Crime Analysis in the city of Chicago for 2017

Jinesh Shah - js9656 and Saurabh Rewaskar - ssr1137

ABSTRACT

Crime incidents have increased over a period of time and it gets really difficult for the police department to figure out crime before hand using manual techniques. So, it becomes increasingly important to determine something which can help in knowing about the crimes before hand. With the help of data mining we can accomplish this task. As we have lots of data recorded on the daily basis for different crime incidents, careful evaluation of that data can lead us to a pattern in data with the help of which we can predict where or at what time the crime might take place.

In this project we are doing the similar kind of analysis on crime data for the city of Chicago in the year 2017. Given a city area we will try and predict whether it will be safe or not? We will also try and find patterns in the data like monthly density of crimes and determine the month with the highest crime rate, or hours which observed the highest criminal activity during the day, etc.

1. OVERVIEW

1.1 Who has done this before

Similar kind of research has also been carried out by the students and professors of Computer Science department in one of the engineering colleges in India. They used the crime data for analysis and predicting the future crime locations along with the time perfections. Their main purpose is to find the crime data characteristics. They tend to use the crime data collected over the years to find some pattern in it and follow that pattern for predicting future location and time where crime can take place. [6]

A group of people belonging to cyber-security department conducted crime data analysis for stating which region is more crime prone as compared to others and visualize the country map and show regions which are crime prone, below map shows regions/city in India which are crime prone:

Figure 1: Crime Prone Areas in India

These people used an approach between computer science and criminal justice which proposed a better approach for crime prediction. Their major focus for predicting this was on crime factors of each day rather than criminal background and other stuffs. [3]

Some people did crime analysis using k means clustering in the past. Their focus was on clustering approach which in their context is the best data mining technique one can use as it groups the similar data together with the help of which one can find the hidden patterns in the data collected. They implemented k-Means algorithm on the collected data using open source statistical data mining package named

rapid miner tool written in java. [1]

A student and a professor from Libya wanted to analyze the recent criminal data to find a pattern in the increasing criminal activity in recent times. They collected data manually from a police department in Libya. They collected both kind of data i.e. crime data and criminal data from which they created training data and testing data using which the model was created and tested. WEKA and Microsoft excels analytic tool were used to accomplish this purpose. [10]

One pair tried and implemented an advanced version of crime analysis by generating a developing an entire framework which can be used for detailed crime analysis. They did not focus on the regions of a city where major crime would take place, instead they segregated the entire process into 3 parts which they named as named-entity extraction, deceptive-identity detection and criminal-network analysis. They initially prioritized eight different types of crimes and then evaluated different data mining technique that will be best suited for mining different types of crimes. Their focus was mainly on new techniques which can be used on structured as well as unstructured data. They did use the clustering and classification techniques in some of the steps of their process. [2]

An individual from the Defence Technology institute of Thailand conducted similar research to analyze patterns and predict trends in crime data to increase the efficiency of solving crimes faster. During the project, he devised clustering techniques to detect co-relation between diverse types of crimes, identifying groups of offenders etc. by analyzing information about the convict and the victim and generating criminal profiles [8]

Individuals from different engineering colleges of Coimbatore and Chennai, India came together to use data mining techniques to analyze the cityâĂŹs crime data from Police Department of Tamil Nadu, India. They computed per capita crime rate by clustering techniques and predicted future crime rates [4]

This group has mined Chicago crime data to identify a set of predictive tasks, train and evaluate models to perform predictions. They have trained a model to predict whether an arrest will be made or not for a given incident and predict how many criminal incidents will be reported in each beat where a beat is the area assigned under every copâĂŹs jurisdiction. [9]

This project was conducted by a group from Indian Institute of Technology, Delhi. This group highlighted the existing systems used by Indian police as e-governance initiatives and initiated and deployed an interactive query based interface as a crime analysis tool to assist the police in their activities. The system designed by the group is used to extract relevant information from the vast crime database maintained by National Crime Record Bureau (NCRB) and find crime hot spots using crime data mining techniques such as clustering etc. [5]

1.2 Challenges and Ethical Concerns

The research group from India collected the data from the information gathered through the city police department and the records preserved in the police department. Their focus was to find the crime variables and its parameters for determining the characteristics in data. Since the data collected by them was highly complex because of the relationships between the data, it became difficult for this group to easily find the key variables i.e. features that can be used by them to determine the pattern in crime. So, they decided to keep only selected attributes and rest of the attributes were discarded which eased the process to little extent in finding the key attributes. [6]

To accomplish the purpose of crime analysis, it was difficult for some group to collect data from a direct source as they said the data available is very limited. They found increase in crime information as a difficult task to deal with because it gets difficult to find the relative and useful information from the bunch of information available. Incomplete and inconsistent data is also a major problem and they faced the same while working on this project, Difficulty in getting crime related data from legal enforcement departments and achieving accuracy in of training data. [3]

Main objective for some of the people was not to analyze the entire crime data but to focus on single type of crime, so they had to clean all the data and keep only those which were related to their work. Also, they had to look out for such data set where in there are enough number of records related to crime they are focusing on. [1]

The data was manually collected from local police department which had inconsistent and missing data which had to be dealt with. Since the data was collected from more than one local police department the structure of this data had to be rearranged in required form. [10]

Attributes which were selected as main attributes included some gender, age and marital status information. Crime was analyzed using these attributes of a person, which can be an ethical issue if prior approval from the selected authority has not been taken.[10]

Since the focus was both on structured and unstructured data they relied on finding major entities from all the data that is available instead of focusing on attributes and missing values, etc. [2]

An individual who was working alone, faced many data mining challenges while dealing with crime data to predict trends and patterns in the data for faster crime solving. Challenges included modeling of crimes for finding suitable algorithms to detect the crime, precise detection, data preparation and transformation, and processing time. Findings of association rule mining had their limitations. To discover co-occurrences between two data sets, support and confidence results had to be greater than user-specified thresholds and association rule findings took a prolonged period to find out hidden knowledge from the data due to its large volumes. The management and analysis with huge data are very difficult and complex. Thus, they used apriori to prune to data first to reduce the processing time. The issues of crime pattern are concerning with finding and predicting the hidden crime. Nowadays, the crime rate is increase continuously, and the crime patterns are always changing. Consequently, the behaviors in crime are difficult to be explained and predicted. The issues on performance are concerning with precision, reliability and processing time. The uncertainty in crime patterns effects the precision of crime detection. Besides that, the algorithms used properly, and the transformed data also effects the processing time [8]. They also mentioned about the research gaps and challenges they faced in a table

DBScan, EM and Association Apriori involving enormous amounts of data was taking a while to process and was consuming a large amount of processing memory thus slowing

down systems. This group had to overcome this by filtering all the tables which they had initially populated to just the useful attributes which contributed the most towards the results. [4]

The exploratory analysis performed by the group on Chicago crime data revealed many patterns in the data including repeated patterns in crime count across the months of the year and the months with highest crime rate frequency, but these analyses did not add any value towards their target predictions of whether an arrest will be made given a crime or number of crimes in each beat per day. So, they had to do trial and error to find out which attributes contributed the most towards the target findings. Also, when the model was trained on a large amount of data, the baselines substantially out performed with approx. 300

To design a common system, data was required from the police records but harnessing of information became next to impossible as many police records mere missing. The exchange of information among police agencies became very time consuming and therefore, not available in time of need. There were neither a staffs nor time for entering data in records manually. Thus, it became increasingly difficult to coordinate information and come to any meaningful crime analysis. CCIS is only collecting the information and creating a huge crime database but there is no analytical tool for analyzing huge building database. Absence of crime analysis tool made it somewhat âĂŸstandaloneâĂŹ system. Therefore, there is need of support systems as crime analysis tool based on current technologies to meet and fulfill the new emerging responsibilities and tasks of the Police [5]

1.3 Business Case

One of the main objective for a group of people performing crime analysis was to analyze specifically one type of crime namely homicide which is crime committed by human by killing another human. [1]

Crimes do occur in different sections of the society in various ways. So, these people wanted to analyze crime data in different manner like tactically, strategically, administratively, investigative, intelligence and operative. This model aims to help Libyan Security Committee to identify the criminal behavior and specifying offense type related to criminal groups in Libya [10]

Their main reason for undergoing the entire process was to create a general framework which can detect named-entity, deceptive-identity and perform criminal-network analysis. These 3 approaches of criminal analysis can be used for various reasons as per requirement. [2]

This experiment was carried out for the Tamil Nadu Police Department following the rise in criminal activity in India. The results of this data mining could potentially be used to lessen and even prevent crime for the forth coming years. Visual and intuitive criminal and intelligence investigation techniques can be developed for crime pattern recognition. [4]

This group from IIT-Delhi proposed, initiated and deployed an interactive tool for the Government of India to analyze trends in crime data. Government of India approved the design, development and implementation of a Government to Government (G2G), model called the Crime Criminal Information System (CCIS). The CCIS was designed to create computerized storage, analysis and retrieval of criminal records. The Crime Criminal Information System today

Year	Homicide	Attempted murder	Child destruction	Causing death by careless driving
1990	10	19	0	7
1990	6	10	0	5
1990	6	8	0	9
1990	6	2	0	15
1990	10	5	0	1

