Chicago Crime Data Analysis and Predictions with Pyspark

Group 14

Arunima Divya*
AM.EN.U4AIE19016
Department of Computer Science & Engineering
Amrita Vishwa Vidyapeetham
Amritapuri Campus, India
arunimadivya@am.students.amrita.edu

Amrita Varshini E R
AM.EN.U4AIE19010
Department of Computer Science & Engineering
Amrita Vishwa Vidyapeetham
Amritapuri Campus, India
amritavarshinier@am.students.amrita.edu

Ann Maria John
AM.EN.U4AIE19013
Department of Computer Science & Engineering
Amrita Vishwa Vidyapeetham
Amritapuri Campus, India
annmariajohn@am.students.amrita.edu

Namitha S
AM.EN.U4AIE19042
Department of Computer Science & Engineering
Amrita Vishwa Vidyapeetham
Amritapuri Campus, India
namithas@am.students.amrita.edu

Devi Parvathy Nair
AM.EN.U4AIE19026
Department of Computer Science & Engineering
Amrita Vishwa Vidyapeetham
Amritapuri Campus, India
deviparvathynair@am.students.amrita.edu

Abstract— **Rapid urbanisation has been hampered by rising crime rates, and it is critical to reduce crime rates in order to improve urban safety. Over the last 20 years, the increase in crime has been even more closely linked to the rise in mass unemployment. Many emerging tools are assisting police forces in locating and tracking crime patterns. The project aims to conduct an exploratory data analysis of the Chicago Crime Dataset based on Apache Spark platform. The data is then used to develop classification models using different machine learning methods aiming to predict the likeness of a crime based on some selected features of the data.**

I. INTRODUCTION

Every year crime rates around the world are growing. Security agencies and the police work hard to identify, catch and reduce these crimes, but due to the great rate at which crime is increasing, they are in a great struggle. Higher crime rates have not only added stress to the government in terms of protecting the people, but it is also stunting the growth of urban areas because of lack of safety. To better protect communities, law enforcement is required to do more analysis on the crimes in their areas. But to do this, a system that can handle large amounts of data of several years must be accessible.

This is where new techniques to study crime is necessary as it's very important that a system that can handle large amounts of data is available to use. The increase of computerized methods allows for better tracking, and data analysts are able to use this to speed up the process of solving crimes. In addition to better technology, data collection has also drastically improved. Cellular phones, GPS devices, and more have enabled police with capturing more information about reported crimes. Current computers cannot easily handle such huge collections of data, so this is where big data comes into play. Thousands of petabytes of data is generated per day around the world, and Big Data analytics is a crucial part of handling the data. Big data is used by various industries to deal with the large inflow of data, while also analyzing and predicting future patterns. Different softwares, such as Apache Spark, Hadoop, MapReduce, MongoDB, etc. are being used to solve major data problems, including crime analysis. We utilize Apache Spark to read the crime dataset provided by Chicago Police Department's CLEAR. With the use of statistical algorithms and machine learning, many models have been generated to predict crime, and many are employed in countries like U.S., UK, and China. Machine learning is a branch of data science that uses data to analyse, detect patterns, and make decisions based on the learned information. Machine learning models are great for predicting future behaviour, whether it is for predicting locations of certain crimes, the time of day during which crimes are at a peak, or what type of crime is mostly

*Corresponding author's contact: 7306393655 likely to happen. In addition to analyzing the crime patterns, we also designed models that would predict what crime is likely to happen. In this project, we aimed to combine the use of Big Data analytics and machine learning in a study of the crime patterns in Chicago over the time range of 2001-2018. We use Apache Spark to read the data, then we involve Python and 4 machine learning algorithms, Logistic Regression, Naive Bayes, Random Forest, and Decision Tree Regression, to analyze crime patterns, and predict what type of crime is likely to occur.

