Crime Pattern Detection, Analysis & Prediction

Sunil Yadav, Meet Timbadia, Ajit Yadav, Rohit Vishwakarma and Nikhilesh Yadav
Department of Information Technology,
University of Mumbai, Shree L.R Tiwari College of Engineering, Thane, India
sunvan.15@gmail.com, meettimbadia@gmail.com, ajitdmyadav@gmail.com,
rohitvish29@gmail.com,nikhileshyadav29@gmail.com

Abstract: Crimes are a social irritation and cost our society deeply in several ways. Any research that can help in solving crimes quickly will pay for itself. About 10% of the criminals commit about 50% of the crimes [9]. The system is trained by feeding previous years record of crimes taken from legitimate online portal of India listing various crimes such as murder, kidnapping and abduction, dacoits, robbery, burglary, rape and other such crimes. As per data of Indian statistics, which gives data of various crime of past 14 years (2001-2014) a regression model is created and the crime rate for the following years in various states can be predicted [8]. We have used supervised, semi-supervised and unsupervised learning technique [4] on the crime records for knowledge discovery and to help in increasing the predictive accuracy of the crime. This work will be helpful to the local police stations in crime suppression.

Keywords: K-means, Naive Bayes, Crime prediction, Regression, Crime suppression, Apriori, Association Rule

I. INTRODUCTION

The crime rates accelerate continuously and the crime patterns are constantly changing [2]. As a result, the behaviors in crime pattern are difficult to explain. This paper illustrates how social development may lead to crime prevention. The aim is to provide a comprehensive review of theory and research with respect to the prevention of the crime in the society and to implement different data analysis algorithms which address the connections between crime and its pattern. The data for the project are collected from the legitimate government sources [7]. The data was converted to .csv format upon which pre-processing of the data was performed. Technologies used for mining various crime pattern and analysis are Weka Tool and R Tool.

Weka Tool:

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. [11].

R Tool:

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering and graphical techniques, and is highly extensible [12].

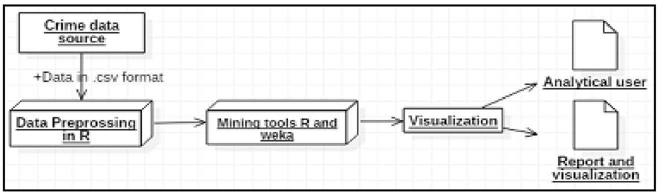


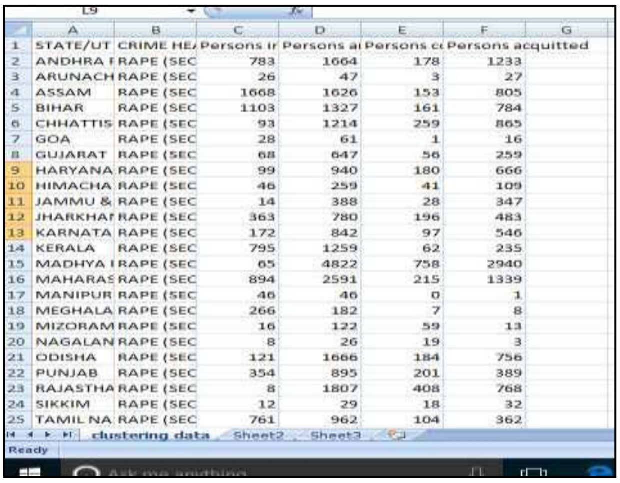
Fig 1.Block diagram(process of mining)

The four algorithms which were executed on the dataset with the help of Weka Tool and R Tool are:

i. *Association Mining (Aprori)*
ii. *Clustering(k-Means)*
iii. *Classification Techniques (Naive Bayes)*
iv. *Correlation & Regression.*

II. ARCHITECTURE AND WORKING

The dataset embraces number of people arrested and number of crimes committed along with various other attributes. Here we are primarily using four data mining algorithms for analysis of crime and to find hidden patterns of crime in India.