Figure 2: Tucson police department survey

is in operation in all the States. In CCIS, the information is collected at district level not at basic unit of a single station. Common Integrated Police Application (CIPA) was developed with objective of automation the process or workflow at a police station and to provide inputs for building CCIS. [5]

1.4 Source of Data

The research group from India collected the data from the city police department of Jalandhar which was received in excel format. The data received was highly complex because of the relationships between the data collected and presented in excel. Out of the total data collected by them two third data was used for training and creating a model and one third of the unused data was then used for testing. [6]

One of the group collected data from different web sites like news sites, social media, RSS feed, blogs etc. The data collected is in unstructured form they used mongoDB to store the data another reason for them to use NoSQL database is to reduce the complexity caused by joins. [3]

A team was doing criminal analysis for 2 specific countries, England and Wales. These people collected their data from the offenses recorded by police in England and Wales by offense and police force area from 1990 to 2011-12 which is really a huge data set for analysis. Some of the data used by them were classified as below: [1]

The data was collected manually from different police departments of Libya and was rearranged structurally as per the requirements of different business perspective. The data consisted of 350 criminal records which were used to test the model, this was part of the 70-30 training and testing model generally used.[10]

The data was collected from Tucson police departmentâĂŹs crime dataset which consisted of 1.3 million criminal records ranging from 1970 to present. Out of all these records only those records were kept for processing which belongs to the eight selected crime types in increasing order of public harm. [2]

One of the group obtained data from various sources such a news articles, social media, criminal records, government, different sensors etc [8] while other group got it from Department of Police, Tamil Nadu, India which was spanned over the years of 2000 to 2009 [4]

There was a research performed similar to what we are going to do with the same data but of more years.[9]

The data was obtained from the Government of Indias vast crime database maintained by National Crime Record Bureau (NCRB). [5]

1.5 Issues in data

The data collected by one of the research group had lots and lots of attributes with redundant data in it. Also, there were missing values in it, to deal with which they straight away rejected those records and cleared away redundant attributes or attributes which were of less or no importance so that they can process useful data. [3]

Data collected from the police department of England and Wales had huge amount of missing records as the data belong to last 21 years, so they used rapid miner tool and applied "replace missing value operator" which is readily available in the tool to get rid of missing values. They specifically did not mention any technique which was used behind the scenes to replace the missing value. [1]

Data was in large volumes and in unstructured format in different databases. This lead to challenge of data preparation, transformation, integration and caused difficulty in analyzing and extracting hidden knowledge from such copious volumes of data. [8]

This experienced a lot of unwanted data in their dataset acquired from the Department of Police, Tamil Nadu. They had to perform a lot of pre-processing tasks to come up with a clean dataset which had no missing values or unwanted data which did not contribute towards the final predictions. Since the volume of the data was enormous and unstructured, it was difficult finding patterns. They performed filtering to remove unwanted data [4]

The data was so vast in the Chicago crime dataset, that computation time was exceeding all leaps and bounds. Counting the nature of conditional probabilities was proving out to be a very time consuming and expensive operation. Also in the data, there were merely any date-beat pairs, thus latent factor model where a bias for the month and bias for the beat would be calculated was proving out to be inefficient to calculate the bias. [9]

The data this group worked on was classified data provided and funded by the Government of India for the deployment of a crime analysis tool. Thus, this data was well structured clean data with no underlying issues except for abundance in volume. [5]

1.6 About data

Some of them collected data from mixture of online websites. Since the data was collected from various from it does not guarantees to be in structured form and cannot use one method on the data obtain from these sites. It must be cleaned thoroughly and transform it to get into the form which we require. [3]

Data was cleaned and only the related records were kept, later they selected the attributes which were useful for the analysis purpose and got rid of the remaining attributes. Once the attributes were selected, the data was transformed

Property
- Murder for gain
- Dacoity
- Prep.& Assembly For Dacoity- explanation
- Robbery
- Burglary
- Theft

Violent
- Murder
- Attempt to commit murder
- C.H.Not Amounting to murder
- Hurt/Grievous Hurt
- Riots

Crime against Women and Child
- Rape
- Dowry Death
- Molestation
- Sexual Harassment
- Cruelty by Husband and her relatives.

Others
- Kidnapping & Abduction of others
- Criminal Breach of Trust
- Arson
- Cheating
- Counterfeiting
- Others IPC crimes

Figure 3: Crime Categorization

to normalize values using "Normalize Operator" of the rapid minder tool for performing the k means clustering algorithm. [1]

Since the data was in unstructured format they selected important attributes using Entity detection process. Some of the entities recognized by this process didnâĂŹt had enough data to be classified they were discarded form consideration in this process like Vehicle name which only occurred in 36 reports out of total considered.[10]

Data was in large volumes and in unstructured format in different databases obtained from various sources thus preprocessing the data is difficult to get it into set format to do further processing [8]

[4]The following yearly attributes were presented and used in the data set for the city crime statistics and predictive analysis of per capita crime was conducted using the following key attributes.

[9] The data obtained by the group was based on crimes in the city of Chicago. The fields and description of their data is given below:

Field	Description
sid	meta data
id	Record id
case_number	Case number
date	Date on which crime was reported
block	Block address of the crime
iucr	Code for the type of crime
primary_type	Category of crime
description	Description of the crime
location_description	short description of the location
arrest	If arrest happened for the crime
domestic	If it was domestic violence
beat	Officers Patrol Area
district	District
ward	Ward
community_area	Community area
fbi_code	FBI code
latitude	Latitude
longitude	Longitude
year	Year in which the crime occurred

1.7 Data Cleanliness

The data obtained by one of the group was initially cleaned and unnecessary attributes were removed. They then transformed the data as per their requirements so that they can use the data in WEKA tool which they used for pre processing, clustering, classification, regression of the data. Also, they used WEKA for visualization of data. [3]

Here they followed a slightly different approach where they initially the data was cleaned, and the clusters were formed which was then presented to the experts so that they can drill much deeper into it. A slightly different approach where in the domain experts were consulted after the cleaning and clustering was done. Before all this the data was cleaned, integrated, transformed, reduced and then discretized.[2]

The data was not clean and had to be filtered before using it for predictive analysis. Since the span of the data was enormous, it had a lot of unwanted and irrelevant attributes and features. After careful consideration, the group filtered out such unwanted data and filtered it to perform further analysis on clean data records. [8]

The data had lot of missing values and it was difficult to handle them for such a large volume of records, so the group ignored the fields with missing values. Pre-processing was performed on the data prior to training the model to infer useful features for building the models [9]

Since the data was obtained from the National Crime Record Bureau, it was structured and well-maintained data from the past few yearsâĂŹ crimes therefore not a lot of preprocessing went into this dataset, and the group focused directly on the interactive CCIS tool deployment [5]

1.8 Methods and Algorithms

Some of them read some of the research papers related to the same process and observed that most of those papers used k-means algorithms for clustering and neural networks for classification purpose. So, they decided to move with the same techniques and apply K means on their data. [6]

One of the group weren't specific about what distance metric they used to calculate the distance, but they used precision, recall and F1 measure to calculate their performance for all the methods and compare the results between them. They also evaluated the performance based on the cost benefit curve, where value above threshold indicates good performance. [6]

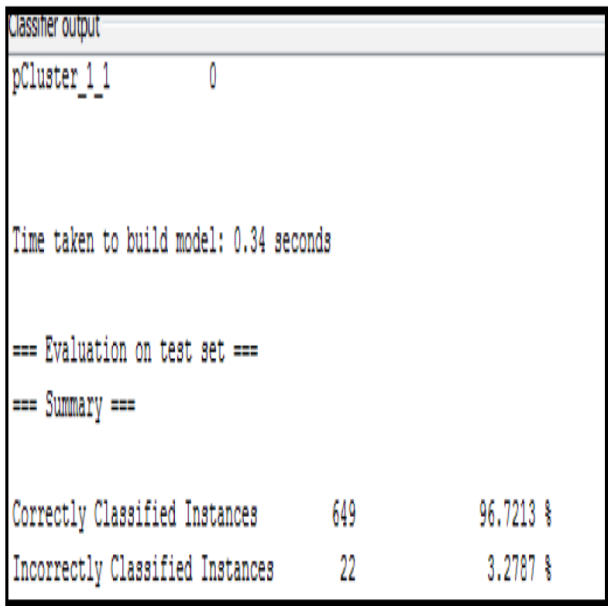


Figure 4: RBF Classifier

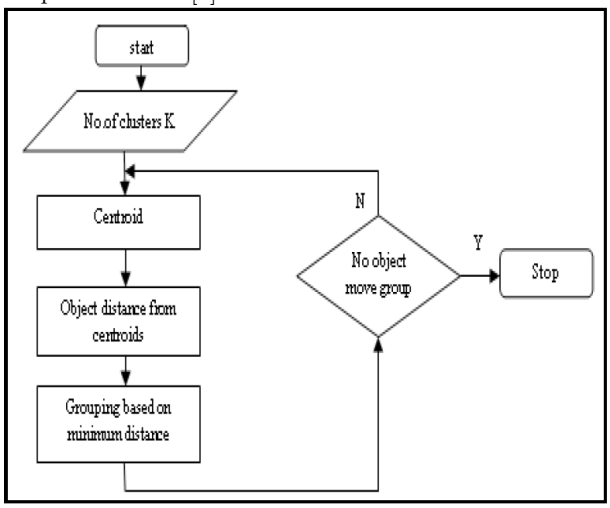
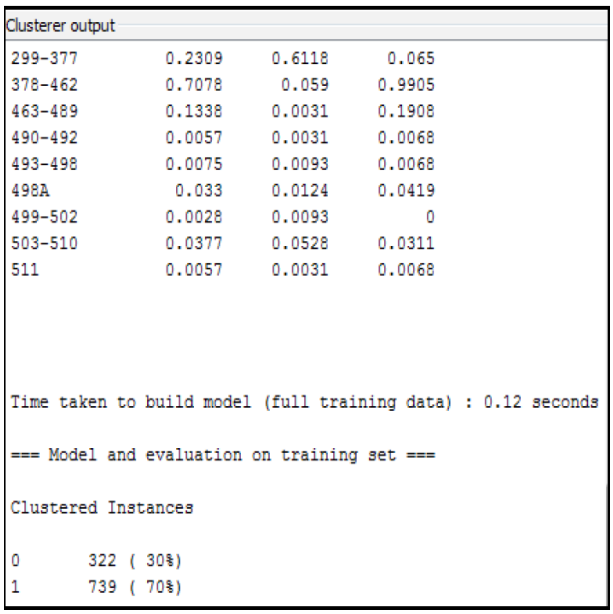