II. LITERATURE REVIEW

The world's rapid urbanisation is changing people's lives in terms of economy and social interaction, which also has been hampered by rising crime rates, and it is critical to reduce crime rates in order to improve urban safety. When cities become big, the major challenges arise because current law enforcement agencies lack the specialised resources and technology needed to differentiate between meaningful current criminal activity. Hence, it is very important to understand and implement ways where we can perform an easy analysis and predictions of crime types and papers. For analysis and crime pattern predictions, Palash Sontakke and Chang-Soo Kim, in their work [1], have used machine learning models and Apache Spark. With the aid of the logistic regression model in the MLlib libraries of Apache Spark, they have presented an overview of crime on an actual timestamp basis as well as a predictive crime trend according to the areas of Chicago city. For the experimentation, crimes that occured in Chicago from the year 2001 and 2018 were taken into account. A prediction model, that considered the features in the data description which later was normalised, was also implemented to predict the most occurring crimes in respective areas. The results showed that the proposed model predicted the crime that is most likely to occur within the city with an accuracy of 75.80%, that is calculated by the hypertuning of the logistic regression model. This proposed model showed its excellence when compared to other implemented prediction models as it was able to predict the repeating crimes in the top neighbouring areas of Chicago rather than just finding the highly occurring crimes without considering the areas as in other implementations.

Tirthraj Chauhan and Rajanikanth Aluvalu have also put forward efforts in their work [2] to propose an idea about how to use big data analytics to analyse crime-related data in order to create crime forecasting models. They have introduced the idea of crime mapping which is a technique for analysing, mapping, and visualising crime events or patterns in order to forecast their occurrence thereby assisting intelligence and other agencies in allocating resources appropriately to deter crime. As the historical data on criminal activity is critical for crime mapping and prediction of criminal hotspots, it is important to analyse them which is a tedious work when using traditional data handling methods. As a result, Big Data Analytics can be used to manage large amounts of unstructured or semi-structured data. Crime mapping was classified into three phases. In the first phase of crime mapping, in order to distribute the data over geographical areas and visualise the patterns of geographically distributed data, the R programming language was used. On the basis of distributed data, the Kernel Density Estimation technique was used to estimate or create clusters. In the second phase, cluster analysis was performed by the Hadoop environment using the GAMMA Test and parallel processing on various clusters. The third phase involved inputting the identified cluster from Hadoop to the neural network for criminal prediction purposes using the regression tree prediction, specification and classification. The output from the third phase forecasted the crime rate at various locations where the risks of crime incidence are large.

Another study conducted to analyze the crime pattern is shown in the work [3] by Khin Nandar Win, Jianguo Chen et al, have highlighted the importance of criminal statistical analysis and is also a great medium for demonstrating the adept utilisation of Apache Spark as a platform. Criminal Statistical Analysis is an essential input for assessing quality of life and the human rights situation in a society. A parallel implementation of a Fuzzy C-Means Clustering based Clustering Algorithm for Criminal Activity is proposed which employs data mining and machine learning technology for analysis. The dataset utilised, the Global Terrorism Database, is a vast open-source database and contains information of terrorist events throughout the world from 1970 to 2017. The unique feature of this database is that it has both domestic and international terrorist incidents which amounts to more than 180000 cases. The paper has observed that the utilisation of Apache Spark cloud computing platform has significantly accelerated crime data mining due to the very large dataset as it has been proven to be adept in tackling such large datasets.

III. METHODOLOGY

To understand the crime patterns, an analysis was conducted on various aspects, such as, time, location, arrests, and so on. Based on this analysis, we selected features that seemed to best provide insight towards a crime type and these features were used by the machine learning models for crime prediction. The selected features were :

- Arrest
- Domestic
- District
- Ward
- Community Area
- FBI Code
- Hour
- Day of the Week
- Month
- Day of the Month

Apache Spark [4] is a data processing system that can handle large data sets rapidly and spread processing tasks across many devices, either on its own or in conjunction with other distributed computing resources. Apache Spark has become one of the world's most common big data distributed processing frameworks. Spark supports SQL, streaming data, machine learning, and graph processing and comes with native bindings for Java, Scala, Python, and R. It also supports SQL, streaming data, machine learning, and graph processing.

Using the PySpark[5] library's variety of machine learning classification methods, we conducted a prediction of the likeness of a crime based on the aforementioned features.

The 4 machine learning methods used for predictions are: Multinomial Logistic Regression, Naive Bayes, Decision Tree Regression, and Random Forest.

