

# Crime Prediction Using Machine Learning and Data Analytics



By

**Muhammad Naqi Haider**

**MSCS600000171469**

Supervisor

**Dr. Rafia Mumtaz**

**Department of Computing**

A thesis submitted in partial fulfillment of the requirements for the degree  
of Masters in Computer Science (MS CS)

In

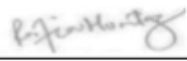
School of Electrical Engineering and Computer Science,  
National University of Sciences and Technology (NUST),  
Islamabad, Pakistan.

(July 2020)

## Approval

It is certified that the contents and form of the thesis entitled "Crime prediction using machine learning and data analytics" submitted by MUHAMMAD NAQI HAIDER have been found satisfactory for the requirement of the degree

Advisor : Dr. Rafia Mumtaz

Signature: 

Date: 27-Jul-2020

Committee Member 1:Asad Shah

Signature: 

Date: 27-Jul-2020

Committee Member 2:Ms. Hirra Anwar

Signature: 

Date: 27-Jul-2020

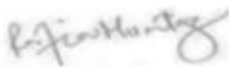
Committee Member 3:MS. Iram Fatima

Signature: 

Date: 28-Jul-2020

## THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis entitled "Crime prediction using machine learning and data analytics" written by MUHAMMAD NAQI HAIDER, (Registration No 00000171469), of SEECs has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: 

Name of Advisor: Dr. Rafia Mumtaz

Date: 27-Jul-2020

Signature (HOD): \_\_\_\_\_

Date: \_\_\_\_\_

Signature (Dean/Principal): \_\_\_\_\_

Date: \_\_\_\_\_

# Dedication

*Dedicated to my parents, siblings and mentors without whose continuous support and guidance this research wouldn't have materialized as swiftly.*

# Certificate of Originality

I hereby declare that this submission is an original work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's de-sign and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: Muhammad Naqi Haider

Signature:

A handwritten signature in black ink, appearing to read 'Muhammad Naqi Haider', written in a cursive style. The signature is positioned above a horizontal line.

# Acknowledgment

I am thankful to Allah the almighty to have directed me throughout this research at each and every step and for every new idea which you incepted in my mind to improve it. Indeed, I could have done very little if it weren't for your guidance and direction. Whoever helped me throughout the course of my research, whether my family or anyone else was your will, so indeed none be worthy of praise except you.

I am greatly thankful to my beloved parents for raising me when I wasn't capable of walking and for continued unconditional support throughout in each and every department of my life.

Also, I would like to express special thanks to my supervisor Dr. Rafia Mumtaz for her support and help throughout my research and encouraging me.

I would also like to pay special thanks to my co-advisors and committee members for their cooperation and tremendous support. Every time I was stuck in some problem, they would come up with the solution. Without their help I would not have been able to complete my thesis effectively.

Finally, I would like to express my gratitude to all the individuals who have rendered valuable assistance to my study.

# Table of Contents

<b><u>CHAPTER 1: INTRODUCTION</u></b> .....	1
1.1 Motivation and Objectives .....	1
1.2 Key Contributions.....	5
1.3 Organization of Thesis.....	6
<b><u>CHAPTER 2: LITERATURE REVIEW</u></b> .....	7
2.1 Research Works.....	7
2.2 Summary of Literature Review.....	12
<b><u>CHAPTER 3: METHODOLOGY</u></b> .....	13
3.1 Data Extraction.....	15
3.2 Data Visualization.....	19
3.3 Outlier Removal.....	25
3.4 Handling Missing Values.....	27
3.5 Reduction of Class Imbalance.....	28
3.6 Feature Extraction.....	29
3.7 Scaling or Standardization.....	31
3.8 Clustering.....	32
3.9 Machine Learning Algorithms.....	35
<b><u>CHAPTER 4: RESULTS AND DISCUSSION</u></b> .....	39
<b><u>CHAPTER 5: CONCLUSION AND FUTURE WORKS</u></b> .....	42
5.1 Conclusion.....	42
5.2 Future Works.....	42
<b><u>REFERENCES</u></b> .....	44

# List of Figures

- Figure 1: Flow Chart of Methodology
- Figure 2: Datasets on Chicago Data Portal
- Figure 3: Number of reported Crimes by Date
- Figure 4: Number of reported crimes by Arrest
- Figure 5: Number of reported crimes by District
- Figure 6: Number of reported Crimes Domestic or not
- Figure 7: Number of reported crimes by Location Description
- Figure 8: Number of reported crimes by Crime Type
- Figure 9: Scatterplot of x-y coordinates before outlier removal
- Figure 10: Scatterplot of x-y coordinates after outlier removal
- Figure 11: Logistic Regression Model
- Figure 12: Counts of Crime categories in sample Dataset
- Figure 13: Crime counts by Day of the Week
- Figure 14: Division by District Beat Community and Ward
- Figure 15: Visualization of k-means
- Figure 16: Clusters created by k-means
- Figure 17: Entropy Explanation for Decision Tree



# List of Tables

- Table 1: Property Crimes in Lahore City
- Table 2: Characteristics of Columns in Dataset
- Table 3: Using Top 10 Crime Types
- Table 4: Using top 3 Crime types
- Table 5: Detailed Results of the Algorithms

# Abstract

Crimes have both short term and long term effects on individuals and on a society as a whole. Modern day law enforcement agencies are making use of data analytics and machine learning algorithms for predictive policing in order to prevent crimes. Using historical data, data analysts can find key factors that contribute to crimes and predict the occurrence of a particular crime type. This problem is popularly referred to as crime classification problem. Due to increasing need of more accurate systems, this thesis presents improved use of machine learning algorithms to predict the crime types and also analyzes their performance given common set of features. The dataset used for this work is Chicago Crime Dataset which is available on City of Chicago Data Portal. Related research works in the past have used data from multiple domains such as geography, socio-economics and data also data from education is also used. Data from multiple domains is treated but had a difficulty in retrieving nonlinear relationships and creating distinction between multiple values of the feature set. In this research work we specially focused on extracting features with the help of visualization and clustering in light of results of multiple iterations while running algorithms. This study also captures the importance of crime prediction systems in society, it presents a literature survey of the systems in place and provides a methodology to predict crimes in advance using machine learning algorithms. The results also show that the dataset suffers from high number of distinct values for features, class imbalance and lack of independent features. Among the three Algorithms of Naïve Bayes, Decision Tree and Random Forest, Random Forest performed the best with 54% accuracy. For future studies it is better using Random Forest Classifier, extracting more independent features from the data and sampling data in order to reduce the class imbalance problem. This study also highlights the importance of crime reporting in Pakistan in order to make an accurate crime prediction system.

# 1. Introduction

## 1.1 Motivation & Objectives

A legally prohibited action or instance of negligence which might be injurious to the public welfare or morals and to the interests of state is called a Crime. It is that illegal act which might result in an individual being punished by the government. In layman terms, a delinquent action that may be punishable by an authority or a state is called a crime. However, the word 'crime' does not have a universally agreed upon definition in modern law although common definitions have been provided for various specific purposes. A publically aware point of view is that crime is a category created by law, that something is declared as a crime only if it falls under a predefined law. According to another proposed definition, a crime or a criminal offence can not only be against an individual but can also be charged against a community, society or a state. Such actions are punishable by the order of constitution.

Cutting crime rates has a great impact on the wellbeing of individuals and the society and it helps a great deal in improving the overall quality of life. Prediction of criminal activities ahead of time can help in this regard but it involves a huge effort in terms of resources. The provision of resources involved in creating an effective crime prediction system is also a considerable factor. In the early stages of the work, we wanted to base this study to focus solely on crime data analytics and prediction systems in Pakistan. But, due to insufficiency of sample data and lack of prediction systems to study in country, we decided to build around the research work on Chicago Crime dataset and in future utilizing the knowledge and expertise to do further research work in Pakistan thus bringing in improvements in the structure in place.

In Pakistan, there is a lack of open source criminal data repositories but multiple government agencies present the summary statistics from week to week on the web portal. A brief overview of the portal is as follows.

For 2019, Crimes increased by 2% in Lahore compared to 2018. The overall count of reported crimes cited by police was roughly around 84,000. This includes all kinds of crimes i.e. Burglary, Robbery, Dacoity, Vehicle Theft and Car snatching. Table 1 shows summarizes property crimes of all kinds for 2018 and 2019. These summary statistics can be accessed on the Punjab Police portal.

**Table 1: Property Crimes in Lahore City**

Property Crimes- Lahore City Figures			
	2018	2019	% change
Total Property Crime	29,301	29,900	2.04%
Burglary	3,429	3,230	-5.57%
Robbery	3,099	3,115	0.51%
Dacoity (Banditry) (robbery by more than three armed people)	57	60	5.26%
Vehicle Theft	4,252	4,300	1.12%
Motor Bike Snatching (jacking)	289	300	3.80%
Car Snatching (jacking)	12	10	-16.66
Motor Bike Theft	3,618	3,700	2.26%

As for our study it is important to get a row by row data for each crime and a set of features associated with the occurrence, the summarized data cannot be used for the prediction tasks. In Pakistan, not only the accessibility of data set is a limitation but generally there is a lack of computerized crime reporting systems which other countries like US have. Across various countries, these reporting systems present a great importance and act as a backbone of crime detection and reporting and as a doorway to gathering conclusive insights for the criminal justice system. The purpose of designing these systems is to cut crime, save time and optimize resources by using data driven insights in decision making and to being proactive by knowing what's coming. Both Unstructured and structured datasets were used by these crime forecasting projects also took advantage from multiple types of dataset. The UI of these systems provide necessary knowledge to the authorities by using surveillance data and incident reports available online.

Public safety systems, efficiency and connectivity can be improved and made smarter by state of the art technologies and capabilities that are now available. Public safety officials can now perform examinations to anticipate and then take action to prevent incidents from happening instead of just

reacting to crimes and emergencies using these new technologies and capabilities. Intelligence can be applied to a data mass by these smarter public safety systems after the collection from different processes. This can be made possible as the intelligence applied to this data can notice patterns of incidents and generate new insights, so that authorities can operate in real time and make well-informed decisions.

Reviewing such prediction systems, studying the algorithms used and making use of the knowledge for the improvements in techniques is the main objective of our study.

There are multiple modules involved in composing a framework for crime prediction and analytics. The first stage in developing this framework is gathering and maintaining the domain data. The crime prediction systems always rely on latest data because occurrence of crime involves a lot of environment factors like weather or education standards of the locality, the combination of which may or may not result in the occurrence of particular crime. There can also be any change in rules, regulations and penalties imposed per the type of crime, therefore, latest samples are always needed by the statistical models so that the performance of deployed models stays reliable at runtime. To address this need many law enforcement agencies have built information management systems which feed information related to criminal activities on daily or weekly basis. The type of records stored by these information systems depend on the usage and data maintenance standards applied by the agencies.

The goal of crime prediction system is to provide law enforcement agencies access to the insights and predict the future occurrence of a crime. The prediction can be based on crime type, location or time. For this research we have focused on tools and techniques use to accurately predict the type of crime that is going to happen given the spatial and temporal characteristics. From the evolution of data mining and big data techniques, more and more departments are using data driven algorithms to forecast crime. The primary objective of these systems is to improve accuracy and effectiveness of deployment of police force in order to efficiently prevent crimes from happening. One working example of these system is Operation Laser. It is deployed by Los Angeles police department. It uses crime, arrest and field data to determine where violent crimes are likely to take place and who will perpetrate them before they actually happen. Another popular tool focused on this area is predpoll. Today more than 60 departments across US use it to forecast crime like burglaries and car break-ins. Officers get a map labelled with 10 to 20 hotspots they are encouraged to visit during their shifts. The more time the police patrols spend in that area, the more likely they are to deter crime. Predpoll forecasts crime base on patterns from the last several years. It analyzes three elements of crime data

the crime type, location and the time and runs them through machine learning algorithms to provide 500 by 500 foot hotspots that officers should keep an eye on. Another popular tool in this regard is developed by IBM. The IDM Crime Management System can integrate and analyze crime management data regardless of the source. The system is specifically designed to support end-to-end law enforcement process of preventing, detecting, responding to and solving crime. The crime management system can help police agencies prevent crime by collecting, integrating and analyzing crime data from multiple sources to provide a picture of when and where crime is likely to happen. When agencies know where the problems are in real-time, they can create dynamic police patrol schedules deploying officers to area where they are most needed. This capability can reduce response time by up to 20 percent. With the increased information at the finger tips, responding officers can respond more quickly and safely to the incidents in the field. After an incident, analysts can use tool included in the crime management center to identify persons of interest, associations, commodity flows and complex networks to build a single common intelligence picture. This capability can speed investigation and reduce manpower requirements.

