A Mini-Project Report

on

Crime Analysis and Prediction Against Women

Submitted in partial fulfillment of the requirements

for the award of degree of

BACHELOR OF TECHNOLOGY

in

Information Technology

by

A. Spoorthi (19WH1A1250)
Riya Fathima (19WH1A1251)
Ch. Supraja (19WH1A1254)
K. Sowmya (19WH1A1256)

Under the esteemed guidance of
Ms. R. Sravani
Assistant Professor

Department of Information Technology

BVRIT HYDERABAD College of Engineering for Women

Rajiv Gandhi Nagar, Nizampet Road, Bachupally, Hyderabad – 500090

(Affiliated to Jawaharlal Nehru Technological University Hyderabad)

(NAAC ‘A’ Grade & NBA Accredited- ECE, EEE, CSE IT)

January, 2023

DECLARATION

We hereby declare that the work presented in this project entitled “**CRIME ANALYSIS AND PREDICTION AGAINST WOMEN**” submitted towards completion of the project in IV year I sem of B.Tech IT at “BVRIT HYDERABAD College of Engineering for Women ”, Hyderabad is an authentic record of our original work carried out under the esteemed guidance of **Ms. R. Sravani, Assistant Professor**, Department of IT.

A. Spoorthi (19WH1A1250)

Riya Fathima (19WH1A1251)

Ch. Supraja (19WH1A1254)

K. Sowmya (19WH1A1256)

VISHNU
UNIVERSAL LEARNING

BVRIT HYDERABAD

College of Engineering for Women

Rajiv Gandhi Nagar, Nizampet Road, Bachupally, Hyderabad – 500090

(Affiliated to Jawaharlal Nehru Technological University, Hyderabad)
(NAAC ‘A’ Grade & NBA Accredited- ECE, EEE, CSE & IT)

CERTIFICATE

This is to certify that the mini-project report on **“CRIME ANALYSIS AND PREDICTION AGAINST WOMEN”** is a bonafide work carried out by **A. Spoorthi (19WH1A1250), Riya Fathima (19WH1A1251), Ch. Supraja (19WH1A1254), K. Sowmya (19WH1A1256)** in the partial fulfillment for the award of B.tech degree in **Information Techonology, BVRIT HYDERABAD College of Engineering for Women, Bachupally, Hyderabad** affiliated to the Jawaharlal Nehru Technological University Hyderabad under my guidance and supervision. The results embodied in the mini-project work have not been submitted to any other university or institute for the award of any degree or diploma.

Internal Guide	**Head of the Department**
Ms. R. Sravani	**Dr. Aruna Rao S L**
Assistant Professor	**Professor & HoD**
Department of IT	**Department of IT**

External Examiner

ACKNOWLEDGEMENT

We would like to express our profound gratitude and thanks to **Dr. K. V. N. Sunitha, Principal, BVRIT HYDERABAD College of Engineering for Women** for providing the working facilities in the college.

Our sincere thanks and gratitude to **Dr. Aruna Rao S L, Professor & Head, Department of Information Technology, BVRIT HYDERABAD College of Engineering for Women** for all the timely support, constant guidance and valuable suggestions during the period of our project.

We are extremely thankful and indebted to our internal guide, **Ms R. Sravani, Assistant Professor , Department of IT, BVRIT HYDERABAD College of Engineering for Women** for her constant guidance, encouragement and moral support throughout the project.

Finally, we would also like to thank our Project Coordinators **Ms S. Rama Devi, Associate Professor and Ms K .Kavitha, Assistant Professor**, all the faculty and staff of the Department of IT who helped us directly or indirectly, parents and friends for their cooperation in completing the project work.

A. Spoorthi (19WH1A1250)

Riya Fathima (19WH1A1251)

Ch. Supraja (19WH1A1254)

K. Sowmya (19WH1A1256)

ABSTRACT

One of the risky features of our society that is steadily intensifying and becoming more complex is crime against women. The main goal of this project is to use clustering and classification algorithms to discriminate between different crimes based on frequency and regularity. Because crime patterns are always evolving, it is challenging to interpret criminal behaviour. According to the amount of crimes committed in each state, the K-means clustering technique is employed in this project to classify the states as safe or dangerous. To forecast crime over the next five years, we also employ linear regression and the ARIMA model. This proposed system can indicate the crime ahead which has a high probability of crime and thus effectively help in significantly reducing the crime rate in various parts of the country.

