CRIME PREDICTION AND POLICING SYSTEM

Riya Srivastava[1], Shikha Nigam[2], Kritika Srivastava[3], Mahima Verma[4], Mr. Ashutosh Kumar Rao[5]

[1-5]*Department of Computer Science and Engineering*
Institute of Technology and Management Gorakhpur
Uttar Pradesh, India

Abstract

This paper deals with the high crime rate and its adverse impact on society by introducing machine learning algorithms to predict future crime hotspots based on previous crime data. It empowers the law enforcement for better policing the area. It helps in utilizing available resources in a more efficient way. Therefore, the method of crime prediction is of paramount importance, and it identifies future crimes and reduces the number of crimes. In this paper, we have designed a model to predict future crime incidents and also predict what type of crime may occur in a given area. First, we analyze how certain crime features like given a date, time and location of crime. Second, we discuss several analytics techniques we used to find meaning in our data, classification models like naïve Bayes, Logistic Regression and Decision Tree. Finally, we select the best model for foreseeing crime type and seriousness and location of the crime.

Keyword: Machine Learning, Naïve Bayes, Logistic Regression, Decision Tree, Crime Prediction, Predictive Policing.

1. INTRODUCTION

Throughout world any society that has flourished has one thing in common i.e., lesscrime rate. It is essential for a society to be crime free in order to flourish/grow.
In order to reduce crime rate there is a need for efficient policing system that is equipped with advanced technologies, so that they tackle with ongoing criminal activities in the area. (Explaining how crime affects the society badly) crime or criminal offence is an act harmful not only to some individual but also to a community, society, or the state.

Crime has three main elements:

- Crime's work. One hidden assumption is the definition of law.
- The offenses thus defined are prosecuted by the agents of the State/Government and not by private persons.
- The result of a conviction is a punishment, not compensation (if any) to the injured party.

Hinder developments in a society. It leads to the death of people and defacement of their property. It increases the costs of the living within the society.

Crime in the society can be reduced by implementing two ways.

1- To make the society aware of crime in their surroundings, so that they couldprotect themselves before some crime happens against them.
2- By equipping the police with better policing system. It will be very helpful ifthey could predict the crime before it occurs by analyzing the crime patternin the area.

This will also reduce the huge load on judicial system and courts that leads to long Delay in prosecution. Courts won't be overloaded with cases. It aims to propose an improved technology for criminal prediction to provide the

police with rich sources of intelligence and information that can be used to predict and prevent criminal act. The main goals is to provide accurate results quickly which can help the government officials to the predict criminals and crime prone areas. This paper therefore aims to apply Naïve Bayes technique that has been developed for Crime-hotspot Prediction.

2. LITRETURE REVIEW

Tayla proposed an approach for crime detection and criminal identification in India using data mining techniques [1]. Their approach is divided into six modules, which include: data extraction, data pre-processing, clustering, Google Map representations, classification, and WEKA implementation.

They used k-means clustering for analyzing crime detection that generates two crime clusters based on similar crime attributes. Then KNN classification was used for criminal identification and prediction, while WEKA was used for crime validation of their results. The works related to crime analysis cover visualization methods [2, 3, 4], statistical methods [5, 6, 7], supervised [8, 9] and unsupervised [9, 10] learning techniques.

The main purpose of this work is to identify and classify text into natural image. Here the system traces the text and finds the connected regions, chaining them together in their relative positions. Uses text classification engine to filter out series with low classification confidence scores. [11]

Authorized entities only provide limited access to incident level data, such as the date and location of events and criminal information, due to the potential for misuse of published information and breach of personal fact. For example, the Atlanta Police Department hides criminal-related information that has to be kept confidential for security issues.

Ajay Bazil Issac, C.P. Chaithanya, N. Manohar, C.P. Chaithanya, C.P. Chaithanya, C.P. Chaithanya, C.P. Chaithany proposed a text detection technique for identifying places in a photograph where there is text. Text detection and categorization in natural images are critical for various types of computer vision applications, such as optical character recognition, distinguishing between machine and human inputs, and spam detection.

