New York City Crimes Detection using Machine Leraning

Ines Bougheriw, Eya Gourar, Karim Omrane et Rami Zouari
Higher School of Communications of Tunis
{eya.gourar, ines.boughariou, karim.omrane, rami.zouari}@supcom.tn

***Abstract*— Crime is one of the dominant and alarming aspect of our society. Everyday huge number of crimes are committed, these frequent crimes have made the lives of common citizens restless. So, preventing the crime from occurring is a vital task. In the recent time, it is seen that artificial intelligence has shown its importance in almost all the field and crime prediction is one of them. However, it is needed to maintain a proper database of the crime that has occurred as this information can be used for future reference. The capability to predict any crime on the basis of time, location and so on can help in protecting citizens from a possible assault of any kind. However, predicting the crime accurately is a challenging task because crimes are increasing at an alarming rate. Thus, the crime prediction and analysis methods are very important to detect the future crimes and reduce them. In Recent time, many researchers have conducted experiments to predict the crimes using various machine learning methods and particular inputs. For crime prediction, KNN, Decision trees and some other algorithms are used. The main purpose is to highlight the worth and effectiveness of machine learning in predicting violent crimes occurring in a particular region in such a way that it can be used to reduce crime rates in the society.**

Index Terms—Machine Learning, Crime Prediction, Random Forest, Decision trees.

I. INTRODUCTION

Crime is increasing considerably day by day. Crime is among the main issues which is growing continuously in intensity and complexity[1]. Crime patterns are changing constantly because of which it is difficult to explain behaviours in crime patterns[2]. Crime is classified into various types like kidnapping, theft murder, rape etc. The law enforcement agencies collects the crime data information with the help of information technologies(IT). With rapid increase in crime number, analysis of crime is also required. Crime analysis basically consists of procedures and methods that aims at reducing crime risk. It is a practical approach to identify and analyse crime patterns. Therefore a crime prediction and analysis tool were needed for identifying crime patterns effectively. This paper introduces a methodology with the help of which it can be predicted that at what place and time which type of crime has a higher probability of occurrence. Classification helps in extracting features and predict future trends in crime data based on similarities. In this study the used methodology is the Random Forest Classifier. The paper organisation is as follows. The introduction of the study is described in Section one. Section II consists of the related works. Section III discusses the methodology for crime prediction method. Section IV discusses its implementation. Section V consists of Conclusion.

II. RELATED WORK

Many crime-predictions algorithms has been proposed. The prediction accuracy depends upon on type of data used, type of attributes selected for prediction. IN[3], data collected from various websites, newsletter was used for prediction and classification of crime using Naive Bayes algorithm and decision trees and found that former performed better. In[4], a thorough study of various crime prediction method like Support Vector Machine(SVM), Artificial neural networks(ANN) was done and concluded that there does not exist particular method which can solve different crime datasets problems. IN[5], various supervised learning techniques, unsupervised learning technique[6] on the crime records were done which address the connections between crime and crime pattern for the purpose of knowledge discovery which will help in increasing predictive accuracy of crime. Clustering approaches were used for detection of crime and classification method were used for the prediction of crime, [7].

III. METHODOLOGY

Predictive modeling was used for making predictions since it has the method which is able to build a model and has the capability to make predictions. This method consists of different algorithms of Machine Learning that can study properties from the data used for training which is used for producing predictions. It is split in two major classes one is Regression and other is classification of patterns. Regression models are based upon analysis of the relationship that are present between trends and variable in order to make predictions about the continuous variables. Whereas, the job of classification is to assign a particular class labels to a data value as output of the prediction. Division of pattern classification is in two ways i.e., Supervised and Unsupervised learning. It is already known in supervised learning that which class labels are to be used for building classification models. In unsupervised learning, these class labels are not known. This paper deals with supervised learning.

A. Data collection and Pre-processing

Data collection is a process in which information is gathered from many sources which is later used to develop

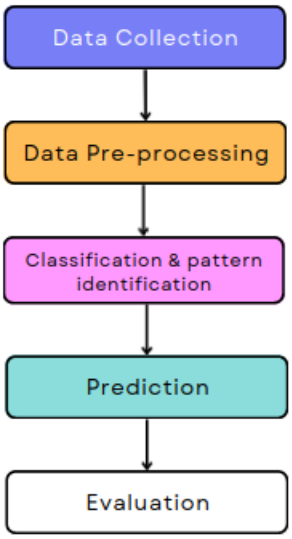

Fig. 1. Architecture

the machine learning models. The data should be stored in a way that makes sense for problem.

Data pre-processing basically involves methods to remove the infinite or null values from data which might affect the performance of the model. In this step the data set is converted into the understandable format which can be fed into machine learning models.

