Analysis of Charlottesville Crime Data using PySpark

Manisha Sudhir
Charlottesville, USA
ms8jd@virginia.edu

Akhilandeshwari Goud Ranga
Charlottesville, USA
ar8aq@virginia.edu

ABSTRACT

As the population of a town increases, crime rates will increase too. Records have proved that crime rates increase by a small percentage every year. Regular occurrences of the crime incidents pose a serious threat to the safety and well-being of the society. This paper offers an extensive analysis of the crimes that have occurred in Charlottesville. The paper used Apache Spark to review and identify the properties of the crimes occurring using Crime data over the past 6 years and then seeks to predict the type of crime that may occur (given a set of features) using Machine Learning classification techniques. This paper intends to be of use to specialists and law enforcement officers in discovering patterns and trends for making forecasts, finding relationship and possible explanations.

INTRODUCTION

A Data set is simply data collected from different sources for some purpose which can be cleaned, analysed, transformed and modeled with the goal of finding useful information and supporting decision making. There are many tools and techniques available today to analyse data sets and draw meaningful conclusions from them. Today, data related to different areas of our life is collected and is readily available and one such data is that of Crime.

The exponentially increasing population leads to increase in crime and in turn generates huge chunk of data. The increasing crime rate is always a cause for concern and we should take advantage of such data and analyse it so that it helps the government in making critical and essential decisions in order to maintain law and order.

Analysing Crime Data has many advantages. The main advantage is that it gives us information like the unsafe areas, high crime rate time, highest occurring crime, etc. Crime Data can also be used to make Predictions. For example, guessing what crime is taking place given the exact day, time and location will help decide what kind of help is required. We explore these things in our Project on the Crime Data of Charlottesville city.

The paper has 3 parts :

ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

A. Performing exploratory analysis of the data to mine patterns in crime:

The main points that are analysed are:

1. Top 10 overall offenses. (Overall offenses occurring from most to least)
2. Count of offenses every hour.
3. Hour at which the offense count is the highest.
4. Top 20 streets with the highest offense count.
5. Offense count at each day of the week (Which day has the highest offenses.)
6. Offense count per month (Which month has the highest offenses)
7. The month that has the highest offense count every year from 2015-2018.
8. Hour that has the highest offense count for each day of the week.
9. The offense that happens the most for every hour of the day.
10. Total number of offense count per year from 2014 - 2019.

We have found that analysing the results of these statements can be extremely useful for understanding why crimes happen and when they happen.

B. Effect of external features (extracted from the following data) on the Charlottesville Crime Dataset:

1. Weather Data
2. US Federal Holidays Data

C. Building a prediction model to predict the type of crime that can take place in the city, in the future:

1. After observing the patterns of crime from the historical data as explained previous point, the next step would be to predict the crimes that can occur in the future.
2. Our goal is to build a prediction model that treats this problem as a multi-class classification problem by using various classification techniques in Machine Learning to classify the unseen data into one of the crime categories (classes) thereby predicting the crime that can occur.
3. This is expected to help the police plan their patrol and effectively contribute to building a smarter city.

For all three parts we use Apache Spark 2.4.4 and Python 3.7.5. For part B. we use NOAA weather data and US Federal Holiday Library for our features. For part C. we used Pyspark MLlib for various classification techniques which we will talk about in further sections of this paper.

RELATED WORK

Crime analysis is a law enforcement function that involves systematic analysis for identifying and analyzing patterns and trends in crime and disorder. Information on patterns can help law enforcement agencies deploy resources in a more effective manner. [4] Uses data mining techniques to analyze and visualise crime rates. The authors use Naive Bayes algorithm to create a model by training crime data related to vandalism, murder, robbery, burglary, sex abuse, gang rape, arson, armed robbery, highway robbery, snatching etc. [2] We also referred to a Medium article which also uses data mining techniques to classify Chicago data crime. The authors did their data analysis using Spark SQL and Haddoop. [1] Talks about different data mining techniques to analyse crime as well elaborates on a novel COPLINK methodology. A Master's thesis [5] talks about a log loss function along with several ML classification techniques for Analysis of San Fransisco Crime data. There isn't too much research in the area of analysis of crime data especially for the city of Charlottesville.

OVERVIEW OF THE DATA SET

The data used in this research project is the Charlottesville crime dataset made available by the Charlottesville Police Department on the Charlottesville Open Data website [3], which is a part of the open data initiative. The dataset consists of the following attributes:

·**RecordID** It is a numerical field. It is a unique identifier for each complaint registered. It is used in the database update or search operations.

