PREDICTION of CRIME RATE ANALYSIS using MACHINE LEARNING APPROACH

Keerthi.R [1], Kirthika.B [2], Pavithraa.S [3], Dr. V.Gowri[4]

[1,2,3]*Department of Information Technology, Velammal Engineering College, Chennai, India*
[4]*Professor, Department of Information Technology, Velammal Engineering College, Chennai, India*

---***---

Abstract - *In recent years, report points out that the crimes in India have seen a spike. The report adds that the cases of murder, rapes, and kidnapping have seen a rise. Most of countries in the world have seen a remarkable increase in the crime rate. There is no particular reason for any trouble for criminal activities. To prevent this problem in police sectors have to 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 analyze 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, Machine learning-Classification method, python, Prediction of Accuracy result.**

1. INTRODUCTION

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. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using python. Process of training and prediction involves use of specialized algorithms. It 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. Data scientists use many different kinds of machine learning algorithms to discover patterns in python that lead to actionable insights. At a high level, these different algorithms can be classified into two groups based on the way they "learn" about data to make predictions: supervised and 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 data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc.
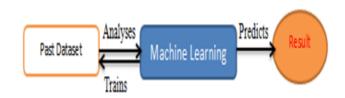

Fig-1: Process of Machine Learning

2. RELATED WORK

A literature review is a body of text that aims to review the critical points of current knowledge on and/or methodological approaches to a particular topic. It is secondary sources and discuss published information in a particular subject area and sometimes information in a particular subject area within a certain time period. Its

ultimate goal is to bring the reader up to date with current literature on a topic and forms the basis for another goal, such as future research that may be needed in the area and precedes a research proposal and may be just a simple summary of sources. Usually, it has an organizational pattern and combines both summary and synthesis.

The 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.

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.

In the past a strong reliance has been put on standard video surveillance in order to achieve this goal. This often creates a backlog of video data that must 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 have been implemented to help reduce the workload. Currently, auto regressive models have been used to better forecast criminal acts, but also have a list of shortcomings. It proposed a solution of using neural networks in combination with a Hybrid Deep Learning algorithm to analyze video stream data. Our system will be able to quickly identify and assess criminal activity which will in turn 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 can be applied to various 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.
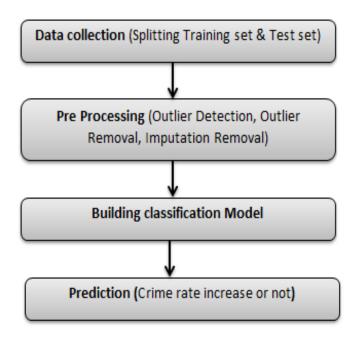
3. Overview of the system

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. Visualization of dataset is done to analyze 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. It have to find

Accuracy of the training dataset, Accuracy of the testing dataset, Specification, False Positive rate, precision and recall by comparing algorithm using python code.

Predictive modeling is the way of building a model that is capable of making predictions. The process includes a machine learning algorithm that learns certain properties. Classification model is to assign discrete class labels to particular data value as output of a prediction. Example of a classification model is - A pattern classification task in weather forecasting could be the prediction of a sunny, rainy, or snowy day. Pattern classification tasks can be divided into two parts, Supervised and unsupervised learning. In supervised learning, the class labels in the dataset, which is used to build the classification model, are known. In a supervised learning problem, it would know which training dataset has the particular output which will be used to train so that prediction can be made for unseen data. This analysis aims to observe which features are most helpful in predicting the crime rate and to see the general trends that may help us in model selection and hyper parameter selection. The goal is to classify whether the crime rate is high or low with to helpful for identifying the fraud areas in real time world. To achieve used machine learning classification methods to fit a function that can predict the discrete class of new input information about crimes. The repository is a learning exercise to:

➢ Apply the fundamental concepts of machine learning from an available dataset and Evaluate and interpret my results and justify my interpretation based on observed dataset.

➢ Create notebooks that serve as computational records and document my thought process and investigate applications of statistics for loan to analyses the data set.

Fig-2: Data Flow Diagram for Machine Learning Model

4. EXISTING SYSTEM

Latest technical developments in sophisticated tools of data analytics and visualization are helping the society in different ways to analyze the data of social relevance. One of such socially relevant activities is crime details of different demographic places. The analysis of the crime data will help decision making agencies to take precautionary steps to control the crime rate over demographic places. Advancements in the field of information technology, publicly available information and services, somehow help criminals to achieve their misdeeds and involve them in much serious crimes than earlier. As a result, crime rate is increasing with a very high rate in developed and under-developed nations. Based on 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 can be modified by using Machine Learning (ML) models for forecasting crime, as the data points will sufficiently increase to apply ML models. This can also increase the accuracy of the predictions. Further, statistical modeling's methods can also be clubbed with ML models and then calculate weighted accuracy for a district, this can make the solution more robust.

5. PROPOSED SYSTEM

Exploratory Data Analysis of crime rate Prediction
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 datasets. 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, smugglings etc.

6. 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.

7. DATA PROCESSING

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. The data which was collected might contain missing values that may lead to inconsistency. To gain better results data need to be preprocessed so as to improve the efficiency of the algorithm. The outliers have to be removed and also variable conversion need 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 the strongest among all. Some variables such as applicant income and co- applicant income are not significant alone, which is strange since by intuition it is 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.

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.

8. TESTING THE DATASET

Steps in testing the dataset, are the following,

1. 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.

2. 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. 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.

4. For example, to import any algorithm and train_test_split class from sklearn and numpy module for use in this program.

5. 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.

6. This method divides dataset into training and test data randomly in ratio of 67:33 / 70:30. Then we encapsulate any algorithm.

7. 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.
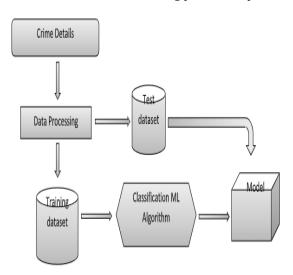
Fig-3: Architecture of Proposed Model

9. 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 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.
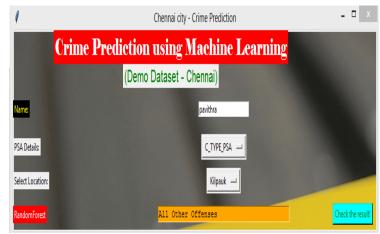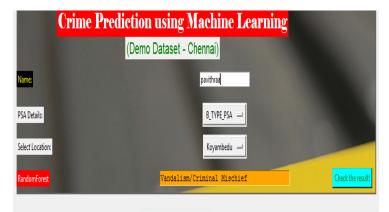
Fig-4: Result-1

Fig-5: Result-2

REFERENCES

[1]A Relationship between Fines and Violent Crimes, Authors: Samuel Smith, Simranjyot Singh Gill and Kedar Gangopadhyay, Issued on: 2018.

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

[3]Intelligent Crime Anomaly Detection in Smart Cities, Authors: Sharmila Chackravarthy, Steven Schmitt, Li Yang, Issued on: 2018.

[4]Sathyadhevan and Gangatharan (2014, August),Crime Analysis And Prediction using Data Mining.In Networks & Soft Computing(ICNSC), Issued on:2014.