National Institute of Technology Delhi

Data Mining(CSB-352)

Crime Prediction and Analysis

Submitted to:

Dr. Rishav Singh

Dept. of Computer Science Engineering

National Institute of Technology, Delhi

Submitted By:

Binshumesh Sachchan (171210021)

Divyanshi Singh (171210022)

G.D.V. Sujith Sagar (171210023)

T.J. Vamsi Kesav (171210055)

Index

Abstract

Crime is a menace that every country is facing because of which an attempt is being made to incorporate Artificial Intelligence in reducing crime. Our aim is to build a model which would work on crime data. The data which has been taken reflects reported incidents of crime that occurred in the City of Chicago from 2012 to 2017. From this data, the aim of the model is to explore crime data in Chicago and showcase the implementation of a predictive model to find which particular region or part of the city is more affected with crime i.e., finding the regions with higher threat levels and danger levels. This can be done by predicting the type of crime that might most likely occur given the location and time. Clusters of regions having similar threat level can be formed and a prediction can be made on given parameters (location, time) about probability of happening of particular crime type for given coordinates or about the safety of that place with the aid of machine learning and data mining tools.

Brief overview of crime in Chicago:

Chicago, the nation's third-biggest city, accounted for 22% of the nationwide increase with 749 murders in 2016, more than the number of murders in the largest city, New York (334), and the second-largest, Los Angeles (294) for the same year, combined. The estimated number of homicides in Chicago increased by 52% in 2016. The number of homicides rose by 8.6% in the United States, making Chicago an outlier, and an interesting case to analyze. The vast majority of these killings happened in five mostly black and Latino neighborhoods on the south and west side where only 9% of the 2.7m city lives.

The graph in Fig1 and Fig2 below shows the graph of number of cases still touches an unsatisfying mark.

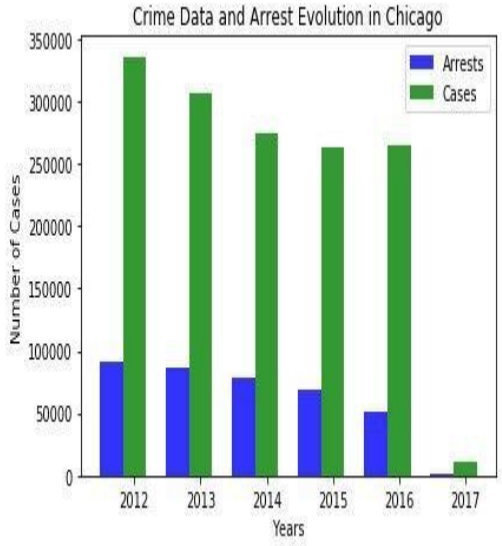


Fig 1

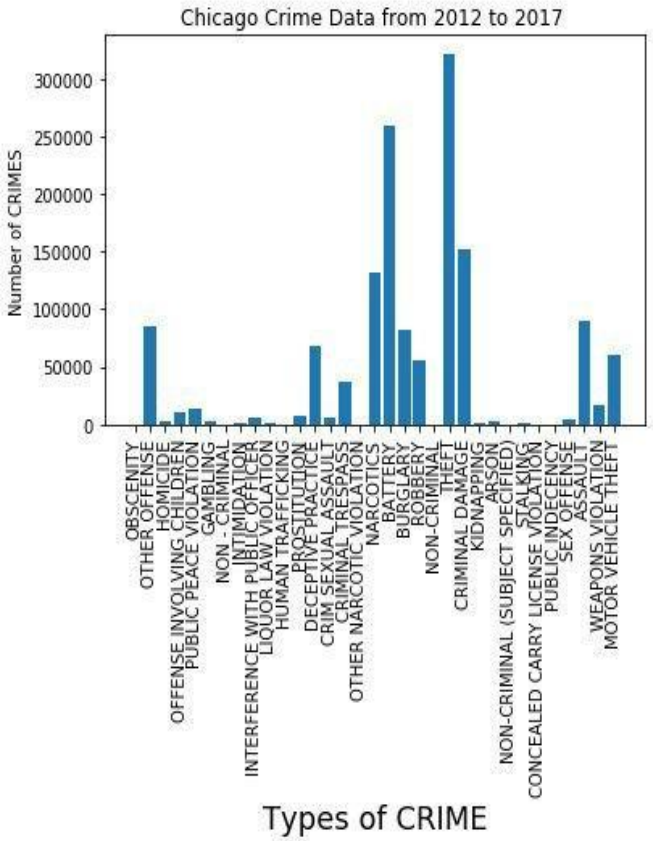


Fig2

Introduction

Rationale:

Many important questions in public safety and protection related to crime, and a better understanding of crime is beneficial in multiple ways: it can lead to targeted and sensitive practices by law enforcement authorities to mitigate crime, and more concerted efforts by citizens and authorities to create healthy neighborhood environments.