Fig. 1: K-means Flow Chart

They also decided to apply new technique called weighted k means and compare this results with the results obtained from k means and found that weighted k means shows much better accuracy when compared to k means algorithm. Along with this clustering technique they also used RBF classification technique famously known as Radial Basis function which can be used in as a single layered and multilayered architecture. Here they used activation functions for triggering neurons. Using this technique, they attained a whopping 96 percent accuracy which can be seen in the below image:

The results obtained by them using k means algorithm is displayed below:

As guaranteed by them, their proposed technique of weighted k means and RBF classification technique yield better re-



```
Clusterer output
299-377          0.2309      0.6118       0.065
378-462          0.7078       0.059      0.9905
463-489          0.1338      0.0031      0.1908
490-492          0.0057      0.0031      0.0068
493-498          0.0075      0.0093      0.0068
498A              0.033      0.0124      0.0419
499-502          0.0028      0.0093           0
503-510          0.0377      0.0528      0.0311
511              0.0057      0.0031      0.0068



Time taken to build model (full training data) : 0.12 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       322 ( 30%)
1       739 ( 70%)
```

Figure 5: Radial Basis

Sr No.	Technique Used	Precision	Recall	F1 Measure.
1	Simple K-Means and RBF Networks	0.962	0.967	0.967
2	Proposed K-Means and RBF Networks.	0.989	0.987	0.987

Figure 6: Comparision

Area sensitivity	Notable event	VIP presence	Criminal group	crime
Yes	Yes	Yes	No	yes
Yes	Yes	No	Yes	no
No	No	No	Yes	no
Yes	No	No	No	no
Yes	Yes	Yes	yes	yes
No	Yes	No	No	no

sults than normal k means clustering technique. This comparison we can see in the below image:

A group which used mongoDB to store the unstructured data selected NaÃŕve Bayes classification algorithm for their model. Using NaÃŕve Bayes method, they created a model by training crime data which includes vandalism, murder, robbery, burglary, sex abuse, gang rape, arson, armed robbery, highway, snatching etc. Another reason for them to use the NaÃŕve Bayes classifier is it solves the zero-frequency problem for them. They achieved 90 percent of accuracy as they classified each word as token and removed the frequently occurring words. For finding crime patterns which occur frequently they used Apriori algorithm which determines association rules and highlight general trends, this helped in finding crime pattern for a place. Now, for prediction they have used decision tree algorithm where in using the training data they created a tree model using binary split. The attributes used can be seen below: [3]

And the decision tree model generated from their training data is:

The results obtained were visualized using statistically data and were plotted using bar graphs as below:

This project had a future scope of criminal profiling and similar areas.

One group preferred using kMeans algorithm as it was widely used among other papers which they referred, and it is easy to implement this algorithm on the tools which are readily available, in this case they used rapid miner open source tool. For the data to be used in this tool they initially cleaned the data as mentioned earlier and then normalized the values in their data. Once this process was completed it was used to find the clusters in the data. They did emphasize on selecting the number of cluster for the clustering algorithm, but did not mention anything about how they

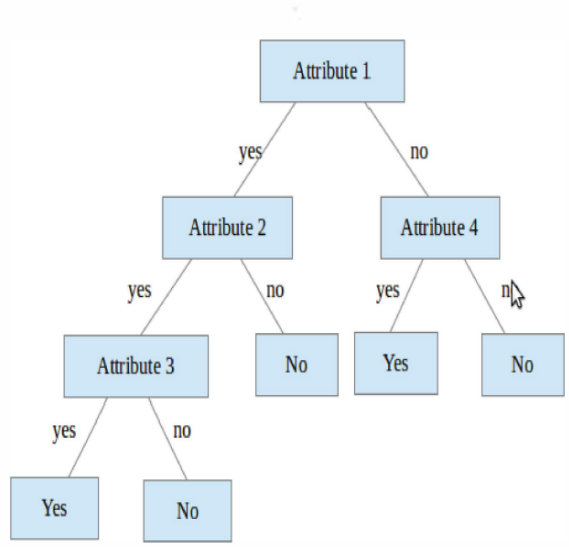


Figure 7: Dec Tree

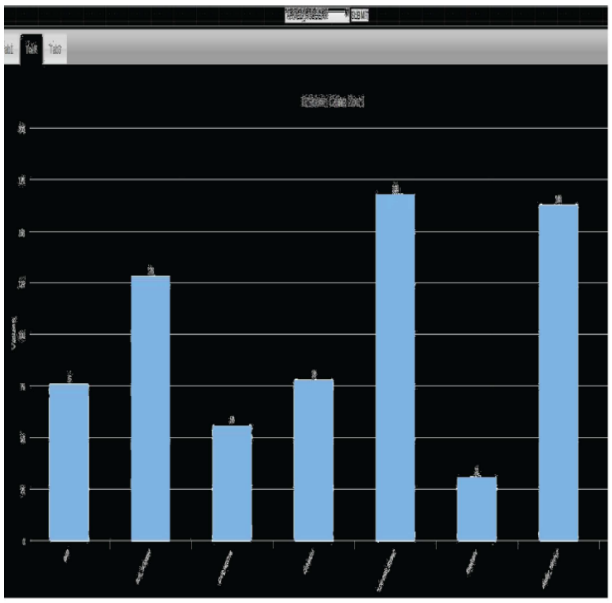


Figure 8: Statistical Result

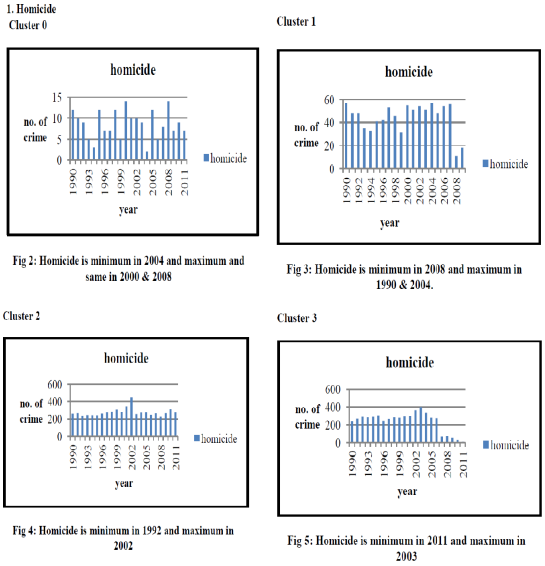


Fig 2: Homicide is minimum in 2004 and maximum and same in 2000 & 2008

Fig 3: Homicide is minimum in 2008 and maximum in 1990 & 2004.

Fig 4: Homicide is minimum in 1992 and maximum in 2002

Fig 5: Homicide is minimum in 2011 and maximum in 2003

Figure 9: Homicide Result

selected the number of clusters used by them. They posted the results of different cluster they got from which we can guess that the number of clusters used by them would be 4. Below are the results and observation of the cluster they found.