In a Chinese town, Chen et alia proposed an auto-regressive coordinated shift toward a normal model for here and now crime expectations. For forecasting future events, the model used a well-known time series analysis technique.

In Salinas, Shingleton JS et al. introduced a method that used a multivariate analysis technique to predict three types of crimes: cruelty, arson, and homicide. They employed a conventional statistical procedure in California, including negative binomial, Poisson regression, and regression models. The results of the experiment performed for all three models shows that they perform similarly. Poisson regression models and Negative binomial regression, on the other hand, need the assumption that data traces the Poisson distribution. If the data cannot be fitted using a Poisson distribution, common minimum squares will suffice.

Shiju Sathyadevan, Deva M.S, suggested that the rate of crime is steadily growing. Because crime is neither random nor systematic, it is impossible to foresee. Modern technologies and high-tech approaches also aid criminals in committing their crimes.

Policymakers and law enforcement organizations all over the world are struggling to discover effective measures to regulate criminal networks, according to PAC Duijn, V Kashirin. Both network resilience and network topology are known to affect the success of disruption methods. Due to the secrecy with which these illegal criminal networks operate, data-driven understanding of the success of various criminal network disruption techniques is scarce.

K. Zakhir Hussain proposed it. Crime analysis is a branch of criminology, which is a work that implies investigating and detecting crimes as well as their connections to criminals. Because of the complexity of relationships between a large number of datasets, criminology is an ideal topic for applying data mining techniques. The initial step in generating further analysis is to identify the characteristics of crime. The knowledge gathered through data mining is a very effective tool for detecting violent criminal behavior. The goal is to apply data mining and a simulation model to try to capture years of human experience in computer models.

According to Nishat Sharma, criminal activities influence the socioeconomic growth and quality of life in every region of the world. As a result, several governments are concerned about it, and they are employing various technological technologies to address the problem.

Kianmehr and Alhajj et al. [9] developed a computational system for use in, St. Louis, Missouri, and, Columbus, Ohio that used SVM with the k-means for clustering to forecast criminal hot spots. Due to the lack of negative instances in a variety of dataset categories contributes to this problem when using the k-means for clustering and dataset division into small sets.

Based on the varied crime rates, these divided sets are identified as either a non-hotspot class or a hotspot class with this system. A dataset is categorized as a positive category (hot spot) if the criminal offense rate records are higher than a pre-defined rate, or as a negative category if the rate is lower (non-hot spot). Their framework can be used for criminal hot spot forecasting and other types of forecasting tasks.

SVM was employed by Wang et alia to forecast criminal lapsing. Training a logistic regression and a multilayer neural network was used to analyze the SVM. Even though the SVM and multilayer impartial system both outperformed the strategic relapse plan, they both outdid it. As a result, the authors combined the three models' expectations and found the best execution.

"Crime Analysis and Prediction Using K-Means Clustering Technique" by Wasim A. Ali, Husam Alalloush, and Manasa K.N proposes a framework for forecasting the likelihood of a crime occurring in a city by analyzing the dataset of crime incidents and visualizing the findings on a Google map for better comprehension. It foresees the areas where the most offense will occur. This study is carried out using the K-means clustering method, which groups related objects into small clusters. This technology might be used by officers to predict criminal cases and take appropriate action to capture perpetrators.

Sivanagaleela B [2] proposed a crime analysis methodological approach for detecting crime regions called "Crime Analysis and Prediction" utilizing Fuzzy C-Mean Clusters. This technique is mainly concerned with determining where the crime might occur. The fuzzy C-Means algorithm is used to cluster the crime-related data for total knowable crimes such as abduction, burglary, murder, theft, robbery, cheating, crimes against women, etc. in the existing Naive Bayes classification system. Based on the clustering technique, this concept is effective for identifying criminal hotspots. Because crime patterns are not static, they are also detected, which aids investigators in solving the crime rate. Because it takes less time, the crime-prone zones are discovered using the fuzzy clustering technique.

