International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 03 | Mar 2019 www.irjet.net p-ISSN: 2395-0072

Prediction of Crime Rate Analysis Using Supervised Classification Machine Learning Approach

Kirthika V[1], Krithika Padmanabhan A [2], Lavanya M[3], Lalitha S D[4]

[1]*Student, Computer Science and Engineering, RMK Engineering College, Tamil Nadu, India*
[2] *Student, Computer Science and Engineering, RMK Engineering College, Tamil Nadu, India*
[3] *Student, Computer Science and Engineering, RMK Engineering College, Tamil Nadu, India*
[4] *Assistant Professor, Computer Science and Engineering, RMK Engineering College, Tamil Nadu, India*

Abstract - *In recent years, report points out that the crimes in India have seen a spike. There is no particular reason for any trouble for criminal activities. Sometimes society, cultural factors, different family systems, political influences and law enforcement are responsible for the criminal activities of an individual. Crime can be found in various categories. To prevent this problem in police sectors, we must predict crime rate using machine learning techniques. The aim is to investigate machine learning based techniques for crime rate by prediction results in best accuracy and explore in this work the applicability of data technique in the efforts of crime prediction with particular importance to the data set. The analysis of dataset by supervised machine learning technique (SMLT) to capture several information's like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments and analyse the data validation, data cleaning/preparing and data visualization will be done on the entire given dataset. Our analysis provides a comprehensive guide to sensitivity analysis of model parameters with regard to performance in prediction of crime rate by accuracy calculation from comparing supervise classification machine learning algorithms*

Key Words: **Dataset, Crime rate analysis, Machine Learning-Classification method, Python, Prediction of Accuracy result.**

1.INTRODUCTION

1.1 Crimes are the significant threat to the humankind. There are many crimes that happens regular interval of time. Crimes have different types are 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.

1.2 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. The problem made me to go for a research about how can solve a crime case made easier. Through many documentation and cases, it came out that machine learning and data science can make the work easier and faster.

1.3 The aim of this project is to make crime prediction using the features present in the dataset. The dataset is extracted from the official sites. With the help of machine learning algorithm, using python as core we can predict the type of crime which will occur in an area. The objective would be to test a model for prediction. The training would be done using the training data set which will be validated using the test dataset. Building the model will be done using better algorithm depending upon the accuracy. The supervised classification and other algorithm will be used for crime prediction.

1.4 Visualization of dataset is done to analyse the crimes which may have occurred in the country. This work helps the law enforcement agencies to predict and detect crimes in India with improved accuracy and thus reduces the crime rate. This helps all others department to carried out other formalities.

2. PROPOSED SYSTEM

2.1 Predictive Model

Predictive modeling is the way of building a model that can make predictions. The process includes a machine learning algorithm that learns certain properties from a training dataset in order to make those predictions.

The different types of predictive models are

a. Decision Trees - A decision tree is an algorithm that uses a tree shaped graph or model of decisions including chance event outcomes, costs, and utility. It is one way to display an algorithm. It builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.

b. Support Vector Machines - A classifier that categorizes the data set by setting an optimal

hyper plane between data. I chose this classifier as it is incredibly versatile in the number of different kernelling functions that can be applied, and this model can yield a high predictability rate.

c. Logistic Regression - Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 or 0

d. K-Nearest Neighbor - K-Nearest Neighbor is a supervised machine learning algorithm which stores all instances correspond to training data points in n-dimensional space. When an unknown discrete data is received, it analyzes the closest k number of instances saved (nearest neighbors) and returns the most common class as the prediction and for real-valued data it returns the mean of k nearest neighbors.

e. Random forests - Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. Random forest is a type of supervised machine learning algorithm based on ensemble learning.