With the advent of the Big Data era and the availability of fast, efficient algorithms for data analysis, understanding patterns in crime from data is an active and growing field of research.

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.

Criminal and sociology scholars are analyzing the pattern of criminal activity and its relationship with the area. Researchers have shown that many crook activities are taking place in a region. This is called a hotspot. Machine learning can be used to become aware of hotspots by way of data pushed approach

Thus, an attempt is made in the direction where analysis of criminal data and detecting pattern can help towards reducing/preventing crimes and create a safer environment for general public.

Goal/Objective:

Goal much of the current work is focused to explore crime data in Chicago and showcase the **implementation of a predictive model for type of crime that might occur at a given location and time.**

This could help the public institutions in 3 main ways:

-Better create public policy for correctional agencies

-Help focus the countermeasures on negatively impacted crime categories according to the prediction

-Help regulation enforcement organization to behavior their operations

Our approach will help regulation enforcement agencies to have assumption ahead about the possibility of crime, which could arise at a place in a given time. This will assist them
to resolve the instances faster than earlier.

Related Work:

Criminal activities are common around the world. Therefore, researchers have completed many works on this subject matter. Researchers have been analyzing the relation among criminal activities and socio-economic variables like unemployment, earnings level, level of schooling and so forth.

Researchers like Sangani et al worked on similar paper on Crime Prediction and Analysis where they used Simple K-Means clustering techniques and algorithm for predicting Crimes. This model tried to figure out crime trends based on crime zone but not based on a certain period. Therefore, they can extend their model by using more datasets with more features such as season, date, time so that this model will be more beneficial of police using as well as safety of normal people.

Crime Prediction using Ensemble Approach by Almaw et al investigates on hidden crime data mining. They used ensembled classification learning methods. They used Naïve Bayes classification and artificial network to test for crime analysis. They identify the crime trends and patterns and predicts the type of crimes might occur next in a specific locality of longitude and latitude in a specific schedule of time and season. Therefore, they can further implement the required combined techniques for improving a better crime prediction of a single classifier model by integrating multiple models by using more datasets with more features.

Some researchers like Tabedzki et al. used the random forest model. By the usage of random forest and k-nearest neighbor, researchers obtained the first-rate accuracy across 39 different categories. Since the information was very noisy, consequently they thought random forest might provide great effects.

Kang et al. analyzed crime occurrences by the usage of multimodal records in which they have applied deep learning. They found out that DNN version provides greater precision values in predicting crime prevalence than other prediction fashions when they are compared with other works. Their present crime predicting methodology for finding occurrences is not able to produce statistics based on the unique form of at a selected time.

Prediction of Hourly Effect of Land Used on Crime by Matijosaitiene et al predicted future crimes based on the time using Manhattan. They used the random forest and logistic regression for their predicting model of crimes based on exact time features when most of the crimes happened. They achieved high accuracy on random forest algorithms. They also tried to analyze hot spot feature to predict for controlling crimes in different areas. In their future work, they can also add specific longitude and latitude of a specific area where crimes mostly committed.

Methodology:

Machine Learning:

The term machine learning refers to the automated detection of meaningful patterns in data. In the past couple of decades it has become a common tool in almost any task that requires information extraction from large data sets. We are surrounded by a machine learning based technology: search engines learn how to bring us the best results, anti-spam software learns to filter our email messages, and credit card transactions are secured by a software that learns how to detect frauds. Digital cameras learn to detect faces and intelligent personal assistance applications on smart-phones learn to recognize voice commands. Cars are equipped with accident prevention systems that are built using machine learning algorithms.

Machine learning is also widely used in scientific applications such as bioinformatics, medicine, and astronomy. One common feature of all of these applications is that, in contrast to more traditional uses of computers, in these cases, due to the complexity of the patterns that need to be detected, a human programmer cannot provide an explicit, fine detailed specification of how such tasks should be executed. Taking example from intelligent beings, many of our skills are acquired through learning from our experience (rather than following explicit instructions given to us). Machine learning tools are concerned with endowing programs with the ability to learn and adapt.