All of the major crime management systems start with gathering data from multiple sources, organizing data in data cleaning and data transforming activities and then feeding the data into one or more machine learning algorithms to predict the outcomes. As the prediction can affect the community in multiple ways, the use of the type of machine learning tools becomes the key in overall performance of the system. In our study we made use of Chicago Crime Dataset and implemented three algorithms namely Naïve Bayes Classifier, Decision Tree and Random Forest Algorithm. We have also incorporated multiple techniques like clustering, outlier removal and discussed the class imbalance to improve the efficiency of these algorithms.

## 1.2 Key Contributions

The research work highlights the limitation of crime reporting and prediction systems in Pakistan by using multiple statistics. There exists multiple crime reporting systems in Pakistan, especially for the province of Punjab, but proper maintenance of these projects is required so that the crime records are inserting regularly as in other crime reporting system of the work. The study also indicates that there are little to no crime prediction systems in Pakistan because the crime reporting systems are in bad shape. This research work also explains the working of existing crime prediction and reporting systems like Predpoll and IBM crime management system in detail so that similar systems can be implemented in Pakistan.

This study discusses the properties of a crime data set in detail by bringing the limitation in a typical crime prediction dataset in terms of features. The thesis highlights that the more features should be extracted in order to improve the accuracy of predictions. It also provides a clustering technique for the feature extraction. It also suggests the use of multi-modal data to improve so that not just the temporal and features should be considered but data related specifically related to community can also be used. This study also demonstrates that the limitation of feature set is the primary cause of low accuracy of predictions and how can we enlarge the feature set in future. The thesis also introduces the use of logistic regression which can impute missing values in any dataset. If the dataset has many missing values, this can certainly affect the prediction accuracy. Traditionally, While mean values can be imputed in place of missing values for the columns which are continuous in nature, if the column that is missing values is a multinomial variable, the values are excluded from the dataset. In this regard, this study suggests the use of Logistic regression predict the missing values where the subject column is multinomial.

The research work provides the comparison of three machine learning algorithms namely Naïve Bayesian Classifier, Decision Tree and Random Forest Classifier on two datasets. Given the dataset at hand this study discusses pros and cons of the learning algorithms, The research work discusses the problem of class imbalance related to crime categories in a crime dataset and how it affect the accuracy of predictions of the three algorithms.

Finally, the research work provides suggestions to for the application of complete framework in Pakistan.

## 1.3 Organization of Thesis

This first chapter provides the motivation and objective of the study. It emphasizes on the importance of crime reporting and crime prediction systems and also traverses through systems already being using around the globe listing important statistic. It also builds the motivation by briefly discussing the lack of crime such systems in Pakistan. Also, the first section reviews the definition of crime used by law enforcement agencies. In the later part of the chapter we discuss multiple crime prediction systems in place and their operation using different machine learning algorithms. We have also summarized the importance of accuracy of algorithms that are a part of crime prediction and reporting systems. The second chapter focuses on the literature review by spanning through a number of research works which were analyzed and a few of them implemented for reuse of methodology, reproducing results from the past and delivering improvements. The 3rd chapter explains the methodology followed for the implementation of knowledge gathered through the literature reviews. The three algorithms Naïve Bayes, Decision Tree and Random Forest Classifier are demonstrated in detail. In addition, the data extraction, transformation and prediction tasks are explained to the reader with the help of data visualizations. This chapter strictly follows the flowchart of methodology so that the algorithms can be implemented afterwards giving forth equivalent results. The 4th chapter discusses the results which are obtained in multiple iterations of improvements in methodology. Last section is dedicated to the references of research articles, journals and others links which are used in this research work and to whom user of this document can refer to while implementing the discussed methodology.



## 2. Literature Review

### 2.1 Research Works

In order to proceed with the research we did a literature survey of a number of international research papers. The research papers can be broadly categorized in two types: Research concerning the data analytics or insights given the crime activity data, Research concerning the use of one or more machine learning algorithms in order to predict the crime type. While the former focuses on techniques which can help in ad-hoc analysis on the historical data, the later strictly focuses on providing the accurate prediction results at real time. Research concerning Machine Learning tools and applications help us understand the application of machine learning in Crime prediction systems.

This section demonstrates in detail the use of clustering approaches for making new features. Clustering of structured data can be used to create new features. There are number of algorithms to achieve this. The most popular algorithm for clustering is k-means clustering. The need for clustering arises for grouping data into segments in order to find high level insights in the data. We used k-means clustering to create a new feature called cluster. There are multiple spatial features present in the Chicago Crime Dataset, k-means used these as a basis to find clusters in data. Catlett et al. (2018) represents an approach which uses auto-regressive models and spatial analysis to detect potential urban regions with high crime rate. The results are reliable in in forecasting crime trends. The research work used spatial clustering in which each cluster represents a region with high crime rate. The idea is to mathematically group similar crime rate regions in same clusters. The algorithm used for this task is DBSCAN which is widely used to find arbitrary shaped clustered in the considered data on the bases of estimated density distribution. These clusters can be used as features to input in a prediction model.

Yang et al (2019) points out that due to fast convergence and simplicity, k-means is a widely used algorithm. One drawback is that the k-value needs to be specified in advance. The K-value affects the speed of convergence and quality of results. For solving this problem, the study suggests the use of four algorithms namely and Canopy, Silhouette Coefficient, Gap Statistic and Elbow Method. Verifications of results are done using easily available Iris dataset. The study also mentions the pros

and cons of using each k-value selection method and also visually explains the clusters obtained after the run.

Oylade et al (2010) suggests that academic progress of students can be monitored using k-means algorithm at the back-end. In the community of higher learning, accurately measuring the academic performance of students is an extremely difficult task. It using k-means to divide students' performances into multiple clusters then uses standard set of statistical algorithms to arrange the clusters. In summary, the k-means model is combined with another deterministic model. A dataset of Nigerian private institution is used for this task.

Haiyan et al (2012).An improved K-means algorithm is proposed by combining the traditional k-means with largest minimum distance algorithm. The study claims that this combinations of algorithms provides more accurate results by making up for the shortcomings of k-means clustering algorithm. The study explains the two disadvantages of k-means algorithm that it is hard for k-means to determine the initial focal point and its susceptible to be trapped in local minima. The above combination not only keeps improves the accuracy of traditional k-means but also increases the speed of convergence. The combination of two algorithms does it by improving the choice of initial focal point. Thus, providing cluster stability and precision. This method can be proven more accurate in the conditions in which we have data which is completely randomly distributed. The research uses congregation data. Using the degree of cluster, the congregation data is classified into five classes. Obvious differences between each class are represented in the result output.

Kanungo et al (2020). In this study filtering algorithm, also called Lloyd's k-means clustering algorithm is implemented more efficiently. Kd-tree is the only data structure used to implement this algorithm. Practical efficiently is established in two ways. Firstly, the data sensitive analysis for algorithm's running time is presented. The results show that the algorithm performs better when there is more separation between clusters. Secondly, empirical studies are also presented using both real world datasets and manually generated dataset. It's applications range from image segmentation to data compression. Given the data points, it requires building kd-tree just for once which makes the algorithm mote time and space efficient. The Kd-tree does not need to be computed at every level of the execution which is the main cause of efficiency of this algorithm. The algorithm could be applied to many variants of largest minimum distance algorithm because it only varies in how it computes nearest centers.

Raval et el (2016) The goal of this research work is to present techniques for brining improvements in determining the initial centroid while clustering. The presented technique is the enhanced version

of k-means in terms of time complexity and accuracy. The technique presented, still needs improvement in determining the initial number of clusters. The algorithm performs much better than k-means with a time complexity of  $O(n)$ .

After we have a required set of features there is a need to dive deep in the machine algorithms to find out which techniques are better for the problem of crime prediction

Patel et al (2012).The research tries to extract a kind of “structure” from the data points. It is a classification algorithm. This study uses Decision tree which is a widely used supervised learning technique. The study also provides performance and results analysis of the decision tree. Data classification with decision tree is very visual and easy as compared to other methods. Because of the pictorial view, the classification and categorization task becomes more useful and easy on the eye. It also reformulates the decision tree algorithm and compares its performance with traditional algorithm. It also compared the classification results of decision tree with k-means. The results shown suggest that the performance of decision tree is better than k-means in classification tasks. Here the dataset at hand is multimodal in nature.

S. Aldossar et al. (2020) aims to build a machine learning model for crime category prediction. Decision Tree and Naïve Bayesian Classifier are applied on the Chicago Crime Data set extracted from Chicago Police Department’s CLEAR. The study has selected the top 9 features from the dataset. The comparison of these two algorithms shows that Decision Tree has performed better than Naïve Bayes Classifier. The methodology uses 70% of data set for training and 30% for validation. The original class labels were also combined in to 4 main categories for better prediction. For example Burglary, Motorcycle theft and robbery is grouped in to one class i.e. Theft. This increases the accuracy of prediction but reduces the overall usefulness of the prediction. The prediction accuracy reached 91.59% for Decision Tree and 83.40% for Naïve Bayes Classifier. The study suggests to use better feature selection methods to further improve the accuracy.

Iqbal et al. (2013) also presents a comparison between two classification algorithms namely Naïve Bayesian Classifier and Decision Tree Algorithm. It uses data from multiple sources in addition to Chicago Crime Data to perform the classification. During the pre-processing step, multiple crime categories are clustered in to three main categories to extract a new target variable which is composed of three nominal values namely High, Medium and Low. The experiment was performed on WEKA using 10-fold-cross-validation. The results show that Decision Tree with the accuracy of 83.95% outperforms Naïve Bayes Classifier 70.81%. This study also suggests to use better methods for feature selection to improve the accuracy further.

McClendon et al. (2015) uses WEKA, an open source machine learning software, to conduct a comparative study of machine algorithms on a crime data sets from California-Irvine repository and data of the state of Mississippi provided by neighborhoodscout.com. Methodology involves implementing Linear Regression, Additive regression and Decision stump algorithms using common set of features. The study concludes that the linear regression algorithm performed the best among selected algorithms. Linear regression model handled the randomness for this dataset very accurately. The accuracy measures used in this study are Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error and Root Relative Squared Error.

Ahishakiye et al. (2017) uses a real and authentic dataset. The dataset is acquired from UCI machine learning repository webpage with a title of "Crime and Communities". This dataset contains a total of 128 attributes and 1994 examples with all numeric and normalized features. Attribute selection is done manually by human understanding of dataset. This reduced 128 features to only 12. The paper does not suggest any method of imputing the missing values. Waikato Environment for Knowledge analysis (WEKA) Tool Kit is used to train Decision tree (J48 Classifier) is used to predict the crime type with an accuracy of 94.25%. The performance measures used are Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error and Root Relative Squared Error. The study favors the used of Decision Tree Algorithm on crime datasets because of better accuracy of prediction and little time execute in comparison to other datasets.

Sebalj et al (2016). The study uses machine learning algorithms to classify students in to two categories. It also aims to find out meaningful variables in the dataset. The dataset is composed of student names and their performance of the first academic year. The previous grades of the students assigned by teachers are also considered as features. The algorithm with high accuracy was deployed for testing in real time at the institution. Decision tree algorithm provided the highest accuracy in this classification task i.e. 79%.

Shahbazi et al. (2019) Classification models can be used in collections industry. The customers are rated on the base of previous bank loans. The problem is essentially of classification in nature. The customers can be divided on the basis of credit rating and then further loans can be given to customers paying on time and collected from the customers which do not get cured on time. Thus, the credit worthiness of the customers can be measured on the basis of classes. Data mining can be used to extract multiple patterns in the huge banking data sets. The study uses Decision tree algorithm to determine the credit rating of the customers. The results show that model ranks the customers very accurately. 89% accuracy was reached by using C5 decision tree algorithm.

Ali et al. (2012) presents a detailed comparison between two machine learning algorithms i.e. Decision tree classifier and Random Forest classifier. The study used 20 varying datasets from UCI data repository to achieve this. The problem addressed is a classification problem. The metrics used for measuring the classification accuracy of the algorithms include precision, recall, F-score, sensitivity and specificity. In the light of the results, the pros and cons of using type of datasets are also discussed. The study concludes that Decision tree is handy with small datasets while Random Forest gives better results using large datasets. For Random Forest Classifier, when the number of instances increased from 286 to 700, the accuracy improved from 69.23% to 96.13%. Using the same number of features but more training examples, the accuracy of Random Forest classifier is significantly improved.