2.2 Functional Diagram of Proposed Work

It can be divided into 4 parts:

a. Data processing and cleaning
b. Random sampling
c. Train model
d. Estimate the performance

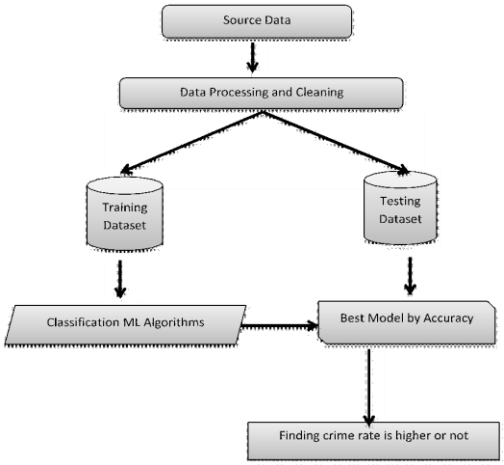

Fig -1: Functional Diagram

a. DATA PROCESSING AND CLEANING

In this step we need prepare data into right format for analysis and cleaning. We may need to transform the variables using one of the approaches

1. Normalization or standardization
2. Missing Value Treatment

b. RANDOM SAMPLING

Training Sample: Model will be developed on this sample. 70% or 67% of the data goes here.

Test Sample: Model performances will be validated on this sample. 30% or 33% of the data goes here.

c. TRAIN MODELS

Validate the assumptions of the chosen algorithm. Develop/Train Model on Training Sample, which is the available data and check Model performance - Error, Accuracy, etc.

d. ESTIMATE THE PERFORMANCE

Score and Predict using Test Sample and check Model Performance: Accuracy, Error, Precision etc.

3. IMPLEMENTATION

In the first step of accumulating information, data from previously available/ current datasets from online sources are gathered together. These datasets are merged to form a common dataset, on which analysis will be done.

3.1 Data collection

The data set collected for predicting crimes is split into Training set and Test set. Generally, 7:3 ratios are applied to split the Training set and Test set. The Data Model, which was created using Random Forest, logistic, Decision tree algorithms, K-Nearest Neighbor (KNN) and Support vector classifier (SVC) are applied on the Training set and based on the test result accuracy, Test set prediction is done.

Variable	Description
Dc_Dist	District boundary
Psa	Police Service Areas(PSA Boundary)
Dis_date	Dispatch date (The officer was dispatched to the scene)
Dis_time	Dispatch time (The officer was dispatched to the scene)
Hour	The generalized hour of the dispatched time
User_gen	General user crime category code
Type_crime	Categories of crime names
Pol_dis	It describing the national policy on discrimination cases registered at the police department
Year	Year of Observations
Month	Month of incidents
area	Area of incidents

Fig -2: Dataset Description

3.2 Data Preprocessing

This process includes methods to remove any null values or infinite values which may affect the accuracy of the system. The main steps include: Formatting, cleaning and sampling. Cleaning process is used for removal or fixing of some missing data there may be data that are incomplete. Sampling is the process where appropriate data are used which may reduce the running time for the algorithm. Using python, the preprocessing is done.

3.3 Feature selection

Features selection is done which can be used to build the model. The attributes used for feature selection are Dc_Dist, Psa, Dis_date, Dis_time, Hour, User_gen, Pol_dis, Year, Month and area.

3.4 Training

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

3.5 Prediction

The dimensions of new features in a numpy array and 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 or 1. 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.

4. RESULTS AND DISCUSSION

The results are obtained after undergoing various processes that comes under machine learning.

Data preprocessing - Data preprocessing includes dropping row without any row and converting any value which consist of value as infinity. Converting string variable into numerical so that it can undergo further processing.

