Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.199

IJCSMC, Vol. 9, Issue. 3, March 2020, pg.221 – 229

GUI BASED PREDICTION OF CRIME RATE USING MACHINE LEARNING APPROACH

Mrs. Prithi S

Assistant Professor

Department of Computer Science and Engineering

Rajalakshmi Engineering College

Thandalam, Chennai-602105.

Aravindan S; Anusuya E; Ashok Kumar M

UG Students

Department of Computer Science and Engineering

Rajalakshmi Engineering College

Thandalam, Chennai-602105.

ABSTRACT: Crime in India is increasing in various forms. Crime rates are increases based on location and time. There is no specific reason for any criminal activity. To prevent this problem, Police sectors have to predict crime rate using machine learning. 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 is carried out by supervised machine learning technique (SMLT) to capture few vital information and to perform data validation, data cleaning and data visualization on the given dataset. The analysis does the prediction of accuracy by comparing the result of different supervised machine learning algorithms. The most accurate algorithm would be taken from the comparison and predict the result.

Keywords: dataset, Machine learning-Classification method, python, Prediction of Accuracy result.

I. INTRODUCTION:

Crime is the significant threat to the humankind. There are many crimes that occurs in regular interval of time. Crimes activities can be robbery, murder, assault, kidnapping. Since crimes are increasing there is a need to solve the cases and to reduce the crime activities. Crime prediction is the major problem for police department as there are huge amount of crime data. By using those data sets we can work on with machine learning to predict crime.

Machine learning is to predict the future from past data. Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. The computer programs using machine learning change according to the data it is exposed to. Simple machine learning algorithm can be implemented using python. Process of training and prediction involves use of specialized machine learning algorithms. The training dataset to feed to the algorithm and the algorithm uses this training data to give predictions on a new test data.

Data scientists use many different kinds of machine learning algorithms to discover patterns in python that lead to actionable insights. These different algorithms can be classified into two : supervised and unsupervised learning. Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories.
Classification predictive modeling is the task of approximating a mapping function from input variables(X) to discrete output variables(y). In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the input data given to it and then uses this knowledge to classify new observations.

Supervised Machine Learning is the majority of practical machine learning uses supervised learning. Supervised learning is where have input variables (X) and an output variable (y) and use an algorithm to learn the mapping function from the input to the output is $y = f(X)$. The goal is to approximate the mapping function so well that when you have new input data (X) that you can predict the output variables (y) for that data. Techniques of Supervised Machine Learning algorithms include logistic regression, multi-class classification, Decision Trees and support vector machines etc. Supervised learning requires Machine Learning Past Dataset Result that the data used to train the algorithm is already labeled with correct answers. Supervised learning problems can be further grouped into Classification problems. This problem has as goal the construction of a succinct model that can predict the value of the dependent attribute from the attribute variables. The difference between the two tasks is the fact that the dependent attribute is numerical for categorical for classification. A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes.

II. LITERATURE SURVEY:

[1] The results of our experiments confirm our alignment with previous studies which debunked the common that unemployment and violent crimes are strongly correlated and then tested whether there was any positive linear relationship between fines and violent crimes. Due to the complex ways in which boundaries are drawn and crimes are defined, we saw no relationship at the local level. At the state level, however, the linear relationship became apparent and statistically significant. The results of our fit were confirmed by overlaps between the top fine states and top violent crime states. It also discussed equitable stop and search treatment with respect to subsets of the population. The causes of violent crime are a highly nuanced topic. It showed that a relationship between areas marked by high fines and high rates of violent crime exists, and there are potential consequences of excess fining in certain areas, it analyzes and discusses the dependence of city and county revenue generated from fines (primarily traffic violations) and their potential effects on the incidence of violent crimes on an aggregated state level. Following the riots, several press articles pointed to Ferguson's elevated levels of municipal court fines (again, usually for traffic violations) and how they reduced the local population's faith in the police and overall city government. It tested whether the practice of collecting significant municipal revenue from low-level offenses had an impact on violent crimes not only in Missouri, but in other states as well.

