Analysis Of Classification Algorithms For Crime Prediction By Using Python

Talha Tariq
Department of Computer Science
Riphah International University,
Lahore, Pakistan
talha9828@gmail.com

Eman Qadeer
Department of Computer Science
Riphah International University,
Lahore, Pakistan
emanabid02@gmail.com

Iqra Mushtaq
Department of Computer Science
Riphah International University,
Lahore, Pakistan
iqramushtaq012@gmail.com

Alisha Mujahid
Department of Computer Science
Riphah International University,
Lahore, Pakistan
alishamujahid123@gmail.com

Amna Asghar
Department of Computer Science
Riphah International University
Lahore, Pakistan
amnaasghar15@gmail.com

***Abstract* — Dominating problem in our society is crime and its prevention is very important. Huge numbers of crimes committed frequently on daily basis. And there is need to make database of crimes that can be used for future references. Understanding of patterns in crime is very important to have complete knowledge about criminal activities. The objective of this paper is to expect which crime category is mostly to happen at which time and places in Chicago. In this paper we uses a Chicago_Crimes_2012_to_2017 dataset to train the algorithm. In first phase data preprocessing techniques are applied on the dataset. Then algorithms that is used for train the dataset are Random Forest Classifier and K-Nearest Neighbors classifier. After that performance of each algorithm is calculated and compare.**

Keyword: Crime Analysis, Classification Algorithms, Crime Prediction, Python, Random Forest, K-Nearest Neighbors

I. INTRODUCTION

Crimes are the significant danger for humankind. There are many crimes that happens regular interval of time. Perhaps it is increasing and spreading at a fast and vast rate. Crimes happen from small village, town to big cities. Crimes are of different types like false imprisonment, assault, battery, robbery, murder, rape, kidnapping and homicide. Since crimes are increasing day by day and there is a need to solve the cases as soon as possible. The crime activities have been increased more rapidly and it is the responsibility of police department to control and reduce the crime activities.

Fig. 1 Crime Analysis

Dominating problem in our society is crime and its prevention is very important. Huge numbers of crimes committed frequently on daily basis. And there is need to make database of crimes that can be used for future references. Understanding of patterns in crime is very important to have better response towards criminal activities..

The objective of this paper is to expect which crime category is mostly to happen at which time and places in Chicago. In this paper we uses a Chicago_Crimes_2012_to_2017 dataset to train the algorithm. In first phase data preprocessing techniques are applied on the dataset.

So, First step of proposed approach is data preprocessing. By using data mining tools and its classification techniques to classify the crime types. After that performance of each algorithm is calculated and compare. Now performance of each algorithm is calculated and outputs of these algorithms are compared.

II. LITERATURE REVIEW

Suliman et al presents a model for analysis of crime and criminal data using simple k-means algorithm for clustering of data and apply Aprior algorithm for data Association rules. 350 crime records used in this work. To preprocess and analyze the data, WEKA and Excel software were used. This work aims to help the Libyan government to identify the criminal behavior and avoidance of the high crime rate these days. The K-means algorithm shows a favorable results [1].

Almanie et al presents a model for interesting patterns for criminal hotspots using Aprori Algorithm and to predict the crime types Decision tree and NB classifier also used. Crime datasets for CO, Denver and Los Angeles was taken. Finally provide an analysis study by combining results of Denver crimes' dataset with demographics information [2].

McClendon et al uses WEKA for comparative study between Crime Unnormalized Dataset and the violent crime patterns from the Communities. And implementes the Additive Regression, Decision Stump and Linear

Regression algorithms. Linear regression algorithm Showed the best results among the other algorithms. This algorithm best for predicting violent crime patterns [3].

Rajkumar et al aims that mainly focusing on crime factors instead of crime occurrences. By applying Naïve Bayesian algorithm on the preprocess data create a predictive model that helps to predict the crime trends. In this paper analysis of some types of crime prediction and criminal using Data Mining techniques [4].

Fredrick et al presents the survey on crime prediction and crime analysis by using data mining techniques. The purpose of this paper that, for the criminal identification supervised and unsupervised learning techniques has been applied. And observed that data mining techniques can be applied for crimes identification [5].

Alkesh et al states that dominating problem in our society is crime and its prevention is very important. Huge numbers of crimes committed frequently on daily basis. And there is need to make database of crimes that can be used for future references. In this paper author uses Chicago crime dataset for crime type prediction. Data preprocessing is done bofore training the model and apply K-Nearest Neighbor (KNN) classifier and various other algorithms will be applied for crime prediction. The model predicts the crime type with accuracy of 78% [6].

