L.J. INSTITUTE OF COMPUTER APPLICATIONS

Master of Computer Application

Project Title

Analysis of Crime Ratio in Different states
of India using Chi-Square Test.

Name	Enrollment number
Undhad Archana	185173693121
Soni Pooja	185173693113
Saiyed Muhammed Munaf	185173693098

Under the guidance of

Prof. Shanti Varma

Project submitted in partial fulfilment of MCA II, Semester 4
Software Project 2 - Data Science (4649304)
Master of Computer Application
Gujarat Technological University

L.J. INSTITUTE OF COMPUTER APPLICATIONS

Near Nagdev Kalyan Mandir, Near Sanand Cross Roads,
Serkhej-Gandhinagr Highway Ahmedabad - 382481
Ph. No. : 079 29296364

CERTIFICATE

Enrollment No: **185173693121** **Seat No:** ___M400755___

This is to certify that ~~Mr~~. / Ms._____Undhad Archana_____ of Master in Computer Applications, Semester IV, Roll No___59___ has satisfactorily completed his/her Project titled __Analysis of Crime Ratio in Different states of India using Chi-Square Test__ under the supervision of

Internal Guide Names **Signature**

Prof. Shanti Varma

Date of Submission:

Director

L.J. INSTITUTE OF COMPUTER APPLICATIONS

Near Nagdev Kalyan Mandir, Near Sanand Cross Roads,
Serkhej-Gandhinagr Highway Ahmedabad - 382481
Ph. No. : 079 29296364

CERTIFICATE

Enrollment No: 185173693113 **Seat No:** M400754

This is to certify that ~~Mr.~~ / Ms. **Soni Pooja** of Master in Computer Applications, Semester IV, Roll No **57** has satisfactorily completed his/her Project titled **Analysis of Crime Ratio in Different states of India using Chi-Square** under the supervision of

Internal Guide Names **Signature**

Prof. Shanti Varma

Date of Submission:

Director

L.J. INSTITUTE OF COMPUTER APPLICATIONS

Near Nagdev Kalyan Mandir, Near Sanand Cross Roads,
Serkhej-Gandhinagr Highway Ahmedabad - 382481
Ph. No. : 079 29296364

CERTIFICATE

Enrollment No: **185173693098** **Seat No:** M400606

This is to certify that Mr. / ~~Ms.~~ **Saiyed Muhammed Munaf** of Master in Computer Applications, Semester IV, Roll No **49** has satisfactorily completed his/her Project titled **Analysis of Crime Ratio in Different states of India using Chi-Square** under the supervision of

Internal Guide Names **Signature**

Prof. Shanti Varma

Date of Submission:

Director

Components

Introduction

This study presents the trend analyses of police crime recorded data in India. The data for our study is drawn from Crime in India, an annually publication by the National Crime Record of India.

Crime exists in India in various forms such as murder, drug trafficking, money laundering, fraud, human trafficking, robbery and prostitution etc.

It results suggest that rates of murder, robbery, theft, and rioting follow declining trends, while rates of rape show an increasing trend between 2001 and 2012.

Accordingly, this paper will attempt an overview of the trends and characteristics of crime and crime control in India and in doing so will provide a general understanding of crime in Indian society.

Problem Statement

By examining the crime data, we can get a better picture of how we can make safe our country. And through this analysis we can predict that which crime happened in large amount during 2000 to 2012 and how we can control that.

Organization of Data

National Crime Records Bureau (NCRB)
Government of India has published this dataset on their website and also has shared on Open Government Data Platform India portal under Govt. Open Data License - India.

List of Attributes

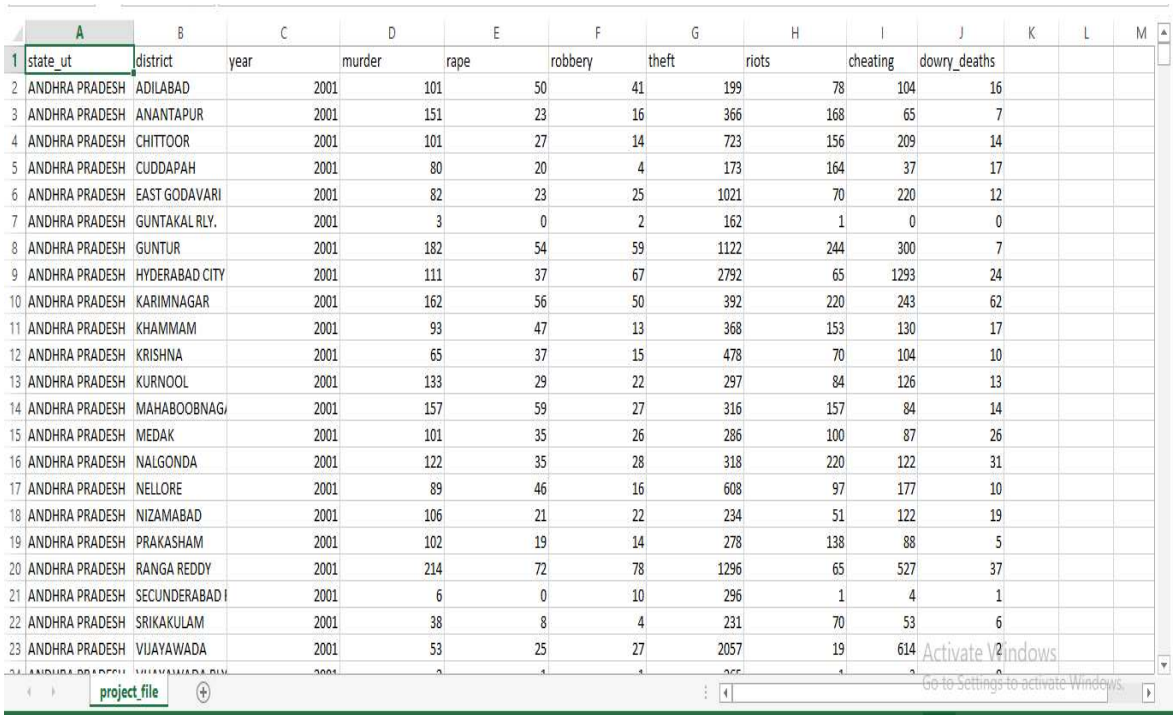
Sr. No	Attribute Name	Meaning of Attributes
1	state_ut	State value of India state
2	District	State wise District value of India
3	Year	Year value between 2001 to 2012
4	Murder	Number of Murder to state in district wise
5	Rape	Number of Rape to state in district wise
6	Robbery	Number of Robbery to state in district wise
7	Theft	Number of Theft to state in district wise
8	Riots	Number of Riots to state in district wise
9	Cheating	Number of Cheating to state in district wise
10	dowry_deaths	Number of Dowry deaths to state in district wise

Cleaning Data

For creating a predictive model, firstly the data must be clean and accurate and null values should not be there.
Data cleansing, data cleaning, data-wash or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table or database
In This dataset we have checked null values. And we clear all the null or Nan values by different techniques.

any (is.na(murder))

After cleaning Data, the data shown in below figure 1.



