Chicago Crime Analysis

Arjun Ravikumar, Tejal Shanbhag , Utkarsh Havle

1. INTRODUCTION

Crimes are something that happens every day in all parts of the world. Some crimes are organized crimes and others are unorganized. There have been multiple risk factors that have been the underlying reason for these crimes happening. Three major categories of crime risk factors: (1) biological factors; (2) socioeconomic factors; and (3) psychological factors, as a result of the individual value system of those involved in criminal activity [1]. Chicago is a city that has a history of having one of the highest violent crime counts in the country. According to [5] Chicago has had 40,000 homicides from 1957 - 2020. For the first time since 1957, the homicides went below 500 for a year in 2019 in Chicago. This might not seem like a large number but we need to consider that these numbers are of just a city and not a state in the US. From the beginning of the 20th century, the Chicago Police department has been tracking the crimes happening in the city and making them accessible to the general public through their data website [4].

We propose to study these relationships between the socioeconomic factors of the residents of the area to the number of crimes happening in that area. The dataset we are utilizing for the same is of Hardship Index [6] and Crimes: 2001-present [2] from the City of Chicago's data portal. The Crimes: 2001-present dataset[2] consists of 7,299,208 rows of crimes between 2001 - February 18th 2021. This dataset[2] has 22 attributes which are id, case_number, date, block, iucr, primary_type, description, location_description, arrest, domestic, beat, district, ward, community_area, fbi_code, x _ coordinate, y_coordinate, year, updated_on, latitude, longitude and location. The hardship index data [6] consists of 78 rows and has a selection of six socioeconomic indicators of public health significance and a "hardship index". This data[6] has 6 attributes which are hardship_index, community_area_name, percent _ households _below _ poverty, percent_aged_25_without_high_school_diploma, percent_aged_16 _ unemployed, percent_aged_under _18_or _ over _64 and per_capita_income_. In this project, we will do data exploration where we try to visualize the data to analyse the trends in the data. We visualize the number of crimes against the year, day of the week, the hour of the day, location. Finally, we predict the severity of the crime using several classifies like naive bayes, random forest, decision tree and multi layer perceptron. Here we also add the fields like percent _ households _below _ poverty, percent_aged_25_without_high_school_diploma, percent_aged_16 _ unemployed, percent_aged_under _18_or _ over _64 and per_capita_income_ from [6] to identify if the addition of these features improve the accuracy of the prediction.

The rest of the study would be conducted as follows: Section 2 provides the motivation for choosing this topic. Section 3 provides information about the design of the project. Section 4 describes the implementation and analysis of the methods and data respectively in detail. Section 5 will describe inferences from the analysis followed by the future work in the project.

2. MOTIVATION

Chicago has always had a very bad record against the prevention of crimes. Likewise, Chicago is never considered a very safe city due to the number of crimes occurring in the city. In the day and age where armed shooting at a public place has become a regular column in the newspaper, we wanted to find out if there was any pattern in the crimes that occur. Theft is the most common crime in Chicago while theft can often be a case of organized well-planned crime it sometimes can be a crime out of necessity. In this project, we study if society and its factors like unemployment, lack of proper education, or crowded households can affect a community of people to commit crimes.

We have selected the city of Chicago for this study mainly because of the number of crimes occurring in Chicago. Chicago also keeps a record of all the individual cases that occurred and also the number of attributes provided for each crime is what made Chicago the only candidate for this study.

3. IMPLEMENTATION AND RESULTS

We divided the project into three main categories. The categories are Data Pre-Processing, Data Analysis, and Data Mining. Figure 1 describes the steps involved in the Data preprocessing phase.

