Crime Prediction using Data Mining Techniques

By

Ammad Azam Tarar

NUST201464086MSEECS61314F

Supervisor

Dr. Mian M. Hamayun

Department of Computing

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

In

NUST School of Electrical Engineering and Computer Science (SEECS)

National University of Science and Technology (NUST), Islamabad, Pakistan.

(2018)

Certificate

Certified that the contents of thesis document titled "Crime Prediction using Data Mining Techniques" submitted by Mr. Ammad Azam Tarar have been found satisfactory for the requirement of degree.

Advisor: _____

Dr. Mian M. Hamayun

Committee Member1: _____

Dr. Anis ur Rahman

Committee Member2: _____

Dr. Sohail Iqbal

Committee Member3: _____

Dr. Asad Waqar Malik

Abstract

This thesis addresses the problem of predicting crime and crime types. It is a challenging issue to predict crime as criminals do not follow predefined patterns. Criminals are always looking for places where there is less chance to get caught. They are aware of areas where there is less police patrolling. Therefore, we need to identify potential criminal activities, which will help police and other law enforcement agencies to predict crime beforehand. Most of the crime data consists of crime type attribute, therefore, we have used supervised learning algorithms for predicting crime. We have also used spatial and temporal information for identification of crime types. Neighborhood information is used for identifying race of majority of the population in a particular locality. We have also incorporated census information in the crime data by adding attributes like literacy rate and average income. In this way, we will be in a position to better predict crime and crime types. We have identified criminal hotspots. This study is beneficial for law enforcement agencies for better patrolling by using past crimes. Police patrolling can be increased or decreased based on predefined crime rates in certain localities. On the other hand, this study may also benefit police by identifying whether they need to hire more people or they need to lay off some? We have used Naïve Bayes, Decision Trees, Artificial Neural Network, Support Vector Machine and Ensemble methods for predicting crime. 10 fold cross validation has been used for testing and the results show that Ensemble methods have the best prediction with accuracy of over 80%. In future, we would like to add more features including user employment history as well as previous criminal records.

Certificate of Originality

I hereby declare that the thesis titled “Crime Prediction using Data Mining Techniques” is my own work. 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 or any other educational institute, except where due acknowledgment, is made in the thesis. Any contribution made to the research by others, with whom I have worked at SEECS-NUST 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 to the extent that assistance from others in the project’s design and conception or in style, presentation and linguistic is acknowledged. I have also verified the originality of contents through plagiarism software.

Signature: _____

Author Name: Ammad Azam Tarar

Acknowledgements

I bow before Almighty Allah to express my gratitude for He is the only one who brightens my mind and heart when I feel standing alone in the darkness. He is the only one who helps me when I fall and who is the only hope when I feel broken. I am nothing but His benevolence makes me what I am today. He blessed me even more than I deserve. Thank you Allah!

My words are not enough to pay special credit and appreciation to my supervisor, Dr. Mian M. Hamayun for his continued support and encouragement. I wish to convey my sincere thanks to him for his patience, motivation, and knowledge sharing. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor than him, for my research.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Anis, Dr. Sohail Iqbal and Dr. Asad Waqar Malik for their insightful comments and encouragement. I offer my sincere appreciation for the learning opportunities provided by my committee.

I place on record, my sincere thank you to my Institution and the Department for providing me a platform where I can learn not only about course books but also about how to excel in every field of life. My Institution makes me feel proud.

I would also like to thank my wife, teachers, my classmates, my friends and my colleagues for being the best human beings I have ever met and to those who forwarded positive criticism to me.

Last but not the least, no words in this world are enough to show gratitude and say thanks to my Parents for supporting me spiritually throughout and for all of the sacrifices they have made for me. Their prayers for me were the only thing that sustained me thus far. They were always there for me to support in the moments when there was no one to answer my queries.

My deepest gratitude!

I truly dedicate my endeavor to my Parents and my Wife

For their endless love, support, encouragement and sacrifices for me.

Table of Contents

List of Figures

List of Tables

1 Chapter 1: Introduction

Crime is a well-known social problem. Crime not only harms individual but it has adverse effects on society as well. Crime affects quality of life and hinders economic development of a country. US crime victims have suffered over $14 billion in 2015 according to FBI in 2017; US lost over $15 billion in economic loss to victims while police budget was estimated as $179 billion.

Crime analysis has shown that crime is proportional to economic development and growth of city, state and country. Hence, many studies have focused on relationship between criminal behavior and socio economic conditions like poverty, income, education, unemployment, population density and race. Historically, race has been an important factor in screening potential criminals. Black people in US are commonly discriminated in this regard. Similarly, highly educated people are least likely to be questioned/ screened at airports by law enforcement agencies. In the same manner, poor and unemployed people are considered more likely to commit crimes due to their economic conditions.

Geographical crime research has shown that every city has certain parts with high crime rates while some parts have low crime rates. A crime hotspot is defined as an area with above average crime rate. Research is still carried out to determine hotspots and improve police patrolling in those areas. Crime clustering techniques like K-Nearest Neighbor are very popular in this regard.

Social media has become a platform where people share information about literally everything. Hash tags from twitter can be used to identify popular discussions. For example, in case of mass shooting in Virginia University, more than 10,000 users tweeted with hash tag of #VirginiaShooting. Similarly, Facebook also allows hash tag in status updates. Recently, Facebook has also incorporated safety feature which allows users to mark themselves as safe during any dangerous incident. Both twitter and Facebook provide Application Programming Interfaces which allows researchers to get access to these kinds of data. In this way, both law enforcement agencies and general public can get access to crime information. Ultimately, it helps public to refrain from visiting dangerous neighborhoods especially during such incidents. Law enforcement agencies also become quickly aware about any criminal activity. In this way, criminal and terrorist activities can be identified and stopped.

Crime forecasting can help police to efficiently use resources. Police patrols can be deployed on hotspots and they can be relocated if crime in certain area reduces. It can also help police to determine when they need to hire new people. Police can also use forecasting to increase or decrease their resources in a certain city. In the long term, this study will help in reducing budget and ultimately, money will be utilized in a better way.

Crime forecasting and prediction is a very important task for police and other law enforcement

agencies. Crime statistics in an area has significant impact on population in that locality and can trigger movements to more safe localities. It is no surprise that people want to live in peaceful areas. With the emergence of artificial intelligence and date mining, law enforcement agencies need to be more aware of crime ratios in different parts of cities. This will also create a healthy competition among law enforcement agencies to reduce crime in their localities. This will ensure reduction of crime.

Machine learning is already being used by US police departments to identify officers who are under lot of stress. It uses all employment record of officers including previous use of gun violence. In this way, potential stressed employees are identified who might be at risk of committing unlawful activities. North Carolina Police department has identified 48 such officials in 2015.

Machine learning and data mining algorithms have become prevalent to combat and analyze crime. Although, many approaches to crime analysis have been proposed but this is still a relatively young field and has tremendous potential for growth. Crime analysis requires analyzing huge amount of data. Therefore, data mining and machine learning are considered to be the best approach for research in this field. Data mining researchers have developed lot of tools and sophisticated algorithm for such kind of analysis.

We have chosen classification techniques to predict crime due to the availability of 'crime type' class variable. Therefore, we have not considered unsolved crime cases as crime type is missing or unavailable for those cases.

1.1 Motivation

Contribution for the betterment of society is the primary motive for selection of crime prediction domain. This domain has the ingredients to bring peace in our lives. Many countries store and maintain crime data but it is only used for registration of crime. In our age Artificial Intelligence, it is the need of the hour to predict and forecast crime by using existing crime data. Research in this topic can help in following domains:

1.2 Crime Hotspots

Crime datasets of particular area can help us to locate crime hotspots and we can uncover hidden patterns of criminal activities. Hotspots are areas with above average rate of crime. Identification of crime hotspots will help police in efficient planning of patrolling. It is generally expected that areas with high rate of patrolling eventually reduces crime in that locality.

1.3 Crime Prediction

Crime datasets can be used in prediction of crime. Crime prediction is helpful for law enforcement agencies to reduce and eliminate crime. Crime data combined with census data can screen potential criminals by identifying race, age, literacy, income etc. In this way, criminal activities can be identified and ahead of time action can be taken before any such event takes place.

1.4 Human Resource Management

Crime prediction and detection of hotspots will empower LEAs to determine the number of personnel required for patrolling and other activities. In this way, LEAs can know when to hire more people. Similarly, patrolling staff can be reduced when crime reduces in a certain area.

1.5 Problem Statement

The aim of this study is to identify, predict and analyze crime data and crime types by improving existing state of the art using spatial and temporal information. We want to identify criminal neighborhoods in US cities of Austin, Boston, Denver and Los Angeles. We have not predicted crime in Pakistan due to unavailability of crime data due to security reasons. We will also predict the potential type of criminal activity using criminal and census data. This study will be applicable to predict crime for other datasets including Pakistan.

1.6 Challenges

Crime datasets consist of neighborhood information but there is no standard mapping of city districts and neighborhoods. Hence, determination of district from neighborhood has proved to be very challenging task. Crime datasets constitutes dozens of crime types but for the sake of this study, we have reduced them to six standard crime types. Conventional techniques were not suitable to standardize crime types. Therefore, non-conventional methods have been used for this study.

1.7 Contribution

We have proposed crime prediction model which can predict type of crime based on crime and census data. Classification models including Naïve Bayes, Decision Trees, Artificial Neural Network, Support Vector Machine and Ensemble methods have been used for prediction of crime. This model has been evaluated on crime datasets of Austin, Boston, Denver and Los Angeles. Significant improvement on existing models is observed due to incorporation of new features including literacy rate, income and race.

The main contribution of this research are as follows:

- New feature of income is added to crime dataset on the basis of neighborhood of crime from US census website.
- Similarly, literacy rate feature is also incorporated in crime data from US census website.
- US cities are divided into districts. District is calculated from crime neighborhood and census data is used to determine dominant race in particular district. In this way, crime forecasting results have been improved.

1.8 Thesis Outline

This thesis is organized as follows: Chapter 2 presents literature review on different state of the art crime forecasting techniques for classification of crime and crime types. Literature review also consists of methodologies for detection and prediction of crime hotspots. Chapter 3 describes the

proposed approach for crime prediction and also analyzes different crime datasets. Feature extraction and incorporation of new features are also explained in this chapter. Evaluation and results of classification algorithms on crime datasets constitutes 4th chapter. Evaluation and thesis summary are part of 5th chapter.

2 Chapter 2: Literature Review:

2.1 Overview

This chapter discusses crime prediction and hotspot identification. We have divided literature review in three parts: (1) Hotspot detection, (2) Crime prediction and (3) Crime pattern detection. In the 1st part we have provided literature review on hotspot detection. In the 2nd phase, we have discussed crime prediction and in the final section, we have reviewed literature on crime pattern detection.

2.2 Hotspot Detection

Crime prediction and pattern discovery is used by K.R. Sai Vineeth [4] to classify states as most dangerous, dangerous, moderate and safe. FP Max is used for finding frequent patterns. It has better performance than FP Growth and Apriori techniques. FP Max reduces complexity and uses less memory. Indian crime data from 2001 to 2012 is used for crime prediction. It contains 31 attributes and over 10 million instances. Correlation is used to generate weightages of crime types to get intensity of crimes (CIP). Pearson correlation coefficient is used for this purpose. Random forest is used for classification of states. It achieves 97% accuracy. Support Vector Machine (SVM) is used for prediction. SVM results in root mean square error of 3.2%. Random forest is preferred by author because of less running time.

Hotspots are areas with above average crime rates. Prediction of hotspots is classified into three categories by Zhang et al [5]. First category uses clustering methods to make map of hotspots and predicts changes in hotspot areas. Second category uses probability to predict crime types which helps to draw hotspots on map. Third category is determined by Time Series analysis which uses historical data for performing regression and/or machine learning methods. All these prediction functions are depended on input data and fail to perform well on unknown/new data. 5 heat levels are determined to classify hotspots. LDA is used for dimension reduction while KNN is used for prediction of heat level. Average root mean square error (RMSE) is around 0.5. Hence, this model predicts crime accurately.

Bagula et al [17] have devised CitiSafe which is a crime prevention tool which is build using FP Growth algorithm. Frequent pattern mining can help law enforcement agencies to predict and prevent crime. CitiSafe can identify crime using spatial and temporal crime data. It is sensitive to new data. It is easy to use and requires no computer science knowledge like SQL. It creates heat maps for crime hotspots. Colors of hotspots determine intensity of crime in a particular region.

Hotspot and heat map analysis is done on Rawalpindi, Pakistan crime data by Malik et al [19]. Crimes are divided into two types; crimes against person and crimes against property. Kernel

density function and Getis-Ord algorithm is used. Socio economic and geographical features are used for analysis. Heat maps will determine areas with type and intensity of crime. Behavior of "Crime against Person" type shows no hotspot due to sparse data. "Crime against Property" shows hotspots with different intensity. In this way, we can find areas where there is high property crime and ultimately, safe neighborhoods can be identified.

2.3 Crime Prediction

Nafiz Mahmud et al [2] have developed CRIMECAST which is a crime analysis system that focuses on history of crime to predict future criminal activities. It is a mathematical simulation that predicts crime using location, time and nature of crime. It is developed using last 30 years of crime data. Firstly, hotspots are detected to help law enforcement agencies to focus on areas with high crime. Artificial neural network is created which includes precedence factor, time factor, season and weather. Precedence factor and time factor are given arbitrary weights. These weights are updated after every iteration. Ultimately, precise output is achieved. Crime season is determined using cultural and environment factors. As a result, precise prediction is achieved using ANN.

Mohammad Al Boni et al [3] have proposed Area specific crime prediction model because criminal activities vary across cities. Law enforcement agencies use global models for resource allocation. These models usually don't work well for specific areas. Author has used two global methods and three area specific models. Area specific models include Pooled model, hierarchical model and Multi-Task model. Chicago crime data is used for evaluation. Data is divided into each zip code area for area specific models. Surveillance plot is used for evaluation. Multi-Task model outperforms both global models. Law enforcement agencies could have patrolled more area using area specific models which would reduce crime significantly. Hence, area specific model gives improved prediction of crimes as compared to global models.

Tayebi et al [6] have proposed Location learning from crime prediction model using Crime Pattern Theory. It is based on the assumption that offenders are afraid to go into unknown territory and most of the crime is committed in most familiar areas. CRIMETRACER models criminal opportunities. It is also used for prediction of hotspots. 2001-2006 crime data of BC, Canada is used. We have divided persons associated with crime into four categories. This includes suspect, charged, chargeable and charge recommended. 80% data is used for training and 20% is used for testing. First, learning of offender activity space is done. Then future crime location is predicted. N highest probability roads are selected and removed because focus is on cold spots. Evaluation is done using recall and precision. CRIMETRACER works better in cold spots than all other models.

Babakura et al [7] have used Naive Bayes and Back Propagation (BP) to predict crime in different US states. 1990 US census data is used which contains 128 attributes and 2000 instances. Socio economic factors are also used to improve crime prediction. First group of data contains race and

2nd group contains marital status. State, population and violent crime per person are added to data. Naïve Bayes gives accuracy of 90% in group 1 and 92% in group 2. BP achieves accuracy of 65.9% in group 1 and 65.9% in group 2. Weka is used for testing. Testing was done using 10 fold cross validation. It is clear that Naïve Bayes is better than BP for crime prediction.