Crime analysis and prediction performed by Dr. Sarvanaguru RA and Alkesh Bharati. K [3] proposed a method that focuses on predicting crime using the attributes present in the dataset using machine learning. The information was gathered from official sites. We can forecast the crime type that might occur in a given location using a machine learning system with Python as the core. The goal is to develop a predictive model. The training and testing dataset would be used to validate the training. Depending on the accuracy, a better algorithm can be used to build the model. For crime prediction, the (KNN) K-Nearest Neighbor classification and other algorithms will be employed.

"Crime Prediction Using Machine Learning," proposed by Rahul Shah [4], and "Crime Prediction Using Machine Learning," proposed by Akash Kumar and Aniket Verma. Many experts have worked on crime rate prediction and analysis using different methods such as k-means clustering, KNN, Naive Bayes, Fuzzy c algorithm, and others. Gandhali Shinde developed a scheme "Crime Prediction Using K-Nearest Neighboring Algorithm" [5], and many experts have worked on crime rate prediction and analysis using different methods such as k-means clustering, KNN, Naive Bayes. The Naive Bayes Algorithm was determined to be one of the best strategies for predicting future crimes among all the methods tested. As a result, they developed a system that combines Bernoulli NB with linear regression.

Implementation of the "prediction-led policing business process" (Perry et alia 2013). PRECOBS are a type of software that focuses on "closest prediction." This implies that only house burglaries that may happen as a result of the initial event are expected. Only a small amount of data, largely from police records, is required for the calibration of the software as well as for predictions. As a response, collecting data is a common aspect of police activity. The approach to prediction was understandable to the police officers who used the software, and they saw it as a useful addition to their arsenal.

Countering criminal activity has been a primary concern for all governments around the world. Many studies have been conducted to identify signs and evaluate crime patterns in an attempt to prevent them from happening again.

A study has been done to see whether there is a link between criminal activity and socioeconomic characteristics.

Some researchers examined mobile network infrastructure and concluded that it may be utilized to anticipate crime hot locations in London.

Sojayee et. al. used a several techniques, with the KNN algorithm outperforming the others, with a prediction accuracy of 89.50 percent.

Wang et alia [16] suggested an algorithm for detecting criminal trends perpetrated by an individual or a group of individuals.

Sadhana and Sangareddy [13] suggested that social networks may be misused as a prospective indicator of criminal behavior. They used Twitter data with sentiment analysis to forecast crime in real-time.

Toppireddy, Saini, and Mahajan use geo-location information to visualize and predict the next likely crime location, whereas two prior studies used document attributes as the data for analyzing and predicting crime..

Hakim et. al. used TF-IDF to create a lexicon to classify Indonesian news [8]. They acquired 7500 news items, each of which is divided into 500 categories.

Wongso et al. [9] attempt to combine various categorization algorithms. They mix (SVD) Singular Value Decomposition and three forms of Nave Bayes with TF-IDF implementation in their research. According to their findings, combining Multinomial Nave Bayes and TF-IDF yields 98.2 percent accuracy, which is the greatest among other combinations.

The majority of prior works on crime analysis have only focused on the most accurate and efficient method of crime prediction. Our research will be focused on determining the best accurate data mining strategy for acquiring criminal data from various categories on the internet. As a result, future academics and police officers will have access to more relevant and up-to-date crime data. In comparison to earlier studies, this study presents a comparison of the approaches used to classify Indonesian news.

There are numerous crime data mining approaches, including regression, classification, and clustering, according to studies. Depending on the kind of data provided, the training set, and the testing set, each method is employed in different contexts.

This section outlines some of the successful crime prediction systems developed by other researchers.

Wang et alia created the (HOT)Hotspot Optimization Tool, a crime-hotspot mapping tool that applies spatial data mining to the field of hotspot mapping. It can capture the differences between two classes in a spatial dataset [3]. HOT was able to precisely map the crime hotspots in a north-eastern city in the United States as part of the study.