Initially we performed preliminary statistical analysis on the collected datasets [6, 2] in order to understand the patterns and issues in the datasets[6, 2]. Then we checked for data quality issues like missing values, redundant records, checking of the importance of attributes and then eliminating them accordingly, etc. The cleaned datasets were then

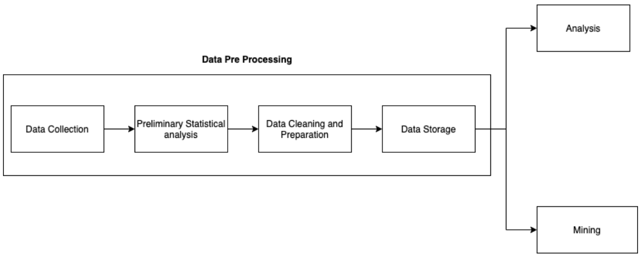
Data Pre Processing
Data Collection
Preliminary Statistical analysis
Data Cleaning and Preparation
Data Storage
Analysis
Mining

Figure 1: Project Architechture

used for knowledge discovery of the Crimes in chicago[2] and Mining.

3.1 Data Pre-Processing

Data preprocessing is an essential part of data mining. This step transforms the unprocessed raw data into an understandable format. Data in the real world has various issues like duplicate records, missing information, etc. Data pre-processing helps to solve these problems and make it usable to get future information from the existing records.

3.1.1 Phase 1 : Data Collection

For this project, the datasets Crimes in Chicago from 2001 to 2021[2], U.S Census Bureau. hardship index dataset [6] and Boundaries - Community Areas [3] was taken from the Chicago city data portal [4]. The crime dataset [2] had the records which is being updated everyday, while the hardship index dataset [6] was collected from 2008 to 2012 and the community area dataset [3] had the boundaries of all the community areas. The crime dataset[2] had in all 7,299,208 records with total 22 columns, the hardship index dataset [6] had in all 78 records with 9 columns and the community area dataset [3] had in all 78 records with 10 attributes. These three datasets [6, 2, 3] were collected in CSV format.

3.1.2 Phase 2 : Preliminary Statistical Analysis

In order to understand the datasets [6, 2] collected for this project, we performed pre processing statistical analysis on the data for understanding the datasets [6, 2] better. While doing the preliminary analysis of the raw data [6, 2] we identified multiple issues with it [6, 2].The datasets [6, 2] had following issues:

1. **Missing Records**

   Missing values were something that we were expecting when we had considered the raw dataset [6, 2]. Hence this is one of the first things which we looked for and found. We performed data visualization of the fields of both datasets [6, 2] to identify missing values. Missing values were easily identifiable after data visualization Figure 2 shows the missing data marked in white in the dataset [2] and complete data marked with black. We noticed that most missing data deals with the location of the incident. Most of the missing data we found were in fields like latitude, longitude, xcoordinate, ycoordinate, ward, and communityarea as seen in Figure 2. Since all of these were randomly missing we suspect these to be a case of missing at random (MAR). The second dataset did only have one missing value which was for the row which contained the census data for the whole of Chicago which is visible in Figure 3. Similar to Figure 2, in Figure 3 missing values are marked in white and complete data is marked in black.

