Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 9, September : 2023

MACHINE LEARNING ALGORITHMS FOR PREDICTING CRIMINAL ACTIVITY NATURE AND FREQUENCY

[#1]PEDDI RAMYA,

[#2]Y.SUSHEELA, ***Assistant Professor,***

[#3]Dr.V.BAPUJI, ***Associate Professor& HOD,***

Department of Master of Computer Applications,

VAAGESWARI COLLEGE OF ENGINEERING, KARIMNAGAR, TELANGANA

Abstract: The problem of crime is one of the most pressing issues facing modern society. It is the single most pervasive and powerful force in today's society. And it's a widespread social problem. As a result, crime prevention must be given top importance. Analysis of criminal cases needs to be done methodically. The analysis is essential in spotting and stopping illicit acts. The investigational patterns and crime trends are easier to spot after conducting this research. The primary purpose of this study is to analyze the efficiency of police work in solving crimes. The software was developed with the express purpose of discovering patterns in criminal activity. A criminal's prognosis is presented in the study based on the inferences drawn from the crime scene. In this work, we describe the approach taken to make predictions about the age and gender of perpetrators. Two major components of crime forecasting are presented in this study. The perpetrator's age and gender should be taken into account. In unsolved cases, the parameters used include study of multiple elements such year, month, and weapon used. The objective of the study's quantitative section is to count the number of open criminal cases. The task of prediction requires the generation of a detailed description of the offender's and victim's ages, sexes, and relationship dynamics. Kaggle provided the dataset used for this study. Predictions are made using a combination of multi-linear regression, KNeighbor's classifier, and neural networks. Machine learning techniques were used in the building and evaluation of the model.

Keywords: Crime Prediction, KNN, Decision Tree. Multilinear Regression; K-Neighbors Classifier, Artificia l Neural Networks.

1. INTRODUCTION

It is the act itself that defines a crime. There has been a breach of trust. The behavior in question is illegal. Law enforcement faces formidable obstacles when tasked with uncovering and evaluating underground criminal activity. Furthermore, a wealth of data pertaining to the occurrence is accessible. Therefore, specific methods ought to be able to aid the probe. Therefore, the recommended methodology should help bring about a positive outcome to the criminal incidence.

The application of machine learning methods can improve crime analysis and forecasting. Regression strategies are available via the machine learning methodology. The use of classification strategies aids in accomplishing the study's primary goal. Regression methods, particularly multilinear regression, are a type of statistical instrument frequently used in data analysis. Using this method, you can easily evaluate the connection between any two numbers. The dependent variable values are predicted from the independent variable values using this method. Classifiers can be developed using a wide variety of approaches, such as the K-Nearest Neighbor classifier. Multiclass target variables are classified using classifiers. When

neural networks are used, accuracy is greatly improved. The neural network has an input layer and an output layer that are both heavily interconnected.

Predictions are made using the aforementioned algorithms about the perpetrator's physical characteristics, including their gender, age, and connection to the incident. As a result, the model is supposed to make the police investigation easier. As a result, it helps provide closure to murder investigations. This paper serves as a model for similar works. An electronic version is available for download from the conference website. If you have any questions about the paper requirements, please get in touch with the conference publications committee via the contact information provided on the conference website. Your final work submission instructions can be found on the conference website.

2. LITERATURE SURVEY

Many kinds of criminal activity can be found in many different places. Numerous academics have proposed a framework for investigating links between crime and economic factors like unemployment, income, and education. Two machine learning models created by Suhong Kim and Param Joshi are the K-nearest neighbor algorithm (KNN) and the decision tree technique. Between 39% and 44% accuracy is achieved in identifying crime types and forecasting crime patterns. Franklin Fredrick is the person in question.

David H used a data mining technique, which entails the analysis of massive existing databases, to improve the dissemination of information. Validating the discovery of new patterns requires a comparison to established data sets. To foresee criminal behavior, Shraddha S. Kavathekar employed association rule mining. Two examples of machine learning techniques that have been brought up are Deep Neural Networks (DNNs) and Artificial Neural Networks (ANNs). Using a feature-level dataset improves a deep neural network's accuracy. In order to identify multi-labeled input, a prediction model was created using deep neural networks (DNN) and fully linked convolutional layers.