A. *Multinomial Logistic Regression*

Multinomial logistic regression(MLR), also known as softmax regression, is a method that extends from logistic regression to work for multi-class classification [6]. It modifies binomial logistic regression, which predicts binomial probability, to predict a multinomial probability. The 2 modifications made are changing the loss function from log loss to cross -entropy loss, and changing output to predict the probability for each class label instead of a single probability [7].The main idea behind MLR is a linear predictor function is generated to construct a score from a set of weights that are linearly combined with the given features using a dot product. MLR is often used for analysis since it does not assume normality, linearity, or homoscedasticity.

B. *Naive Bayes*

A simple Bayesian network classifier, Naive Bayes, is one of the most commonly used machine learning methods for classification [8]. The Naive Bayes classifier is robust and has proven to be a great classifier in various real-life classification tasks. Naive Bayes is a probabilistic classifier built off of Bayes' theorem. The Bayes theorem is written as shown in Equation 1 [9]:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \tag{1}$$

The probability of A happening given B can be determined by the aforementioned theorem by taking B as the evidence and A as the hypothesis. For example, A is the class variable(crime type), which represents what the type of crime is given the tokens, and B would represent the parameters/features. The classifier assumes strong independence, thus presuming that the absence or presence of a certain feature is independent of any other feature of the class. The major drawback of the Naive Bayes method is that is assumes conditional independence with all the features. Considering how some features may be reliant on each other, for example, how the district and location description are related, this may lead to issues as these features cannot easily be considered as independent of one another.

C. *Decision Tree Regression*

Decision trees build regression or classification models in the form of a tree structure. It examines an object's characteristics and trains a model in the shape of a tree to forecast future data and generate measurable continuous production [10]. It incrementally divides a dataset into smaller and smaller subsets while also developing an associated decision tree. The end result is a tree with leaf nodes and decision nodes. Decision tree algorithm is a tree-structured classifier with three different kinds of nodes. The Root Node is the initial node that represents the entire sample and can be further subdivided into nodes. Interior Nodes represent data set attributes, while branches represent decision laws. Finally, the result is represented by the Leaf Nodes. This algorithm is extremely useful for dealing with decision-making issues [11].

D. *Random Forest*

Random Forest (RF) classification [12] is one of the most popular ensemble algorithms which has been proved highly efficient in developing highly accurate models for big data. This method is derived from the concept [13] of how a combination of learning models into one can guarantee a much better output. Here, in the RF algorithm, this idea is realised using the multiple decision trees which are generated at the time of training and the final output of the RF model is nothing but the merger of all the outputs from each of the generated decision trees. Moreover, another reason which leads RF models to be more accurate is the randomness added to the model while growing the trees. The best feature is chosen from a random subset of features every time a node is split. A considerably good number of trees in an RF model assures lesser chance of overfitting, a common problem faced while executing most of the learning-based algorithms.

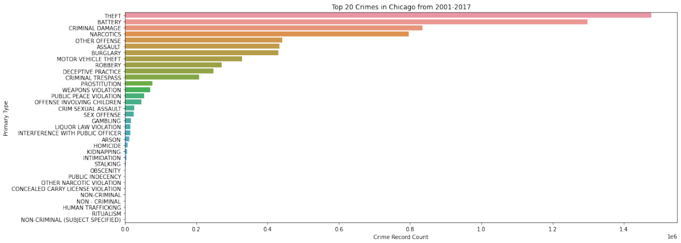

Fig. 1. Top 20 crimes with count

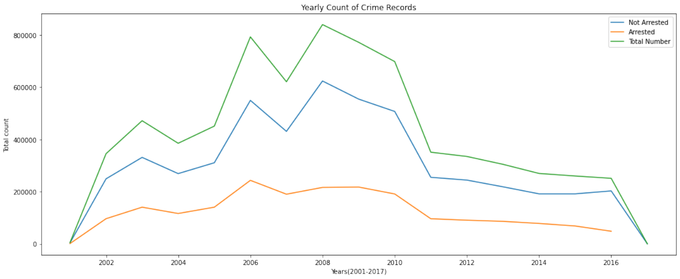

Fig. 2. Yearly crime count arrested, not arrested, and total

IV. DATASET

The dataset used in this study was the Chicago Crime Dataset [17], which was obtained from the Chicago Police Department's CLEAR(Citizen Law Enforcement Analysis