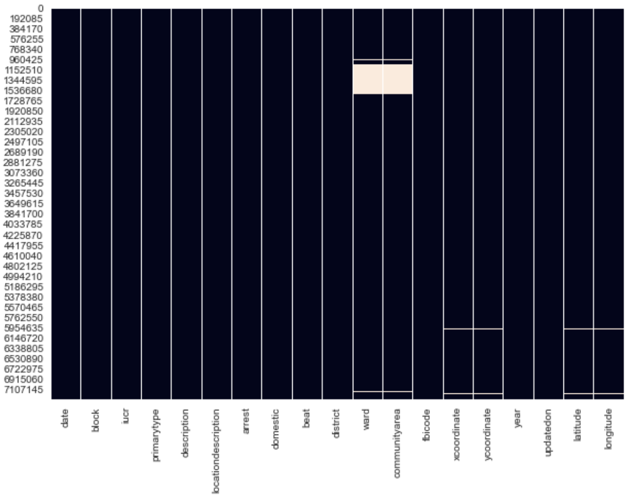

Figure 2: Crimes dataset [2] missing values

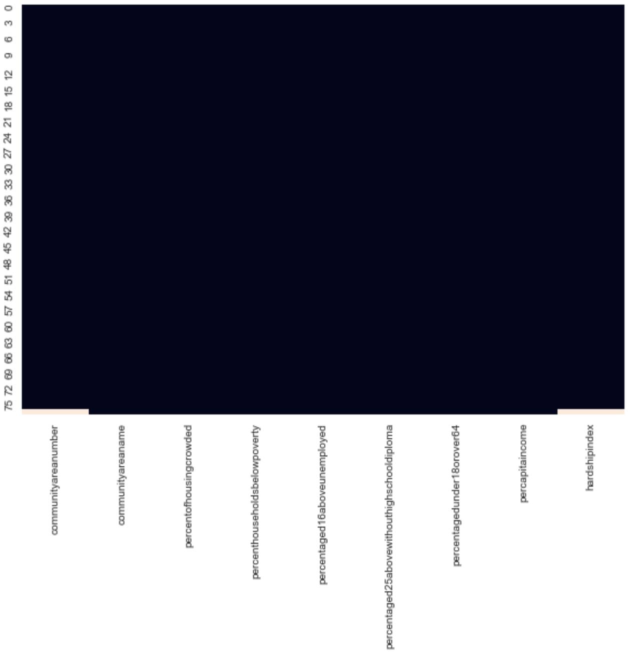

Figure 3: Hardship Index dataset [6] missing values

2. **Incorrect values**

   Incorrect values were tough for us to identify. These were not as evident to us as the missing values were but visualization helped us to identify these issues. We noticed that some of the crime districts are marked in locations of another district.

3.1.3 Phase 3 : Data Cleaning and Preparation

To perform the correct data cleaning steps, understanding issues with the collected data is important. We identified the issues in the datasets [2, 6] and have handled it in this step for data preparation.

1. **Unique primary key attributes**:

   There were some attributes in the datasets [2, 6] which had fully unique values in them for each row and were used for documentation purpose. Attributes like attributes ID, Case Number are as such attributes which were very specific to the each case and therefore did not provide any information which would help in data mining. Therefore we dropped these attributes. There was a duplicate column Location which had values of Longitude and Latitude coupled together in tuple format. As these values were redundant , we eliminated the Location column.

2. **Missing Values**:

   As mentioned in Phase 2 ,handling the missing values,was a primary concern for us. We first analysed the correlation among the missing attributes as seen in Figure 4 and then considered the next steps to rectify them accordingly. The dataset [2] had many missing values for the attributes telling the information about the location of the crime. From our study we found out that the most of the missing values are in ward and communityarea which had missing values for the same rows.
   There were 586329 missing values for ward and communityarea which was almost 8% of the total data. Since this was a significant number of values we could not just remove the rows which contained missing values. We couldn't also replace these values with adjacent values or with a constant value as we needed the communityarea field for joining the databases [2, 6]. Thus we had considered another dataset [3] specifically for filling the missing values. This dataset contained the communityarea and ward numbers and their boundaries in location. Next step was for us to find the communityarea and ward from the location given in [2]. We filled the communityarea and ward fields by checking the longitude and latitude of each incomplete row and then checked with all the community areas in the [3] dataset and found corresponding community area and wards. Once we did this then we found we could also approximate for the missing latitude and longitude from the same database [3] using the communityarea value from [2]. We approximated the missing latitude and longitude by finding the center of the communityarea in which it belongs and then further adding a random small value to the center to add some noise to the data. Using the above techniques we could fill all the missing values in these fields.

   From the Figure 4, the geographical attributes Longitude, Latitude, X coordinate, Y coordinate values are highly co-ordinated. The co-relation of 1 clearly tells that if any one of them is missing then others are missing as well.Therefore we removed the subset of these rows.