LIST OF FIGURES

CONTENTS

1. Introduction

Crimes in India are increasing at a very tremendous rate. Over the past years, they have seen plenty of growth and have moved on the way to success, and India is one in every one of them. India is one of the countries which has tried to balance between the advancement and their culture. We Indians pray for women one hand then attempt to suppress their voice on the opposite hand. The rise in the number of crimes against women within the past few decades indicates the statement. In a country where the economy is booming and is growing in each and every particular state and sector. In spite of all this, there has been a huge increase in the number of crimes against women. According to the report of WHO on crime against women published on 29th Nov 2017, one out of each 3 women across the globe faces some crime a minimum of once in their life. So, if we have a tendency to look at the statistics, around 35 you look after the girl bearing this and are mostly done by their partners or knowns.

1.1 Objective

The objective would be to train a model for prediction. The training would be done using the training data set which will be validated using the test dataset. Building the model will be done using a better algorithm depending upon the accuracy. The K-means clustering and a few classification algorithms will be used for crime prediction. Visualization of the dataset is done to analyze the crimes which may have occurred in the particular state. This work helps the law enforcement agencies to predict and detect crimes in India with improved accuracy and thus reduces the crime rate.

1.2 Problem Definition

In country like India crimes against women has been always a serious issue leading to various legal norms and measures against it. With each passing year crime reports are generated leading to huge amount of data. Such data can be used to generate analysis and statistics which will help government and non-government organizations to initiate certain schemes and policies
crime analysis include:

1. Extraction of crime patterns by analysis of available crime and criminal data
2. Prediction of crime based on existing data
3. Detection of crime.

1.3 Aim of the Project

The aim of this project is to make crime predictions using the features present in the dataset. The dataset is extracted from the official sites. With the help of machine learning algorithms, using python as core we can predict the type of crime that will occur in a particular area.

2. Literature Survey

2.1 Related Work

Keerthi.R, et al.[1] The authors have built various analytical process which are data cleaning and processing, Eliminating missing value, exploratory analysis and finally building the model and evaluation to utilize the resources identify the hotspots of crimes and allocate vigilante resources such as policeman, police cars, weapons etc. reschedule patrols according to the vulnerability of a place. Through that they could avoid crimes ensure better civilization through avoiding happening crimes such as murder, rapes, thefts, drug, smugglings etc.

Lavanyaa, et al.[2] The authors have identified the Dominant part of the examination works in Crimes against women by utilizing the 'WEKA Tools' for their usage; in this way they acquired the better outcomes by utilizing MATLAB. Credulous bayes was frequently utilized and mainstream calculation to arrange and anticipate in the ground of information mining. The Next most utilized calculation is Apriori. The abnormal state of precision has been exploited by Naïve Bayes calculation. So as to investigate the better grouping calculations to foresee Crimes in different urban areas and nations, interconnected examine credentials comprise to be assembled. Investigation demonstrates the preeminent piece of utilizing tools and calculations. Regard we found that clustering in WEKA utensils, Euclidean distance calculation gives the improved exactness in the metropolitan urban areas violations rate to decrease and foresee.

Priyanka Das, et al.[3] This paper demonstrates an unsupervised approach of extracting relations from newspapers based on criminological data. The proposed work demonstrates an unsupervised approach of extracting relations from newspapers based on criminological data. The proposed clustering technique identifies significant crime patterns that can help both in the criminology and criminal justice industry and eventually it will help the law enforcement agencies to analyze crime at a faster pace.

B. Sivanagaleela, et al.[4] proposed the project to identify the crime areas based on the clustering technique. They stated that crime patterns identified are not static. So they have identified the crime areas and which type of crime occurred very frequently in which place using the fuzzy clustering technique.

Bhajneet Kaur, et al.[5] In this paper the authors have detected and predicted crime against women, using various data mining techniques by many researchers the authors have used the Indian Crime dataset and few used the techniques on US and England datasets. Most of the techniques used by the researchers are classification and clustering for crime pattern and detection. In the classification some authors depicted the Naïve bayes, some used decision trees and others used Bayesnet, J48, JRip OneR. Correlation and regression techniques are also used for the crime analysis against women. In this paper review has been done on various techniques of data mining used for crime against women.

