CRIME TYPE AND OCCURRENCE PREDICTION USING MACHINE LEARNING ALGORITHM

Kanimozhi N [1] , Keerthana N V[2], Pavithra G S[3], Ranjitha G[4] ,Yuvarani S [5]

[1]*Assistant Professor,Kongu Engineering College*

[2]*Assistant Professor,Velalar College of Engineering and Technology*

[3,4,5]*UG Scholar, Kongu Engineering College*

[1]*kanimozhi6465@gmail.com*

[2] saranya.sowndarajan@gmail.com

[3]pavithrags.18bsr@kongu.edu

[4]ranjithag.18bsr@kongu.edu

[5]yuvaranis.18bsr@kongu.edu

***Abstract* - In this era of recent times, crime has become an evident way of making people and society under trouble. An increasing crime factor leads to an imbalance in the constituency of a country. In order to analyse and have a response ahead this type of criminal activities, it is necessary to understand the crime patterns. This study imposes one such crime pattern analysis by using crime data obtained from Kaggle open source which in turn used for the prediction of most recently occurring crimes. The major aspect of this project is to estimate which type of crime contributes the most along with time period and location where it has happened. Some machine learning algorithms such as Naïve Bayes is implied in this work in order to classify among various crime patterns and the accuracy achieved was comparatively high when compared to pre-composed works.**

Keywords: Crime, Analyse, Crime patterns, Kaggle, Estimate, Naïve Bayes, Accuracy

I. Introduction

Crime has become a major thread imposed which is considered to grow relatively high in intensity. An action stated is said to be a crime, when it violates the rule, against the government laws and it is highly offensive. The crime pattern analysis requires a study in the different aspects of criminology and also in indicating patterns. The Government has to spend a lot of time and work to imply technology to govern some of these criminal activities. Hence, use of machine learning techniques and its records is required to predict the crime type and patterns. It imposes the uses of existing crime data and predicts the crime type and its occurrence bases on the location and time. Researchers undergone many studies that helps in analysing the crime patterns along with their relations in a specific location. Some of the hotspots analysed has become easier way of classifying the crime patterns. This leads to assist the officials to resolve them faster. This approach uses a dataset obtained from Kaggle open source based on various factors along with the time and space where it occurs over a certain period of time. We implied a classification algorithm that helps in locating the type of crime and hotspots of the criminal actions that takes place on the certain time and day. In this proposed one to impose a machine learning algorithms to find the matching criminal patterns along with the assist of its category with the given temporal and spatial data.

II. Literature Survey

Crime are of different type that occurs at different locations around the various geographical location. Many research scholars have been suggesting a mechanism to analyse the relationship between crime and social variables that includes unemployed individuals, earning amount, level of education and so on.

Suhong Kim and Param Joshi [1] proposed two different machine learning models which is used for prediction, K nearest neighbour algorithm (KNN)

and decision tree approach. The accuracy obtained ranges between 39 to 44 percent when predicting crime patterns and finding the crime type. Benjamin Fredrick David. H [2] imposed a data mining technique that involves evaluating and inspect large pre-existing datasets in accordance to deliver more information. The extraction of new patterns is cross checked with predefined datasets available.

Shraddha S. Kavathekar [3] used association rule mining in predicting crimes. Some Machine learning algorithms including Deep Neural Network (DNN) and Artificial Neural Network (ANN) have been implied. A deep neural network works more accurately using the feature level dataset. Using DNN, entirely connected convolution layers has been used in building the prediction model, mainly for multi-labelled data classification. It was implemented using Tenserflow that is an API mainly designed for Deep learning technique with the dropout layers. These findings suggest that when there is more count of missing values, there is a need for pre-processing because crimes do not occur in the same manner but focuses on some particular areas. Artificial Neural Network [ANN] is based on the prognosis by trend analysis in solving problems. It comprises of enormous amount of processing constituent that works altogether in building a model. Chandy and Abraham [4] proposed a random forest classifier in extracting the features for data processing using cloud computing. The extracted features are request number, user identification, expiry time, time of arrival nd memory requirement. After feature extraction, the prediction of work load is done by using the trained data that has been perceived from the learning stage that allows to learn the details of the extracted features from user's request.

