ANALYSIS AND COMPARISON OF THE CRIME DATA WITH MACHINE LEARNING MODELS USING JUPYTER NOTEBOOK

Bharadwaj Varma Chennamraju
Computer Science, Lakehead University
Thunder Bay, Canada

***Abstract*—Crime is one of the major issues which is continuing to grow in intensity and complexity. The rapid change in urbanization and development in cities and towns have increased the number of crimes caused over the years. To find the pattern of crime, significant data plays a vital role. The main objective of the project is to reduce the criminal activities by analyzing the crime datasets of Toronto, Chicago, Ireland, Indore, and Youtube data based on crime videos and applying machine learning prediction models such as Decision Tree, K-Nearest Neighbor, and Random Forest on Toronto, Chicago, and Indore crime dataset to find out which prediction model has the best accuracy.**

***Index Terms*—Crime, Analysis, Machine learning, Prediction.**

I. INTRODUCTION

Crime is one of the major issues which is continuing to grow in intensity and complexity. The rapid change in urbanization and development in cities and towns have increased the number of crimes caused over the years. According to official documentation, there has been a remarkable rise in offenses and crime in various cities. To find the pattern or behavior of crime, significant data plays a vital role. Also, analysis of the crime data over the years on different datasets helps to find a very good pattern of criminal behavior.

Data analysis is a process of obtaining raw data and converting it into useful information. The goal is to discover useful information by inspecting, cleansing, transforming, and modeling the data. Machine learning performs effectively on a specific task without using explicit instructions, relying on patterns and inferences instead. It mainly focuses on prediction models, based on known properties learned from the training data. Deep learning, being a subset of the machine learning method, is used in artificial neural networks. Deep learning is mainly used in transforming the input data into a slightly more abstract and composite representation.

II. PROBLEM STATEMENT

The effect of crimes on victims is enormous. A victim might experience several kinds of effects such as emotional effect, physical injury, and mental injury. The most common types of crimes are theft, assault, battery, robbery, and homicides. Data analysis, along with machine learning and deep learning models, will help find the pattern of crimes caused concerning different locations. The dataset has various crime-related attributes such as latitude, longitude, offense caused, major crime indicator, date, etc. Analysis of crime data would help solve a real-world issue. For comparison, Toronto, Ireland, Chicago, and Indore are used. The officially updated government dataset is used and it is updated every day.

The main objective of this project is analyzing the crime data that had occurred over the years, understanding the crime pattern, classifying crime based on location, visualizing the dataset, and applying prediction modeling techniques between top crimes.

III. RELATED WORK

In paper [1], the potentiality to predict the crime pattern based on location and time will help law enforcement in preventing the crime and arresting the offenders. This paper gives an overview of crime prediction methods such as support vector machine, multivariate time series, and artificial neural network. Thorough review and summarization of each model are done to provide an accurate prediction of crimes.

In paper [2], analysis of Vancouver crime data is done using machine learning predictive model. It mainly focuses on two prediction models that are K-nearest-neighbor and boosted decision tree. Dataset has the crime data of the last 15 years where they approached in two different ways. According to the evaluation results, both prediction models gave an accuracy between 39% to 44% which is low but can be increased by tuning the algorithm and reducing the data.

In paper [3], classification is done to the dataset which is taken from different states of the United States of America for the prediction of Crime Category. The dataset was gathered from the Law enforcement data from the 1990 US LEMAS survey, socio-economic data from 1990 US census, and the 1995 FBI UCR. Nave Bayesian and Decision Tree algorithms are implemented for predicting crime and comparison is done between these two algorithms. The result from the analysis shows that Decision Tree outperformed Nave Bayesian algorithm by achieving an accuracy of 83.9519% whereas Nave Bayesian achieved only 70.8124% of accuracy.

