LAUTECH Journal of Engineering and Technology 16 (2) 2022: 166-179

CRIME RATE PREDICTION USING THE RANDOM FOREST ALGORITHM

[1] Abdulraheem M., [2] Awotunde J. B., [3] Oladipo I. D., [4] Adeleke M. O., [5] Ndunagu J. N., [6]Ayantola J. A., and [7] Mohammed A.

[1]Department of Computer Science, Faculty of Communication and Information Sciences, University of Ilorin, Ilorin, 240003, Kwara State, Nigeria. Email: muyideen@unilorin.edu.ng
[2]Department of Computer Science, Faculty of Communication and Information Sciences, University of Ilorin, Ilorin, 240003, Kwara State, Nigeria. Email: awotunde.jb@unilorin.edu.ng
[3]Department of Computer Science, Faculty of Communication and Information Sciences, University of Ilorin, Ilorin, 240003, Kwara State, Nigeria. Email: odidowu@unilorin.edu.ng
[4]Department of Statistics, Faculty of Physical Sciences, University of Ilorin, Ilorin, 240003, Kwara State, Nigeria. Email: adeleke.mo@unilorin.edu.ng
[5]Department of Computer Science, National Open University of Nigeria, Abuja, Nigeria, Email: jndunagu@noun.edu.ng
[6]Department of Computer Science, Osun State Polytechnic, Iree Osun State, Nigeria. Email: tosoayantola24@gmail.com
[7]Department of Computer Science, Faculty of Communication and Information Sciences, University of Ilorin, Ilorin, 240003, Kwara State, Nigeria. Email: alimoh22062000@gmail.com

ABSTRACT

An act that creates crimes punishable by law is characterized as a crime. Rape, fraud, terrorism, kidnapping, burglary, murder, and other crimes are common in Nigeria. Examples are cybercrime, bribery and corruption, robbery, money laundering, among other crimes. Crime is a harmful and widespread social issue that affects individuals all around the world. The rate of crime has risen dramatically in recent years. To cut down on crime, at any rate, law enforcements must take preventative actions. To protect society against crime, modern systems and new technologies are required. Although accurate real-time crime study is on aid in reducing crime rates, they are nonetheless useless. As crime occurrences are dependent on, this is a difficult subject for the scientific community to solve. Therefore, this paper proposes machine learning algorithm to indicate the frequency and pattern of crimes based on the data collected and to show the extent of crime in a particular region. Various visualization approaches and machine learning algorithms are used in this study to anticipate the crime distribution over a large area. In the first stage, raw datasets were processed and visualized according to the requirements. Then, to extract knowledge from these massive datasets, machine learning methods were deployed and uncover hidden patterns in the data, which were then utilized to investigate and report on crime patterns, It is beneficial to crime analysts. Investigate these crime networks using a variety of interactive crime visualizations. As a result, it is helpful in crime prevention.

Keywords: *Crime rate, Radom Forest, Corruption, Machine learning, Crime network, Kidnapping, Cybercrime*

INTRODUCTION

According to data released by the Nigerian National Bureau of Statistics in 2015, Lagos, Abuja, Delta, Kano, Plateau, Ondo, Oyo, Bauchi, Adamawa, and Gombe states were among the top ten states in terms of crime rate. It's an important topic, and we're interested in it because of the implications. Fines are attracted to it (which ranges from fine to death). Officials in the criminal justice system must take precautionary measures. There is a need for improved technologies and novel ways for enhancing crime analytics in order to protect their communities.

Accurate real-time crime forecasting aids in the reduction of crime rates but they remain a difficult

challenge for scientists to solve because crime occurrences are influenced by a variety of complicated elements. For estimating the crime distribution over an area, multiple visualization approaches and machine learning algorithms are used in this study. In the initial stage, depending on the necessity, raw datasets were processed and shown.

Then, machine learning techniques were used to extract information from these enormous databases and discover newly undiscovered connections. This information was then used to report and uncover crime tendencies This might help crime analysts study these crime networks using various interactive visualizations for crime prediction, as a result, it's helpful in preventing crime, ToppiReddy, Bhavna, and Mahajan, (2018).

It is impossible to predict crime because it is not systematic. Criminals are still able to carry out their crimes despite the use of new technologies and high-tech techniques to combat crime are a common societal issue that has an impact on people's personal happiness and financial progress. It is regarded as an important factor in determining whether people should relocate to another city and which regions should be avoided when traveling. Daily, the crime rate continues to climb. Because crime is neither systematic nor random, it is impossible to foresee.

Furthermore, modern technologies and high-tech methods assist crooks in their illicit activities. Furthermore, crooks benefit from contemporary technologies and high-tech procedures in their illegal activities. Although the identities of crime victims cannot be predicted, the location and chance of the crime being committed may. Criminal justice and law enforcement experts have always overseen solving crimes. Law enforcement officers' benefit from the assistance of computer data analyzers in reducing the time it takes to investigate a crime as the use of electronic systems to monitor crimes grows, so does the use of electronic systems to track crimes. To aid in the faster resolution of crimes, a machine learning system has been developed. (Vaidya, Mitra, Kumbhar, Chavan, & Patil, 2018).

In today's world, criminals are becoming more technologically adept in their illegal activities, and reviewing huge volumes of data related to crime and terrorist activities is a challenge for intelligence and law enforcement organizations. As a result, agencies must adopt strategies to arrest criminals and keep ahead of them in the never-ending race between law enforcement and criminals. To undertake crime analysis, appropriate fields must be picked, and data mining, which is the process of collecting or mining information from massive amounts of data, when applied to a large-scale crime dataset, and data mining methodologies provide helpful information to police agencies, Agarwal, Nagpal & Sehgal, (2013).