Keivan et al [8] have used Support Vector Machine (SVM) to analyze crime data. Model is developed to identify hotspots. K means clustering is used to consider data within certain distance form center. Points with crime rate more than selected threshold will be labelled as hotspot. Crime data of 49 neighborhoods in Ohio, USA is selected. It consists of 20 attributes. N fold cross validation is used for testing. Linear SVM gives 63% accuracy. Polynomial SVM gives 60.5% accuracy while Gaussian SVM produces 68.5% accuracy.

Shiju et al [9] have proposed Crime Analysis and prediction using Naïve Bayes and Decision Tree algorithm. Apriori is used for finding frequent patterns. It helps to find crime pattern in neighborhoods. Data consists of four attributes. Naïve Bayes correctly predicts 90% of data while decision trees accuracy is far less. Hence, Naïve Bayes is better for crime prediction.

Pervaz et al [10] have devised Spatial and temporal model for crime prediction using data of Dhaka. City is divided into grids and it is expected that crime in one grid will affect neighboring grids. Probability of crime in each grid during different time is calculated. Crimes are divided into multiple major categories. Joint probability of factors including time, location, crime type and day is calculated. Low, moderate and high thresholds are used for classification purpose. This model achieves sensitivity of 79% and specificity of 69%. Result is improved if we give more impact to time period relating to crime in surrounding grid. Android alerts application is also developed for grids with high crime prediction.

2.4 Crime Pattern Detection

Kumari et al [1] have used different clustering approaches on criminal data of Tamilnadu. Data contained 1760 instances and 9 attributes. It spans from 2000-2014. K-means clustering algorithm found patterns and relationships in criminal data. Results of K-means are improved by using GMAPI. It plots crime on map and different coloring schemes are used to differentiate crime density. Hierarchical clustering and DBSCAN are used for crime detection. Confusion matrix is used to evaluate performance of above mentioned clustering approaches. DBSCAN has better precision, recall and F-Measure. Hierarchical clustering is 2nd while K-means is last. Hence, DBSACAN performs best for crime detection.

Kadar et al [11] have developed CityWatch which is a mobile based system which helps in prevention of crime. A spatio-temporal model is created which will help to predict location and time of future crimes. Criminal and demographic data is used to build this model. It helps to understand behavior and patterns of crime. Victimization model is created which will help to predict location and demographics of future victims. Logistic regression is used to predict victims.

A co-offending network is proposed by Tayebi et al [12] which helps to identify criminals who work in group. Criminology theory is used to create co-offending model which will identify co-offending networks. Social relation, geographical relation and experience are used to build the model. Police arrests data of British Columbia, Canada is used for evaluation. ROC is used for evaluation. Naive Bayes, Decision Trees and Random Forests are used for prediction. 10 fold cross validation is used for testing. Random forest performs best; decision tree is second while Naive Bayes is last.

Bogomolov et al [13] have developed Geographical crime prediction is done using human behavior and demographic data. Mobile phone activity is used to determine human behavior. London crime and geographical data is used for analysis. Geographical data is divided into different parts (SmartCells) and mobile activity data is acquired through telecom towers in each location. Every crime event is linked to one Smart cell. Entropy is used to predict human behavior such as mobility, spending patterns and other socio-economic factors. Mean, median and standard deviation are also calculated for each aspect of behavior. Pearson correlation is used to understand relationship between features. Naive Bayes, Decision Tree, SVM, neural networks and logistic regression are used with 5 fold cross validation for testing. Decision Tree performs better than all other classifiers for prediction of hotspots.

Crime patterns are detected using clustering algorithms by Nath et al [14]. Clustering is used because there are many unsolved crimes and they can't be classified. Missing data, outliers and noise are removed using standard data mining techniques. Attributes are assigned weight according to expert knowledge. K-means clustering is used. It results in 4 clusters which refer to 4 crime types. Graph of crime patterns along with features will help law enforcement agencies to prevent future crimes.

Demographic data is used to predict crime by Wang et al [15]. 2000 US census data is used for Chicago city. Demographic features include population, poverty, population density, diversity, race distribution and residential stability. Pearson correlation is used to understand relationship between features. Crime data is acquired through Chicago police website. Chicago is divided into 77 community areas. Crime for each area is calculated to normalize population. Demographic data from neighboring areas of Chicago is calculated to find geographical influence. There is positive correlation between geographical influence and crime rate if particular neighborhood has high crime rate. Taxi data is used to calculate relationship between community areas. Crime rate and taxi data is also positively correlated. Evaluation is done using Mean Relation Error (MRE) and Mean Absolute Error (MAE). Negative Binomial Regression outperforms linear regression by 6%.

Fuzzy association rule mining is used by Buscak et al [16] to discover crime patterns. Rare association rule mining has high confidence and low support. US crime data and census data is used for finding frequent patterns. Attributes with missing values are removed and similar attributes are also removed. Data is grouped into regions and every region constitutes communities. Fuzzy Apriori was used with 60% support and minimum support for each region is calculated.

Average support helps in removing rules with no interest. Rules with high confidence indicate that robberies occur less in areas with more retired people. Violent crimes are low in communities with average age of less than 30. Fuzzy association mining helps to find crime patterns in communities, states and regions.

Dynamic Bayesian Network (DBN) has been created between police patrolling and criminals by Zhang et al [18]. Data collector is software component which acquires crime data from police and maps it. As a result criminal hot areas are identified. Hot areas are those where there is no police patrol. Police can cool these areas by patrolling. Data collector also provides live security camera footage to police patrols. Finally, data collector provides a schedule for police patrolling which will help to reduce hot areas.

Elliptical Hotspot Detection (EHD) is proposed by Tang et al [20] which finds areas with high concentration of crime. Fast EHD is proposed which uses pruning algorithm that maximizes advantages of lookup table. Crime data of downtown, Denver is used to compare Fast EHD with other ellipse based algorithms. Fast EHD detects two hotspots and both detect same center. But Fast EHD has higher precision and it works better on boundaries.

3 Chapter 3: Proposed Approach

3.1 Overview

In our proposed solution, we have used different machine learning techniques including Naïve Bayes, Decision Tress, Support Vector Machine, Artificial Neural Networks and Ensemble Methods. The main contributions of this study are as follows:

- New features (literacy, income etc.)are incorporated for prediction of crime
- Crime types are predicted using spatial and temporal data
- Comparison of different machine learning techniques for crime prediction

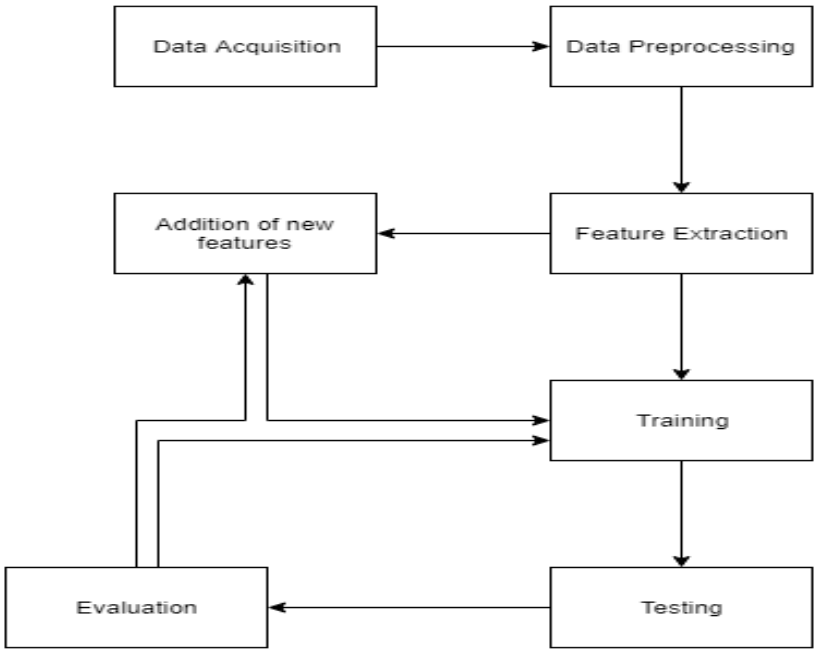


Figure 1 Data Mining Process Model

3.2 Crime Acquisition Data

We have used crime data of four US cities which include Denver, Austin, Boston and Los Angeles. Austin data consists of 95,000 crime instances and 9 attributes. Denver data consists of 186,000 crime instances and 19 attributes. Boston data has 118,000 instances while Los Angeles dataset consist of 196,000 instances. We have incorporated census information in given datasets based on neighborhoods and districts.

3.3 Data Analysis

3.3.1 Austin

Crime data for Austin contains date and time attribute for each crime committed. We have extracted hour from each crime instance. In this way, we have calculated hour attribute for each crime instance. Following figure show the crime committed in each hour of day:

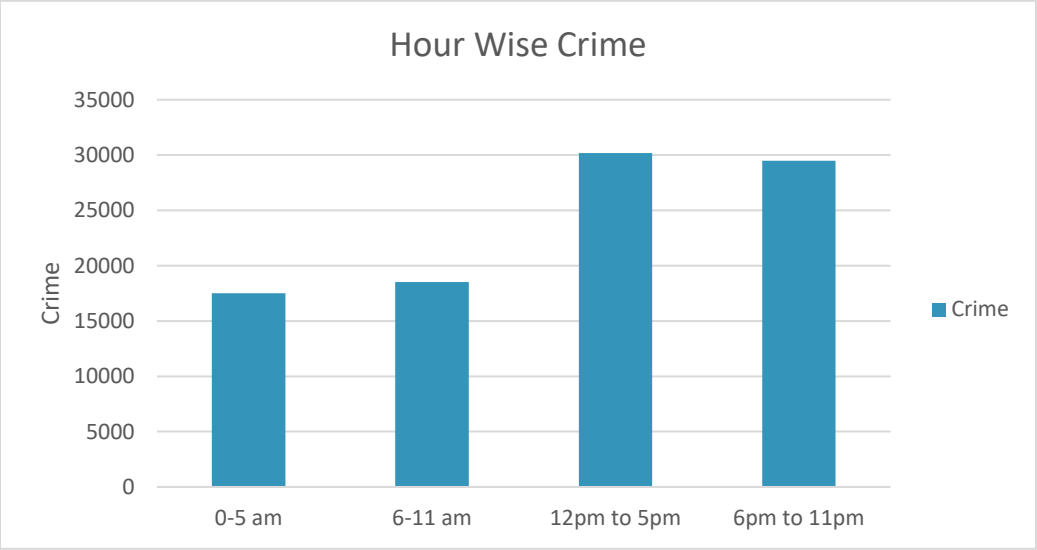


Figure 2 Austin Hour Wise Crimes

We have also extracted day from date time attribute for each crime instance. In this way, we have calculated how many crimes are committed on which day. It is evident from below figure that most crimes are committed on Friday. Following figure illustrates these statistics:

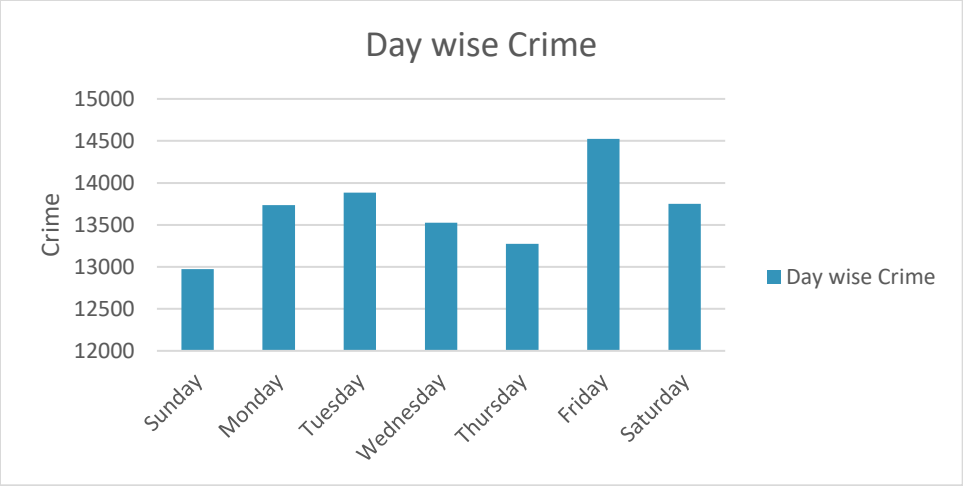


Figure 3 Austin Day wise Crimes

We have broadly divided crimes into six categories which are assault, drug, public disorder, theft, white collar crime and other crimes. Amount of different crime types is described in the below pie chart:

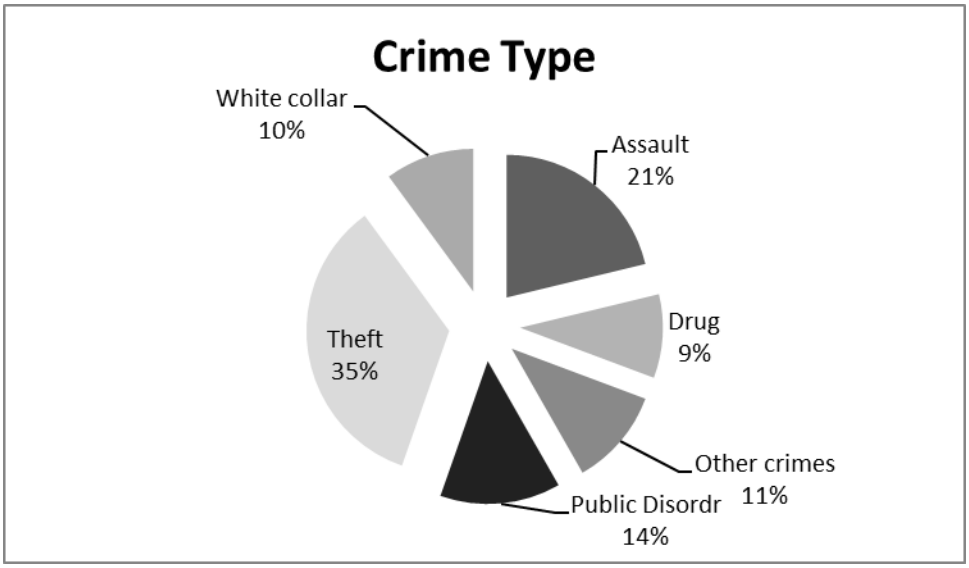


Figure 4 Austin Crime Types

Literacy rate for each neighborhood is calculated from census data. Then, it is incorporated in crime dataset through neighborhood and district attributes. Literacy rate is shown in the below pie chart:

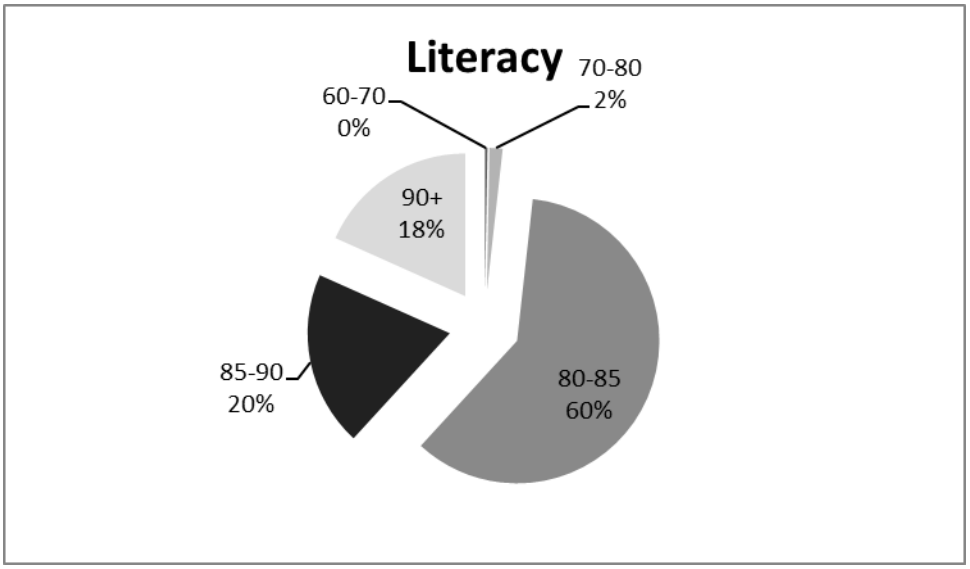


Figure 5 Austin Literacy Rates

Average income for each neighborhood is calculated from census data. It is incorporated in crime data through neighborhood and district attributes. Average income for each neighborhood is illustrated in following figure:

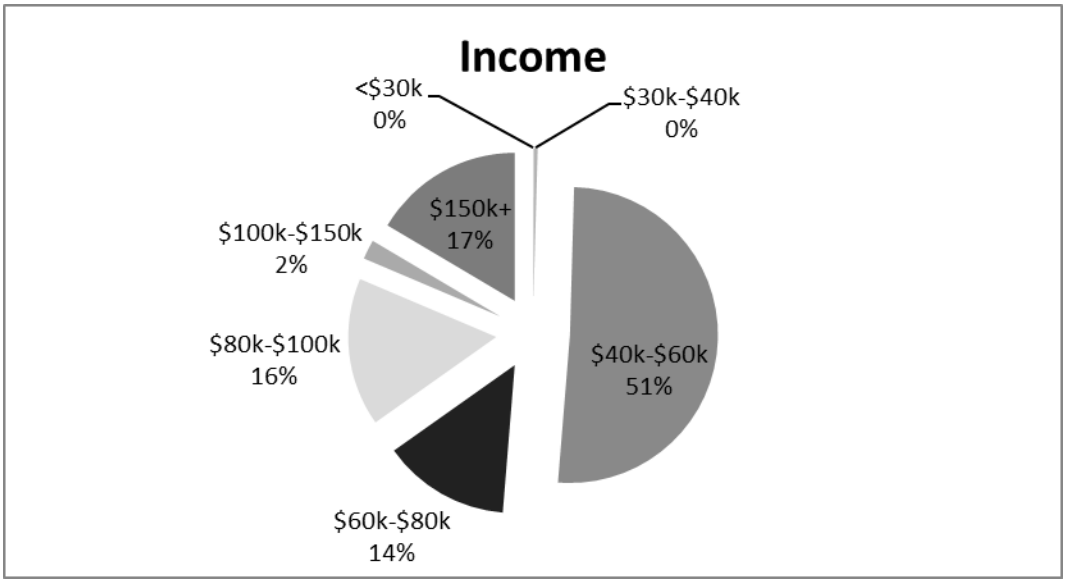


Figure 6 Austin Average Income

3.3.2 Denver

Crime data for Denver contains date and time attribute for each crime committed. We have extracted hour from each crime statistic. In this way, we have calculated hour attribute for each crime. Following figure

shows the percentages of crime committed in each hour of day:

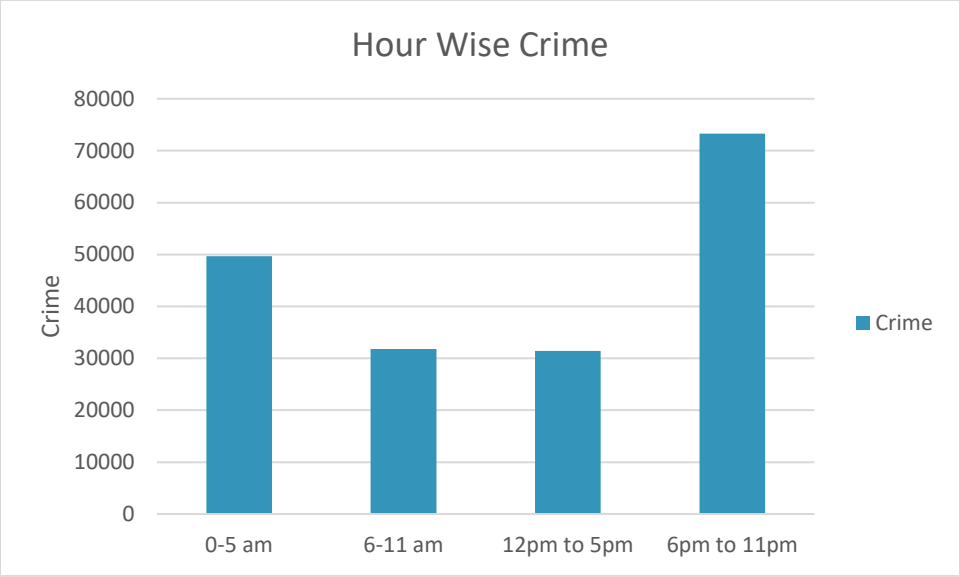


Figure 7 Denver Hour wise Crimes

We have also extracted day from date time attribute for each crime instance. In this way, we have calculated how many crimes are committed on which day. Following figure illustrates these statistics:

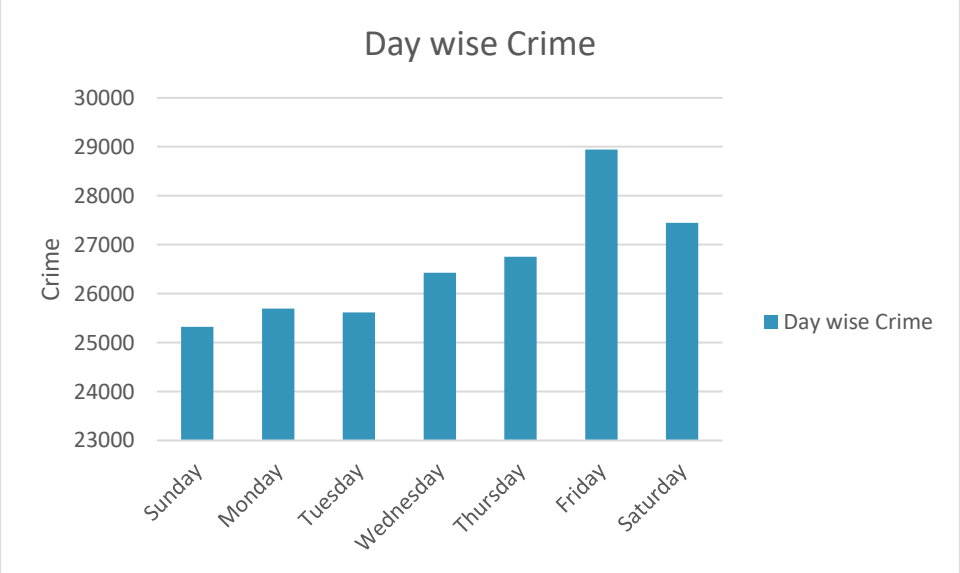


Figure 8 Denver Day wise Crimes

We have broadly divided crimes into six categories which are assault, drug, public disorder, theft,

white collar crime and other crimes. Amount of different crime types is described in the below pie chart:

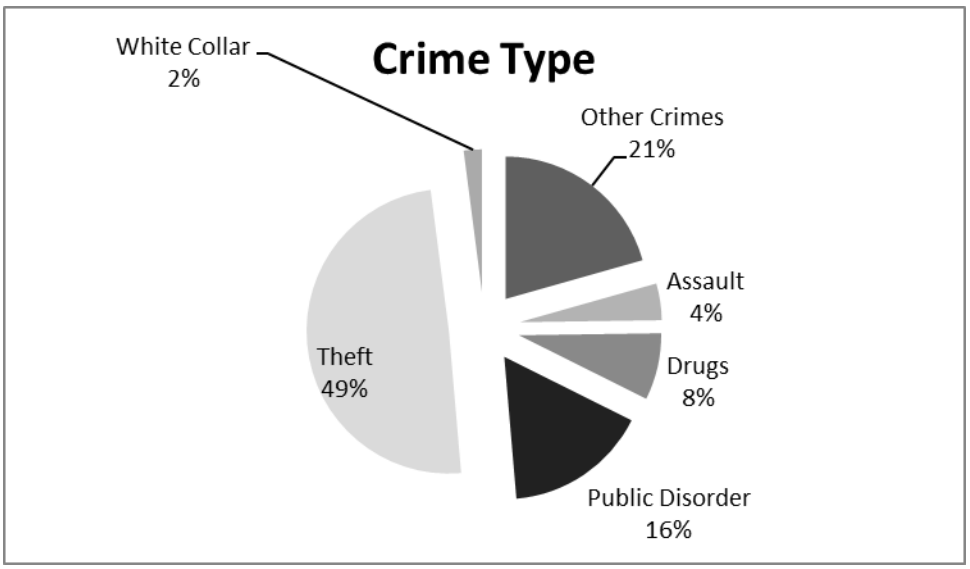


Figure 9 Denver Crime Types

Average income for each neighborhood is calculated from census data. It is incorporated in crime data through neighborhood and district attributes. Average income for each neighborhood is provided in following figure:

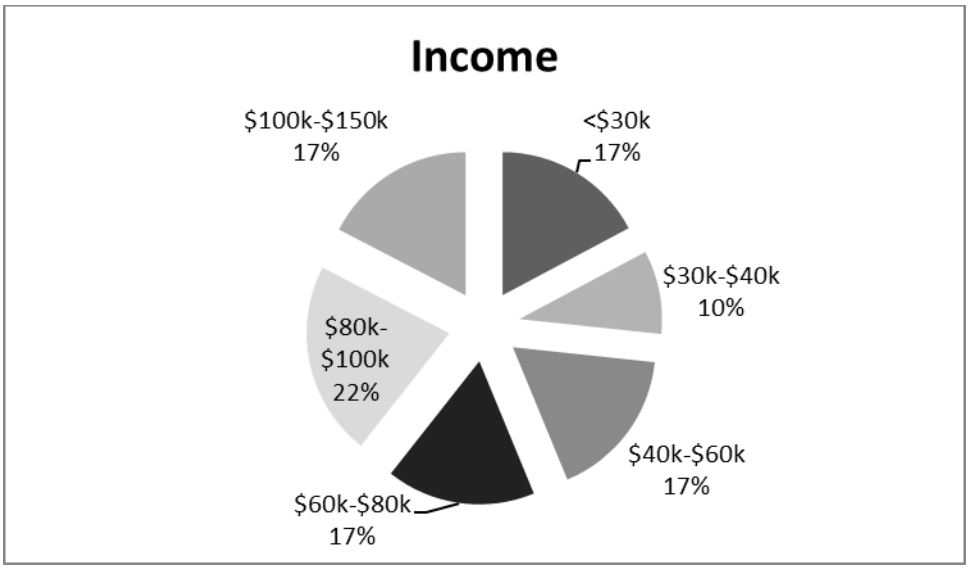


Figure 10 Denver Average income

Literacy rate for each neighborhood is calculated from census data. Then, it is incorporated in crime dataset through neighborhood and district attributes. Literacy rate is explained in below pie chart:

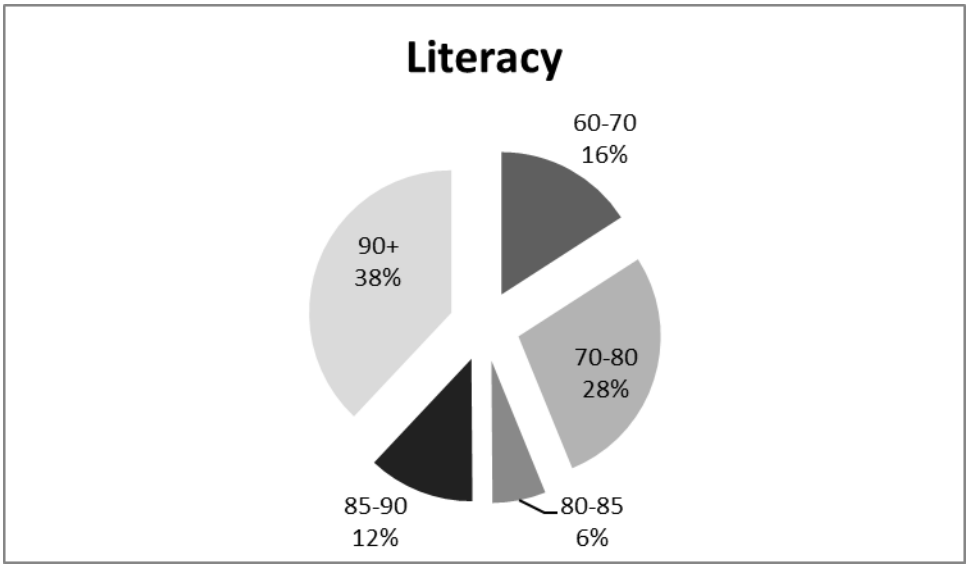


Figure 11 Denver Literacy Rate

3.3.3 Boston

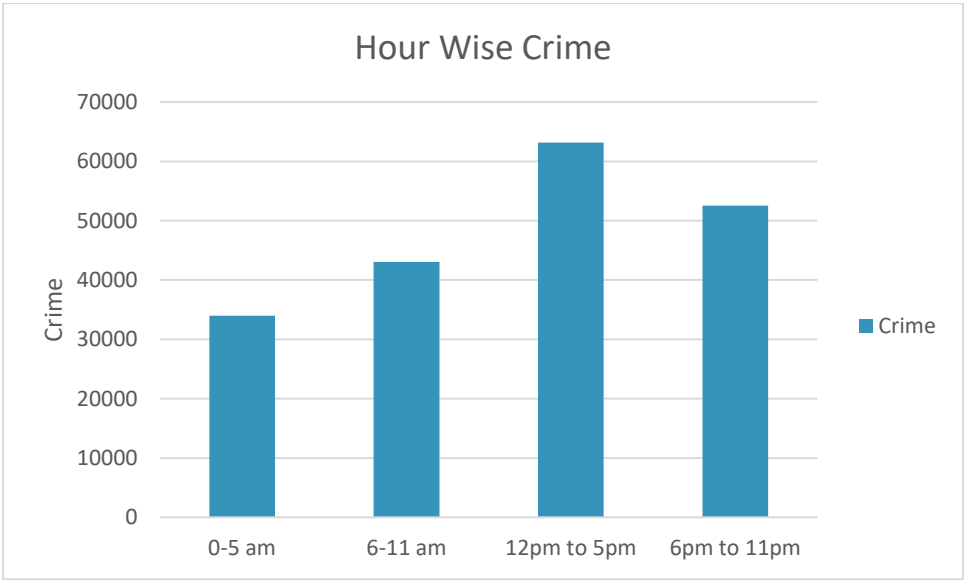


Figure 12 Hour wise crime in Boston

We have also extracted day from date time attribute for each crime instance. In this way, we have calculated how many crimes are committed on which day. Following figure illustrates these statistics:

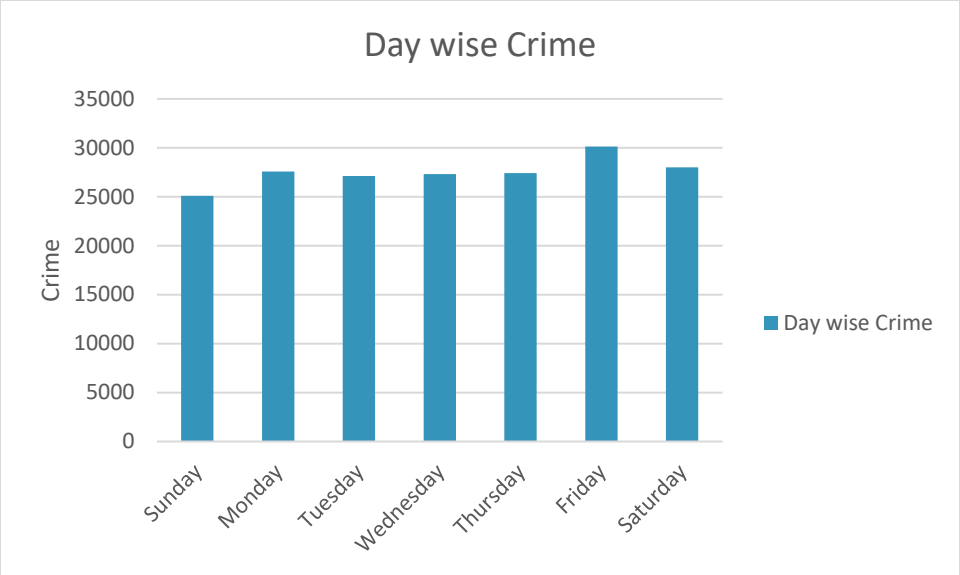


Figure 13 Boston Week wise Crimes

Average income for each neighborhood is calculated from census data. It is incorporated in crime data through neighborhood and district attributes. Average income for each neighborhood is provided in following figure:

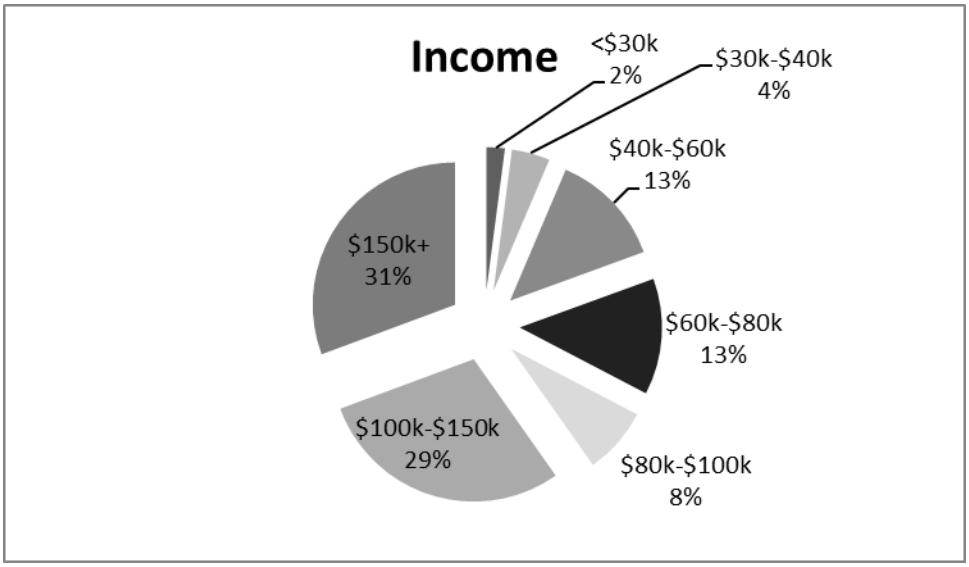


Figure 14 Boston average Income

Literacy rate for each neighborhood is calculated from census data. Then, it is incorporated in crime dataset through neighborhood and district attributes. Literacy rate is explained in below pie chart:

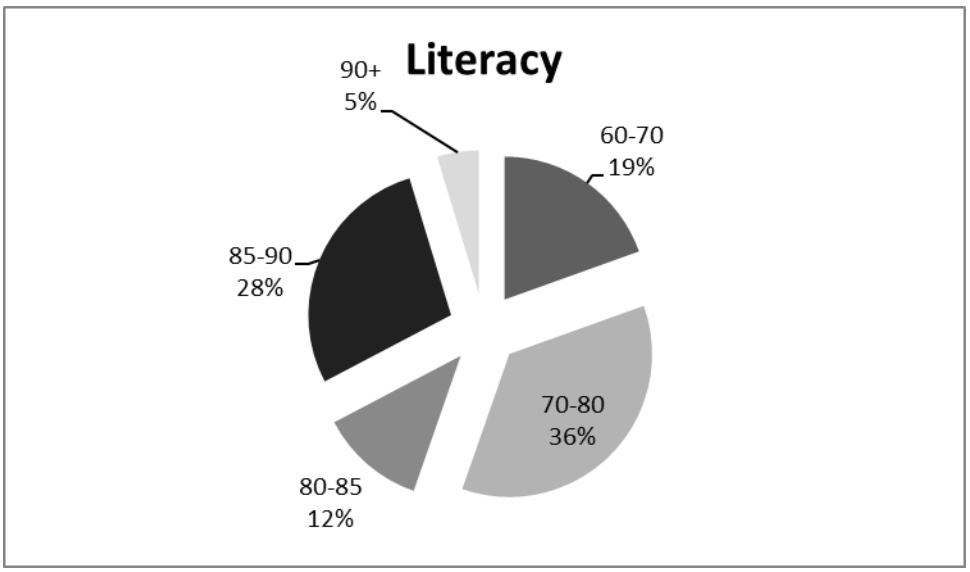


Figure 15 Boston Literacy Rate

We have broadly divided crimes into six categories which are assault, drug, public disorder, theft, white collar crime and other crimes. Amount of different crime types is described in the below pie chart:

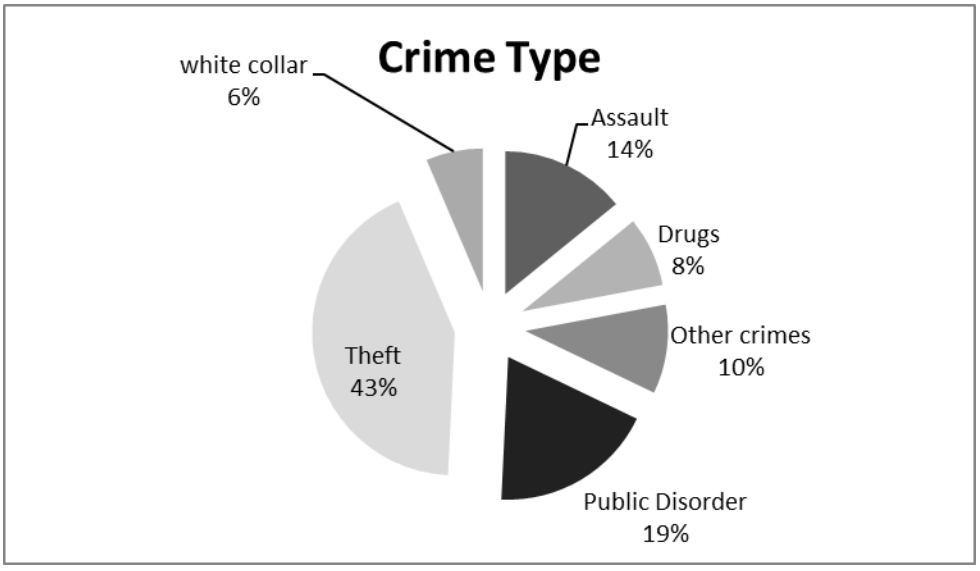


Figure 16 Boston Crime Types

3.3.4 Los Angeles

Hour wise crime is Boston is illustrated in below figure:

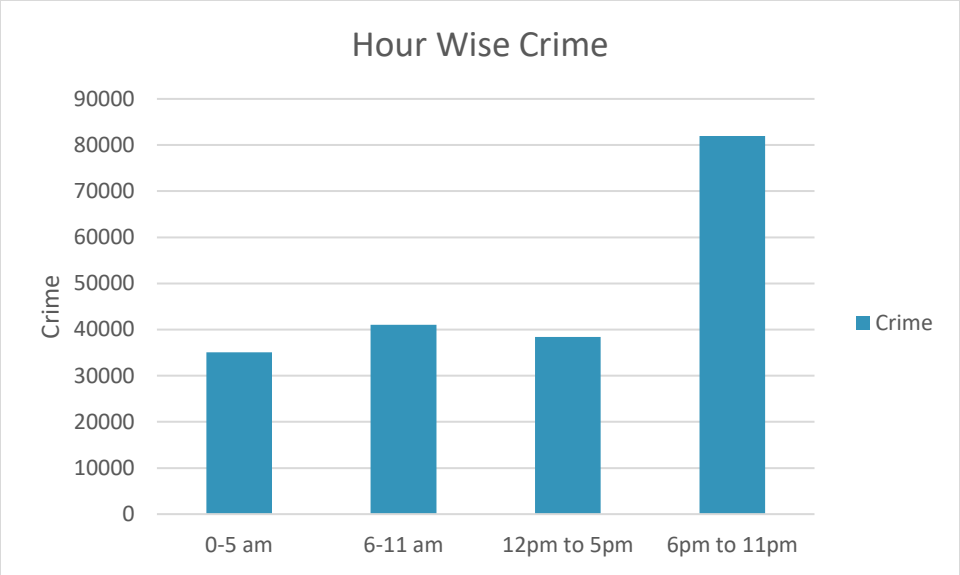


Figure 17 Hour wise crime in LA

Following figure illustrates day wise crime in LA. Same pattern of most crimes committed on Friday is also true for LA.

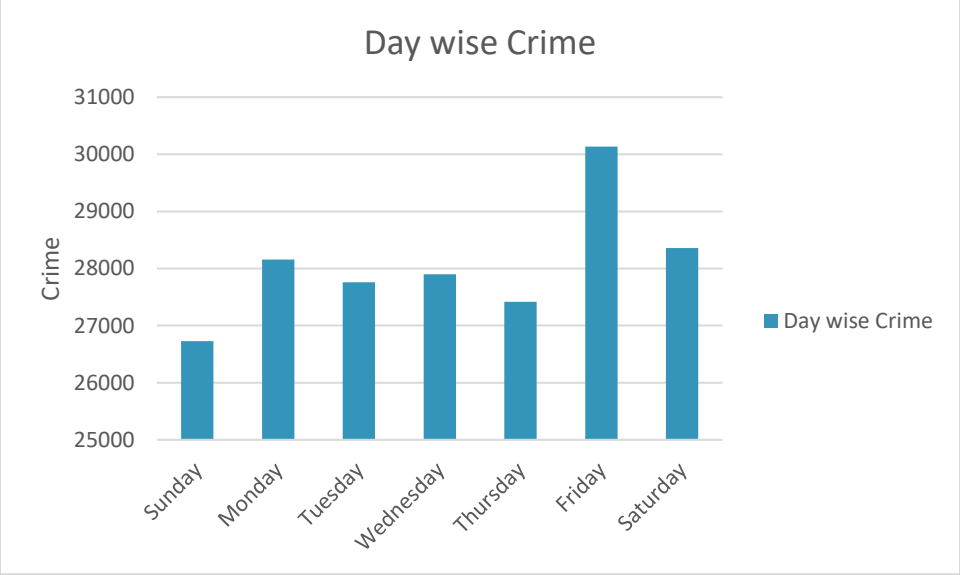


Figure 18 Day wise crime in LA

We have broadly divided crimes into six categories which are assault, drug, public disorder, theft, white collar crime and other crimes. Amount of different crime types is described in the below pie chart:

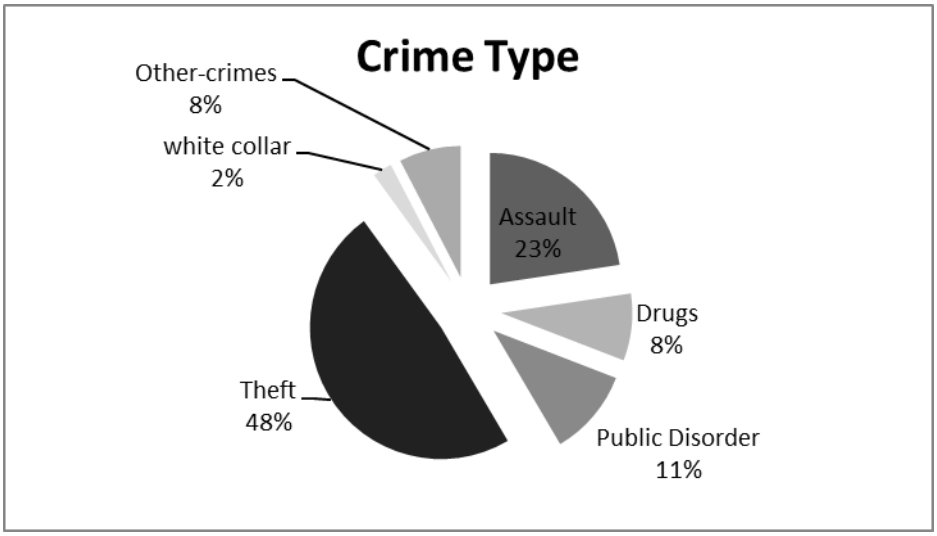


Figure 19 Los Angeles Crime Types

Literacy rate for each neighborhood is calculated from census data. Then, it is incorporated in crime dataset through neighborhood and district attributes. Literacy rate is explained in below pie chart:

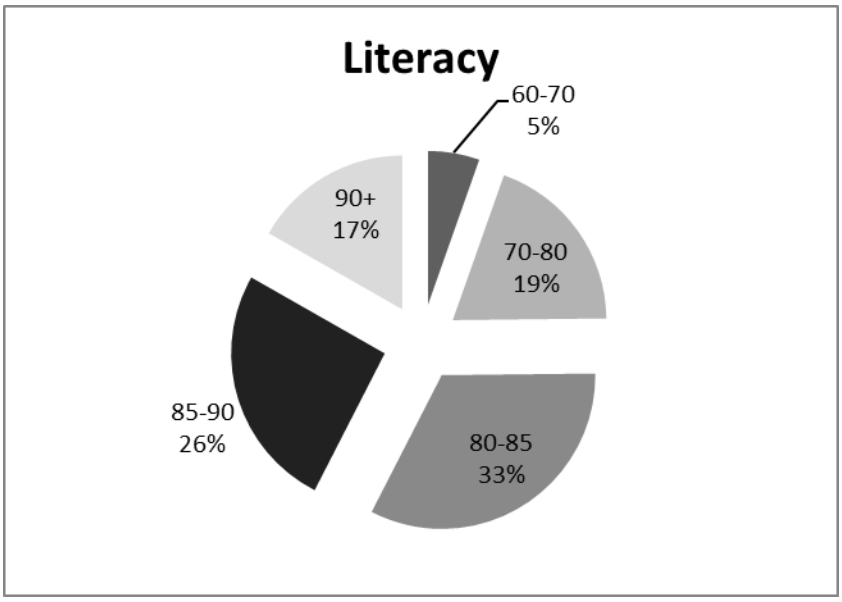


Figure 20 Los Angeles Literacy rate

Average income for each neighborhood is calculated from census data. It is incorporated in crime data through neighborhood and district attributes. Average income for each neighborhood is provided in following figure:

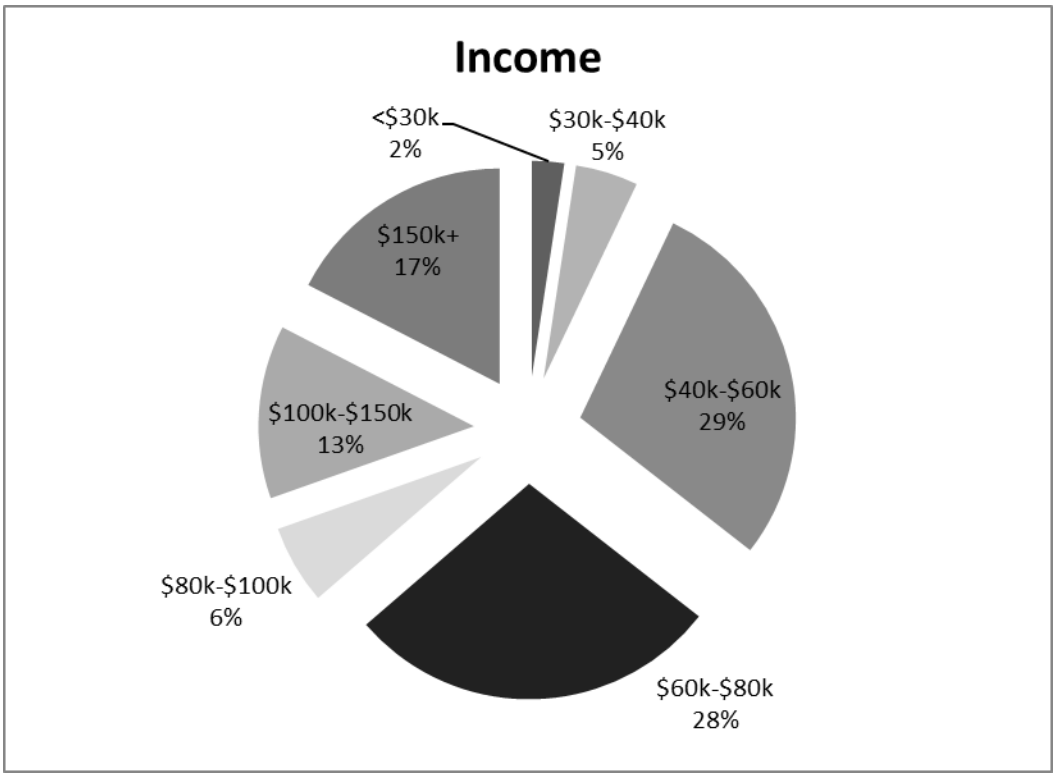


Figure 21 Los Angeles average Income

3.4 Models

3.4.1 Naïve Bayes:

Naïve Bayes is the most commonly used classification algorithm for supervised data. It is based on Bayes theorem with certain assumptions. It is based on the assumption that attributes don't depend on other attributes for classification of class variable. Therefore, naïve Bayes calculates probabilities of all attributes independently. Probability is calculated as follows:

P (A/B) =P (A)*P (B/A)/P (B)

Here "A" is class variable and "B" is attributes. We have used Weka for testing which is a very popular tool for data mining. 10 fold cross validation is used for testing. We have used Naïve Bayes classification algorithm on both datasets of Denver and Austin to predict crime types.

3.4.2 Decision Tree:

Decision tree is supervised learning algorithm which is used for classification. There are multiple implementations of decision tree. We have used ID3 which is proposed by Ross Quinlan. In ID3 algorithm, information gain (IG) of each attribute is calculated. Attribute with small IG is selected as first branch of decision tree. Then, data is reduced due to selection of attribute. This process is

repeated until all attributes are selected. Smaller trees are preferred because large trees might result in over fitting.

3.4.3 KNN:

K Nearest Neighbor (KNN) is a supervised learning algorithm which is used for classification. It is also called lazy learner or instance based learner because all computation is delayed until classification. We have used Euclidean distance as metric. KNN is very sensitive to noise and irrelevant features. Therefore, we have reduced attributes to five for KNN. KNN is very slow because it calculates distance between every point.

3.4.4 Apriori:

Apriori algorithm is used for finding frequent patterns in data. It is very important in crime analysis because it can help to find crime hotspots. It uses bottom up approach to find frequent subsets which are called candidates. Finally, these groups of candidates are tested against the data. Confidence and support are the two most used metrics in Apriori algorithm. Support is defined as total number of occurrences of particular value in data. Confidence tells us how often a rule is found to be true. Mathematically, confidence is defined as:

Conf(X->Y)= Support(X union Y) / Support(X)

We have implemented this model using Weka. We have tested it for different values of support and confidence. Ultimately, we have found minimum support as 0.0015 and confidence as 0.4.

3.4.5 Ensemble Methods:

Ensemble methods are special kind of algorithms which increase the accuracy by combining different machine learning algorithms. There are two types of ensemble methods i-e average and boosting. Average ensemble methods use multiple machine learning techniques and uses average voting for prediction. On the other hand, boosting ensemble methods use weak models and combine them to create a powerful model which results in reduction of bias.

In this study, Random Forest algorithm is used which is a type of average boosting model. Random Forest uses multiple decision trees using random features, which results in reduction of variance. Randomness can result in slight increase of bias but on the average, this model greatly improves the performance in final prediction.

Adaboost is also used for prediction in this study. It is a type of boosting ensemble model. Adaboost combines the output of many weak models which results in boosting of final prediction. Adaboost gathers information about each classifier and training through every iteration, which ultimately helps to correctly predict samples in the data.

3.5 Data Preprocessing

R is used for preprocessing the datasets of Austin, Denver, Los Angeles and Boston. There were many "string" attributes in the given datasets. They need to be converted into numeric because machine learning algorithms require nominal features for solving classification problems. "Month", "Day", "Hour" and "Crime Type" are in string format. Integer value is assigned to each unique variable. "Hour", "Month" and "Year" are separated by parsing the "DateTime" attribute.

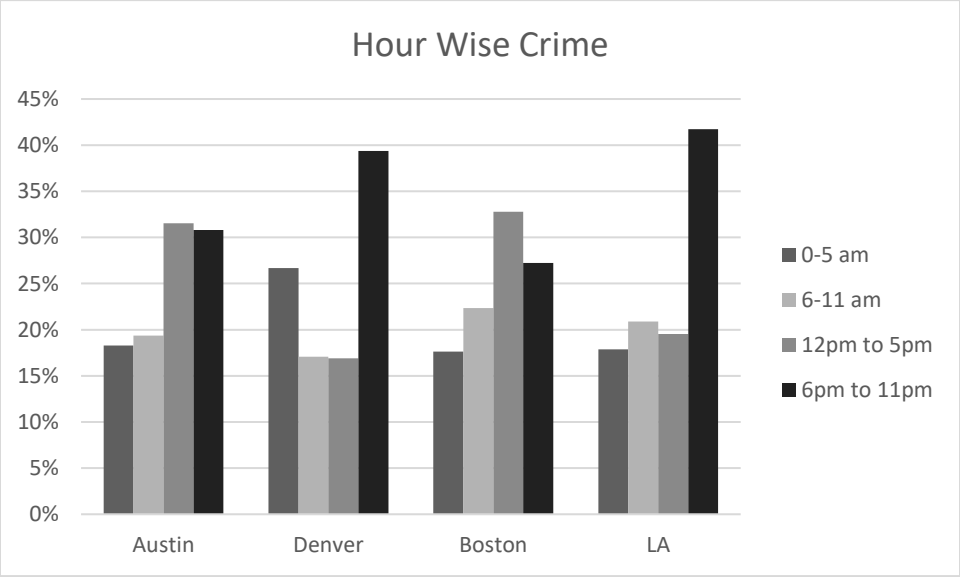
66% data is used for training classifiers while 33% is used for testing the models. Training datasets consists of all features along with class attribute which is type of crime. Testing data also consists of all features but there is no class attribute. Therefore, model will predict class attribute which will determine performance of models.

3.6 Extracting New Features:

Many new features are extracted from given datasets which will result in better training and prediction of data.

3.6.1 Hour:

Although, we have extracted hour from data but values ranging from 1 to 24 are not very meaningful. Therefore, we have divided hours into 4 types. Type one ranges from 12am to 5am. Type 2 is from 6am to 11am, type 3 is from 12pm to 5pm and type 4 is from 6pm to 11pm. In this way, crime time can become a better feature which will result in improvement in prediction of crime types. Following diagram pictorially depicts total crime in different hour slabs:



3.6.2 Literacy:

District wise literacy rate is incorporated in the datasets for all three cities. Literacy rate is present in the form of percentage for each district. Literacy rate is divided into different categories as

specified in the following table:

Table 1 Literacy rate

Literacy Rate	<50 %	50 to 60 %	60 to 70 %	70 to 80 %	80 to 85 %	85 to 90 %	90 to 100 %
Value	1	2	3	4	5	6	7

Below figure depicts literacy in each city. It is clear from following figure that Austin has highest literacy rate with 87% and Los Angeles has the lowest literacy rate with 75%. Whereas, literacy rate in Boston is 85% and Denver has literacy rate of 86%.

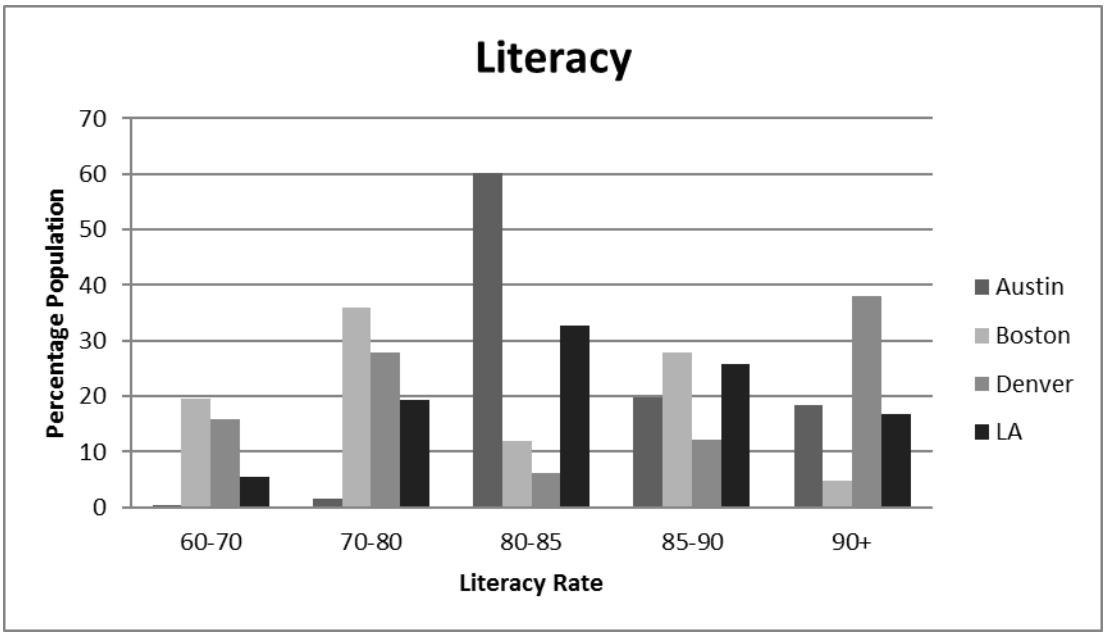


Figure 22 Literacy rate comparison

3.6.3 Tax:

Tax information for each district consists of number of people who filed tax returns in specified period. Tax information is transformed by categorizing it which is specified in below table:

Table 2 Tax returns

Tax Returns	<50 %	50 to 60 %	60 to 70 %	70 to 80 %	80 to 85 %	85 to 90 %	90 to 100 %
Value	1	2	3	4	5	6	7

3.6.4 Income:

District wise average income for each dataset is acquired from census data 2010 of USA. It is transformed as follows:

Table 3 Income Tax rate

Income	<$30,000	$30,000 to 40,000	$40,000 to $60,000	$60,000 to $80,000	$80,000 to $100,000	$100,000 to $150,000	$150,000 +
Value	1	2	3	4	5	6	7

4 Chapter 4: Experimental Results and Evaluation:

4.1 Overview

We have used 10 fold cross validation for evaluation of the models. 66% of data was used for training while 33% data is used for testing. Each tuple consists of different features including neighborhood, district, literacy, income, tax, date, time and hour of crime. Once model is developed, we have predicted type of criminal activity. We have computed True Positive (TP) rate, False Positive (FP) rate, Precision, Recall, F- measure and ROC Area to measure statistical relationships.

4.2 Model and Results

Decision Tree

Austin

Table 4 Austin Decision Tree

Class	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Theft	0.640	0.151	0.595	0.640	0.617	**0.782**
Assault	0.631	0.101	0.634	0.631	0.632	**0.792**
Disorder	0.644	0.080	0.630	0.644	0.637	**0.804**
White collar	0.352	0.025	0.360	0.352	0.356	**0.820**
Drug	0.406	0.042	0.379	0.406	0.392	**0.834**
Other	0.596	0.105	0.653	0.596	0.623	**0.787**

Table 5 Austin Decision Tree Confusion Matrix

Class	Theft	Assault	Disorder	White collar	Other	drug
Theft	15645	2430	2123	764	2393	1082
Assault	2756	13033	1516	454	1964	920
Disorder	2148	1364	10698	349	1464	584
White collar	1039	483	232	1265	540	36
Other	3344	2458	1997	629	14136	1157
Drug	1315	762	385	31	906	2320

Denver

Table 6 Denver Decision Tree

Class	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Theft	0.943	0.204	0.761	0.943	0.842	**0.875**
Assault	0.338	0.010	0.734	0.338	0.463	**0.654**
Disorder	0.755	0.050	0.750	0.755	0.752	**0.855**
White collar	0.178	0.005	0.702	0.178	0.284	**0.581**
Drug	0.535	0.021	0.736	0.535	0.620	**0.75**
Other	0.802	0.062	0.759	0.802	0.780	**0.872**

Table 7 Denver Decision Tree Confusion Matrix

Class	Theft	Disorder	Drug	Other	Assault	White collar
Theft	70724	1380	667	1802	273	132
Disorder	4457	22827	635	1833	312	160
Drug	4480	1612	9797	1863	363	182
other	4442	1519	738	28976	317	157
Assault	4418	1591	788	1868	4526	180
White collar	4432	1520	684	1814	374	1910

Boston

Table 8 Boston Decision Tree

Class	TP Rate	FP Rate	Precision	Recall	F-measure
Assault	0.775	0.031	0.810	0.775	0.792
Drug	0.764	0.016	0.810	0.764	0.786
Other	0.770	0.023	0.810	0.770	0.789
Disorder	0.785	0.042	0.810	0.785	0.797
Theft	0.864	0.122	0.810	0.864	0.836
White collar	0.762	0.012	0.810	0.762	0.785

Table 9 Boston Decision Tree Confusion Matrix

Class	Assault	Drug	Other	Disorder	Theft	White collar
Assault	21949	436	641	1186	3761	336
Drug	468	12185	356	659	2088	187
Other	663	343	17270	933	2959	264
Disorder	1122	581	853	29228	5009	447

Theft	2529	1309	1923	3561	65892	1009
White collar	367	190	279	517	1639	9562

Los Angeles

Table 10 Los Angeles Decision Tree

Class	TP Rate	FP Rate	Precision	Recall	F-measure
Assault	0.761	0.059	0.800	0.761	0.780
Drug	0.733	0.018	0.800	0.733	0.765
Other	0.732	0.016	0.800	0.732	0.764
Disorder	0.737	0.023	0.800	0.737	0.767
Theft	0.868	0.158	0.800	0.868	0.833
White collar	0.724	0.005	0.800	0.724	0.760

Table 11 Los Angeles Decision Tree Confusion Matrix

Class	Assault	Drug	other	Disorder	Theft	White collar
Assault	35535	798	734	1060	8353	217
Drug	946	12953	268	387	3045	79
other	877	269	12001	358	2821	73
Disorder	1224	376	346	16760	3940	102
Theft	5563	1710	1574	2273	76163	465
White collar	274	84	77	112	881	3749