	A	B	C	D	E	F	G	H	I	J
1	state_ut	district	year	murder	rape	robbery	theft	riots	cheating	dowry_deaths
2	ANDHRA PRADESH	ADILABAD	2001	101	50	41	199	78	104	16
3	ANDHRA PRADESH	ANANTAPUR	2001	151	23	16	366	168	65	7
4	ANDHRA PRADESH	CHITTOOR	2001	101	27	14	723	156	209	14
5	ANDHRA PRADESH	CUDDAPAH	2001	80	20	4	173	164	37	17
6	ANDHRA PRADESH	EAST GODAVARI	2001	82	23	25	1021	70	220	12
7	ANDHRA PRADESH	GUNTAKAL RLY.	2001	3	0	2	162	1	0	0
8	ANDHRA PRADESH	GUNTUR	2001	182	54	59	1122	244	300	7
9	ANDHRA PRADESH	HYDERABAD CITY	2001	111	37	67	2792	65	1293	24
10	ANDHRA PRADESH	KARIMNAGAR	2001	162	56	50	392	220	243	62
11	ANDHRA PRADESH	KHAMMAM	2001	93	47	13	368	153	130	17
12	ANDHRA PRADESH	KRISHNA	2001	65	37	15	478	70	104	10
13	ANDHRA PRADESH	KURNOOL	2001	133	29	22	297	84	126	13
14	ANDHRA PRADESH	MAHABOOBNAG/	2001	157	59	27	316	157	84	14
15	ANDHRA PRADESH	MEDAK	2001	101	35	26	286	100	87	26
16	ANDHRA PRADESH	NALGONDA	2001	122	35	28	318	220	122	31
17	ANDHRA PRADESH	NELLORE	2001	89	46	16	608	97	177	10
18	ANDHRA PRADESH	NIZAMABAD	2001	106	21	22	234	51	122	19
19	ANDHRA PRADESH	PRAKASHAM	2001	102	19	14	278	138	88	5
20	ANDHRA PRADESH	RANGA REDDY	2001	214	72	78	1296	65	527	37
21	ANDHRA PRADESH	SECUNDERABAD I	2001	6	0	10	296	1	4	1
22	ANDHRA PRADESH	SRIKAKULAM	2001	38	8	4	231	70	53	6
23	ANDHRA PRADESH	VIJAYAWADA	2001	53	25	27	2057	19	614	2

project_file

Figure 1: data after cleaning.

The several types of techniques used for cleaning data are firstly Excel's inbuild filtering method is used for cleaning the data. And python dropna( ) is used for dropping and deleting the null value rows from the datasets.

"data = data.dropna( )"

Methods / Techniques

Application and Identification of technique and tool used:

Matplot Library:

Matplot is a python library used to create 2D plot graph and plots by using python scripts.

Pyplot Library:

It has a module named pyplot which makes things easy for plotting by providing feature to control line styles, font properties, formatting axes etc. It supports a very wide variety of graph and plots namely-histogram, bar charts, power spectra, error charts etc.

Pandas Library:

Pandas is an open –source python library used for high-performance data manipulation and data analysis using its powerful data structures.

Numpy Library:

It is used along with NumPy to provide an environment that is an effective open source alternative for MatLab.
Numpy is a Python package which stands for 'Numerical python'. It is a library consisting of multidimensional array objects and a collection of routines for processing of array.

Seaborn Library:

The main idea of Seaborn is that it provides high-level commands to create a variety of plot types useful for statistical data exploration, and even some statistical model fitting.

Scipy Library:

SciPy is an Open Source Python-based library, which is used in mathematics, scientific computing, Engineering, and technical computing.

SciPy contains varieties of sub packages which help to solve the most common issue related to Scientific Computation.

Sklearn Library:

Scikit-learn is a library in Python that provides many unsupervised and supervised learning algorithms.

The functionality that scikit-learn provides include:

Regression, Classification, Clustering, Model selection, Pre-processing

Data Visualization

Preliminary Analysis:

In this project statistical analysis done using inbuild describe() function.

```
" murder_summary=[ ]
  murder_summary=data['murder'].describe( )
  print(murder_summary)"
```

Crime Summery Table

Crime Name	Mean	standard Deviation	Quantile Value			Maximum
			Q1	Q2	Q3	
Murder	46.823659	42.755817	18.00000	37.00000	63.00000	542.000000
Rape	27.816331	30.897866	8.000000	19.000000	39.000000	568.000000
Robbery	28.878330	50.937364	6.000000	16.000000	34.000000	1131.000000
Theft	407.180412	778.770851	86.000000	206.000000	430.000000	13195.000000
Riots	90.139235	137.592242	10.000000	44.000000	116.000000	3181.000000
Cheating	89.286612	183.973214	12.000000	35.000000	93.000000	3155.000000
Dowry Deaths	10.608584	14.186601	1.000000	5.000000	15.000000	168.000000

In this we have calculated mean, median, standard deviation, Q1, Q2, Q3, Max value.
This all is calculated for all columns like for murder, rape, robbery, theft, riots, cheating, dowry deaths.
Through this we understand that from these seven crimes the highest crime rate of crime is theft and second highest crime is Riots in India.

Graphical Representation with Line Graph:

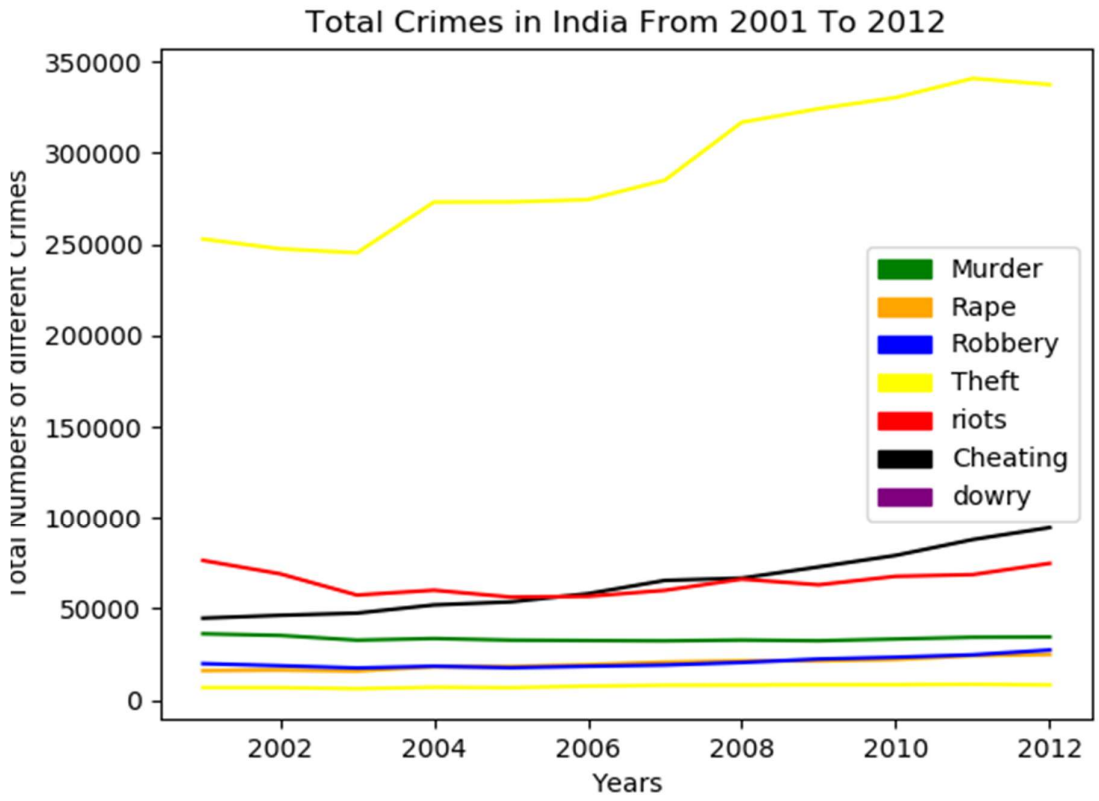


Figure 2: line graph for all crimes.