B. Model selection

Decision trees: Decision trees are one of the most popular and powerful tool for classification and prediction. It has a structure like a tree, where all of the intermediate node represents a test on a peculiarity and the end product of test is denoted by every branch, and label of class are held by every leaf node.

Random Forest is a machine learning algorithm that can be used for both regression and classification tasks. It is an ensemble of decision trees, which means that it uses multiple trees to make predictions. Each tree in the Random Forest is trained on a random subset of the data, and the final prediction is made by averaging the predictions of all the trees.

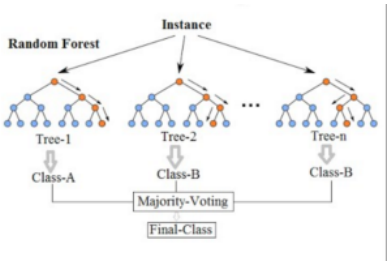

Fig. 2. Random Forest

C. Training and testing

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

IV. IMPLEMENTATION

A. Data collection

NYPD Complaint Data Historic : This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the end of last year (2019). The data contains 6901167 complaint and 35 columns including spatial and temporal information about crime occurences along with their description and penal description.

B. Exploratory Data Analysis

In order to get a hold of the data used in this study, an exploratory data analysis (EDA) is needed, in which graphics and visualizations are used to explore and analyze the data set. The goal is to investigate and learn about repeating patterns and possible explanations to why crimes are taking places the way they do.

Since the final goal of this study is to predict any crime on the basis of simple informations and to protect citizens from possible danger, we need to analyze the profiles of victims to have an idea what categories are more prone to be victims of crimes and such. For example, by plotting number of victims by race, it is clear to say that black people are more targeted than any other ethnic group, and combining that with sex specification we can plot the number of victims by race and sex fig4 and find that black females are the most prone to get assaulted.

Moreover, we may exploit Geo-spatial information and combine it with victims specifications as well. fig3

C. Data Pre-processing

In this step all the null values are removed. The categorical attributes are converted into numeric using Label Encoder which can be understood by the classification models. There exists some samples which are

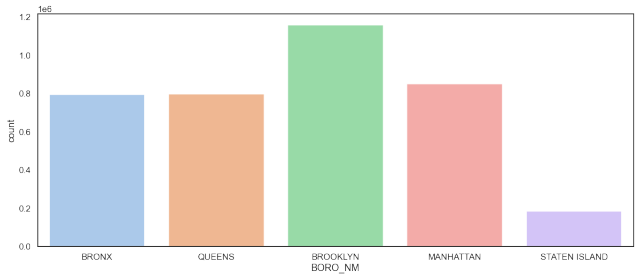

Fig. 3. Occurrence of incidents by borough

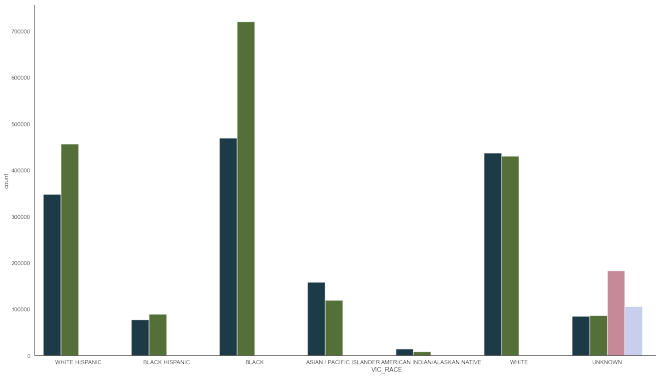

Fig. 4. Number of victims by Race and Sexe

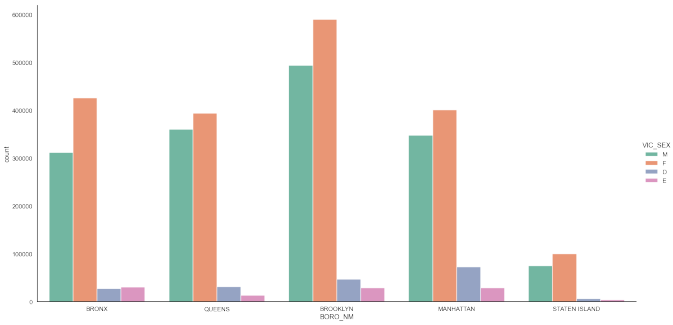

Fig. 5. Total number of victims by Sexe gender in the different boroughs

considered to be outliers, those samples have been removed after checking location of each point and if not in New York range then it was removed. There also exist 20 features ('SUSP_AGE_GROUP', 'IN_PARK', 'SUSP_SEX', 'SUSP_RACE', etc.)are considered redundant as they does not exist in testing values so they were removed.