Rohit Patil, Muzamil Kacchi, Pranali Gavali and Komal Pimparia [5] suggests an Apriori algorithm for frequent patterns and the result obtained from K-means is used. Due to increase in crime rate over these recent years, system has to handle an enormous amount of data which requires more time to analyse them manually. Hence, advance machine learning approaches like K means clustering has been used. A literature survey on Spatial and Temporal Hotspot prediction of crime [6] proposed a study to categorize and evaluate the location and time of the crime hotspot detection techniques by performing (SLR) Systematic Literature Review. Fuzhan Nasiri, Zakikhani, Kimiya and Tarek Zayed [7] suggested a failure prediction model that helps in detecting the corrosion in the pipelines of gas transmission. Most of the prediction model depend absolutely on the experimental tests data or involving some of the limited historical data records. This helps in ignoring the corrosion from various geographical circumstances. Nikhli Dubey and Setu K. Chaturvedi [8] imposed pertinent analysis of data mining approaches for the detection of the impeding future crime. A Computational mechanism to classify the crime using machine learning techniques [9] proposed a malleable computational implementation tool to analyse the crime rate in a country helps in classifying cybercrimes. Hyeon-Woo Kang and Hang-Bong Kang [10] suggested a fusion method based on Deep Neural Network in predicting the criminal activities from the feature level data with sufficient parameters.

III. Existing System

In pre-work, the dataset obtained from the open source are first pre-processed to remove the duplicated values and features. Decision tree has been used in the factor of finding crime patterns and also extracting the features from large amount of data is inclusive. It provides a primary structure for further classification process. The classified crime patterns are feature extracted using Deep Neural network. Based on the prediction, the performance is calculated for both trained and test values. The crime prediction helps in forecasting the future happening of any type

of criminal activities and help the officials to resolve them at the earliest.

IV. Drawbacks

1. The pre-existing works account for low accuracy since the classifier uses a categorical values which produces a biased outcome for the nominal attributes with greater value.
2. The classification techniques does not suited for regions with inappropriate data and real valued attributes.
3. The value of the classifier must be tuned and hence there is a need of assigning an optimal value.

V. Proposed System

The data obtained is first pre-processed using machine learning technique filter and wrapper in order to remove irrelevant and repeated data values. It also reduces the dimensionality thus the data has been cleaned. The data is then further undergoes a splitting process. It is classified into test and trained data set. The model is trained by dataset both training and testing .It is then followed by mapping. The crime type, year, month, time, date, place are mapped to an integer for ensuring classification easier. The independent effect between the attributes are analysed initially by using Naïve Bayes. Bernouille Naïve Bayes is used for classifying the independent features extracted. The crime features are labelled that allows to analyse the occurrence of crime at a particular time and location. Finally, the crime which occur the most along with spatial and temporal information is gained. The performance of the prediction model is find out by calculating accuracy rate. The language used in designing the prediction model is python and run on the Colab – an online compiler for data analysis and machine learning models.

VI. Advantages

1. The proposed algorithm is well suited for the crime pattern detection since most of the featured attributes depends on the time and location.
2. It also overcomes the problem of analysing independent effect of the attributes.
3. The initialization of optimal value is not required since it accounts for real valued, nominal value and also concern the region with insufficient information.
4. The accuracy has been relatively high when compared to other machine learning prediction model.

VII. Module Description

1. Data pre-processing
2. Mapping
3. Naïve Bayes classification
4. Crime prediction
5. Evaluation
6.

A. Data Pre-Processing

Data obtained from the open source must be first pre-processed in order to overcome unnecessary violations. The dataset has been chosen for Denver city with enormous amount of crime data over six years. The machine learning technique filter and wrapper is implied to find the missing integral in specified attribute values. Data cleaning play a vital in training a prediction model and also in the performance of the commenced process.

Filtering the instance and removal of irrelevant context from datasets are done. The filtering methods contributes in measuring the significance of the features. The correlation with the dependent values is considered in the feature selection. The wrapper method imposed is used in measuring how useful is the feature subset by training a prediction model on it actually. The data after pre-processing is split into test and trained attributes.