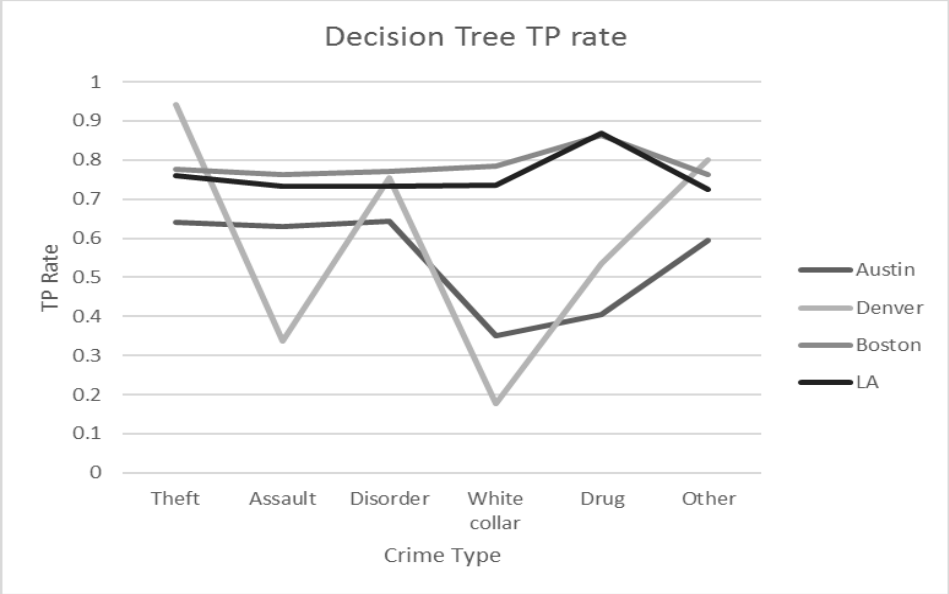


Figure 23 Decision Tree TP Rate Comparison

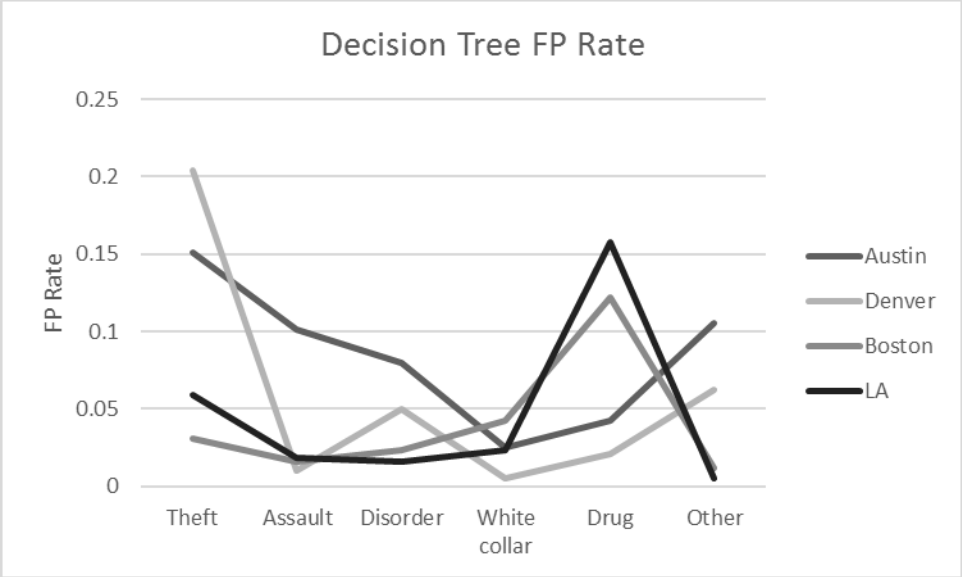


Figure 24 Decision Tree FP Rate Comparison

Naïve Bayes

Denver

Table 12 Denver Naive Bayes

Class	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Theft	0.941	0.190	0.770	0.941	0.847	**0.874**
Assault	0.423	0.010	0.720	0.423	0.546	**0.705**
Disorder	0.766	0.045	0.760	0.766	0.768	**0.861**

White collar	0.260	0.005	0.700	0.260	0.389	**0.625**
Drug	0.577	0.019	0.770	0.577	0.660	**0.777**
Other	0.813	0.059	0.770	0.813	0.791	**0.875**

Table 13 Denver Naive Bayes Confusion Matrix

Class	Theft	Disorder	Drug	Other	Assault	White collar
Theft	70704	1371	679	1801	366	187
Disorder	4222	23330	619	1799	318	173
Drug	4196	1414	10870	1790	377	186
other	4236	1354	674	29558	360	174
Assault	4211	1410	647	1721	5969	160
White collar	4524	1420	628	1718	362	2948

Austin

Table 14 Austin Naive Bayes

Class	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Theft	0.635	0.125	0.638	0.635	0.637	**0.820**
Assault	0.605	0.060	0.738	0.605	0.665	**0.829**
Disorder	0.658	0.066	0.678	0.658	0.668	**0.835**
White collar	0.519	0.037	0.355	0.519	0.421	**0.904**
Drug	0.646	0.072	0.363	0.646	0.465	**0.881**
Other	0.618	0.107	0.656	0.618	0.637	**0.819**

Table 15 Austin Naive Bayes Confusion Matrix

Class	Theft	Assault	Disorder	White collar	Other	drug
Theft	15650	1587	1843	1174	2528	1873
Assault	2423	12542	1299	738	2170	1556
Disorder	1798	886	10982	497	1508	1022
White collar	923	187	130	1894	458	59
Other	741	225	219	43	799	3703
Drug	53	27	23	23	243	35

Boston

Table 16 Boston Naive Bayes

Class	TP Rate	FP Rate	Precision	Recall	F-measure
Assault	0.686	0.045	0.730	0.686	0.708
Drug	0.673	0.023	0.730	0.673	0.700
Other	0.680	0.033	0.730	0.680	0.704
Disorder	0.698	0.059	0.730	0.698	0.714
Theft	0.802	0.165	0.730	0.802	0.764
White collar	0.670	0.017	0.730	0.670	0.698

Table 17 Boston Naive Bayes Confusion Matrix

Class	Assault	Drug	Other	Disorder	Theft	White collar
Assault	19781	620	910	1686	5345	477
Drug	665	10982	505	936	2967	265
other	942	487	15564	1326	4206	376
Disorder	1594	825	1212	26342	7118	636
Theft	3594	1860	2733	5061	59384	1433
White collar	522	270	397	734	2329	8617

Los Angeles

Table 18 Los Angeles Naive Bayes

Class	TP Rate	FP Rate	Precision	Recall	F-measure
Assault	0.705	0.074	0.750	0.705	0.727
Drug	0.673	0.023	0.750	0.673	0.709
Other	0.672	0.020	0.750	0.672	0.709
Disorder	0.677	0.029	0.750	0.677	0.712
Theft	0.831	0.190	0.750	0.831	0.789
White collar	0.663	0.006	0.750	0.663	0.704

Table 19 Los Angeles Naive Bayes Confusion Matrix

Class	Assault	Drug	Other	Disorder	Theft	White collar
Assault	33314	997	918	1326	10442	271
Drug	1183	12144	335	483	3806	99

Other	1096	337	11250	448	3526	92
Disorder	1530	470	433	15712	4925	128
Theft	6954	2138	1968	2841	71403	582
White collar	342	105	97	140	1102	3515

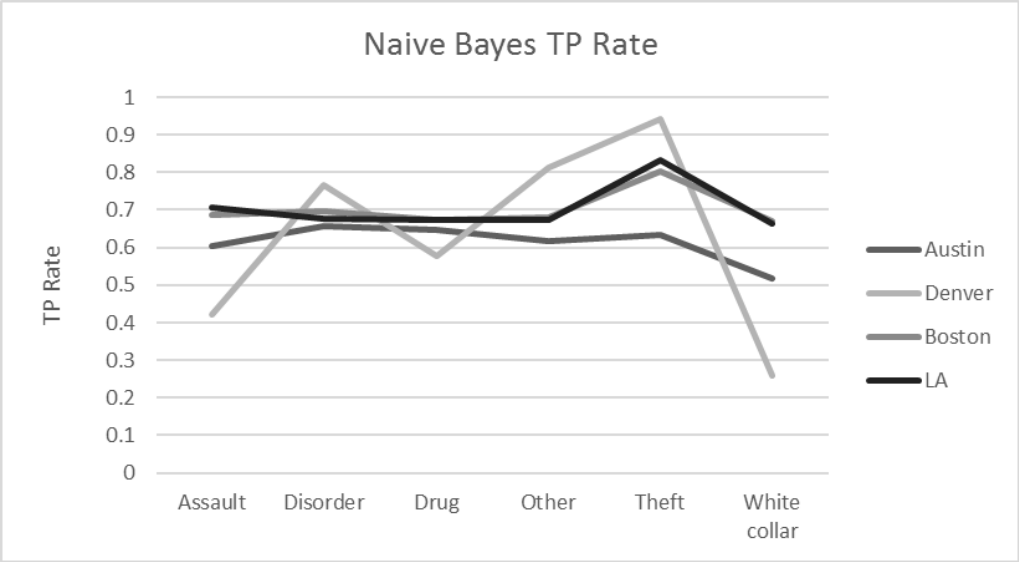


Figure 25 Naive Bayes TP Rate Comparison

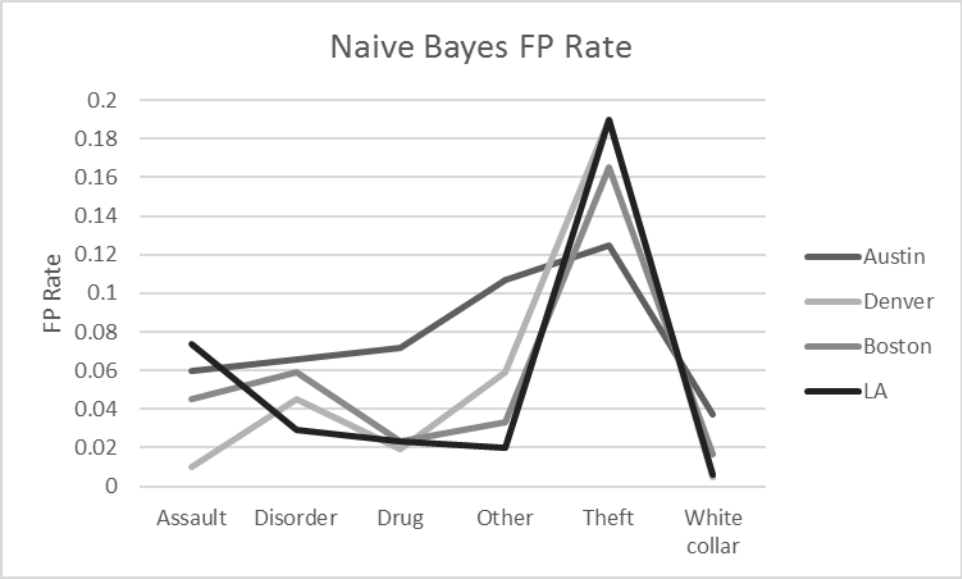


Figure 26 Naive Bayes FP Rate Comparison

ANN

Denver

Table 20 Denver ANN

Class	Theft	Disorder	Drug	Other	Assault	White collar
Theft	70706	1370	677	1801	366	188
Disorder	4222	23321	618	1798	323	179
Drug	4201	1417	10854	1792	380	189
other	4237	1356	674	29549	362	178
Assault	4215	1414	647	1725	5947	170
White collar	4262	1420	626	1724	365	2933

Table 21 Denver ANN Confusion Matrix

Class	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Theft	0.941	0.190	0.770	0.941	0.847	**0.876**
Disorder	0.766	0.045	0.770	0.766	0.768	**0.860**
Drug	0.576	0.019	0.770	0.576	0.659	**0.778**
Other	0.813	0.059	0.770	0.813	0.791	**0.877**
Assault	0.421	0.010	0.768	0.421	0.544	**0.706**
White collar	0.259	0.005	0.764	0.259	0.387	**0.629**

Austin

Table 22 Austin ANN

Class	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Theft	0.615	0.114	0.652	0.615	0.633	**0.812**
Assault	0.613	0.069	0.710	0.613	0.658	**0.822**
Disorder	0.657	0.066	0.678	0.657	0.667	**0.829**
White collar	0.364	0.027	0.352	0.364	0.358	**0.875**
Other	0.631	0.128	0.619	0.631	0.625	**0.810**
Drug	0.636	0.074	0.354	0.636	0.455	**0.862**

Table 23 Austin ANN Confusion Matrix

Class	Theft	Assault	Disorder	White collar	Other	drug
Theft	15173	1809	1844	855	3049	1925
Assault	2206	12711	1302	517	2407	1584

Disorder	1623	1008	10965	348	1733	1016
White collar	864	266	128	1330	991	71
Other	2632	1765	1702	680	15016	2010
Drug	731	300	220	33	804	3642

Boston

Table 24 Boston ANN

Class	TP Rate	FP Rate	Precision	Recall	F-measure
Assault	0.775	0.031	0.810	0.775	0.792
Drug	0.764	0.016	0.810	0.764	0.786
Other	0.770	0.023	0.810	0.770	0.789
Disorder	0.785	0.042	0.810	0.785	0.797
Theft	0.864	0.122	0.810	0.864	0.836
White collar	0.762	0.012	0.810	0.762	0.785

Table 25 Boston ANN Confusion Matrix

Class	Assault	Drug	Other	Disorder	Theft	White collar
Assault	21949	436	641	1186	3761	336
Drug	468	12185	356	659	2088	187
other	663	343	17270	933	2959	264
Disorder	1122	581	853	29228	5009	447
Theft	2529	1309	1923	3561	65892	1009
White collar	367	190	279	517	1639	9562

Los Angeles

Table 26 Los Angeles ANN

Class	TP Rate	FP Rate	Precision	Recall	F-measure
Assault	0.772	0.056	0.810	0.772	0.791
Drug	0.745	0.017	0.810	0.745	0.776
Other	0.744	0.015	0.810	0.744	0.776
Disorder	0.749	0.022	0.810	0.749	0.778
Theft	0.875	0.152	0.810	0.875	0.841
White collar	0.737	0.005	0.810	0.737	0.772

Table 27 Los Angeles ANN Confusion Matrix

Class	Assault	Drug	Other	Disorder	Theft	White collar
Assault	35979	758	698	1007	7936	206
Drug	899	13115	254	367	2893	75
Other	833	256	12150	340	2680	70
Disorder	1163	358	329	16969	3743	97
Theft	5285	1625	1495	2159	77115	442
White collar	260	80	74	106	837	3796