Criminals have long been a source of annoyance to society in all corners of the globe, and efforts are on to eradicate crime from the earth. Detecting criminals after they have committed a crime is the focus of current policing strategies. However, because of technological improvements, it can recognize criminal tendencies and exploit these patterns to prevent crimes before they occur. Convert crime data into an algorithm problem that will aid investigators in solving crimes more quickly. Date, type of crime, quantity, gender, and location are among the information categories found in these crime reports. As the usage of electronic systems to track crimes has increased, computer data analysts have begun to aid law enforcement officials and detectives in speeding up the process of investigating crimes Win & Phyo, (2019).

The threat of criminality to humanity's survival is grave. On a regular basis, a great number of crimes

occur. It's conceivable that it's rapidly spreading and expanding throughout a large area. Crime can strike anywhere, from a tiny village to a large city, at any time. Robbery, murder, rape, assault, battery, false imprisonment, kidnapping, homicide, and other crimes are examples. Because crime is on the rise, it's vital to get cases resolved as soon as feasible. The rate of rise in criminal activity has accelerated, and it is the role of the police department to control and limit it. Due to the large amount of crime data, the police department's main concerns are crime prediction and criminal identification. Case resolution technologies that allow for faster resolution are in high demand. Machine learning and data science have been shown to make work easier and faster through a variety of documentation and instances. The goal is to use the features in the datasets to develop crime predictions. The information is taken from the official websites. The type of crime that will occur in a specific location can be predicted using a machine learning algorithm with Python as the core (He & Zheng, (2021).

Crimes such as burglary and arson have dropped, whereas crimes such as murder, sex abuse, and gang rape have increased, according to the Crime Records Bureau. Although the victims of these crimes cannot be anticipated, the location and likelihood of the acts can.

Although the accuracy of the expected outcomes cannot be guaranteed, the results show that this application does help to reduce crime rates by providing security in crime-prone areas to some level. As a result, to construct such a comprehensive crime analytics platform, crime records must be collected and reviewed Sathyadevan, Devan, & Gangadharan, (2014).

Every day, the population grows, resulting in a rise in crime in various places or regions Sathyadevan, Devan, & Gangadharan, (2014). Officials will be unable to precisely estimate crime rates because of this. In addition, officials may unable to foresee what crime will occur in the future because they are focused on a variety of topics. Police officers try to minimize crime rates, but not in a comprehensive way. It may be difficult for them to forecast future crime rates. Regarding crimes, a significant deal of work has been done. Large datasets have been evaluated, and information such as location and nature of crime has been extracted, to aid citizens in following police enforcement. These datasets have previously been used to pinpoint crime hotspots based on their location. Even though crime hotspots have been identified, there is no information available on the incident itself, such as date and time, or any reliable methods for forecasting future crimes, Win & Phyo, (2019*).*

Using a series of real-world crime statistics, this study tries to uncover spatial and temporal criminal hotspots. The study will attempt to pinpoint the most likely crime hotspots as well as the frequency with which they occur.

The aim of this paper is to develop a crime prediction system using Random Forest Algorithm. The main contributions are:

i. Design crime prediction system.
ii. Implement the system.
iii. Compare the proposed system with existing models.

The trend of urbanization and development of large cities and towns has shifted dramatically. Criminal activity is also on the rise. This extraordinary increase in criminal offenses and crime in cities is a source of considerable concern and alarm for us all. Robberies, murders, rapes, and other crimes are common. The thefts that occur on a regular basis, Burglaries, robberies, murders, homicides, rapes, stealing, pick pocketing, and drug-related offenses are all on the rise. Abuse, unlawful trafficking, smuggling, and auto theft, among other things,

have become commonplace. Sleepless nights and restless days are expected of citizens.

In the company of anti-social and bad people, they feel incredibly nervous and vulnerable. The crooks have been acting in a systematic and occasionally violent manner.

RELATED WORK

The crime rate continues to rise daily. Since crime is neither systematic nor random, it is impossible to foresee. Furthermore, current technologies and high-tech ways aid criminals in committing their crimes. According to the Crime Records Bureau, some crimes, such as burglary and arson, have decreased, while others, such as murder, sex abuse, and gang rape, have increased drastically. Even if we can't identify who will be the victims, we can estimate where they will most likely occur. Although the expected outcomes cannot be guaranteed to be 100 percent accurate, the results/conclusions imply that our program helps to reduce crime rates by providing enough security in crime-prone areas.

So, to create such a strong tool, we must first gather and analyze criminal records. Criminal data and records are few, information was gathered from a variety of sources, including news sites, blogs, websites, and social media. This information is utilized to create a database of criminal records. As a result, the key issue ahead of us is to design a better, more efficient crime pattern detection program that can successfully identify criminal trends Sathyadevan, Devan, & Gangadharan, (2014).

Policymakers and law enforcement agencies around the world, according to Duijn, Kashirin, and Sloot (2014), are having a difficult time devising effective strategies or methods to combat criminal activity. Both network topology and network resilience are known to affect the success of disruption methods. However, because these criminal acts are carried out in secret, data-driven knowledge about the efficacy of various criminal network disruption tactics is scarce.

According to Shama (2017), criminal activities exist in every corner of the world, posing a threat to people's quality of life as well as socioeconomic progress and development. Many governments are concerned about it, and they are employing various technological technologies to address the problem. Crime analysis is a branch of criminology that investigates patterns of criminal conduct and tries to find indicators of such behavior. Machine learning agents interact with data and identify patterns in several methods, which makes it excellent for predictive research. Law enforcement agencies use a variety of patrolling strategies based on the information they collect to keep an area safe. A machine learning specialist can study and assess the pattern of occurrence of a crime based on past criminal activity records, and can identify hotspots based on time, kind, and other characteristics. Classification is a strategy that allows us to predict some nominal class labels. Many diverse domains have employed classification, including the financial market, weather forecasting, corporate intelligence, healthcare, and so on.