·**Offense** It is a text field. Specifies the category of the crime. Originally, there are 122 distinct values (such as Assault, Larceny/Theft, Prostitution, etc.) in this field. It is also the dependent variable we will try to predict for the test set.

·**IncidentID** It is a numerical field. Denotes the incident number of the crime as recorded in the police logs.

·**BlockNumber** It is a numerical field. It denotes the block number/ area code of where the crime happened.

·**StreetName** It is a text field. Gives the street Name of where the crime occured.

·**Agency** It is a text field. It signifies which police department recorded the Offense occurrence. In this data set CPD is the value in all rows. CPD stands for 'Charlottesville Police Department.'

·**DateReported** It is a Date-Time field. Specifies the exact date of the crime and also provides the exact time stamp of when the crime was reported.

·**HourReported** It is a time field. It gives the hour and minutes at which the crime was reported.

This dataset consisted of 29,362 rows and the size of the dataset is approximately 0.2 MB. It consists of data from the year (October) 2014 to the year (October) 2019.

Features Added/Explored

Two additional features were Explored:

·**WeatherData** We imported the daily weather data from NOAA (National Oceanic and Atmospheric Administration). We obtained TMAX (Maximum Temperature), TMIN (Minimum Temperature) and SNOW (amount of snow in inches) from the dates of 1st October 2014 to 1st October 2019.

·**US Federal Holidays** We obtained the Federal Holidays data using the inbuilt pandas function "pandas.tseries.holiday.USFederalHolidayCalendar()"

PROPOSED METHOD

A. Analysis and Querying Data

We have written SQL queries to determine answers to the following questions:

1. Top 10 overall offenses. (Overall offenses occurring from most to least.
2. Count of offenses every hour.
3. Hour at which the offense count is the highest.
4. Top 20 streets with the highest offense count.
5. Offense count at each day of the week (Which day has the highest offenses.)
6. Offense count per month (Which month has the highest offenses)
7. The month that has the highest offense count every year frim 2015-2018.
8. Hour that has the highest offense count for each day of the week.
9. The offense that happens the most for every hour of the day.
10. Total number of offense count per year from 2014 - 2019.
11. Does Maximum and Minimum temperatures and Snow have any effect on prediction of Crime?
12. Do Holidays have any effect on the prediction of Crime?

We have found that asking these questions can be extremely useful for understanding the data and moreover, analysing why crimes happen and when they happen. The next step would be to see how accurately we are able to make predictions by training on the existing data.

B. Data Pre-processing

An important step before prediction is pre-processing of the Data. We used Apache Spark for data preprocessing. Using Apache Spark has a lot of advantages, especially in terms of distributed and parallel processing. It can also significantly decrease the processing time required to process such a huge volume of data. Although the use of Apache Spark does not necessarily make a huge difference for this dataset because of it's small size, this technique can be easily modified to work on data of a much larger scale spanning multiple distributed nodes.

The following were the various stages in the data pre-processing for this Project.

I. Data Cleaning

Data Cleaning includes checking if there are any missing values in the dataset and then handling them. We have handled the missing values in our Project in 2 ways. 1. Dropping the records having any missing values by using pyspark 'dropna' function. 2. Filling the missing values with the mean of the column. We have achieved this by using pyspark 'fillna' and 'approxQuantile' functions.

II. Extracting Features from other attributes

To use Date Features more meaningfully in the prediction, we have extracted the following features from the DateReported column of our dataset 1. DayOfWeek 2. DayOfMonth 3. MonthOfYear 4. Hour

III. Additional Features

We have added some additional features to our dataset. 1. First, we obtained weather data for 6 years (from 2014 to 2019 October). This included features like Maximum Temperature, Minimum Temperature and Snow everyday. 2. Next, we obtained the list of US Federal Holidays in these 6 years using pandas function. 3. Finally, we joined this data with the Charlottesville Crime Dataset and wrote everything into a new file.

IV. Data Transformation

Data transformation is one of the most important data preprocessing techniques. Usually, the data is originally present in the form that makes more sense if it is transformed. Data Transformation phase included the following steps: 1. One-Hot Encoding - Handling categorical data is important for some of the machine learning Algorithms like KNN. We have used pandas get dummies function to convert categorical columns into dummy columns where there is a column for each categorical value in the column. 2. Combining Features - We have used pyspark VectorAssembler to combine all the features into a single column. We have done this because pyspark MLlib machine learning functions need this as an input form performing classification. 3. Scaling - Some Algorithms like KNN which are dependant on distance require feature scaling. We have achieved this using pyspark MinMaxScaler function.