Akash et al Uses existing datasets and implements different approaches to data mining analysis and prediction. Police stations and other criminal agencies maintain big databases of data and predict criminals built proceeding crime statistics. On the basis of subjects, we have used the model of data mining to predict the criminology and reasons behind the crime occurrences. The proposed algorithm with higher accuracy is able to predict more significant features [7].

Quader et al states that understanding of patterns in crime is very important to have better response towards criminal activities. The objective of this paper is to expect which crime category is mostly to happen at which time and places in Chicago. Crime dataset is taken from Chicago Police Department (2001-2017) to analyze the patterns and implements different classifiers like Decision Tree Random Forest, and different methods such as AdaBoost, Extra Trees and Bagging to calculate the accuracy given by each classifier [8].

III. RESEARCH QUESTION

In [8], authors uses Python data mining tool. And use its classification techniques namely Decision Tree, Random forest and several ensemble methods such as Bagging, AdaBoost and ExtraTree Classifier technique to predict the crime types that might occur based on location and time. The main purpose of this paper is to use different classifiers on crime datasets for classification of the crime types occurring based on location and time.

- The first limitation of their work they performed experiment on 17 years (2001 to 2017) data.
- There is need to check that which attributes correctly classify the crime types.
- There is need to find efficient classifiers to classify the crime types correctly.

IV. METHODS AND MATERIALS

This paper uses a Chicago_Crimes_2012_to_2017 dataset for training. The algorithms that are used for classification are Random Forest Classifier and K-Nearest Neighbors classifier. In first phase data preprocessing techniques are applied on the dataset.

- Remove Null Values .
- Remove irrelevant/not meaningfull attributes classification.
- Splitting the Date to Day, Month, Year, Hour, Minute, Second.
- Convert Categorical Attributes to Numerical.
- Encode target labels into categorical variables.
- Feature Selection using Filter Method.
- Split Dataframe to target class and features.
- Split dataset to Training Set & Test Set 80/20%.

And technique is applied on dataset by using different classification algorithms like K-Nearest Neighbours and Random Forest. After that performance of each algorithm is calculated and compare. Now performance of each algorithm is calculated and outputs of these algorithms are compared.

A. Methodology

The figure 1 shows the methodology to achieve the objectives of this paper.

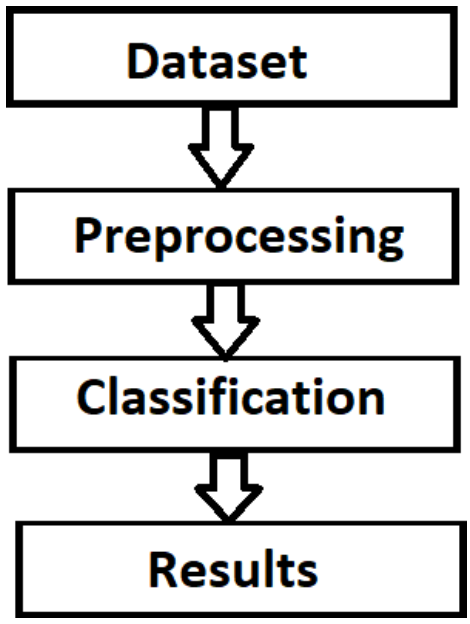

Fig. 2 Depicts the methodology of work

B. Data Collection

Chicago_Crimes_2012_to_2017 dataset is taken from kaggle repository.

```
In [4]: df = pd.read_csv('Chicago_Crimes_2012_to_2017.csv')
        df.head()
```

Out[4]:

	Unnamed: 0	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest
0	3	10508693	HZ250496	05/03/2016 11:40:00 PM	013XX S SAWYER AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE	APARTMENT	True
1	89	10508695	HZ250409	05/03/2016 09:40:00 PM	061XX S DREXEL AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE	RESIDENCE	False
2	197	10508697	HZ250503	05/03/2016 11:31:00 PM	053XX W CHICAGO AVE	0470	PUBLIC PEACE VIOLATION	RECKLESS CONDUCT	STREET	False
3	673	10508698	HZ250424	05/03/2016 10:10:00 PM	049XX W FULTON ST	0460	BATTERY	SIMPLE	SIDEWALK	False
4	911	10508699	HZ250455	05/03/2016 10:00:00 PM	003XX N LOTUS AVE	0820	THEFT	$500 AND UNDER	RESIDENCE	False