They did not compare the results obtained by their method with some other method but only concluded with the results obtained from their method. They used data visualization technique to visualize the results obtained by them. [1]

This process of analyzing criminal records were completed using clustering and association rule technique. They even used the Apriori algorithm for finding some of the best rules in the data, some of the rules obtained from this data is as shown below:

Best rules found:

1. MARRIED=NO 12 ==> GENDER=M 12<conf:(1)> lift:(1) lev:(0) [0] conv:(0)
2.CRIMEADDRESS=TRIPOLI 5 ==> GENDER=M 5 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
3. CRIMEADDRESS=TRIPOLI 5 ==> MARRIED=NO 5 <conf:(1)> lift:(1.08) lev:(0.03) [0] conv:(0.38)
4.CRIMEADDRESS=TRIPOLI MARRIED=NO 5 ==> GENDER=M 5 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
5. CRIMEADDRESS=TRIPOLI GENDER=M 5 ==> MARRIED=NO 5 <conf:(1)> lift:(1.08) lev:(0.03) [0] conv:(0.38)
6. CRIMEADDRESS=TRIPOLI 5 ==> GENDER=M MARRIED=NO 5 <conf:(1)> lift:(1.08) lev:(0.03) [0] conv:(0.38)
7. CRIMEADDRESS=BENGHAZI 4 ==> GENDER=M 4 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
8. CRIMEADDRESS=JAFARA 4 ==> GENDER=M 4 <conf:(1)> lift:(1 lev:(0) [0] conv:(0)
9. CRIMEADDRESS=JAFARA 4 ==> MARRIED=NO 4 <conf:(1)> lift:(1.08) lev:(0.02) [0] conv:(0.31)
10. CRIMEADDRESS=JAFARA MARRIED=NO 4 ==> GENDER=M 4 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)

The clusters obtained using kMeans algorithm were evaluated using by finding confusion matrix and calculating the

squared error function for the same i.e. they used an objective function here and minimized it for the clusters obtained. Clustering results obtained from this technique is shown below:

Cluster centroids:

Attribute	Full Data (13)	Cluster# 0 (9)	1 (4)
CRIMEID	13	13	65
CRIMETYPE	MOLESTATION	MOLESTATION	DACOITY
CRIMEADDRESS	TRIPOLI	JAFARA	TRIPOLI
CRIMEDATE	12OCT12	05NOV12	12OCT12
GENDER	M	M	M
MARRIED	NO	NO	NO
AGE	19	19	30

Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 9 (69%)

1 4 (31%)

The results were then visualized using data visualization for criminal age vs number of crimes which can be seen below:

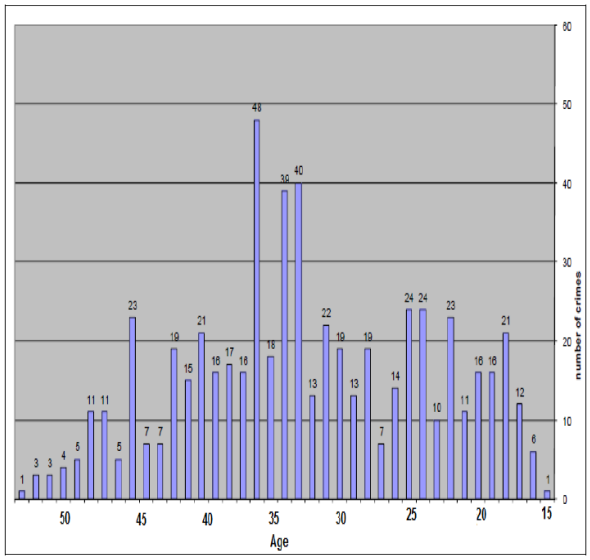


Fig.3: criminal age vs. number of crimes

The tools used in this process comprised of googleâĂŹs app engine, WEKA and the actual data set containing more than 7 attributes and 340 records. [10]

One of the team implemented an entire framework which can be used for different purpose of criminal analysis based on the requirement. Entity extraction is useful in analyzing important attributes from the data, this can be used automatically to identify person, address, vehicles etc. related to crime events. Clustering mechanism was used for this entity based detection, they also stated that some of the researchers used different kind of algorithm such as space âĂŞ based algorithm to automatically analyze these important attributes. In general clustering crime analysis can automate major part of crime analysis, but is limited when high computation is required. Association rule is also used by some of the researchers for discovering frequently occurring items sets and patterns in data. They used a new type of method namely AI Entity Extractor system, which used a three steps process to identify important attributes.

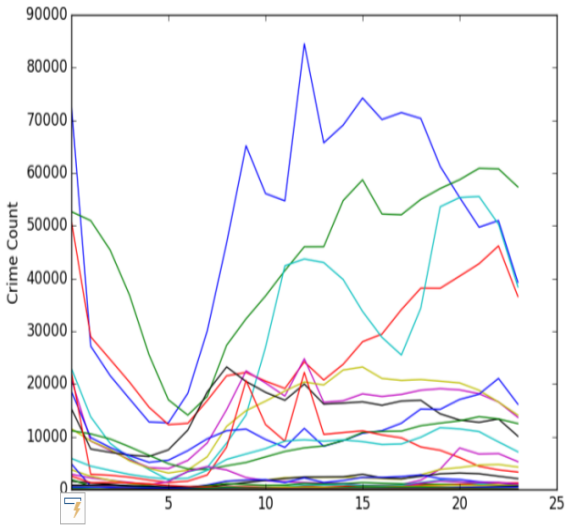
Another part of the entire framework was to find out suspects who provide false information every time and have multiple entries in the data set. Deceptive identity detection is the name of the technique used in this process. Out of the total number of records available 44 suspects with 120 duplicate records were found using technique of string comparator. In string comparator, they find the similarities between the values in different fields which is a sort of clustering. They normalized the similarity values between 0 and 1, and calculate the overall similarity between 2 records using Euclidean distance. For evaluating their results, they used hold-out validation method wherein they used two third of data as training and one third of data as testing. They got 97.4 and 94 percent accuracy with training and testing data respectively in this process.

Another approach in this process is crime-network analysis where similarity between suspects were found, similarity in which they form the groups for carrying out various illegal activities. In this process they decided to find the subgroups by using hierarchical clustering algorithm. This technique can be used for various other purpose like finding the behavioral patterns of the criminals, identity of criminals using clustering methods etc. [2]

[8] This research from Thailand used association rule mining, classification using decision trees, k-Nearest Neighbors and neural networks and clustering using k-means, hierarchical clustering and expectation maximization to predict trends in the data.

[4] Classification and clustering using Weka and Association by implementing Apriori Associator. Data visualization to predict crime trends in the data. Preprocessing has been used to keep the data set ready for the process. Entity extraction has been used to automatically identify person, address, vehicle, and personal properties from police narrative reports. Clustering techniques like Simple K-Means and DB Scan and EM have been used to cluster the city crime data mining depends on the crimes. Classification techniques such as NaÃŕve Bayes and Randomizable-classifiers have been used to detect criminal data from the city crime data base. Social network analysis has been used to analyze criminals roles and associations among entities in a criminal network

[9] Exploratory analysis to identify useful features and predictive model to predict if an arrest was made or not. Since positive and negative correlations in the trends of the selected features from the data proved to be more accurate that models such as NaÃŕve Bayes and conditional probability, linear regression and SVM were used for predictive analysis to save computation time.



Number of all types of crimes for a given hour of the day

Hamming loss was used to calculate a score which predicted if an arrest will be made given a new crime. Linear regression model was used to find out how many incidents will happen in a beat per day.

For assessing validity of this classification, the Hamming Loss is used where the total score is calculated as:

1- Hammingloss(x,y)

For this application, this can be interpreted as the percentage of correct predictions as to whether there was an arrest for the crime.

Two clustering techniques, K-means and DBScan algorithm are used to automatically associate different objects in crime records. Also entity extraction has been carried out to automatically identify people, their addresses, vehicles and personal properties from police narratives. Deviation detection has been used to trace abnormal activities which can turn out to be criminal activities like fraud detection, network intrusion etc. Classification has been used to detect crimes involving email spamming and unsolicited emails. String comparator is also used to criminal profiles and their associations to detect entities in a criminal network. [7]

1.9 How do you plan to solve

We have collected the crime data set for the city of chicago for year 2017. In this data we have removed the redundant attributes like longitude and latitude as it was already present in location attribute. We have also removed the outliers we found or the records containing the missing data. After this we will segregate the data into training and testing set using the traditional 70-30 rule and we are also planning to use 10 fold cross validation.

In our project the hypothesis is that given a region in terms of block, district or location we will try and predict it into 2 class, Safe or Not Safe. Also we will predict which time during the day that region is most safe and at which time it is the safest.

To achieve our goal, we will have to create new features based in the data which we have and based on those features we will use Decision Tree classification algorithm to classify our data. We are also planning to segregate our data based on kMeans clustering technique where the cluster will range from safest, safe, not safe, worst.

1.10 Result Validation

For the classification process, we will validate our results based on the accuracy, precision and recall values. To calculate these values, we will need the confusion matrix from which we will get the numbers for true positive, true negative, false positive and false negative. In clustering technique we can check SSE values for different number of clusters from 2 to 4. Since we are classifying the records as Safe and Not Safe, 2 clusters would be a good option, but that needs to be verified using SSE values.

1.11 Algorithms to be used

We will be using Decision Tree, KNN for classification process. Before that we are planning to use One R rule to determine the most important attribute in our data. Now, for the clustering approach we will use kMeans as a main technique and we might use agglomeration clustering when we further find sun region in a given region.Apart from this, we might use DBScan algorithm to find highly dense region in terms of crime.

2. IMPLEMENTATION

2.1 Algorithms Used

Initially we cleaned the data, preprocessed it and created different data sets as per our requirements which have been explained further in the report.

Multiple algorithms were performed for this iteration:

1. ZeroR to get a biased decision based on the data we have.
2. OneR classifier to get the best attribute in our data,
3. kMeans clustering technique was performed, where we tried all the possible values of k and calculated its sum of squared error(SSE) to find the ”KNEEâĂİ point in our graph to determine a good value for k.
4. Decision Tree Classifier using J48 was implemented by altering various parameters like minNumObject and confidenceFactor we observed the complexity of our model in terms of number of leaves, size of a tree and itâĂŹs accuracy to classify the test data. The observation was made to find the best fit model where the complexity of our tree was less and the accuracy, not much affected.
5. Naive Bayes Classifier, where we have used 2 different data sets with sets of attributes as per our requirement and compared the accuracy of different models built using 10
6. k - Nearest Neighbors was used to classify whether there will be an arrest for a given crime at a particular location and at a particular time.