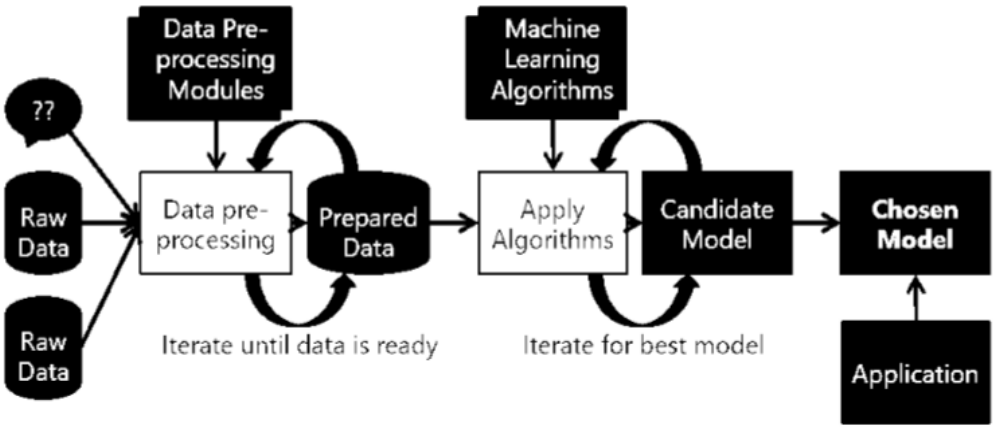


Fig 1.1-Machine learning process

Software Tools required are:

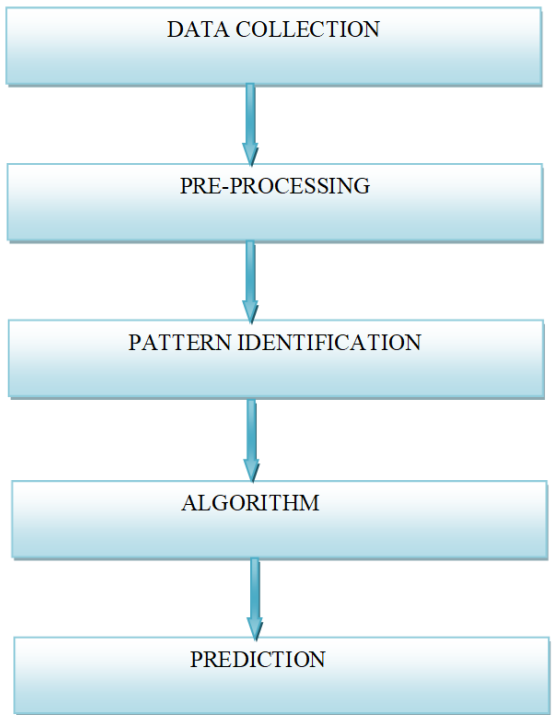
- Google colab (can also work on any other python supporting notebook)
- Python3.7
- Pandas
- Sklearn

Our Approach:

This model uses a dataset from Chicago, which involves extraordinary classes of crimes occurring, based totally on different factors along with places and crimes over years. We have used different classifiers to locate hotspots of crook activities primarily based on the time of day. The goal of this challenge is to apply machine learning procedures to find a criminal sample with the aid of its category with the given precise time and location.

The algorithm used to train the dataset are KNN, Random Forest, Logistic Regression, XG Boost, Decision Tree with ensembling method of Bagging are used. The following steps are followed for all the implemented algorithms:

Flowchart depicting the approach followed:



Dataset:

Description-

- The data is taken from Kaggle (a data science platform). Originally, data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system.
- For analysis, data is taken for the year 2012-2017.
- The dataset contains 1418365 observations and 23 columns (before pre-processing).

Features (before pre-processing)-

- **ID -** Unique identifier for the record.
- **Case Number -** The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.
- **Date -** Date when the incident occurred.
- **Block -** The partially redacted address where the incident occurred, placing it on the same block as the actual address.
- **IUCR -** The Illinois Uniform Crime Reporting code. This is directly linked to the Primary Type and Description.
- **Primary Type -** The primary description of the IUCR code.
- **Description -** The secondary description of the IUCR code, a subcategory of the primary description.
- **Location Description -** Description of the location where the incident occurred.
- **Arrest -** Indicates whether an arrest was made.
- **Domestic -** Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.
- **Beat -** Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district.
- **District -** Indicates the police district where the incident occurred.
- **Ward -** The ward (City Council district) where the incident occurred.
- **Community Area -** Indicates the community area where the incident occurred. Chicago has 77 community areas.
- **FBI Code -** Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).