3. MATERIAL AND METHODS

In this project, mainly three machine learning classifiers are used such as Logistic Regression (LR), Naive Bayes (NB) and Decision Tree (DT). These are implemented by using Python language.

3.1. Generating Dataset:

In this section, synthetic dataset is generated from which training and testing dataset is drawn. Certain parameters that are used in analyzing crime patterns and predicting future possible crime-prone hotspots are: type of crime, location of crime incident, date and time of crime committed.

Synthetic dataset is generated in following three steps:

Step 1- creating geographical map of entire area that is considered for reducing crime.
Step 2- plotting the crime incident area over the geographical map.
Step 3- distribution of crime type and location on the map.

Fig1. Framework of crime prediction and policing system

3.2. Classification techniques

Classification is the process of recognizing, understanding, and grouping ideas and objects into preset categories or "sub-populations." Classification is a form of "pattern recognition," in which classification algorithms applied to the training data to find similar patterns (similar word or emotion, number sequence, etc.) in a future set of data.

The classification techniques that are considered are as follows:

1- Logistic Regression
2-Naive Bayes
3-Decision Tree

Logistic Regression: Logistics regressions are the calculation used to predict a binary outcome: either the something happens, or does not. This can be exhibited as Yes/No, Pass/Fail, Alive/Dead, etc. A logistics regression is the statistical models that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist.

A logistic regression is a generalized linear model with a canonical link function. In the terms of easy computations, a logistics regression is the best's model among generalized linear models for binary response data. Interpreting the result of this model is very simple and logical.

Naive Bayes: It calculates the probability of whether a data point falls under a certain category. Here we present the naive Bayesian model that has been formalized for the criminal prediction problem. Naive Bayesian Classifier is a simple probabilistic classifier that works by applying Bayes' theorem with Naive assumptions about feature independence. The Naive Bayesian Classifier has proven to be quite useful in the modeling of complex real-world problem situations.

In essence, naive Bayes is a conditional probability model: a problem example to be classified is given, represented by the vector X (x1... xn), which represents some n features (independent variables). Is; It assigns a possible p (ck

|x1... xn) to each of the possible outcomes for this example.

On top of the description the problem with a variety of features n is huge or if an attribute can take a large number of values, it is not possible to base such a model on probability of the table. We tend to design techniques for additional permissiveness on these lines. Using the naive Bayes hypothesis, the restrictive probability can be decomposed as:

P (Ck |X) = P (Ck). P (x| Ck) /P(x)

Decision Tree: It is a simple yet widely used classifier with three types of nodes. Root and other nodes hold test conditions for features, and each leaf node is assigned with a class label. Decision trees are one of the most popular and powerful tools for classification and prediction. It is the structure is likes a trees. It is a graphical representation to get all possible solutions to a problem/decision based on the given conditions. Decision trees are used either to calculate the probability that a given record belongs to each category's or to the classify record (which is done by assigning records to the most similar class).

3.3. Training and Testing:

In this step, after validating the assumptions of the algorithm we have chosen. The model is trained based on the given training sample. After training, the performance of the model is checked on the basis of the error and accuracy. Finally, the trained model is tested with some unseen data and the performance of the model is checked on the basis of various performance parameters depending on the problem.
In this step, after validating the assumptions of the algorithm we have chosen. The model is trained based on the given training sample. After training, the performance of the model is checked on the basis of the error and accuracy. Finally, the trained model is tested with some unseen data and the performance of the model is checked on the basis of various performance parameters depending on the problem.

4. PERFORMANCE EVALUATION

Accuracy is defined as the ratio of the total number of predicted transactions that are correct.
Accuracy=(TP+TN)/(TP+FP+TN+FN) (1)

Sensitivity The proportions of the positives observed values correctly predicted as positive.
It is also called as True Positive Rate (TPR)

Sensitivity (Recall) = (TP)/(TP+FN) (2)

Specificity is defined as; with how much accuracy is the negative (legitimate) case are classifieds and, in our case, it gives the accuracy on prediction of legitimate transactions classification. It is also called as True Negative

Rate (TNR)

Specificity= (TN)/ (FP+TN) (3)

Precisions, the proportions of positives (fraud) prediction that is actually correct

Precision= (TP) / (TP+FP) (4)

F-measure gives the accuracies of the test which mean that it is give the accuracies of experiments performed. It is uses the both precision and recall to compute its value. The best value for f1 score is considered at value.