5 rows × 23 columns

Fig. 3 Dataset

```
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1456714 entries, 0 to 1456713
Data columns (total 23 columns):
Unnamed: 0              1456714 non-null int64
ID                      1456714 non-null int64
Case Number             1456713 non-null object
Date                    1456714 non-null object
Block                   1456714 non-null object
IUCR                    1456714 non-null object
Primary Type            1456714 non-null object
Description             1456714 non-null object
Location Description    1455056 non-null object
Arrest                  1456714 non-null bool
Domestic                1456714 non-null bool
Beat                    1456714 non-null int64
District                1456713 non-null float64
Ward                    1456700 non-null float64
Community Area          1456674 non-null float64
FBI Code                1456714 non-null object
X Coordinate            1419631 non-null float64
Y Coordinate            1419631 non-null float64
Year                    1456714 non-null int64
Updated On              1456714 non-null object
Latitude                1419631 non-null float64
Longitude               1419631 non-null float64
Location                1419631 non-null object
dtypes: bool(2), float64(7), int64(4), object(10)
memory usage: 236.2+ MB
```

Fig. 4 Dataset info

C. Preprocessing

In first phase data preprocessing techniques are applied on the dataset.

- Remove Null Values and irrelevant/not meaningfull attributes classification.

```
In [6]: # Preprocessing
        # Remove NaN Value
        df = df.dropna()
```

```
In [7]: # Remove irrelevant/not meaningfull attributes
        df = df.drop(['Unnamed: 0'], axis=1)
        df = df.drop(['ID'], axis=1)
        df = df.drop(['Case Number'], axis=1)

        df.info()
```

Fig. 5 Remove unnecessary values

- Splitting the Date to Day, Month, Year, Hour, Minute, Second.

Month	Day	Hour	Minute	Second
5	3	23	40	0
5	3	21	40	0
5	3	23	31	0
5	3	22	10	0
5	3	22	0	0

Fig. 6 Splitting the date

- Convert Categorical Attributes to Numerical.

```
# Convert Categorical Attributes to Numerical
df['Block'] = pd.factorize(df["Block"])[0]
df['IUCR'] = pd.factorize(df["IUCR"])[0]
df['Description'] = pd.factorize(df["Description"])[0]
df['Location Description'] = pd.factorize(df["Location Description"])[
df['FBI Code'] = pd.factorize(df["FBI Code"])[0]
df['Location'] = pd.factorize(df["Location"])[0]
```

```
Target = 'Primary Type'
print('Target: ', Target)
```

```
Target:  Primary Type
```

Fig. 7 Categorical Attributes to Numerical

- Encode target labels into categorical variables.

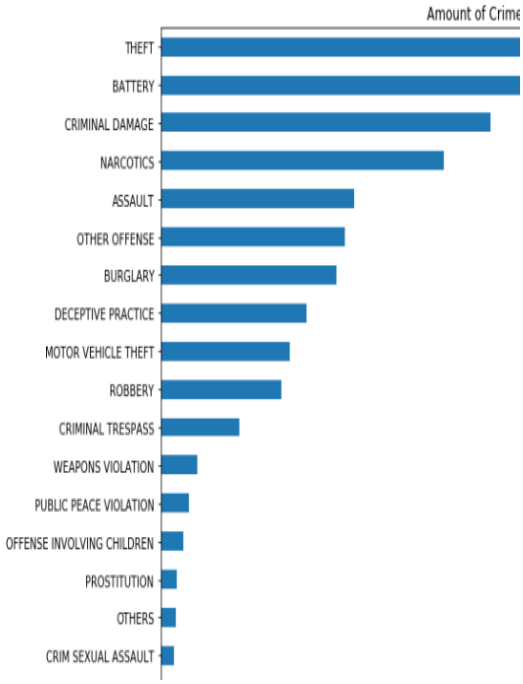

Fig. 8 Target Variables

	Primary Type	Amt
0	ARSON	2175
15	LIQUOR LAW VIOLATION	1928
14	KIDNAPPING	1075
30	STALKING	774
13	INTIMIDATION	643
21	OBSCENITY	169
4	CONCEALED CARRY LICENSE VIOLATION	84
19	NON-CRIMINAL	80
26	PUBLIC INDECENCY	61
18	NON - CRIMINAL	38
23	OTHER NARCOTIC VIOLATION	30
11	HUMAN TRAFFICKING	20
20	NON-CRIMINAL (SUBJECT SPECIFIED)	4

Fig. 9 Total amount of crimes

```
#Encode target labels into categorical variables:
df['Primary Type'] = pd.factorize(df["Primary Type"])[0]
df['Primary Type'].unique()
```

```
array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16,
       17, 18, 19, 20], dtype=int64)
```

Fig. 10 Encoded target labels

- Feature Selection using Filter Method.
- Split Dataframe to target class and features.