- **X Coordinate -** The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.
- **Y Coordinate -** The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.
- **Updated On -** Date and time the record was last updated.
- **Location -** The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal.
- **Year -** Year the incident occurred.
- **Latitude -** The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
- **Longitude -** The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.

Implementation:

Pre-Processing-

- Rows containing any null values are removed.
- Some columns are dropped as they were giving the redundant information, and others are retained to see if they can provide any valuable information during prediction.
- From Date column hour, minute, month, Year have been extracted and the date column dropped.
- We can see "other offense", which accounts for around 6% of the total crime records, involve lots of different categories of crime that cannot be categorized easily. Therefore, we drop the rows with other offense.
- Encoding the categorical values present in column Primary Type, Location Description, FBI code, IUCR, Block as the values present in both columns were nominal so that's why Binary encoder is used. We used binary encoder instead of one hot label encoder as in binary encoder number of new columns generated are less.

Target/Feature Selection-

The model is focused in predicting following targets with the data:

Prediction the type of crime that would most likely occur in the given location and time:

Features for its prediction are:

- Block
- IUCR
- Location Description
- Domestic
- Beat
- District
- Ward
- Year
- Month
- Day
- Hour

Target for its prediction are:

- Primary Type

Refer the following table to get type of crime corresponding to the integer value obtained:

Crime Type	Encoded Integer Value
Battery	1
Public Peace Violation	12
Theft	15
Weapons Violation	16
Robbery	13
Motor Vehicle Theft	8
Assault	0
Deceptive Practice	6
Criminal Damage	4
Criminal Trespass	5
Burglary	2
Crim Sexual Assault	3
Narcotics	9
Sex Offense	14
Offense Involving Children	10
Interference with public officer	7
Prostitution	11

Data Splitting:

The process of splitting the data set was done as:

- Extract the X(features) and the y(target) from data frame
- Split the data using train_test_split from Scikit.
- Test size is kept as 20%.

Model Selection:

The following models have been used and their accuracies can be compared to get the best working algorithm.

- KNN Model
  A powerful classification algorithm used in pattern recognition K nearest neighbors stores all available cases and classifies new cases based on a similarity measure (e.g. distance function). One of the top data mining algorithms used today. A non-parametric lazy learning algorithm (An Instance based Learning method).
  KNN: Classification Approach
  -An object (a new instance) is classified by a majority votes for its neighbor classes.
  - The object is assigned to the most common class amongst its K nearest neighbors. (measured by distance function)

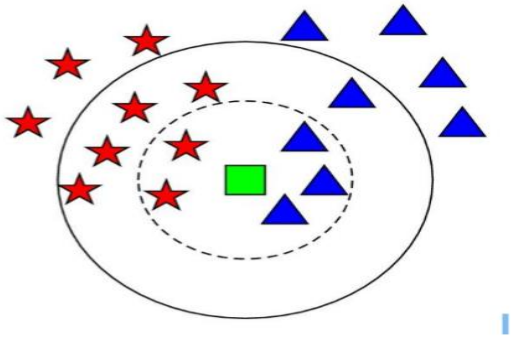


Fig 4.1.1 Principle diagram of KNN

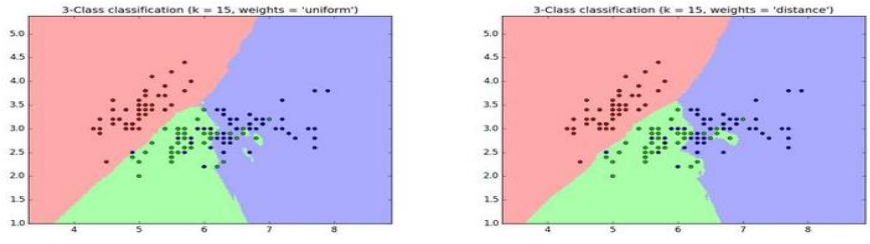
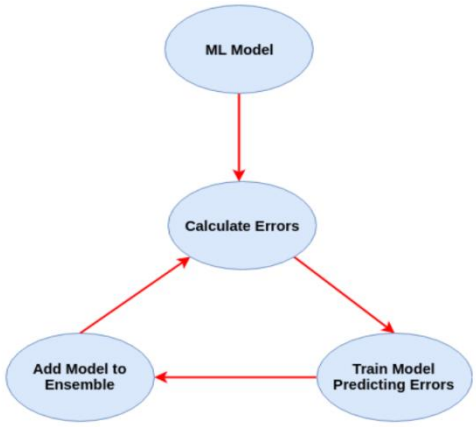