	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
5	17	1	25/04/2015	12	600	Thefts	2011	4.0	Thiruvottiyur
6	23	K	10/02/2009	14	800	Other Assaults	2012	2.0	Royapuram
7	77	A	02/04/2009	18	500	Burglary Non-Residential	2013	4.0	Madhavaram
8	35	D	18/03/2009	1	2600	All Other Offenses	2014	3.0	Puzhal
9	23	L	14/06/2009	20	2600	All Other Offenses	2015	6.0	Ennore
10	22	P	19/01/2009	16	400	Aggravated Assault Firearm	2005	1.0	Pulianthope
11	1	J	09/02/2009	22	800	Other Assaults	2009	2.0	MKBnager
12	22	3	06/10/2015	18	600	Thefts	2006	10.0	Sembium
13	22	3	09/10/2015	0	600	Thefts	2007	10.0	AnnaNagar
14	77	A	03/05/2015	20	600	Thefts	2008	5.0	Thirumangalam
15	2	1	30/11/2015	8	600	Thefts	2010	11.0	Koyambedu

Fig -3: Preprocessed Dataset

After dividing the data set intro training set and testing set the model is trained using algorithm as mentioned in the table. The accuracy is calculated using the function score_accuracy imported from metric from sklearn. The accuracy is mentioned in the table below.

Algorithm	Precision	Recall	F1-Score	Accuracy (100%)
DT	1	1	1	98
SVC	0.76	0.60	0.45	59.97
LR	0.81	0.80	0.79	79.71
KNN	0.87	0.87	0.86	86.73
RF	1	1	1	98

Fig -4: Results

As we can see from the results obtained from the table the algorithm which can be used for the predictive modeling will be Decision Trees or Random Forest algorithms with accuracy of 98%, the highest among the rest of the algorithm. The least which can be used will be SVM. For further modelling using unseen data there is no need for using other algorithm.

5. CRIME VISUALIZATION

This section deals with the analysis done on the dataset and plotting them into various graphs like bar, pie, scatter. Analysis done are

a. crime categories by total areas
b. crime codes in percentage values
c. Relationship diagram for co-relation dataset columns
d. Classify the crime rate by PSA
e. Classify the crime rate by Year

This graph shows which crimes have occurred most in the city. The x coordinate denotes the Types of crimes committed and y coordinate denotes the area code.

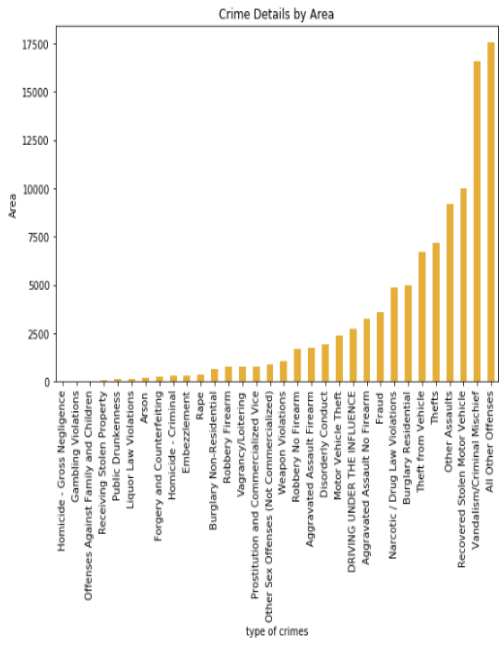

Fig -5: Most occurring crimes in the city

The graph below shows the percentage of different crimes happening in the city. 17.37 % of crime are vandalism and 16.38 % are thefts.

