CRIME RATE PREDICTION USING RANDOM FOREST ALGORITHM

BY

MOHAMMED ALI

(17/52HA069)

A SEMINAR REPORT SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE, FACULTY OF COMMUNICATION AND INFORMATION SCIENCES, UNIVERSITY OF ILORIN, ILORIN, NIGERIA

NOVEMBER, 2021

CERTIFICATION

This is to certify that this seminar report was written and submitted by **MOHAMMED ALI (17/52HA069)** to the Department of Computer Science, Faculty of Communication and Information Sciences, University of Ilorin, Ilorin, Nigeria as part of the requirement for the award of Bachelor degree in Computer Science.

---------- ----------

Dr. J. B. Awotunde **DATE**

(Supervisor)

---------- ----------

Dr. Mrs. S. A. Salihu **DATE**

(Seminar Coordinator)

Table of Contents

I. BACKGROUND TO THE STUDY

Crimes are a typical social issue influencing the personal satisfaction and the monetary development of a general public. It is viewed as a fundamental factor that decides if individuals move to another city and what spots ought to be kept away from when they travel.

Day by day, the crime rate is increasing considerably. Crime cannot be predicted since it is neither systematic nor random. Also, the modern technologies and hi-tech methods help criminals in achieving their misdeeds. According to Crime Records Bureau, crimes like burglary, arson etc. have been decreased while crimes like murder, sex abuse, gang rape etc. have been increased.

Crimes are treacherous and common social problem faced worldwide. Crimes affect the quality of life, economic growth, and reputation of a nation. There has been an enormous increase in crime rate in the last few years. In order to reduce the crime rate, the law enforcements need to take the preventive measures. With the aim of securing the society from crimes, there is a need for advanced systems and new approaches for improving the crime analytics for protecting their communities.

Accurate real-time crime predictions help to reduce the crime rate but remains challenging problem for the scientific community as crime occurrences depend on many complex factors. In this work, various visualizing techniques and machine learning algorithms are adopted for predicting the crime distribution over an area. In the first step, the raw datasets were processed and visualized based on the need. Afterwards, machine learning algorithms were used to extract the knowledge out of these large datasets and discover the hidden relationships among the data which is further used to report and discover the crime patterns that is valuable for crime analysts to analyze these crime networks by the means of various interactive visualizations for crime prediction and hence is supportive in prevention of crimes. *(ToppiReddy, Bhavna & Mahajan, 2018).*

Crimes cannot be predicted since it is not systematic. Despite of the modern technologies and hi-tech methods used to curb the crime; the criminals are successful in achieving their misdeeds.

Although, the victims of the crimes cannot be predicted but the place and probability of the occurrence of the crime can very well be predicted.

The process of solving crimes has been the prerogative of the criminal justice and law enforcement specialists. The increase in the use of computerized systems to track crimes, the law enforcement officers are helped by computer data analysts to speed up the process of solving crimes. Machine Learning algorithm is developed to help in solving crimes at a faster rate *(Omkar, Mitra, Kumbhar, Chavan & Rohini 2018).*

In present scenario, criminals are becoming technologically sophisticated in committing crime and one challenge faced by intelligence and law enforcement agencies is difficulty in analyzing large volume of data involved in crime and terrorist activities therefore agencies need to know technique to catch criminal and remain ahead in the eternal race between the criminals and the law enforcement.

So appropriate fields need to be chosen to perform crime analysis and as data mining refers to extracting or mining knowledge from large amounts of data, data mining is used here on high volume crime dataset and knowledge gained from data mining approaches is useful and support police forces. *(Jyoti, Nagpal & Sehgal, 2013).*

Criminals are nuisance for the society in all corners of world for a long time now and measures are taken to eradicate crimes from the world. Current policing strategies work towards detecting the criminals, specifically after the crime has occurred. But, with the help of technological advancement, it uses historic crime data or records to recognize crime patterns and use these patterns to detect crimes beforehand. Convert crime information into algorithm problem such that it can help the detectives in solving crimes faster. These crime reports have the following kinds of information categories namely: date, type of crime, quantity, gender and location etc. The increasing use of the computerized systems to track crimes, computer data analysts have started helping the law enforcement officers and detectives to speed up the process of solving crimes. *(Than & Phyo, 2019).*

Crimes are the significant threat to the humankind. There are many crimes that happen in regular intervals of time. Perhaps it is increasing and spreading at a fast and vast rate. Crimes happen from small village, town to big cities. Crimes are of different type; robbery, murder, rape, assault, battery, false imprisonment, kidnapping, homicide etc. Since crimes are increasing, there is a need to solve the cases in a much faster way. The crime activities have been increased at a faster rate and it is the responsibility of police department to control and reduce the crime activities. Crime prediction and criminal identification are the major problems to the police department as there are tremendous amount of crime data that exist. There is a need of technology through which the case solving could be faster.