Fig 4.1.2 Shows graphical representation of KNN

Accuracy for prediction in our model = 85.47%

- XG Boost

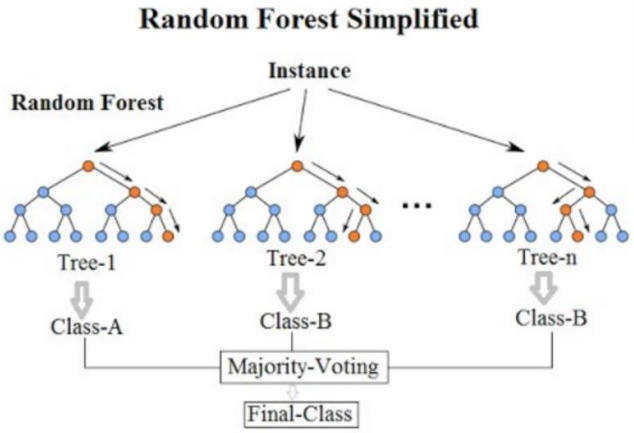
  Boosting trains models in succession, with each new model being trained to correct the errors made by the previous ones. Models are added sequentially until no further improvements can be made. The advantage of this iterative approach is that the new models being added are focused on correcting the mistakes which were caused by other models. **Gradient Boosting** specifically is an approach where new models are trained to predict the residuals (i.e errors) of prior models.



Accuracy for prediction in our model = 91.18%

- Random Forest Classifier

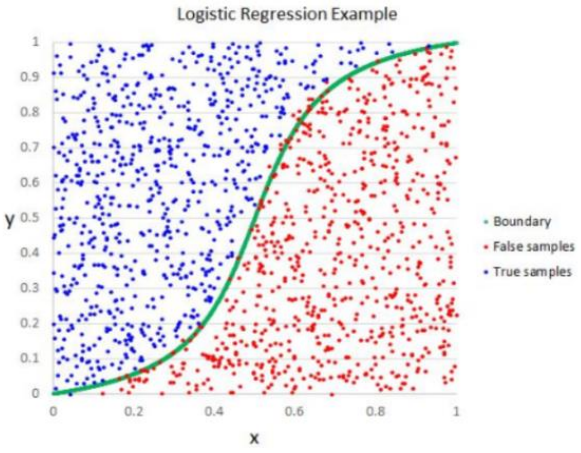
  Random Forests is a very popular ensemble learning method which builds a number of classifiers on the training data and combines all their outputs to make the best predictions on the test data.
  Thus, the Random Forests algorithm is a variance minimizing algorithm that uses randomness when making split decision to help avoid overfitting on the training data.
  Each tree then works like regular decision trees: it partitions the data based on the value of a particular feature (which is selected randomly from the subset), until the data is fully partitioned, or the maximum allowed depth is reached. The final output is obtained by aggregating the results .



Accuracy for prediction in our model = 99.99%

- Logistic Regression

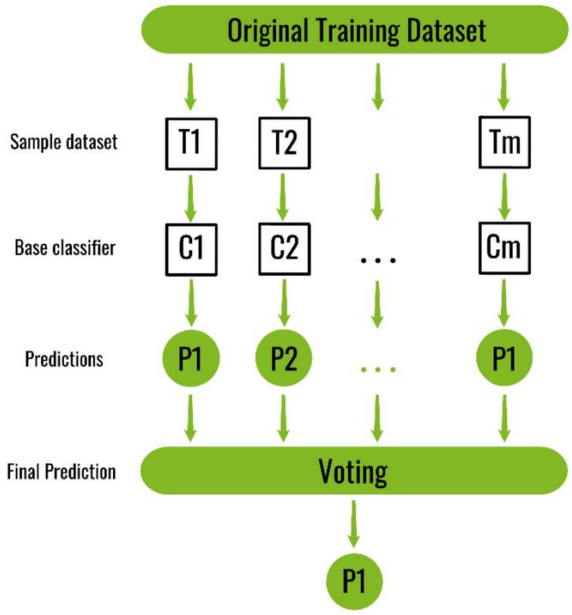
Logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.
There can be binary,multiple or ordinal Logistic RegressionTo predict which class a data belongs, a threshold can be set. Based upon this threshold, the obtained estimated probability is classified into classes.



