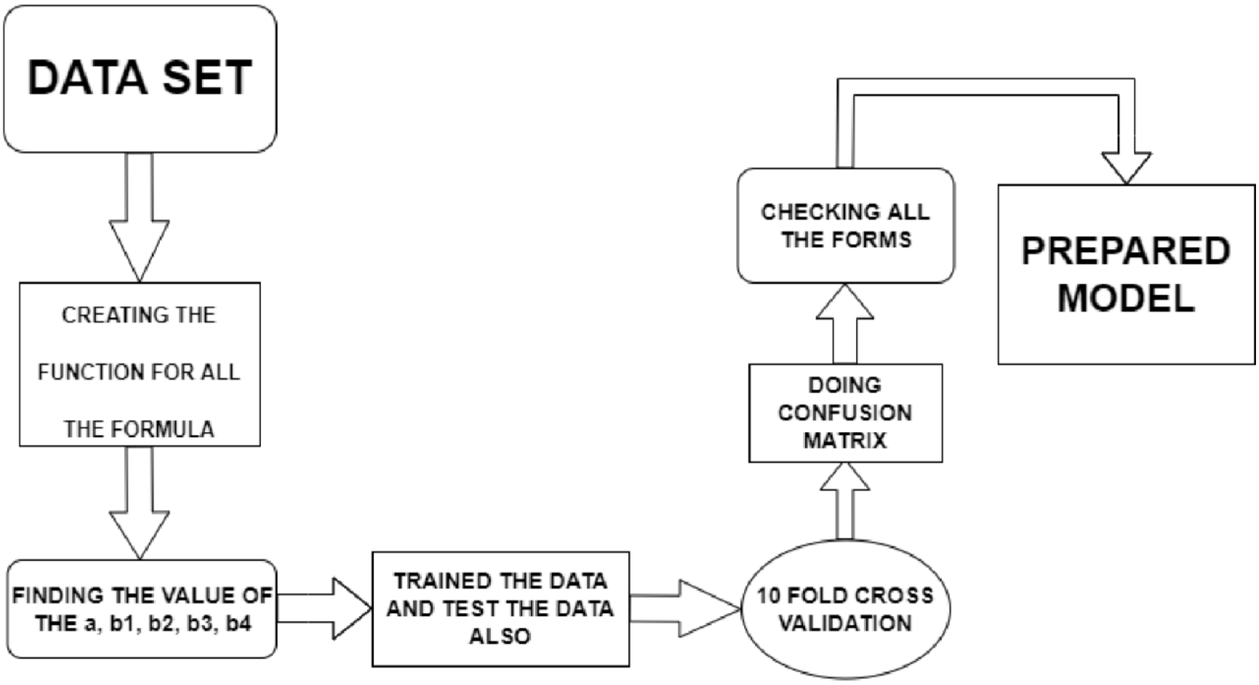
$$\text{F1\_Score} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

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

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

where- ŷ - Predicted value of y
ȳ - Mean value of y

FLOWCHART:


DATA SET
CREATING THE FUNCTION FOR ALL THE FORMULA
FINDING THE VALUE OF THE a, b1, b2, b3, b4
TRAINED THE DATA AND TEST THE DATA ALSO
10 FOLD CROSS VALIDATION
DOING CONFUSION MATRIX
CHECKING ALL THE FORMS
PREPARED MODEL

Step 2: Selecting suitable methods of multiple linear regression:

Multiple regression is a statistical technique that can be used to analyze the relationship between a single dependent variable and several independent variables. The objective of multiple regression analysis is to use the independent variables whose values are known to predict the value of the single dependent value.

Step 3: Interpreting the output:

From the confusion matrix output we can interpret the value of Matthews Correlation Coefficient(MCC), instead, is a more reliable statistical rate which produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (True Positive, False Negative, True Negative and False Positive).Proportionally both to the size of positive elements and the size of negative elements in the dataset.

A true positive is an outcome where the model correctly predicts the positive class. Similarly, a true negative is an outcome where the model correctly predicts the negative class.