Latitude	Longitude	Lat_Lon	PATROL_BORO	STATION_NAME
40.685041	-73.921777	(40.685040958, -73.921776995)	PATROL BORO BKLYN NORTH	NaN
40.636991	-74.134093	(40.63699139, -74.134092508)	PATROL BORO STATEN ISLAND	NaN
40.823876	-73.891863	(40.823876276, -73.891862968)	PATROL BORO BRONX	NaN
40.845707	-73.910398	(40.845707148, -73.910398033)	PATROL BORO BRONX	NaN
40.763992	-73.828426	(40.763991557, -73.828425559)	PATROL BORO QUEENS NORTH	NaN

Fig. 6. Data before pre-processing

There are features from data that are not numbers, so they are converted into numbers so that we can train the models on them by using Label Encoding and One-hot Encoding. One-hot Encoding might produce very high number of dimensions due to lot of data labels in very feature but even then this is better because problem with the Label Encoding technique is that it assume higher the categorical value, better is the category which results in more errors.

	PATROL_BORO_0	PATROL_BORO_1	PATROL_BORO_2	PATROL_BORO_3
7	0	0	0	1
17	0	0	1	0
27	0	0	1	1
29	0	1	0	0
31	0	1	0	1

Fig. 7. Data after pre-processing

D. Training

For training the data splitted in the ratio of 80% for training and 20% for testing. As a result we train size of around 32k data points and test size of around 8k data points. After trying different combinations of parameters for the model, different models were trained and their f-score, accuracy were calculated.

E. Model evaluation and Metrics

For evaluating classification models that were implemented for the purpose of classification and prediction, the metrics used are accuracy, f-score.
Precision is a measure which identifies positive cases from all the predicted cases.

$$q = \frac{tv}{tv + fv} \tag{1}$$

Next is recall it measure which correctly identifies positive cases from all the actual positive cases.

$$rc = \frac{tv}{tv + fnv} \tag{2}$$

accuracy is one of the most commonly used metric which measure all the correctly identified value without caring about the wrongly identified values. So, instead of using accuracy the measure that is used to check the performance is F-score.

$$accuracy = \frac{tv + tnv}{tv + fv + tnv + fnv)} \tag{3}$$

F-Score is the harmonic mean of Recall and precision which gives a better measure of incorrectly classified cases than that of Accuracy Metric.

$$F - Score = \frac{(2 * (rc * q))}{(rc + q)} \tag{4}$$

Here tv stands for true positive, fv stands for false positive, fnv stands for false negative, tnv stands for true negative,rc stands for recall and q stands for precision.

TABLE I
ERFORMANCE OF MODEL DURING TRAINING

Model	accuracy	F-score
Random Forest	0.98	19.315%

V. CONCLUSIONS

Crime prediction is one the current trends in the society. Crime prediction intends to reduce crime occurrences. It does this by predicting which type of crime may occur in future. Here, analysis of crime and prediction are performed with the help of Random Forest. From the results obtained we can see that for this data set it is working best with optimal training and good accuracy. However which model will work best is totally dependant on the dataset that is being used.

REFERENCES

[1] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi and A. Pent-land, “Once upon a crime: towards crime prediction from demographics and mobile data”,IEEE,Proceedings of the 16th international conference on multimodal interaction, 2014, pp. 427-434.

[2] Ubon Thansatapornwatana,"A Survey of Data Mining Techniques for Analyzing Crime Patterns", Second Asian Conference on Defense Technology ACDT, IEEE,Jan 2016, pp. 123–128.

[3] Shiju Sathyadevan,Devan M. S.,Surya S Gangadharan, First,"Crime Analysis and Prediction Using Data Mining" International Conference on Networks Soft Computing (ICNSC), 2014.

[4] Sunil Yadav, Meet Timbadia, Ajit Yadav, Rohit Vishwakarma and Nikhilesh Yadav,"Crime pattern detection,analysis and prediction,International Conference on Electronics, Communication and Aerospace Technology(ICECA), 2017.

[5] Amanpreet Singh,Narina Thakur,Aakanksha Sharma,"A review of supervised machine learning algorithms",3rd International Conference on Computing for Sustainable Global Development,2016

[6] Bin Li,Yajuan Guo,Yi Wu,Jinming Chen,Yubo Yuan,Xiaoyi Zhang,"An unsupervised learning algorithm for the classification of the protection device in the fault diagnosis system",in China International Conference on Electricity Distribution (CICED),2014

[7] R. Iqbal, M. A. A. Murad, A. Mustapha, P. H. Shariat Panahy, and N. Khanahmadliravi, "An experimental study of classification algorithms for crime prediction," Indian J. of Sci. and Technol., vol. 6, no. 3, pp. 4219- 4225, Mar. 2013.