Accuracy for prediction in our model = 0.32% (until limit reached)

- Bagging

A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. Each base classifier is trained in parallel with a training set which is generated by randomly drawing, with replacement, N examples (or data) from the original training dataset – *where N is the size of the original training set*. Training set for each of the base classifiers is independent of each other. Bagging reduces overfitting (variance) by averaging or voting, however, this leads to an increase in bias, which is compensated by the reduction in variance though.



Accuracy for prediction in our model = 92.51%

Result and Discussion:

This model uses dataset which contains both the mixture of categorical and numeric values. The features with non-numeric value are initially converted to have numeric value using fit transform. Several algorithms are chosen to apply on model such as Random forest, KNN, XG Boost, Logistic Regression, Decision Tree with ensembling method such as Bagging. The main motive of this model is to use algorithms on these datasets to classify the type of crime occurring based on time and location. The chosen algorithms are applied where it provides a simple and fast way of learning a function. This is where the algorithm maps data x to outputs y, where x is a mixture of numeric variables and y is the numeric value for classification. The applied algorithm gives better performance for any classification problem.

According to the algorithms applied following results are obtained.

Algorithm Used	Accuracy
KNN	85.47%
Random Forest	99.99%
XG Boost	91.18%
Bagging	92.51%
Logistic Regression	32.09% (until reaching limit)

Highest accuracy is acquired with the implement of Random Forest Classifier showing that tree-s based classifiers are giving much better predictive results. The main reason of difference in actual and predicted values are noise, variance, and bias.
Logistic Regression however shows error because of technical limit.

Conclusion:

The idea behind this project is that crimes are relatively predictable; it just requires being able to sort through a massive volume of data to find patterns that are useful to law enforcement. This kind of data analysis was technologically impossible a few decades ago, but the hope is that recent developments in machine learning are up to the task. We used different algorithms to predict the type of crime that might occur based on location and time.

Thus, the dataset used, provides the maximum correct result with higher accuracy when implemented with different tree classifiers. The results in this paper provides similar results when implemented with tree-based algorithms. Therefore, this model expects to get more variation in the results when implemented with other classifying algorithms such as Logistic Regression.

Future Work:

The use of AI and machine learning to detect crime via sound or cameras currently exists, is proven to work, and expected to continue to expand.

The use of AI/ML in predicting crimes or an individual's likelihood for committing a crime is promise but still more is unknown. The biggest challenge will probably be "proving" to world or politicians in specific that it works. When a system is designed to stop something from happening, it is difficult to prove the negative. Companies that are directly involved in providing governments with AI tools to monitor areas or predict crime will likely benefit from a positive feedback loop. Improvements in crime prevention technology will likely spur increased total spending on this technology.

Possible avenues through which to extend this work includes:

- Time-series modeling of the data to understand temporal correlations in it, which can then be used to predict surges in different categories of crime
- Additional parameters can be used for more accurate prediction
- Generating parameters from existing data by performing various algebraic operations.
- Adding non-linearity to data.

Future Scope in this area is generation of alert or notification system depending on user location to inform them about the area and safety status which in turn would not only help the government but society as a whole as the goal of any society shouldn't be to just catch criminals but to prevent crimes from happening in the first place.

References:

- Alkesh Bharati and Dr Sarvanaguru RA.K. 2018. Crime Prediction and Analysis Using Machine Learning.
- https://innovate.mygov.in/innovation/paasbaan-crime-prediction-and-classification-in-indore-city/
- https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-018-0171-7
- K. Das, A. Ashrafi and M. Ahmmad, “Joint Cognition of Both Human and Machine for Predicting Criminal Punishment in Judicial System," 2019 4th International Conference on Computer and Communication Systems
  (ICCCS), Singapore, 2019.
- Al Boni, M. and Gerber, M.S., 2016, December. Area-specific crime prediction models. In 2016 15th IEEE International Conference on Machine Learning and Applications