2.2 Major Issues

1. In some researches most of the work has been implemented using Naive Bayes Algorithm and K-Means Clustering Algorithm.
2. Analysis and Prediction together has not been discussed together.
3. There is no dedicated portal implemented to demonstrate the analysis and prediction of crimes against women in India.

3. System Analysis and Design

3.1 Proposed System

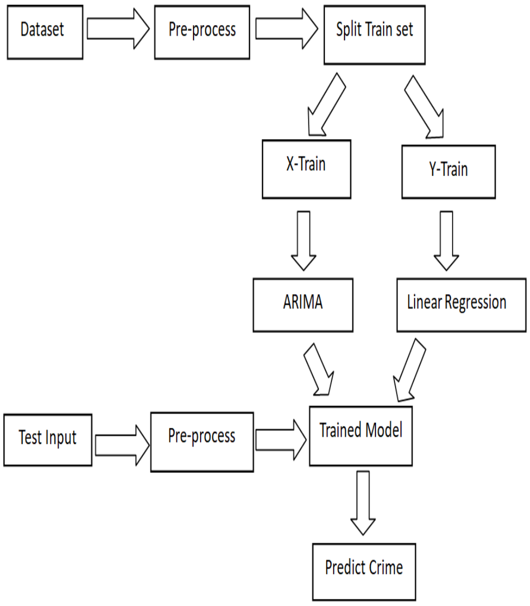


Figure 3.1.1: Block diagram of proposed system

In the proposed system the crimes are predicting based on the location and the types of crimes that are occurring in some areas. Linear Regression and ARIMA model tend to give good accuracy so these models are used in this paper to predict crimes. The data set contains different types of crimes that being committed in India according to the state and year respectively. We take the type of crime as an input and gives the area in which crimes are committed as output. The data pre-processing involves data cleaning, feature selection, dropping null values, data scaling by normalizing and standardizing. After data preprocessing the data is free of null values which m ay alter the accuracy of the model significantly . After data pre-processing the models chosen i.e., Linear Regression and ARIMA are trained by splitting the data into as train and test data. As the output required is a categorical value classification models are used here. Python language is used for the data prediction.

3.2 Architecture Design

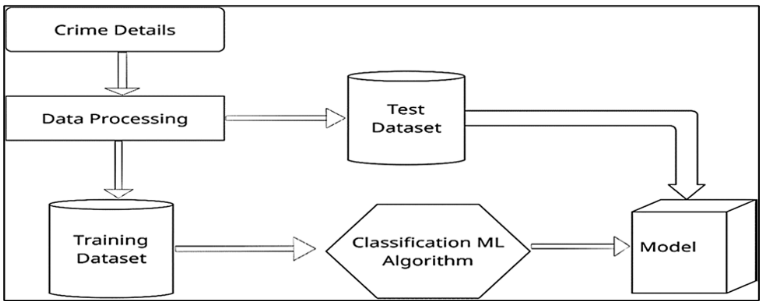


Figure 3.2.1: Architecture

3.3 Activity Diagram

The activity diagram is used to demonstrate the flow of control within the system rather than the implementation. It models the concurrent and sequential activities. The activity diagram helps in envisioning the workflow from one activity to another. It emphasized the condition of flow and the order in which it occurs. The flow can be sequential, branched, or concurrent.