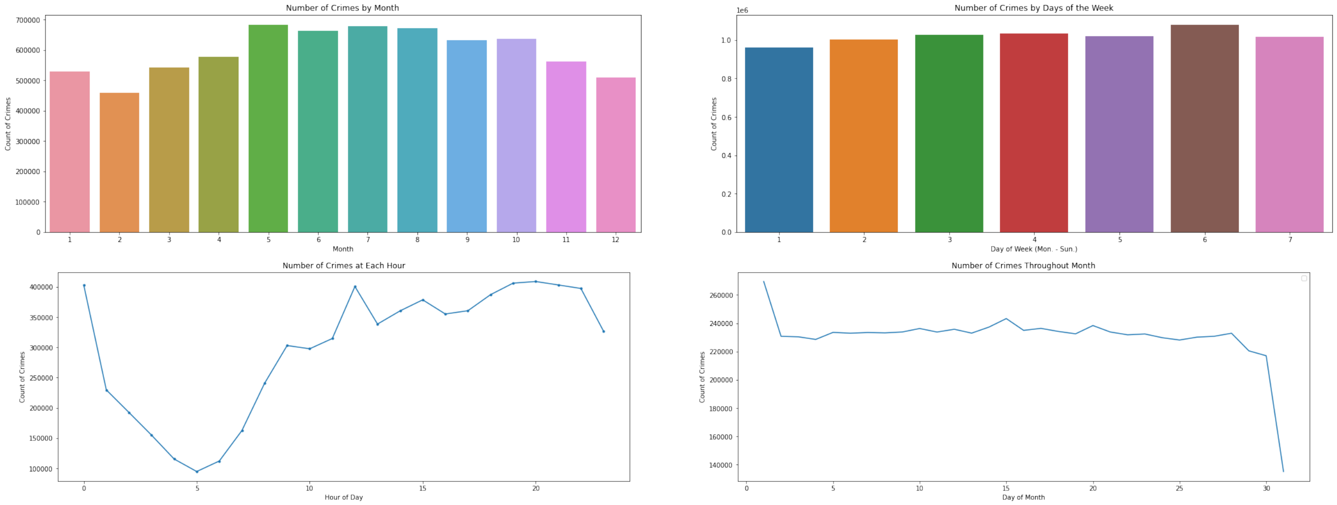

Fig. 3. Time-series evolution of crime count

and Reporting) system. The dataset consists of reported incidents of crime that occurred between 2001 and 2017. There are over 7 million records of data within the dataset. There are 23 columns that detail the incident reported. Several features, such as, ID number, case number,crime description, location coordinates, primary type, district, date of incident, etc.. are listed to provide more insight regarding each incident.

V. Results

For crime pattern analysis, we have used the Python programming language in order to implement Apache Spark, which is an open-source platform that is scalable, fast, and works with programming languages like Java, Scala, Python, and R. We conducted crime analysis on an annual, monthly, weekly, daily, and hourly basis during the entire process. The primary goal of this investigation is to determine the seriousness of illegal activity and how it has changed over time in Chicago. The first analysis as shown in Figure 1 depicts the total number of cases of each crime type reported from the year 2001 upto 2017. This analysis shows that the higher number of reported cases in the city is theft and there are much less cases reported for crimes like domestic violence, human trafficking etc. The crimes with significantly higher counts after theft were battery, criminal damage, and narcotics.

Figure 2 shows how the arrests of crimes have changed throughout the years from 2001 to 2017. From this graph, we can see that the relative difference between arrests and non-arrests seems to have remained stable. Over the years 2006-2010, there was noteworthy increase in total recorded cases, showing the high crime rates of Chicago when criminal activity was at a peak.

The progression of the total number of crimes over time is described in Figure 3. From these graphs, we can analyse that there are more than 500000 crime cases reporting each month. The months of May to August appear to be the busiest for offenders, with the months February and December having the lowest crime count out of all months. On a weekly basis, there are more than 10 million cases reported for each day over the years with little variance except where Saturday is the day having the highest crime rate. When observing crime activity over the course of a day, there are less crimes committed at early hours of 2 to 7 AM, and a considerable increase is observed from 12 PM to 12 AM. The rate at which crimes occur in each month is almost constant throughout the days.