Tenserflow, an API created for the express purpose of facilitating the use of Deep Learning techniques that make use of dropout layers, was employed in the system's development. Since crime does not occur randomly but rather shows concentration in certain locations, the results imply that pre-processing is crucial in circumstances when there is a higher number of missing data. For the purposes of both issue solving and future prediction, Artificial Neural Networks (ANNs) rely heavily on trend analysis. A large number of processors in the system work together to generate the model. For feature extraction in cloud computing-based data processing,

Chandy and Abraham suggested using a random forest classifier. Information like the request number, user ID, expiration time, arrival time, and memory requirements can be accessed. Workload prediction is performed utilizing a dataset that has been taught and obtained during the learning phase, after feature extraction has been performed. The system can then learn the details of the derived features in response to user input.

The Apriori technique is proposed by Rohit Patil, Muzamil Kachi, Pranali Gavali, and Komal Pimparia to find repeating structures. The K-means algorithm's findings are also factored into the evaluation. The increasing number of criminal activity over the past few years has placed a strain on the current system, making it more time-consuming to manually analyze large amounts of data. This has led to the use of sophisticated machine learning techniques, such as K-means clustering. The purpose of this research is to conduct a systematic literature review (SLR) to identify and assess the strategies used to detect

crime hotspots, with an emphasis on pinpointing the best times and locations to intervene. An in-depth analysis of previous studies on the topic of foreseeing crime hotspots across space and time led to the recommendation of employing a Systematic study Review approach. A model for forecasting failures in gas transmission pipelines was developed by Nasiri, Zakikhani, Kimiya, and Zayed.

The focus of the model was on the detection of corrosion. Most prediction models rely heavily on sample data from experiments or small samples of the past. Because of this, corrosion brought on by different environmental conditions can be disregarded. Nikhli Dubey and Setu K. Chaturvedi did extensive research into many data mining algorithms to efficiently identify possible future crimes. The use of a computational process built on machine learning methods allows for the classification of cybercrimes. This system is useful because it can be easily implemented on a computer and used to study the incidence of cybercrime in a country. To predict criminal behavior utilizing feature level data with an appropriate number of parameters, Kang & Kang (year) proposed a fusion technique based on deep neural networks.

3. SYSTEM DESIGN

Filtering and wrapping are two examples of machine learning pre-processing techniques used to cleanse the collected data of extraneous information. In addition, it successfully reduces the data's dimensionality, which improves the clarity of the data. The information is then divided in a second phase. There are two collections of information: the trained set and the test set. Both the training and testing datasets are used to help the learning process.The third step is the mapping phase. Numbers are used to indicate incident categories, years, months, days, times, and locations in the mapping process. The Nave Bayes approach is used to first determine the independence of the link between the attributes. For the goal of categorizing the collected independent features, the Bernoulli Naive Bayes method is used. To examine the expression of criminal activity within a given temporal and spatial environment, it is necessary to classify the criminological dimensions. The most common crimes have been pinpointed by the combination of geographical and chronological information. To evaluate the performance of the prediction model, the accuracy rate calculation is used. The prediction model was built with the help of Python and Colab, a web-based IDE made specifically for machine learning and data analysis projects.

Module Description

A. Data Pre-Processing

Pre-processing open source content is crucial for reducing unnecessary infractions. Denver was selected for inclusion in the dataset because of its substantial collection of crime data over a six-year period. Using machine learning techniques, notably the filter and wrapper approaches, it is expected that the missing integral can be identified in the available attribute values. Both the first step and the training of a prediction model require clean data to function optimally. Data cleansing refers to the steps used to remove unnecessary information from a collection of records.

In order to determine how relevant particular features are, filtering techniques are used. When choose which feature to use, it's important to think about how it connects to the values you're trying to predict. Using a trained prediction model, the wrapper method evaluates the usefulness of the feature subset. After the data has been cleaned up, it is separated into training and test sets.