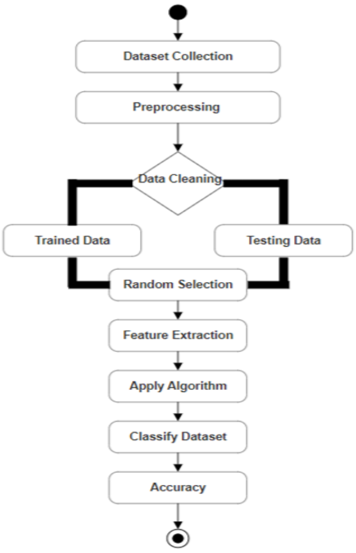


Figure 3.3.1: Activity Diagram

3.4 Use case Diagram

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

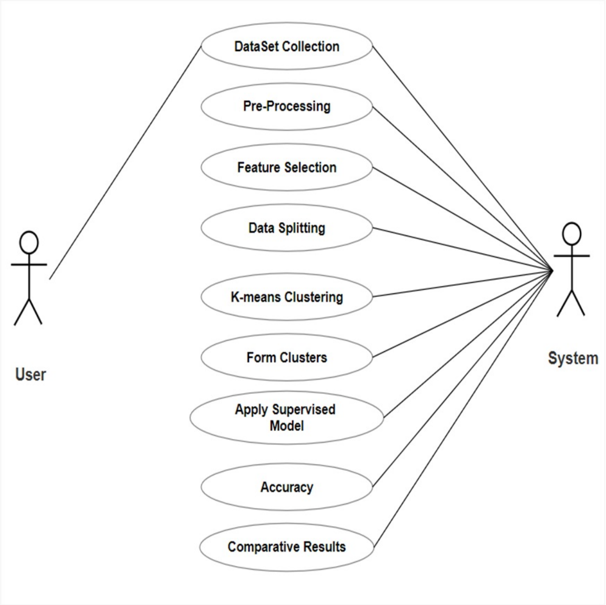


Figure 3.4.1: Use case Diagram

4. Implementation

In the proposed work the data is first pre-processed by removing the missing value and columns which are not needed for the analysis and prediction model. Then we are splitting the data into X and Y attributes, where Y contains column class and its corresponding row entities and X attribute contains all other column and their corresponding row entries with the help of sklearn split data library. The test and train data is split into 3:1 ratio in which the training data set size would be 0.80 of the total dataset size and the testing dataset size would be 0.20 of the total data set. After dividing the training and testing dataset, Machine learning algorithms ARIMA model and Linear Regression which are compatible and relevant for our datasetwas applied. ARIMA was implemented to the train data which was 80 percentage of the total dataset. Then , we tested the model using the test data which consists of 20 percentage of the total dataset . Accuracy score of the predictive model was observed to be 0.7692307692307693 which was 70.923 percentage accuracy. Linear Regression was implemented on the prediction model and it was observed an accuracy of 83 percentage. Linear Regression gave better accuracy result when compared with ARIMA model.

4.1 Module

1. Data Collection
2. Data Preprocessing
3. Data Analysis
4. Training Model
5. Testing Model
6. Results

4.1.1 Data Collection

Data collection is a process in which information is gathered from many sources which is later used to develop the machine learning models. The data should be stored in a way that makes sense for problem. In this step the data set is converted into the understandable format which can be fed into machine learning models.

This dataset contains statistics of various crimes cases against women from the year 2001-2014 with State or Union Territories. The dataset contains the various crimes stats like Rape, Domestic Violence,Dowry, etc. with the year and place of the crime.

4.1.2 Data Pre-processing

Data pre-processing includes methods to remove any anomalies in data such as missing value, inconsistent values or duplicate values.There were many repeated rows. In some places, different letter cases were used for the same word as in Assam and ASSAM. Other places have improper spacing, and shorthands were used for the State/UT. Delhi was repeated with the suffix UT.Some rows contain the total for that particular district which we don't want. So we are dropping all the columns which contain the word total.

4.1.3 Data Analysis

Using matplotlib library from sklearn, analysis of the crime dataset is done by plotting various graphs These visualizations would provide the comparative study between the various crimes as well as between crimes in different region.

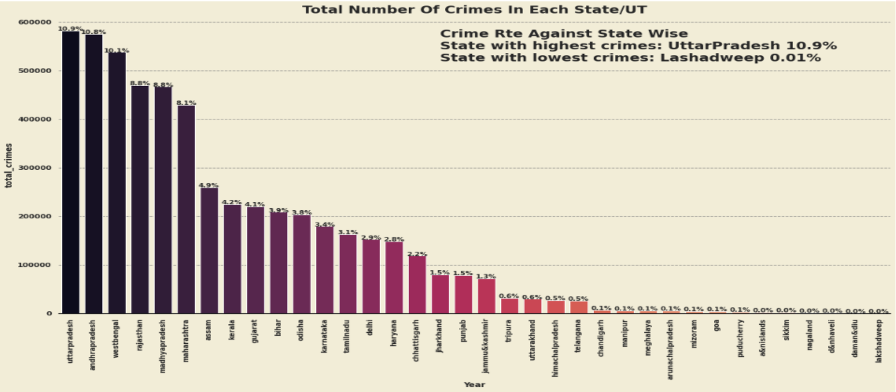


Figure 4.1.3.1: Data Analysis

4.1.4 Training Model

This method divides the dataset into train and test sets randomly. The data is split in the ratio 8:2 such that 80 percentage of the data is to be used to train the prediction model and the subsequent 20 percentage is used to test predictions of the model

4.1.5 Testing Model

Once we train the model with the training dataset, it's time to test the model with the test dataset. This dataset evaluates the performance of the model and ensures that the model can generalize well with the new or unseen dataset. At this stage, we can also check and compare the testing accuracy with the training accuracy, which means how accurate our model is with the test dataset against the training dataset.