[2] India's population is estimated to be around one billion. The high population density, combined with other factors such as lack of jobs, poverty, and illiteracy will result in a higher violence rate. The crime and violence rate vary from state to state. States like Uttar Pradesh, Bihar etc records high crime rates according to 2017 statistics. Like other counties increase in crime rate is a major concern in India also. From the reports of National Crime Record Bureau (NCRB), states that most of crime incidents recorded are in urban area. In India, crime rate (case reported per lakh population) has increased from 166.7 to 215.5 in years from 1953 to 2013. By analyzing the data, crime rates got highly fluctuated in the years 1970-2005. The statistics indicate that crime rate in India is steadily increasing for the past 8-9 years. Source of data is from the National Crime Record Bureau of India. As a part of modeling, data is divided into training data for the years 1953 to 2008 and test data for the years 2009 to 2013. By examining the model, it's clear that the forecast values are within the 95% confidence interval of the test data and accuracy measurements are also significant. Hence the time series model suitable for crime forecasting. This paper concluded that time series model can be applied for crime forecasting. The result obtained from both the models conclude that they are significant for forecasting all test data which are lying between a 95% confidence interval and accuracy measurements for training data shows that they are numerically significant. In future, we are trying to analyze crime against women, children so that we can predict how much police strength is convenient to decrease the crime rate.

[3] In the past a powerful reliance has been placed on standard video surveillance in order to achieve this goal. This often creates a backlog of video data that has to be monitored by a supervising official. For large urban areas, this creates an increasingly large workload for supervising officials which leads to an increase in error rate. Solutions are implemented to help reduce the workload. Currently, auto regressive models are wont to better forecast criminal acts, but also have a list of shortcomings. It proposed an answer of using neural networks in together with a

Hybrid Deep Learning algorithm to research video stream data. Our system are going to be able to quickly identify and assess criminal activity which will can successively reduce workloads on the supervising officials. When implemented across smart city infrastructure it will allow for a efficient and adaptable crime detection system. Our system will be applied to varied video surveillance systems to act as an alert system, which would reduce the overall workload on security officials. Automation and smart, adaptive security systems are a way to increase detection rates in hopes of curbing crime rates in large hard to monitor areas.

III. EXISTING SYSTEM:

Latest technical developments in sophisticated tools of data analytics and visualization are helping the society in numerous ways to investigate the info of social relevance. One among such socially relevant activities is crime details of various demographic places. The analysis of the crime data will help higher cognitive process agencies to require precautionary steps to regulate the rate over demographic places. Advancements within the field of knowledge technology, publicly available information and services, somehow help criminals to attain their misdeeds and involve them in much serious crimes than earlier. As a result, rate is increasing with a really high rate in developed and under-developed nations. supported the previous year crime details in Indian states, It present statistical models through Weighted Moving Average, Functional Coefficient Regression and Arithmetic-Geometric Progression based prediction of the crime in coming years. Difference between actual records and our predicted values for both years gives the accuracy of the proposed approaches between the range 85% and 90%. In future, this work are often modified by using Machine Learning (ML) models for forecasting crime, because the data points will sufficiently increase to use ML models. This will also increase the accuracy of the predictions. Further, statistical modeling's methods may also be clubbed with ML models so calculate weighted accuracy for a part, this will make the answer more robust.

Drawbacks

- The accuracy results are not more than 90%. Its cannot work on top features and find-out the Recall, Precision, Confusion matrix and compare it with our old result.
- It's cannot work on using the popular machine learning algorithm to find out the features importance.

IV. PROPOSED SYSTEM:

Machine learning is a computer system's method of learning by way of examples. There are many machine learning algorithms available to users that can be implemented on data sets. However, there are two major types of learning algorithms: supervised learning and unsupervised learning algorithms. Supervised learning algorithms work by inferring information or "the right answer" from labeled training data. The algorithms are given a particular attribute or set of attributes to predict. Data preprocessing 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.