Through many documentation and cases, it came out that machine learning and data science can make the work easier and faster. The aim of this is to make crime prediction using the features present in the datasets. The datasets are extracted from the official sites. With the help of machine learning algorithm, using python as core, the type of crime that will occur in a particular place can be predicted. *(Babu & Avinas, 2020).*

According to Crime Records Bureau, crimes like burglary, arson etc. have been decreased while crimes like murder, sex abuse, gang rape etc. have been increased. Even though the victims of these crimes cannot be predicted but the place and probability for the crimes can be predicted.

The predicted results cannot be assured of 100% accuracy but the results show that this application helps in reducing crime rate to a certain extent by providing security in crime sensitive areas. So, for building such a powerful crime analytics tool, crime records have to be collected and evaluated *(Devan, 2014).*

II. STATEMENT OF PROBLEM

Population increases every day and by that, crimes are also going to be increased in different areas or regions. *(Devan, 2014).* By this, the crime rates cannot be accurately predicted by the officials. The officials as they focus on many issues might not be able to predict what crime will occur in the future. The officials/police officers although try to reduce the crime rates but not in a full-fledged manner. The crime rate prediction in the future may be difficult for them.

There has been countless of work done related to crimes. Large datasets have been reviewed, and information such as location and the type of crimes have been extracted to help people follow law enforcements. Existing methods have used these databases to identify crime hotspots based on locations.

Even though crime locations have been identified, there is no information available that includes the crime occurrence, that is, date and time along with techniques that can accurately predict what crimes will occur in the future (*Than & Phyo, 2019).*

This study aims to find spatial and temporal criminal hotspots using a set of real-world datasets of crimes. The project will try to locate the most likely crime locations and their frequent occurrence time.

III. AIM AND OBJECTIVES

AIM

The aim of this project is to predict the rates of crimes in different locations based on the available Datasets.

OBJECTIVES

The main objectives of this work are to:

- Design the proposed system for crime prediction.
- Implement the proposed system.
- Compare the proposed system with existing models.

IV. LITERATURE REVIEW

This Chapter explains more about each of the algorithms that are used in this study and also gives report on some of the related work.

1. Data Mining

Data mining is an information extraction knowledge discovery process from bulks of datasets. The derived information helps in various ways which include business and financial management, budgeting, population control, medical assistance, prevention and management of risks, improvement of productivity and so on.

There are six common classes of operations in data mining namely; Anomaly detection, Association rule, Clustering, Classification, Regression and Summarization. Some of the major roles carried out in a data mining process are as follows:

i. Extract, transform and load data into a data warehouse.

ii. Store and manage data in a multidimensional database.

iii. Provide data access to Organization using application software.

iv. Present analyzed data in easily understandable forms e.g. graphs.

1.1 Classification

This is the task of generalizing known structure and apply them to a new data. It is a data mining technique that assigns items in a collection to target classes. The aim of classification is to accurately predict the target class for each case in the data. Example of the algorithm in this category are Neural Networks, C4.5 Algorithm, ID3 Algorithm, K-Nearest Neighbors Algorithm, Naives Bayes Algorithm, SVM Algorithm, J48 Decision tree.

1.2 Regression

The work of Regression is to find a function which models the data with the least error i.e. to estimate and analyze relationships among datasets. Some types of Regression include; Linear Regression, Standard Multiple Regression, Stepwise Multiple Regression, Hierarchical Regression and Setwise Regression, Logistic Regression, Polynomial Regression, Ridge Regression, Lasso Regression and Elastic Net Regression (Mike, 2018).

1.3 Anomaly Detection

Anomaly detection (also referred to as deviation detection or outliers) is the identification of unusual data records, that might be interesting or some errors in data that require further investigation.

1.4 Clustering

Clustering is the task of discovering groups and structures of data that are similar, without using known structures in the data. Some of the techniques used for Clustering are Artificial Neural Network (ANN), Nearest neighbor search, Neighborhood components analysis, Latent class analysis and Affinity propagation.

1.5 Naïve Bayes

The Naïve Bayes classifier technique is based on the **Bayesian theorem** and it is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naïve Bayes can often outperform more sophisticated classification methods. In machine learning, Naïve Bayes classifiers are a family of simple "probability classifiers" based on applying Bayes theorem with strong independence assumptions between the features.

According to Srivastava and Sharma (2016), Naïve Bayes is one of the most effective algorithms use for classification in data mining. It is use for supervised data and used as a probabilistic

learning method. It is easy to build, understand and debug. It has a very good advantage which is, a very small amount of training data is needed to estimate the parameters that are required for classification.

2. Ensemble Method

In Machine learning and Statistics, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. Unlike a statistical ensemble in statistical mechanics, which is usually infinite, a machine learning ensemble consists of only a concrete finite set of alternative models, but typically allows for much more flexible structure to exist among those alternatives.