Through this Line graph we can read statistics of Indian IPC Crimes and through this graph we can understand that Theft Crime is very High in India Among All Other Crimes and this crime is Continue Rising from 2001 to 2012 no changes are there in crimes.

Its show that every year theft crime keeps increasing. And this crime is very higher from all other crimes.

We can say that theft crime is increasing because of unemployment in our country.

Crime Data Visualization by Murder Wise:

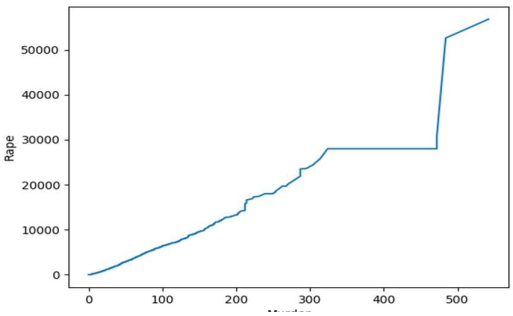


Figure 3: Murder wise Rape crime graph

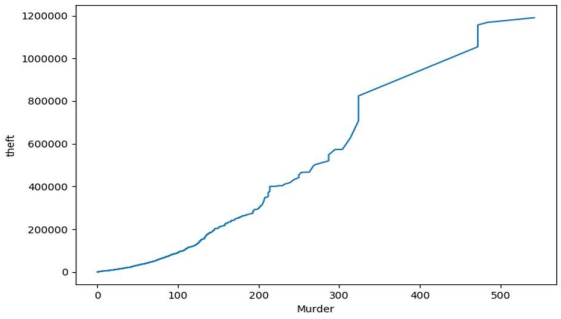


Figure 4: Murder wise Theft graph

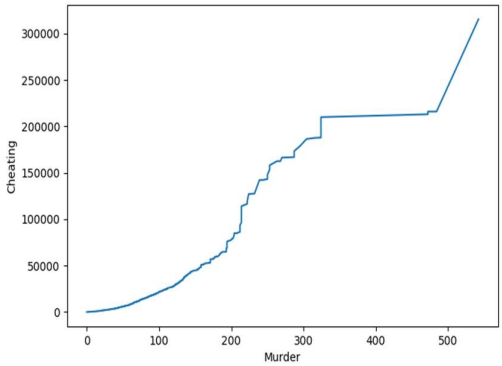


Figure 5: Murder wise Cheating graph

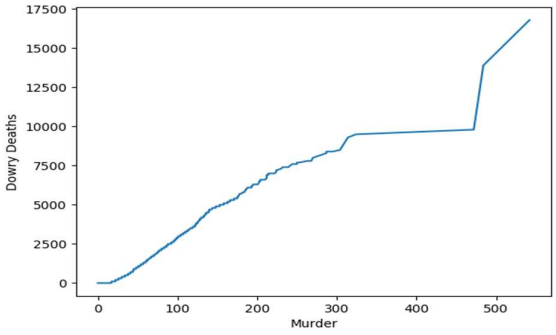


Figure 6: Murder wise Dowry death graph

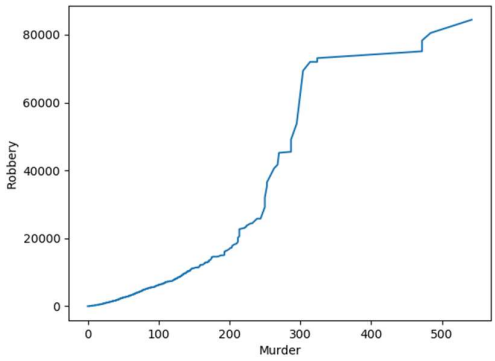


Figure 7: Murder wise Robbery graph

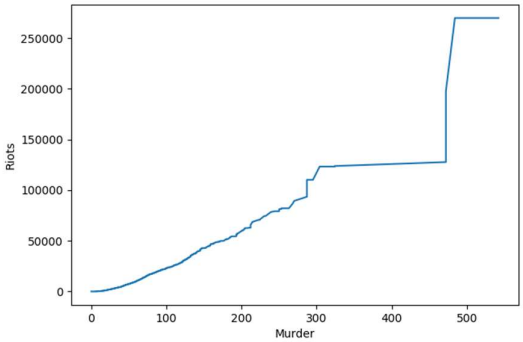
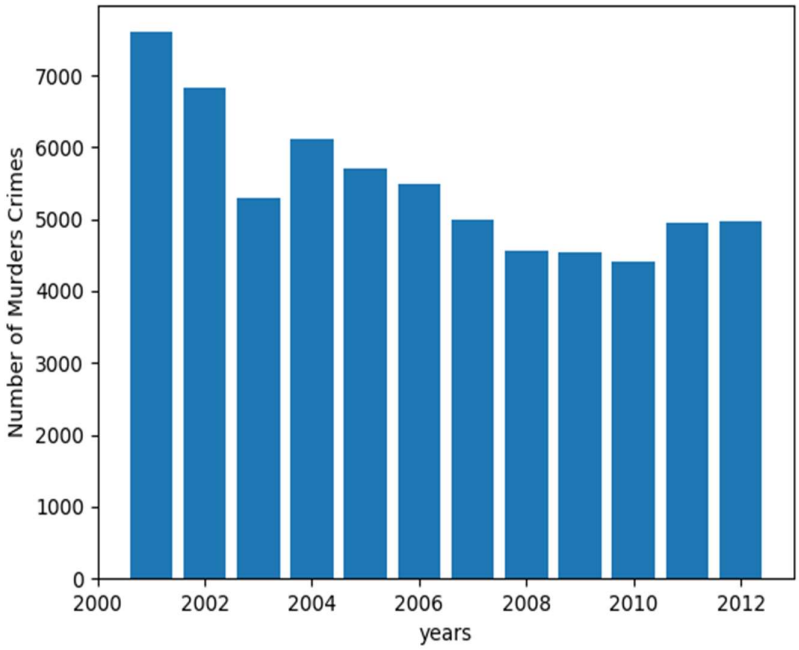


Figure 8: Murder wise Riots graph

- So, this graph Show Relation between Murder and the different crimes like robbery, rape, theft, riots.
- Which means for example it checks like if rape is increasing then it affects to murder crime or not. Murder is increasing because of rape increasing.