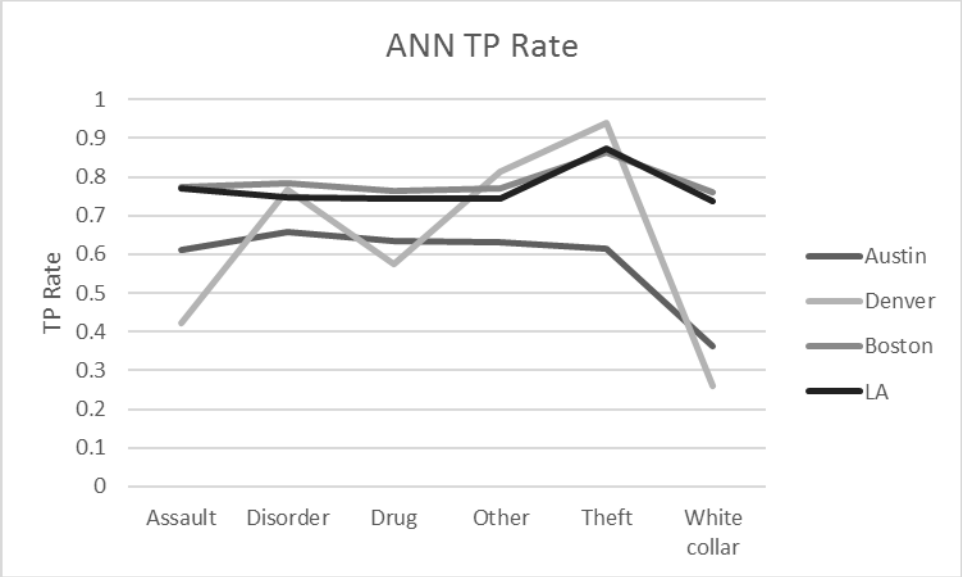


Figure 27 ANN TP Rate Comparison

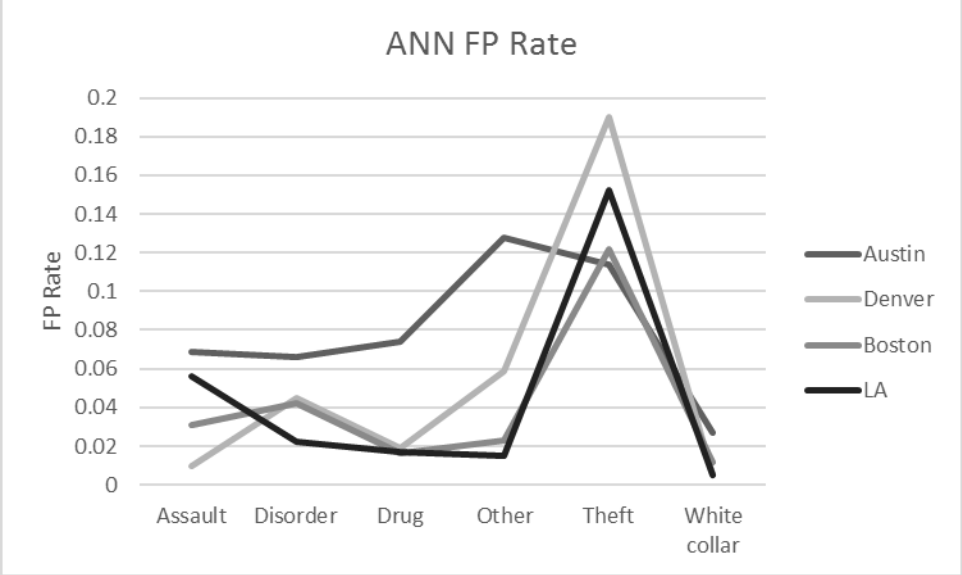


Figure 28 ANN FP Rate Comparison

KNN

Austin

Table 28 Austin KNN

Class	Theft	Assault	Disorder	White collar	Other	drug
Theft	15778	2431	2131	767	2428	1120
Assault	2765	13099	1504	450	1966	944
Disorder	2145	1352	10790	346	1462	598
White collar	1063	484	233	1284	550	37
Other	3355	2438	1992	621	14214	1187
Drug	1313	754	383	31	905	2344

Table 29 Austin KNN Confusion Matrix

Class	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Theft	0.640	0.151	0.596	0.640	0.617	0.790
Assault	0.632	0.100	0.636	0.632	0.634	0.799
Public Disorder	0.646	0.079	0.633	0.646	0.639	0.811
White collar	0.352	0.024	0.365	0.352	0.358	0.842
Other	0.597	0.105	0.653	0.597	0.624	0.792
Drug	0.409	0.043	0.375	0.409	0.391	0.849

Denver

Table 30 Denver KNN

Class	Theft	Disorder	Drug	Other	Assault	White collar
Theft	70774	1381	694	1802	323	134
Disorder	4472	23060	644	1814	326	145
Drug	4484	1581	10432	1814	367	155
Other	4471	1505	751	29162	342	125
Assault	4556	1553	772	1818	5299	120
White collar	4967	1493	681	1770	378	2041

Table 31 Denver KNN Confusion Matrix

Class	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Theft	0.942	0.207	0.755	0.942	0.838	0.877
disorder	0.757	0.048	0.754	0.757	0.756	0.862

Drug	0.554	0.021	0.747	0.554	0.636	0.777
Other	0.802	0.060	0.764	0.802	0.782	0.877
Assault	0.375	0.010	0.753	0.375	0.501	0.703
White collar	0.180	0.004	0.750	0.180	0.291	0.630

Boston

Table 32 Boston KNN

Class	TP Rate	FP Rate	Precision	Recall	F-measure
Assault	0.676	0.046	0.720	0.676	0.697
Drug	0.662	0.024	0.720	0.662	0.690
Other	0.669	0.034	0.720	0.669	0.693
Disorder	0.688	0.061	0.720	0.688	0.703
Theft	0.794	0.170	0.720	0.794	0.755
White collar	0.658	0.018	0.720	0.658	0.688

Table 33 Boston KNN Confusion Matrix

Class	Assault	Drug	Other	Disorder	Theft	White collar
Assault	19510	642	944	1748	5543	**495**
Drug	689	10831	524	971	3077	**275**
Other	977	506	15351	1375	4361	**390**
Disorder	1653	856	1257	25981	7381	**659**
Theft	3727	1929	2834	5248	58571	**1486**
White collar	541	280	411	762	2415	**8499**

Los Angeles

Table 34 Los Angeles KNN

Class	TP Rate	FP Rate	Precision	Recall	F-measure
Assault	0.661	0.087	0.710	0.661	0.685
Drug	0.627	0.026	0.710	0.627	0.666
Other	0.625	0.023	0.710	0.625	0.665
Disorder	0.631	0.033	0.710	0.631	0.668
Theft	0.801	0.214	0.710	0.801	0.753
White collar	0.616	0.007	0.710	0.616	0.660

Table 35 Los Angeles KNN Confusion Matrix

Class	Assault	Drug	Other	Disorder	Theft	White collar
Assault	31537	1157	1065	1538	12112	315
Drug	1372	11496	388	561	4415	115
Other	1271	391	10650	519	4091	106
Disorder	1775	546	502	14874	5713	148
Theft	8066	2480	2283	3296	67594	675
White collar	397	122	112	162	1278	3327

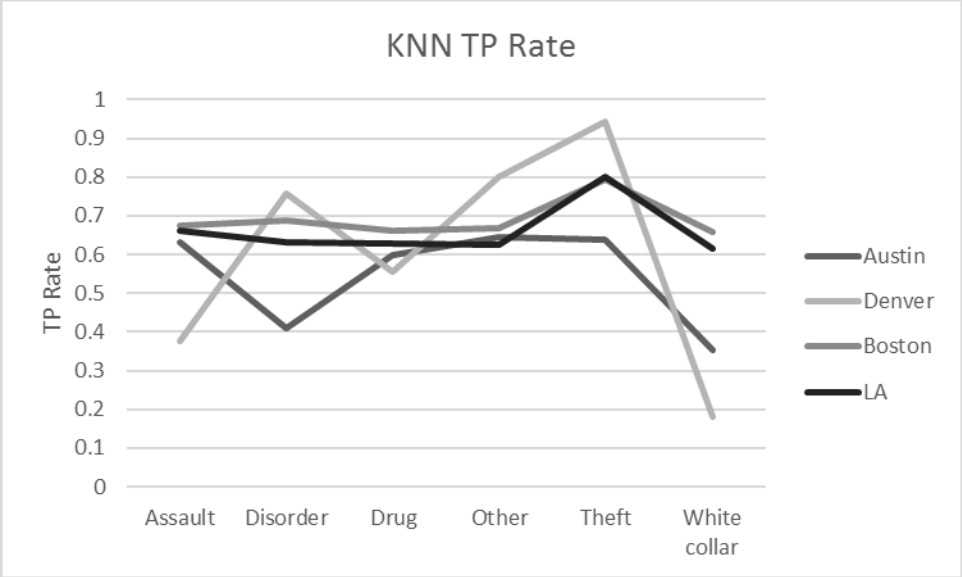


Figure 29 KNN TP Rate Comparison

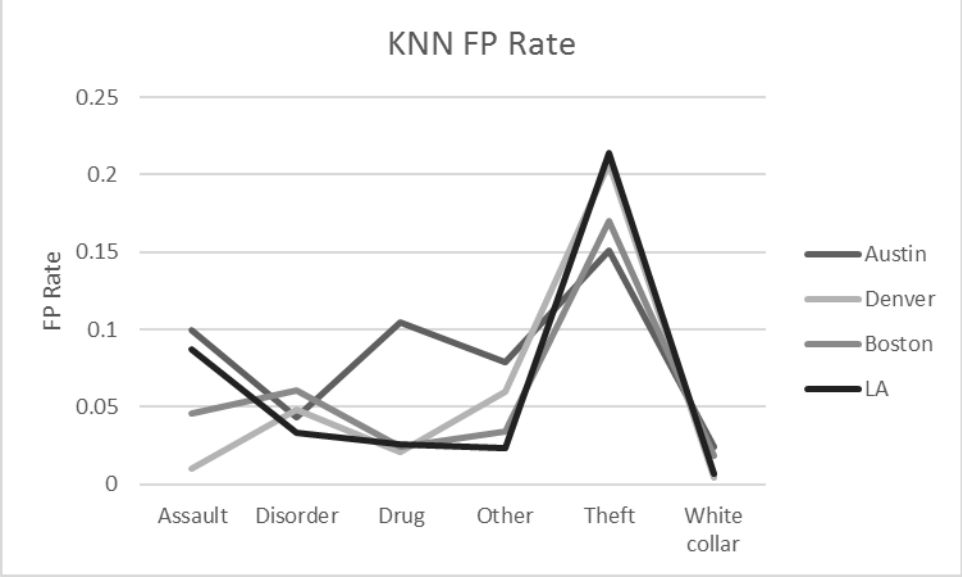


Figure 30 KNN FP Rate Comparison

SVM

Austin

Table 36 Austin SVM

Class	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Theft	0.638	0.137	0.621	0.638	0.629	0.751
Assault	0.599	0.058	0.744	0.599	0.664	0.770
Public Disorder	0.673	0.066	0.678	0.673	0.675	0.803
White collar	0.580	0.049	0.323	0.580	0.415	0.765
Other	0.557	0.053	0.773	0.557	0.647	0.752
Drug	0.777	0.099	0.334	0.777	0.467	0.839

Denver

Table 37 Austin SVM Confusion Matrix

Class	TP Rate	FP Rate	Precision	Recall	F-measure
Assault	0.674	0.010	0.760	0.674	0.715
Drug	0.680	0.020	0.760	0.680	0.718
Other	0.707	0.059	0.760	0.707	0.732
Disorder	0.697	0.045	0.760	0.697	0.727
Theft	0.841	0.189	0.760	0.841	0.798
White collar	0.671	0.005	0.760	0.671	0.713

Table 38 Austin SVM Confusion Matrix

Class	Assault	Drug	Other	Disorder	Theft	White collar
Assault	5892	153	483	362	1810	39
Drug	147	10728	880	658	3296	71
Other	400	756	29174	1790	8963	193
Disorder	316	597	1888	23027	7075	153
Theft	957	1808	5723	4283	69785	463
White collar	40	75	239	179	894	2909

Boston

Table 39 Boston SVM

Class	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Theft	0.943	0.204	0.761	0.943	0.842	**0.875**
Assault	0.338	0.010	0.734	0.338	0.463	**0.654**

Disorder	0.755	0.050	0.750	0.755	0.752	**0.855**
White collar	0.178	0.005	0.702	0.178	0.284	**0.581**
Drug	0.535	0.021	0.736	0.535	0.620	**0.75**
Other	0.802	0.062	0.759	0.802	0.780	**0.872**

Table 40 Boston SVM Confusion Matrix

Class	Assault	Drug	Other	Disorder	Theft	White collar
Assault	21678	459	674	1249	3959	354
Drug	492	12035	374	693	2198	196
other	698	361	17056	982	3115	278
Disorder	1181	611	898	28868	5272	471
Theft	2662	1378	2024	3749	65079	1062
White collar	386	200	294	544	1725	9444

Los Angeles

Table 41 Los Angeles SVM

Class	TP Rate	FP Rate	Precision	Recall	F-measure
Assault	0.764	0.033	0.800	0.764	0.782
Drug	0.753	0.017	0.800	0.753	0.776
Other	0.758	0.024	0.800	0.758	0.779
Disorder	0.774	0.044	0.800	0.774	0.787
Theft	0.857	0.127	0.800	0.857	0.827
White collar	0.750	0.013	0.800	0.750	0.774

Table 42 Los Angeles SVM Confusion Matrix

Class	Assault	Drug	other	Disorder	Theft	White collar
Assault	35979	758	698	1007	7936	206
Drug	899	13115	254	367	2893	75
other	833	256	12150	340	2680	70
Disorder	1163	358	329	16969	3743	97
Theft	5285	1625	1495	2159	77115	442
White collar	260	80	74	106	837	3796

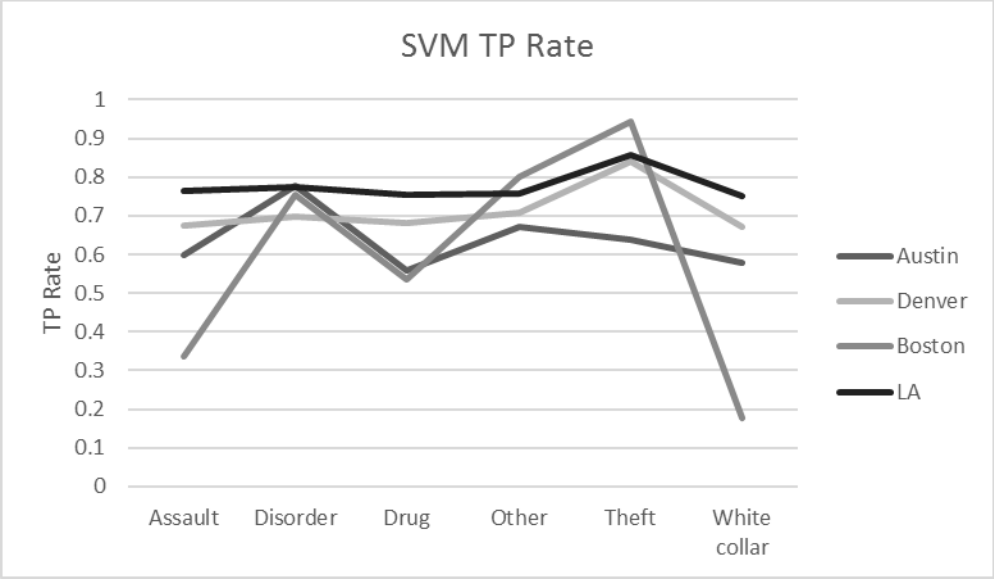


Figure 31 SVM TP Rate Comparison

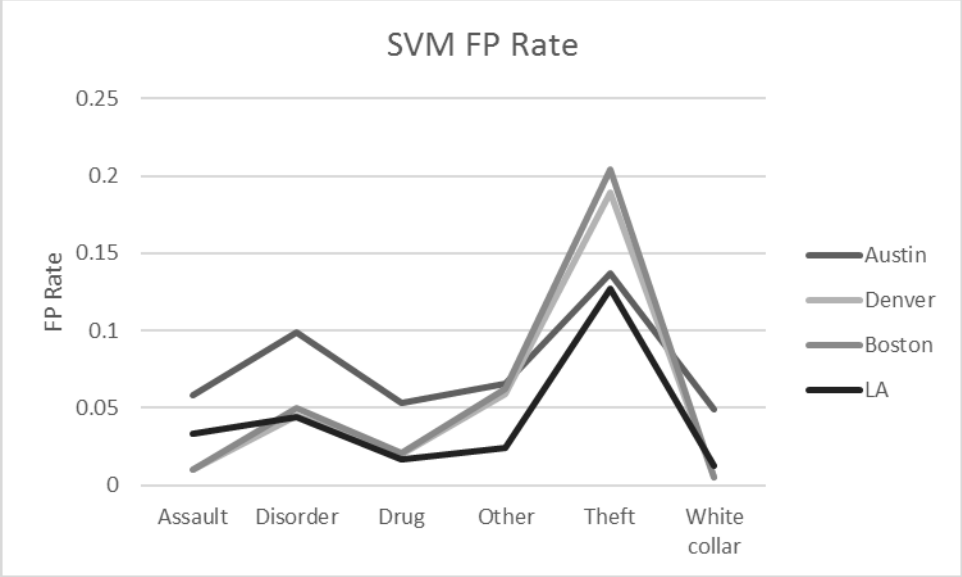


Figure 32 SVM FP Rate Comparison

Ensemble Method

Austin

Table 43 Austin Ensemble

Class	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Theft	0.813	0.204	0.771	0.943	0.842	**0.875**
Assault	0.338	0.010	0.794	0.338	0.463	**0.654**
Disorder	0.755	0.050	0.760	0.755	0.752	**0.855**