To study criminal behavior, Jung & Suh (2019) employed information mining approaches. In this study, a technique for assessing criminal investigations was proposed (CIA). This instrument was utilized by law enforcement to aid in the resolution of violent crimes. This research focuses on different types of crime scenes. The analysis was completed from both an investigative and a behavioral standpoint. It supplied information about unknown criminals, as well as suggestions for investigation, interview, and trial techniques.

According to Mosleh, Bouguila, & Hamza (2012), Text detection is a technique for detecting places in

a photograph where there is text. For a range of computer vision applications, such as optical character recognition, distinguishing between human and machine inputs, and spam detection, text detection and categorization in natural images is crucial. Due to a range of concerns such as low image quality, indistinct words, standard font, image with more color strokes than the backdrop color, and blurred pictures due to natural problems such as rain, sunshine, snow, and so on, detecting text in natural photography has become a challenge. This study's main purpose is to detect and identify text in natural photographs. The method locates the related sections and ties them together in their relative positions after identifying the text. Using a text classification engine, filters chains with low classification confidence ratings.

According to Shah, Khalique, Saddar, & Mahoto (2017), Vancouver is Canada's most populous city. It is one of Canada's most ethnically diversified cities. Preventing crime is a key job because it is one of the society's most serious and prevalent challenges. Despite the fact that Vancouver is regarded as the safest city in the world, vehicle thefts and other crimes remain a problem. Machine learning algorithms have risen in popularity, allowing for historical data-based crime prediction. The purpose of this study is to assess and forecast crime in states using machine learning algorithms. Its goal is to create a model that can help detect the number of crimes committed by kind in a particular state. In this experiment, various machine learning models such as K-NN and boosted decision trees were used to predict crimes Area. A comprehensive geographical analysis may be carried out to understand the pattern of crimes. To assist law enforcement officials in better detecting and forecasting crimes, a variety of visualization tools and plots are used. This would help to reduce crime rates and improve security in such sensitive areas in an indirect way.

Since 2010, Microsoft's Azure cloud computing platform has been in use. You can construct a model, quickly build a web service, and deploy it to a range of devices using Azure Machine Learning Studio, which includes over 600 services. In addition, unlike other cloud platforms, machine learning libraries, and tools, it offers a user-friendly GUI interface. At Microsoft Ignite, Microsoft announced Azure Machine Learning designer, a drag-and-drop workflow tool in Azure Machine Learning studio that simplifies and accelerates building, testing, and deploying machine learning models for the entire data science team, from beginners to pros (azureml.net). Azure Machine Learning Studio allows data input, output, and visualization natively, and it comes preloaded with popular machine learning algorithms. It offers the advantage of being able to quickly develop a model by dragging and dropping blocks, unlike conventional machine learning tools and libraries. Moreover, scripts developed in the R and Python programming languages can be inserted and used in block form, with the outcomes checked using visualization. Anyone who understands how to use this simple structure can simply design and deploy predictive models (Kang et al., 2018).

In their paper, Venturini and Baralis (2016) used spectrum analysis to find spatio-temporal trends in crime. The goal is to search for seasonal tendencies in crime and determine if they occur across all sorts of crimes or if they vary by type of crime. The findings of the temporal analysis reveal that patterns differ not just by month, but also by type of crime. As a result, the authors properly underline the importance of accounting for volatility in models based on this data. They used the Lomb-Scargle periodogram to depict the crime's periodicity since it is better at handling uneven or missing data. The AstroML Python package was used to do this. When the algorithm is applied to the data in their study, they explain how

each sort of crime performs in detail. The authors also recommend that academics concentrate on monthly and weekly crime patterns.

The Support Vector Machine technique, according to Kianmehr and Alhajj (2006) can be used to anticipate hotspots and give a proportion of data. In a subset of the crime datasets, each data point is classified according to a preset degree of crime rate (The quantity or proportion of data crimes). Positive or hotspot class data points have a rate that is larger than or equal to the set rate, whereas negative or non-hotspot class data points have a rate that is less than or equal to the set rate. In SVM classification, the parameter supplied will be utilized as the training. As a consequence, the performance of one-class SVM for predicting crime hotspot of regions was compared using linear, polynomial, and gaussian kernel functions. The information is sourced from available web databases (two datasets, Columbus, and ST. Louis). Adopting this approach seems very slow and computationally high in price, despite its advantages.

Shrivastav (2012) used the Fuzzy time series to detect crime in the community, where there is a criminal trend. A simpler approach and model were utilized in this method to reduce the processing cost, which included simple arithmetic operations. The statistics from the preceding seventeen years were used in this plan (murder cases in the city of Delhi). As a result, real-life examples and years from the past are used as criteria. Twenty (Scheme-III), ten (Scheme-II) and five (Scheme-I) intervals were used to partition the data (Scheme-III). The results of Schemes II and III are fairly accurate; however, the outcomes of Scheme I are significantly over forecasted, with an average absolute inaccuracy of about 5%. So, this formula is restricted to binary transaction data input, such as 1 or 0.