V. Grouping Labels

We grouped labels that belonged to the same subcategory together. Labels such as "Larceny- all other", "Larceny-from Motor Vehicles", etc. have been grouped into one label "Larceny". This reduced the number of Unique categories from 123 to 92 unique categories.

C. Prediction

In the next phase of our project we have used different Machine Learning techniques to predict the type of crime that has the possibility of occurring in the future given the features Block Number, Street Name, DayOfWeek, DayOfMonth, MonthOfYear, Hour, Snow and Holiday(Whether it is a federal Holiday or not).

Machine Learning Classification Models

The following are the machine learning classification models that were used to predict the type of crime given a set of input features:

1. Logistic Regression - Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. In regression analysis, logistic regression is estimating the parameters of a logistic model. Here logistic regression is used to predict the type of crime that could occur, given a specific feature set.

2. OneVsRest - The strategy of OneVsRest classifier is that it fits one classifier per class. For each classifier, the class is fitted against all the other classes. One advantage of this classifier is its interpretability.

3. Naive Bayes - Naive Bayes classification algorithm is a simple probabilistic classifier based on applying "Bayes theorem". It is a simple classifier which assumes that the attributes are conditionally independent to each other. The advantage of Naive Bayes classifier is that it is easy and fast to predict class of test data set.

4. Decision Tree - Decision Tree classifiers use decision trees for prediction and have many advantages. In a decision tree classifier, an input is tested against only specific subsets of the data, determined by the splitting criteria or decision functions. This eliminates unnecessary computations. Another advantage of Decision Trees is that we can use a feature selection algorithm in order to decide which features are worth considering for the decision tree classifier. The trees to make a prediction about the value of a target variable. The decision trees are basically functions that successively determine the class that the input needs to be assigned.

5. Random Forest - Decision Tree is built on the entire Dataset whereas Random Forest selects specific features to build multiple Decision Trees. A Random Forest Classifier generates multiple decision trees on different subsamples of the data while training, and then predicts the accuracy or loss score by taking a mean of these values. This helps to control over-fitting. In the Random Forest algorithm, the split for each node is determined from a subset of predictor variables which are randomly chosen at the given node.

6. Multi Layer Perceptron - A multilayer perceptron is similar to a logistic regression classifier but instead of feeding the input to the logistic regression, we feed it to an intermediate laye/hidden layer which has a nonlinear activation function like tanh or sigmoid.

D. Technologies and Platforms Used

Apache Spark

Apache Spark is a high performance, cluster-computing, open source framework for data analytics and is capable of handling big data. It is known for its ability to implicitly handle parallelism and for being fault tolerant. Spark's Python API, which is known as PySpark exposes the Spark programming model to Python. On top of the Spark core data processing engine are libraries for SQL, machine learning, graph computation, and stream processing. These libraries can be used together in many stages in modern data pipelines and allow for code reuse across batch, interactive, and streaming applications. Spark is useful for ETL processing, analytics and machine learning workloads, and for batch and interactive processing of SQL queries, machines learning inferences, and artificial intelligence applications.

We are using the PySpark API by Apache Spark. PySpark is a higher-level abstraction module over the PySpark Core. It is majorly used for processing structured and semi-structured datasets. It also provides an optimized API that can read the data from the various data source containing different files formats. The advantage of using Apache Spark is that, large amount of data is processed quickly and the existing code can be easily modified to support distributed big data in the future, if necessary.