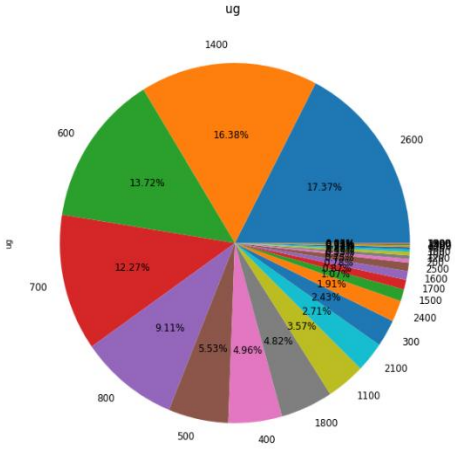

Fig -6: Percentage of different crimes happening in the city

The graph below shows the relationship diagram for co-relation dataset columns

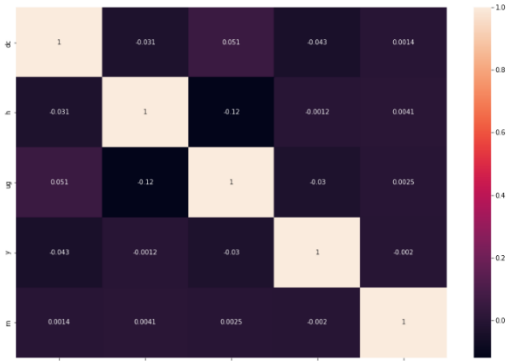

Fig -7: Relationship diagram for co-relation dataset columns

The graph below shows the crime rate by PSA. The x coordinate denotes PSA and y coordinate denotes the crime rate.

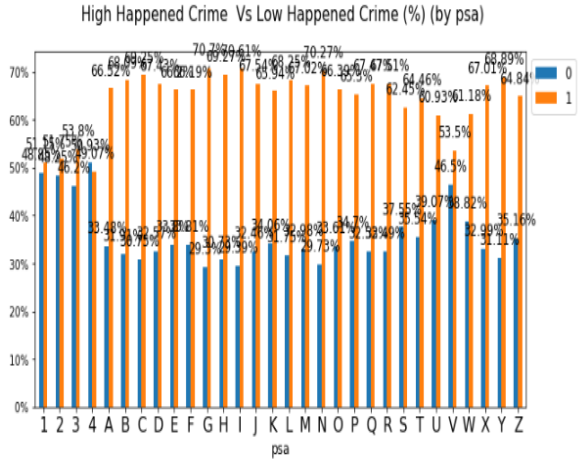

Fig -8: Crime rate by PSA

The graph below shows the crime rate by year. The x coordinate denotes year and y coordinate denotes the crime rate.

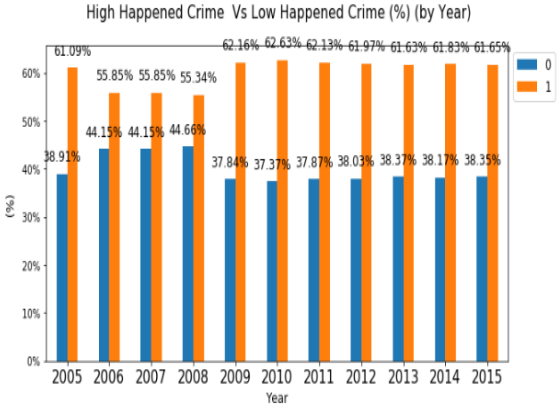

Fig -9: Crime rate by year

6. CONCLUSIONS

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 of decision tree algorithm/ Random forest method. 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. The model predicts the type of crime with accuracy of 100. 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.

REFERENCES

[1] Shamsuddin, N. H. M., Ali, N. A., & Alwee, R. (2017, May). An overview on crime prediction methods. In Student Project Conference (ICT-ISPC), 2017 6th ICT International (pp. 1-5). IEEE.

[2] Al Boni, M., & Gerber, M. S. (2016, December). Area Specific Crime Prediction Models. In Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on (pp. 671-676). IEEE.

[3] Sivaranjani, S., Sivakumari, S., & Aasha, M. (2016, October). Crime prediction and forecasting in Tamilnadu using clustering approaches. In Emerging Technological Trends (ICETT), International Conference on (pp. 1-6). IEEE.

[4] Sathyadevan, S., & Gangadharan, S. (2014, August). Crime analysis and prediction using data

mining. In Networks & Soft Computing (ICNSC), 2014 First International Conference on (pp. 406-412). IEEE.

[5] Zhao, X., & Tang, J. (2017, November). Exploring Transfer Learning for Crime Prediction. In Data Mining Workshops (ICDMW), 2017 IEEE International Conference on (pp. 1158-1159). IEEE.

[6] Tayebi, M. A., Gla, U., & Brantingham, P. L. (2015, May). Learning where to inspect: location learning for crime prediction. In Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on (pp. 25-30). IEEE.