## 2.2 Literature Review Summary

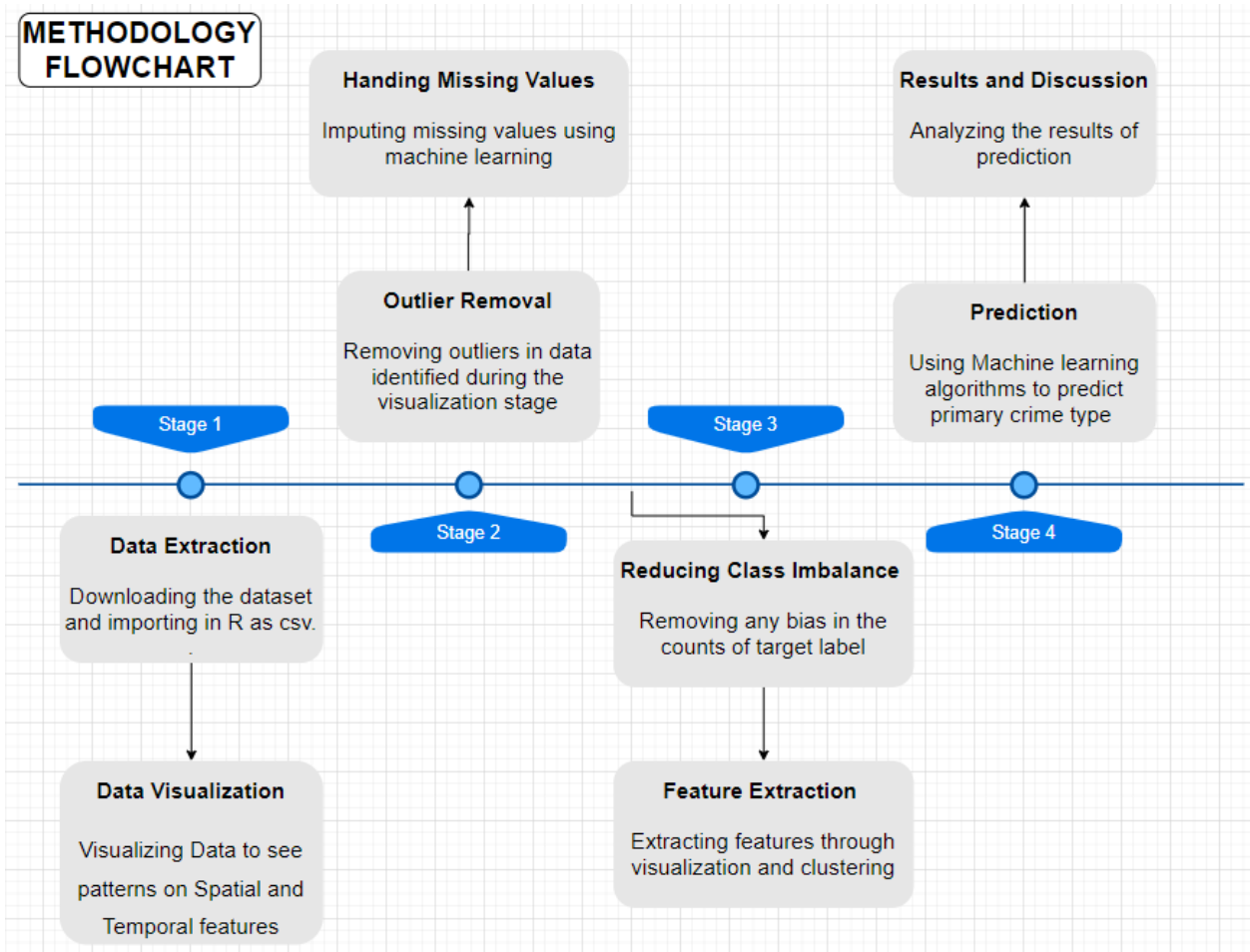
The above literature proves that given the dataset at hand feature engineering and data transformation is a very important part of the framework for the model to perform with high accuracy and better predictions. It suggests using k-means clustering because it is very efficient in terms of computation resource allocation. For K-means to determine the clusters there is a need to feed optimal number of required clusters. Brute force algorithm can be used to determine the optimal number of clusters. Brute force algorithm segments the data in to multiple sets multiple times and determines the optimal number of clusters on the basis of variance in each cluster compared to the variance of the dataset.

After we have the required set of features, a good selection of machine learning algorithms is necessary. Naïve Bayes Classifier is an easy and fast to predict algorithm. It also performs well in multiclass classification problems. When assumption of independence holds, a Naïve Bayes Classifier performs better compares to other models like logistic regression and you need less training data. Given the feature set at hand, Naïve bayes Classifier also handles categorical features very well. From the above literature we also get an idea that the independence assumption of Naïve Bayes Classifier can also effect the model badly because it is impossible for us to get all the features in the dataset as independent features. The second algorithm discussed in the literature review is Decision Tree Classifier, Compared to other algorithms decision trees requires less effort for data preparation during pre-processing and they do not require normalization and scaling of data. The Decision tree method is also very intuitive and easy to explain because it can visualized as a tree. From the research work on decision tree we came to know that a small change in the data can largely impact the decision tree, to cope with this problem, it is necessary to tune the decision tree so that it is neither over-fitted nor much generalized. The other disadvantage of using decision tree is the time complexity of this algorithm. Due to the evaluation of entropy function or information gain at every possible value of a feature, each value treated as a category and evaluation is done using the brute-force way, thus, taking more time than other algorithms. From the research works mentioned above another way to improve the decision tree is to use Random Forest Classifier which is essentially employs the use of multiple decision trees to find a good balance between generalization and over-fitting. Due to use of multiple trees the computational and space complexity of Random Forest Classifiers is greater than Decision trees.

### 3. Methodology

This chapter explains the methods followed for carefully analyzing data and implementation of algorithms for the prediction tasks. The user can follow the methodology flow chart which provides an aerial perspective of the implementation of multiple algorithms to achieve the desired results. The section starts with data extraction process and enlists all the columns of data and the model data dictionary. The model data dictionary helps to form business logic of the features which are used in visualizations and prediction section. Summary statistics of the dataset listing the columns imported, unique values and types in the dataset are also represented in a form of a table. The next section is reserved for data visualization. In this section we have analyzed crime counts by using multiple pivots. In data visualization, the crime counts are listed against multiple spatial temporal features. The section of outlier detection explains the detection of outliers on the basis of geographic coordinates. As the primary crime type is a multinomial variable, there is an implication of class imbalance in treating primary crime type as a target variable for prediction. This is problem is discussed in class imbalance section and the treatment is also proposed in the light of multiple research articles. After we have the dataset it is important to gather best features which can help improve the accuracy of prediction. The feature extraction chapter covers making new features using visualization and through k-means clustering algorithm. The optimization of k-means clustering algorithm is also discussed in detail using mathematical model equations. Last two sections of methodology present the actual application of three machine learning algorithms namely Naïve Bayes Classifier, Decision Tree Classifier and Random forest classifier. The pros and cons of using these three algorithms, given the data set at hand, are also discussed in the results and discussion section.

**Figure 1: Flow Chart of Methodology**





## 3.1. Data Extraction

For crime classification we have used Chicago Crime Data set available on Chicago Data Portal. The dataset contains reported crimes in the City of Chicago from 2001 to present excluding most recent seven days. The Data is extracted from CLEAR (Citizen Law Enforcement and Reporting) system which is fully owned by Chicago Police Department. Each row in the dataset represents an occurrence of a crime with following properties.

- **ID:** A unique identifier for record. Most of the systems use this column to maintain the database schema so that the records are not duplicated.
- **Case Number:** A key for each case assigned in Chicago Police Department for each incident. This represents RD number which is Records Division Number.
- **Date:** This is the timestamp of the occurrence of Crime.
- **Block:** In order to protect the privacy of the people and places influenced by the crime, the use of exact addresses has been avoided while entering the records. Instead, the exact addresses have been redacted partially up to the block level.
- **IUCR:** This is code used by Illinois Uniform Crime Reporting. This has a one-to-one mapping with the primary crime type and description of the crime.
- **Primary Type:** It is the category of crime which will act as a target label for classification problem discussed in this study.
- **Description:** A description of crime type. This has one-to-one mapping which the primary type and IUCR.
- **Arrest:** This represents if the arrest was made afterwards.
- **Domestic:** Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.
- **Beat:** Represents the smallest geographical area where the incident occurred. Each beat has a dedicated Police Patrol Car. Multiple beats are joined to create a police sector which make up a district.
- **District:** A group of police sectors form a district.
- **Ward:** The City Council district where the incident occurred.
- **Community Area:** This represents community area where the incident occurred. Chicago has 77 community areas.
- **FBI Code:** Represents the crime classification as Indicated by FBI's National Incident Based Reporting System (NIBRS).

- **X Coordinate** – This represent the x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. To fulfill the privacy standards, this location is shifted from the actual location through partial redaction but comes under the same block.
- **Y Coordinate** – This represent the y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. To fulfill the privacy standards, this location is shifted from the actual location through partial redaction but comes under the same block.
- **Year** - Year In which the incident occurred.
- **Updated On** – This represents the Timestamp when the record was updated in the system.
- **Latitude** - The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but comes under 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 comes under the same block.
- **Location:** Contains the location at which the incident occurred. The location is kept in such a format that it supports creation of maps and other geographical operations present on Chicago Data Portal.

The City of Chicago's open data portal lets us find city data and facts about the neighborhood. It provides numerous create maps and graphs about the city, and allows to freely download the data any type of analysis. The support of the portal maintains the website and updates the data sets on regular basis. Some of the data sets are even updated on a daily basis. The Crimes dataset of the city of Chicago is the subject of this study. There are many different types of data sets available at the portal. Following snapshot represents a few.

**Figure 2: Datasets on Chicago Data Portal**



The dataset shows reports of incidents of felony (exceptions being murders with existing for each victim) occurring in the City of Chicago from 2001 to present date, without the inclusion of the past seven days. Data is taken from Chicago PD's CLEAR (Citizen Law Enforcement Analysis and Reporting) systems. For privacy protection purposes, the addresses of the victims are only shown at block level and specific locations are unidentifiable. The bases of the crimes might be upon preliminary information given to the Police Department by reporting parties yet to be verified. Even

though there is always the possibility of mechanical or human error, the introductory crime classifications may be altered at a later date based upon supplementary information from investigation. Hence, the Chicago PD does not insure (either explicitly or implicitly) the accuracy, completeness, timeliness, or correct sequencing of the data and therefore the data should not be used for comparison later in time.

The data is extracted from the portal in .csv format and the imported into R notebook for further analysis. Following table represents the columns present in the data set and the raw summary statistics.