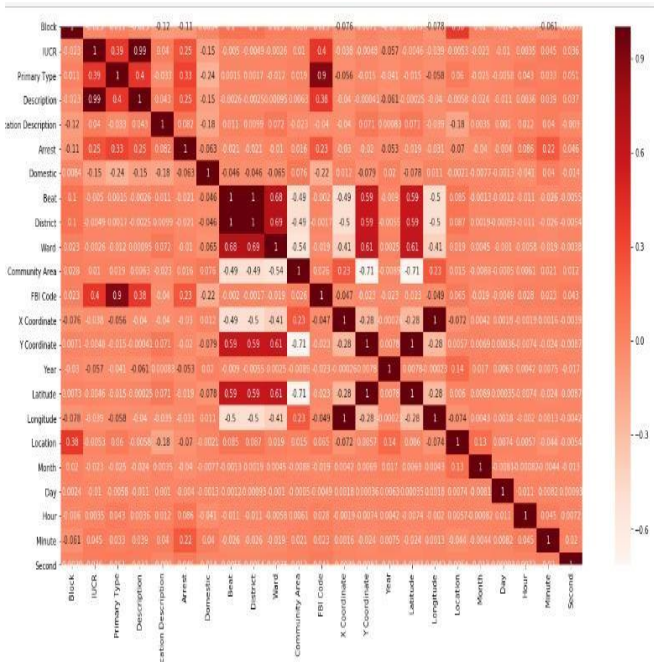

Fig. 11 Pearson Correlation

- Split dataset to Training Set & Test Set 80/20%.

```
#Split dataset to Training Set & Test Set
x, y = train_test_split(df,
                        test_size = 0.2,
                        train_size = 0.8,
                        random_state= 3)

x1 = x[Features]    #Features to train
x2 = x[Target]      #Target Class to train
y1 = y[Features]    #Features to test
y2 = y[Target]      #Target Class to test

print('Feature Set Used    : ', Features)
print('Target Class        : ', Target)
print('Training Set Size   : ', x.shape)
print('Test Set Size       : ', y.shape)
```

```
Feature Set Used    :  ['IUCR', 'Description', 'FBI Code']
Target Class        :  Primary Type
Training Set Size   :  (1134692, 23)
Test Set Size       :  (283673, 23)
```

Fig. 12 Splitting Dataset

V. RESULTS AND DISCUSSIONS

This work uses Python data mining tool. And techniques are applied on dataset by using different classification algorithms like K-Nearest Neighbours and Random Forest. After that performance of each algorithm is calculated and compare. In this phase of proposed approach two classification techniques are applied on the preprocessed news dataset one after another. And the result of each algorithm is calculated and analyzed. Then compare the result of these algorithms with each other.

```
========== Random Forest Results ==========
Accuracy    :  0.9997461866303807
Recall      :  0.9997461866303807
Precision   :  0.9997471563043292
F1 Score    :  0.9997461866303807
Confusion Matrix:
[[51838     0     0     0     0     0     0     0     0
      0     0     0     0     0     0     0     0     0]
 [    0  2581     0     0     0     0     0     0     0
      6     0     0     0     0     0     0     0     0]
 [    0     0 64548     0     0     0     0     0     0
      0     0     0     0     0     0     0     0     0]
 [    0     0     0  3340     0     0     0     0     0
      3     0     0     0     0     0     0     0     0]
 [    0     0     0     0 11214     0     0     0     0
      0     0     0     0     0     0     0     0     0]
 [    0     0     0     0     0 11876     0     0     0
```