An ensemble itself is a supervised learning algorithm, because it can be trained and then used to make predictions. Empirically, ensembles tend to produce better results when there is a significant diversity among the models. Many ensemble methods, therefore, seek to promote diversity among the models they combine. Although perhaps non-intuitive, more random algorithms (e.g. random decision trees) can be used to produce a stronger ensemble than deliberate algorithms (like entropy-reducing decision trees).

2.1 Bootstrap Aggregating

This is sometimes referred to as **Bagging.** It is to decrease the variance of the prediction model by generating additional data in the training stage. This is produced by random sampling with replacement from the original set.

2.2 Stacking

Stacking, also known as stacked generalization, is an ensemble method where models are combined using another machine learning algorithm. The basic idea is to train machine learning

algorithms with training dataset and then generate a new dataset with these models. Then this new dataset is used as input for the combiner machine learning algorithm (Necati, 2016).

2.3 Voting

Voting is one of the easiest ensemble methods. It is easy to understand and implement. It is majorl0079 used for classification. The first step is to create multiple classification models using some training dataset and same algorithm, or using the same dataset with different algorithms, or any other method.

2.3.1 Majority Voting

Every model makes a prediction(votes) for each test instance and the final output prediction is the one that gets more than half of the votes. If none of the predictions get more than half of the votes, we may say that the ensemble method could not make a stable prediction for this instance.

2.3.2 Weighted Voting

Unlike majority voting, where each model has the same rights, the importance of one or more models can be increased. In weighted voting, there is need to take the count of the prediction of the better models ‘multiple times and a reasonable set of weights is used.

2.4 Random Forest

Random Forest (or random forests) is a trademark term for an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. It is a collection of slightly different trees.

A random forest is a meta estimator that fits a number of classifiable decision trees on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. Each decision tree is constructed using a Random subset of the training data.

It is a supervised classification algorithm. It creates a forest by some ways and makes it random. There is a direct relationship between the number of trees in the generated forest and the results it can get. The larger the number of trees, the more accurate the result. But an important thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach (Dan, 2012). Decision tree is a decision support tool. It uses a tree-like graph to show the possible consequences. When a training dataset is supplied with targets and features into the decision tree, it will formulate some set of rules. These rules can be used to perform predictions (Synced, 2017).

The difference between Random Forest algorithm and the decision tree algorithm is that in Random Forest, the process of finding the root node and splitting the feature nodes will run randomly (Synced, 2017).

There are two stages in Random Forest algorithm, first is random forest creation, the second is to make a prediction from the random forest classifier created in the first stage.

The whole process is described in step (a) to (h).

The pseudocode of Random Forest is described in step (a) to (e) below:

a. Randomly select "K" features from total "m" features where k << m.
b. Among the "K" features, calculate the node "d" using the best split point.
c. Split the node into child nodes using the best split.
d. Repeat the (a) to (c) steps until "I" number of nodes has been reached.
e. Build forest by repeating steps (a) to (d) for "n" number of times to create "n" number of trees.

Prediction is then made with the random forest classifier created.

The pseudocode for random forest prediction is shown in step (f) to (h).

f. Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target).
g. Calculate the votes for each predicted target.
h. Consider the high voted predicted targets as the final prediction from the random forest algorithm.

2.5 Related Work

Day by day, crime rate keeps increasing. Crime cannot be predicted because it is neither systematic nor random. Also, the modern technologies and hi-tech methods help criminals in achieving their bad deeds. According to Crime Records Bureau, some crimes like burglary, arson etc. have been decreased while crimes like murder, sex abuse, gang rape etc. have highly increased. Even though we cannot predict who the victims are but at least we can predict the places that has probability for the occurrence. The predicted results cannot be assured of 100% accuracy but the results/conclusions show that our application helps in reducing crime rate to a certain level by providing adequate security in crime sensitive areas. So, for building such a powerful tool, we have to collect crime records and evaluate it. Since the availability of criminal data or records is limited, we are collecting crime data from various sources like news sites, blogs, web sites, social media etc. This data is used as a record for creating a crime record database. So, the main challenge in front of us is developing a better, efficient crime pattern detection tool to identify crime patterns effectively. *Sathyadevan, Devan(2014),*

Duijn, Kashirin(2016), proposed that policymakers and law enforcement agencies across the world find it very difficult to find effective strategies or methods to control criminal activities. The effectiveness of disruption strategies is known to depend on both network topology and network

resilience. However, as these criminal acts operate in secrecy, data-driven knowledge concerning the effectiveness of different criminal network disruption strategies is very limited.