**Table 2: Characteristics of Columns in Dataset**

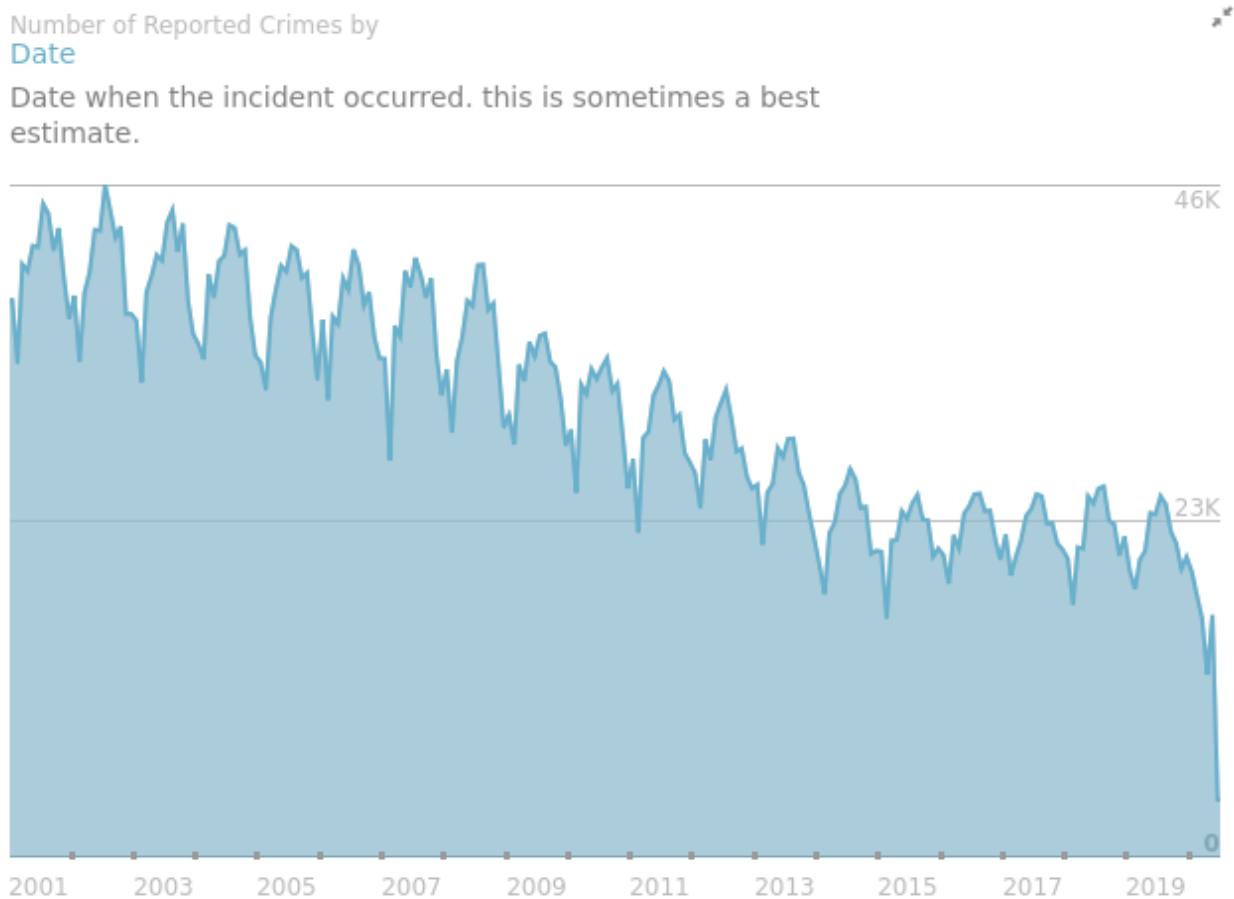
<b>ColumnName</b>	<b>Distinct Values</b>	<b>Type</b>	<b>Missing Values</b>
<b>ID</b>	5717422	Categorical	0
<b>Case.Number</b>	5716995	Categorical	0
<b>Date</b>	2359427	Timestamp	0
<b>Block</b>	60549	Categorical	0
<b>IUCR</b>	402	Categorical	0
<b>Primary.Type</b>	37	Categorical	0
<b>Description</b>	512	Categorical	0
<b>Location.Description</b>	213	Categorical	0
<b>Arrest</b>	3	Categorical	0
<b>Domestic</b>	3	Categorical	0
<b>Beat</b>	305	Categorical	5
<b>District</b>	25	Categorical	52
<b>Ward</b>	51	Categorical	614759
<b>Community.Area</b>	79	Categorical	613211
<b>FBI.Code</b>	27	Categorical	0
<b>X.Coordinate</b>	78132	Categorical	68309
<b>Y.Coordinate</b>	129370	Categorical	68309
<b>Year</b>	21	Categorical	5
<b>Updated.On</b>	3338	Categorical	0
<b>Latitude</b>	823892	Categorical	68309
<b>Longitude</b>	823390	Categorical	68309
<b>Location</b>	824982	Categorical	68309

## 3.2 Data Visualization

### 3.2.1 Number of Reported Crimes by Date

We can see the steady decreasing trend towards the end of the chart. For year 2020, the crime counts got significantly reduced due to corona virus worldwide pandemic.

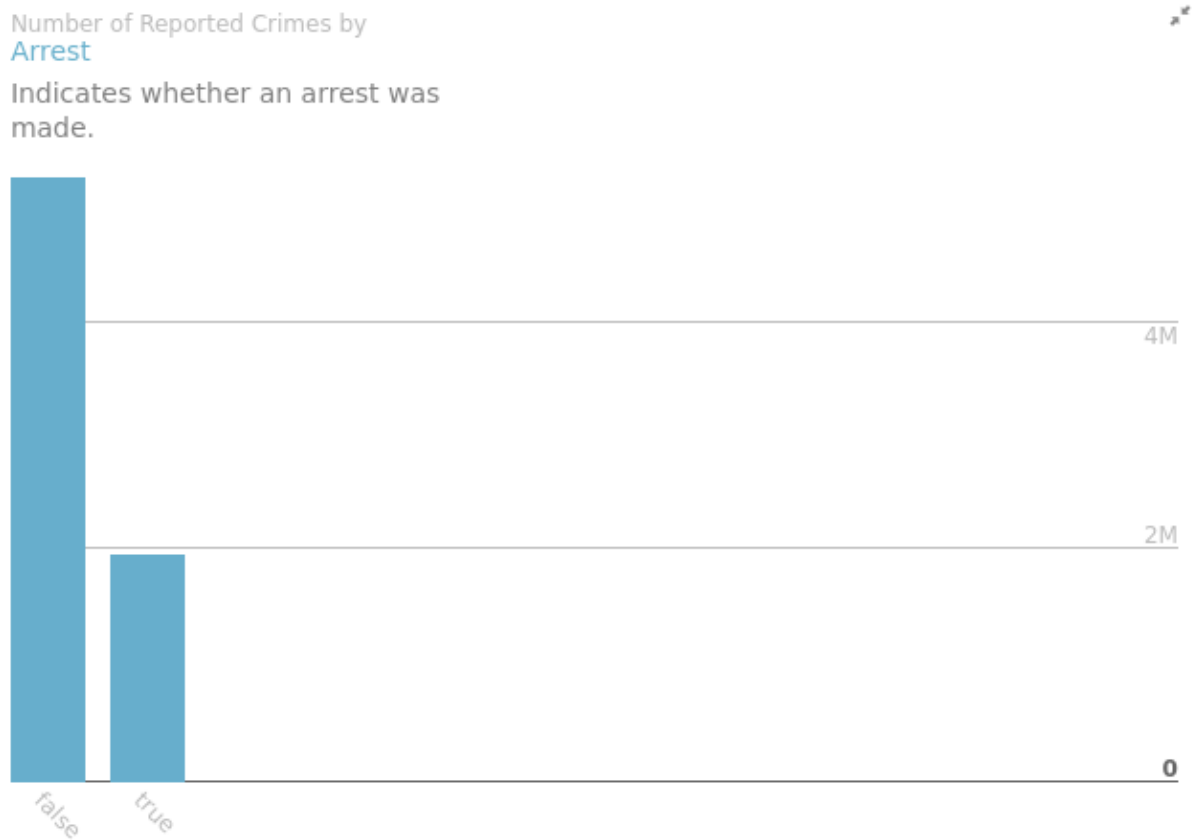
**Figure 3: Number of reported Crimes by Date**



## 3.2.2 Number of Reported Crimes by Arrest

The Arrest column represents if the arrest was made in the result of the crime. We can see that for more than 75% of the cases the arrest is not made.

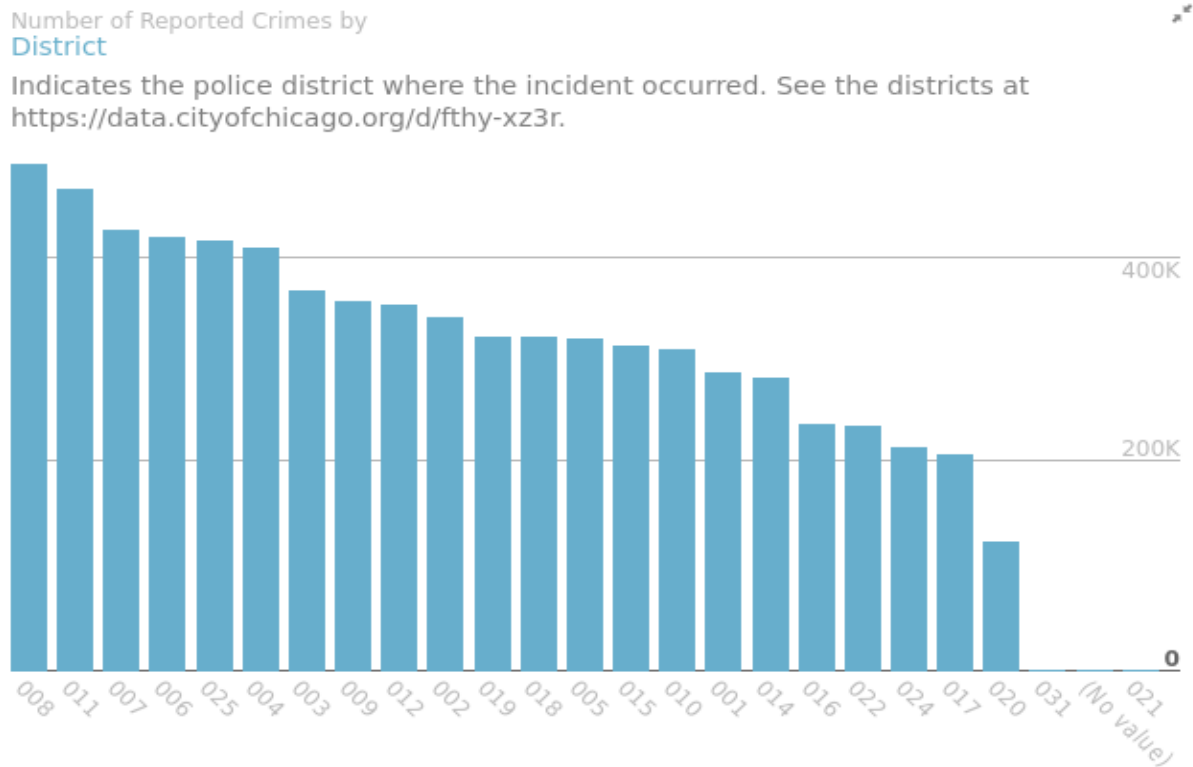
**Figure 4: Number of reported crimes by Arrest**



### 3.2.3 Number of Reported Crimes by District

The area of Chicago is divided into 25 districts which are mentioned in the chart with the crime counts.

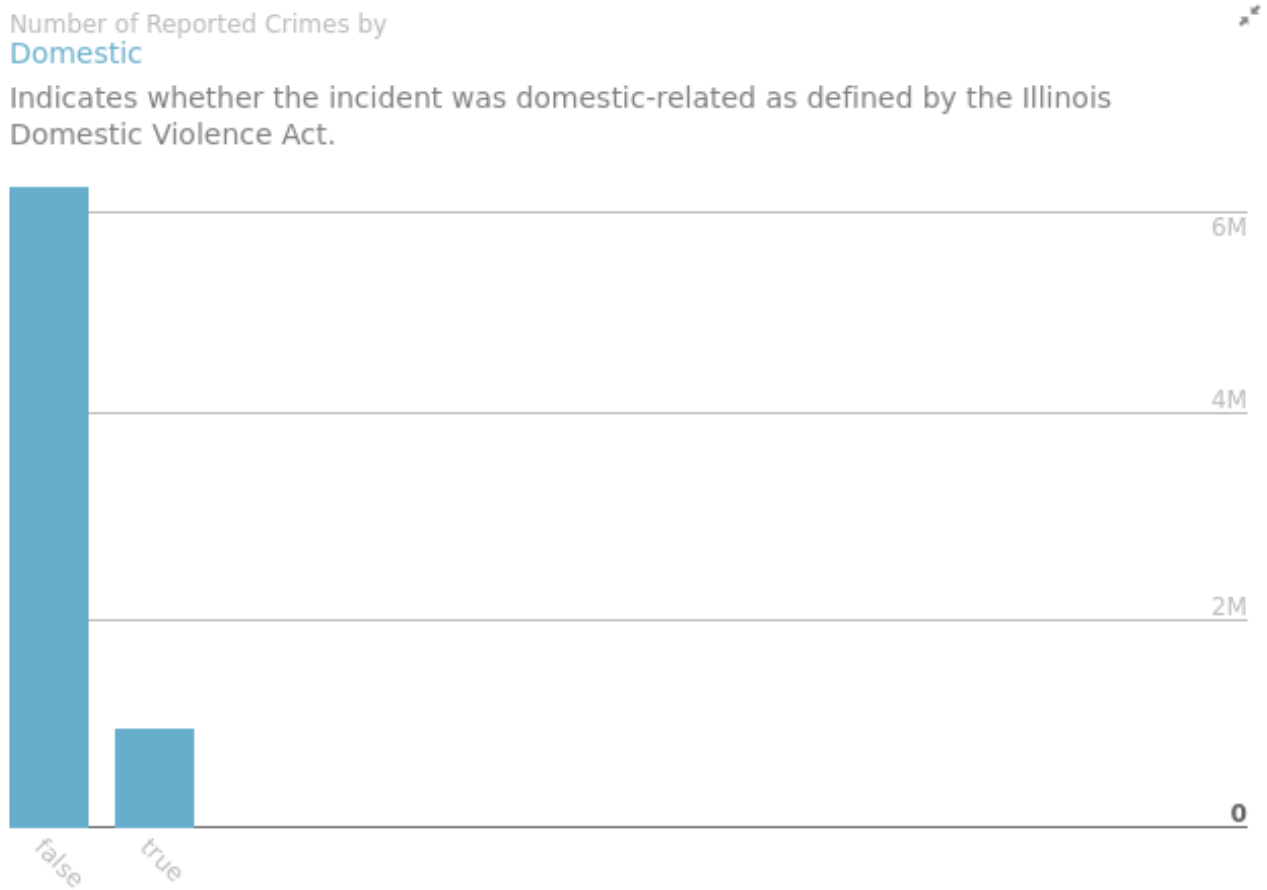
**Figure 5: Number of reported crimes by District**



## 3.2.4 Number of Reported Crimes by Domestic

Domestic is a flag which represents if the crime is domestic or not. As the chart represents, majority of the crimes are domestic in nature.