	A	B	C	D	E	F	G
1	STATE/UT	CRIME HE/	Persons ir	Persons ai	Persons cc	Persons acquitted	
2	ANDHRA I	RAPE (SEC	783	1664	178	1233	
3	ARUNACH	RAPE (SEC	26	47	3	27	
4	ASSAM	RAPE (SEC	1668	1626	153	805	
5	BIHAR	RAPE (SEC	1103	1327	161	784	
6	CHHATTIS	RAPE (SEC	93	1214	259	865	
7	GOA	RAPE (SEC	28	61	1	16	
8	GUJARAT	RAPE (SEC	68	647	56	259	
9	HARYANA	RAPE (SEC	99	940	180	666	
10	HIMACHA	RAPE (SEC	46	259	41	109	
11	JAMMU &	RAPE (SEC	14	388	28	347	
12	JHARKHAI	RAPE (SEC	363	780	196	483	
13	KARNATA	RAPE (SEC	172	842	97	546	
14	KERALA	RAPE (SEC	795	1259	62	235	
15	MADHYA I	RAPE (SEC	65	4822	758	2940	
16	MAHARAS	RAPE (SEC	894	2591	215	1339	
17	MANIPUR	RAPE (SEC	46	46	0	1	
18	MEGHALA	RAPE (SEC	266	182	7	8	
19	MIZORAM	RAPE (SEC	16	122	59	13	
20	NAGALAN	RAPE (SEC	8	26	19	3	
21	ODISHA	RAPE (SEC	121	1666	184	756	
22	PUNJAB	RAPE (SEC	354	895	201	389	
23	RAJASTHA	RAPE (SEC	8	1807	408	768	
24	SIKKIM	RAPE (SEC	12	29	18	32	
25	TAMIL NA	RAPE (SEC	761	962	104	362	

Fig 2.Data set for K-mean

The data mining techniques used are as follows:-

1: *Association mining (Apriori Algorithm) along with*

Clustering (k-mean):-

The paper tends to help specialist in discovering patterns, trends, making forecasts, finding relationships and possible explanations, mapping criminal networks and identifying possible suspects [3].

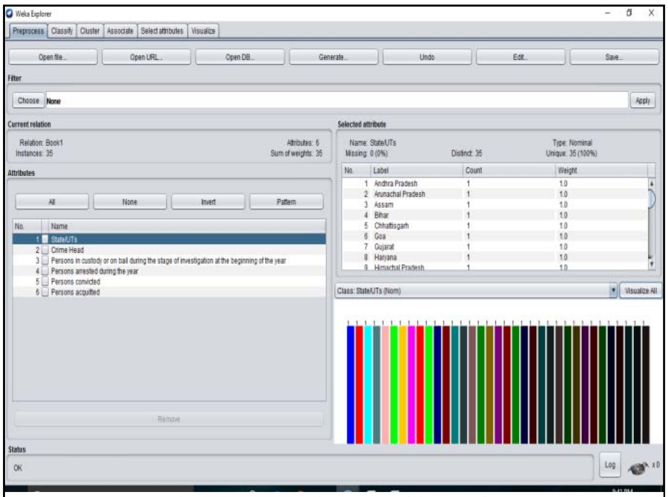
K-means is used to create number of clusters according to values high and low.

Two clusters are to be created:

Cluster 1: High number of people involved in crime.

Cluster 2: Low number of people involved in crime.

The data is first imported from the file named as book.csv into weka for preprocessing, later k-means is applied on this data set using the same graphic user interface (GUI) of Weka.



Weka Explorer

Filter: None

Current relation

Relation: Book1
Instances: 35
Attributes: 6
Sum of weights: 35

Attributes

No.	Name
1	State/UTs
2	Crime Head
3	Persons in custody or on bail during the stage of investigation at the beginning of the year
4	Persons arrested during the year
5	Persons convicted
6	Persons acquitted

Selected attribute

Name: State/UTs
Missing: 0 (0%)
Distinct: 35
Type: Nominal
Unique: 35 (100%)

No.	Label	Count	Weight
1	Andhra Pradesh	1	1.0
2	Arunachal Pradesh	1	1.0
3	Assam	1	1.0
4	Bihar	1	1.0
5	Chhattisgarh	1	1.0
6	Goa	1	1.0
7	Gujarat	1	1.0
8	Haryana	1	1.0
9	Himachal Pradesh	1	1.0

Class: State/UTs (Nom)

Status

OK

Fig 3.Data set imported in Weka

The imported data set is visualized with class attribute being STATE/UT. The visualization chart shows the distributing of attribute STATE/UT with other five attributes in the data set, each color in the visualization chart represents a particular state of India.