Crime Data Visualization by Using Bar Chart

Figure 9: Bar graph for Murder crime from 2000 to 2012



This Bar Chart Show Years Wise Murder Crime Records, and Year 2010 has a smaller Number of Murder. So, we identify that years wise Murder crime is decreasing

Crime Data Visualization by Using Bar Chart

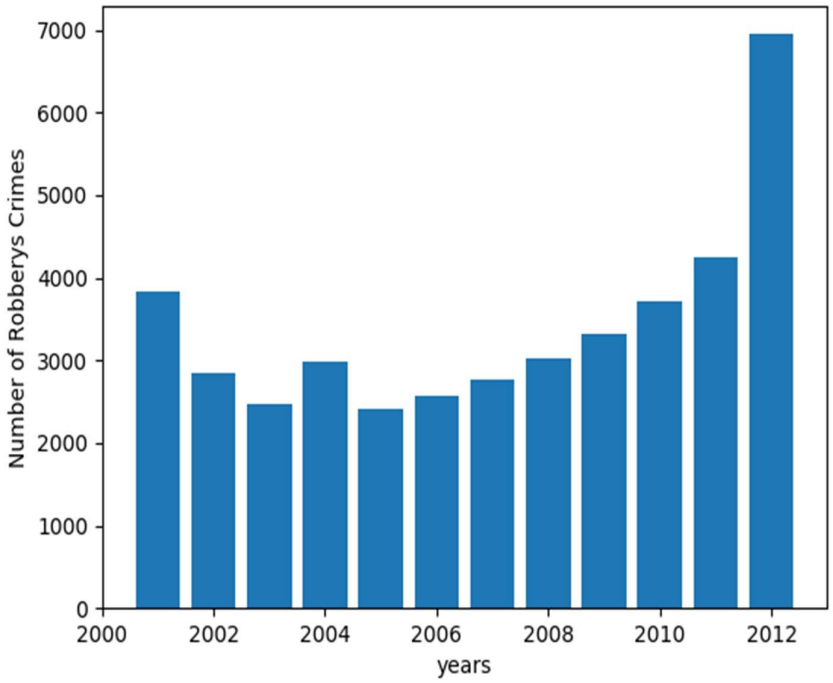


Figure 10: Bar graph for Robbery crime from 2000 to 2012.

By reading this bar graph we identify that every year from 2005 to 2012 Robbery crime is keep increasing. And in 2012 robbery crime is more frequently increasing from 2011.
The robbery crime increases in India very speedy from 2011 to 2012. This period has more robbery crime.

Visualizing Data with Pairs Plots

Plot pairwise relationships in a dataset.
By default, this function will create a grid of Axes such that each variable in data will by shared in the y-axis across a single row and in the x-axis across a single column. The diagonal Axes are treated differently, drawing a plot to show the univariate distribution of the data for the variable in that column.

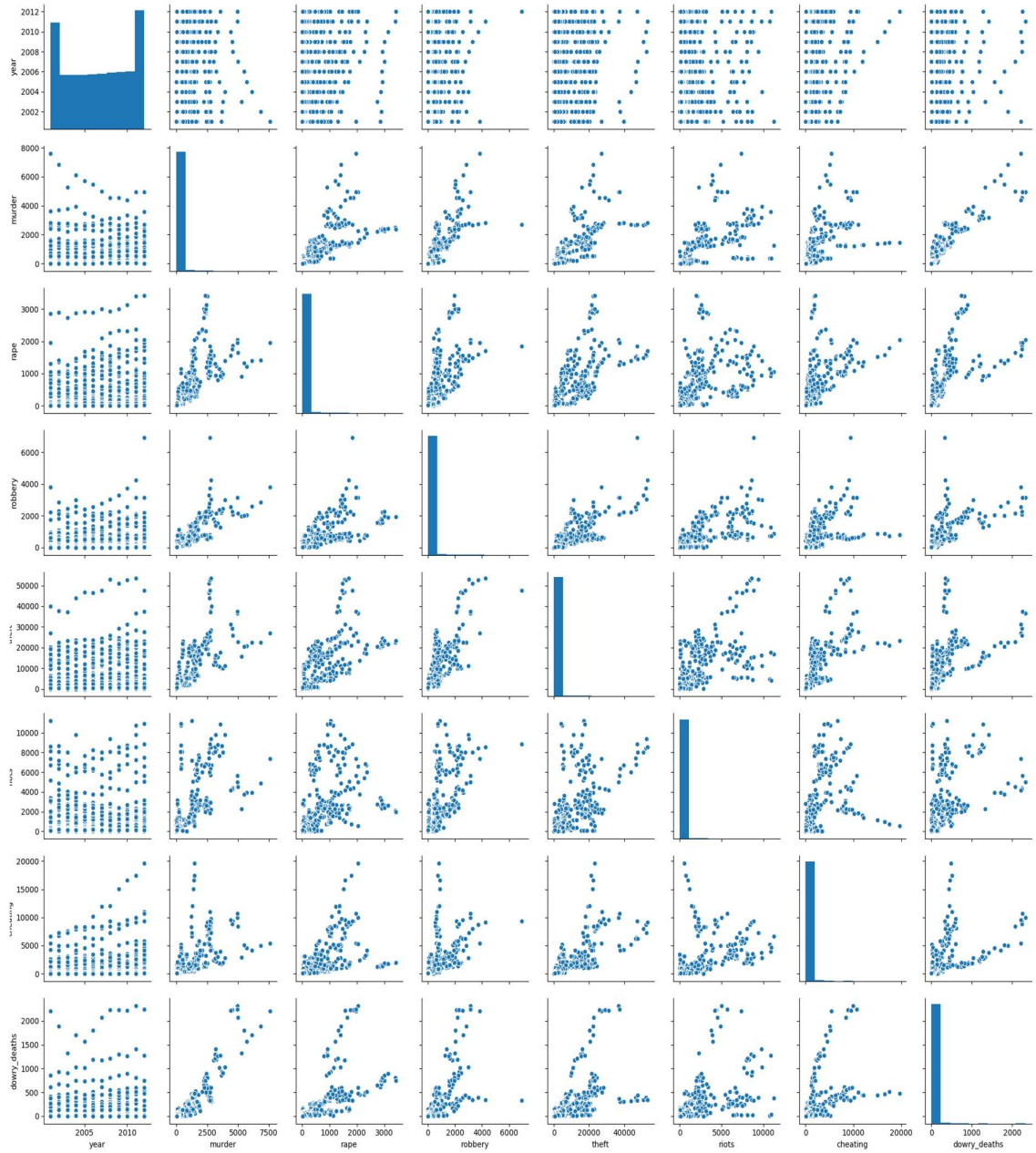


Figure 11: Pari plot graph for whole dataset

Visualizing Data with Heatmap Graph

A heatmap contains values representing various shades of the same colour for each value to be plotted. Usually the darker shades of the chart represent lower values than the lighter shade. For a very different value a completely different colour can also be used.

The below figure is a two-dimensional plot of values which are mapped to the indices and columns of the chart.

Also note that it's now easier to compare magnitudes of negative vs positive values