**Figure 6: Number of reported Crimes Domestic or not**

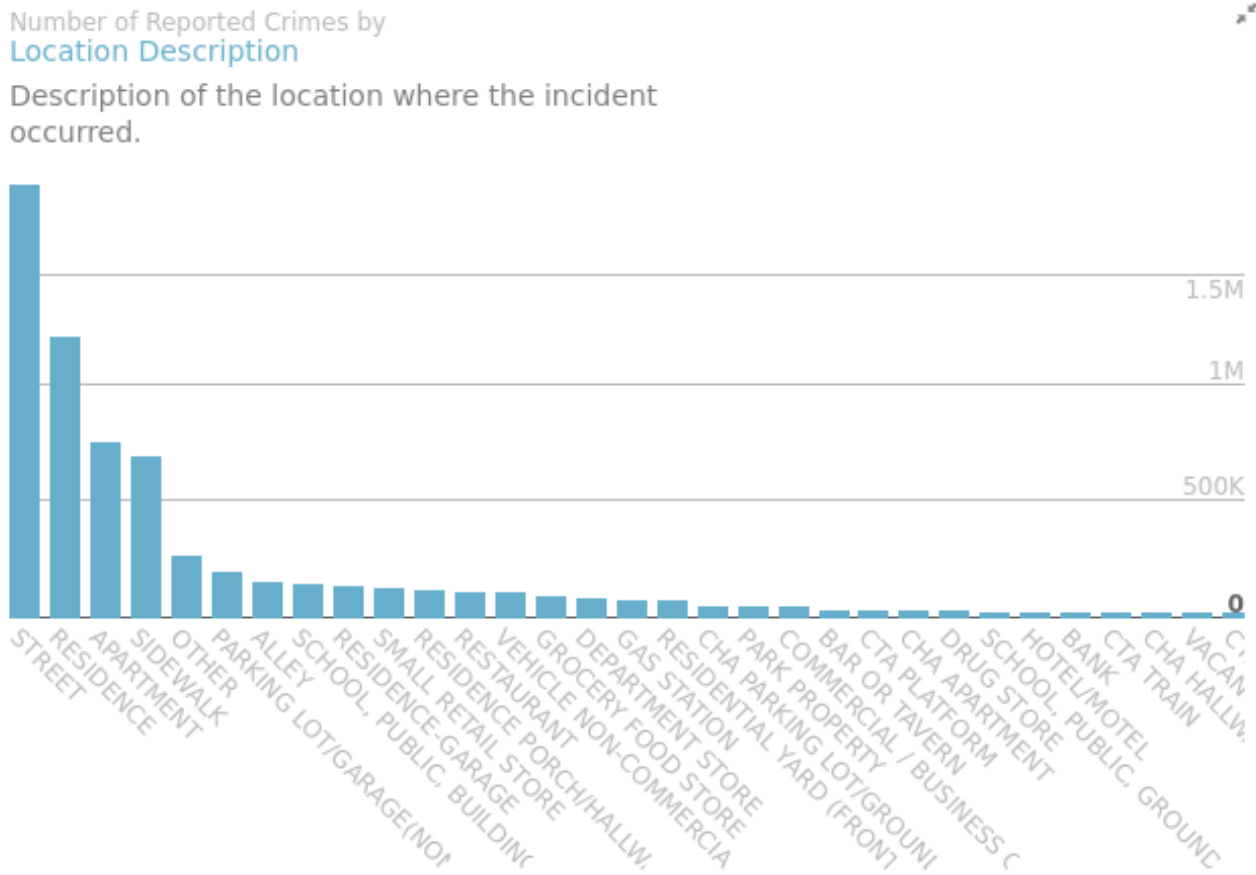




## 3.2.5 Number of Reported Crimes by Location Description

The description column represents the location description of the crime. As we can see majority of the crimes occur on the street, residence, apartment or a sidewalk.

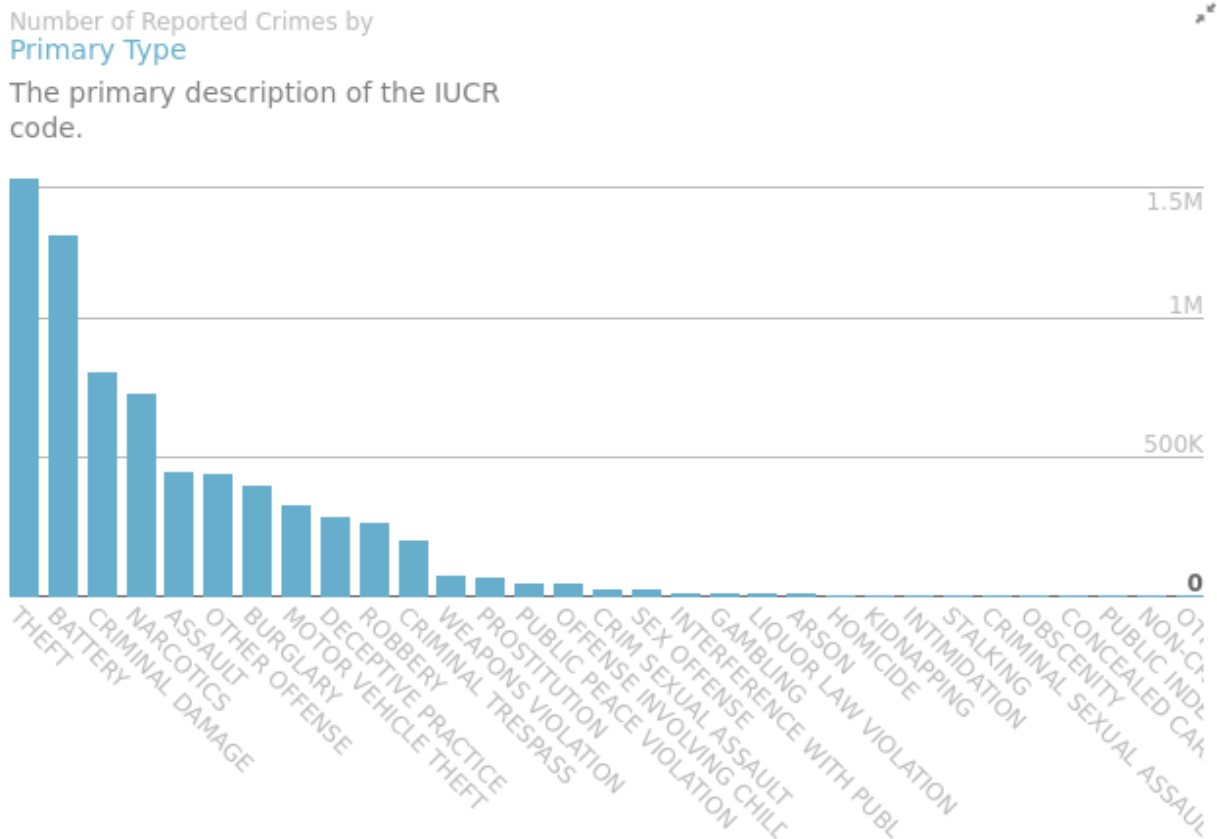
**Figure 7: Number of reported crimes by Location Description**



## 3.2.6 Number of Reported Crimes by Primary Type

Primary.type is the classification of crime in to category. This is the main class label for training and testing during this study. The chart represents the counts of each crime type. As we can see the counts are dominated by Theft, Battery, Criminal Damage and Narcotics. This may induce a class imbalance in the prediction results which we will mention in the algorithm section of this study.

**Figure 8: Number of reported crimes by Crime Type**

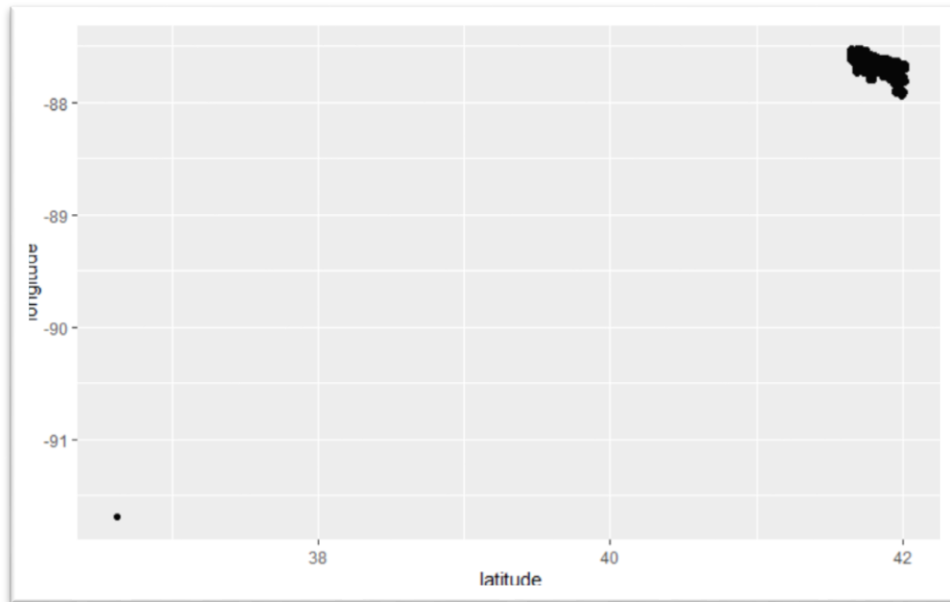


### 3.3 Outlier Removal

All of the columns present the dataset has been tested to find anomalous data or outliers. Two outlier sample points were excluded from the data on the basis after making a scatter plot as below. The scatterplot represents latitude on x-axis and longitude on y-axis. The outlier data points were around the latitude of 36 and longitude of -92. Since the city of Chicago geographical location lies between the latitude of -88 to -87 and longitude of 41.5 to 43.5, the values data points were considered as outliers. The mean values of latitude and longitude were substituted in place of outlier sample points.

