|  |
| --- |
|  |

|  |
| --- |
| TECHNICAL REPORT  Electrical & Computer Engineering & Computer Science (ECECS) |

|  |  |
| --- | --- |
| FALL 2022 |  |



Contents

[Project Name 2](#_Toc96341549)

[Executive Summary 2](#_Toc96341550)

[Technical Report 3](#_Toc96341551)

[Highlights of Project 3](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341552)

[Submitted on: 3](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341553)

[Abstract 4](#_Toc96341554)

[Methodology 5](#_Toc96341555)

[Results Section 5](#_Toc96341556)

[Discussion 6](#_Toc96341557)

[Conclusion 6](#_Toc96341558)

[Contributions/References 7](#_Toc96341559)

|  |
| --- |
| US CRIME RATE DATA ANALYSIS |

|  |
| --- |
| Executive Summary The United States of America also known as the United States or informally America, is the most populous country in North America. Crime rates in the United States have been tracked since the country's inception and as such, it is host to every type of crime that you can possibly imagine. From terrorist attacks to multiple homicide incidents to violent sexual assaults and every property or drug crime under the sun, all of these and more can be found in the US. Public corruption, embezzlement, and other white collar crimes also take place frequently within different cities. City ordinance violations of many types as well as federal crimes like human trafficking, burglaries, rapes, theft are also found within the country. |
| person at a table writing in a notebook with people around  **Questions?**  **Team Members:** Contact:  Pavani Billapati [pbill@unh.newhaven.edu](mailto:pbill@unh.newhaven.edu)  Vishnupriya Doddapaneni [vdodd5@unh.newhaven.edu](mailto:vdodd5@unh.newhaven.edu)  Trilok Kumar tpidi1@unh.newhaven.edu Gowthami Middela gmidd1@unh.newhaven.edu |

|  |  |
| --- | --- |
| Crime exists in the US in various forms such as murder, theft, drug trafficking, money laundering, fraud, human trafficking, poaching etc. These crimes have different statistical trends that change over time. This project aims to provide a comprehensive picture of violent crime in the United States. We investigated overall trends and characteristics of crime, as well as crime rates, using Kaggle's crime data analysis dataset. Recent data, however, show an alarming increase in serious crime such as murder, rape, and so on. This project investigates the factors responsible for the increase in crime rates and concludes that race, gender, and age factors have played and continue to play an important role in crime and crime practices in the United States. The crime attributes in the dataset that could be predicted by applying various machine learning algorithms as considered by the FBI are Rape, Murder, Larceny, Robbery, Assault, Burglaries, Auto theft and Arsons. Other columns in the dataset include information about community names, county codes, community codes, percent of the population considered urban, age-based population, gender based population, race based population, and other factors that can be used to predict, and confusion matrix for the logistic regression.  TECHNICAL REPORT  dataset. In these where we are cleaning the dataset and loading into pandas data frame where feature selection by using the clustering algorithms and plotting the actual data and visualizing the data and splitting the data for finding the optimal number of clusters. In these we use the logistic regression where it is often used for research in order to calculate the predictive factors in the dataset. logistic regression is employed in data science as a supervised machine learning classification model. It can be useful in predicting category trends to within a high range of accuracy. Logistic regression is used for the train the dataset and visualizing the selected features by plotting their Histograms where used for graphical representation of data points organized into user specified ranges, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins. Histograms used for the plotting the features to analyze the label and its frequency using the age as the input The predictions made by logistic regression algorithm show the occurrence possibility of a crime whether a crime will occur or not, if a crime occurs, will it be a violent or a non-violent crime or if a crime occurs, is the cause of the crime murder or not. These predictions might help the local police departments as well as the FBI solve many cases with efficiency and accuracy. |  |
| Crimes prediction is one of the most important topics in recent years that aim to protect people's lives. These analytical studies for criminal hotspots are frequently demanded by law enforcement agencies hence, there is a huge requirement and demand for enhanced geographic information systems and innovative spatial data mining techniques in order to enhance crime detections and better protect their communities.  ABSTRACT  In this project, we propose a methodology to predict criminal patterns within the United States using a dataset from Kaggle. In order to achieve the study objectives, the methodology passes through several stages until the final results are reached, In the final stages we will get the appropriate estimation of the crime rate whether it is violent or not of the given state when we enter through the age and the race as the input in the website. The crime attributes in the dataset that could be predicted by applying various machine learning algorithms as considered by the FBI are Rape, Murder, Larceny, Robbery, Assault, Burglaries, Auto theft and Arsons.  In this project we use the logistic regression, it is a process of modeling the probability of a discrete outcome given an input variable. The most common [logistic regression models](https://www.sciencedirect.com/topics/computer-science/logistic-regression-model) a binary outcome; something that can take two values such as true/false, yes/no, and so on. Multinomial logistic regression can model scenarios where there are more than two possible discrete outcomes. Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category. |

Introductory Section

Day by day the crime rate is increasing considerably. Crime cannot be predicted since it is neither systematic nor random. Also the modern technologies and hi-tech methods help criminals in achieving their misdeeds. According to Crime Records Bureau crimes like burglary, arson etc. have been decreased while crimes like murder, sex abuse, gang rape etc. have been increased. Even though we cannot predict who all may be the victims of crime but can predict the place that has probability for its occurrence.