In paper [4], the forecasting of crime is mainly based on the autoregressive model. For this, Chicago crime dataset is considered to analyze crime patterns and trends. Components such as seasonal, trend, observed, and random were observed using the ARIMA model. Based on 2013 and 2014 crime data, testing is done to find the accuracy of the prediction model. According to the evaluation results, for one year it gave an accuracy of 84% and 80% for a two year ahead forecasting.

IV. Methodology

A. Data Acquisition

The datasets are available on the government website of Toronto, Chicago, Ireland, Indore, and Youtube based on crime videos. The datasets are stored in the form of .csv file in Microsoft Excel.

B. Data Cleaning

As the datasets were taken from the government websites, there was not much cleaning required to do as they were updated perfectly. Some of the attributes in a few datasets were not taken into consideration. Different datasets had different attributes that were not needed for data analysis.

C. Data Visualization

As the datasets are cleaned, visualization is performed using Jupyter Notebook. To use Jupyter Notebook, Anaconda should be downloaded and installed.

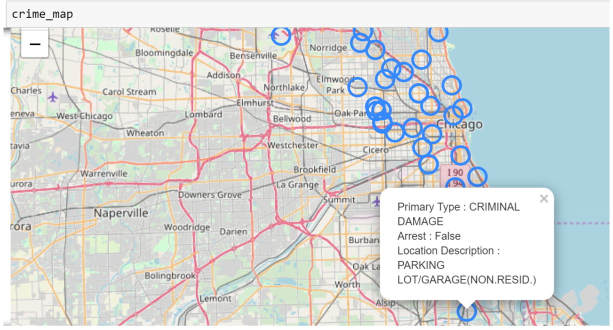

Fig. 1. Mapping of crimes in Chicago.

Fig. 1 gives information about the type of crime based on location in Chicago. Fig. 2, shows the different types of crimes that have occurred in various places.

Fig.3 gives information about Theft offenses at the top five stations in Ireland. When compared with the other stations from the year 2013 to 2018, Pearse Station and Blanchardstown have a high rate of theft-related offenses.

The Youtube dataset is collected from crime videos of Toronto posted on Youtube. Fig 4 gives information about the different type of crimes that occurred in Toronto.

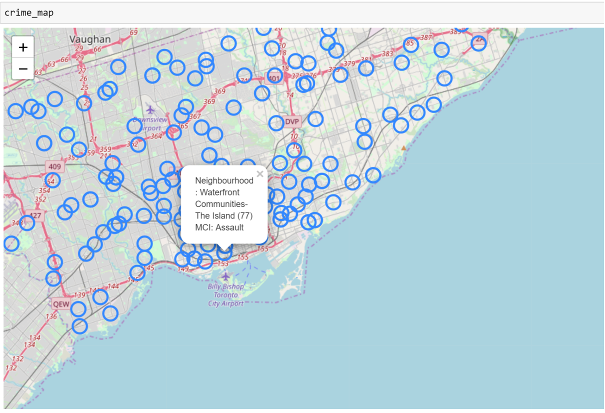
crime_map
+
−
Vaughan
DVP
QEW
Billy Bishop Toronto City Airport
Neighbourhood : Waterfront Communities-The Island (77)
MCI: Assault

Fig. 2. Crime mapping in Toronto.

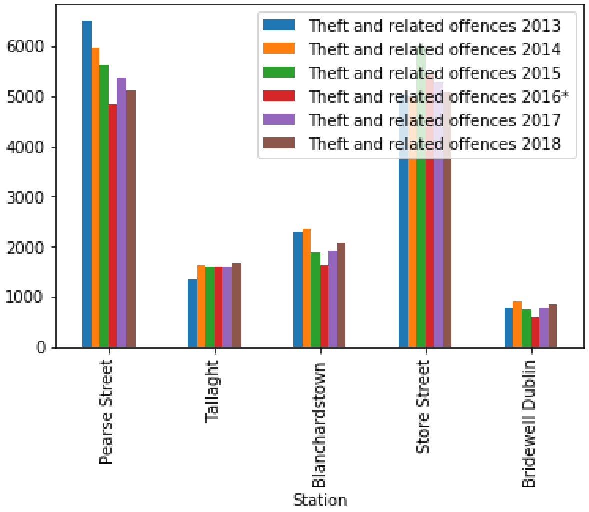

Fig. 3. Theft offences in Ireland.