Fig. 13 Random Forest Classifier

```
========== K-Nearest Neighbors Results ==========
Accuracy    :  0.9999576977717302
Recall      :  0.9999576977717302
Precision   :  0.9999577081573986
F1 Score    :  0.9999576977717302
Confusion Matrix:
[[51838     0     0     0     0     0     0     0     0
      0     0     0     0     0     0     0     0     0]
 [    0  2585     0     0     0     0     0     2     0
      0     0     0     0     0     0     0     0     0]
 [    0     0 64548     0     0     0     0     0     0
      0     0     0     0     0     0     0     0     0]
 [    0     0     0  3343     0     0     0     0     0
      0     0     0     0     0     0     0     0     0]
 [    0     0     0     0 11214     0     0     0     0
      0     0     0     0     0     0     0     0     0]
 [    0     0     0     0     0 11875     0     1     0
      0     0     0     0     0     0     0     0     0]
 [    0     0     0     0     0     0 17629     0     0
      0     0     0     0     0     0     0     0     0]
 [    0     0     0     0     0     0     0 16903     0
```

Fig. 14 K-Nearest Neighbours Classifier

Table 1: *Comparison Table*

Algorithms	Accuracy
Random Forest Classifier	99.97%
K-Nearest Neighbours	99.99%

The table shows that K-Nearest Neighbour technique has the best accuracy of 99.99% among other classifiers. And the result of these algorithms are analyzed and compared. The algorithm K-Nearest Neighbour showed that the predicted results is closer to real results. Thus, the dataset used, provides result with higher accuracy when implementes different classifiers. K-Nearest Neighbour works best and Random Forest works least well for predicting crimes.

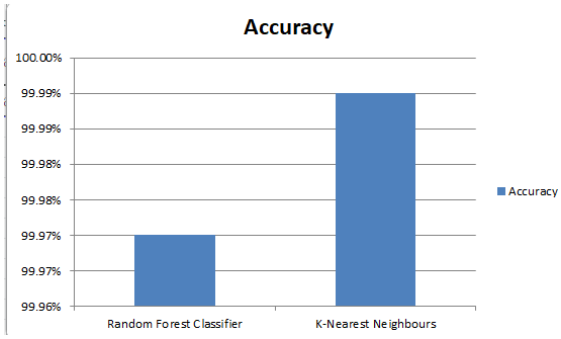

Fig. 15 Comparison Graph

VI. CONCLUSION AND FUTURE SCOPE

A. *Conclusion*

There are many problems arise to handle the huge amount of data. Different classification algorithms like K-Nearest Neighbours and Random Forest. After that performance of each algorithm is calculated and compare. Proposed work analyzed result on the basis of accuracy. Hence, K-Nearest Neighbour works best and Random Forest works least well for predicting crimes.

B. *Future Scope*

- By finding efficient classifier we can apply that algorithm for other cities crime dataset.

REFERENCES

[1] Suliman et al. (2014). Crime Data Analysis Using Data Mining Techniques to Improve Crimes Prevention, Volume 8, 2014

[2] Almanie et al. (2015). Crime Prediction Based On Crime Types And Using Spatial And Temporal Criminal Hotspots (IJDKP) Vol.5, No.4, July 2015

[3] McClendon, L. and Meghanathan, N. (2015). Using Machine Learning Algorithms To Analyze Crime Data (MLAIJ) Vol.2, No.1, March 2015

[4] Rajkumar et al. (2019). Crime Analysis And Prediction Using Data Mining Techniques, March - 2019 [ISSN: 2455-1457]

[5] Benjamin, H. & Suruliandi, A. (2017) Survey On Crime Analysis And Prediction Using Data Mining Techniques, ICTACT Journal On Soft Computing, APRIL 2017, Volume: 07, Issue: 03

[6] Bharati1, A. & Dr Sarvanaguru, RA.K. (2018) Crime Prediction and Analysis Using Machine Learning, Volume: 05 Issue: 09 |Sep 2018

[7] Akash et al (2019). Mining Criminal Dataset Using Gradient Boosting Algorithm

[8] Quader et al (2019). Predicting Crime Using Time and Location Data.

[9] Sangani et al. 2019. Crime Prediction and Analysis, 2nd International Conference on Advances in Science & Technology.

[10] Almaw, A. and Kadam, K. 2018. Survey Paper on Crime Prediction using Ensemble Approach, International Journal of Pure and Applied Mathematics, 118 (8), 133-139.

[11] Tayal, D.K., Jain, A., Arora, S., Agarwal, S., Gupta, T. and Tyagi, N., 2015. Crime detection and criminal identification in India using data mining techniques. AI & society, 30(1), pp.117-127.

[12] Hyeon-Woo Kang and Hang-Bong Kang. 2018. Prediction of crime occurrence from multimodal data using deep learning.