A false positive is an outcome where the model incorrectly predicts the positive class. And a false negative is an outcome where the model incorrectly predicts the negative class.

- True Positive: the truth is positive, and the machine predicts that it is positive.
- True Negative: the truth is negative, and the machine predicts that it is negative
- False Negative: the truth is positive, but the machine predicts that it is negative
- False Positive: the truth is negative, but the machine predicts that it is positive

Step 4: Developing equation of multiple linear regression:

In this research, the hypotheses that used:

To predict the number of police required in order to minimize the crime using these formula we can easily test the actual value with the given value

❖ Y = a + b1X1 + b2X2 + ... + bnXn where a and b can be calculated using this formula:-

$$a=\frac{\sum y^{*}\sum x^{2}-\sum x^{*}\sum(x^{*}y)}{n\sum x^{2}-(\sum x)^{2}}$$

$$b_i=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^{2}}$$

Where -

Y = Police required list(Stands for the dependent variable) n = Amount of data taken for test

X1 =Population list(Stands for the independent variable) X2 = Crime list(Stands for the independent variable)

4. RESULTS AND DISCUSSIONS:-

After all the data was analyzed and calculated , we got the results as follows :

Table 1. 10 FOLD CROSS VALIDATION TEST CASES ACCURACY

TEST CASES Accuracy	Sensitivity	Specificity	Precision	F1 Score	Standard Deviation	R_Square	Mean of given data	Mean of predicted data
TC-1-> 98.49	100.0	80.0	98.39	99.19	0.057	0.113	0.057	0.090
TC-2-> 97.98	100.0	69.23	97.89	98.93	0.051	0.0045	0.056	0.085
TC-3-> 97.48	100.0	66.66	97.355	98.65	0.052	0.176	0.056	0.086
TC-4-> 97.98	100.0	71.42	97.88	98.93	0.055	0.1996	0.058	0.088
TC-5-> 97.98	100.0	69.23	97.89	98.93	0.054	0.1926	0.056	0.086
TC-6-> 98.99	100.0	84.61	98.93	99.46	0.046	0.2264	0.058	0.086
TC-7-> 97.98	100.0	75.0	97.86	98.91	0.057	0.0941	0.057	0.089
TC-8-> 97.98	100.0	66.66	97.90	98.94	0.052	0.1036	0.056	0.086
TC-9-> 97.48	100.0	61.53	97.38	98.67	0.052	0.2157	0.056	0.085
TC-10-> 97.48	100.0	68.75	97.34	98.65	0.053	0.1909	0.057	0.085

10 FOLD CROSS VALIDATION TEST CASES ACCURACY:-In this method we have one data set which we divide randomly into 10 parts. We have used 9/10 of those parts for training and reserve 1/10 for testing. We have repeated this process for 10 times and each time reserving a different tenth for testing.(1)

Table 2 : CONFUSION MATRIX AND ACCURACY,SENSITIVITY & SPECIFICITY

	½	2/3	4/5
CONFUSION MATRIX:	351 97	526 24	333
	513 35	98 16	42
ACCURACY (in %) :	86.75	93.98	94.22
SENSITIVITY (in %) :	90.93	97.05	97.37
SPECIFICITY (in %) :	84.1	80.33	75.0
PRECISION (in %) :	78.34	95.63	95.96
F1-SCORE (in %) :	84.17	96.33	96.66

According to Table 3, it shows Confusion matrix for each set of data such as ½ th & ⅔ th and ⅘ th
Data set along with that it shows 83.63% accuracy for ½ th of given data, 92.32% accuracy for ⅔ th of given data, 93.47% accuracy for ⅘ th of given data set and it also so 84.12% (Sensitivity for ½ th data set), 96.25%(Sensitivity for ⅔ th data set)& 96.08%(Sensitivity for ⅘ th data set).This Table also represented the Specificity for given set of data such as 83.45% for ½ th ,79.62% for ⅔ th & 80.3% for ⅘ th. (2)

5. CONCLUSIONS:-

This paper using multiple linear regressions (MLR) to predict the number of police required to minimize the number of crime.We have collected the data from different sources Based on that we made a relationship between a dependent variable with an independent variable after that we perform cross validation for more accuracy(we are taking 200 datas from one subset at a time and comparing them with the rest one thousand eight hundred datas) .After checking the cross validation we move to the Confusion matrix where we compares the actual target values with those predicted values with the help of machine learning model.Using these model we predict the accuracy as well as sensitivity and specificity for ½ th ,⅔ th,⅘ th set of data.