4.1.6 Results

Using k-means clustering total districts are divided into three clusters. Using ARIMA and Linear Regression models we predict the crime rates in coming four years.

4.2 Proposed Algorithms Names

4.2.1 K-Means Clustering

For clustering, k-means clustering algorithm is used which is unsupervised technique. The K-means clustering algorithm uses iterative refinement to produce a final result. The algorithm inputs are the number of clusters and the data set. The dataset is a collection of features for each data point. The dataset consist of

States and union territories along with its districts, seven different crimes such as rape,dowry deaths, assault, insult to modesty, importation of girls, kidnapping, Cruelty by husband and his relatives. The algorithms starts with initial estimates for the centroids, which can either be randomly generated or randomly selected from the data set. So here the number of clusters, k=3. Three different clusters are formed based on the number of crimes in that particular place. The districts with highest number of crime rate will be grouped together in a cluster, similarly the one with lowest and average number of crimes. The algorithm then iterates between two steps:

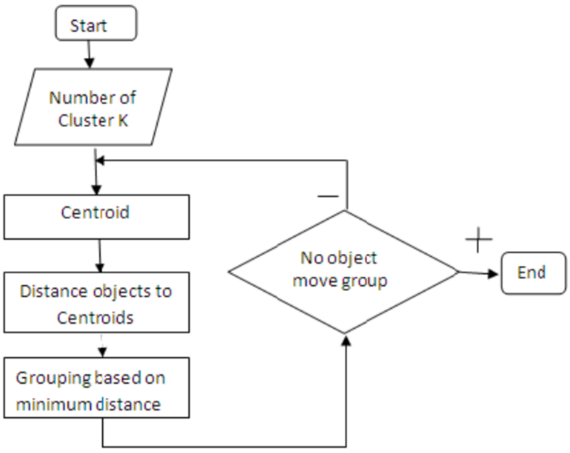

Start
Number of Cluster K
Centroid
Distance objects to Centroids
Grouping based on minimum distance
No object move group
−
+
End

Figure 4.2.1.1: Flow Chart of K-Means Algorithm

(a) Data assignment step: Each centroid defines one of the clusters. In this step,each data point is assigned to its nearest centroid, based on the squared Euclidean distance. Since the data is present in the form of 1d array the y coordinate becomes zero and the distance can be calculated by just finding the difference between the centroids and the dataset values.

(b) Centroid update step: In this step, the centroids are recomputed. This is done by taking the mean ofall data points assigned to that centroid's cluster. The algorithm iterates between steps one and two until a stopping criteria is met(i.e., no data points change clusters, the sum of the distances is minimized, or some maximum number of iterations is reached). Here the maximum number of iteration is kept one. This algorithm is guaranteed to converge to a result.The result may be a local optimum (i.e. not necessarily the best possible outcome), meaning that assessing more than one run of the algorithm with randomized starting centroids may give a better outcome.