3. **Subdividing attributes**:

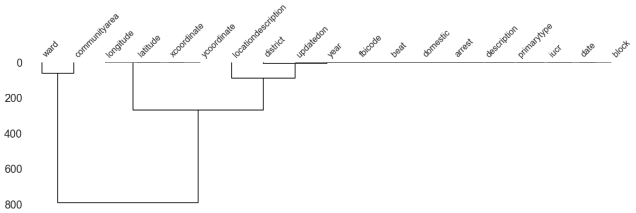

Figure 4: Attribute correlation in [2] dataset

   Each record in the [2] dataset had an attribute date associated with it. This attribute contains the date of the incident and the time at which the incident occurred. This was in a "month-day-year hour:minute:second" format, we intended to extract useful information from this attribute and create into multiple attributes. Therefore we created the attributes hour, day, month and weekday. We expected that some crimes will happen more often during then night hours than during day hours and some crimes will happen during particular days of the week. We did the same to find patterns in the data and also for classification task.

4. **Grouping to broader categories**:

   As there were different types of crimes in the dataset [2] we grouped them into broader categories. Similarly the attribute locationdescription describes the location where the crime occurred. In order to reduce the dimensionality of the data we grouped them into similar categories. Further for the classification we combined the crime types into two categories severe and non-severe. Here we considered "crimes involving physical harm to humans" and "crimes involving safety concerns to humans" as severe crimes and the remaining as non severe crimes.

5. **Encoding nominal data**:

   After the data visualisation step we converted the data into encoded numeric values for the classification. Attributes like crime type and location description were converted to encoded values. crime type was also made into two classes severe and non-severe as we wanted to perform binary classification on the data.

3.1.4 Phase 4 : Data Storage

We chose to save the data into a MySQL database after every process for easy access to data. After completing the initial data cleaning step of removing missing values and removing redundant columns we save the data into a MySQL database with two tables one for the crimes dataset and another for the hardship index dataset. This cleaned dataset from MySQL is used for data visualization. After which the MySQL database is accessed to encode the nominal values. Then this is saved into another MySQL database with two tables that have the data of the crimes before the hardship index values are joined with the crimes dataset and the second table with the crimes after the hardship index values are joined with the crimes dataset. This MySQL database is used only for the classification process, to compare the accuracy of the data with and without the hardship index values.

Common Category	Individual Values
Crimes involving physical harm to humans	homicide, kidnapping, human trafficking, offense involving children, battery, crime sexual assault, sex offense, assault, criminal sexual assault, domestic violence
Crimes involving monetary benefit	criminal damage, deceptive practice, burglary, motor vehicle theft, theft, robbery, gambling
Crimes involving safety concerns to humans	stalking, prostitution, intimidation, interference with public officer, obscenity, public indecency, arson
Crimes involving violations	weapons violation, concealed carry license violation, public peace violation, liquor law violation, narcotics, other narcotic violation, criminal trespass
Other non-criminal offenses	non-criminal, non - criminal, non-criminal (subject specified), other offense, ritualism

Table 1: Crime categories

3.2 Analysis

3.2.1 Crime Based Analysis

The pie chart in figure 5, shows the group of Crimes involving monetary benefits are the highest in Chicago. The lowest percentage is for crimes involving safety concerns to humans.

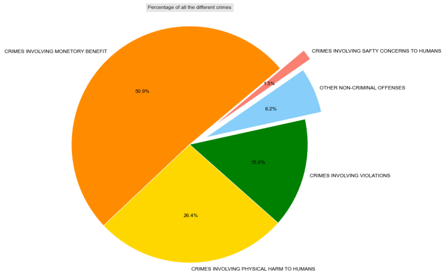

Figure 5: Percentage of crimes in Chicago using [2] dataset

The pie chart in figure 6, shows that 75% of criminals are not arrested.Though most of the criminals are not arrested, the crimes involving violations have the highest arrest rate. Whereas crimes which cause harm to humans have least arrest rates. Figure 7 tells the arrests happening for various crimes.