Nishat(2017) proposed that Criminal activities are present in every region of the world affecting quality of life and socio-economic growth and development. It is a major concern of many governments who are using different advanced technology to tackle such issues. Crime Analysis, a sub branch of criminology, studies the behavioral pattern of criminal activities and tries to identify the indicators of such events. Machine learning agents work with data and employ different techniques to find patterns in data making it very useful for predictive analysis. Law enforcement agencies use different patrolling strategies based on the information they get to keep an area safe. A machine learning expert can learn and analyze the pattern of occurrence of a crime based on the reports of previous criminal activities and can find hotspots based on time, type or other factors. This technique is known as classification and it allows us to predict some nominal class labels. Classification has been used on many different domains such as financial market, weather forecasting, business intelligence, healthcare, etc.

Yu et al (2017) provides the static maps with no interactive features. To overcome these limitations, the proposed framework provides the visualization techniques that consider the type of crime to identify the crime hotspots and helps to check these locations with the interaction features using Google maps.

Ahishakiye et. al and Iqbal et. al (2017) used the attributes population of country, Median Household income, percentage of people who are unemployed with age greater than 16, type of crime, etc. which only predicts whether in an area there will be high, medium or low percentage of violent crimes that can happen in future. The methods proposed by them didn't predict the type of crime that can happen.

Nasridinov et. al (2013) also proposed a method for classifying the crime rate as high, medium or low. None of them has classified the type of crime that can happen and its probability of happening.

Hyeon-Woo and Hang-Bong (2017) proposed the predictive method based on deep neural network. With their model they achieved the accuracy of 74.35%. The scientists used the algorithm of random forest regressor to measure the impact of urban factors on homicides and predict the future number of crimes of this type. However, the scientists note that the model developed works well only with small datasets.

Yanqing et al.(2018) have postulated a link between street lights and spatial criminal patterns. The researchers have determined the relationship between traffic and crime. The authors determined that the weather also influence the criminal activity.

Anneleen et al.(2017) explored 3 approaches to predictive modeling: logistic regression, neural network, and ensemble model. The obtained models were tested on 3 types of crime: home burglary, street robbery and battery. The results with the highest accuracy have been achieved using the logistic regression. Several models, developed not only in the environmental but also in temporal context, were proposed by researchers.

Shiju-Sathya(2014) proposed Apriori algorithm for the identification of criminal trends and patterns. This algorithm is also used to identify association rules in the database that highlight general trends. This paper also suggested the naïve Bayes algorithm by training crime data to create the model. The result showed after testing that the Naïve Bayes algorithm have 90% precision.

Zakir-Hussain et al.(2017) used the methods of information mining to analyze criminal conduct. This paper proposed tool for analyzing criminal investigation (CIA). Within the law enforcement community, this instrument was used to assist resolve violent offenses. This study is about the various type of crime scene. Both from an investigative and a behavioral perspective, the analysis

was done. It provided insight into the unknown criminals as well as recommendation for investigation and interview and trial strategies.

Chaithanya, Manohar and Issac(2014), describes Text detection is the method of locating areas in a picture wherever, text is present. Text detection and classification in natural pictures is very important for several computer vision applications like optical character recognition, distinguish between human and machine inputs and spam removal. Currently the challenge in text identifying is to detect the text in natural pictures due to many factors like, low- quality image, unclear words, typical font, image having a lot of color stroke than the background color, blurred pictures due to some natural problems like rain, sunny, snow, etc. The main aim of this work is to identify and classify the text in natural pictures. Here system detects the text and finds the connected regions, chain them together in their relative position. Uses a text classification engine to filter chains with low classification confidence scores.

Shah(2017) proposed that Vancouver is most populated city in Canada. It is most ethnically diverse cities in Canada. Crime is one of the biggest and dominating problem in our society and its prevention is an important task. Even though Vancouver known to be the safest city, it is observed that vehicle breakings and many more thefts is still a problem. There has been tremendous increase in machine learning algorithms that have made crime prediction feasible based on past data. The aim of this project is to perform analysis and prediction of crimes in states using machine learning models. It focuses on creating a model that can help to detect the number of crimes by its type in a particular state. In this project various machine learning models like K-NN, boosted decision trees were used to predict crimes. Area Wise geographical analysis can be done to understand the pattern of crimes. Various visualization techniques and plots are used which can help law

enforcement agencies to detect and predict crimes with higher accuracy. This will indirectly help reduce the rates of crimes and can help to improve securities in such required areas.

Azure is Microsoft's cloud computing platform that has been in service since 2010. The Azure platform provides more than 600 services, and you can create a model through Azure Machine Learning Studio, easily build a web service, and apply it to various devices. In addition, unlike existing cloud platforms and machine learning libraries and tools, it provides an easy-to-access GUI environment in consideration of user convenience. At Microsoft Ignite, it announced Azure Machine Learning designer's general availability, the drag-and-drop workflow capability in Azure Machine Learning studio, which simplifies and accelerates building, testing, and deploying machine learning models for the entire data science team, from beginners to professionals (azureml.net). Azure Machine Learning Studio natively supports data input, output, and visualization, and representative machine learning algorithms that data scientists love are prepared. Unlike existing machine learning tools and libraries, it has the advantage of being able to easily create a model by dragging and dropping blocks. In addition, scripts written in R and Python languages can be inserted and utilized in block form, and the results can be checked through visualization. Thanks to this easy structure, anyone who knows how to use it can easily create and deploy predictive models *(Kang et al., 2018).*