F-measure= 2. (*Precision. Recall*) / (*Precision + Recall)* (5)

Area under the curve (AUC) AUC represent the degrees or measures of the separability that is how much model is the capable of differentiating between the classes

AUC= (Sensitivity + Specificity)/2 (6)

5. EXPERIMENTAL RESULT

S No.	Name of the method system	Performance Metric	Performance Value
1	Decision Tree	Accuracy of prediction	58.1
2	Naïve Bayes	Accuracy of prediction	51.6
3	Logistic Regression	Accuracy of prediction	67.8

6. CONCLUSION

This research work was done to compare different algorithms of machine learning and identify the most efficient algorithms. Logistic regression demonstrated optimal performance for all test cases compared to Naive Bayes and Decision Trees. It exhibits better F-measurement, sensitivity, precision and specificity than Naïve Bayes and Decision Tree algorithms. Logistic regression and Naïve Bayes being supervised machine learning.

Algorithm gives better results. Crime is a common social problem that affects the quality of life and economic development of the society. With the rise in crimes, law enforcement agencies are demanding advanced systems and new approaches to improve crime analysis and better protect their communities. With the increases in crime, laws enforcement agencies are the continuing to the demand.
Advanced systems and new approaches to the improve crime analytics and better protect their communities. This work proposes a crime event foresee strategy by comparing different model.

7. REFERENCE

[1]- D. K. Tayal, A. Jain, S. Arora, S. Agarwal, T. Gupta, and N. Tyagi, "Crime detection and criminal identification in India using data mining techniques," Ai & Society, vol. 30, pp. 117-127, 2014.

[2]- Brunsdon C, Corcoran J, Higgs G (2007) Visualizing space and time in crime patterns: a comparison of methods. Compute Environ Urban Syst 31:52–75

[3]- Vural MS, Go¨k M, Yetgin Z (2013) Generating incident-level artificial data using GIS based crime simulation. In: International conference on IEEE electronics, computer and computation (ICECCO' 2013). Ankara, pp 239–242

[4]- Xiang Y, Chau M, Atabakhsh H, Chen H (2005) Visualizing criminal relationships: comparison of a hyperbolic tree and a hierarchical list. Decis Support Syst 41:69–83

[5]- Tollenaar N, Van Der Heijden PGM (2013) which method predicts recidivism best? a comparison of statistical, machine learning and data mining predictive models. J R Stat Soc Ser A Stat Soc 176:565–584

[6]- Enzmann D, Podana Z (2010) Official crime statistics and survey data: comparing trends of youth violence between 2000 and 2006 in cities of the Czech Republic, Germany, Poland, Russia, and Slovenia. Eur J Crim Policy Res 16:191–205

[7]- Holst A, Bjurling B (2013) A Bayesian parametric statistical anomaly detection method for finding trends and patterns in criminal behaviour. In: Intelligence and security informatics conference (EISIC) 12–14 Aug. 2013. Uppsala Sweden IEEE, pp 83–88

[8]- Jain LC, Seera M, Lim CP, Balasubramaniam P (2014) A review of online learning in supervised neural networks. Neural Comput Appl 25:491–509

[9]- Corsini P, Lazzerini B, Marcelloni F (2006) combining supervised and unsupervised learning for data clustering. Neural Comput Appl 15:289–297

[10]- Mackenzie MD (1995) CDUL: class directed unsupervised learning. Neural Comput Appl 3:2–16

[11]- Ashish Sharma, Dinesh Bhuriya, Upendra Singh. "Survey of Stock Market Prediction Using Machine Learning Approach", ICECA 2017.