Crime rate prediction using k-means

Rehnuma Reza Deepty(Roll:11)
Albina Alam(Roll:15)
Tanzim Kabir(Roll: 29)

26 February 2019

1 Experiment Name

Crime rate prediction using k-means.

2 Problem Statement

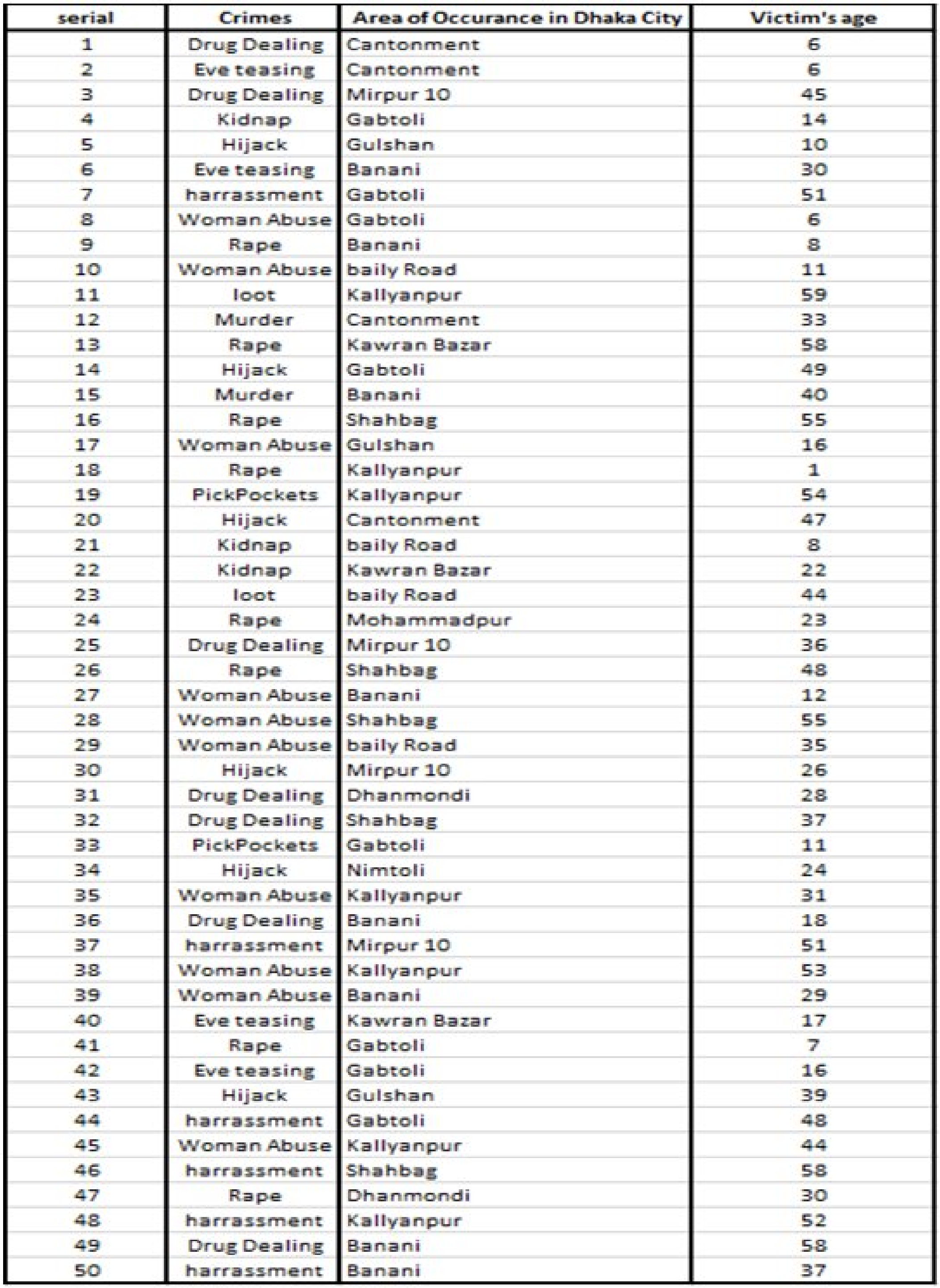
This data mining system will be able to analyze the characteristics of crime occuring in various areas and among various victims in Dhaka city. The system will predict which crimes occur most often in areas, which areas host maximum occurrence of a specific crime, age group of victims of a certain crime, areas in which a certain age group is highly prone to crimes, which crimes occur most often to victims in a specific age range, and which age group of people are prone to crime in a certain area. The result should be a comparative analysis of the data presented as relations among the various crimes, areas and age groups, e.g. people between 16 - 30 years of age are most prone to crime in Mohammadpur, the highest occuring crime in Banani is pick pocketing, eve teasing occurs the most in Mirpur, most child abuse victims are of the age 1 - 15 years, people of age 30 - 50 are most prone to murder, most victims of crime in Jatrabari are above the age of 50, etc.

3 Objectives

To predict possible crime rates in various areas on various demographics by reaching multiple decisions based statistical analysis of the dataset.