B,. Mapping

Separating incident details, such as the type of crime and the time and date it was done, is the first order of business. The information is then

transformed into an integer format to make labeling easier. After the data has been obtained, it is analyzed and used to make charts. Because of Python's strength as a machine learning programming language, it is used to carry out the proposed work. The matplotlib software allows for the easy generation of a graph that displays the frequency and distribution of criminal activity. By visualizing the most common criminal acts, the graph improves the accuracy with which they can be predicted.

C. Naïve Bayes

The grouping of Naive Bayes is often used for crime prediction because of the integration of spatial and temporal data. Since the values of the selected criminal traits are subject to their own influences, the research begins by focusing on the independence of these values. Data on crimes like robbery, burglary, homicide, sexual abuse, armed robbery, chain snatching, gang rape, and cross-state robberies are used for training purposes before being incorporated into the model development process. Naive Bayes has been challenged in the literature by a number of other methods.

A sample with a Gaussian shape The Naive Bayes classifier is often used to determine which features have continuous values. The mean and standard deviation are determined by the trained data in order to establish a normal distribution.

Many classifiers for acquired data's categorical attributes use multinomial Naive Bayes.

For the aim of crime prediction, the Bernoulli Naive Bayes method is used to operationalize the independent feature effects of the specified characteristics.

D. Crime Prediction

The expected crime category is determined by extrapolating the underlying criminal characteristics. The qualities subsequently exert an influence on the nominal values..

One possible approach to provide a comprehensive explanation is by utilizing a specific tuple as an illustrative example.

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

Based on the evidence that has been gathered, it is probable that the following event will take place.

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}

After the establishment of the independent event, the conditional probability is calculated. By employing this methodology, it is possible to make anticipations regarding the category of criminal activity. The utilization 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:-