For crime prediction, Chitsazan and Rahmani (2013) used an artificial neural network (ANN). This method was developed by concentrating mostly on geographical locations. In terms of crime prediction, they outperform traditional policing boundaries. As a result, ANN may be trained utilizing crime data from geographical clusters, making predictive modeling easier. As a result, a scanning technique based on geographic crime occurrences was employed to identify high-criminality crime hotspot clusters. With each cluster, ANN demonstrates its ability to model broad trends. The data included 18,498 violent occurrences (criminal damages, violence against people), as well as time, day, month, location, and weather information. When the mean standard error (MSE) of ANN (9.94) and random walk are compared, ANN (9.94) outshines random walk (22.50). As a result, in terms of accuracy, ANN outperforms a random walk. Nonetheless, the training procedures for this system are lengthy.

B. Chandra et al. (2008) proposed an unsupervised strategy Using multivariate time series, assess many crime data at various time intervals. This technique makes extensive use of the parametric Minkowski model and dynamic time wrapping (DTW). In numerous criminal circumstances, similar crime trends have been identified, and the information has been used to forecast future crime trends. To demonstrate the utility of this strategy, data from the Indian National Crime Records Bureau's crime dataset was used (29 districts). The different dimensions data weights are the model's parameter. They discovered that DTW and the Euclidean (equal weight age) model are equivalent, as are DWT and the parametric Minkowski model (different weight age). This strategy is used when the dimensions of data do not have the same weight age for efficiently detecting similar crime patterns and studying crime trends. However, using this

method to manage a missing number in order to offer more precise results is problematic.

One of the crime prediction models introduced by Liao et al. (2013) is the Bayesian Network, which is based on Bayesian learning theory. This strategy can be used to create useful mathematical models for understanding repeated crime behavior. There are numerous elements that could influence the selection of criminals for future crime scenes. A serial crime dataset from Gansu, China, was used to test the model. As a result, the parameters employed are victim characteristics (age, gender, occupation, and race) as well as crime area characteristics (a school, a hospital, a hotel, a bus stop, a house). The proposed zones (red, yellow, and green) will also assist police in apprehending offenders. However, this is contingent on the geographical parameters specified. This approach, however, is entirely dependent on parameter choices. As a result, it's possible that some deviation will occur during the experiment, hence extra parameters other than geography should be incorporated to increase the model's accuracy.

MATERIALS AND METHODS

Overall Framework of crime rate prediction

The main framework of the proposed model is depicted in Figure 3.1. From data gathering to data preprocessing, including data mapping (assigning numerical values to categorical data) and data standardization (to transform data in a way that they are either dimensionless or have similar distribution). The data is trained and tested with a Random Forest classifier, and the results are displayed.

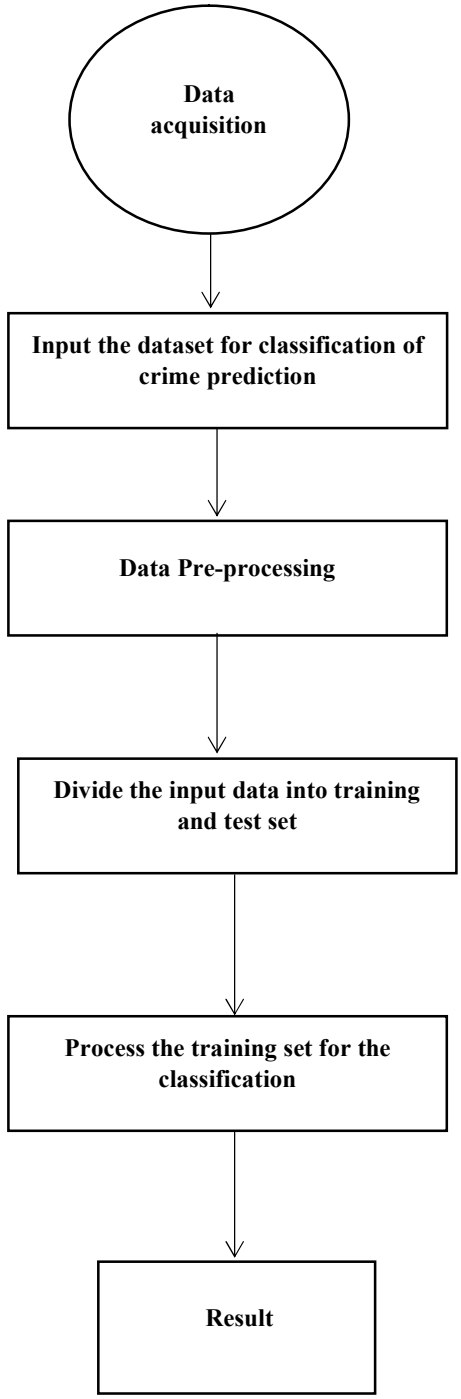

Figure 3.1: Proposed framework

Dataset

The crime data from San Francisco used in this paper is open source and freely available on data.world. The dataset file contains 25 columns and 161,224 rows. Buffalo is one of New York's most popular cities. On kaggle.com, the data was

retrieved from the internet. This dataset is updated daily and provides an overview of crime reports in Buffalo. https://data.buffalony.gov/d/d6g9-xbgu is the source. The administrator of data world developed this dataset.

Data preprocessing

Feature Selection

Because quality decisions must be founded on quality data, data preparation is an important part of the knowledge discovery process. Cleaning, integrating, altering, and reducing data are all examples of data preparation. When these data processing processes are used before mining, the overall quality of the patterns mined can be much improved, and the time spent mining can be greatly reduced. The goal of data reduction is to find the least number of attributes that will result in a probability distribution for the data classes that is as close as feasible to the original. Due to the large dimensionality and redundancy of the data used in this study, a feature selection machine learning preprocessing strategy was used.