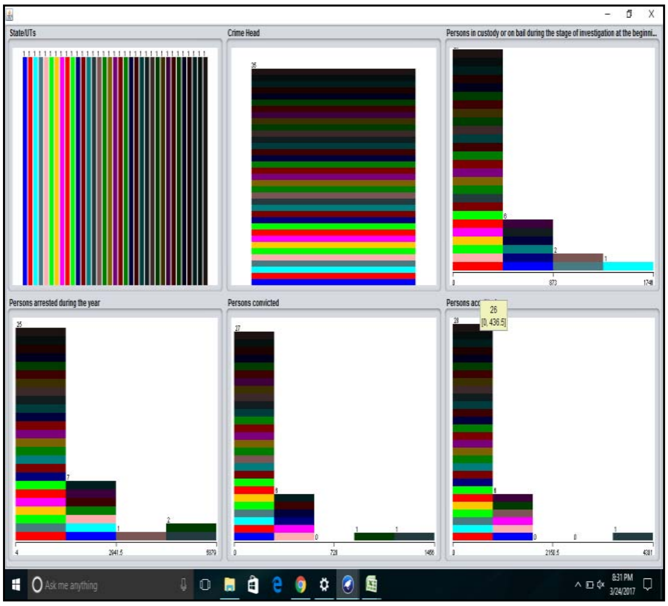
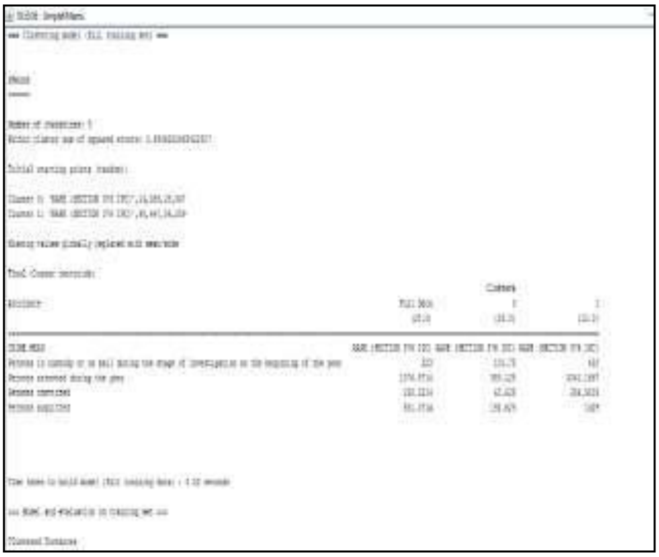


Fig 4.Visualization of dataset in Weka

The result of the k-means will acts as an input to the Apriori algorithm for discovering the association among a number of Other attributes.

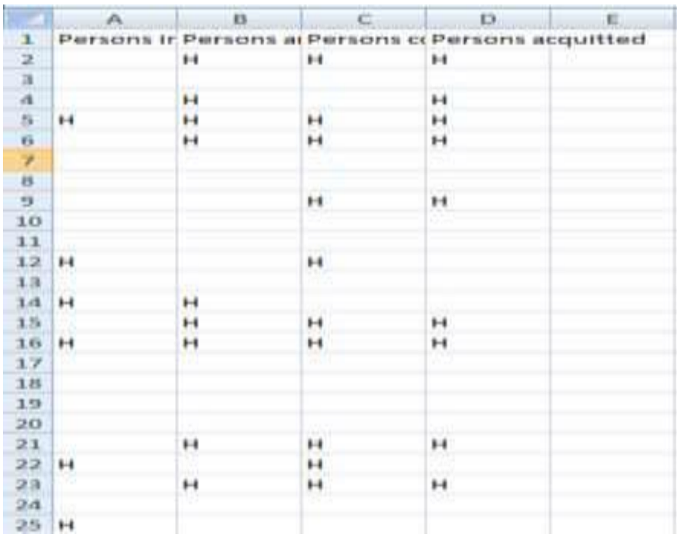


```
=== Clustering model (full training set) ===


kMeans
======

Number of iterations: [illegible]
Within cluster sum of squared errors: [illegible]

Initial starting points (random):

Cluster 0: [illegible]
Cluster 1: [illegible]

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                                          Cluster#
Attribute                                                    Full Data           0           1
                                                                  (35)        ([illegible])  ([illegible])
==============================================================================================
Crime Head                                                   [illegible] [illegible] [illegible]
Persons in custody or on bail during the stage of investigation at the beginning of the year   [illegible] [illegible] [illegible]
Persons arrested during the year                             [illegible] [illegible] [illegible]
Persons convicted                                            [illegible] [illegible] [illegible]
Persons acquitted                                            [illegible] [illegible] [illegible]




Time taken to build model (full training data) : [illegible] seconds

=== Model and evaluation on training set ===

Clustered Instances
```

Fig 5. Result of k-means

Apriori Algorithm: - Apriori algorithm [10] is a type of association mining, used to find frequent item set.