Venturini et al.(2016), in their paper have discovered spatio-temporal patterns in crime using spectral analysis. The goal is to observe seasonal patterns in crime and verifying if these patterns exist for all the categories of crime or if the patterns change with the type of crime. The temporal analysis thus performed highlights that the patterns not only change with month but also with the type of crime. Hence, the authors rightly stress the fact that models built upon this data would need to account for this variation. They have used the Lomb-Scargle periodogram to highlight the

seasonality of the crime as it deals better with uneven or missing data. The AstroML Python package was used to achieve this. In their paper they have described in detail how every category of crime performs when the algorithm is applied to the data. Further, the authors suggest that researchers should focus on the monthly and weekly crime patterns.

Kianmehr and Alhajj(2006) proposed that Support Vector Machine method can be used in hot-spots prediction to predefine a level of crime rate and given the percentage of data. A subset of the crime datasets is selected (percentage or number of data crimes) and classify each of selected data point based on the predefined level of crime rate. The data point which has above than predefined rate is positive or classified as hotspot class and the data point which have below than predefined rate is negative or non-hotspot class. The parameter selected will be used as the training in SVM classification. As a result, linear, polynomial and gaussian kernel function have been chosen to compare the performance of one-class SVM for predicting the hotspot crime of location. The datasets are from the Internet public datasets (2 datasets, Columbus and ST. Louis). However, using this method is still slow and computationally expensive even though it has been modified.

Shrivastav(2012) applied the Fuzzy time series in order to discover a crime pattern in community. To reduce the computational overhead in this method, the simplified process and model that include simple arithmetic operations has been used. This method utilized seventeen years historic data of crime incidents (cases of murder in Delhi City). Hence, the parameters used are actual registered cases and years. The three different sets of data obtained by partitioning it into five intervals (Scheme-I), ten intervals (Scheme-II) and twenty intervals (Scheme-III) implemented. The prediction of results obtained from Scheme-II and Scheme-III are quite acceptable while the result of scheme-I incline to slightly over forecast with an average absolute error (-0.323). However, this method works on binary transaction data only for example 0 or 1.

Chitsazan and Rahmani(2013) used artificial neural network (ANN) for crime prediction. This method was introduced by focusing on geographical areas that outperform traditional policing boundaries for crime prediction. Therefore, ANN can be educated using geographical clusters of crime data to ease predictive modeling. Hence, the scanning algorithm based on geographical crime incidents used to identify clusters with relatively high level of crime hotspots. Generally, ANN reveals a capacity to model the trends with each cluster. In this study, the dataset used are 18,498 violent incidents (criminal damages, violence against the person) included a number of variables related to time, day, month, location and weather. The result shows that the comparison of mean standard error (MSE) between ANN (9.94) and random walk (22.50). Thus, ANN predicts accurately more than random walk. Nevertheless, this method takes a long time in training phases.

B. Chandra et al(2008) introduced an unsupervised method using multivariate time series for analyzing a large number of crime data in different time points. This method usually is based on parametric Minkowski model and dynamic time wrapping (DTW). It has been proposed to find similar crime trends among various crimes cases and used the obtained information for future crime trends prediction. The effectiveness of this method has been proved using Indian crime dataset provided by the Indian National Crime Records Bureau (29 districts). The parameter of this model is the different of dimensions data weights. Their results showed the comparison between DTW with Euclidean (equal weight age) and DWT with parametric Minkowski model (different weight age). This method proposed for finding similar crime trends and predict crime trends efficiently where the dimensions of data do not have equal weight age. However, this method is difficult to handle a missing value in order to get more accurate results.

Liao et al(2013) introduced Bayesian Network based on Bayesian learning theory as one of crime prediction model. This method can be used to build effective mathematical models to understand

the behaviors of serial crimes. There are many potential factors affecting the selective of criminals for the future crime location. The model has been testing using serial crime dataset found in Gansu, China. Hence, the parameters used are characteristics of the victims (age, gender, jobs and race) and characteristics of crime areas (residence, school, bus stop, hotel and hospital). Furthermore, the given predicted areas (yellow, green and red color) will help police to arrest criminals but it depends on geographical factors selected. Nevertheless, this method is totally depending on selection of parameter. Therefore, it is may lead to some deviation during the experiment so more factors unrelated to the geography should be considered to improve the accuracy of the model.