4 Data Set preparation

Data set prepared for crime rate prediction system is organized in a table called "crime_list" below in figure 1



serial	Crimes	Area of Occurance in Dhaka City	Victim's age
1	Drug Dealing	Cantonment	6
2	Eve teasing	Cantonment	6
3	Drug Dealing	Mirpur 10	45
4	Kidnap	Gabtoli	14
5	Hijack	Gulshan	10
6	Eve teasing	Banani	30
7	harrassment	Gabtoli	51
8	Woman Abuse	Gabtoli	6
9	Rape	Banani	8
10	Woman Abuse	baily Road	11
11	loot	Kallyanpur	59
12	Murder	Cantonment	33
13	Rape	Kawran Bazar	58
14	Hijack	Gabtoli	49
15	Murder	Banani	40
16	Rape	Shahbag	55
17	Woman Abuse	Gulshan	16
18	Rape	Kallyanpur	1
19	PickPockets	Kallyanpur	54
20	Hijack	Cantonment	47
21	Kidnap	baily Road	8
22	Kidnap	Kawran Bazar	22
23	loot	baily Road	44
24	Rape	Mohammadpur	23
25	Drug Dealing	Mirpur 10	36
26	Rape	Shahbag	48
27	Woman Abuse	Banani	12
28	Woman Abuse	Shahbag	55
29	Woman Abuse	baily Road	35
30	Hijack	Mirpur 10	26
31	Drug Dealing	Dhanmondi	28
32	Drug Dealing	Shahbag	37
33	PickPockets	Gabtoli	11
34	Hijack	Nimtoli	24
35	Woman Abuse	Kallyanpur	31
36	Drug Dealing	Banani	18
37	harrassment	Mirpur 10	51
38	Woman Abuse	Kallyanpur	53
39	Woman Abuse	Banani	29
40	Eve teasing	Kawran Bazar	17
41	Rape	Gabtoli	7
42	Eve teasing	Gabtoli	16
43	Hijack	Gulshan	39
44	harrassment	Gabtoli	48
45	Woman Abuse	Kallyanpur	44
46	harrassment	Shahbag	58
47	Rape	Dhanmondi	30
48	harrassment	Kallyanpur	52
49	Drug Dealing	Banani	58
50	harrassment	Banani	37

Figure 1: Crime dataset

5 Expected Output

1. SELECT ’areas’ FROM ’crime_list’ WHERE ’crime’=”Murder”
   Expected output= Cantonment, Banani.

2. SELECT ’crime’ FROM ’crime_list’ WHERE ’area’=”Mohammadpur”
   Expected Output= Rape, Drug dealing.

3. SELECT ’victim-age’ FROM ’crime_list’ WHERE ’area’=”Mohammadpur”
   Expected Output=23, 35.

4. SELECT area, crime, count(*) FROM ’crime_list’ GROUP BY ’crime’.
   Expected output =

Crime	**Area**
Murder	Cantonment, Banani
Drug dealing	Cantonment, Mirpur 10,dhanmondi, shahbag,banani
Eve teasing	Cantonment, banani , kawran bazar, gabtoli
Hijack	Gulshan, Gabtoli, Cantonment,Mirpur 10
Women abuse	Gulshan, Banani, Shahbag,Baily road,Kallyanpur

Table 1: Query 4

5. SELECT ’age’, ’crime’, COUNT(*) FROM ’crime_list’ GROUP BY ’crime’.
   Expected output =

Crime	**Age(years)**
Kidnap	7-23
Murder	20-40
Rape	1-50
Harassment	15-30
Hijack	20-55

Table 2: Query 5

6 Algorithms, Procedures and Queries

K-means algorithm is an unsupervised learning algorithm that is used to create clusters. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. The algorithm has a loose relationship to the k-nearest neighbor classifier. k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes. Applying the 1-nearest neighbor classifier to the cluster centers obtained by k-means classifies new data into the existing clusters. This is known as nearest centroid classifier or Rocchio algorithm.

The data is first clustered into multiple suitable clusters. The clustered data is then used with k-means algorithm to determine classification of data. The classified data is then called in various orders to get the desired output.

The following queries and others similar to these will be used to derive the outputs:

- SELECT 'area' FROM 'crime_list' WHERE 'crime'="murder"
- SELECT 'crime' FROM 'crime_list' WHERE 'area'="Mohammadpur"
- SELECT 'age' FROM 'crime_list' WHERE 'area'="Mohammadpur"
- SELECT 'area','crime',COUNT(*) FROM 'crime_list' GROUP BY 'crime
- SELECT 'crime','age',COUNT(*) FROM 'crime_list' GROUP BY 'age'

7 Steps

1. Gathering data set.
2. Determining queries
3. Performing computation using K-means.
4. Generating output.

8 Obtained Output

- The most prevalent crime in Gabtoli is Harrassment.
- People of ages 1-15 years are most prone to Kidnap.

9 Discussion

The obtained output did not exactly match the expected output. This is due to minor inconsistencies in the algorithm used. Due to the data being very spread out, a proper comprehensive output was difficult to achieve. Dividing the data into smaller classes can improve the outcome of this project.