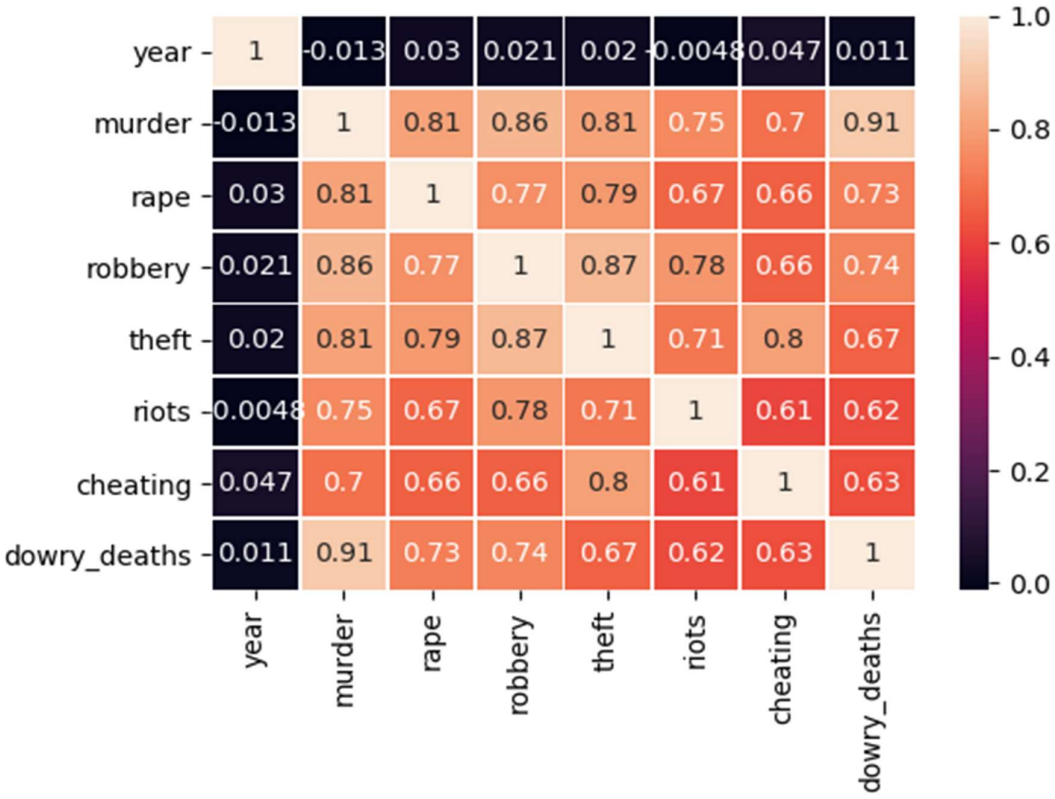


Figure 12: Heatmap to check dependency

So we identify the relation between all variables this shows relations variable to variable with the help of heatmap.

In this heatmap we can identify that dowry deaths and murder has highly dependency with each other and murder and year has negative dependency.

Visualizing Data with Distplot Graph

Histograms and KDE can be combined using dispolt **Distplot()**. This function combines the matplotlib hist function (with automatic calculation of a good default bin size) with the seaborn **kdeplot()** and **rugplot()** functions. It can also fit scipy.stats distributions and plot the estimated PDF over the data.
Given below Graph shown Murder to correspond with rape and robbery Crime.

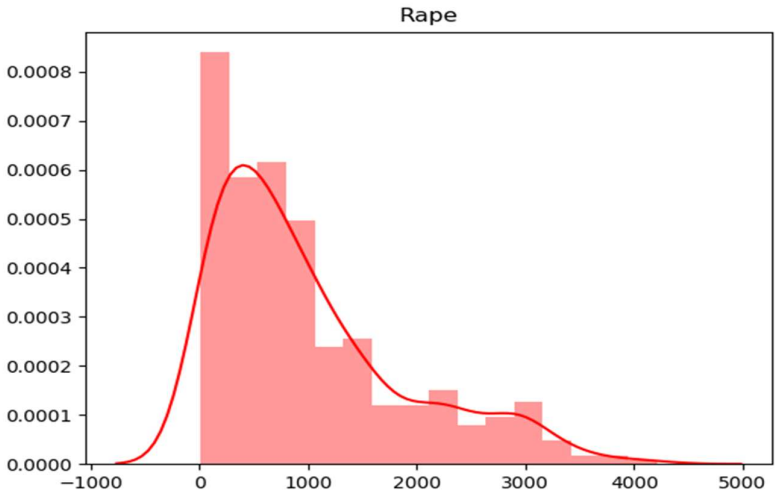


Figure 13: Rape corresponds with murder

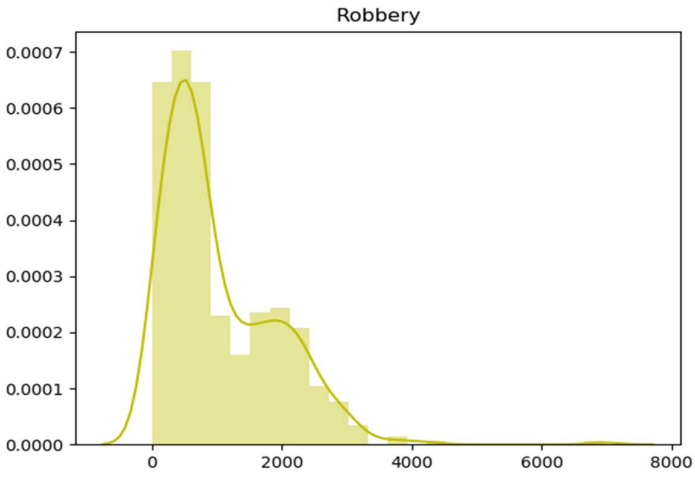


Figure 14: Robbery corresponds with murder

This dist. plot graph shows us the corresponds of one variable with another variable. It helps us to plot the estimated PDF over the data.

Visualizing Data with Boxplot

We can detect Outliers and handle it in data. It can distort predictions and affect the accuracy, if you don't detect and handle them appropriately especially in regression models then our whole analysis became wrong.
Data point that falls outside of 1.5 times of an interquartile range above the 3rd quartile and below the 1st quartile then it identify an outlier.
Data point that falls outside of 3 standard deviations. we can use a z score and if the z score falls outside of 2 standard deviation

The impact of an outlier:

Causes serious issues for statistical analysis
Skew the data.
Significant impact on mean
Significant impact on standard deviation.

we can identify an outlier by plotting

using box plots
using scatter plots
using Z score
using the IQR interquartile range

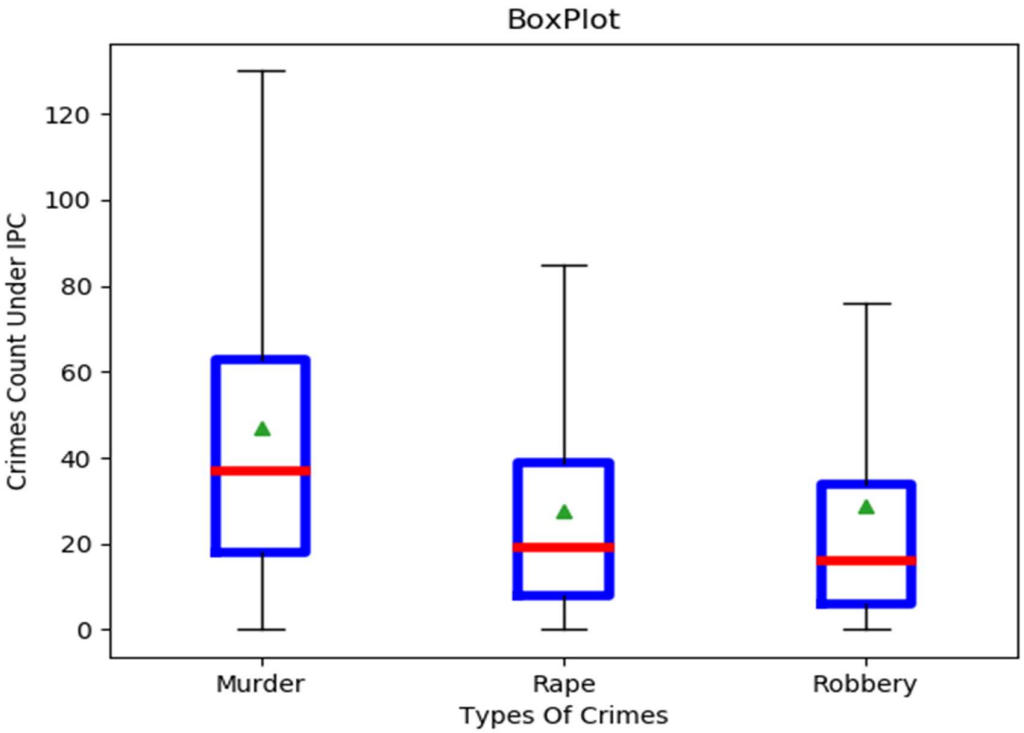


Figure 15: Box plot for murder, rape and robbery.