The result obtained after k-means is used as dataset for Apriori as the clusters are now divided into high and low value now we can get the association between various attributes.

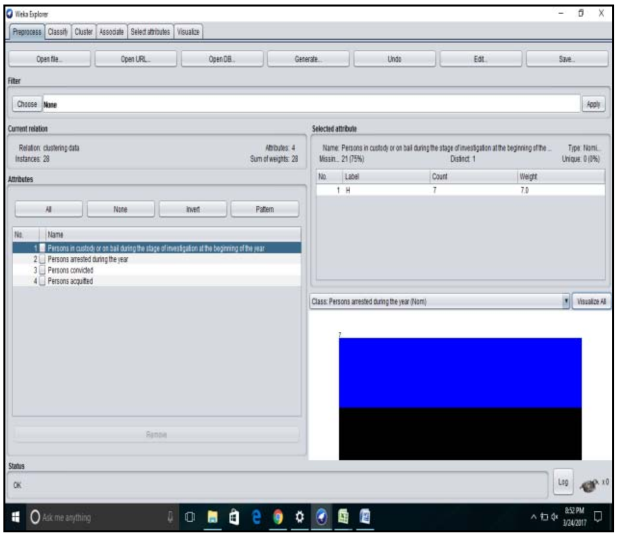


	A	B	C	D	E
1	Persons ir	Persons ai	Persons c	Persons acquitted	
2		H	H	H	
3					
4		H		H	
5	H	H	H	H	
6		H	H	H	
7					
8					
9			H	H	
10					
11					
12	H		H		
13					
14	H	H			
15		H	H	H	
16	H	H	H	H	
17					
18					
19					
20					
21		H	H	H	
22	H		H		
23		H	H	H	
24					
25	H				

Fig 6. Data set for Apriori

The dataset consists of attributes such as person arrested, person convicted, person acquitted. The "H" in the dataset represents the high number of person involved in a particular crime.

As apriori is to be applied on the output of k-means after some data preprocessing, the data set for apriori is imported into weka and association will be found using weka



None

Current relation

Relation: clustering data
Instances: 28
Attributes: 4
Sum of weights: 28

Attributes

No.	Name
1	Persons in custody or on bail during the stage of investigation at the beginning of the year
2	Persons arrested during the year
3	Persons convicted
4	Persons acquitted

Selected attribute

Name: Persons in custody or on bail during the stage of investigation at the beginning of the ...
Type: Nomi...
Missin... 21 (75%)
Distinct: 1
Unique: 0 (0%)

No.	Label	Count	Weight
1	H	7	7.0

Class: Persons arrested during the year (Nom)

Status

OK

Fig 7. Data set imported in weka

The imported dataset is visualized in weka, the visualization chart shows the distribution of crime as high or low of particular attributes with class attribute which is persons arrested during the year.

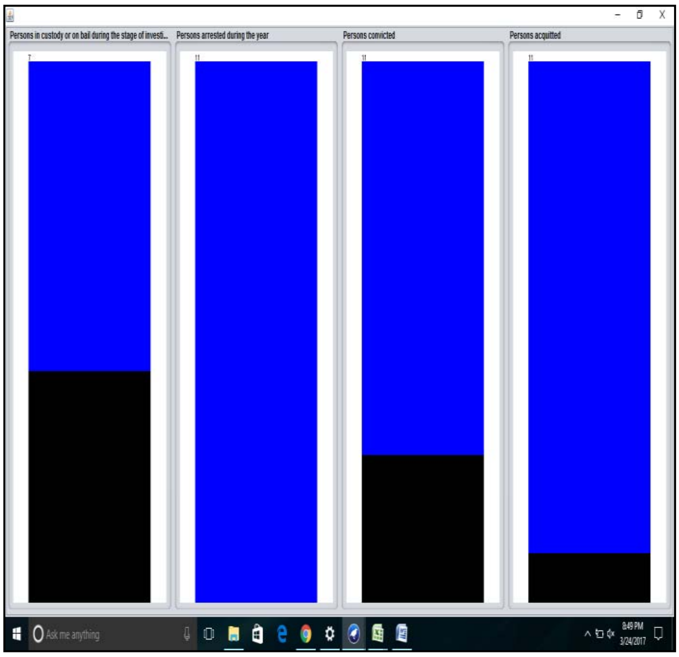
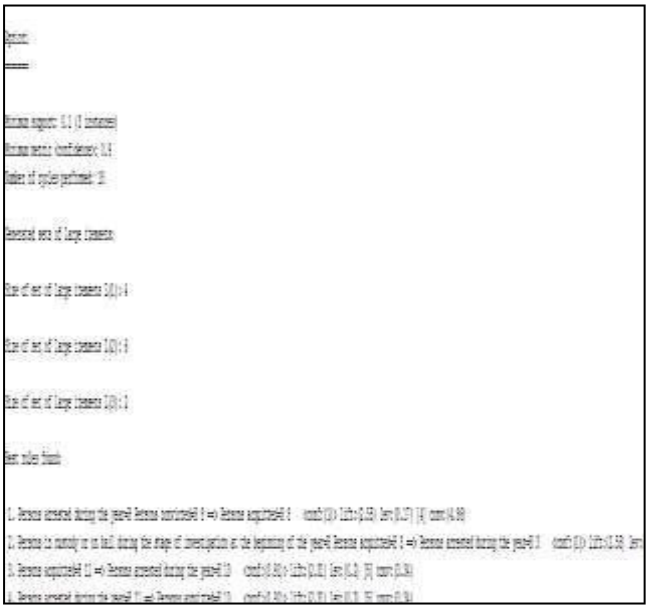