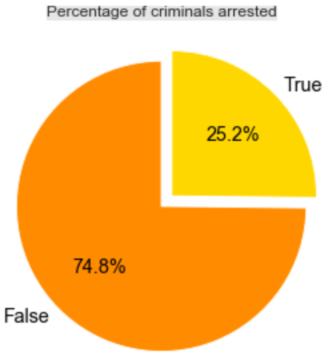

Figure 6: Percentage of criminals arrested

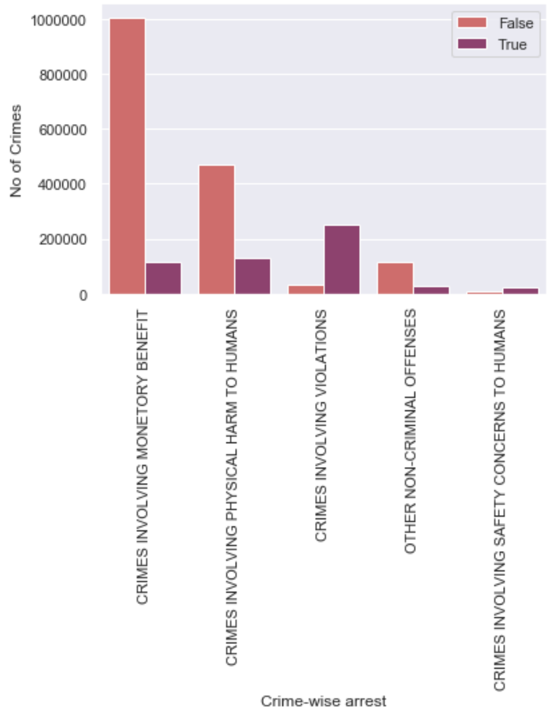

Figure 7: Crime Wise Arrest

3.2.2 Location Based Analysis

We analyzed the crimes occurring at various locations in the city. This analysis depicted that crimes done for monetary benefits like theft,motor vehicle theft were highest in the public areas and residential areas.Whereas the all the crimes were lowest in the Government locations. Figure 8 tells the counts of various crimes occurring at different locations.

The heatmap in figure 9 describes the various crimes occurring in different districts of Chicago. As from 9 Crimes are highest in district 8 and 18. District 31 has lowest crimes of all categories. Crimes involving monetary benefits and physical harm are highest overall in the city. Also district 11 can be termed as unsafes of them all because the count

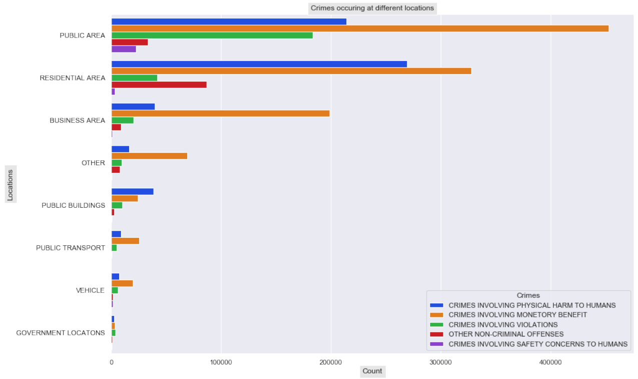

Figure 8: Crimes occurring at various locations using the [2] dataset

of all categories of crime are at higher rate.

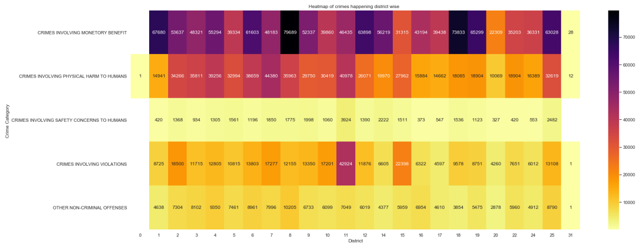

Figure 9: Heat map of crimes occurring district-wise

3.2.3 *Time Series Analysis*

We studied crimes occurring at different time intervals in the city of Chicago. From the figure 10, we can see that the crimes happening rises in the summer season with the peak in the month of July. Considering the climate of Chicago , the months of summer is the time when people tend to be out more than the winter months. So that might be the reason the count of crimes go up.