Feature selection is the process of picking a subset of the original features. Feature selection is one of the most important and widely used approaches in data preprocessing for data mining. In most real-world situations, relevant qualities are unknown a priori. As a result, feature selection is critical for identifying and discarding redundant or irrelevant data. It can be used in both supervised and unsupervised learning environments. The number of input variables should be minimized to reduce the computational cost of modeling and, in some situations, to increase the model's performance.

Each input variable's link with the goal variable is analyzed using statistics, and the input variables with the strongest association to the target variable are chosen. Although the statistical measures utilized are influenced by the type of data in both the input and output variables, these processes can be swift and successful. As a result, when performing filter-based feature selection, a machine learning expert may find it difficult to determine an acceptable statistical measure for a dataset.

The feature selection approach is broken down into four steps, which are explained below:

i. Generate candidate subset: Because the original feature set has n features, the total number of competing candidate subsets that must be constructed is 2n, which is an enormous number even for a small n. Subset generation is a search strategy that generates candidate feature subsets for examination using a search strategy. Heuristic (forward selection, random (Las Vegas algorithm)), complete (e.g., breadth first search, branch & bound, beam search, best first) (LVW), Random generation + sequential selection), genetic algorithm (GA), and random (Las Vegas algorithm) (LVW), Random generation plus sequential selection) (RGSS), simulated annealing (SA)) are all examples of genetic algorithms are examples of several sorts of search strategies.
ii. Subset evaluation function to evaluate the subset obtained in the preceding phase using a filter or wrapper approach (generate candidate subset). The sole difference between the Filter and Wrapper approaches is how they analyze a subset of features. The learning induction algorithm has no bearing on the filter strategy. The worth of feature subsets is estimated via an induction process in wrapper systems for feature selection. Wrappers usually outperform filters because they are customized to the specific interaction between an induction method and its training data.
iii. Halting Condition: Because the number of subsets can be enormous, a halting condition is required. As a stopping criterion, a generating procedure/evaluation function could be employed. The following are examples of

stopping criteria based on the generating procedure:

- Whether or not a predetermined amount of features is chosen
- Whether or not a predetermined number of iterations have been completed

The following are some examples of evaluation-based halting criteria:

- Whether or not adding (or removing) a feature result in a superior subset
- Is it possible to obtain an optimal subset based on an evaluation function?

iv. Validation method for validating the correctness of the feature subset chosen. When using artificial/real-world datasets for induction, it's common to compare the output of the original feature set to the feature chosen as input by filters/wrappers.

Another way to validate is to extract critical features using several feature selection techniques, then compare the results with classifiers on each relevant attribute subset.

Random Forest Algorithm Pseudo Code

Input *training set (1v1, 1v2, 1v3,…,1vk)*

Output: *Class of attack*

Generate trees

1. *For I in number of trees*
2. *Using the bagging method with implacement for j in the number of nodes, select sample sets. For j in number of nodes*
3. *Choose 'k' features at random from the 'M' features,*
4. *For each random features in k set*
5. *Calculate gini index*
6. *End for*
7. *Make a 'd' node for the feature with the lowest gini index.*
8. *End for*
9. *End for*

Deciding the class of attack

10. *For each tress*
11. *For each class_label*
12. *Compute class attack*
13. *End for*
14. *End for*
15. *Compute mode of class_labels*
16. *Assign the instance the class attack with the highest frequency.*

EXPERIMENTAL RESULTS

Simulation tools

Google Collaboratory, random forest, Scikit learn, Pandas, numpy, and Matplotlib were used in the experiment.

Performance evaluation and measurement terms

The effectiveness of the Artificial Neural Network model created in this study will be evaluated using the following terms:

1. System accuracy: Accuracy is a statistic that describes how well the model performs across all classes. It is beneficial when all classes are of equal priority. It is calculated by dividing the total number of forecasts by the number of correct estimates.

$$\frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FP + FN}}$$

2. True Positive **TP:** When compared to all positive data points, the True Positive Rate is the percentage of positive data points that are correctly identified as positive.

True Positive rate:

$$\frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}}$$

3. True Negative **TN:** It's the number of normal and legitimate cases that have been successfully classified as glaucoma.

$$\text{True Negative rate}: \frac{TN}{FP + FN}$$

4. Training Time: It's the time taken to build the model, measured in seconds

Dependencies Used

Tensorflow, numpy, pandas, matplotlib, ipykernel, keras, and scikit-learn are among the dependencies employed in this work. Anaconda was used to install these dependencies, which were then called by the jupyter notebook.

The first step was to download and install tensorflow, an open-source artificial intelligence framework that creates models using data flow graphs. Keras is a Python interface to an open-source artificial neural network software framework. Keras is a user interface for the TensorFlow library that makes creating neural networks a lot easier. Scikit-learn is Python's most popular and reliable machine learning library. It offers a single Python interface for a variety of efficient machine learning and statistical modeling algorithms, categorization, regression, clustering, and dimensionality reduction are just a few examples. Ipykernel is a robust interactive Python shell that acts as a backend for Jupyter and other interactive frontends by executing Python code. Numpy (Numerical Python) is a Python package that allows you to manipulate massive multidimensional arrays and matrices. It also makes sophisticated mathematical formulas in Python, such as the fourier transform and linear algebra, easier to utilize. Pandas is a data analysis and manipulation package written in Python. It offers data manipulation and time series operations as well as data structures and operations. Matplotlib is a charting toolkit for the Python programming language and its numerical extension, numpy. It's used to make graphs.

System Requirements

To carry out this study, the following software and hardware components were used:

Hardware Requirements

During the course of this study, the software developed required the minimum hardware configurations for an effective and efficient operation: Intel core i3, 64-bit Windows 10 operating and system 4GB RAM.