White collar	0.178	0.005	0.772	0.178	0.284	**0.581**
Drug	0.535	0.021	0.716	0.535	0.620	**0.75**
Other	0.802	0.062	0.799	0.802	0.780	**0.872**

Table 44 Austin Ensemble Confusion Matrix

Class	Theft	Disorder	Drug	Other	Assault	White collar
Theft	70724	1380	667	1802	273	132
Disorder	4457	22827	635	1833	312	160
Drug	4480	1612	9797	1863	363	182
other	4442	1519	738	28976	317	157
Assault	4418	1591	788	1868	4526	180
White collar	4432	1520	684	1814	374	1910

Denver

Table 45 Denver Ensemble

Class	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
Theft	0.943	0.204	0.761	0.943	0.842	**0.875**
Assault	0.338	0.010	0.734	0.338	0.463	**0.654**
Disorder	0.755	0.050	0.750	0.755	0.752	**0.855**
White collar	0.178	0.005	0.702	0.178	0.284	**0.581**
Drug	0.535	0.021	0.736	0.535	0.620	**0.75**
Other	0.802	0.062	0.759	0.802	0.780	**0.872**

Table 46 Denver SVM Confusion Matrix

Class	Theft	Disorder	Drug	Other	Assault	White collar
Theft	70724	1380	667	1802	273	132
Disorder	4457	22827	635	1833	312	160
Drug	4480	1612	9797	1863	363	182
other	4442	1519	738	28976	317	157
Assault	4418	1591	788	1868	4526	180
White collar	4432	1520	684	1814	374	1910

Boston

Table 47 Boston Ensemble

Class	TP Rate	FP Rate	Precision	Recall	F-measure
Assault	0.784	0.030	0.820	0.784	0.802

Drug	0.776	0.015	0.820	0.776	0.797
Other	0.781	0.022	0.820	0.781	0.800
Disorder	0.796	0.040	0.820	0.796	0.808
Theft	0.872	0.117	0.819	0.872	0.845
White collar	0.774	0.012	0.820	0.774	0.796

Table 48 Boston Ensemble Confusion Matrix

Class	Assault	Drug	Other	Disorder	Theft	White collar
Assault	22220	413	607	1124	3663	318
Drug	443	12336	337	624	1978	177
Other	628	325	17483	884	2804	250
Disorder	1063	550	808	29589	4745	424
Theft	2396	1240	1822	3374	66706	956
White collar	348	180	264	490	1552	9680

Los Angeles

Table 49 Los Angeles Ensemble

Class	TP Rate	FP Rate	Precision	Recall	F-measure
Assault	0.795	0.050	0.830	0.795	0.812
Drug	0.770	0.015	0.830	0.770	0.799
Other	0.769	0.014	0.830	0.769	0.798
Disorder	0.774	0.020	0.830	0.774	0.801
Theft	0.889	0.138	0.830	0.889	0.859
White collar	0.762	0.004	0.830	0.762	0.795

Table 50 Los Angeles Ensemble Confusion Matrix

Class	Assault	Drug	Other	Disorder	Theft	White collar
Assault	36868	678	624	901	7100	185
Drug	804	13439	228	329	2588	67
other	745	229	12450	304	2398	62
Disorder	1041	320	294	17388	3349	87
Theft	4729	1454	1338	1932	79019	396
White collar	233	72	66	95	749	3890

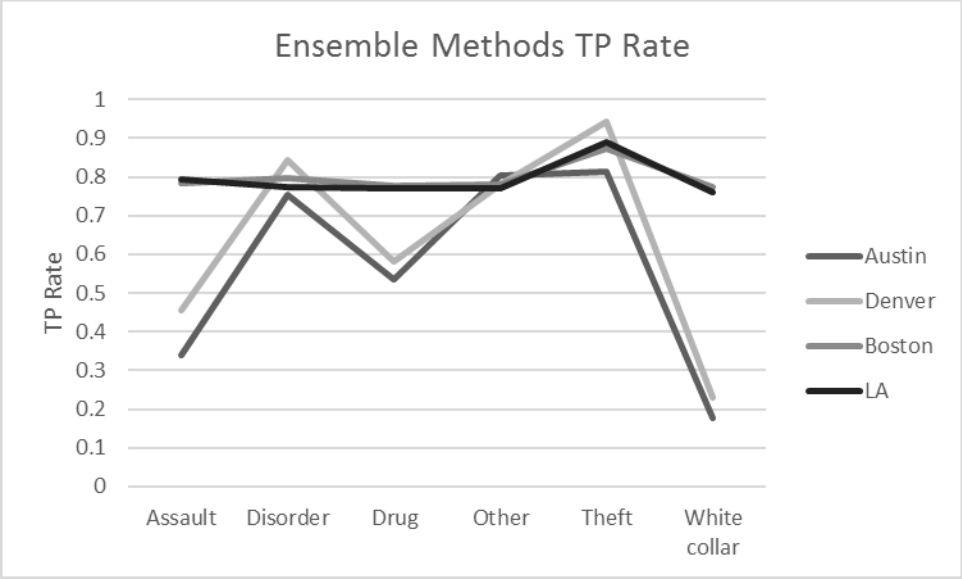


Figure 33 Ensemble Methods TP Rate Comparison

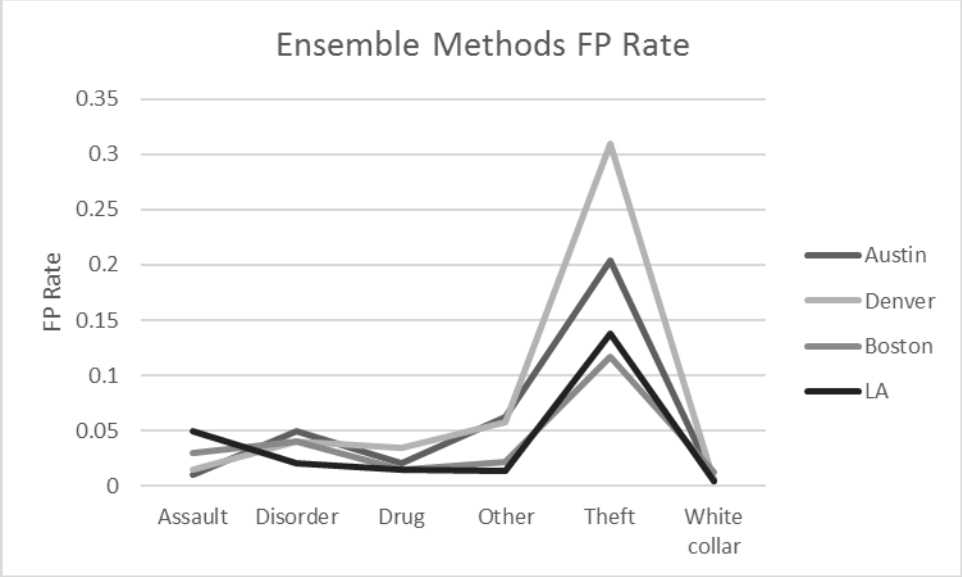


Figure 34 Ensemble Methods FP Rate Comparison

4.3 Evaluation

We have compared results of our model with existing state of the art crime prediction research. Below are the results of our model on different crime datasets using 10 fold cross validation:

Table 51 Evaluation of datasets

Data	Naïve Bayes	Decision Tree	SVM	Ensemble	KNN	ANN
Austin	69	75	77	77	73	76

Boston	73	81	80	82	72	81
Denver	71	77	76	79	70	77
LA	75	80	81	83	71	81

We can see from above table that results of ensemble methods are best with average accuracy of 80.25%. On the other hand, naïve Bayes results in least precision with only 72% average accuracy.

We have compared on results with Almanie et al [21] who have predicted type of criminal activity using different crime datasets of US cities. Below table provides average accuracy of different models on their crime data.

Table 52 Almanie et al crime prediction results

Data	Naïve Bayes	Decision Tree	SVM	Ensemble	KNN	ANN
Almanie data	42	44				
Almanie data results without new features	42	44	46	47	41	48

In the above table, we have used crime data from Almanie et al [21]. They had only used Naïve Bayes and Decision tree for crime prediction. We have also applied other techniques including SVM, Ensemble methods, KNN and ANN on their dataset. It is evident from above table that there is no significant increase in crime prediction accuracy using existing features.

5 Chapter 5: Conclusion and Future Work

Crime prediction plays a very important role for police and other law enforcement agencies. Various studies have been undertaken in this regard. Due to its complex and irregular nature, it is very hard to predict crime accurately. False negatives are very damaging in this regard. Therefore, contemporary models try to minimize false negatives.

We have proposed inclusion of literacy, income and other census information to determine and predict crime. This is very useful because these criteria play an important role in determining poverty and the tendency of criminal activities. Most of the sociological studies closely related crime with poverty. We are not only predicting whether crime will happen or not but we are also predicting the expected type of crime including assault, theft, white collar crime etc. We would like to further drill down criminal activity in future by including more crime types. For example, currently, we are classifying both homicide and murder in same category. We would like to separate them into different categories for better understanding and decision making.

Our study is only based on a few sociological parameters which include literacy, income, demographics etc. In future, we would like to include other parameters including nature of employment and user criminal history. Currently, we are only working on historical crime data for prediction of crime. We would also like to work on those people who have not committed any crime but they are identified as potential criminals by our research. This will further help law enforcement agencies to reduce criminal activities.

6 References

[1] S. Sivaranjani, Dr. Kumari. Aasha, "Crime Prediction and Forecasting in Tamilnadu using Clustering Approaches" Internation Conference on Emerging Technological Trends, 2016

[2] Nafiz Mahmud, Khalif Ibn Zinnah, Yeasin Ar Rahman, Nasim Ahmed, "CRIMECAST: A Crime Prediction and Strategy Direction Service" 19th International Conference on Computer and Information Technology, December 18-20, 2016, North South University, Dhaka, Bangladesh

[3] Mohammad Al Boni, Matthew S. Gerber, "Area-Specific Crime Prediction Models", 15th IEEE Conference on Machine Learning and Applications, 2016

[4] K.R. Sai Vineeth, Ayush Pandey, Tribikram Pradhan, "A Novel Approach for Intelligent Crime Pattern Discovery and Prediction", International Conference on Advanced Communication Control and Computing Technologies, 2016

[5] Qiang Zhang, Pingmei Yuan, Qiyan Zhou, Zhiming Yang, "Mixed Spatial-Temporal Characteristics Based Crime Hot Spots Prediction" 20th International Conference on Computer Supported Cooperative Work in Design, 2016

[6] Muhammad A. Tayebi, Martin Ester, Uwe Glasser, Patricia L. Brantingham, "CrimeTracer: Activity Based Crime Location Predictor" Advances in Social Networks Analysis and Mining, IEEE 2014

[7] Abba Babakura, Md Nasir Sulaiman, Mahmud A. Yousuf, "Improved Methods of Classification Algorithms for Crime Prediction" International Symposium on Biometrics and Security Technologies, 2014

[8] Keivan Kianmehr, Reda Alhajj, "Crime Hot-spots Prediction Using Support Machine" IEEE, 2006

[9] Shiju Sathyadevan, Devan MS, Surya Gangadharan, "Crime Analysis and Prediction using Data Mining" IEEE, 2014

[10] MD Rizwan Pervaz, Turash Musharaff, Mohammad Eunus Ali, "A Novel Approach to identify Spatio-Temporal Crime Pattern in Dhaka City" ICTD, 2016

[11] Cristina Kadar, Irena Pletilosa Cvijikj, "CityWatch: The Personalized Crime Prevention Assistant" MUM 25-28 November, 2014

[12] Mohammad A. Tayebi, Martin Ester, Uwe Glasser, "Spatially Embedded Co-offence Prediction Using Supervised Learning" KDD, August 24-27, 2014

[13] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Nuria Oliver, Fabio Painesi, Alex Pentland, "Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data" ICMI, November 12-16, 2014

[14] Shyam Varan Nath, "Crime Pattern Detection Using Data Mining" Internation Conference on Web Intelligence and Intelligent Agent Technology, 2006

[15] Hongjian Wang, Daniel Kifer, Corina Graif, Zhenhui Li, "Crime Rate Inference with Big Data" KDD, August 13-17, 2016

[16] Anna L. Buczak, Christopher M. Gifford, "Fuzzy Association Rule Mining for Community Crime Pattern Discovery" ISI-KDD, 25th July,2014

[17] Omowunmi Isafiade, Antoine Bagula, "Efficient Frequent Pattern Knowledge for Crime Situation Recognition in Developing Countries" ACM Dev, 4-7 December, 2013

[18] Chao Zhang, Manish Jain, Ripple Goyal, Arunesh Sinha, Milind Tambe, "Learning Prediction and Planning against Crime: Demonstration based on Real Urban Crime Data" 14th Internation Conference on Autonomous Agents and Multi-Agent Systems, May 4-8, 2014, Turkey

[19] Sunia Malik, Hammad Afzal, Imran Siddiqi, Awais Majeed, "Analyzing Socio-Economic and Geographical factors for Crime incidents using Heat Maps and Hot Spots" MedPRAI, 2016

[20] Xun Tang, Emre Efelioglu, Shashi Shekhar, "Elliptical HotSpot Detection: A Summary of Results" International Workshop on Analytics for Big Geospatial Data, 2015

[21] Tahani Almanie, Rsha Mirza, Elizabeth Lor, "Crime Prediction Based on Crime Types and Using Spatial and Temporal Criminal Hotspots" International Journal of Data Mining and Knowledge Maagement Process, 2015

[22] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2835847/

[23] Data.gov.in

[24] UCI Machine Learning Repository.: Available from: http://archive.ics.uci.edu/ml/datasets.html

(2012)
[25] https://www.census.gov/
[26] http://www.austintexas.gov/department/crime-information
[27] https://www.denvergov.org/content/denvergov/en/police-department/crime-information/crime-statistics-maps.html