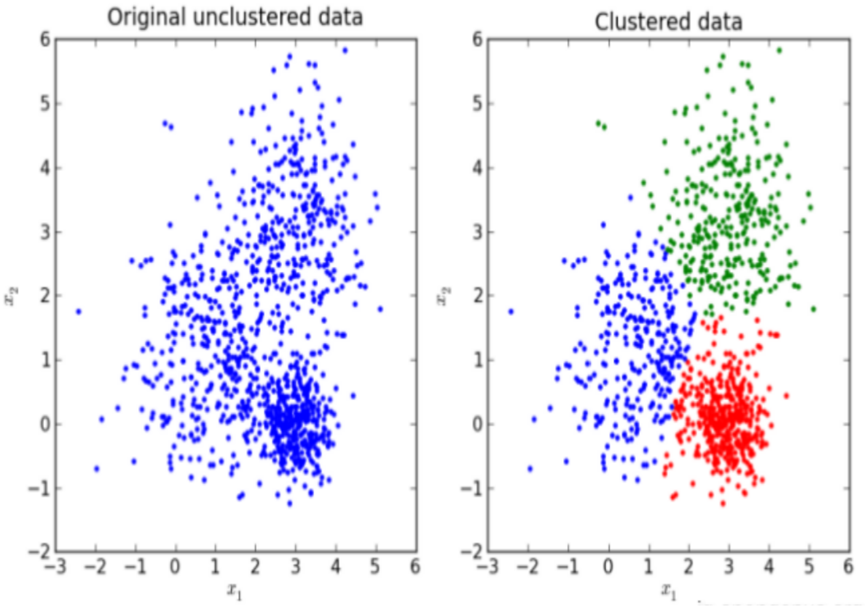


Figure 4.2.1.2: K-Means Clusters

4.2.2 Linear Regression

In order to forecast the crime rate for future years, linear regression technique is being used. This technique consists of a dependent and an independent variable. The linear regression line has an equation of the form y = mX + c, where m is the slope of the line, c is the coefficient of the line, X is the independent variable and y is the dependent variable. Here, the independent variable (X) is Year and the dependent variable (Y) will be the rate of specific crime from the dataset. The core idea is obtaining a line that best fits the data which can be used to predict any new feature value. The best fit line is the one for which total prediction error (all data points) are as small as possible. This best fitting line is called the regression line.

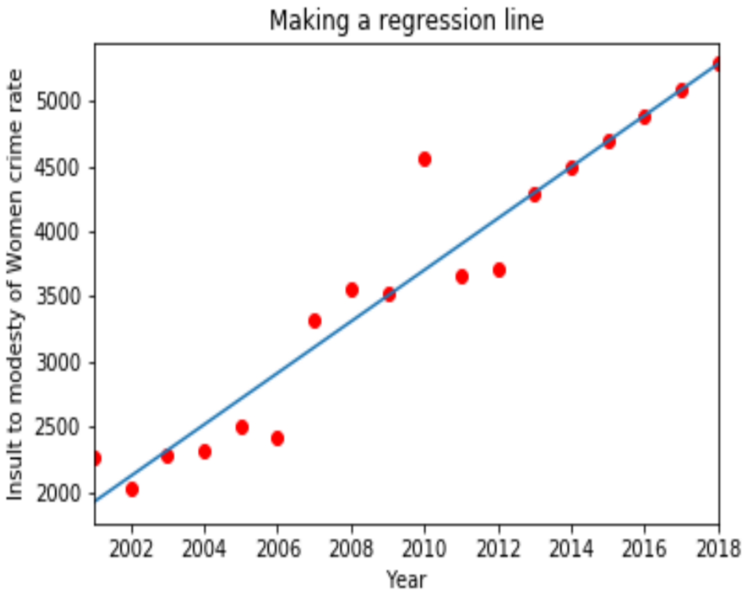


Figure 4.2.2.1: Linear Regression

4.2.3 Data Visualization

Data visualization is essential for representing insights from data in a graphical manner. With the large amount of data in dataset, one of the greatest challenge is to easily communicate the hidden patterns and findings in an easy and understandable manner. To visualise the data, there are many visualisation techniques available. Some of techniques that we are utilizing for the project are line chart, bar chart, correlation matrix, scatterplot.