2.2 Challenges Faced

The required data was not in the format we wanted to perform classification, clustering and exploratory analysis upon. The date attribute was in time stamp format which was of

Longitude	948019
Location_Coeff	947999

Figure 10: Difference between Location-Coeff and Latitude.

no use to differentiate identify the crime pone hours,days and months in our data and so on.

Initially our data set had 192502 records and total of 22 attributes consisting of crime data for the first 9 months of 2017 in the City of Chicago. Getting rid of unimportant attributes and selecting important attributes, was the initial challenge we faced.

To reduce the number of records while ensuring its equal proportion in terms of months, ward, arrests etc was a challenge. To solve this, we created different subsets to perform different analysis from the main data.

Selecting a criteria to classify a given ward as "SAFE" and "UNSAFE" was difficult. We then decided to count the number of crimes that took place in a particular ward and if found to be exceeding a predetermined threshold, declared the ward as "UNSAFE".

Nominal attributes were required for the classification using dec tree, for this purpose attributes had to be converted to Nominal from Numerical.

Newly created feature, location-coefficient was of numerical type. It was difficult converting this attribute to nominal form because of duplicates in it. Also, converting and maintaining the data in CSV format for frequent use was bit difficult.

2.3 Interesting Facts

In a data set consisting of equal number of records for every month, the resulting decision tree built had the "Month" attribute as its root node. Here, an attribute with equal number of records in each class was the one with the greatest amount of entropy and hence became the root node. Surprisingly, for the processed data, SSE did not reduce greatly from the start to the "KNEE" point.

Minimum number of objects heavily affected the number of leaves in the decision tree and the size of the tree. We observed, as the number of minimum number of objects in the final class went up, the number of leaves in the tree reduced drastically.

In Naive Bayes, while classifying a particular region as "SAFE" or "UNSAFE", great amount of accuracy was observed for the testing data even when only 10

New feature location-coeff was created by dividing the latitude and longitude, in this method, there is a great chance that the value of location-coeff would be same for different set of inputs as division for different value can be same, but strangely very few values were matching, as few as 19 records.

None of the attributes had negative correlation coefficient, and X-Y co -ordinate had a correlation coefficient of nearly

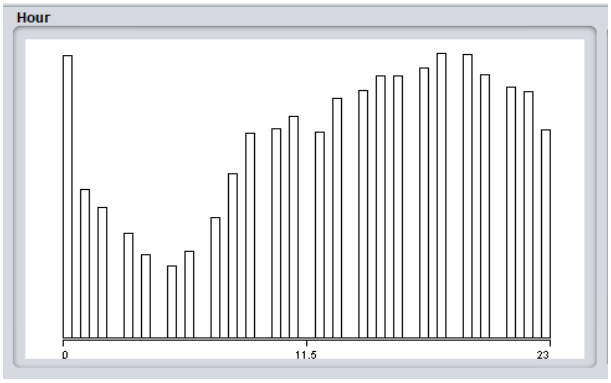


Figure 11: Hourly crime rate in a single day.

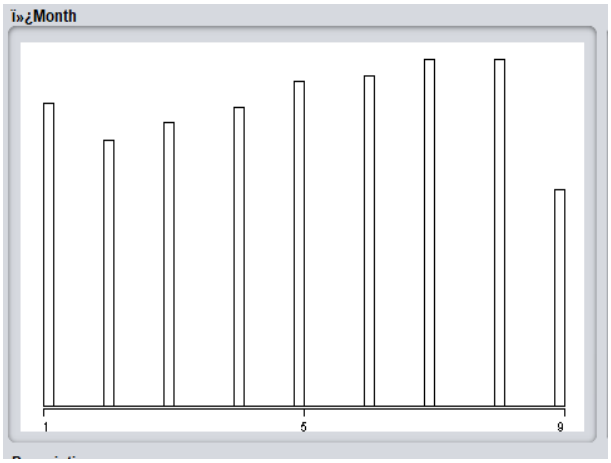


Figure 12: Monthly crime rate

1.

In a single day, crime rates rose dramatically during the evening hours and went down only by the wee hours of dawn. we can see this in figure above(fig 11).

Crime rate increased between summer and fall period of the year.(fig 12)

Observed location-coefficient can be observed in the below image, which was obtained from latitude and longitude.(fig 13)

Wards with the largest number of crimes, did not experience the largest number of arrests.(fig 14)

Number of arrests for any particular month remained almost the same throughout the year.(fig 15)

Trend was observed for different hours of the day, throughout the year.(fig 16)

Similar trend was also observed for different days of the month, throughout the year.(fig 17)

Some of the primary types of criminal activity accounted for greater frequency of occurrence as compared to others, throughout the year.(fig 18)

The location-coefficient feature surely shows different clusters when hourly comparison is made with it and we see that most of the crimes happened in the midtown of Chicago.(fig 19)

Note: Images of above facts are in next page.