Crimes Prediction ways:

- To utilize the resources identify the hotspots of crimes and allocate vigilante resources such as policeman, police cars, weapons etc. reschedule patrols according to the vulnerability of a place.
- Through that avoid crimes Ensure better civilization through avoiding happening crimes such as murder, rapes, thefts, drug, smuggling etc.

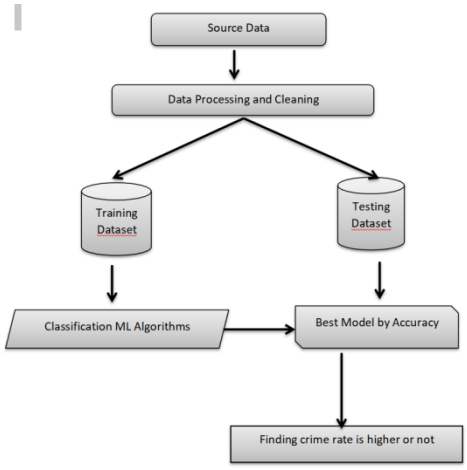

Fig 1: Workflow Diagram

The above figure 1 denotes the work flow diagram of the proposed system.

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.

2. Data Preprocessing

This process includes methods to remove any null values or infinite values which can 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 is also data that are incomplete. Sampling is that the process where appropriate data are used which can reduce the period of time for the algorithm. Using python, the preprocessing is completed. The data which was collected might contain missing values which will result in inconsistency. To realize better results data have to be preprocessed so on improve the efficiency of the algorithm. The outliers should be removed and also variable conversion have to be done. Based on the correlation among attributes it was observed that attributes that are significant individually include property area, education, loan amount, and lastly credit history, which is that the strongest among all. Some variables like applicant income and co- applicant income are not significant alone, which is strange since by intuition it's considered as important.

The correlation among attributes can be identified using plot diagram in data visualization process. Data preprocessing is the most time consuming phase of a data mining process. Data cleaning of loan data removed several attributes that has no significance about the crimes. Data integration, data reduction and data transformation are also to be applicable for loan data. For easy analysis, the data is reduced to some minimum amount of records. The dataset obtained from online is maintained and updated by the Indian police department.

3. Construction of a Predictive Model

Machine learning needs data gathering have lot of past data's. Data gathering have sufficient historical data and raw data. Before data pre-processing, raw data can't be used directly. It's used to preprocess then, what kind of algorithm with model. Training and testing this model working and predicting correctly with minimum errors. Tuned model involved by tuned time to time with improving the accuracy. Finally, once model is ready, deployed and model to do the predictions and the aims and objectives due to the inconsistency in historical data on bank accountant therefore perform an analysis of the given dataset and describe how to repair it automatically**.** The below figure 2 denotes the process of dataflow diagram.

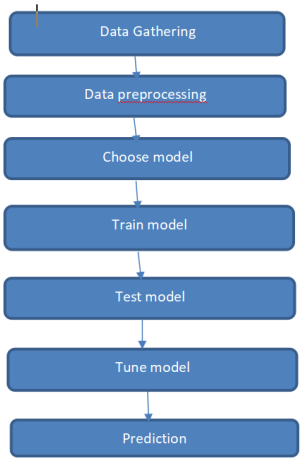

Fig 2: Process of dataflow diagram

4. 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 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.

5. Testing the Dataset

- 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.

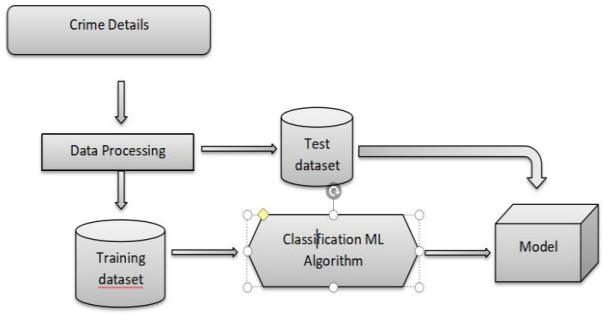

Fig 3: Architecture of Proposed model

The above figure 3 denotes the proposed models architecture.