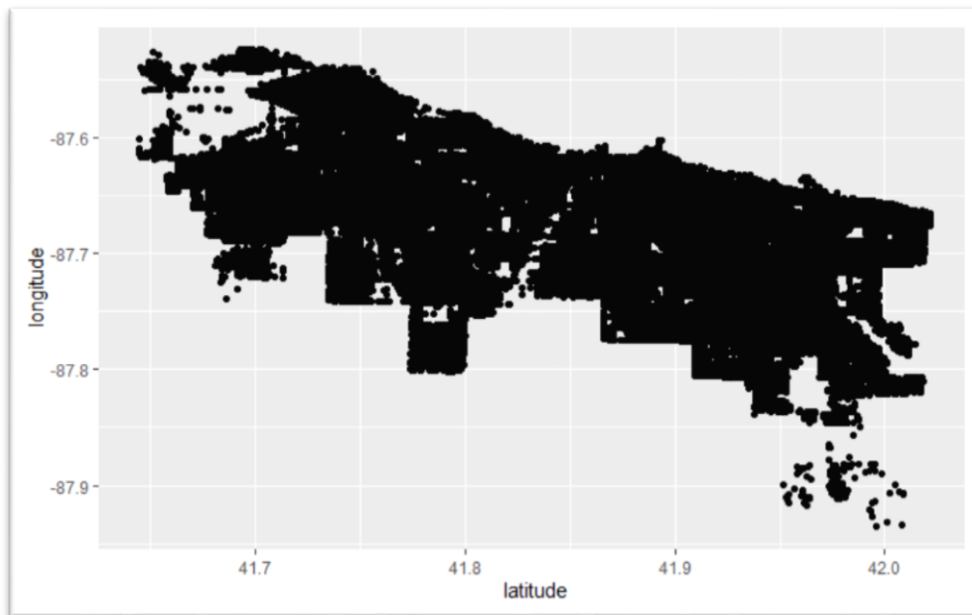
Scatter plot of longitude and Latitude before outlier Extraction is represented in Figure 9.

**Figure 9: Scatterplot of x-y coordinates before outlier removal**



Scatter plot of longitude and Latitude after placing mean values in place of outliers. The scatter plot is now very similar to actual City of Chicago map.

**Figure 10: Scatterplot of x-y coordinates after outlier removal**

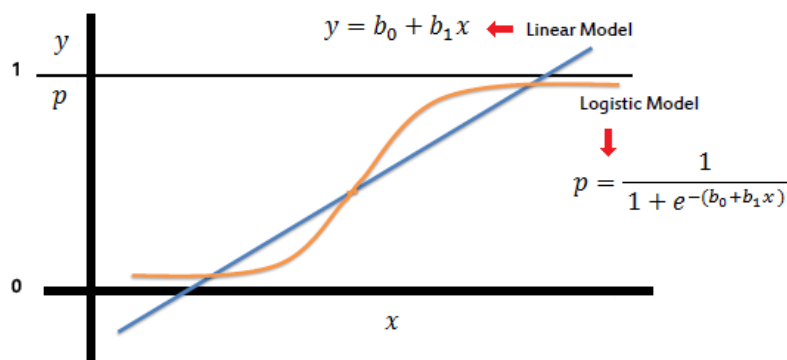


## 3.4 Handling Missing Values

As mentioned in table 3.1. There were thousands of missing value in the data. Since the variables are categorical with many distinct values, district was the only variable picked for the imputation. The District variable had significantly lesser missing values and proves to have much predictive power. The values are predicted using logistic regression. Rest of the columns are used to predict the missing categorical values of the district variable.

The logistic function is a sigmoid function where the inputs are log odds and the output is the probability. R package psych uses a wrapper around logistic regression to output the categorical variable instead of probability. The following diagram represents a simple logistic regression model for binary classification.

**Figure 11: Logistic Regression Model**



**Linear Model:**  $y = b_0 + b_1x$

**Logistic Model :**  $p = \frac{1}{1 + e^{-(b_0 + b_1x)}}$

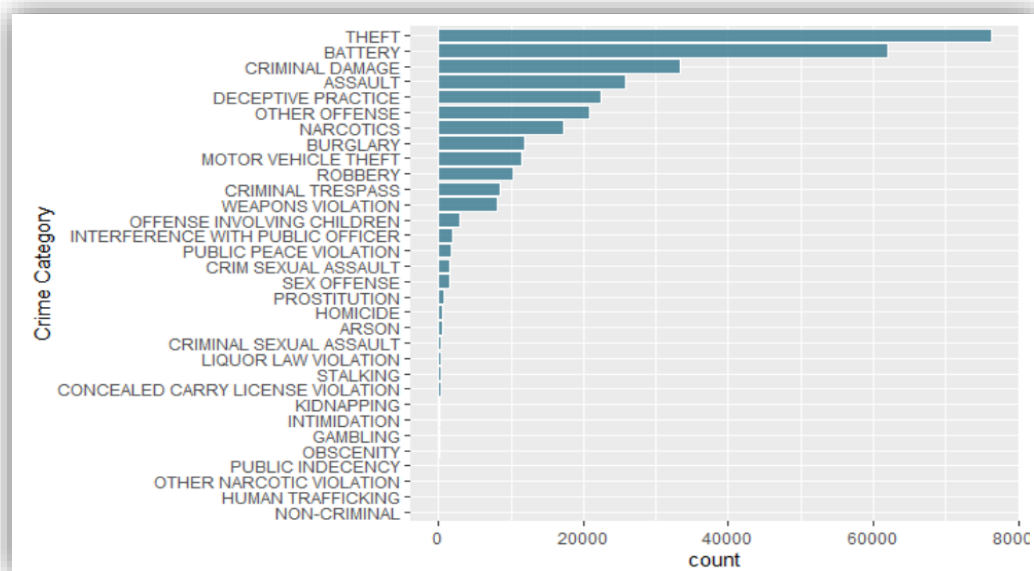
Y represents the target which is district in our case. The regression equation is converted into a sigmoid function which limits the predicted values between 0 and 1 so that the predicted values can act as a probability of each class.

## 3.5 Reduction of Class Imbalance

Class imbalance can reduce the performance of machine learning algorithms like Random Forest, Naïve Bayes and Decision Tree. Stephen et al. (2002) answers some question in this regards. The study explains the relationship between the three factors that are involved i.e. size of training dataset, class imbalance level and complexity of the concept. Moreover, it summarizes the past works done in order to deal with the problem of class imbalance. These methods include re-sampling of dataset and some cost modifying methods. The obtained results on artificial domains are compared to real-world domains. It also investigates why the class imbalance affects not only decision the mentioned three algorithms but also affects Support Vector Machines and Neural Networks.

Represents the counts of distinct crime type which is the target class for the classification problem in our study.

**Figure 12: Counts of Crime categories in sample Dataset**



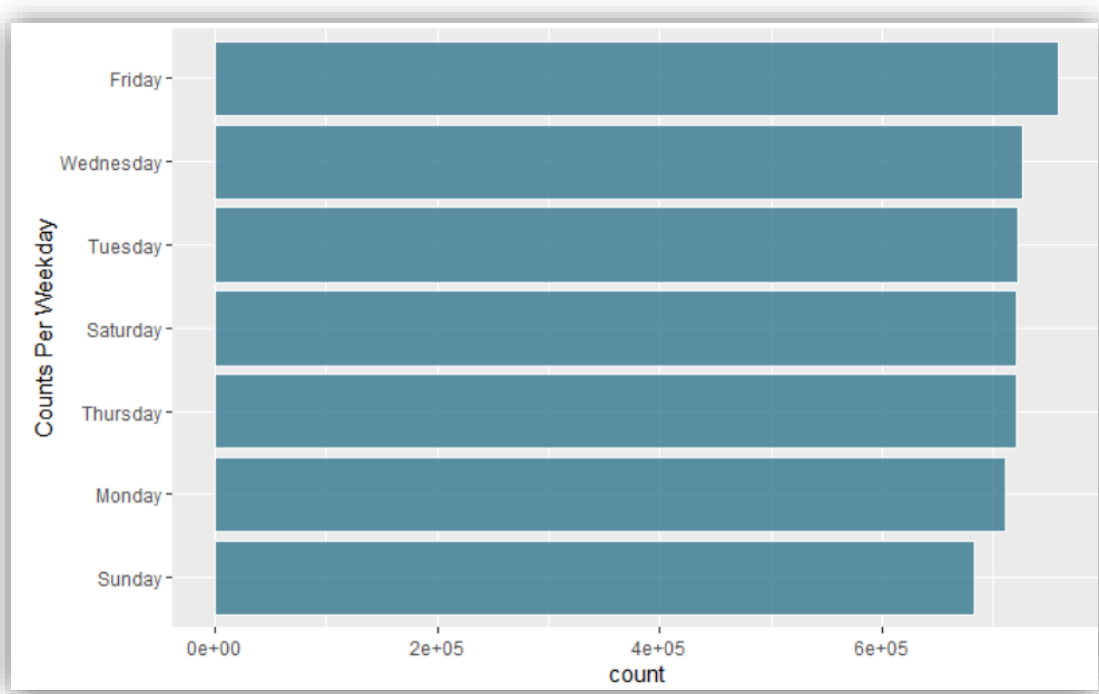
## 3.6 Feature Extraction

The features of the existing data set can be divided into two categories. i.e. Temporal and Spatial. The features that represent time are referred to as temporal features while the features which are a reflection of location or represent spatial coordinates are called spatial features. The property of both of the types in the given dataset is that they are multinomial or categorical in nature from which multiple features can be extracted.

### 3.6.1 Temporal Features

After looking at the distinction between crimes conversion rates on the basis of Day of the week and month, the date timestamp column of the dataset is transformed to extract day of the week and month of every crime and are used as two added features to reach better accuracy of the prediction models. Following graphic represents a distinction between crime rates on multiple days of the week.

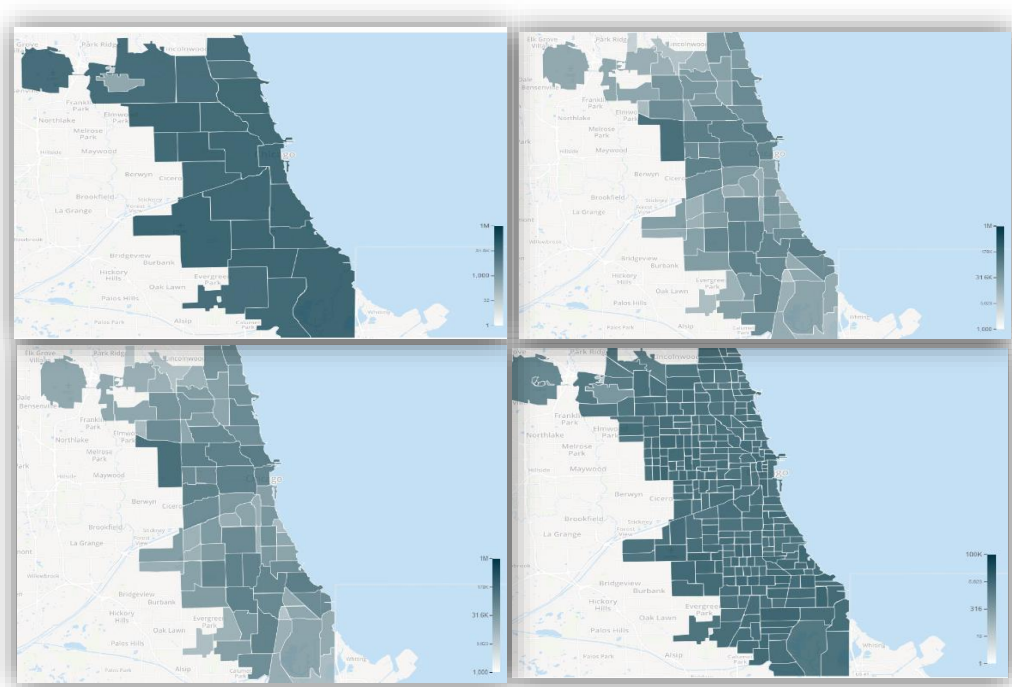
**Figure 13: Crime counts by Day of the Week**



## 3.6.2 Spatial Features

Moving from low to high granularity, the Chicago administration has divided the city into multiple districts, community areas, wards and beats etc. in order to manage the policing and reporting of crimes. These division of the city are readily available in the dataset to be used as features in the prediction process. Following are the graphics which explain division of the city moving from low to higher granularity scale.

**Figure 14: Division by District Beat Community and Ward**





## 3.7 Scaling or standardization

The x-coordinate and y-coordinate present in the dataset can also act as temporal features but due to higher number of distinct values, it is easier to make clusters of locations using k-means clustering.

Rescaling of features to represent Normal Distribution having 0 mean and 1 standard deviation, is a way to normalize dataset.

$$Z = \frac{x - \mu}{\sigma}$$

Above is the formula to calculate the z-score. X is the data point for which the z-score has to be calculated. Sigma represents the standard deviation of all the data points in the sample.

## 3.8 Clustering

First problem is to find the optimal number of clusters which can represent data accurately and then an algorithm which outputs those clusters to be used as features in the model.