2.4 Best Algorithms

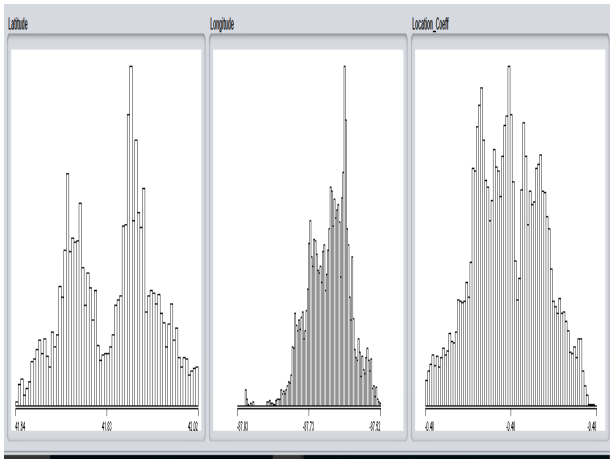


Figure 13: Latitude, Longitude and Location-Coeff

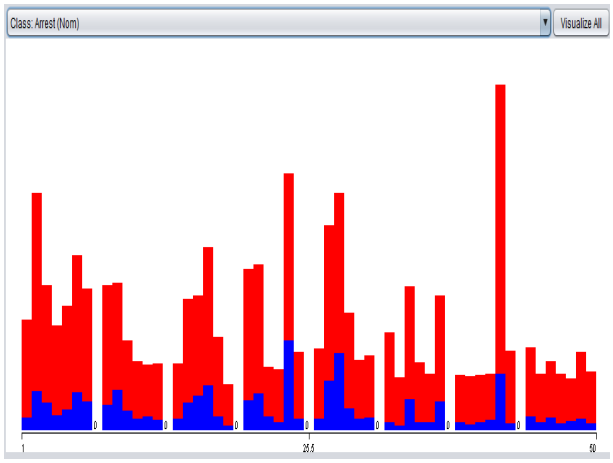


Figure 14: Ward wise arrest. Red - Not Arrest, Blue - Arrest

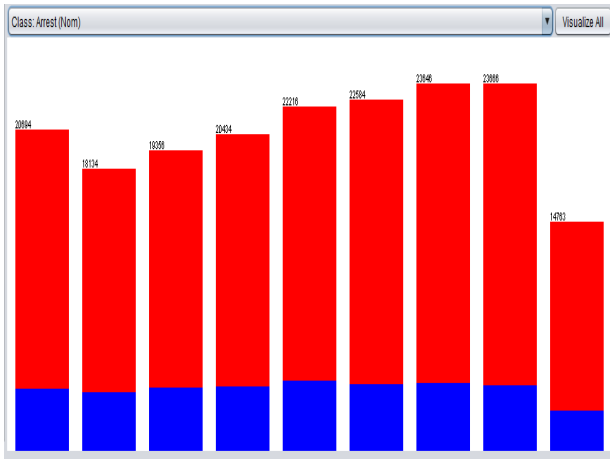


Figure 15: Month wise arrest. Red - Not Arrest, Blue - Arrest

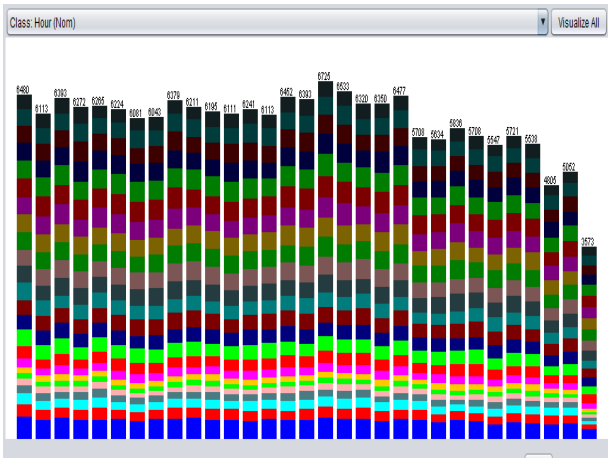


Figure 16: Hourly analysis of each day in a month

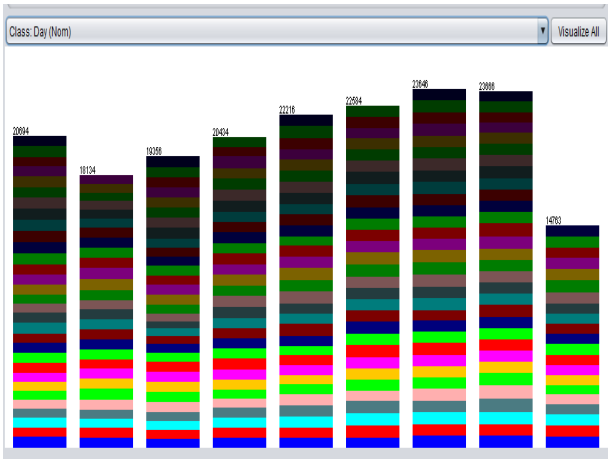


Figure 17: Monthly analysis of each day in a month

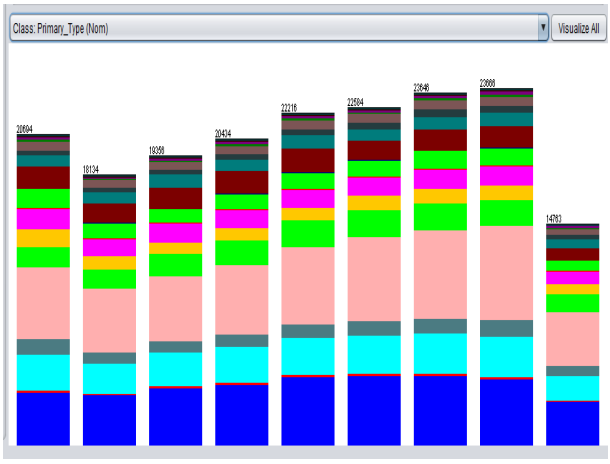


Figure 18: Frequency of primary types.

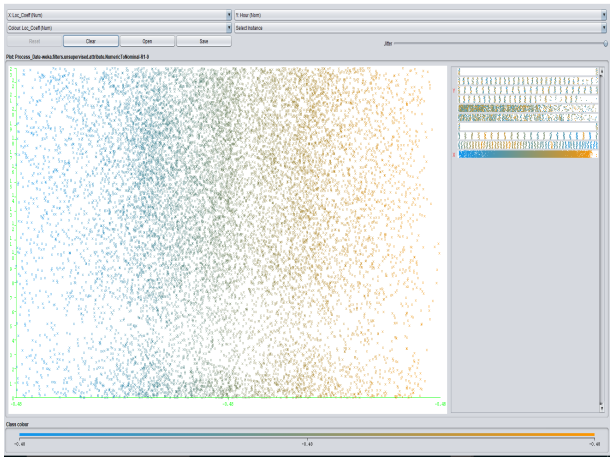


Figure 19: Hourly analysis against Location Coefficient

Most algorithms yield required results with good accuracy for every model. Naive Bayes classification gave pretty good amount of accuracy even when less amount of data was used to build the model. Naive Bayes classification was used to predict whether there will be an arrest or whether a particular ward is safe or not. In both the cases, the accuracy was pretty good given the observed attributes.

Decision tree classifier using J48 also gave accuracy of about 89

2.5 Data Mining Significance

With the increase in criminal activities, it becomes imperative to perform crucial analysis which can assist authorities counter the criminals swiftly without consuming much resources. If this criminal data is converted into information, we can extract vital insights from the data and this wisdom can boost the crime solving front. Data mining is the process of extracting meaning from raw data to predict relevant trends. There are two aspects to this process however. One being proper categorization of data and the other, understanding the true meaning of data which can confirm the trends.

For our project, weâĂŹve analyzed the given data to predict relevant acumen whether a given ward is Safe or Unsafe based on the frequency of crimes in the ward, given a crime whether an arrest will be made or not among other predictions and this was possible by data mining techniques which helped us classify and visualize the data accurately for our required research.

2.6 Data Cleaning

We selected the crime data of 2017 for the city of chicago from, $https://data.cityofchicago.org/Public-Safety/Crimes-2017/d62x-nvdr$. It had total of 192502 records along with 22 attributes. The attributes are:

ID: Unique identifier
CaseNum: Case number for crime report.
Date: Timestamp of the event occurred.
Block:Block where the crime occurred.
IUCR: Illinois Unified Crime Reporting code.
PrimaryType: Major category of crime.
Description: Subcategory of crime.
Location Description: Location where it happened.
Arrest: Whether there was an arrest or not.
Domestic: whether it is a domestic crime or not.
Beat: Police beat where crime occurred.
District: District where crime occurred.
Ward: Ward where crime occurred.
Community Area: Community area code
FBI Code: FBO code of conduct for the place and crime.
X Coordinate: Geographical location of crime.
Y Coordinate: Geographical location of crime.
Year: Year when the crime took place.
Updated On: Case entry last updated on date.
Latitude: Geographical location of crime.
Longitude:Geographical location of crime.
Location: Combination of latitude and longitude.

Out of these attributes, not all the attributes were required for our analysis. ID and case number are the unique values for identifying the records, hence were discarded. For location of the crime we had block, district, ward, FBI code, community code out of all theses we decided to go with ward as each of them covered good amount of area, not too big as district and not too small as block. So block, FBI code, Community Code were discarded. Though we kept district because if we want to classify some instance district wise, we can do that. Ward was the major attribute used for location identification.

IUCR is a code for the local police department and hence not required for trend analysis. Domestic tells whether a given crime is domestic or not, since we did not plan to do any analysis with this attribute, it was discarded. Community Area code and FBI code also does not had any significance, hence removed. Year and Updated On are used for maintenance purpose, hence removed.

Observing the date attribute, we cannot generally get any pattern from this field in the given format, but if we break down this attribute into something like the month in which crime took place, day or the hour of the day when it took place, then we can definitely find a patterns and trends and give more accurate information to the police to help them secure the place like which month is more prone to crimes, which days are more prone to crimes or at what time of the day we can expect more crimes.

So, we went ahead and broke this date attribute further into three different attributes namely, "Month", "Day" and "Hours". Before doing this break, the time which was present in 12 hours format along with "AM" and "PM" was also converted to 24 hours format which makes the the processing work less while actual analysis. We used python for creating this attributes and created a new data set named "Process-Data".

Below is the snapshot of the python code that was used to create this new data set.

We then went ahead and found the correlation coefficient between X , Y coordinates and Latitude and Longitude, since both of them were used for geographical analysis we had a hint that this might be something same. Not to our surprise, the correlation coefficient between these variables were almost near to 1 and hence we got rid of X Coordinate and Y Coordinate attribute.

Location attribute which was just a combination of latitude and longitude was removed and a new feature called location-coeff was created by dividing latitude by longitude. This work was done in using excel.

As the dataset had huge amount of data, we reduced our

```
with open('C:\Jinesh\Semester 3\Big Data Analytics\Project\Crime_Data.csv') as file:
    dateLst=[]
    fLine=file.readline().strip().split(",")
    frstLine=fLine[0:2]
    frstLine.extend(("Month","Day","Hour"))
    frstLine=frstLine+fLine[3:]
    print(frstLine)
    file=file.readlines()[0:]
    fle=open("C:\Jinesh\Semester 3\Big Data Analytics\Project\Process_Date.csv","w+")
    fle.write(",".join(frstLine)+"\n")
    for line in file:
        date=line.split(",")[2].split(" ")
        date[1]=str(int(date[1][0:2])+12)+date[1][2:]
        if date[1][0:2]=="24":
            date[1]="00"+date[1][2:]
        date=date[0:2]
        temp=line.split(",")[0:2]
        temp.append(date[0][0:2])
        temp.append(date[0][3:5])
        temp.append(date[1][0:2])
        temp=temp+(line.split(",")[3:])
        fle.write(','.join(temp))
        dateLst.append(temp)
        print(temp)
    fle.close()
```

Figure 20: Python Script for Process Data

	X_Coordinate	Y_Coordinate	Latitude	Longitude
X_Coordinate	1			
Y_Coordinate	-0.542417769	1		
Latitude	-0.545149776	0.999994671	1	
Longitude	0.999915397	-0.5314954	-0.534249847	1

Figure 21: Correlation Coefficient Matrix

K
Loc_Coeff
-0.478196393
-0.476232969
-0.476992471
-0.477207727
-0.478552664
-0.47667424
-0.476242111
-0.476408028

Figure 22: Location Coefficient

```
with open('C:\Jinesh\Semester 3\Big Data Analytics\Project\Process_Date.csv') as file:
    firstLine=file.readline()
    data=[]
    fle=open('C:\Jinesh\Semester 3\Big Data Analytics\Project\Every_Month_Data.csv','w+')
    fle.write(firstLine)
    for line in file:
        line=line.split(",")
        data.append(line)
    actualData=[]
    for ind in range(1,10):
        tmp=[x for x in data if x[0]==str(ind)][0:1500]
        for record in tmp:
            fle.write(",".join(record))
    fle.close()
```