While analysing the crimes happening by hour, from figure 11 we can see that most of the crimes begin to happen in the evening and late at night. We can easily term the hours after sunset to be the unsafe hours in Chicago.

3.2.4 *Crime patterns over 24 hrs*

We oberved various crimes from all the different categories to check how the trend is over 24 hrs in the day.Following are the trends in each of them.

1. **Motor Vehicle Theft**

   As the crimes with monetary benefits are the highest, as shown in figure 5. We further analysed motor vehicle theft which was categorized to be part of crimes with monetary benefits. From fig 12, its observed that vehicle theft rises after evening 5pm with highest thefts occurring after 8pm till midnight.

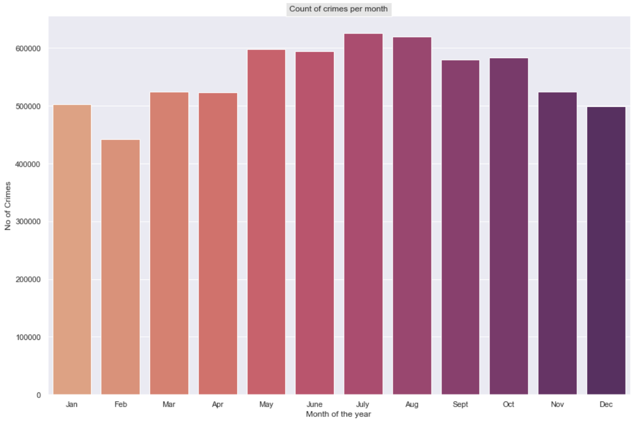

Figure 10: Crimes in Chicago month-wise using [2] dataset

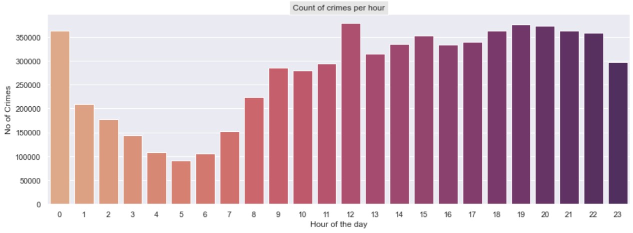

Figure 11: Crimes in Chicago month-wise using [2] dataset

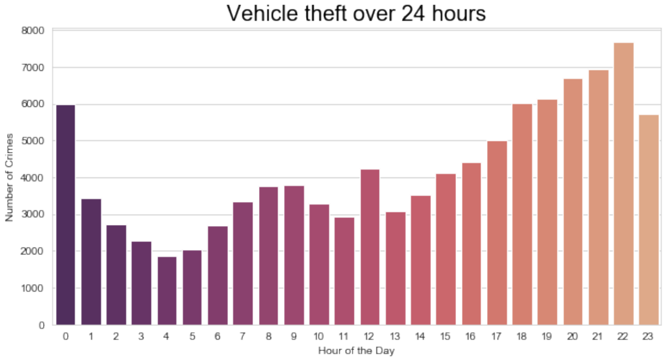

Figure 12: Vehicle theft in Chicago hour-wise using [2] dataset

2. **Interference with public officer** We also checked for the crime of interference with public officer round the clock. Looking at the figure 13 most of this crime starts at 8-9 am. This crime has its peak at the evening at 7pm and have a higher trend from 6pm to 11pm. We suspect that this is the time when most traffic is present and chances of a driving ticket is high around this time.