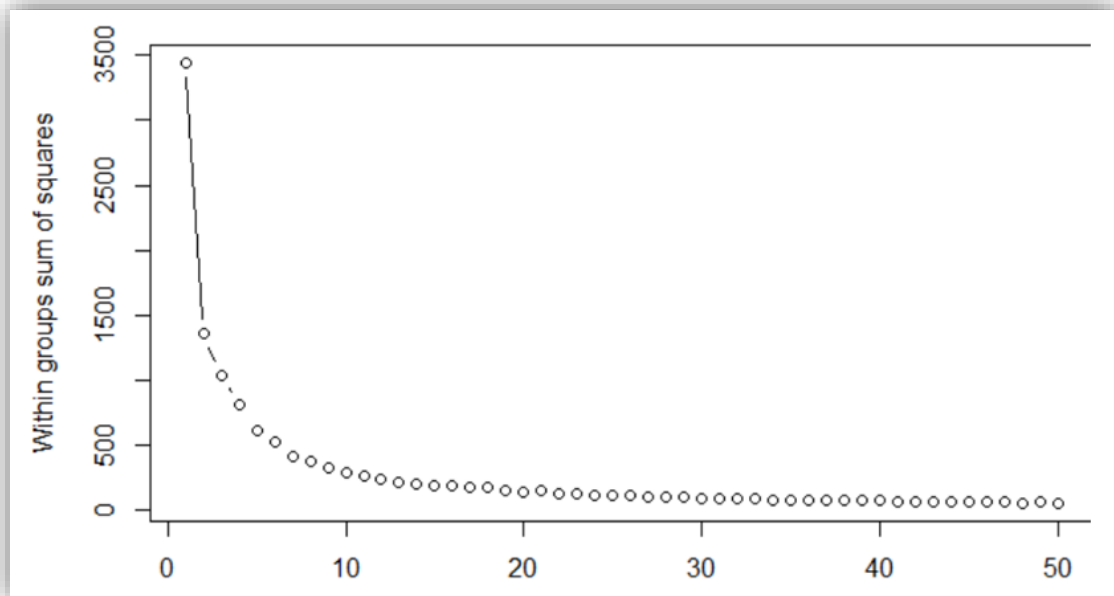
Gap Statistic method is used to determine optimal number of clusters for k-means algorithm. The methodology involves clustering the observed data by varying the number of clusters from  $k = 1$  to  $\max(k)$  and then computing the intra cluster variation or intra-cluster variance.

$$Gap(k) = 1B \sum b = 1B \log(W * kb) - \log(Wk) \quad Gap(k) = 1B \sum b = 1B \log(Wkb *) - \log(Wk).$$

The algorithm is ran multiple times in order to extract the min value of k for which the intra-cluster variance is the lowest.

Following visualization represents a run of optimal cluster measurement algorithm. Due to the limitation of computation power the k-means algorithm and to save time, value of k is chosen to be 14.

**Figure 15: Visualization of k-means**



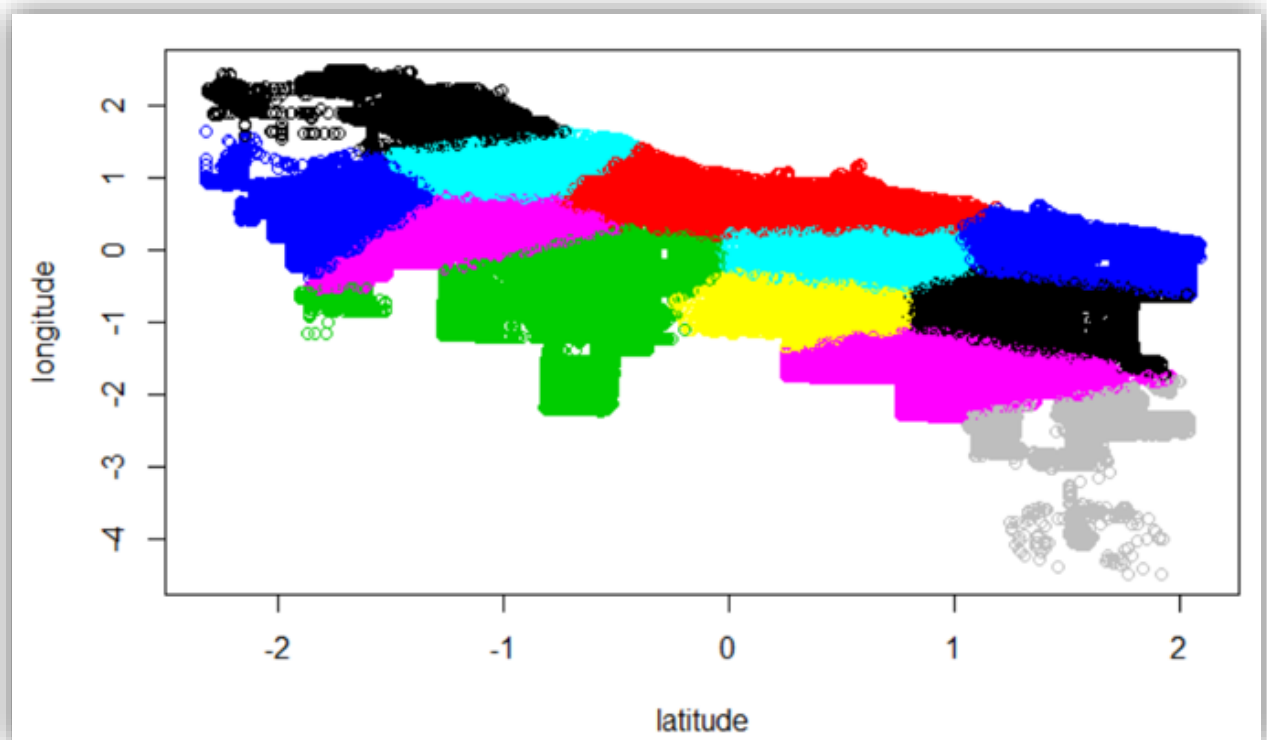
K-Means is a simple algorithm which can give reliable results. It works best when the sample points are separated in a linear way and we have also made sure that the coordinate features are free of outliers. K-means clustering involves minimizing the squared error function which is the objective function while clustering on the basis of coordinates.

$$J = \sum_{i=1}^k \sum_{j \in C_i} \|x_j - v_i\|^2$$

$\|x_i - v_j\|$  represents a Euclidean distance between a point two sample points and centroid of the clusters. It is iterated over 14 clusters in order to output the clusters. K represents the value of 14 which is the optimal number of clusters given by the Gap statistics method.

Following is the visualization of clustering having latitude on x-axis and longitude on y-axis.

**Figure 16: Clusters created by k-means**



Due to the type values chosen at x-axis and y-axis the graph shows quite a resemblance with distribution of districts by Chicago administration but the clusters formed as the result are do not represent district beats or wards at all. The cluster column is added in the data set as a separate feature for prediction tasks.

## 3.9. Machine Learning Algorithms

The following three algorithms are used for the predicting the primary type of crimes.

- 1) Naïve Bayes Classifier
- 2) Decision Tree Classifier
- 3) Random Forest Classifier

The working of the mentioned algorithms is discussed results and the following discussion is documented in the next section.

### 3.9.1 Naïve Bayes Classifier

Naïve Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector representing n independent features:

$$x = (x_1, x_2, x_3, \dots, x_n)$$

Where n represents the total number of features and each individual feature is represented by x

Naïve Bayes classifier assigns probability to each class of the outcome variable:

$$p(C_k|x_1, \dots, x_n)$$

Where C represent the target variable and k represent the number of classes. Naïve Bayes uses the following formula to calculate the probability of each class.

$$p(C_k|x) = \frac{[p(C_k)p(x|C_k)]}{p(x)}$$

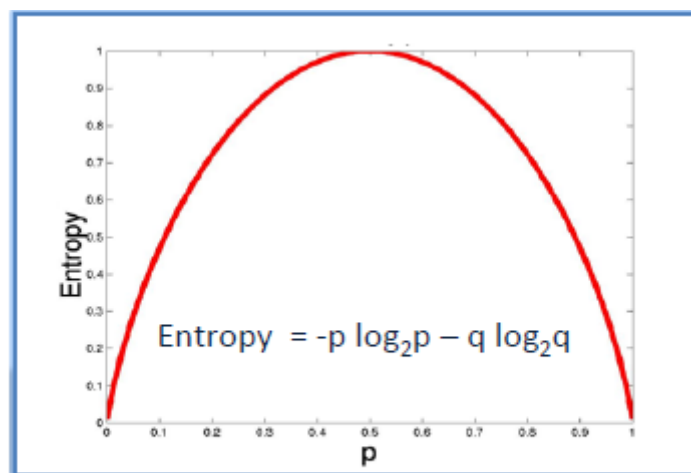
Where left hand side of the equation represents the Posterior probability which is a function of (prior x likelihood)/ evidence.

## 3.9.2 Decision Tree Classifier

Decision tree uses a tree structure where different computations are applied on each node in order to get the separation in the data. Decision tree goes from observations about an item represented by internal or non-leaf nodes to conclusions about the item's target value represented by leaf nodes. Classification trees are the subject of our study since they are effective in prediction models where the class label can take multiple discrete values. The other type is the regression trees where the target label can only take continuous values. Due to more simplicity in application and as a better tool of visualization, decision trees are widely used as a machine learning algorithm.

A decision tree can be used in decision analysis using explicit representation. Decision Analysis making decisions with the help flow chart or a diagram. Decision tree can be trained using the data related to problem and the layout obtained from the trained tree can be used to make informed decisions. ID3 Decision Tree uses entropy or logloss to partition the data into subsets which are homogenous in nature. The entropy indicates the randomness in the subset. The value ranges from 0 to 1. 0 meaning the equal division of the target label and 1 means complete randomness.

**Figure 17: Entropy Explanation for Decision Tree**



$$Entropy = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$$

The information gain is referred to as a decrease in entropy or a decrease in randomness after a dataset is split on a feature. The decision tree is constructed by finding the nodes which can give high information gain.

Following is the formula for information gain:

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

Where T represents a feature and X represents a category of the feature. As the randomness of a category decreases the information gain of the node increases.

## 3.9.3 Random Forest Classifier

Random forest classifier is a variation of decision tree. Highly irregular patterns can be learned by growing deep decision trees, these decision trees over fit the dataset, which means they have high variance but low bias. Random forests acts as a way of normalizing multiple deep decision trees. Random Forest Classifier is trained on different subsets of the same training set in order to reduce the variance. Following formula for Bagging is used to average out predictions from multiple decision trees.

$$f' = \frac{1}{B} \sum_{b=1}^B f_b(x')$$

Where  $x$  represents a subset of the training set using the  $f$  decision tree.  $B$  is the total number of partitions of a training set.



## 4. Results and Discussion

This sections list downs the result after multiple training and test iterations and it also explains multiple reasons tied to results of the three algorithms. The dataset is divided into two parts i.e. Training set and validation set. Due to sufficiency of training examples, validation set is kept separate from training in order to analyze our results without any training bias so that validations are completely out of sample. It also helps to guard against over-fitting. 80% training and 20% validation set were created from the dataset by random sampling of data.

As specified in class imbalance section, firstly the models are trained and tested by taking the samples of top 10 crime types and then the models are trained and tested on the dataset comprising of top 3 crime categories. Following are the prediction accuracies of the three algorithms in the first stage.

**Table 3: Using Top 10 Crime Types**

<b>Algorithm</b>	<b>Accuracy</b>	<b>95% CI</b>
<b>Naïve Bayes Classifier</b>	0.2906	0.2867 to 0.2944
<b>Decision Tree Classifier</b>	0.3176	0.3137 to 0.3216
<b>Random Forest Classifier</b>	0.3324	0.3263 to 0.3385

$$Accuracy = \frac{NumberofCorrectedPredictions}{Totalnumberofpredictionmade}$$

Accuracy is the measured as a function of Number of correct predictions made by the algorithm divided by the total number of rows in the test set. Whereas Confidence Interval represents the uncertainty of the estimates given by algorithms. It is an interval statistic which is used to quantify the uncertainty of an estimate. To take an example. Naïve bayes classifier predicted the Accuracy of 29.06% and the accuracy varies between 28.67% and 29.44%. This indicates the estimates can vary (+-)0.39% at 95% confidence interval. In the first stage the dataset suffered from a high class

imbalance problem due to which the accuracies were not as desired. This problem occurs in many datasets where the target label is multinomial and class distribution of data are highly imbalanced. Equal Number of instances of every class given in the training is the ideal case for machine learning algorithms to perform well. This is not the case in the crime dataset used. The frequency distribution of the crime type is skewed towards the top 3 classes. Multiple techniques can be used to deal with this problem.