Figure 23: Python Script for Every Month Data

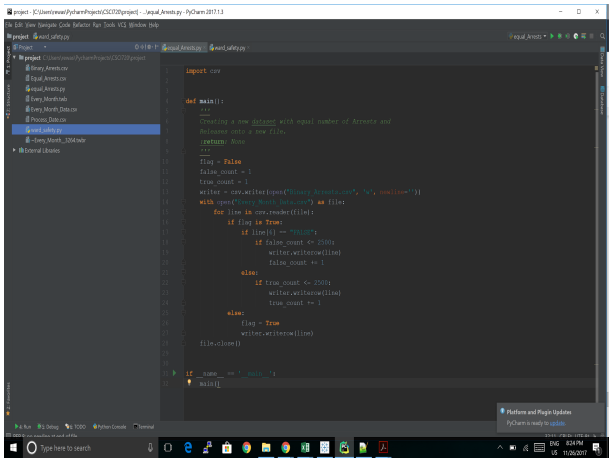


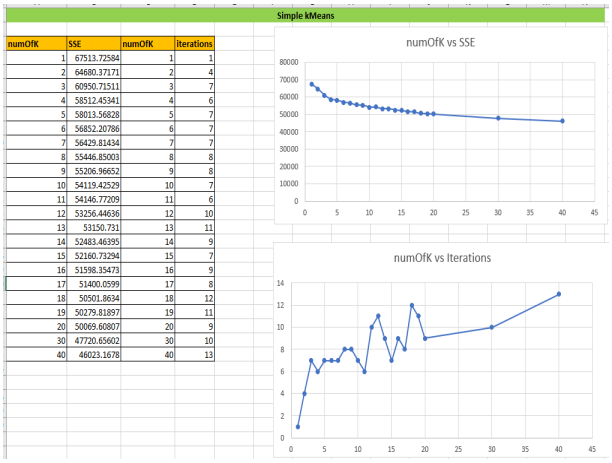
Figure 24: Python Script for Equal Arrest Data

data set to 15000 records which roughly covers first month of crime records. Now to analyse the data for patterns and trends month wise, we created another data set where in 1500 records of each month were collected and a separate data set âĂIJEvery-month-data.csvâĂİ was created. Here also we used python to create this new data set. The python script used for this is shown below:

As the number of crime that took place in the city of chicago is huge, we did not see the similar numbers in arrest and hence majority of the records in the data does not include arrest, Since this was our target variable for some analysis and to make this analysis of pattern unbiased, another data set was created which had equal number of arrests and not arrest records. Again python was used for creating this data set whose script is below.

To predict whether a given ward is ”SAFE” or ”UNSAFE”, we included a new feature attribute called ”Safety” where we assigned a given ward as ”SAFE” or ”UNSAFE”. To do this we counted the number of crimes that has taken place in a particular ward and decide a threshold of 299 on that value above which a ward is ”UNSAFE” and below which it is ”SAFE”.

2.7 Results



numOfK	SSE	numOfK	Iterations
1	67513.72584	1	1
2	64680.37171	2	4
3	60950.71511	3	7
4	58512.45341	4	6
5	58013.56828	5	7
6	56852.20786	6	7
7	56429.81434	7	7
8	55446.85003	8	8
9	55206.96652	9	8
10	54119.42529	10	7
11	54146.77209	11	6
12	53256.44636	12	10
13	53150.731	13	11
14	52483.46395	14	9
15	52160.73294	15	7
16	51598.35473	16	9
17	51400.0599	17	8
18	50501.8634	18	12
19	50279.81897	19	11
20	50069.60807	20	9
30	47720.65602	30	10
40	46023.1678	40	13

Figure 25: kMeans Results

MinNumObj	100
Percentage Split	70-30(Train, Test)
Total Instances	15000
TargetField	Arrest

Confidence Factor	Correctly_Classified(%)	Number of Leaves	Size of Tree
0.05	89	539	543
0.15	89.2889	539	543
0.25	89.2889	539	543
0.35	89.3111	953	959
0.45	89.5333	305	307
0.55	89.9778	402	405
0.65	89.9778	402	405

MinNumObj	500
Percentage Split	70-30(Train, Test)
Total Instances	15000
TargetField	Arrest

Confidence Factor	Correctly_Classified(%)	Number of Leaves	Size of Tree
0.05	88.1556	235	237
0.15	88.1556	235	237
0.25	88.1556	208	209
0.35	89.6889	208	209
0.45	89.6889	208	209
0.55	89.6889	208	209
0.65	89.6889	208	209

Figure 26: Decision Tree J48 Results

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.03 seconds

=== Summary ===

Correctly Classified Instances        3995               88.7778 %
Incorrectly Classified Instances       505               11.2222 %
Kappa statistic                          0.6129
Mean absolute error                      0.1456
Root mean squared error                  0.295
Relative absolute error                 47.0172 %
Root relative squared error             74.331  %
Total Number of Instances             4500

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.608    0.044    0.770      0.608    0.680      0.619    0.894     0.778     TRUE
                 0.956    0.392    0.909      0.956    0.932      0.619    0.894     0.966     FALSE
Weighted Avg.    0.888    0.324    0.882      0.888    0.883      0.619    0.894     0.929

=== Confusion Matrix ===

   a    b   <-- classified as
 536  345 |   a = TRUE
 160 3459 |   b = FALSE
```

Figure 27: Naive Bayes Result

```
Time taken to build model: 0.07 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       13332               88.8859 %
Incorrectly Classified Instances      1667               11.1141 %
Kappa statistic                          0.6041
Mean absolute error                      0.1439
Root mean squared error                  0.295
Relative absolute error                 46.8014 %
Root relative squared error             75.2405 %
Total Number of Instances            14999

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.594    0.042    0.768      0.594    0.670      0.611    0.890     0.765     TRUE
                 0.958    0.406    0.910      0.958    0.933      0.611    0.890     0.965     FALSE
Weighted Avg.    0.889    0.337    0.883      0.889    0.883      0.611    0.890     0.927

=== Confusion Matrix ===

    a     b   <-- classified as
 1689  1156 |     a = TRUE
  511 11643 |     b = FALSE
```

Figure 28: Naive Bayes, 10 fold cross validation

Simple kMeans

On the processed-data containing 15000 records rough for the entire month of January, we performed Simple kMeans algorithm for different values of k starting from 1 to 20 and for 30 and 40. For each and every value of k we observed the sum of squared error (SSE) and number of iterations it took for getting proper clusters.

From the above figure we observed that the SSE decreased by good amount in the start and after "KNEE" point reduction in SSE was very small. For our data value of k as 11 looks perfectly fine and gives good clustering.

Decision Tree J48

We then performed the decision tree classifier algorithm on the same data set, where we used J48 for making the tree.

Here we played with different inputs in WEKA, for creating a decision tree. Initially for the 15000 records we kept the minimum number of objects in the class to be 100 and used WEKA's split mode where in the model was created using 70 percent of the data and remaining of the data was used for testing on that model. Also we kept on varying the confidence factor for each run which helps in pruning the result. We observed that as confidence factor increased the accuracy increased by extremely small amount and the complexity of the tree also decreased i.e. The number of leaves and size of the tree. The target variable here was whether there will be an arrest or not.

It was also observed that the complexity of the tree is less, when the number of minimum objects is significantly larger in terms of data. For different model the accuracy observed was around 89 percent which is significantly good.

Naive Bayes

Naive Bayes Classification was used on two different data set keeping in mind 2 different target values, one where we predicted whether an arrest will be made or not and with other data set where we predicted whether a given ward is SAFE or UNSAFE.

Initially we performed the Naive Bayes classification on processed data to predict whether there will be an arrest or not, for this we used 70 percent of the data as training data and rest 30 percent as testing data. For this we achieved an accuracy of around 88 percent on the testing data.

We also verified our results for the same data and same target using "10 fold cross validation" to get the similar accuracy.

On another data set, we classified a ward as SAFE or UN-

```
Time taken to build model: 0.05 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.06 seconds

=== Summary ===

Correctly Classified Instances        4494               99.8667 %
Incorrectly Classified Instances         6                0.1333 %
Kappa statistic                          0.9972
Mean absolute error                      0.005
Root mean squared error                  0.0306
Relative absolute error                  1.0742 %
Root relative squared error              6.308  %
Total Number of Instances             4500

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.999    0.002    0.999      0.999    0.999      0.997    1.000     1.000     UNSAFE
               0.998    0.001    0.999      0.998    0.998      0.997    1.000     1.000     SAFE
Weighted Avg.  0.999    0.002    0.999      0.999    0.999      0.997    1.000     1.000

=== Confusion Matrix ===

    a    b   <-- classified as
 2808    2 |    a = UNSAFE
    4 1686 |    b = SAFE
```

Figure 29: Naive Bayes, 70 percent training data

```
Time taken to build model: 0.03 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.03 seconds

=== Summary ===

Correctly Classified Instances        7494               99.9333 %
Incorrectly Classified Instances         5                0.0667 %
Kappa statistic                          0.9986
Mean absolute error                      0.0064
Root mean squared error                  0.0316
Relative absolute error                  1.3781 %
Root relative squared error              6.552  %
Total Number of Instances             7499

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    0.001    0.999      1.000    0.999      0.999    1.000     1.000     UNSAFE
               0.999    0.000    0.999      0.999    0.999      0.999    1.000     1.000     SAFE
Weighted Avg.  0.999    0.001    0.999      0.999    0.999      0.999    1.000     1.000