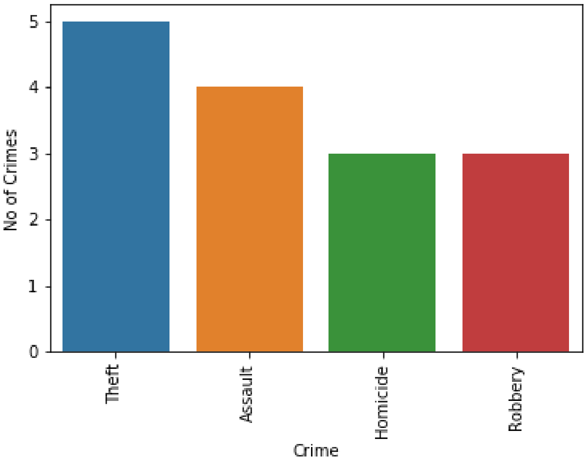

Fig. 4. Different type of crimes based on Youtube.

V. Implementation

A. JUPYTER NOTEBOOK AND ITS LEARNING MODELS

Jupyter Notebook is a non-profit, open-source software that allows the user to run the code, look at the outcome, visualize the data, and see the results without leaving the environment. It also helps in handling the data by making a handy tool for performing an end to end data science. Data cleaning, transforming, statistical modeling, building, visualizing, and training machine learning models can be done by all kind of machine learning users.

In Machine Learning, models are categorized into supervised and unsupervised. Generally, the problems are divided into regression and classification. Regression problem is used when the data contains numerical values, and the classification problem is used when the data contains categorical values. Based on the type of problem, prediction modeling and evaluation can be done using Jupyter Notebook.

B. Data Modelling

Datasets used for modeling are Toronto, Chicago, and Indore. For each dataset classification model is applied for prediction. Algorithms such as Decision Tree, K-Nearest Neighbor, and Random Forest classification models are used for modeling the datasets.

a) Decision Tree: Decision Tree classification model is applied on datasets of Toronto, Chicago, and Indore. For Toronto dataset, the accuracy is about 58.356%. For Chicago, the accuracy is about 74.944%. For Indore, the accuracy is about 98.067%.

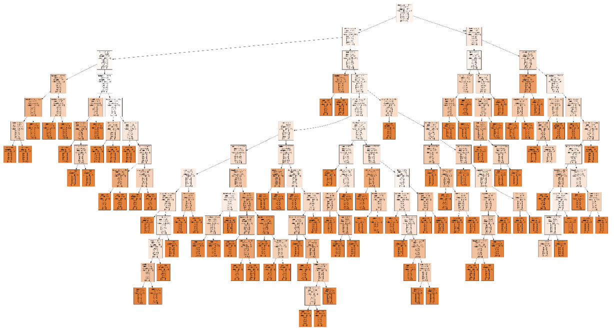

Fig. 5. Decision Tree model of Indore.

b) K-Nearest Neighbor: K-Nearest Neighbor classification model is applied on datasets of Toronto, Chicago, and Indore. For Toronto dataset, the accuracy is about 56.764%. For Chicago, the accuracy is about 60.26%. For Indore, the accuracy is about 93.236%.

c) Random Forest: Random Forest classification model is applied on datasets of Toronto, Chicago, and Indore. For Toronto dataset, the accuracy is about 65.225%. For Chicago, the accuracy is about 74.944%. For Indore, the accuracy is about 98.067%.