6. Comparing Algorithm with prediction in the form of best accuracy result

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created. When have a new dataset, it is a good idea to visualize the data using different techniques in order to look at the data from different perspectives. The same idea applies to model selection. You should use a number of different ways of looking at the estimated accuracy of your machine learning algorithms in order to choose the one or two to finalize. A way to do this is to use different visualization methods to show the average accuracy, variance and other properties of the distribution of model accuracies.

In the next section you will discover exactly how you can do that in Python with scikit-learn. The key to a fair comparison of machine learning algorithms is ensuring that each algorithm is evaluated in the same way on the same data and it can achieve this by forcing each algorithm to be evaluated on a consistent test harness.

In the example below different algorithms are compared:

- Logistic Regression
- Random Forest
- K-Nearest Neighbors
- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Classifier

The K-fold cross validation procedure is used to evaluate each algorithm, importantly configured with the same random seed to ensure that the same splits to the training data are performed and that each algorithm is evaluated in precisely the same way. Before that comparing algorithm, Building a Machine Learning Model using install Scikit-Learn libraries. In this library package have to done preprocessing, linear model with logistic regression method, cross validating by KFold method, ensemble with random forest method and tree with decision tree classifier. Additionally, splitting the train set and test set. To predicting the result by comparing accuracy.

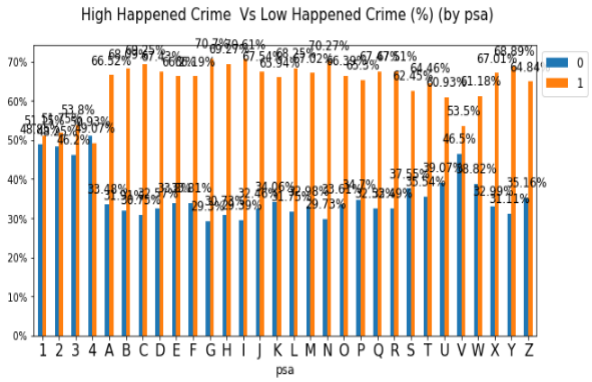

Fig 4: Classify the crime rate by PSA

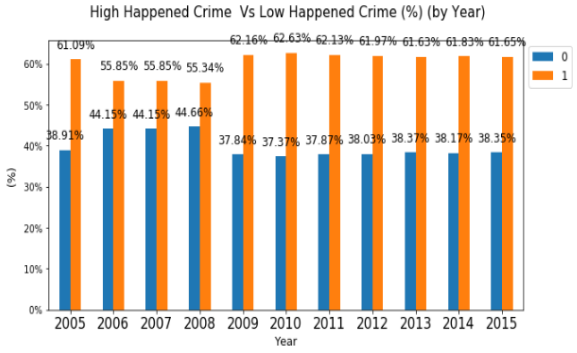

Fig 5: Classify the crime rate by Year

The above Figures 4 & 5 denotes the crime rate by PSA and crime rate by year.

Prediction result by accuracy

Logistic regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. We need the output of the algorithm to be classified variable data. Higher accuracy predicting result is logistic regression model by comparing the best accuracy.

Cross validation process

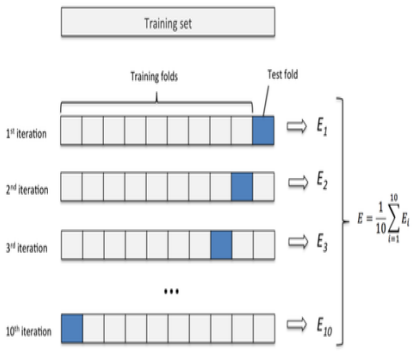

Fig: Cross validation process

Above figure is the Cross validation process. Over-fitting is a common problem in machine learning which can occur in most models. K-fold cross-validation can be conducted to verify that the model is not over-fitted. In this method, the data-set is randomly partitioned into *k mutually exclusive* subsets, each approximately equal size and one is kept for testing while others are used for training. This process is iterated throughout the whole k folds.