=== Confusion Matrix ===

    a    b   <-- classified as
 4741    2 |    a = UNSAFE
    3 2753 |    b = SAFE
```

Figure 30: Naive Bayes, 50 percent training data

SAFE using Naive Bayes classification, here we used different proportion of training data and testing data to observe whether the accuracy is affected or not and to our surprise we got the similar accuracy for 10-90 ratio of training-testing data as we got for 50-50 and 70-30 data proportions. Not only this the accuracy of our data was around 99 percent on the testing data even when only 10 percent of the training data was used.

kNN

We also implemented k nearest neighbor algorithms on the same data set for odd values of k from 1 to 21 and 30 and observed the accuracy of the classification of testing data which increased till a particular value of k and then decreased helping us find the perfect value of k.

Here we used 70-30 proportion for training and testing data set and used Arrest as a target variable. We observed an accuracy of about 87 percent for value of k as 5.

This algorithm can also be used for cleaning the data which can enhance your results and improve on bad results which might have been caused due to outliers.

OneR and Decision Tree on Equal Arrest From the analysis by far, it was difficult to correlate the Arrest prediction with its correctness since the number of releases where

```
Time taken to build model: 0.01 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.04 seconds

=== Summary ===

Correctly Classified Instances       13393               99.2148 %
Incorrectly Classified Instances       106                0.7852 %
Kappa statistic                          0.983
Mean absolute error                      0.0267
Root mean squared error                  0.085
Relative absolute error                  5.7347 %
Root relative squared error             17.6835 %
Total Number of Instances            13499

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.996    0.015    0.991      0.996    0.994      0.983    1.000     1.000     UNSAFE
               0.985    0.004    0.993      0.985    0.989      0.983    1.000     0.999     SAFE
Weighted Avg.  0.992    0.011    0.992      0.992    0.992      0.983    1.000     1.000

=== Confusion Matrix ===

    a    b   <-- classified as
 8572   32 |    a = UNSAFE
   74 4821 |    b = SAFE
```

Figure 31: Naive Bayes, 10 percent training data

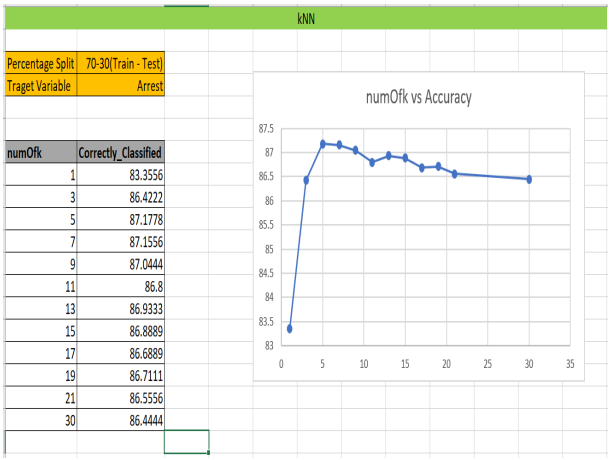


Figure 32: kNN Result

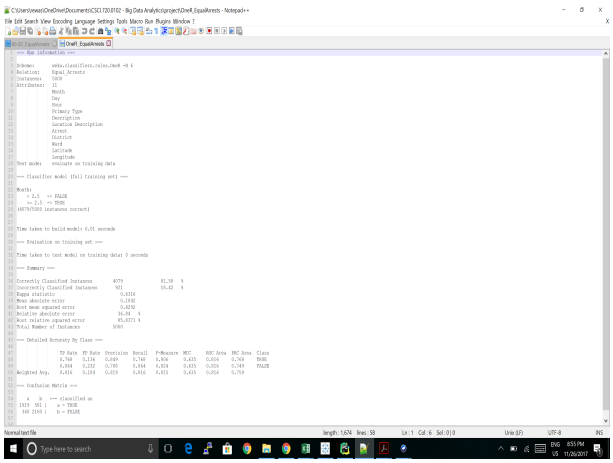


Figure 33: OneR rule

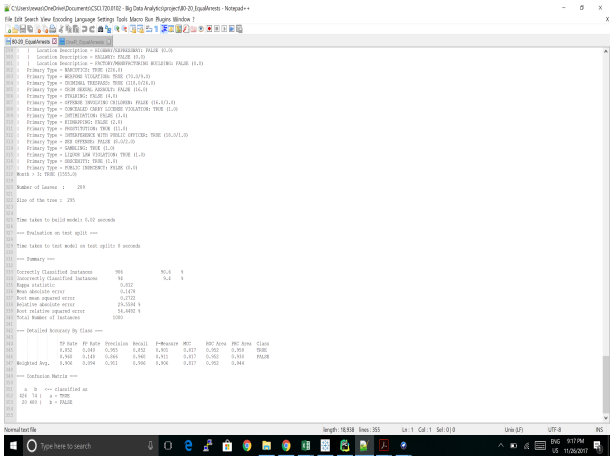


Figure 34: Decision Tree on equal arrest data

humongous compared to number of arrests made. Thus, we created a new data set with equal numbers of arrests and releases which helped us get a better insight and achieve higher accuracy with testing data. This data consisted of 5000 records, 2500 of each case.

Performing OneR rule classifier on the Equal Arrests data helps us find the most important attribute in the data i.e. Month on the basis of which the classification commences. We get a decent accuracy of 81

The same equal arrests data was passed through the J48 Decision Tree classifier. Again, attribute "Month" was found to be at the root node. Equal arrests dataset comprised of 5000 data records of which 80

2.8 Algorithm Description and Parameters Used

We performed two vital operation on the processed data set, clustering and classification algorithm. Clustering to find similar records in one cluster which shows similar characteristic and classifications to classify the test and validation data using the model created using training data.

For all the algorithms we used the following attributes:
Month: Month, when crime took place
Day: Day at which crime took place
Hour: hour of the day when crime took place
Primary Type: Major type of crime
Description: Subcategory of crime
Location Description: Location where crime took place
Arrest: Whether there was an arrest or not.
District: District in which crime took place
Ward: Ward in which crime took place
Safety: Whether a given ward is SAFE or UNSAFE
Loc-Coeff: Location coefficient.

For clustering similar records in one cluster we firstly used kMeans algorithm which takes k random point from the data set as k cluster center and then calculates the distance between all the data points and cluster center then assigns that data point to a cluster which is nearer to it. This process is completed for all the data points and once that is done, cluster centers are recalculated by averaging the values of data points in that cluster. This process is continued till there is no or minimal shift in the cluster center.

We apply this method for different values of k and find the k which defines the KNEE point i.e. from that point the SSE reduction is very less. This value of k best describes the number of clusters. We use the percentage split method of WEKA for this clustering.

Now, for classification we used various different methods of classification with various parameter and different data sets to predict different target value.

Initially in Dec Tree classification where at every stage we split the records at one attributes based on the amount of entropy in that attribute and we stop based on the stopping criteria we decided earlier. Here we used 70-30 ratio of training- testing data and observed the complexity of the tree by varying amount of confidence factor and minimum number of objects. The complexity decreased with more number of minimum number of objects and confidence factor whereas the accuracy remained same near about.

In Naive Bayes Classification, we calculate the probability of each value of target class given the observation and one with more probable is selected as target class. Here for different amount of training and testing data set we created and tested the model to get a superb accuracy for each one of them. We classified for a given ward, month, day , hour whether there will be an arrest or not and given a month, day, hour and ward, whether that ward is safe or not.

We implemented similar kind of algorithm called k- nearest neighbor algorithm which classifies a given data point based on the majority of k nearest data points. Here also we classified whether there will be an arrest or not for a given crime record for different odd values of k and observed its accuracy.

Now, for the data set containing equal number of arrests and release, we have implemented OneR rule and J48 Decision Tree with an 80:20 Training:Testing Split. The description of the algorithms and the parameters used for it are given below:

OneR : OneR rule was performed to find out the most dominant attribute in the data which turned out to be the number of months. Additional parameters included minBucketSize = 6 with a batchSize of 100 upto 2 decimal points.

J48 Decision Tree : J48 decision tree classifier helped us achieve a model which was fast and decently accurate with the testing data. Splitting criterion being information gain or difference in entropy, we set the minNumObj = 100 and confidenceFactor to 0.4

3. REFERENCES

[1] J. Agarwal, R. Nagpal, and R. Sehgal. Crime analysis using kmeans clustering.

[2] H. Chen, W. Chung, J. J. Xu, and G. W. Y. Qin. Crime data mining : A general framework and some examples.

[3] S. Devan and S. Gangadharan. Crime analysis and prediction using data mining.

[4] A. Malathi, S. S. Baboo, and A. Anbarasi. An intelligent analysis of a city crime data using data mining.

[5] C. B. Manish Gupta and G. M. P. Crime data mining for indian police information system.

[6] A. Paper and R. Kumar. Analysis and design of an algorithm using data mining techniques for matching and predicting crime.

[7] Y. S and S. B. N. Datamining techniques to analyze and predict crimes.

[8] U. Thongsatapornwatana. A survey of data mining techniques for analyzing crime patterns.

[9] J. Wheeler, N. Moreno, and A. Kanak. Predictive policing on crime data.

[10] Z. S. ZUBI and A. A. MAHMMUD. Crime data analysis using data mining techniques to improve crime prevention.