Result:

Box Plot of murder, rape and robbery crimes by state.
Based on this box plot we can identify that have maximum number crime among these three crimes is Murder.
Through this plot we can check the wrong data inserted. The wrong data of fake we can see the value which is Outlier.

Chi square Test

A chi-square test is a technique to compares the observed distribution of data to an expected distribution of data.
There are two types of chi-square tests: Chi-square test for goodness-of-fit and Chi-square tests of association and independence.
In this study we use test of independence. Test of independence is used to determine whether the crime and state has independence or not.
So, we have taken three crimes and four states from datasets to predict chi-square.
States are Rajasthan, Gujarat, Bihar and Uttar Pradesh. And three crimes taken are rape, robbery and murder.

DISCUSSION OF P AND CHI-SQUARE VALUE

Step 1:
Prepare final data table to put values in chi-square to predict independence.

State	Rape	Robbery	Murder
Gujarat	4249	14383	13775
Rajasthan	15798	9071	15844
Bihar	13124	23666	41245
Uttar Pradesh	19058	30767	65443

Table 1: table format created before running chi-square test

Hypothesis:

H0: There is no statistically significant relationship between States and Crimes.
Ha: There is a statistically significant relationship between States and Crimes.

Step 2:
Run Chi-square test and calculate expected percentage frequency for each cell

Run Output:

```
Python 3.7.0 (v3.7.0:1bf9cc5093, Jun 27 2018, 04:59:51) [MSC v.1914 64 bit (AMD64)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
==== RESTART: H:\sem-mca-4\project\review_mid_project\Chi-Square-Test.py ====
[[4249, 14383, 13775], [15798, 9071, 15844], [13124, 23666, 41245], [19058, 30767, 65443]]
===Chi2 Stat===
15140.675662959542


===Degrees of Freedom===
6


===P-Value===
0.0


===Contingency Table===
[[ 6352.9995646   9473.97187555 16580.02855985]
 [ 7981.29019266 11902.17597955 20829.53382779]
 [15297.81593556 22813.01556172 39924.16850272]
 [22596.89430717 33697.83658318 58973.26910965]]
>>>
```


Figure 16: output of chi-square test.

Step 3:
Creating observed and expected values table

State	Crime	Observed	Expected
Gujarat	Rape	4249	6352.9995646
	Robbery	14383	9473.97187555
	Murder	13775	16580.02855985
Rajasthan	Rape	15798	7981.29019266
	Robbery	9071	11902.17597955
	Murder	15844	20829.53382779
Bihar	Rape	13124	15297.81593556
	Robbery	23666	22813.01556172
	Murder	41245	39924.16850272
Uttar Pradesh	Rape	19058	22596.89430717
	Robbery	30767	33697.83658318
	Murder	65443	58973.26910965

Table 2: Chi-square Observed and Expected Table

Result

The results show that probability of awarding crime in India.
Here p value is less than (0.5) α alpha, hence can reject H0.
This analysis is done by the help of chi-square to check independence.

Calculating degree of freedom using following the Rule
DOF = (Number of rows−1) ∗ (Number of columns−1)
Degree of freedom df = (4-1) * (3-1) = 6 the following outcomes are drawn

Formula for calculating or prediction chi-square test.

$$X^2 = \frac{(observed - expected)^2}{(expected)}$$

Now we are ready to look into the Chi-squared distribution table. The cut off for a p-value of 0.06 was 12.592. Our $\mathbf{X}^2$ statistic was so large that the p-value is approximately zero. So, we have evidence against the null hypothesis.

With a p-value < 0.05, we can reject the null hypothesis. There is definitely some sort of relationship between 'states' and the ‘crimes' column. We don't know what this relationship is, but we do know that these two variables are not independent of each other

Decision Tree Algorithm

Decision tree implementation using python:

Decision tree is one of the most powerful and popular algorithms. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables

Data Slicing:

Before training the model, we have to split the dataset into the training and testing dataset.
To split the dataset for training and testing we are using the sklearn module *train_test_split.*
First of all, we have to separate the target variable from the attributes in the dataset.

```
X = balance_data.values[:, 2:6]
Y = balance_data.values[:, 0]
```

Above are the lines from the code which separate the dataset. The variable X contains the attributes while the variable Y contains the target variable of the dataset.
Next step is to split the dataset for training and testing purpose.

```
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size = 0.3, random_state = 100)
```

Above line split the dataset for training and testing. As we are splitting the dataset in a ratio of 70:30 between training and testing so we are pass *test_size* parameter's value as 0.3.
random_state variable is a pseudo-random number generator state used for random sampling.

Terms used in code:

Gini index and information gain both of these methods are used to select from the *n* attributes of the dataset which attribute would be placed at the root node or the internal node.

Gini index:

$$\text{Gini Index} = 1 - \sum_j P_j^2$$

Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified.
It means an attribute with lower Gini index should be preferred.
Sklearn supports "Gini" criteria for Gini Index and by default, it takes "Gini" value.

Entropy:

If a random variable x can take N different value, the i^{th} value x, with probability $p(x_i)$, we can associate the following entropy with x:

$$H(x) = -\sum_{i=1}^{N} P(x_i) \log_2 p(x_i)$$

Entropy is the measure of uncertainty of a random variable, it characterizes the impurity of an arbitrary collection of examples. The higher the entropy the more the information content.

Accuracy score:

Accuracy score is used to calculate the accuracy of the trained classifier.

Confusion Matrix:

Confusion matrix is used to understand the trained classifier behaviour over the test dataset or validate dataset.

Data Information:

```
>>>
=============== RESTART: H:\sem-mca-4\project\desicion tree.py ===============
Dataset Lenght:  9017
Dataset Shape:  (9017, 10)
Dataset:           state_ut       district      ...       cheating  dowry_deaths
0  ANDHRA PRADESH       ADILABAD      ...            104            16
1  ANDHRA PRADESH      ANANTAPUR      ...             65             7
2  ANDHRA PRADESH       CHITTOOR      ...            209            14
3  ANDHRA PRADESH       CUDDAPAH      ...             37            17
4  ANDHRA PRADESH  EAST GODAVARI      ...            220            12

[5 rows x 10 columns]
```

Figure 17: Information of data loaded

Result using Gini Index:

```
Results Using Entropy:
Predicted values:
['UTTAR PRADESH' 'NAGALAND' 'PUNJAB' ... 'MADHYA PRADESH' 'UTTAR PRADESH'
 'MADHYA PRADESH']
Confusion Matrix:  [[  0   0   0 ...   0   0   0]
 [  0   0   0 ...  75   0   0]
 [  0   0   0 ...   3   0   0]
 ...
 [  0   0   0 ... 172   0   0]
 [  0   0   0 ...   7   0   0]
 [  0   0   0 ...  52   0   0]]
Accuracy :  16.81448632668145
```

Figure 18: Result using Gini Index.