So, In this research paper we have predicted the police requirement of a country based on its crime rates and population and seen the results accordingly. Now coming back to the social context, if the number of police gets increased then the number of crimes will be decreased. With less number of crime the safety of common people will be ensured, the socio-economic stability will remain unchanged. And along with it the annual budget for law enforcement will also become less and govt. can use that budget in other important projects accordingly.

REFERENCE:-

1. Chaudhuri, A. K., Ray, A., Banerjee, D. K., & Das, A. (2021). An Enhanced Random Forest Model for Detecting Effects on Organs after Recovering from Dengue. methods, 8(8).
2. Chaudhuri, A. K., Sinha, D., Banerjee, D. K., & Das, A. (2021). A novel enhanced decision tree model for detecting chronic kidney disease. Network Modeling Analysis in Health Informatics and Bioinformatics, 10(1), 1-22.
3. Chaudhuri, A. K., Das, A., & Addy, M. (2020, February). Identifying the Association Rule to Determine the Possibilities of Cardio Vascular Diseases (CVD). In International Conference on Advanced Machine Learning Technologies and Applications (pp. 219-229). Springer, Singapore.
4. Chaudhuri, A. K., Das, A., Sinha, D., & Banerjee, D. K. (2021). Application of data mining techniques for avoiding underestimation of an event. Asian Journal For Convergence In Technology (AJCT) ISSN-2350-1146, 7(1), 179-189.
5. Addy, M., Chaudhuri, A. K., & Das, A. (2020, March). Role of Data Mining techniques and MCDM model in detection and severity monitoring to serve as precautionary methodologies against ‘Dengue’. In 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA) (pp. 16). IEEE.
6. Kim, S., Joshi, P., Kalsi, P. S., & Taheri, P. (2018, November). Crime analysis through machine learning. In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 415-420). IEEE.
7. Vaquero Barnadas, M. (2016). Machine learning applied to crime prediction (Bachelor's thesis, Universitat Politècnica de Catalunya).

DOI: 10.17148/IJARCCE.2022.11389

8. Zhang, X., Liu, L., Xiao, L., & Ji, J. (2020). Comparison of machine learning algorithms for predicting crime hotspots. IEEE Access, 8, 181302-181310.
9. Alves, L. G., Ribeiro, H. V., & Rodrigues, F. A. (2018). Crime prediction through urban metrics and statistical learning. Physica A: Statistical Mechanics and its Applications, 505, 435-443.
10. Zaidi, N. A. S., Mustapha, A., Mostafa, S. A., & Razali, M. N. (2019, September). A classification approach for crime prediction. In International Conference on Applied Computing to Support Industry: Innovation and Technology (pp. 68-78). Springer, Cham.
11. 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.
12. Tamilarasi, P., & Rani, R. U. (2020, March). Diagnosis of crime rate against women using k-fold cross validation through machine learning. In 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1034-1038). IEEE.
13. Babakura, A., Sulaiman, M. N., & Yusuf, M. A. (2014, August). Improved method of classification algorithms for crime prediction. In 2014 International Symposium on Biometrics and Security Technologies (ISBAST) (pp. 250-255). IEEE.
14. Chaudhuri, A. K., Ray, A., Banerjee, D. K., & Das, A. Selection of Variables in Logistic Regression Model with Genetic Algorithm for Stroke Prediction.
15. Chaudhuri, A. K., Ray, A., Banerjee, D. K., & Das, A. (2021). A Multi-Stage Approach Combining Feature Selection with Machine Learning Techniques for Higher Prediction Reliability and Accuracy in Cervical Cancer Diagnosis. International Journal Of Computing and Digital System