Software Requirements

The software requirements include: Google Collaboratory, Google chrome browser, Jupyter Notebook, and Tensorflow, matplotlib, numpy, pandas, scikit-learn python libraries

4.5 experimental result of Random Forest for crime rate prediction in San Francisco

Figure 2 shown the analysis of crime by district in San Francisco

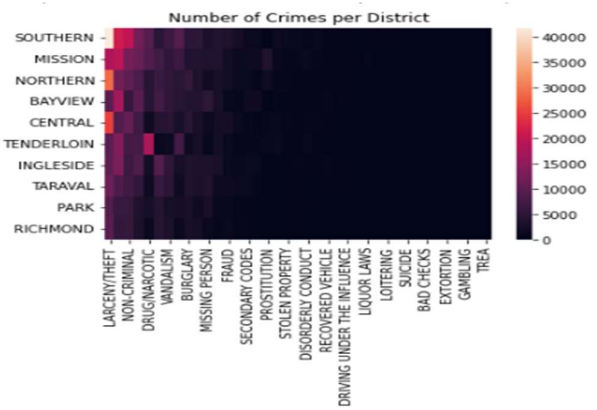

Figure 2: analysis of crime by district in San Francisco

Figure 2 shown the correlation heat map of the training and testing data.

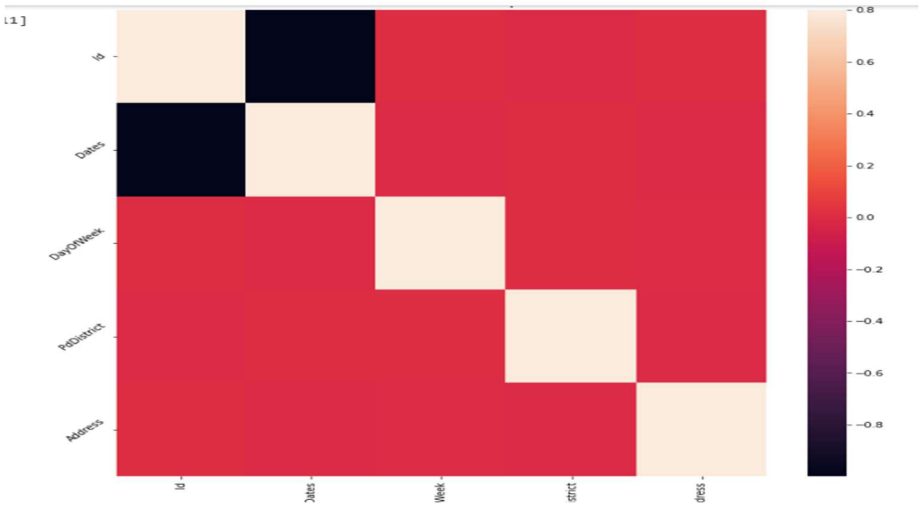

Figure 3: correlation heat map of the training and testing data

Figure 4 shown the line chart of the crime occurrence per hour in san Francisco.

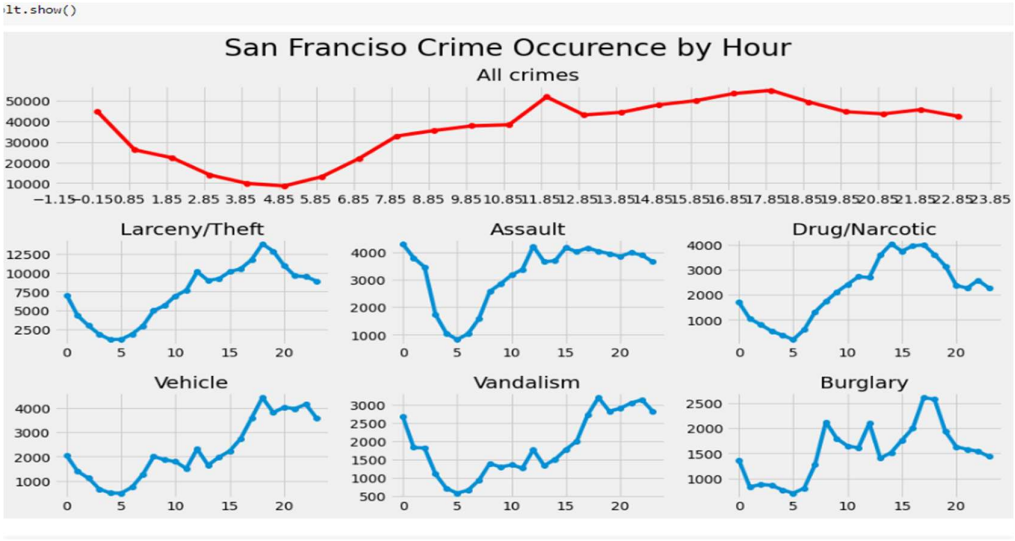

Figure 4: crime occurrence per hour in San Fransisco by hour

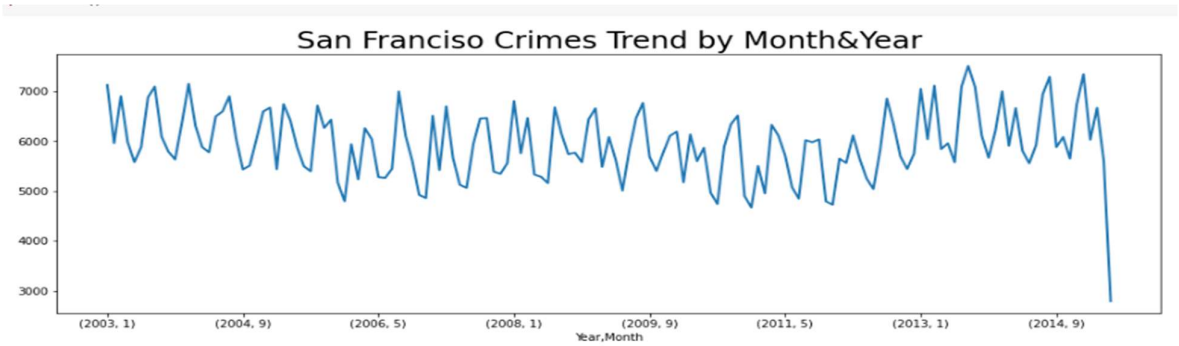

Figure 5 shown the line chart of the san Francisco crime trend by month and year.