Report :	precision	recall	f1-score	support
A & N ISLANDS	0.00	0.00	0.00	4
ANDHRA PRADESH	0.31	0.35	0.33	91
ARUNACHAL PRADESH	0.00	0.00	0.00	58
ASSAM	0.00	0.00	0.00	106
BIHAR	0.00	0.00	0.00	162
CHANDIGARH	0.00	0.00	0.00	2
CHHATTISGARH	0.00	0.00	0.00	82
D & N HAVELI	0.00	0.00	0.00	6
DAMAN & DIU	0.00	0.00	0.00	4
DELHI UT	0.00	0.00	0.00	53
GOA	0.00	0.00	0.00	8
GUJARAT	0.09	0.55	0.15	111
HARYANA	0.00	0.00	0.00	70
HIMACHAL PRADESH	0.00	0.00	0.00	53
JAMMU & KASHMIR	0.14	0.51	0.22	88
JHARKHAND	0.00	0.00	0.00	79
KARNATAKA	0.00	0.00	0.00	115
KERALA	0.00	0.00	0.00	75
LAKSHADWEEP	0.00	0.00	0.00	2
MADHYA PRADESH	0.24	0.73	0.36	201
MAHARASHTRA	0.00	0.00	0.00	154
MANIPUR	0.00	0.00	0.00	40
MEGHALAYA	0.00	0.00	0.00	27
MIZORAM	0.00	0.00	0.00	26
NAGALAND	0.00	0.00	0.00	32
ODISHA	0.00	0.00	0.00	126
PUDUCHERRY	0.00	0.00	0.00	8
PUNJAB	0.20	0.51	0.29	105
RAJASTHAN	0.00	0.00	0.00	133
SIKKIM	0.00	0.00	0.00	17
TAMIL NADU	0.00	0.00	0.00	148
TRIPURA	0.00	0.00	0.00	16
UTTAR PRADESH	0.26	0.60	0.36	250
UTTARAKHAND	0.00	0.00	0.00	48
WEST BENGAL	0.00	0.00	0.00	80
micro avg	0.19	0.19	0.19	2580
macro avg	0.04	0.09	0.05	2580
weighted avg	0.07	0.19	0.10	2580

Figure 18.2: Result using Gini Index

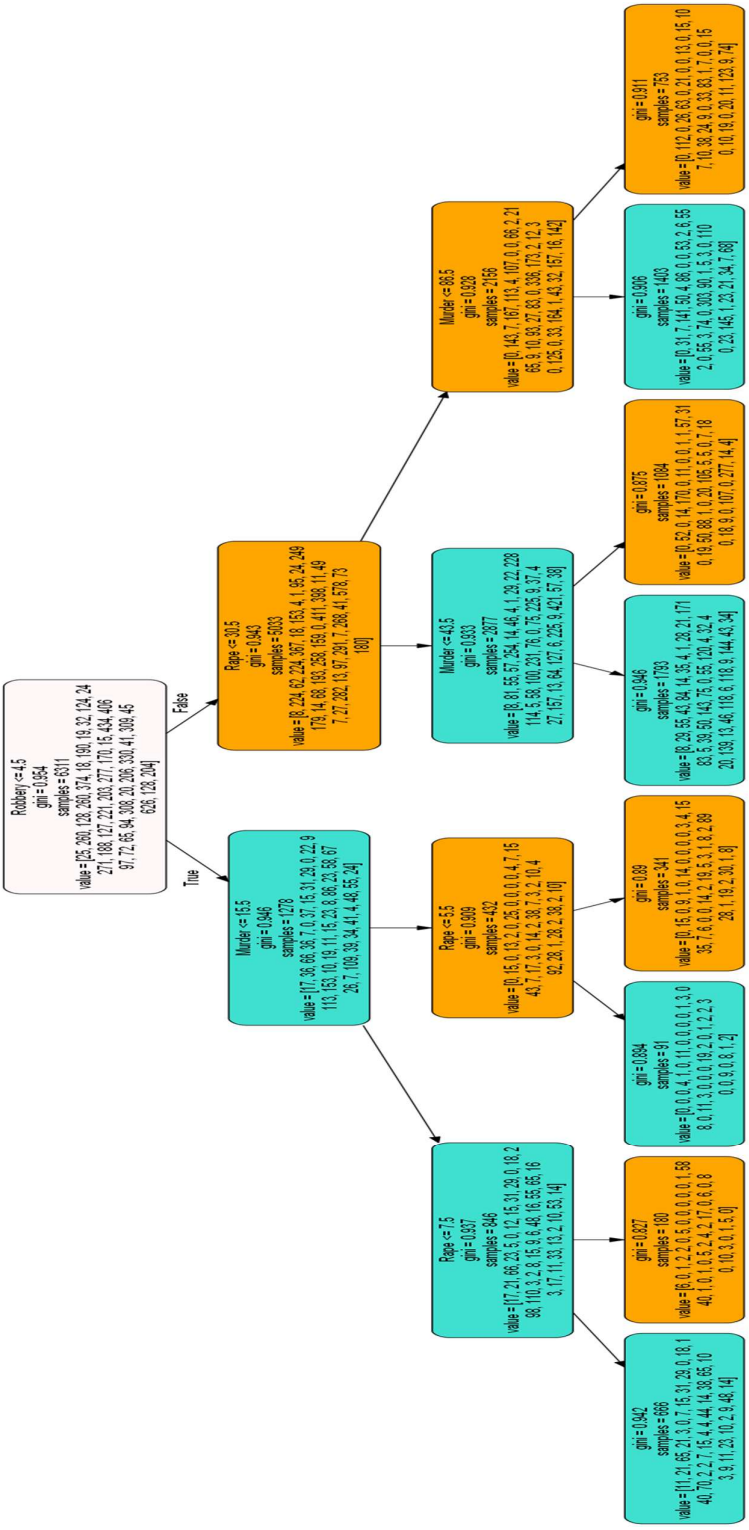


Figure 19: Gini Index Tree graph

Results Using Entropy:

```
Results Using Entropy:
Predicted values:
['UTTAR PRADESH' 'GUJARAT' 'GUJARAT' ... 'GUJARAT' 'ARUNACHAL PRADESH'
 'GUJARAT']
Confusion Matrix:  [[  0   0   1 ...   0   0   0]
 [  0  39  12 ...  16   0   0]
 [  0   0  27 ...   0   0   0]
 ...
 [  0  60   2 ... 101   0   0]
 [  0   1  19 ...   3   0   0]
 [  0  42   6 ...   1   0   0]]
Accuracy :  16.511627906976745
```

