International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2020 IJSRCSEIT | Volume 6 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT206267

An Effective Mechanism for Detecting Crime Rate in Chennai Location Using Supervised Machine Learning Approach

Dr. R. Poorvadevi[1], G. Sravani[2], V. Sathyanarayana[2]
[1]Assistant Professor, CSE Department, SCSVMV University, Kanchipuram, Tamil Nadu, India
[2]UG Scholar, CSE Department, SCSVMV University, Kanchipuram, Tamil Nadu, India

ABSTRACT

In the current era of digital world, the crime is the important challenge among the distinct user. People are applying various techniques to prevent and reduce the crime. But there is no specific solution is optimal for crime issues. It is need to be tracking the all sets of crimes which is managed and stored in the crime specific database. The proposed work brings the solution to identify the occurrences of the crime for Chennai region and also tracking the location and type of threats over the criem can be detected in the public user group. This mechanism will be achieved the effective outcomes by applying the supervised meachine learning approach.

Keywords : Crime Dataset, Data Analysis, Machine Learning, Accuracy, Data predictor, Self evaluation platform

I. INTRODUCTION

Crimes are the significant threat to the humankind. There are many crimes that happens regular interval of time. Perhaps it is increasing and spreading at a fast and vast rate. Crimes happen from small village, town to big cities. Crimes are of different type – robbery, murder, rape, assault, battery, false imprisonment, kidnapping, homicide. Since crimes are increasing there is a need to solve the cases in a much faster way.

The crime activities have been increased at a faster rate and it is the responsibility of police department to control and reduce the crime activities. Crime prediction and criminal identification are the major problems to the police department as there are tremendous amount of crime data that exist. There is a need of technology through which the case solving could be faster.

II. Procedure for Paper Submission

2.1 Domain Overview

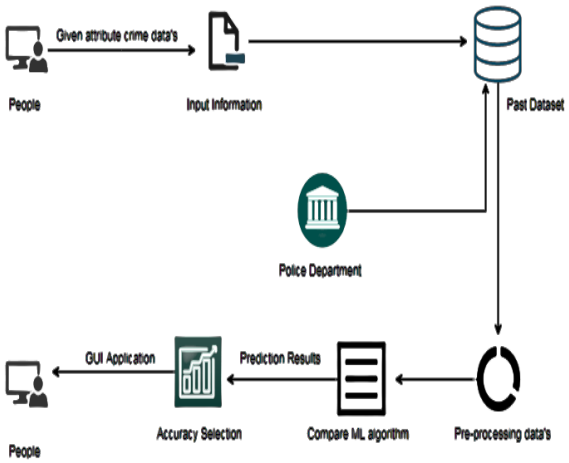
Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data.

The basics of Machine Learning, and implementation of a simple machine learning algorithm using python. Process of training and prediction involves use of specialized algorithms. We feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data

Machine learning can be roughly separated in to three categories. There are supervised learning, unsupervised learning and reinforcement learning.

Supervised learning program is both given the input data and the corresponding labeling.to learn data has to be labeled by a human being beforehand. Unsupervised learning is no labels. It provided to the learning algorithm. This algorithm has to figure out the clustering of the input data. Finally, Reinforcement learning dynamically interacts with its environment and it receives positive or negative feedback to improve its performance.

2.2 Design architecture



2.3 Problem Description

In today's world, every establishment is facing ever growing challenges which need to be coped up quickly and efficiently. With continually increasing population, crimes and crime rate analyzing related data is a huge issue for governments to make strategic decisions so as to maintain law and order. This is really necessary to keep the citizens of the country safe from crimes. The best place to look up to find room for improvement is the voluminous raw data that is generated on a regular basis from various sources by applying Data Science with machine learning concept.

III. DATA Analysis

3.1 Data source

The dataset has been obtained from the kaggle website. The various crimes have been uploaded with the particular data.the crimes are being taken only in the Chennai region, tamilnadu. (https://www.kaggle.com/siddharthaduggirala2/crime-analysis-in-india/data)

The dataset is now supplied to machine learning model on the basis of this dataset the model is trained. In the first step of accumulating information, data from previously/ current datasets from online sources are gathered together. These datasets are merged to form a common dataset, on which analysis will be done.