Fig 8. Visualization of Apriori dataset

The blue region in the chart represents the high crime and black region represents the low crime of particular attribute in the dataset.

The result of Apriori showed an association between the person arrested during the year and person acquitted in the same year. This result states that if person arrested are more, then person released are also more, hence more people turn out to be innocent..



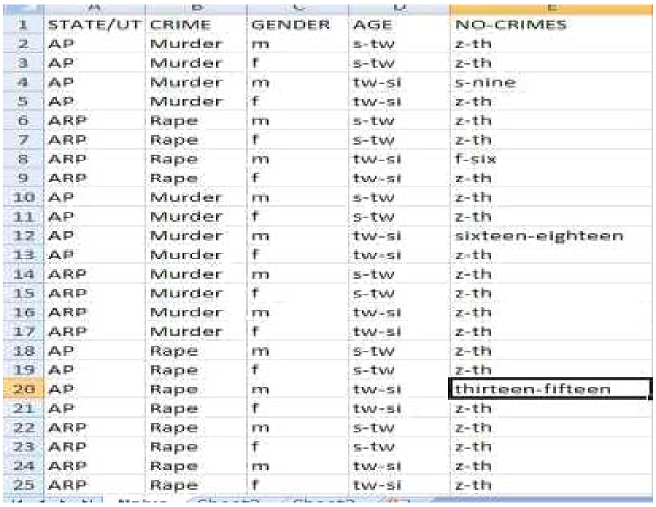
```
Apriori
=======

Minimum support: 0.1 (3 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 4

Size of set of large itemsets L(2): 3

Size of set of large itemsets L(3): 1

Best rules found:

1. Persons arrested during the year=H Persons convicted=H ==> Persons acquitted=H    conf:(1) [illegible]
2. Persons in custody or on bail during the stage of investigation at the beginning of the year=H Persons acquitted=H ==> Persons arrested during the year=H    conf:(1) [illegible]
3. Persons acquitted=H ==> Persons arrested during the year=H    conf:[illegible]
4. Persons arrested during the year=H ==> Persons acquitted=H    conf:[illegible]
```

Fig 9. Result of Apriori

2. Naive Bayes Algorithm:-

Classification is one of the classic data mining techniques, which is used to classify each item in a set of data into one of Predefined set of classes or groups [6]. The idea is to define the Criteria use for the segmentation of the whole database, once this is done, individual dataset can then fall into one or more groups naturally.

With the help of classification, existing dataset can easily be understood and it also helps to predict how new individual dataset will behave based on the classification criteria. Data mining creates classification models by observing already classified data and finding a predictive pattern among those data. [1]



	A	B	C	D	E
1	STATE/UT	CRIME	GENDER	AGE	NO-CRIMES
2	AP	Murder	m	s-tw	z-th
3	AP	Murder	f	s-tw	z-th
4	AP	Murder	m	tw-si	s-nine
5	AP	Murder	f	tw-si	z-th
6	ARP	Rape	m	s-tw	z-th
7	ARP	Rape	f	s-tw	z-th
8	ARP	Rape	m	tw-si	f-six
9	ARP	Rape	f	tw-si	z-th
10	AP	Murder	m	s-tw	z-th
11	AP	Murder	f	s-tw	z-th
12	AP	Murder	m	tw-si	sixteen-eighteen
13	AP	Murder	f	tw-si	z-th
14	ARP	Murder	m	s-tw	z-th
15	ARP	Murder	f	s-tw	z-th
16	ARP	Murder	m	tw-si	z-th
17	ARP	Murder	f	tw-si	z-th
18	AP	Rape	m	s-tw	z-th
19	AP	Rape	f	s-tw	z-th
20	AP	Rape	m	tw-si	thirteen-fifteen
21	AP	Rape	f	tw-si	z-th
22	ARP	Rape	m	s-tw	z-th
23	ARP	Rape	f	s-tw	z-th
24	ARP	Rape	m	tw-si	z-th
25	ARP	Rape	f	tw-si	z-th