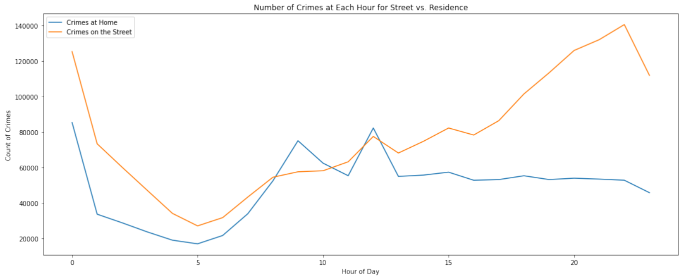

Fig. 4. Hourly crime count on street vs. residence

On an hourly basis, Figure 4 plots the total number of crimes that occurred on residential areas and streets. In the graph, we can see that the frequency of crimes occurring at these kinds of domestic locations are quite high and the highest number of crimes is observed in the streets with a frequency ranging between 14000 and 40000 cases reported within a day. Crimes at both locations are at a minimum at 5 AM, but towards night, crimes increases significantly on streets while remaining constant in residential locations.

Selected communities within Chicago with the highest crime rate and the count of the highest number crimes occurring in a month is given in Figure 5. From this graph, we can see that Austin reports the highest criminal cases, with Narcotics being the top crime type, followed by battery. The top crimes types occurring in these regions include battery, criminal damage, narcotics, other offences and theft

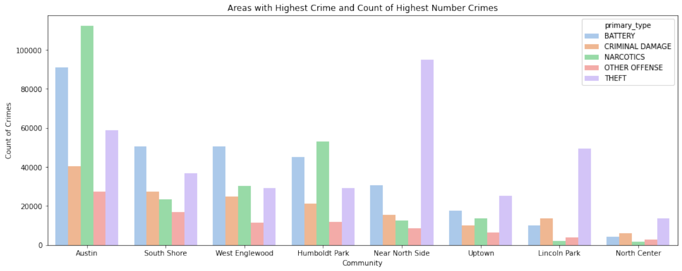

Fig. 5. Count by crime type in top crime affected areas of Chicago

with counts ranging between less than 5000 and upto almost 120000 cases. In the 8 shown communities, narcotics is the most common crime in 2, battery in 2, and theft has the highest count in the remaining 4 communities.

Four famous machine learning models i.e., Multinomial Logistic Regression, Naive Bayes, Decision Tree Regression and Random Forest are used to estimate what kind of crime will most likely occur in the city of Chicago based on the characteristics chosen. These characteristics include 'location_description', 'arrest', 'domestic', 'community_area', 'district', 'ward', 'fbi_code', 'hour', 'week_day', 'year_month', and 'month_day'.

The requisite string type characteristic is converted to a vector index, and each vector value is then normalised to have a unit norm. The dataset of crimes in Chicago is divided in such a way that 80% of the dataset is used to train the models taken and the remaining 20% to test the model.

TABLE I
ACCURACY OF MACHINE LEARNING MODELS

Machine Learning Algorithm	Accuracy
Multinomial Logistic Regression	48.625
Naive Bayes	35.914
Decision Tree Regression	44.165
Random Forest	54.109

Table 1 shows the prediction accuracy obtained after the estimation of the crime pattern by the models. From the below table, we can observe that the Random Forest model performed consistently better than other visualised models with the prediction accuracy of 54.109%. When taking a deeper look into the predictions of each crime type, it was observed that the model was exceptionally good at predicting theft, which may have been a result of the large amount of data on theft cases. Crimes with very low records, such as domestic violence, ritualism, etc., had a poor prediction accuracy as there was not enough data to train the models for such crimes. The importance of spread out data is visible here because it is necessary that enough data is available for all crime types in order to better predict more types of crimes. Irrespective of other predictive models that are concerned with locating the highest number of crimes regardless of community areas and predicting the numeric frequency of crimes in the coming years, in this project, we have attempted to predict the periodic crime trend in Chicago city.