Dataset Description

Table shows details of the datasets:

Variable	Description
dc-dist	District Boundary
Psa	Police Station
Dis-date	Dispatch Date
Dis-time	Dispatch Time
Hour	The generalized hour
Month	The generalized month
Year	The generalized year
User_gen	Crime code
Type_crime	Various crimes
Pol_dis	Police department
Area	Area where crime happend

3.2 Training the Dataset

- The first line imports iris data set which is already predefined in sklearn module and raw data set is basically a table which contains information about various varieties.
- For example, to import any algorithm and train_test_split class from sklearn and numpy module for use in this program.
- To encapsulate load data () method in data dataset

variable. Further divide the dataset into training

	Dc_Dist	psa	dis_date	hour	user_gen	type_crime	Year	Month	Area
0	18	3	02/10/2009	14	800	Other Assaults	2009	10.0	FlowerBazaar
1	14	1	10/05/2009	0	2600	All Other Offenses	2006	5.0	HighCourt
2	25	J	07/08/2009	15	800	Other Assaults	2007	8.0	Harbour
3	35	D	19/07/2009	1	1500	Weapon Violations	2008	7.0	PortMarine
4	9	R	25/06/2009	0	2600	All Other Offenses	2010	6.0	Washermenpet
...	...	...	...	...	...	...	...	...	...
101199	14	4	23/02/2010	19	500	Burglary Residential	2011	5.0	Teyampet
101200	14	2	01/03/2010	17	500	Burglary Residential	2012	11.0	FlowerBazaar
101201	14	2	01/03/2010	18	500	Burglary Residential	2013	3.0	HighCourt
101202	14	4	01/03/2010	21	500	Burglary Residential	2014	3.0	Harbour
101203	14	2	02/03/2010	12	500	Burglary Residential	2015	4.0	PortMarine

101204 rows × 9 columns

data and test data using train_test_split method. The X prefix in variable denotes the feature values and y prefix denotes target values.

- This method divides dataset into training and test data randomly in ratio of 67:33 / 70:30. Then we encapsulate any algorithm.
- In the next line, we fit our training data into this algorithm so that computer can get trained using this data. Now the training part is complete.

3.3 Testing the Data Set

Now, the dimensions of new features in a numpy array called 'n' and it want to predict the species of this features and to do using the predict method which takes this array as input and spits out predicted target value as output.

So, the predicted target value comes out to be 0. Finally, to find the test score which is the ratio of no. of predictions found correct and total predictions made and finding accuracy score method which basically compares the actual values of the test set with the predicted values.

3.4 Data Validation Process

The dataset needs to be preprocessed to fill the empty cells, delete unnecessary coulmns and add some other features in the dataset to obtain the accuracy.

The below table is a preprocessed one:

3.3 Exploratory Analysis of Pre-processing

The dataset has been analysed by comparing one column to another. Exploratory analysis techniques will enchance the easy way of coulmns comparision.

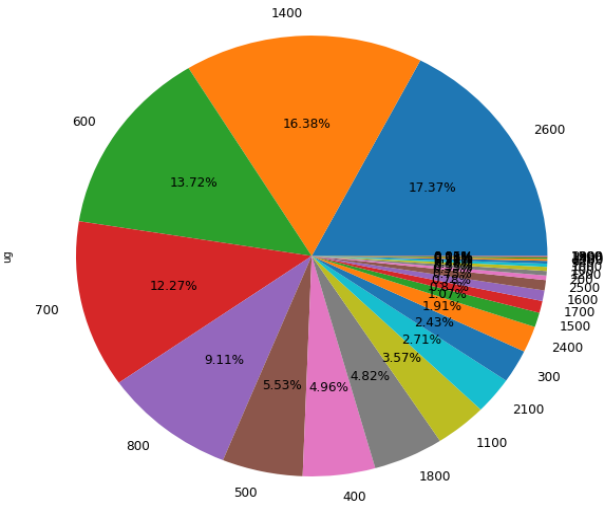
Below diagram shows the exploratory analysis:

user_gen	100	200	300	400	500	600	700	800	900	1000	...	1700	1800	1900	2000	2100	2200	2300	2400	2500	2600
type_crime																					
Aggravated Assault Firearm	0	0	0	1773	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
Aggravated Assault No Firearm	0	0	0	3247	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
All Other Offenses	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	17583
Arson	0	0	0	0	0	0	0	0	208	0	...	0	0	0	0	0	0	0	0	0	0
Burglary Non-Residential	0	0	0	0	636	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
Burglary Residential	0	0	0	0	4963	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
DRIVING UNDER THE INFLUENCE	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	2738	0	0	0	0	0
Disorderly Conduct	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1935	0	0
Embezzlement	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
Forgery and Counterfeiting	0	0	0	0	0	0	0	0	0	235	...	0	0	0	0	0	0	0	0	0	0
Fraud	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
Gambling Violations	0	0	0	0	0	0	0	0	0	0	...	0	0	9	0	0	0	0	0	0	0
Homicide - Criminal	301	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
Homicide - Gross Negligence	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
Liquor Law Violations	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	142	0	0	0	0
Motor Vehicle Theft	0	0	0	0	0	0	2400	0	0	0	...	0	0	0	0	0	0	0	0	0	0
Narcotic / Drug Law Violations	0	0	0	0	0	0	0	0	0	0	...	0	4874	0	0	0	0	0	0	0	0

3.5 Data Visualization Process

In the data visualization the dataset will be in the form of graphs (barchart, piechart, scatterplot, heatmap etc). The data visualization makes esier to understand the data even in the huge amount.

Below is the data visuallizied in the form of piechart:



IV. ALGORITHMS

In this analysis machine learning algorithms are used to analyze the dataset and to obtain the best accuracy.

The three algorithms used are:

1. Logistic Regression
2. Random Forest
3. K-Nearest Neighbour

4.1 Logistic Regression

Back in the ancient times (the '50s), David Cox, a British Statistician, invented an algorithm to predict the probabilities of events given certain variables.

Logistic Regression assigns a certain probability (from 0 to 1) to a binary event, given its context.

To use it, we'll first create the input vectors, where each vector corresponds to an athlete, and each of a vector's fields is a (numerical) feature of that athlete (for instance, their Weight or Height).

We'll then try to predict the probability of one of the fields being 1 or 0 (in our case, 1 could mean female and male, for instance).

It does this by

- Performing an affine transformation in the input features–that is, multiplying them by a matrix, and adding a bias vector to the product. We call the elements of the matrix and bias vector 'weights'.
- Composing that operation with the sigmoid function, which 'crunches numbers' from the whole real domain to just the (still infinite) numbers between 0 and 1. This gives a notion of probability to the result.
- 'Penalizing' the model with a cost function (in our case, we'll use cross entropy): If we want the model to learn a certain thing, we'll have to penalize it for not learning it..
- Finding the gradient of that cost function (which we wish to minimize) as a function of the model's weights.
- Updating the weights just a bit, dictated by a constant called 'learning rate' towards the opposite direction so the cost function decreases in the next iteration.

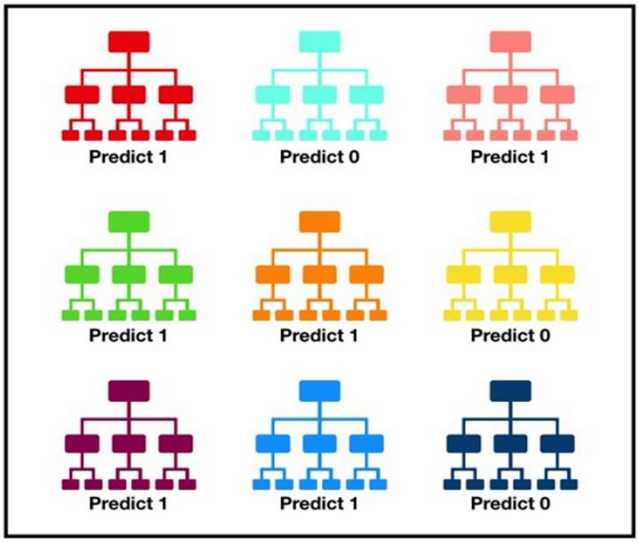
This cycle is performed over and over, iterating the whole inputs set many times. Each of these iterations is called an 'epoch'.