Figure 20: Result using Entropy

Report :	precision	recall	f1-score	support
A & N ISLANDS	0.00	0.00	0.00	4
ANDHRA PRADESH	0.10	0.43	0.16	91
ARUNACHAL PRADESH	0.12	0.47	0.19	58
ASSAM	0.00	0.00	0.00	106
BIHAR	0.00	0.00	0.00	162
CHANDIGARH	0.00	0.00	0.00	2
CHHATTISGARH	0.00	0.00	0.00	82
D & N HAVELI	0.00	0.00	0.00	6
DAMAN & DIU	0.00	0.00	0.00	4
DELHI UT	0.00	0.00	0.00	53
GOA	0.00	0.00	0.00	8
GUJARAT	0.09	0.53	0.16	111
HARYANA	0.00	0.00	0.00	70
HIMACHAL PRADESH	0.20	0.40	0.26	53
JAMMU & KASHMIR	0.00	0.00	0.00	88
JHARKHAND	0.00	0.00	0.00	79
KARNATAKA	0.00	0.00	0.00	115
KERALA	0.00	0.00	0.00	75
LAKSHADWEEP	0.00	0.00	0.00	2
MADHYA PRADESH	0.26	0.62	0.36	201
MAHARASHTRA	0.00	0.00	0.00	154
MANIPUR	0.00	0.00	0.00	40
MEGHALAYA	0.00	0.00	0.00	27
MIZORAM	0.00	0.00	0.00	26
NAGALAND	0.00	0.00	0.00	32
ODISHA	0.00	0.00	0.00	126
PUDUCHERRY	0.00	0.00	0.00	8
PUNJAB	0.19	0.52	0.27	105
RAJASTHAN	0.00	0.00	0.00	133
SIKKIM	0.00	0.00	0.00	17
TAMIL NADU	0.00	0.00	0.00	148
TRIPURA	0.00	0.00	0.00	16
UTTAR PRADESH	0.23	0.40	0.30	250
UTTARAKHAND	0.00	0.00	0.00	48
WEST BENGAL	0.00	0.00	0.00	80
micro avg	0.17	0.17	0.17	2580
macro avg	0.03	0.10	0.05	2580
weighted avg	0.06	0.17	0.09	2580

Figure 20.2: Result using Entropy

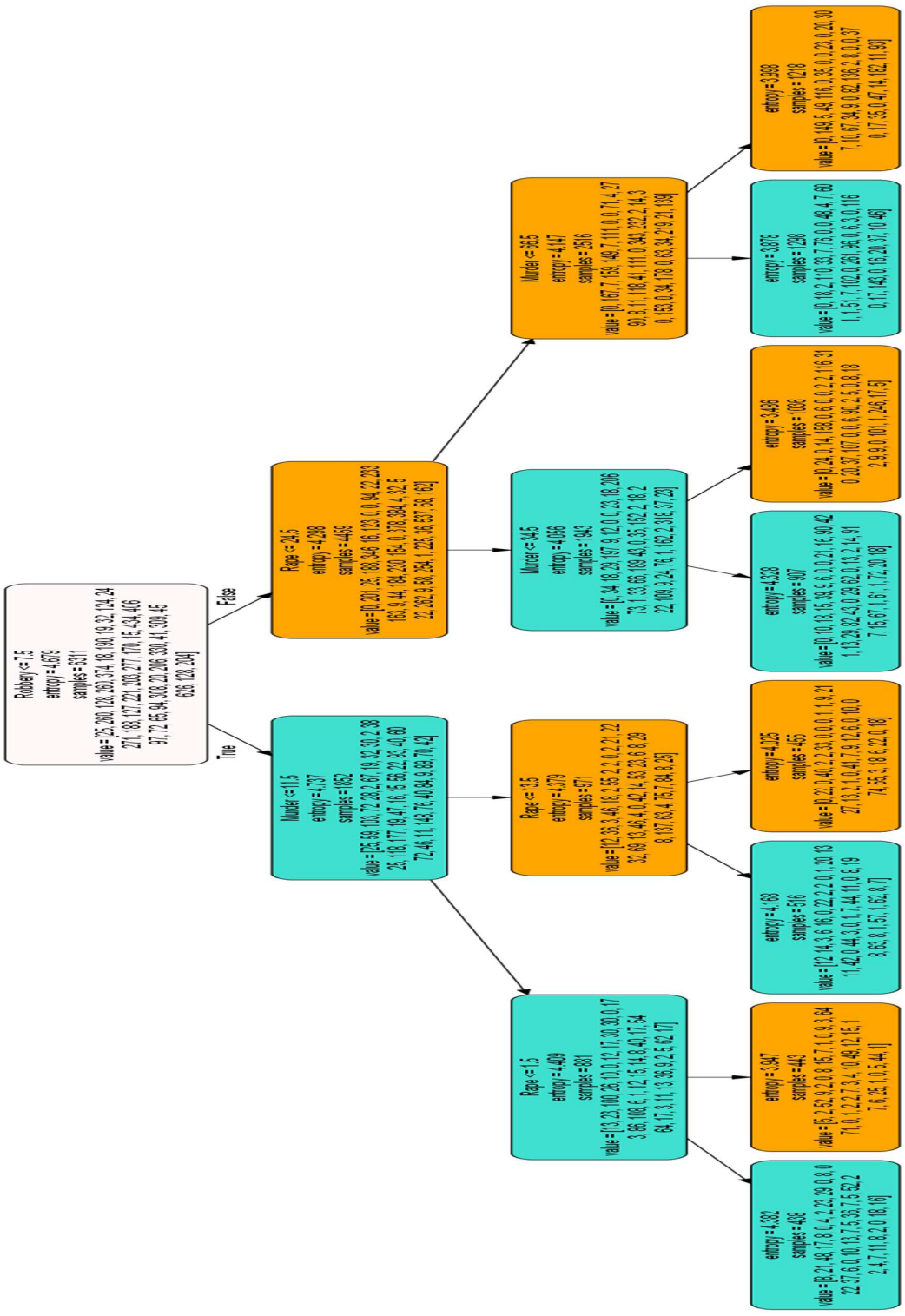


Figure 21: Entropy Tree graph

CONCLUSION

We generated many graphs and found interesting statistics that helped in understanding India crimes datasets that can help in capturing the factors that can help in keeping society safe.
The graphs include Bar, Line, Heatmap, Pari plot, Dist plot, Decision tree and scatter graphs each having its own characteristics.
The objective to test chi-square value at 95% confidence is achieved.
The results of test clearly show that combination of crime according to state.

Through Line graph we can read statistics of Indian IPC Crimes and through this graph we can understand that Theft Crime is very High in India Among All Other Crimes and this crime is Continue Rising from 2001 to 2012 no changes are there in crimes.
Its show that every year theft crime keeps increasing. And this crime is very higher from all other crimes.
We can say that theft crime is increasing because of unemployment in our country.

In Chi Square with a p-value < 0.05, we can reject the null hypothesis. There is definitely some sort of relationship between 'states' and the ‘crimes' column. We don't know what this relationship is, but we do know that these two variables are not independent of each other

Now we are ready to look into the Chi-squared distribution table. The cut off for a p-value of 0.06 was 12.592. Our $\mathbf{X}^2$ statistic was so large that the p-value is approximately zero. So, we have evidence against the null hypothesis.

Using a Decision algorithm, it divides the value and split in small parts and through this we get help for making decision.

Our decision is for control theft and robbery crimes by creating more and more recruitment for jobs. Our thought is that the reason of theft and robbery is because of unemployment in our country.
And by this crimes murder is dependent on all of them.
*

BIBLIOGRAPHY

Dataset: https://www.kaggle.com/rajanand/crime-in-india

Graph :- https://plot.ly/python/ , https://matplotlib.org/3.1.0/tutorials/introductory/pyplot.html#sphx-glr-tutorials- introductory-pyplot-py , https://www.programcreek.com/python/example/96215/seaborn.set

Pandas : http://pandas.pydata.org/pandas-docs/stable/

NumPy:- https://www.numpy.org/devdocs/

Data Modelling (sklearn): https://scikit-learn.org/stable/documentation.html

Decision Tree Algorithm: https://scikit-learn.org/stable/modules/tree.html