Nasridinov et al(2015) proposed that decision tree is one of the supervised methods used to handle the complex classification and produce reasonable classification tree. It provides the tree structure of classification and divides a dataset into smaller subsets. The decision tree helps police to discover the crime pattern and predict the future trends. The final result represented as a tree with decision nodes and leaf nodes which decision nodes have two or more branches and leaf nodes represent a decision or classification. The dataset used in their testing and training phases were extracted from the Internet. The parameters used in dataset are heartbeat rate, facial emotion, state and voice tone. The result of crime cases has been classified into two classes such as neutral and danger. However, this method does not work well on all type of datasets.

Antolos et al(2013) applied a logistic regression model to examine the relationship between the predicting factors of crimes and burglary occurrence probability. The factors used including time of day (category data), day of the week (discrete), barriers (physical structures that will interfere an individual's egress from a targeted residence), connectors (the amount of access bridges, streets and pathways relative to the targeted residence) and repeat victimization (the occurrence of the offense at the same residence following the initial offense over the calendar year, same category

data). The datasets for testing were collected in burglary incidents in 2010 from a local police report of Florida City. The model has been shown the various degrees of significance in terms of predicting the occurrence within different ranges. However, the limitation of this approach is difficult to identify the probability of burglary activity and specific locations.

V. RESEARCH METHODOLOGY

1. Introduction

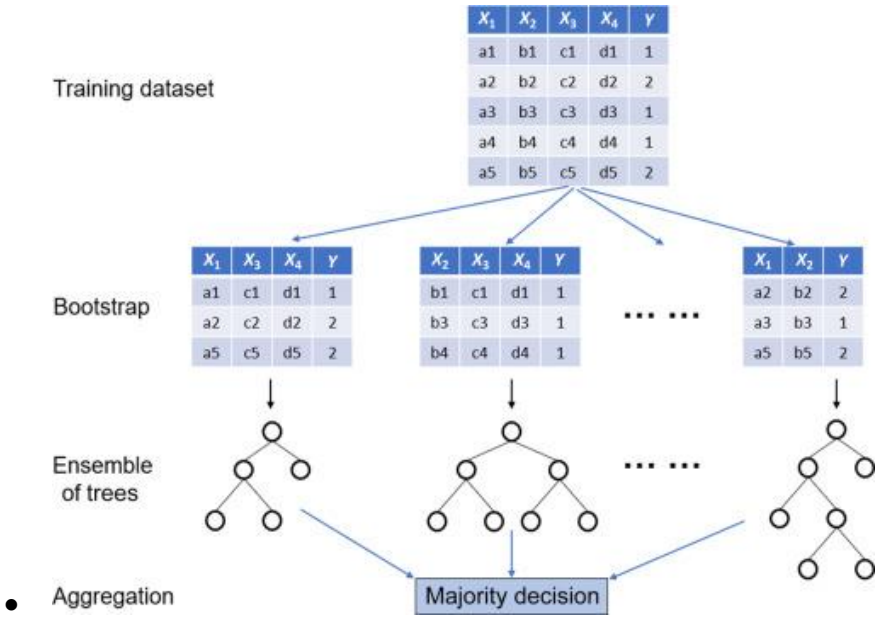
This Chapter explains in details how this project carried out prediction using the Random Forest algorithm. The Chapter also contains description of the crime rate dataset used.