Jupyter Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more

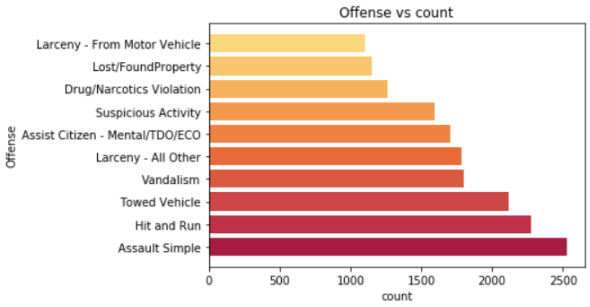
PERFORMANCE EVALUATION AND RESULTS

A. Performing exploratory analysis of the data to mine patterns in crime:

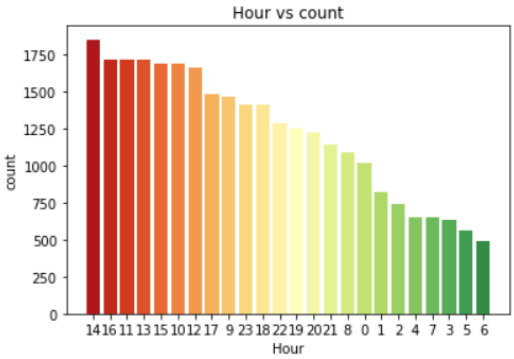
The following are the results of Analysis we have done on Charlottesville Crime Data. It shows the results for all the questions mentioned in the Analysis and Querying Data section

1. Top 10 overall offenses:

   The following is a graph showing the overall top 10 highest occurring offenses. From the figure we can see that 'Assault Simple' is the highest occurring offense over the past 4 years.



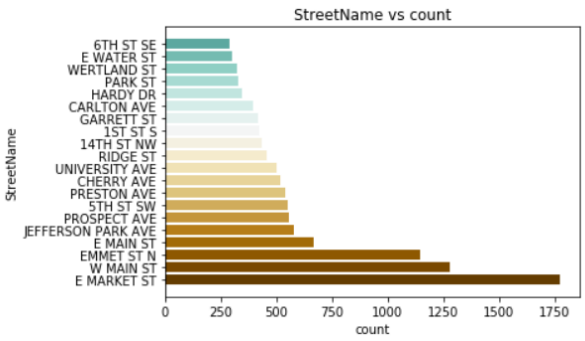
2. Count of offenses every hour: The following is a graph showing the offense count each Hour of the day in descending order.



3. Hour at which the offense count is the highest: From the results of the table we can see that the hour that has the highest offense count is hour 14 (i.e. 2pm) which has an offense count of 1853.

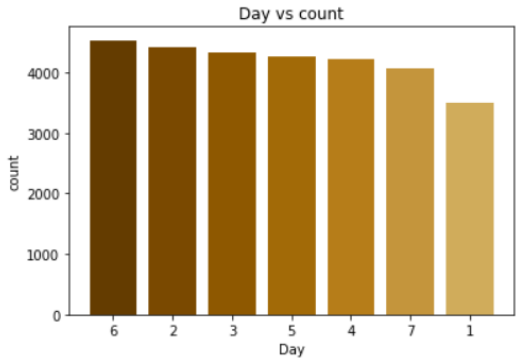
```
Offense happens most at Hour:
+----+-----+
|Hour|count|
+----+-----+
|  14| 1853|
+----+-----+
only showing top 1 row
```

4. Top 20 streets with the highest offense count: The following is a graph showing the count of offenses occurring in each street. The graph displays the offense count for top 20 streets in Charlottesville. E Market Street has the highest offense count of 1773.



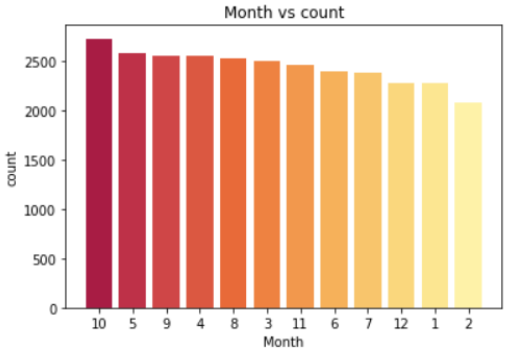
5. Offense count at each day of the week (Which day has the highest offenses): The following is a graph showing the offense count each Day of the Week in descending order. We can also see that the offenses happen most on day 6 which is Saturday. It has an Offense count of 4536.

```
Offense happens most on Day:
+---+-----+
|Day|count|
+---+-----+
|  6| 4536|
+---+-----+
only showing top 1 row
```



6. Offense count per month (Which month has the highest offenses): The following is a graph showing the offense count each Month of the Year in descending order. We can also see that the offenses happen most in October with an offense count of 2733.