3. **Homicide** We also observed the trend Homicides over 24hrs. From figure 14 homicides have peak at midnight and are high until 2am - 3 am.Homicide also rises in the evening after 7 pm when its dark and this could be possible they are planned crimes

4. **Stalking**

   By analysing the trend of stalking in figure 15, it starts

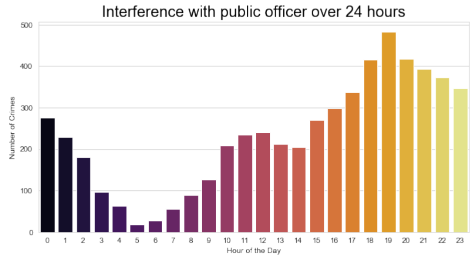

Figure 13: Interference with public officer Chicago hour-wise using [2] dataset

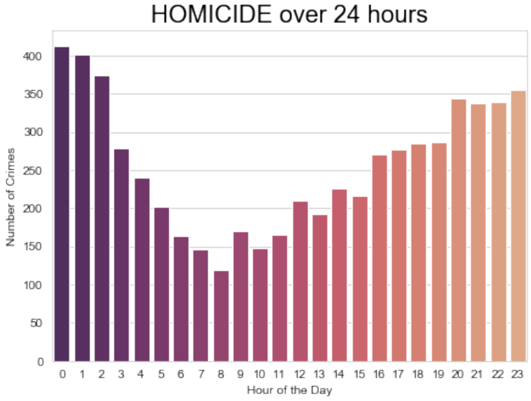

Figure 14: Homicides in Chicago hour-wise using [2] dataset

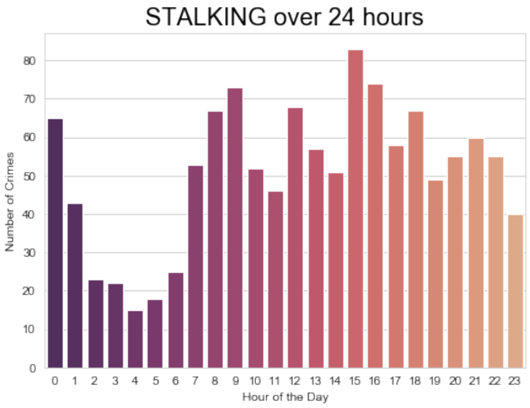

Figure 15: Stalking in Chicago hour-wise using [2] dataset

rising from the morning from 7am.It also rises at 12 pm, this hour can be considered as lunch hour and people go out take lunch so that's when stalking increases.Its at peak during afternoon at 3pm and thereafter it is on the higher end till midnight.

3.3 Classification

We aimed to predict the severity of the crime and to evaluate if the addition of hardship index will improve the accuracy of the prediction. We considered if a crime was severe or not by considering that any crime which harms a human physically or which affects the safety of humans is a severe crime and all the other crimes non-severe crimes. The attributes we used for the basic classification are mentioned in Table 2. These attributes were carefully selected considering that they do not have a one-to-one relation with the outcome and that they will provide some information in predicting the output. For the next phase, we included values from the dataset hardship index along with the attributes selected earlier. We matched the communityarea of each record and concatenated the attributes mentioned in Table 3 with the attributes in Table 2. Further, we used multiple classification algorithms to identify if concatenating the hardship index improved the performance of the models.

The train and test set was divided randomly within the crime dataset at an 80-20 ratio. Here 80% of the data is used for training the model and the remaining 20% is used for validating the model. Accuracy of the models after the classification process shows that Random Forest gave the maximum accuracy as in Table 4. The accuracy of the models after the classification also shows that in most cases the better accuracy was received when the hardship index was not concatenated with the crime dataset. Even though the difference between the accuracy of the models is very small the pattern of data without hardship index having a higher accuracy was generally observed over multiple executions.

4. LESSONS LEARNED

We started the project with a plan of removing any row which has a missing value. We did not understand the importance of the data until we tried to impute all the missing values in the data with accurate values from another dataset. This made the accuracy of the models improve by over 8%.