Figure 5: San Francisco crime trend by month and year

Figure 6 shown the histogram of grouping of best features in the test data

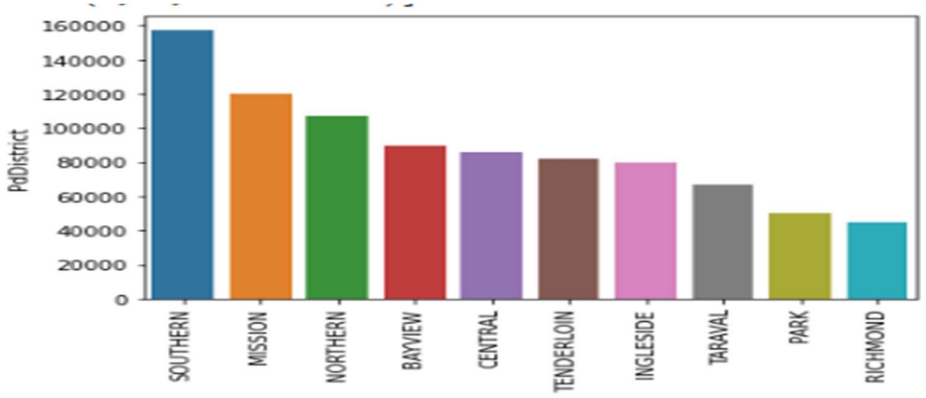

Figure 6: Grouping the features in test set

Figure 7 shown the histogram of grouping of best features in the training data

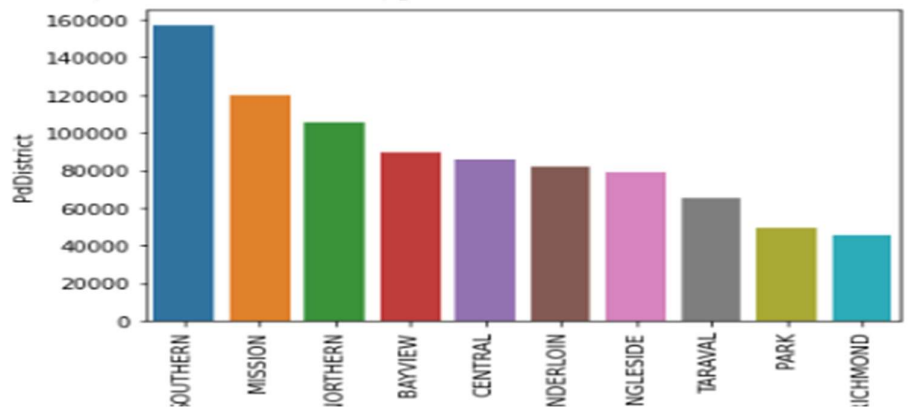

Figure 7: Grouping the features in testing set

Figure 8 shown the performance evaluation using confusion matrix.

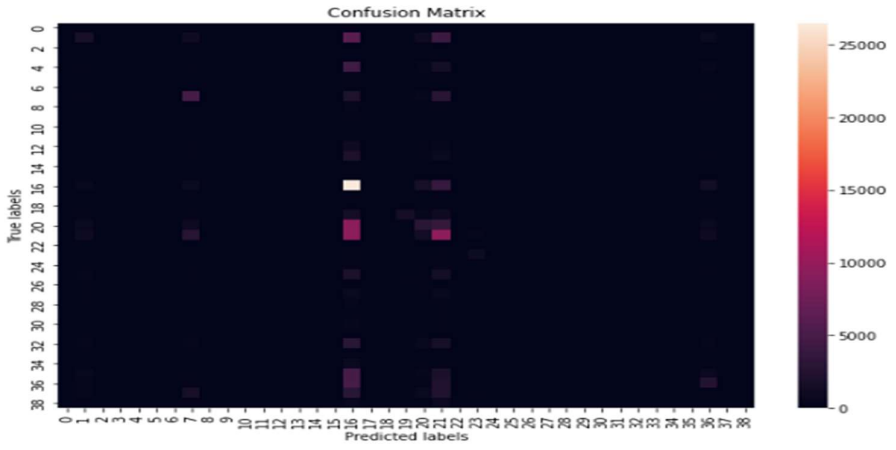

Figure 8: Performance evaluation of the model using confusion matrix

RESULT AND DISCUSSION

In this survey paper, we looked at a study that looked at a range of crime-related variables and found that, in certain situations, information that has long been assumed to influence crime rates really had no effect. The threat of crime and violation is intended to be controlled. Computationally, accurate crime predictions and future trend analysis can help to improve metropolis safety. Big data makes it difficult to detect and forecast crime early and accurately since people have a limited capacity for processing complicated information. Many computational opportunities and problems are presented by the precise calculation of the crime rate, types, and hotspots from historical trends. Despite significant research efforts, a better predictive algorithm is still required to lead police patrols in the direction of criminal activity. There is not enough prior research to accurately foresee and predict crime using learning models.