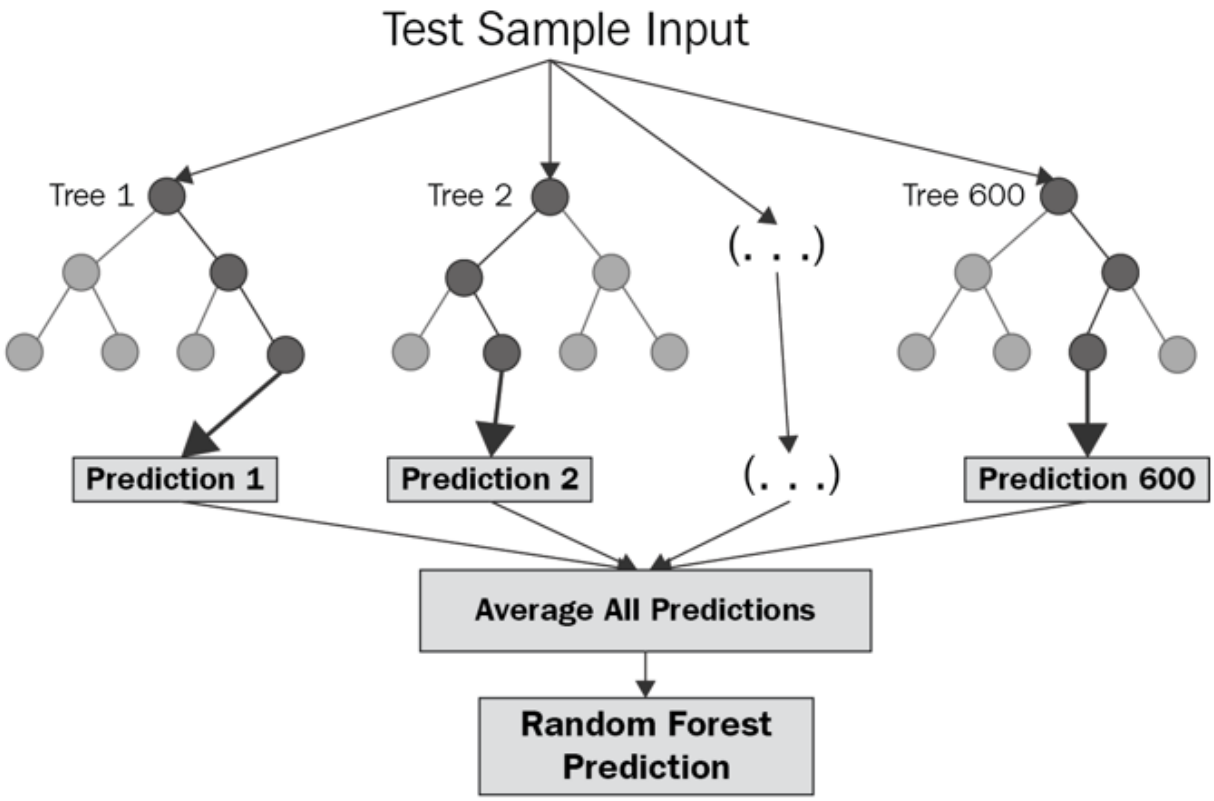
2. Description of the Dataset

The Buffalo crime dataset used in this project is open source and readily available on data.world. There are 25 columns and 161,224 rows in the dataset file. Buffalo is one of the most popular cities in New York. The data was downloaded from the internet on kaggle.com. This dataset is updated daily and offers a preliminary look at crime reports in the city of Buffalo. The source is https://data.buffalony.gov/d/d6q9-xbgu. This dataset was created by dataworldadmin.

3. Methodology

The Flow chart of the work carried out in this project is presented in the figure below.





The pseudocode for random forest algorithms can split into two stages;

1. Random forest creation pseudocode.

2. Pseudocode to perform prediction from the created random forest classifier.

Beginning with random forest creation pseudocode;

Random Forest pseudocode:

Randomly select “k” features from total “m” features.

Where k << m

Among the “k” features, calculate the node “d” using the best split point.

Split the node into child nodes using the best split.

Repeat 1 to 3 steps until “l” number of nodes has been reached.

Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.

The beginning of random forest algorithm starts with randomly selecting “k” features out of total “m” features. In the image, you can observe that we are randomly taking features and observations.

In the next stage, the randomly selected “k” features is used to find the root node by using the best split approach.

In the next stage, the child nodes will be calculated using the same best split approach.

Finally, stages 1 to 4 are repeated to create “n” randomly created trees. These randomly created trees form the random forest.

Random forest prediction pseudocode:

To perform prediction using the trained random forest algorithm uses the below pseudocode;

Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target).

Calculate the votes for each predicted target.

Consider the high voted predicted target as the final prediction from the random forest algorithm.

To perform the prediction using the trained random forest algorithm, the test features need to be passed through the rules of each randomly created trees. Suppose let’s say 100 random decision trees are formed to form the random forest.

Each random forest will predict a different targets (outcomes) for the same test feature. Then by considering each predicted target votes will be calculated. Suppose the 100 random decision trees are predicting some 3 unique targets x, y, z then the votes of x is nothing but out of 100 random0020

decision tree how many trees prediction is x.

Likewise, for the other 2 targets (y, z), If x is getting high votes, Let’s say out of 100 random

decision tree 60 trees are predicting, the target will be x. Then the final random forest returns the x as the predicted target. This concept of voting is known as majority voting.

DATA COLLECTION

The data was downloaded from the internet on kaggle.com. The repository contains a lot of dataset which can be used for research. The dataset downloaded was the past crime cases in Buffalo, a popular city in New York. There are 25 columns and 161,224 rows in the dataset file.

METHODS

Random Forest Classifier

Random Forest (or random forests) is a trademark term for an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. It is a collection of slightly different trees.

A random forest is a meta estimator that fits a number of classifiable decision trees on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. Each decision tree is constructed using a Random subset of the training data.

It is a supervised classification algorithm. It creates a forest by some ways and makes it random. There is a direct relationship between the number of trees in the generated forest and the results it can get. The larger the number of trees, the more accurate the result. But an important thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach (Dan, 2012). Decision tree is a decision support tool. It uses a tree-like graph to show the possible consequences. When a training dataset is supplied with targets and features into the decision tree, it will formulate some set of rules. These rules can be used to perform predictions (Synced, 2017).

The difference between Random Forest algorithm and the decision tree algorithm is that in Random Forest, the process of finding the root node and splitting the feature nodes will run randomly (Synced, 2017).

There are two stages in Random Forest algorithm, first is random forest creation, the second is to make a prediction from the random forest classifier created in the first stage.