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

4. SYSTEM ANALYSIS

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 9, September : 2023

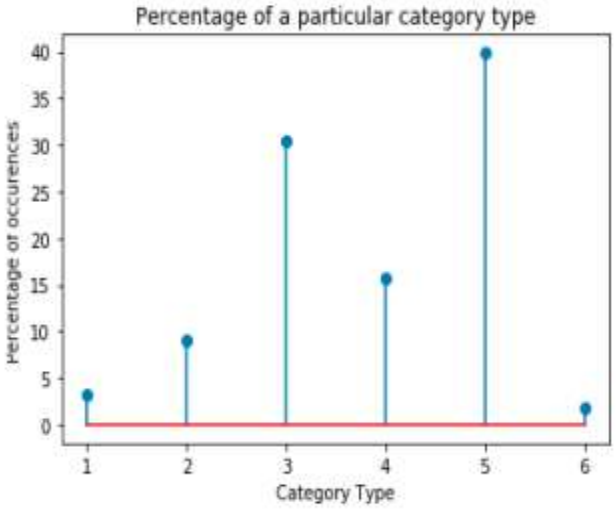

Fig 1. Plotting the highest crime type

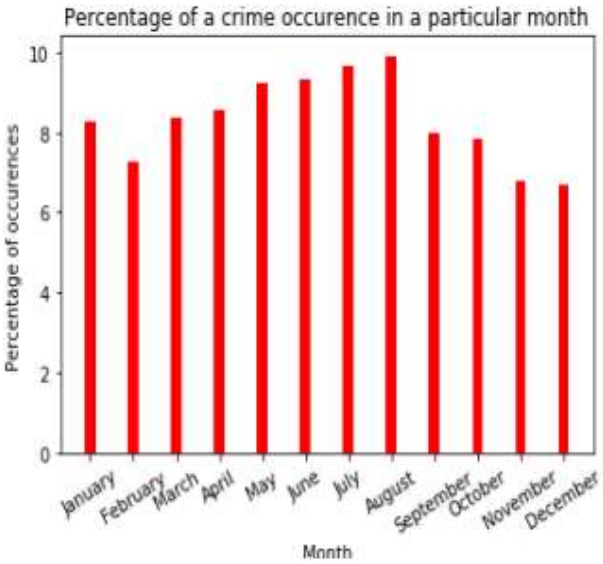

Fig 2. Plotting the highest occurrence month

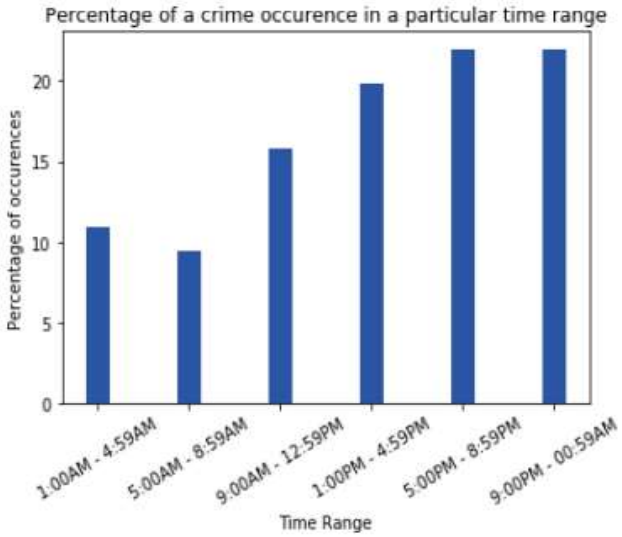

Fig 3. Plotting the highest occurrence time range

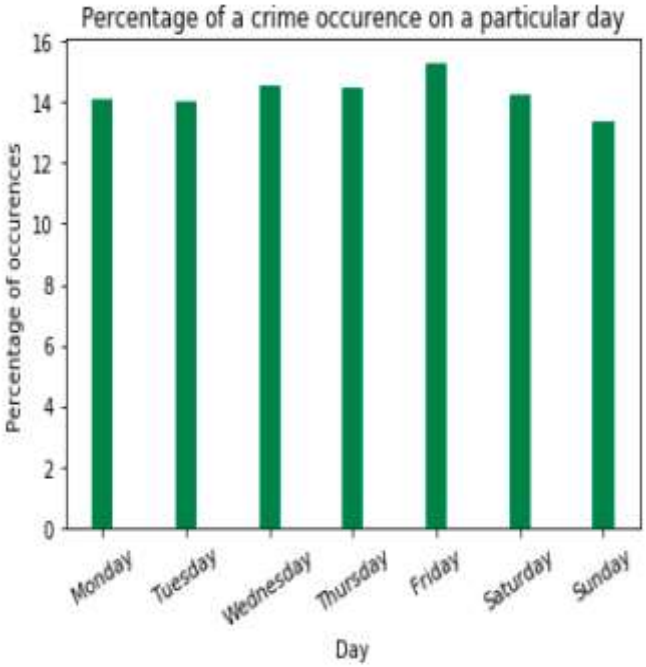

Fig 4. Plotting the highest occurrence day

RESULTS

The purpose of the performance review is to improve upon the accuracy of the projected forecast compared to the previously used model. During the training phase, cross validation is a typical technique used since it allows the data to be trained on many different sets of training data. This research aims to evaluate the performance of cross-validation in establishing the reliability of complete splits. Data arguments such as the model name, target set, and cross-validation (CV) parameters, which help determine the occurrence of data splits, are required by Python in order to generate an accurate calculation of the accuracy value. The average precision is then determined, along with its mean and standard deviation. The improved accuracy of 93.07% over previous prediction models is significant.

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

Table1. Performance measure for Naïve Bayes classifier

5. CONCLUSION

Multinomial Naive Bayes (NB) and Gaussian Naive Bayes (NB) classifiers are used in this research to overcome the difficulty of working with both nominal and real-valued data. Minimal training time is required to generate real-time projections. In addition, it successfully overcomes the limitation of dealing with a continuous collection of target variables, a problem that the prior approach was unable to solve. To this end, Naive Bayesian Classification can be used to make accurate predictions and identifications of typical criminal behavior. The algorithm's efficiency is also evaluated using a number of standard measures. The average precision, recall, F1 score, and accuracy are four metrics of critical relevance that must be taken into account whenever an algorithm is being analyzed. When applied, machine learning algorithms have the ability to greatly improve the precision.

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 multimodal data using deep learning”, Peerreviewed journal, published on April 2017.