INCIDENT_ID	OFFENSE_ID	OFFENSE_CODE	OFFENSE_CODE_EXTENSION	OFFENSE_CATEGORY_ID
2018869789	2018869789239900	2399	0	theft-other
202111218	202111218570700	5707	0	criminal-trespassing
20176005213	20176005213239900	2399	1	theft-bicycle
201960122400	201960122402308800	2308	0	theft-from-bldg
2018861883	201886188350160 0	5016	0	violation-of-restraining-order

Table 1. Dataset Collection

FIRST_OCCURRENCE_DATE	LAST_OCCURRENCE_DATE	REPORTED_DATE
12/27/2018 3:58:00 PM	NIL	12/27/2018 4:51:00 PM
01-06-2021 9.20.00 PM	NIL	01-07-2021 12.23.00 AM
06-08-2017 1.15.00 PM	06-08-2017 5.15.00 PM	06-12-2017 8.44.00 AM
12-07-2019 1.07.00 PM	12-07-2019 6.30.00 PM	12-09-2019 1.35.00 PM
12/22/2018 8:15:00 PM	12/22/2018 8:31:00 PM	12/22/2018 10:00:00 PM

Table 2. Crime Dataset with occurrence date and time

NEIGHBORHOOD_ID	IS_CRIME	IS_TRAFFIC
montbello	1	0
Gateway-green-valley-ranch	1	0
wellshire	1	0
belcaro	1	0
cherry-creek	1	0

Table 3. Neighbourhood dataset

B. Mapping

The crime features such as crime type, the date on which the crime has been occurred including the time of occurrence are first segregated. It is then mapped to an integer for easy labelling. The labelled details are further analysed and used are used in graph plotting. Python is chosen as programming language in implementing the proposed work since it is well suited for machine learning process. The package matplotlib is imported in order to plot the graph to show the occurrence of the criminal activities. The crime which occurred the most can be plotted in the graph which contributes for further prediction process.

NEIGHBORHOOD_ID	IS_CRIME
montbello	1
gateway-green-valley-ranch	2
wellshire	3
belcaro	2
cherry-creek	2

Table 4. Mapping crime type

NEIGHBORHOOD_ID	IS_CRIME	CRIME_OCCURENCE_MONTH
montbello	1	6
gateway-green-valley-ranch	2	10
wellshire	3	3
belcaro	2	1
cherry-creek	2	6

Table 5. Finding crime occurrence type and month in a dataset

CRIME_OCCURENCE_DAY	CRIME_OCCURENCE_TIME	CRIME_OCCURENCE_YEAR
3	6	3
3	3	4
5	5	3
2	5	5
4	5	4

Table 6. Finding crime occurrence day, time, year count in dataset

C. Naïve Bayes Classification

The reason behind the application of Naïve Bayes is that crime prediction usually concerns with the temporal and spatial data. The independent effect among the attribute values is first analysed since the selected crime attributes possess an independent effect upon them. They are used in creating a model by providing a training using crime data that are related to robbery, burglary, murder, sexual abusing, armed robbery, chain snatching, gang rape and highway robbery. Some of the extended techniques of Naïve Bayes has been implied.

1. Gaussian Naïve Bayes is related to real valued attribute selection. It is otherwise stated as normal distribution that is done by calculating the standard deviation and mean from the trained data.
2. Multi-nominal Naïve Bayes is applied for multiple classifier that corresponds to the categorical features in the trained value.
3. Bernouille Naïve Bayes is used for the working of independent feature effects of the selected attributes for crime prediction.

D. Crime Prediction

The expected crime type is predicted by extending the supported crime features. The features are then applied to nominal values. It could be explained clearly by taking a single tuple as an instance.

Considering a tuple:

1. {Gateway town, 20th October 2020 , 2: 30 PM, Friday} => {Larceny – a crime involves the theft of a particular's property}

Considering probable occurrence based on the feature extracted:

1. {Gateway town} => {Theft has occurred}
2. {October} => {Theft has occurred}
3. {2020} => {Theft has occurred}
4. {2:30 PM} => {Theft has occurred}
5. {Friday} => {Theft has occurred}

The independent occurrence has been formed and the conditional probability is calculated. By doing so, we could predict the crime type.