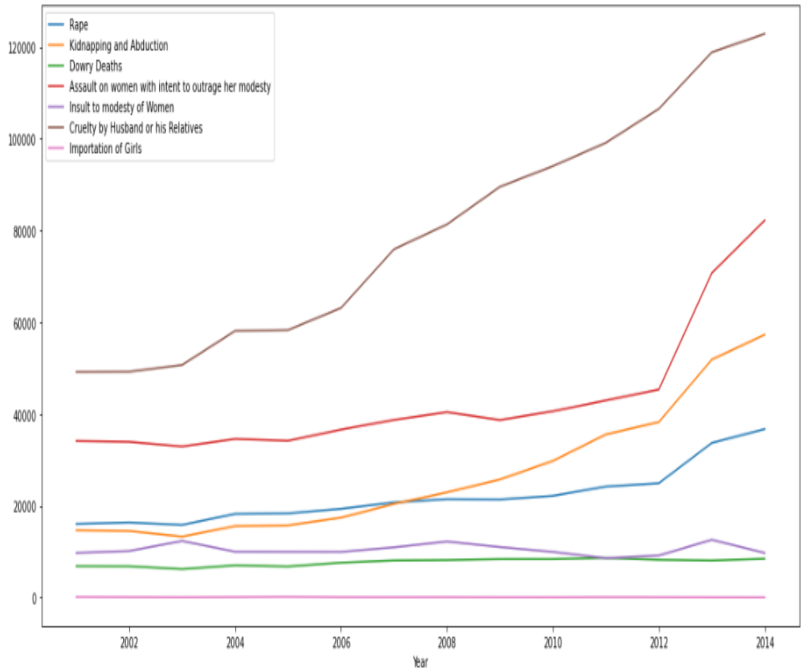


Figure 4.2.3.1: Data Visualization

4.2.4 Arima Model

A type of statistical models for analysing and forecasting time series data is the Autoregressive Integrated Moving Average Model. It conforms to conventional data structures and hence provides a simple yet effective way for forecasting, hence by this model we can use analyse the data and predict the area in which this crime may probably occur with the features selected. ARIMA is a method for forecasting or predicting future outcomes based on a historical time series. It is based on the statistical concept of serial correlation, where past data points influence future data points. The “AR” in ARIMA stands for autoregression, indicating that the model uses the dependent relationship between current data and its past values. In other words, it shows that the data is regressed on its past values. The “I” stands for integrated, which means that the data is stationary. Stationary data refers to time-series data that's been made "stationary" by subtracting the observations from the previous values. The "MA" stands for moving average model, indicating that the forecast or outcome of the model depends linearly on the past values. Also, it means that the errors in forecasting are linear functions of past errors. Note that the moving average models are different from statistical moving averages.