As a result, the RF method was utilized in this study to better fit the crime data, and the San Francisco crime dataset was used to test the model's effectiveness. In both sets of data, the RF model performed about as well as could be expected in terms of mean absolute error (MAE) and root mean square error (RMSE). More than 35 different types of crimes are predicted through exploratory data analysis, which also shows that the crime rate in Los Angeles is slightly higher than that in San Francisco and that fewer crimes were committed in February than in other months. The overall crime rate in San Francisco will likely continue to rise gradually over the coming years before beginning to fall. The ARIMA model predicts a significant drop in crime and the crime rate in Los Angeles. Also, the key regions of both cities' crime projection statistics were determined. These findings offer early crime detection, hotspots with greater crime rates, and enhanced future trends. Because the RF model has been deemed the most effective machine learning algorithm for the prediction of crime data as detailed in the linked literature, it was taken into consideration in this work when creating a prototype crime prediction model. According to the experimental findings, the RF model accurately identified the unknown category of crime data with a precision of 94.92%, which is respectable enough for the system to be used for crime prediction in the future.

CONCLUSION

This study's main purpose is to estimate crime rates in various locations based on specific factors such as density, country, crime rate, and centrality. We looked at a dataset from a specific country in this study. We've decided on the United States of America as our destination. A Random Forest model was used to forecast the crime rate. A graph is shown following the Random Forest implementation. Crime is a significant issue that both the community and the rest of the world must address and manage. Many people, societies, regions, and the entire planet are affected by crime. It is critical, but challenging, to predict crime and extract valuable information from huge amounts of crime data. Crime may be reduced if advanced forecasts of the situation can be made. It can be minimized if it is not stopped. This is an area where a lot of work is being done. On the other hand, the forecasting system can yet be enhanced. A survey is being done in attempt to improve crime predictions using outstanding data gathering and data mining technologies. Crime trends and patterns are detected. Predict which types of violations will occur next in a given district over the course of a certain time period and season. As a result, crime prediction assists consumers in avoiding certain neighborhoods at specific times of the day, month, and season while also saving money. If people had this kind of knowledge, they

would be able to make better judgments about where they live and visit.

REFERENCES

A.K. Shrivastav, "Applicability of Soft Computing Technique for Crime Forecasting: A Preliminary Investigation," in International Journal of Computer Science & Engineering Technology, vol. 9, no. 9, 2012, pp 415-421.

Agarwal, J., Nagpal, R., & Sehgal, R. (2013). Crime analysis using k-means clustering. International Journal of Computer Applications, 83(4).

B. Chandra, M. Gupta and M.P. Gupta, "A Multivariate Time Series Clustering Approach for Crime Trends Prediction," in International Conference on In Systems, Man and Cybernatics, 2008, pp. 892-896.

He, J., & Zheng, H. (2021). Prediction of crime rate in urban neighborhoods based on machine learning. Engineering Applications of Artificial Intelligence, 106, 104460.

Jung, Y., & Suh, Y. (2019). Mining the voice of employees: A text mining approach to identifying and analyzing job satisfaction factors from online employee reviews. Decision Support Systems, 123, 113074.

K. Kianmehr and R. Alhajj, "Crime Hot-spots prediction using support vector machine," in IEEE International Conference on Computer Systems and Applications, 2006, pp. 952-959.

Kang, M. S., Kang, H. J., Yoo, K. B., Ihm, C. H., & Choi, E. S. (2018). Getting started with Machine Learning using Azure Machine Learning studio. Seoul, Korea: Hanti media.

L. Venturini and E. Baralis, "A spectral analysis of crimes in san francisco," in Proceedings of the 2nd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics. ACM, 2016, p. 4.

M. Chitsazan and G. Rahmani, "Groundwater level simulation using artificial neural network: a case study from Aghili plain, urban area of Gotvand, south west Iran," in Journal of Geopersia, Vol. 3, No. 1, 2013, pp. 35-46.

Mosleh, A., Bouguila, N., & Hamza, A. B. (2012, September). Image Text Detection Using a Bandlet-Based Edge Detector and Stroke Width Transform. In BMVC (pp. 1-12).

P. A. C. Duijn, V. Kashirin, and P. M. A. Sloot, "The relative ineffectiveness of criminal network disruption," Sci. Rep., vol. 4, 2014.

R. Liao, X. Wang, L. Li and Z. Qinh, "A Novel Serial Crime Prediction Model Based on Bayesian Learning Theory," in International Conference on Machine Learning and Cybernatics, 2013, pp. 17571762.

Sathyadevan, S., Devan, M. S., & Gangadharan, S. S. (2014, August). Crime analysis and prediction using data mining. In 2014 First international conference on networks & soft computing (ICNSC2014) (pp. 406-412). IEEE.

Shah, S., Khalique, V., Saddar, S., & Mahoto, N. A. (2017). A framework for visual representation of crime information. Indian Journal of Science and Technology, 10(40), 1-8.

Shama, N. (2017). A machine learning approach to predict crime using time and location data (Doctoral dissertation, BRAC University).

ToppiReddy, H. K. R., Saini, B., & Mahajan, G. (2018). Crime prediction & monitoring framework based on spatial analysis. Procedia computer science, 132, 696-705.

Vaidya, O., Mitra, S., Kumbhar, R., Chavan, S., & Patil, M. R. (2018). Crime rate prediction using data clustering algorithms. International Research Journal of Engineering and Technology (IRJET), 2395-0056.

Win, T., & Phyo, E. E. (2019). Predicting of crime detection using K-means clustering algorithm. METHODOLOGY, 1(1), 2.