Usage of symbols:

1. m represents Month
2. t represents Time
3. a represents Area
4. d represents Day
5. y presents Year
6. c represents Type

The Formula using the chain in order to find the conditional probability:-

P(c|m, y, a, t, d) = [P(m|c, y, a, t, d) * P(y|c, a, t, d) * P (t|d, c) * P(d|c) * P(c)] / [P(m|y, a, t, d) * P(y|a, t, d) * P (a|t, d) * P(t|d)]

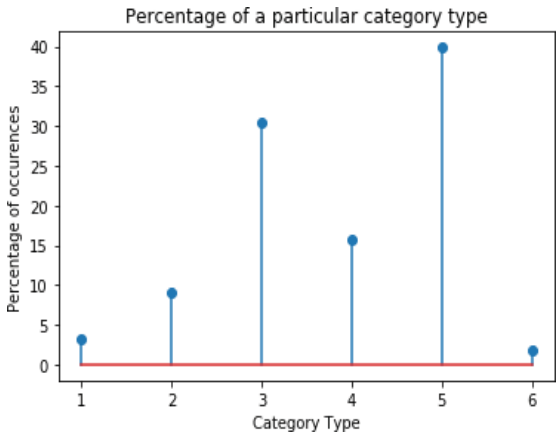

Fig 1. Plotting the highest crime type

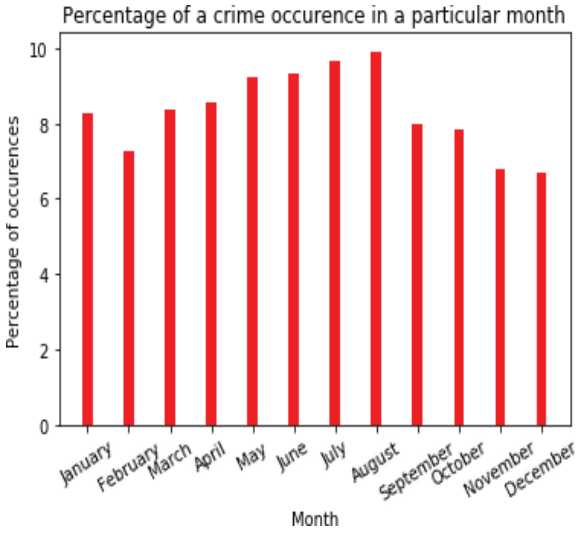

Fig 2. Plotting the highest occurrence month

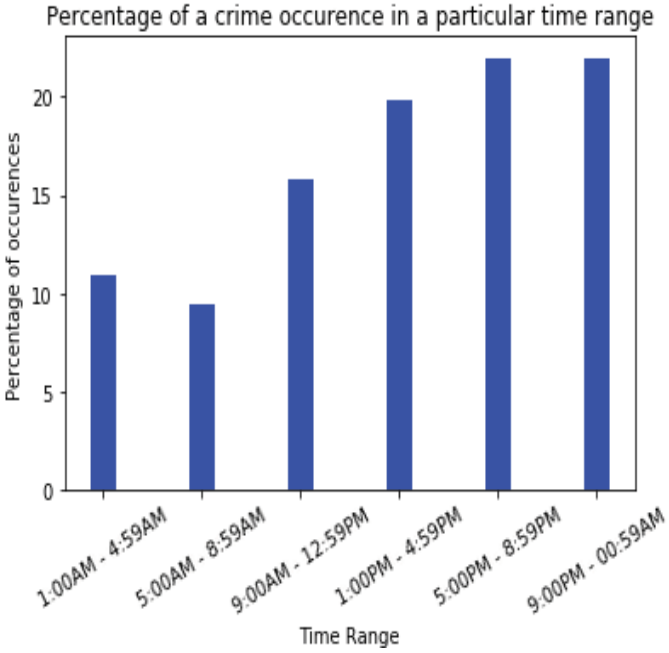

Fig 3. Plotting the highest occurrence time range

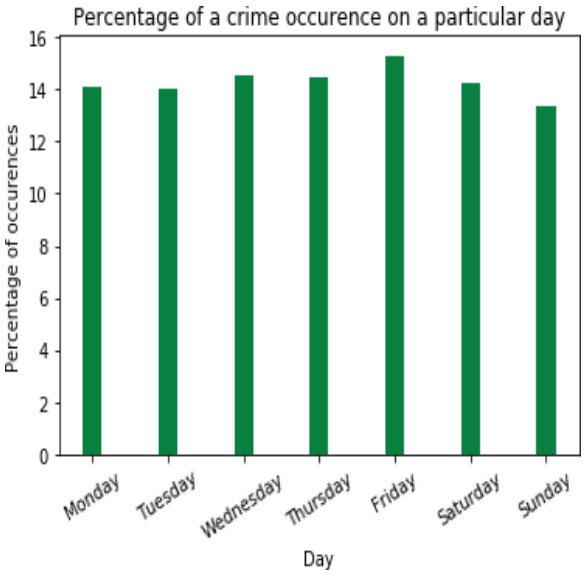

Fig 4. Plotting the highest occurrence day

F. Evaluation

The performance of the implied prediction is then evaluated in order to achieve a high degree of accuracy when compared to the pre-existing model used. The training is done with cross validation that helps in training the data on different set of training data. It will evaluate the accuracy for overall splits in the cross validation implied. In python, in order to calculate the value of accuracy we need to pass the data arguments such as model name, target set and cv that helps in signifying the split occurrence. Finally, the mean and the standard deviation of the average precision is calculated. The accuracy of 93.07% has been achieved who gives a great increase in compared to existing prediction models.

EVALUATION METRICS	CROSS VALIDATION
Accuracy	93.07%
Precision	92.53%
Recall	85.76%
F1 score	92.12%

Table 7. Performance measure for Naïve Bayes classifier

VIII. Conclusion

In this paper, the difficulty in dealing with the nominal distribution and real valued attributes is overcome by using two classifiers such as Multi-nominal NB and Gaussian NB. Much training time is not required and serves to be the best suited for real-time predictions. It also overcomes the problem of working with continuous target set of variables where the existing work refused to fit with. Thus the crime that occur the most could be predicted and spotted using Naïve Bayesian Classification. The performance of the algorithm is also calculated by using some standard metrics. The metrics include average precision, recall, F1 score and accuracy are mainly concerned in the algorithm evaluation. The accuracy value could be increased much better by implementing machine learning algorithms.

IX. Future Work