True Positive Rate(TPR) = TP / (TP + FN)

False Positive rate(FPR) = FP / (FP + TN)

Accuracy

The Proportion of the total number of predictions that is correct otherwise overall how often the model predicts correctly defaulters and non-defaulters.

Accuracy calculation

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same.

Precision is The proportion of positive predictions that are actually correct.

Precision = TP / (TP + FP)

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

Recall is The proportion of positive observed values correctly predicted. (The proportion of actual defaulters that the model will correctly predict)

Recall = TP / (TP + FN)

Recall(Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

General Formula

F- Measure = 2TP / (2TP + FP + FN)

F1-Score Formula

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

7. GUI creation

Finally, we create one Graphical User Interface. The below figure is the example GUI interface.

Fig: GUI Interface

V. CONCLUSION AND FUTURE WORK:

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 is will be find out. 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.

The future references that can be made are 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.

VI. 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.

REFERENCES

[1] SamuelSmith, Simranjyot Singh Gill and Kedar Gangopadhyay,"A Relationship between Fines and ViolentCrimes"-2018.

[2] Manish Kumar, Athulya S and Mary Minu MB,"Forecasting of Annual Crime Rate in India:A case Study"-2018.

[3] Sharmila Chackravarthy, Steven Schmitt and Li Yang,"Intelligent Crime Anomaly Detection in Smart Cities using Deep Learning"-2018

[4] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland, "Once upon a crime: towards crime prediction from demographics and mobile data," Proc. of the 16th Intl. Conf. on Multimodal Interaction, pp. 427-434, 2014. (a) (b) (c) (a) (b) (c) (d) 419

[5] H. Adel, M. Salheen, and R. Mahmoud, "Crime in relation to urban design. Case study: the greater Cairo region," Ain Shams Eng. J., vol. 7, no. 3, pp. 925-938, 2016.

[6] "Overall crime rate in Vancouver went down in 2017, VPD says," CBC News, Feb. 15, 2018. [Online] Available: https://www.cbc.ca/news/canada/british-columbia/crime-rate-vancouver2017-1.4537831. [Accessed: 09- Aug- 2018].

[7] J. Kerr, "Vancouver police go high tech to predict and prevent crime before it happens," Vancouver Courier, July 23, 2017. [Online] Available: https://www.vancourier.com/news/vancouver-police-go-high-tech-topredict-and-prevent-crime-before-it-happens-1.21295288. [Accessed: 09- Aug- 2018]

[8] J. Han, Data mining: concepts and techniques, Morgan Kaufmann, 2012.

[9] R. Iqbal, M. A. A. Murad, A. Mustapha, P. H. Shariat Panahy, and N. Khanahmadliravi, "An experimental study of classification algorithms for crime prediction," Indian J. of Sci. and Technol., vol. 6, no. 3, pp. 4219- 4225, Mar. 2013.

[10] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime data mining: a general framework and some examples," IEEE Computer, vol. 37, no. 4, pp. 50-56, Apr. 2004.

[11] T. Beshah and S. Hill, "Mining road traffic accident data to improve safety: role of road-related factors on accident severity in Ethiopia," Proc. of Artificial Intell. for Develop. (AID 2010), pp. 14-19, 2010.

[12] Q. Zhang, P. Yuan, Q. Zhou, and Z. Yang, "Mixed spatial-temporal characteristics based crime hot spots prediction," IEEE 20th Intl. Conf. on Comput. Supported Cooperative Work in Des. (CSCWD), Nanchang, China, May 2016.

[13] M. Al Boni and M. S. Gerber, "Area-specific crime prediction models," 15th IEEE Intl. Conf. on Mach. Learn. and Appl., Anaheim, CA, USA, Dec. 2016.