Fig 10. Data set for Naive Bayes

Naive Bayes is a type of classification algorithm used for prediction it works on Bayesian principle, here we have provided a data set for prediction of the number of crime committed by a particular age group. The sample dataset includes:

STATE/UT-

AP: Andhra Pradesh
ARP: Arunachal Pradesh

AGE GROUP-

s-tw: 7-12
tw-s:12-16

NO-CRIMES-

z-th:0-3
f-six:4-6
s-nine:7-9

```
=== Confusion Matrix ===
 a b c d e <-- classified as
 1 0 0 0 0 | a = z-th
 0 0 0 0 0 | b = s-nine
 0 0 0 0 0 | c = f-six
 0 0 0 0 0 | d = sixteen-eighteen
 0 0 0 0 0 | e = thirteen-fifteen
```

Test case for Data set:-

Test case used is of the state Andhra Pradesh where crime type is murder, gender is male, and age group is between 7-8 years

Then number of crime committed by that age group is predicted as follows:

@relation Naive-test

@attribute STATE/UT {AP, ARP} @attribute CRIME {Murder, Rape} @attribute GENDER {m, f} @attribute AGE {s-tw, tw-si}

@attribute NO-CRIMES {z-th, s-nine, f-six, sixteen-eighteen, thirteen-fifteen}

@data

AP, Murder, m, s-tw, z-th

The most probabilistic answer for test case having:

State: Andhra Pradesh
Crime: Murder
Gender: Male
Age Group: 7-8 years

Result:
Number of Murder committed: 0-3

3. Correlation & Regression:-

Correlation is a statistical technique used to determine the degree to which two variables are related.

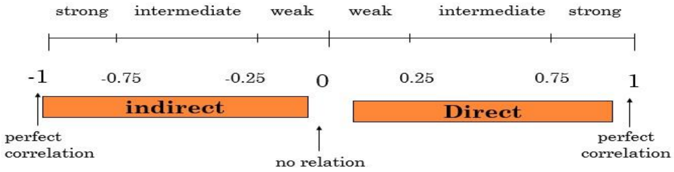


Fig 11: Correlation

If the result of correlation is 1 that means there is perfect relation between the two attributes, if result is 0 then there is no relation between the two attributes, hence there must be strong relation between the attributes to get significant result.

The aim of linear regression is to model a continuous variable *Y* as a mathematical function of one or more *X* variable(s), so that we can use this regression model to predict the *Y* when only the *X* is known. This mathematical equation can be generalized as follows:

$$Y = \beta_1 + \beta_2 X + \epsilon \qquad (1)$$

Where, β_1 is the intercept and β_2 is the slope. Collectively, they are called *regression coefficients*. ϵ is the error term; the part of *Y* the regression model is unable to explain [13].

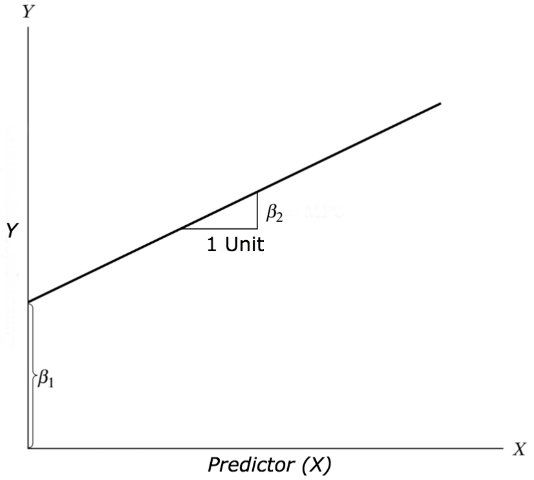
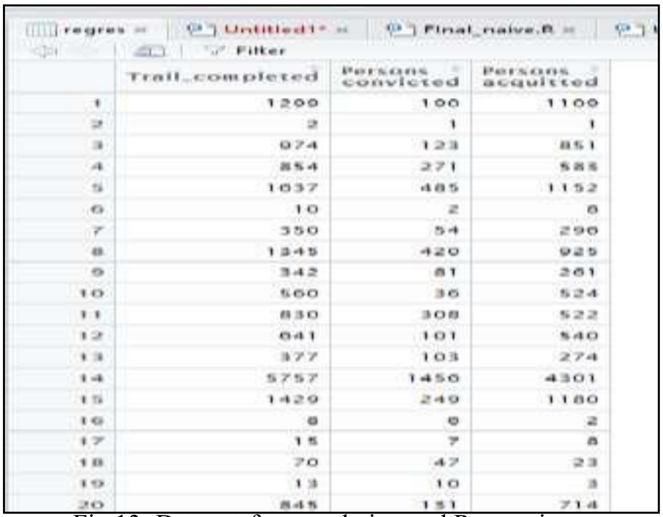