Though it overcomes the problem of the existing work, it has some limitations. In the situation of absence of class labels, then the probability of the estimation will be zero. As a future extension of the proposed work, the application of more machine learning classification models proves to increase accuracy in crime prediction and will enhance the overall performance. It helps in providing a better study for the future improvement by taking the income information into consideration for neighborhoods places in order to foresee if any relationship between the income levels of a particular in the neighborhood places and their crime rate.

X. References

[1] Suhong Kim, Param Joshi, Parminder Singh Kalsi,Pooya Taheri, “Crime Analysis Through Machine Learning”, IEEE Transactions on November 2018.

[2] Benjamin Fredrick David. H and A. Suruliandi,“Survey on Crime Analysis and

Prediction using Data mining techniques", ICTACT Journal on Soft Computing on April 2012.

[3] Shruti S.Gosavi and Shraddha S. Kavathekar,"A Survey on Crime Occurrence Detection and prediction Techniques", International Journal of Management, Technology And Engineering , Volume 8, Issue XII, December 2018.

[4] Chandy, Abraham, "Smart resource usage prediction using cloud computing for massive data processing systems" Journal of Information Technology 1, no. 02 (2019): 108-118.

[5] Learning Rohit Patil, Muzamil Kacchi, Pranali Gavali and Komal Pimparia, "Crime Pattern Detection, Analysis & Prediction using Machine", International Research Journal of Engineering and Technology, (IRJET) e-ISSN: 2395-0056, Volume: 07, Issue: 06, June 2020

[6] Umair Muneer Butt, Sukumar Letchmunan, Fadratul Hafinaz Hassan, Mubashir Ali, Anees Baqir and Hafiz Husnain Raza Sherazi, "Spatio-Temporal Crime Hotspot Detection and Prediction: A Systematic Literature Review", IEEE Transactions on September 2020.

[7] Nasiri, Zakikhani, Kimiya and Tarek Zayed, "A failure prediction model for corrosion in gas transmission pipelines", Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, (2020).

[8] Nikhil Dubey and Setu K. Chaturvedi, "A Survey Paper on Crime Prediction Technique Using Data Mining", Corpus ID: 7997627, Published on 2014.

[9] Rupa Ch, Thippa Reddy Gadekallu, Mustufa Haider Abdi and Abdulrahman Al-Ahmari, "Computational System to Classify Cyber Crime Offenses using Machine Learning", Sustainability Journals, Volume 12, Issue 10, Published on May 2020.

[10] Hyeon-Woo Kang and Hang-Bong Kang, "Prediction of crime occurrence from multi-modal data using deep learning", Peer-reviewed journal, published on April 2017.