Attributes	Description
arrest	If an arrest was recorded or not
domestic	If the case is considered domestic-related or not
district	Police district where the crime occurred
ward	The ward where the crime occurred
communityarea	The community where the crime occurred
year	The year where the crime occurred
latitude	The latitude of location where the crime occurred
longitude	The longitude of location where the crime occurred
hour	The hour in 24 hours at which the crime occurred
month	The month at which the crime occurred
severity	If the crime was severe or not
location	The location at which the crime occurred
weekday	The day of the week in which the crime occurred
day	The day of the month in which the crime occurred

Table 2: Attributes from Crime [2] dataset used for classification

Attributes	Description
croudedhousing	Percentage of citizens living in a crowded home
belowpoverty	Percentage of citizens who are below the poverty line
unemployed	Percentage of citizens who are above 16 and below 64 who are unemployed
uneducated	Percentage of citizens who do not have a diploma
nonworkingage	Percentage of citizens who are above 16 and below 64

Table 3: Attributes from Hardship Index[6] dataset used for classification

Model	With hardship index	Without hardship index
Naive Bayes	67.34 %	77.56%
Random Forest	**78.52**%	78.51%
Decision Tree	68.73%	68.83%
Multi-layer Perceptron	77.53%	77.54 %

Table 4: Accuracy of different classification models

We understood the value of data cleaning and data preprocessing as without proper preprocessing techniques the classification task would have given much lower accuracy than the current accuracy. We have also discovered interesting patterns in the data when the data was visualized with different axes.

Although grouping the crimes and hardship index did not give any significant improvement in the accuracy of the model to predict severe and non-severe crimes, the inclusion of the community area dataset [3] had a significant improvement in the accuracy of the classification models. This also shows us that more data is always better.

5. CONCLUSION

We had started the project with three different datasets all of which were raw. One of the datasets [2] had a large number of missing values and redundant attributes. We started with removing all the redundant attributes and filled the missing values with appropriate values. This data was used for visualizing different patterns in the data we started the visualization and found interesting patterns in the data. After this step, we went to add hardship index dataset [6] to the original crime dataset [2] to see if the crimes could be justified with the financial state of the citizens in the area. However, we found that the addition of the hardship index didn't help much with the accuracy of the classification models. However, we found out that filling the missing values with accurate and appropriate values will boost the accuracy of the model instead of removing these rows completely.

6. CURRENT STATE AND FUTURE WORK

The model which we had created is a very limited model without considering many external factors. However, we know that crimes in a place are not only caused by one factor but multiple factors some of them may be political or international. As future work, we plan to explore more events that happened during the time in study and consider the effects of those in crimes. We also plan to further elaborate our study into other states and explore if there are any patterns of crimes that exist across multiple states. Crimes or justice violations in one state or place can very often spread across the rest of the country fast. This correlation of crimes among different neighboring and non-neighboring states is something which we plan to study further.

7. REFERENCES

[1] O.-R. Lobonţ, A.-C. Nicolescu, N.-C. Moldovan, and A. Kuloğlu. The effect of socioeconomic factors on crime rates in romania: a macro-level analysis. *Economic Research-Ekonomska Istraživanja*, 30(1):91–111, 2017.

[2] Chicago Police Department. Crimes - 2001 to present. `https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2`. Accessed: 2021-02-19.

[3] City of Chicago. Boundaries - community areas. `https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6`. Accessed: 2021-02-19.

[4] City Of Chicago. Chicago data portal. "https://data.cityofchicago.org". Accessed: 2021-02-19.

[5] Kyle Bentle & Jonathon Berlin & Ryan Marx & Kori Rumore. 40,000 homicides: Retracing 63 years of murder in chicago. https://www.chicagotribune.com/news/breaking/ct-history-of-chicago-homicides-htmlstory.html. Accessed: 2021-02-20.

[6] U.S Census Bureau. hardship index: Census data - selected socioeconomic indicators in chicago, 2008 – 2012. https://data.cityofchicago.org/Health-Human-Services/hardship-index/792q-4jtu. Accessed: 2021-02-15.