VI. CONCLUSION

The analysis performed in this paper provided insights into the pattern changes of crimes in Chicago over the years 2001 - 2017, whether it be time-wise or location-wise. The prediction research conducted in this paper offers predictions crime in a specific field by analysing and employing various machine learning algorithms on a selected array of features . We conducted this experiment using crime data collected from the Chicago Police Department's CLEAR framework, which is publicly accessible, and we were effective in performing predictive crime pattern analysis. In this paper, Random Forest model has proven to be the best model for predicting the results with an accuracy of 54.109% which can be observed by comparing this accuracy with the other prediction algorithms. The comparatively poor algorithm turned out to be Naive Bayes with an accuracy of 35.914%. More data will be obtained in the future, and computer capabilities can be improved, allowing for the creation of more efficient and accurate models that can better aid police in catching crime.

REFERENCES

[1] P. Sontakke and C.-S. Kim, "Crime Pattern Analysis based on Machine Learning and Big Data using Apache Spark," International Journal of Information Communication Technology and Digital Convergence, vol. 3, no. 1, pp. 10–16, Jun. 2018.

[2] T. Chauhan and R. Aluvalu, "Using Big Data Analytics for developing Crime Predictive Model," First International Conference on Research Entrepreneurship (ICRE 2016), vol. 1, Jan. 2016.

[3] K. N. Win, J. Chen, G. Xiao, Y. Chen and P. Fournier Viger, "A Parallel Crime Activity Clustering Algorithm Based on Apache Spark Cloud Computing Platform," 2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 2019, pp. 68-74, doi: 10.1109/HPCC/SmartCity/DSS.2019.00025.

[4] "Spark Overview," Overview - Spark 3.1.1 Documentation. [Online]. Available: https://spark.apache.org/docs/latest/. [Accessed: 13-May-2021].

[5] "Python Programming Guide," Python Programming Guide - Spark 0.9.0 Documentation. [Online]. Available: https://spark.apache.org/docs/0.9.0/python-programming-guide.html. [Accessed: 15-May-2021].

[6] J. Brownlee, "Multinomial Logistic Regression With Python," Machine Learning Mastery, 31-Aug-2020. [Online]. Available:https://machinelearningmastery.com/multinomial-logistic-regression-with-python/. [Accessed: 10-May-2021].

[7] D. Bohning, "Multinomial logistic regression algorithm," Annals of the Institute of Statistical Mathematics, vol. 44, no. 1, pp. 197–200, Mar. 1992.

[8] "Learn Naive Bayes Algorithm: Naive Bayes Classifier Examples," Analytics Vidhya, 18-Oct-2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/. [Accessed: 15-May-2021].

[9] R. Gandhi, "Naive Bayes Classifier," Medium, 17-May-2018. [Online]. Available: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c. [Accessed: 15-May-2021].

[10] "Decision Tree Algorithm Tutorial With Example In R," Edureka, 25-Nov-2020. [Online]. Available: https://www.edureka.co/blog/decision-tree-algorithm/. [Accessed: 15-May-2021].

[11] "Decision Tree Algorithm Tutorial With Example In R," Edureka, 25-Nov-2020. [Online]. Available: https://www.edureka.co/blog/decision-tree-algorithm/. [Accessed: 15-May-2021].

[12] "Random Forest - an overview | ScienceDirect Topics", Sciencedirect.com, 2021. [Online]. Available: https://www.sciencedirect.com/topics/engineering/random-forest. [Accessed: 15- May- 2021].

[13] "A complete guide to the random forest algorithm", Built In, 2021. [Online]. Available: https://builtin.com/data-science/random-forest-algorithm. [Accessed: 15- May- 2021].

[14] A. Mukherjee, S. De, S. Bhattacharyya and J. Platos, "Chicago Crime Data Analysis Using PIG in Hadoop," 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), 2018, pp. 242-247, doi: 10.1109/ICRCICN.2018.8718725.

[15] "Multi-Class Text Classification with PySpark", Medium, 2021. [Online]. Available: https://towardsdatascience.com/multi-class-text-classification-with-pyspark-7d78d022ed35. [Accessed: 15- May-2021].

[16] S. Kim, P. Joshi, P. S. Kalsi and P. Taheri, "Crime Analysis Through Machine Learning," 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2018, pp. 415-420, doi: 10.1109/IEMCON.2018.8614828.

[17] Currie32, “Crimes in Chicago,” Kaggle, 28-Jan-2017. [Online]. Available: https://www.kaggle.com/currie32/crimes-in-chicago. [Accessed: 15-May-2021].