Fig 12: Sample regression diagram

The dataset belong to rape Crime against Women of all the states of India.



	Trail_completed	Persons convicted	Persons acquitted
1	1200	100	1100
2	2	1	1
3	974	123	851
4	854	271	583
5	1637	485	1152
6	10	2	8
7	350	54	296
8	1345	420	925
9	342	81	261
10	560	36	524
11	830	308	522
12	641	101	540
13	377	103	274
14	5757	1456	4301
15	1429	249	1180
16	8	6	2
17	15	7	8
18	70	47	23
19	13	10	3
20	845	131	714

Fig 13: Data set for correlation and Regression

The data set of crime rape contains attributes such as number of Trial completed, number of persons convicted during the trails as well as number of persons acquitted on the charge of crime rape.

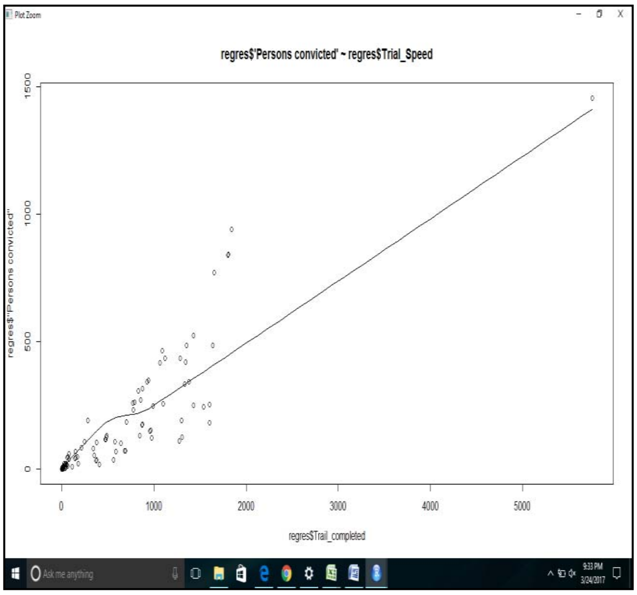
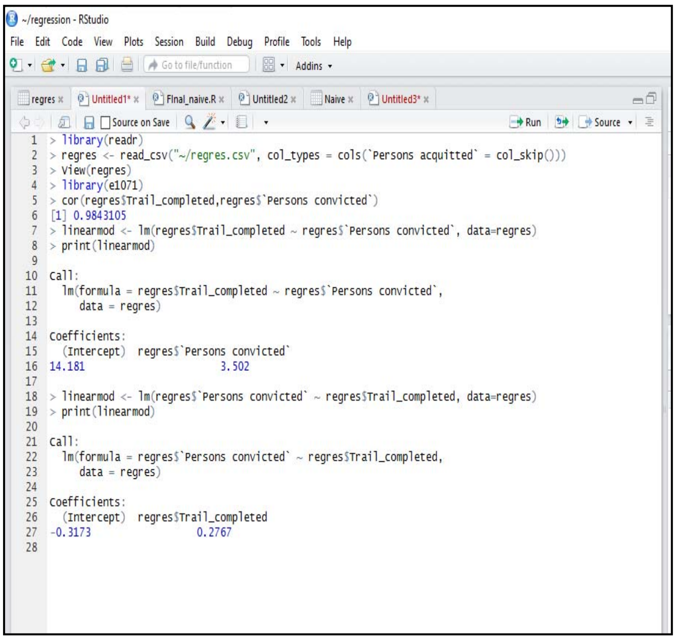


Fig 14: Scattered plot of data set

Result of correlation:
X= Trials completed during year. Y=person convicted.

Correlation: 0.98 (Strong relation) There is a strong relation between X & Y.