The predicted results cannot be assured of 100% accuracy but the results shows that our application helps in reducing crime rate to a certain extent by providing security in crime sensitive areas. So for building such a powerful crime analytics tool we have to collect crime records and evaluate it. It is only within the last few decades that the technology made spatial data mining a practical solution for wide audiences of Law enforcement officials which is affordable and available. Since the availability of criminal data or records is limited we are collecting crime data from various sources like web sites, news sites, blogs, social media, RSS feeds etc. This huge data is used as a record for creating a crime record database. So the main challenge in front of us is developing a better, efficient crime pattern detection tool to identify crime patterns effectively. The main challenges we are facing are: • Increase in crime information that has to be stored and analyzed.

• Analysis of data is difficult since data is incomplete and inconsistent.

• Limitation in getting crime data records from Law Enforcement department.

• Accuracy of the program depends on accuracy of the training set.

Review of available research

Since colonial times, violent crime in the United States has been declining over time. However, crime rates in the United States were greater in the early twentieth century than in regions of Western Europe. In 1916, for example, 198 killings were documented in Chicago, a metropolis of little more than 2 million people at the time. This amount of crime was not unusual in comparison to other American cities such as New York, but it was significantly higher in comparison to European cities such as London, which had three times the population but just 45 homicides in the same year.

For identifying trends in burglary, the Cambridge police department has developed a comparable tool called Series Finder in nations like England. To accomplish this, they studied the offender’s method of operation and gleaned several crime patterns that the perpetrator adhered to. The algorithm creates the offender’s modus operandi. A criminal’s “M.O”. is a set of behaviors that are used to identify a pattern of behavior.

The information included the point of entry (front door, window etc.), the day of the week, the property’s features (apartment, home), and its closeness to prior break-ins geographically. Most of the crimes within these patterns were retrieved by Series Finder using nine known burglary crime series, and nine additional crimes were also discovered. More than 80% of the expected outcome was accurate.

## 

## Methodology

The crime dataset we used contains real data and was obtained from Kaggle under the repository ‘Crime data’. The dataset features a significant number of numerical columns, and also contains some categorical data which we changed to numerical values using lab encoder since it makes it easier to apply algorithms. Rape, murder, larceny, robbery, assault, burglaries, autotheft, and arson are among the criminal features in the dataset.

Other columns in the dataset contain information on community names, county codes, percentage of the population deemed urban, age based population, gender based population, race based population, and other indicators that may be used to predict crimes. We utilized various characteristics from the dataset as predictors to train the model and constructed a binary label based on the selected features like county code, community code, age, percentage of unemployment and to predict the occurrence of the crime. We used logistic regression to construct the model.

The dataset was first loaded into the pandas data frame and then some columns were renamed for our convention. Many columns had the character ‘?’ which was replaced by 0 as a part of data cleaning. And, checks were placed to ensure that there were no ‘?’ values at all after the data was cleaned.

After carefully examining the information, we discovered that forecasting the occurrence of a crime may be a helpful and worthwhile use case. However, to do so, we needed to build a label called ‘violent crime occurrence’ that was based on the mean value from the column violent crimes per population. After calculating the mean and comparing it to the available data in the column ‘violentcrimeperpop’ a judgement of ‘yes’ or ‘1’ was made that a crime had occurred if the value in the relevant column was larger than the mean value, or ‘no’ or ‘0’ if the value was less than the mean. As, a result, a binary variable was formed.

**Results Section**

By using logistic regression, we performed data slicing, visualizing selected featured by plotting their histograms, plotting the features to analyze the label and its frequency and trained the model. The accuracy we obtained after testing was 72.5%. We further created confusion matrix to make further conclusion, where the model accuracy for logistic regression was 0.725.

Text

Description automatically generated

## 

## A screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generated

## 

Chart, treemap chart

Description automatically generated

## Discussion

We considered various factors like population, household, race, age, divorces, Occupation, Immigrants, House vacancy, Rent and other factors to determine violent crimes. We cleaned the data by replacing ‘?’ values with 0 to get proper accuracy. While executing the project we omitted the columns which did not contribute much to the prediction accuracy. We also incurred that the crimes were violent in states with more number of unemployment and where the population was high. We further noticed that the burglaries was happening more in the rich households whose income was relatively high.

## Conclusion

* In this project, crime analysis was performed using Kaggle dataset. Generally, the number of recorded offences decreases every year. There are significant differences between the states in terms of size and population i.e., they share different fractions of the population and area; therefore, different crime indicators were proposed to better compare different states(i.e., crime density and crime rate per 1000 citizens).
* We investigated the variables linked with various crimes using logistic regression. Further analysis indicates that the size of the observation region and other criteria must be improved to improve the prediction model.
* The local police force is expected to profit from this project by using the model in a more confined region while observing the same factors as mentioned in this project.
* A logistic regression model was developed to understand relation between serious crime rates and some demographic factors. Generally, crimes rates mostly decrease with the increase of the fraction of the old population.

## 

## Contributions/References

* <https://www-sup.stanford.edu/books/title/?id=2686>
* <https://www.researchgate.net/publication/230306257_An_alternative_psychology_of_criminal_behavior>
* <https://www-sup.stanford.edu/books/title/?id=2686>
* <https://usafacts.org/state-of-the-union/crime/>
* <https://en.wikipedia.org/wiki/Crime_in_the_United_States>