The whole process is described in step (a) to (h).

The pseudocode of Random Forest is described in step (a) to (e) below:

a. Randomly select "K" features from total "m" features where k << m.
b. Among the "K" features, calculate the node "d" using the best split point.
c. Split the node into child nodes using the best split.
d. Repeat the (a) to (c) steps until "I" number of nodes has been reached.
e. Build forest by repeating steps (a) to (d) for "n" number of times to create "n" number of trees.

Prediction is then made with the random forest classifier created.

The pseudocode for random forest prediction is shown in step (f) to (h).

f. Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target).
g. Calculate the votes for each predicted target.
h. Consider the high voted predicted targets as the final prediction from the random forest algorithm.

Python Programming Language

Python is probably one of the easiest-to-learn programming languages in widespread use that fall within the Object Oriented family. It is very nice to use.

Python was adopted because of its:

(i) Flexibility.

(ii) Readability.

(iii) Platform independence.

(iv) Great Community Support.

(v) Great Llibrary Ecosystem.

Anaconda

Anaconda is a premium open source distribution of Python programming language for large-scale data processing, predictive analytics, and scientific computing, it aims to simplify package management and deployment.

Jupyter Notebook

The Jupyter Notebook is a free, open-source, interactive web application that allows user to create and share documents that contain live codes, equations, visualizations and narrative texts. Its uses include: numerical simulation, data visualization, data cleaning and transformation, statistical modeling, machine learning, and so on.

Pandas – Python Data Analysis Library. pandas are open-source, BSD-licensed libraries for the Python programming language that provide high-performance, simple-to-use data structures, and data analysis tools.

Numpy – NumPy is a scientific computing fundamental package in Python. It contains:

(i) a powerful N-dimensional array object.

(ii) sophisticated (broadcasting) functions.

(iii) tools for integrating C/C++ and Fortran code.

(iv) capabilities in linear algebra, Fourier transform, and random numbers.

NumPy can be used as a multi-dimensional container of generic data in addition to its apparent scientific applications. It is possible to define any number of data kinds. This enables NumPy to work with a wide range of databases with ease and speed.

sci-kit learn – Data mining and data analysis tools that are easy to use.

VI. PURPOSE OF THE STUDY

With the rapid urbanization and development of big cities and towns, the graph of crimes is also on the increase. This phenomenal rise in offences and crime in cities is a matter of great concern and alarm to all of us.

There are robberies, murders, rapes and what not. The frequent and repeated thefts, burglaries, robberies, murders, killings, rapes, shoplifting, pick pocketing, drug- abuse, illegal trafficking, smuggling, theft of vehicles etc., have made the common citizens to have sleepless nights and restless days.

They feel very insecure and vulnerable in the presence of anti-social and evil elements. The criminals have been operating in an organized way and sometimes even have nationwide and international connections and links.

VII. REFERENCE

I. Ihaka R., R:Past and Future history. Computing Science and statistics, 392396(1998).

II. Wang, B., Zhang, D., Zhang, D., Brantingham, P. J., & Bertozzi, A. L.(2017). Deep Learning for Real Time Crime Forecasting. arxiv preprint arXiv:1.707.03340.

III. Loecher, M. (2014). RgoogleMaps: overlays on Google map tiles in http://cran.r-project.org/web/packages/RgoogleMaps/inc.html

IV. Yu, R., Song, M., & Cui, E. San Francisco Crime Analysis and Classification

V. Ahishakiye, E., Taremwa, D., Omulo, E. O., Nairobi-Kenya, G. P. O., & Niyonzima, I. (2017). Crime Prediction Using Decision Tree (J48) Classification Algorithm. analysis, 6(03).

VI. Nasridinov, A., Ihm, S. Y., & Park, Y. H. (2013). A decision tree-based classification model for crime prediction. In Information Technology Convergence (pp. 531-538). Springer, Dordrecht.

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

VIII. H.-W. Kang and H.-B. Kang, "Prediction of crime occurrence from multi-modal data using deep learning," PLoS One, vol. 12, no. 4, p. e0176244, 2017.

IX. Y. Xu, C. Fu, E. Kennedy, S. Jiang, and S. Owusu-Agyemang, "The impact of street lights on spatial-temporal patterns of crime in Detroit, Michigan," Cities, no. October 2017, pp. 0–1, 2018.

X. Rummens, W. Hardyns, and L. Pauwels, "The use of predictive analysis in spatiotemporal crime forecasting : Building and testing a model in an urban context," Appl. Geogr., vol. 86, pp. 255–261, 2017.

XI. Kerr, J.: Vancouver police go high tech to predict and prevent crime before it happens. VancouverCourier,July23,2017.[Online]Availablehttps://www.vancourier.com/news/van cou ver-police-go-high-tech-topredict-and-prevent-crime-before-it-happens-1.21295288. Accessed 09 Aug 2018