In the second stage we included just the top 3 categories of Theft, Battery and Criminal damage. Random sampling is done again to get unbiased sample. 80% - 20% training-test split was used again for this iteration.

**Table 4: Using top 3 Crime types**

Algorithm	Accuracy	95% CI
Naïve Bayes Classifier	0.5135	0.508 to 0.5191
Decision Tree Classifier	0.5333	0.5278 to 0.5388
Random Forest Classifier	0.5503	0.5442 to 0.5564

The detailed results are listed in the following Table 5.

**Table 5: Detailed Results of the Algorithms**

	Naïve Bayes Classifier			Decision Tree Classifier			Random Forest Classifier		
	BATTERY	C DAMAGE	THEFT	BATTERY	C DAMAGE	THEFT	BATTERY	C DAMAGE	THEFT
<b>Sensitivity</b>	0.6746	0.0885	0.5664	0.6458	0.0036	0.6699	0.5874	0.0244	0.6786
<b>Specificity</b>	0.5562	0.9432	0.7165	0.5950	0.9987	0.6261	0.6255	0.9852	0.5866
<b>Precision</b>	0.4607	0.2701	0.6187	0.4726	0.3929	0.5927	0.4685	0.2817	0.5714
<b>Recall</b>	0.6746	0.0885	0.5664	0.6458	0.0036	0.6699	0.5874	0.0244	0.6786
<b>F1-Score</b>	0.5475	0.1333	0.5914	0.5458	0.0071	0.6289	0.5213	0.0448	0.6204

The results in the second stage improved around 15-20% for the three algorithms with the random forest classifier providing the best estimates. From the F-1-Score we can see that all three algorithms performed well on prediction the 'THEFT' class but due to randomness of the data, complex underlying pattern for 'C DAMAGE' is not possible to be predicted by these algorithms as accurately as the other two class. Overall on both datasets, it can be said that Random Forest Classifier performed better than the other two algorithms. As this is real world dataset, it is impossible to have all the independent features. To take an example, although the spatial features of beat, ward, community area and district provide different details but these features are not independent as they are subdivisions of the geography. These spatial features are independent of the temporal features (month, day) but these temporal features are very less in numbers. The clustering algorithm used to build a new feature as represented in feature extraction section, is also built using temporal features of latitudes and longitudes. These coordinates also have a high correlation with districts and wards. To summarize, the features used for model building have high dependence on each other which is the main reason of naïve Bayes Classifier giving low accuracy in both stages. New features can be extracted in the future using multi modal data of city of Chicago. The Decision Tree classifier although has a property of finding the randomness efficiently but it suffers from over-fitting and lack of features to train with. Decision tree gave out the max test accuracy of 53.33% in second stage. The underlying problem is the lack of features and high number of distinct values of the features. One way to generalize the training of Decision tree is to use Random Forest Classifier which is specifically used for real world problems where the problem of over-fitting arises. Random forests consist of multiple single trees each based on a random sample of the training data. Random Forests are generally more accurate than single decision trees. Although the depth of decision tree can be adjusted to avoid the over-fitting, the problem of high number of distinct values is not solved using this approach. It is better to grow on multiple decision trees to greater depth and generalize the results of the grown trees. While the decision trees are often pruned, a random forest tree is fully grown and unpruned thus, providing the ability to split feature space in to many smaller regions without introducing over-fitting. As evident from the results, Random Forest Classifier performed much better the Decision Tree and Naïve Bayes Classifier.

## 5. Conclusion and Future Works

Reducing the crime rates has a great impact on the wellbeing of individuals and the society and it helps a great deal in improving the overall quality of life. Use of Crime prediction can significantly help in this regard. The crime reporting systems act as a backbone of Crime prediction system. These systems are in a very good shape in countries like US but there is a lack of crime reporting structure in Pakistan which is the main cause of less developments in projects which can give criminal insights and predict occurrence of crimes in Pakistan.

### 5.1 Conclusion

Logistic regression can be used to impute missing values in the data set where the columns are multinomial. The column with the missing values can be treated as a target label and the prediction of missing values can be done given rest of the columns as features for this prediction. Class Imbalance is a major problem affecting the accuracy. This problem is inherent in the dataset in which the target class is multinomial and the distinct values of the class are not evenly distributed. The study took top 3 classes which showed improved results. Moreover, random sampling methods can be used in future to reduce the class imbalance and improve the accuracy of predictions. There is also a need to use multi modal data sets in Pakistan. These dataset are very hard to find in Pakistan but can be of great use. There are other data-sets available in Chicago data repository, common keys can be identified and multiple datasets can be joined to produce better results. Even after the experiments on other datasets, the government needs to put effort in reporting and managing datasets in Pakistan so that multiple studies can be performed.

### 5.2 Future Works

Dependence of features is a problem inherent in crime datasets. The performance of Naïve Bayes Classifier suffers greatly because of the mathematical independence assumption. In future, extracting more features using multimodal data can prove to be very successful in this regards. High number of distinct values causes a reduction in the prediction accuracy of Decision tree. It is hard to tune the decision tree given many distinct values of individual columns i.e. finding a balance between over-fitting and generalization. In future, this can be achieved using Random Forest Classifier which trains multiple decision trees to explore the feature space in detail with less over-fitting.

In summary, there is a need of better crime reporting. Improved crime reporting in Pakistan can help in creating real time datasets on which multiple techniques can be applied in future and improvements in prediction can be studied using crime data encompassing the dynamic in Pakistan.

Since occurrence of crime is largely depended on natural factors, it is difficult to build a diverse feature set which gives high accuracy in the real world problem of crime prediction. Multiple techniques like clustering can be used for feature extraction and multi-modal data like education, traffic and financial data can be used to include more natural factors causing crime in the community. Use of temporal features like day of week or time of day and spatial features like district, ward or beat is not enough for predicting crime with high accuracy.

# References

1. Alexander Stec and Diego Klabjan, June 2018, Forecasting Crime with Deep Learning, Northwestern University
2. Alimadad, A., Ghaseminejad, A. H., Borwein, P., Christopher, C., Brantingham, P., Li, J., Brantingham, P., Pollard, N., Dabbaghian-Abdoly, V., Rutherford, A., Ferguson, R., van der Waall, Alexa and Fowler, E. (2008). Using Varieties of Simulation Modeling for Criminal Justice System Analysis. Ch 19 In Liu, L. & Eck, J. Artificial Crime Analysis Systems: Using Computer Simulations and Geographic Information Systems. Premier Reference Source. USA.
3. Atkin, Howard N. 2002. "Intelligence Led Policing." In Intelligence 2000, Revising the Basic Elements, edited by Marilyn Peterson, Bob Morehouse, and Richard Wright, 13-21. Sacramento, CA: Law Enforcement Intelligence Unit.
4. Bennett, Wayne W., and Karen M. Hess. 2001. Management and Supervision in Law Enforcement, 3rd ed. Belmont, CA: Wadsworth
5. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Héigeartaigh, S., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J., Yampolskiy, R., & Amodei, D, The malicious use of artificial intelligence: Forecasting, prevention, and mitigation, 20 Feb 2018.
6. Bshayer S. Aldossari, Futun M. Alqahtani, Noura S. Alshahrani, Manar M. Alhammam, Razan M. Alzamanan, Nida Aslam, Irfanullah, A Comparative Study of Decision Tree and Naive Bayes Machine Learning Model for Crime Category Prediction in Chicago, ICCDE 2020: Proceedings of 2020 the 6th International Conference on Computing and Data Engineering, January 2020.
7. C. McCue, and A. Parker, "Connecting the Dots: Data Mining and Predictive Analytics in Law Enforcement and Intelligence Analysis," The Police Chief, 2003.
8. Charlie Catlett University of Chicago Argonne National Laboratory, Chicago, IL, USA , Eugenio Cesario ICAR-CNRR Rende (CS), Italy, Domenico Talia DIMES - University of Calabria Rende (CS), Italy, Andrea Vinci ICAR-CNRR Rende (CS), Italy. June 2018, A Data-Driven Approach for Spatio-Temporal Crime Predictions in Smart Cities.
9. Dennis, L., Fisher, M., Slavkovik, M., & Webster, M. Formal verification of ethical choices in autonomous systems. Robotics and Autonomous Systems, 77, 1–14, December 2015.
10. Farzad Shahbazi, Using Decision Tree Classification Algorithm to Design and Construct the Credit Rating Model for Banking Customers, OSR Journal of Business and Management (IOSR-JBM) e-ISSN: 2278-487X, p-ISSN: 2319-7668. Volume 21, Issue 3. Ser. II, March. 2019, Pg 24-28
11. Gogarty, B., & Hagger, M. The laws of man over vehicles unmanned: The legal response to robotic revolution on sea, land and air. Journal of Law, Information and Science, 19, 73–145.
12. H. Chen, W. Chung, Yi Qin, M. Chau, J. Jie Xu, G. Wang, R. Zheng, and H. Atabakhsh, "Crime Data Mining: An Overview and Case Studies", in Proceedings National Conference on Digital Government Research, 2003, pp. 50-51.
13. Hoffman, Bruce, David W. Brannan, Eric Herren, and Robert Matthiessen. 2004. Preparing for Suicide Terrorism: A Primer for American Law Enforcement Agencies and Officers. Santa Monica, CA: Rand.
14. J. Eck, S. Chainey, J. Cameron, M. Leitner, and R. Wilson (2005) "Mapping Crime: Understanding Hot Spot" , London's Global University website. [Online]. Available: <http://eprints.ucl.ac.uk/11291/1/11291.pdf>

15. Jesia Quader Yuki, Zaisha Zamal, Predicting Crime Using Time and Location Data, DOI: 10.1145/3348445.3348483: July 2019
16. Martínez-Miranda, E., McBurney, P., & Howard, M. J. Learning unfair trading: A market manipulation analysis from the reinforcement learning perspective, In Proceedings of the 2016 IEEE conference on evolving and adaptive intelligent systems, EAIS, 2016, pp. 103–109.
17. O. J. Oyelade, O. O. Oladipupo, I. C. Obagbuwa, Application of k Means Clustering algorithm for prediction of Students Academic Performance, International Journal of Computer Science and Information Security (IJCSIS), Vol. 7, No. 1, pp. 292-295, January 2010, USA
18. Oatley, G., C., Zeleznikow, F. J. and Ewart, B.W. (2004). Matching and Predicting Crimes. In: Macintosh, A., Ellis, R. and Allen, T. (eds.), Applications and Innovations in Intelligent Systems XII. Proceedings of AI2004, The Twentyfourth SGAI International Conference on Knowledge Based Systems and Applications of Artificial Intelligence, Springer: 19-32. ISBN.
19. Pagallo, U, From automation to autonomous systems: A legal phenomenology with problems of accountability, In Proceedings of the 26th international joint conference on artificial intelligence (IJCAI-17) (pp. 17–23).
20. Prajakta Yerpude(1) & Vaishnavi Gudur(2), Predictive Modelling Of Crime Dataset Using Data Mining, Department of Computer Science, Illinois Institute of Technology, Chicago, Illinois.
21. Ratcliffe, Jerry H. April 2003. "Intelligence-led Policing." Trends and Issues in Crime and Criminal Justice. No. 248. Canberra: Australian Institute of Criminology.
22. Redmond, M. and Baveja, A. (2002). Computing, Artificial Intelligence and Information Technology A data-driven software tool for enabling cooperative information sharing among police departments. European Journal of Operational Research, 141, 660–678.
23. Rizwan Iqbal, Masrah Azrifah, Aida Mustapha, An Experimental Study of Classification Algorithms for Crime Prediction, March 2013, Indian Journal of Science and Technology, 6(3):4219-4225
24. Unnati R. Raval, Chaita Jani, Implementing & Improvisation of K-means Clustering Algorithm, International Journal of Computer Science and Mobile Computing (IJCSMC), Vol. 5, Issue. 5, May 2016, pg.191 – 203
25. Waltz, D. L. (1996). Artificial Intelligence: Realizing the Ultimate Promises of Computing. NEC Research Institute and the Computing Research Association,
26. Zuev, A. and Fedyanin, D. (2012). Models of Opinion Control for Agents in Social Networks. Problemy Upravleniya, 2, 37–45