This means that there is a significant relation between Trials Completed during the year and person convicted on rape charge.



```
> library(readr)
> regres <- read_csv("~/regres.csv", col_types = cols(`Persons acquitted` = col_skip()))
> View(regres)
> library(e1071)
> cor(regres$Trail_completed,regres$`Persons convicted`)
[1] 0.9843105
> linearmod <- lm(regres$Trail_completed ~ regres$`Persons convicted`, data=regres)
> print(linearmod)

Call:
  lm(formula = regres$Trail_completed ~ regres$`Persons convicted`,
     data = regres)

Coefficients:
  (Intercept)  regres$`Persons convicted`
14.181                        3.502

> linearmod <- lm(regres$`Persons convicted` ~ regres$Trail_completed, data=regres)
> print(linearmod)

Call:
  lm(formula = regres$`Persons convicted` ~ regres$Trail_completed,
     data = regres)

Coefficients:
  (Intercept)  regres$Trail_completed
-0.3173                     0.2767

```

Fig 15: Result of correlation and regression

From output,

$$Y = \beta_1 + \beta_2 X + \epsilon \qquad (2)$$

Here if x is 10 i.e. if trail completed is 10 then

Y (Person convicted) = 0.2767*10-0.3173=2.449 (2.1)

Therefore for every 10 rape case trials completed only 2.5~ 3 people are convicted of the rape charges.

Thus by regression we can predict the number of people who acquitted the crime against the number of Trials completed during the year.

III. CONCLUSION AND FUTURE SCOPE

The biggest hurdle in the project was data acquisition and data staging. As a future scope extension of crime detection and analysis will be to generate the crime hot-spots that will help in deployment of police at most likely places of crime for any given window of time, to allow most effective utilization of police resources. The developed model will reduce crimes and will help the crime detection field in many ways that is from arresting the criminals to reducing the crimes by carrying out various necessary measures.
As in near future these methods can be applies on full data set which consists of 42 crime heads having 14 attributes to them, thus when analyzed could provide more unexpected dependencies of attributes over each other.

Geospatial in Crime Pattern Detection:

Crime is neither systematic nor entirely random [5].The system could be enriched to support crime mapping over India through which terrestrial model can be created declaring various crimes and the degree of such crimes performed. This terrestrial model would help us to compare the various crime rates in the diverse states of India and to cultivate new strategies to abate the crime rate in that particular area.

IV. REFERENCES

[1] Mugdha Sharma, *Z-crime: A data mining tool for the detection ofsuspicious criminal activity based on decision tree*, IEEE, 2014,ISBN:978-1-4799-4674-7/14

[2] Ubon Thansatapornwatana, *A Survey of Data Mining Techniques forAnalyzing Crime Patterns* Second Asian Conference on DefenseTechnology ACDT, IEEE, 2016, ISBN: 978-1-5090-2258-8/16

[3] Dr.Zakaria Suliman Zubi, Ayman Altaher Mahmmud, *Using DataMining Techniques to Analyze Crime patterns in the Libyan National Crime Data*, Recent advances in image, audio and signal processing.ISBN: 978-960-474-350-6

[4]Shiju Sathyadeven, Deven M.S, Surya Gangadharan. S, Crime Analysisand prediction using data mining, IEEE, 2014

[5]Chung- Hsien Yu, Max W. Ward, Melissa Morabito, Wei Ding *Crimeforecasting using data mining technique,*11thIEEE

InternationalConference on Data mining workshop, IEEE, 2011,ISBN:978-0-7695-4409-0/11

[6]Fatih OZGUL, Claus ATZENBECK, Ahmet CELIK, Zeki ERDEM, *Incorpating data sources and methologies for crime data mining,* IEEE,2011, ISBN: 978-1-61284-4577-0085-9/11

[7]National Crime Records Bureau Website. [online]

[8]Arunima S. Kumar, Raju K. Gopal, *Data Mining Based CrimeInvestigation System: Taxonomy and Relevance*, IEEE, 2015, ISBN: 978-1-4799-8553-1/15

[9]Shyam Varan Nath, *Crime Pattern Detection Using Data Mining,* International Conference on Web Intelligence and Intelligent Agent (WI-IAT 2006 Workshops), IEEE, 2006, ISBN:0-7695-2749-3/06

[10]Jiawei Han, Micheline Kamber, Jian Pei, *Data Mining: Concepts and Techniques 3rd ed.*

[11]http://www.cs.waikato.ac.nz/ml/weka/.[online]

[12]https://www.r-project.org/about.html.[online]

[13]http://r-statistics.co/Linear-Regression.html.[online]