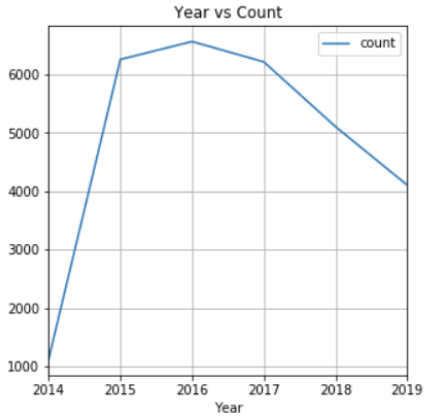
```
Offense happens most in Month:
+-----+-----+
|Month|count|
+-----+-----+
|   10| 2733|
+-----+-----+
only showing top 1 row
```



7. The month that has the highest offense count every year from 2015-2018: The following graph shows the month having the highest offense count in the years 2015 to 2018. The months September, October, October and May have the highest Offense count in the years 2015, 2016, 2017 and 2018 respectively.

```
Month that has most offenses in a year
+----+-----+-----+
|Year|Month|count|
+----+-----+-----+
|2015|    9|  600|
|2016|   10|  638|
|2017|   10|  598|
|2018|    5|  471|
+----+-----+-----+
```

8. Total number of offense count per year from 2014 - 2019. The following figure shows the trend of the number of offense occuring per year. 2019 and 2014 have a comparatively low offense count due to missing months in the data set for each of the years.



9. Hour that has the highest offense count for each day of the week. The following table gives the hour that has the highest offense count (along with the number of offenses that has occurred in that hour) for each day of the week. For example, on Monday (1), 4pm (hour 16) has the highest offense count of 210.

```
Hour that has most offenses during a Day in the week
+---+----+-----+
|Day|Hour|count|
+---+----+-----+
|  1|  16|  210|
|  2|  13|  322|
|  3|  14|  317|
|  4|  14|  313|
|  5|  14|  284|
|  6|  13|  283|
|  7|  23|  207|
+---+----+-----+
```

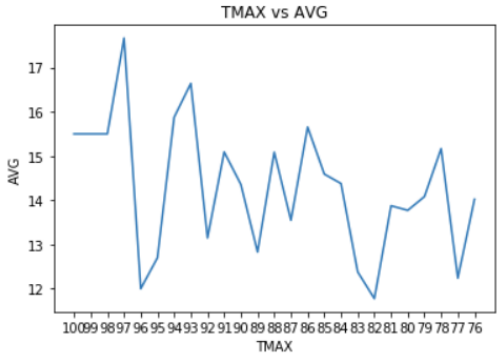
10. The offense that happens the most for every hour of the day: The following graph shows the Offense that happens the most every hour of the Day. Assault Simple happens the most between 7 pm and 5 am, Vehicles mostly get towed between hours 5 am and 11 am, Hit and Run happens most at hours 11 am to 2 pm and 3 pm to 6 pm.

Top offense every hour

Hour	Offense	count
0	Assault Simple	135
1	Assault Simple	124
2	Assault Simple	106
3	Assault Simple	81
4	Assault Simple	103
5	Assault Simple	91
6	Towed Vehicle	58
7	Towed Vehicle	105
8	Towed Vehicle	182
9	Towed Vehicle	190
10	Towed Vehicle	194
11	Towed Vehicle	221
12	Hit and Run	147
13	Hit and Run	147
14	Hit and Run	146
14	Larceny - All Other	146
15	Hit and Run	167
16	Hit and Run	176
17	Hit and Run	165
18	Hit and Run	153
19	Assault Simple	111
20	Assault Simple	135
21	Assault Simple	115
22	Assault Simple	151
23	Assault Simple	181

TMAX	AVG
100.0	15.5
99.0	15.5
98.0	15.5
97.0	17.666666666666668
96.0	12.0
95.0	12.7
94.0	15.875
93.0	16.64
92.0	13.147058823529411
91.0	15.090909090909092
90.0	14.355555555555556
89.0	12.829787234042554

32.0	9.333333333333334
31.0	9.4
30.0	11.0
29.0	7.666666666666667
28.0	9.166666666666666
27.0	12.5
26.0	7.25
25.0	16.666666666666668
24.0	3.5
23.0	9.5
22.0	9.0
19.0	12.5
17.0	13.0



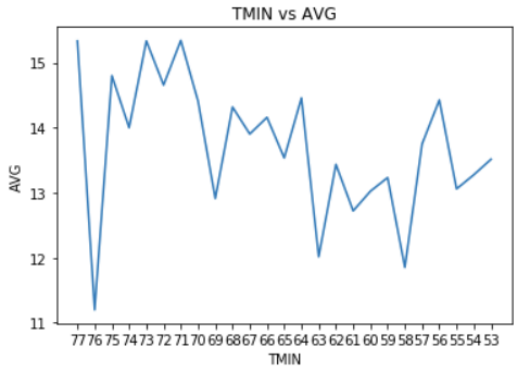
B. Effect of external features on the Charlottesville Crime Dataset:

This section shows the results of analysing the effect of the additional features added on the Offense count.