Eventually the function converges into a value that is usually a local minimum of the cost function. It is, however, not guaranteed to be a global minimum.

This whole process is called 'training' the model. After the model has trained, we will look at how well it performs by measuring its accuracy: how many of the predictions it made were true, divided by how many predictions it made in total. In statistics terms, how many true positives and true negatives the model had, over the whole set of predictions.

4.2 Random forest

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction (see figure below).



4.3 K-nearest neighbour

The K – nearest neighbors (KNN) algorithm is a simple, effective and esay to implement the process with superviser machine learning approach. This will be used for solving classification and regression problems.

V. ENVIRONMENTAL REQUIREMENTS

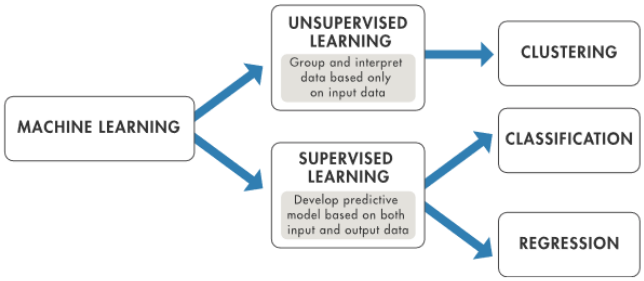
Software Requirements:

Operating system: windows
Tool : anaconda with jupyter notebook

Hardware requirements:

Processor : Pentium IV/III
Hard disk : minimum 80 GB
RAM : minimum 2 GB

VI. MACHINE LEARNING



Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop a conventional algorithm for effectively performing the task.

VII. APPLICATION

To prove, how effective and accurate machine learning algorithms can be at predicting violent crimes, there are other applications in the territory of law enforcement such as determining criminal, creating criminal profiles, and learning crime trends. Utilizing these applications can be a long and tedious process for law enforcement officials who have to sift through large volumes of data. However, the precision in which one could infer and create new knowledge on how to slow down crime is well worth the safety and security of people.

VIII. FUTURE WORK

- Police department wants to automate the detecting the crime from eligibility process (real

time) based on the crime rate of areas.

- To automate this process by show the prediction result in web application or desktop application.
- To optimize the work to implement in Artificial Intelligence environment.

IX. CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score will be find out for future. This brings some of the following insights about crime rate.

It has become easy to find out relation and patterns among various data's. It, mainly revolves around predicting the type of crime which may happen if we know the location of where it has occurred in real time world. Using the concept of machine learning we have built a model using training data set that have undergone data cleaning and data transformation.

Data visualization generated many graphs and found interesting statistics that helped in understanding Indian crimes datasets that can help in capturing the factors that can help in keeping society safe.

X. REFERENCES

[1] Kaggle.com. (2020). *Crime Analysis in India*. [online] Available at: https://www.kaggle.com/siddharthaduggirala2/crime-analysis-in-india/data [Accessed 20 Feb. 2020H.

[2] Crime pattern detection, analysis & prediction. An overview on crime prediction methods.

[3] Crime prediction and forecasting in Tamilnadu using clustering approaches. In Emerging Technological Trends (ICETT),

[4] http://ieeexplore.ieee.org/document/820367 the reference to detect the crimes

Cite this article as :

Dr. R. Poorvadevi, G. Sravani, V. Sathyanarayana, "An Effective Mechanism for Detecting Crime Rate in Chennai Location Using Supervised Machine Learning Approach", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 2, pp. 326-331, March-April 2020. Available at doi : https://doi.org/10.32628/CSEIT206267
Journal URL : http://ijsrcseit.com/CSEIT206267