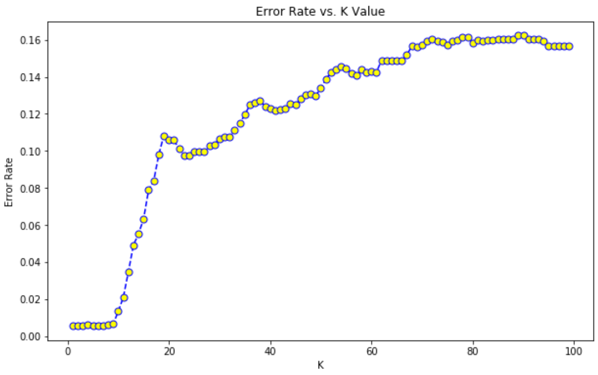

Fig. 6. Error Rate vs. K Value.

Conclusion and Future Scope

According to analysis, different cities have high rates of various crimes. On comparing the datasets, theft and assault crimes occurred more when compared to other crimes. Also, outside crimes are more than an apartment or in any commercial area.Church-Yonge Corridor area and North State Street has the highest crime rate in Toronto and Chicago, respectively when compared to other regions. In Ireland, Pearse station has high crime rate of assault, fraud, and theft while Tallaght station has high crime rate of burglary. On comparing the classification models, Random Forest predictive model outperformed K-Nearest Neighbor and Decision Tree classification models.

The future scope for this project is linking the prediction models to a web-application for predicting the type of crime. Also, linking the analysis to a real-time application to analyze the crime-prone areas and alerting through the app when passing through an area helps an individual to be more cautious.

Acknowledgment

The satisfaction and joy in taking the privilege to express my gratitude for the successful completion of my project would be meaningless without mentioning the people who made it possible. I express my sincere thanks and gratitude to supervisor Dr. Jinan Fiaidhi, Professor, Department of Computer Science, Lakehead University, for guiding and encouraging me during the project work. I also express my sincere thanks and gratitude to our Department of Computer Science, Lakehead University, for having extended helping hand always. Finally, I want to take this golden opportunity to thank my family members and friends who have been the most significant source of support for me.

References

[1] Nurul Hazwani Mohd Shamsuddin, Nor Azizah Ali, Razana Alwee, *"An overview on crime prediction methods"*, 6th ICT International Student Project Conference, 2017.

[2] Suhong Kim, Param Joshi, Parminder Singh Kalsi, and Pooya Taheri, *Crime Analysis Through Machine Learning*, IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2018.

[3] Iqbal, Rizwan and Azmi Murad, Masrah Azrifah and Mustapha, Aida and Panahy, Payam Hassany Shariat and Khanahmadliravi, Nasim *An experimental study of classification algorithms for crime prediction* , Indian Journal of Science and Technology, 6 (3). pp. 4219-4225. ISSN 0974-6846; ESSN: 0974-5645, 2013.

[4] Eugenio Cesario, Charlie Catlett, Domenico Talia, *"Forecasting Crimes using Autoregressive Models"*, IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), 2016.

[5] S.Sivaranjani , Dr.S.Sivakumari, Aasha.M, *"Crime Prediction and Forecasting in Tamilnadu using Clustering Approaches"*, International Conference on Emerging Technological Trends [ICETT], 2016.

[6] Qiang Zhang, Pingmei Yuan, Qiyun Zhou,Zhiming Yang *"Mixed Spatial-Temporal Characteristics Based Crime Hot Spots Prediction"*, IEEE 20th International Conference on Computer Supported Cooperative Work in Design (CSCWD), 2016.

[7] Abba Babakura, Md Nasir Sulaiman and Mahmud A. Yusuf, *"Improved Method of Classification Algorithms for Crime Prediction"*, International Symposium on Biometrics and Security Technologies (ISBAST), 2014.

[8] Ayisheshim Almaw, Kalyani Kadam, *"Crime Data Analysis and Prediction Using Ensemble Learning"*, Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018.

[9] Dr.J.Kiran, Kaishveen.K, *"Prediction Analysis of Crime in India Using a Hybrid Clustering Approach"*, 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018.