1. Effect of Maximum Temperature of the Day on Offenses: 1. The first table shows the top portion of the table showing the average Offense count at a certain Maximum temperature. 2. The second table shows the bottom portion of the table showing the average Offense count at a certain Maximum temperature. 3. The graph shows how the Average Offense count changes with decreasing temperature. We see that there is no strong pattern and we can infer that there is no strong correlation between maximum temperature of the day and the Offense count.

2. Effect of Minimum Temperature of the Day on Offenses: 1. The first table shows the top portion of the table showing the average Offense count at a certain Minimum temperature. 2. The second table shows the bottom portion of the table showing the average Offense count at a certain Minimum temperature. 3. The graph shows how the Average Offense count changes with decreasing temperature. We see that there is no strong pattern and we can infer that there is no strong correlation between Minimum temperature of the day and the Offense count.

TMIN	AVG
77	15.333333333333334
76	11.2
75	14.8
74	14.0
73	15.333333333333334
72	14.65625
71	15.341463414634147
70	14.407407407407407
69	12.912280701754385
68	14.317460317460318
67	13.902777777777779
66	14.158730158730158
16	9.571428571428571
15	12.7
14	9.0
13	9.2
12	10.153846153846153
11	16.75
10	10.333333333333334
9	9.0
8	5.333333333333333
6	12.333333333333334
1	7.0
-1	13.0



3. Effect of Federal Holidays on Offenses: The figure below shows the average Offense count on holidays and non-holidays. We see that the Offense count on non-federal holidays is slightly higher than on the federal holidays.

```
Average Offenses on non-Holidays: 16.20304568527919
Average Offenses on Holidays: 12.68
```

4. Effect of Snow on the Day on Offenses: The following table shows the average offense count on days when there is snow and on days where there is no snow fall. We infer that the offense count is pretty low on days when there is snowfall.

SNOW2	AVG
0	13.808206106870228
1	8.4

C. Results of building a prediction model:

Using various Machine Learning Classification techniques mentioned in Section C we are trying to predict the type of crime that can occur give a set of features. The following table shows the classification accuracies for two pre-processing techniques (i.e. Dropping rows that have null values and Filling out the columns that have null values). We split our dataset into 80 percent training and 20 percent testing.

Classification Algorithm	Classification Accuracy (Drop null values)	Classification Accuracy (Fill null values)
Logistic Regression	50.0467	51.417
Naive Bayes	49.2444	49.8038
Decision Tree	46.56	42.9678
Random Forest	44.7459	43.7147
Multi Layer Perceptron	48.36	47.1665
OneVsRest	50.13	51.0685

CONCLUSION & FUTURE WORK

Analysing Crime Data has many advantages. The main advantage is that it gives us information like the unsafe areas, high crime rate time, highest occurring crime, etc. In this research, a detailed analysis of various types on crimes in Charlottesville was conducted. We found some interesting insights in this data that was mentioned in this paper. Also, prediction models were trained using 6 machine learning classification techniques. Although accuracy rates are not high, we can further improve by further grouping labels and training. We can also try more pre-processing techniques. As a part of the future work, using the pre-processing done in this research, Neural networks can be trained and their results can be compared with the existing ones. It would help us see if there are more factors that contribute to the crime, like population data, housing data and transportation data to name a few.

REFERENCES

[1] Yi Qin-Michael Chau Jennifer Jie Xu Gang Wang Rong Zheng Homa Atabakhsh Hsinchun Chen, Wingyan Chung. Crime Data Mining: An Overview and Case Studies.

[2] https://medium.com/@stafa002/my-notes-on-chicago-crime-data-analysis ed66915dbb20.

[3] https://opendata.charlottesville.org/datasets/d1877e350fad45d192d23

[4] Shiju Sathyadevan, Devan S., and Surya Gangadharan. 2014. Crime Analysis and Prediction Using Data Mining. DOI:http://dx.doi.org/10.1109/CNSC.2014.6906719

[5] Isha Pradhan San Jose State University. 2018. Exploratory Data Analysis And Crime Prediction In San Francisco.