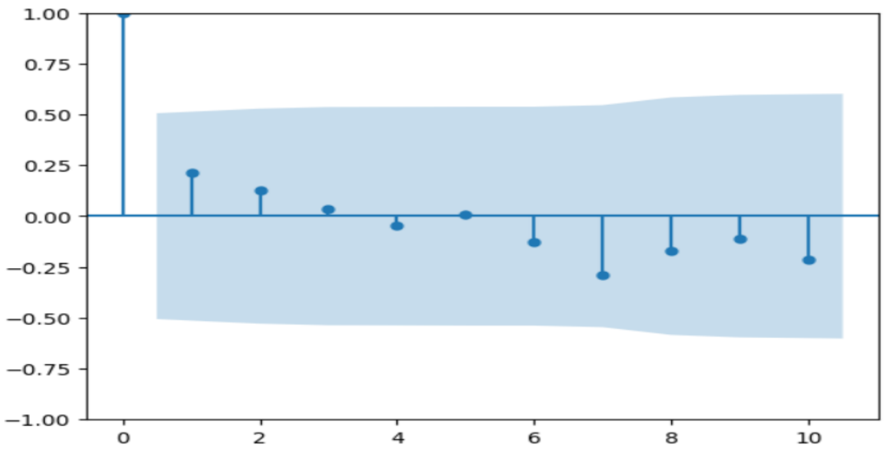


Figure 4.2.4.1: AutoCorrelation

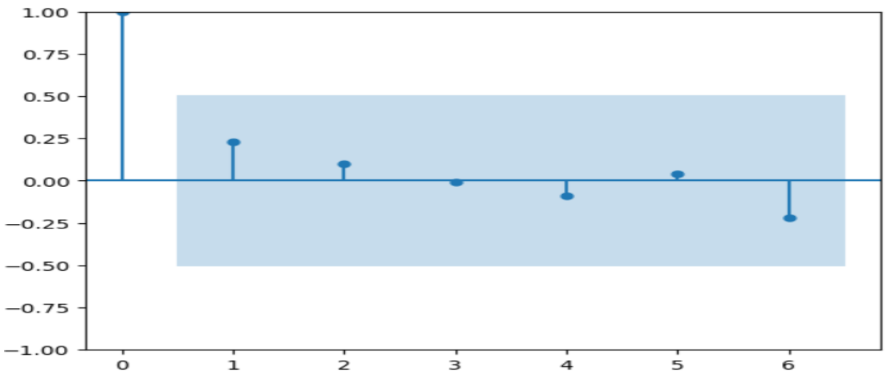
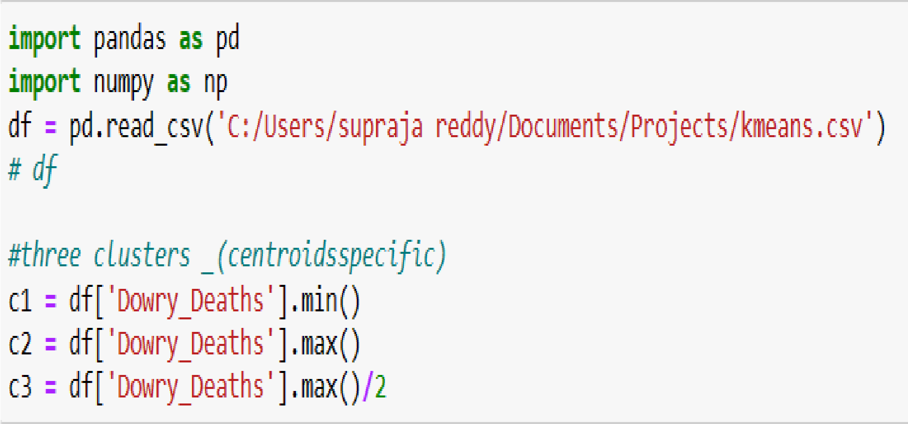


Figure 4.2.4.2: Partial AutoCorrelation

4.3 Results

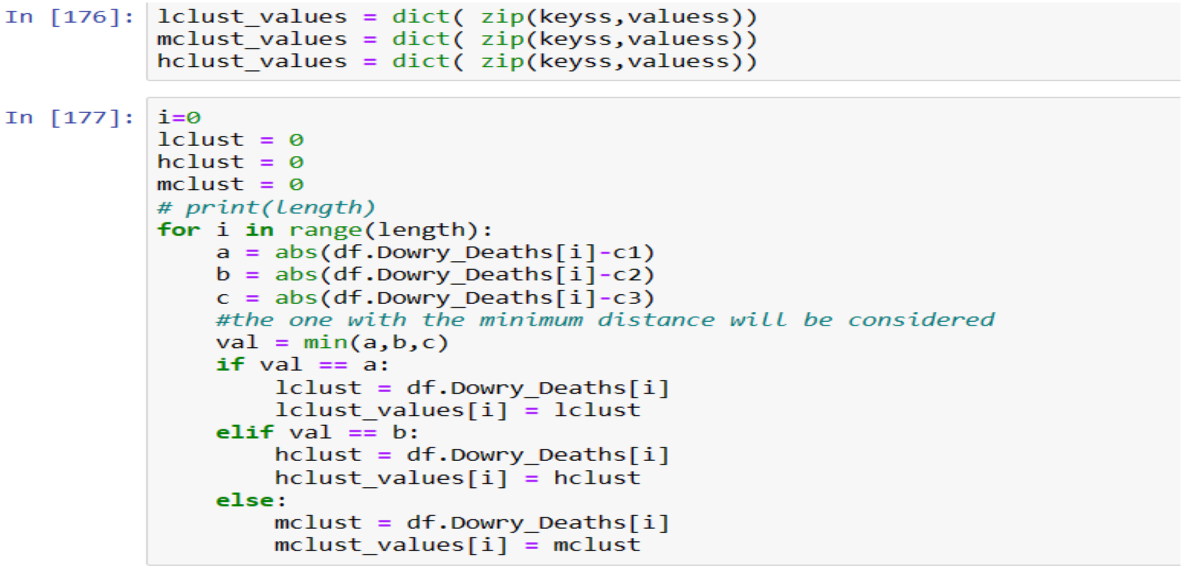
Using k-means clustering total districts are divided into three clusters i.e Red zone, Green zone and blue zone. The red zone indicates maximum crime rate in that particular area, green zone indicates minimum number of crime rates and blue zone indicates average number of crime rates. Using ARIMA and Linear Regression models we predict the crime rates in coming four years.



```
import pandas as pd
import numpy as np
df = pd.read_csv('C:/Users/supraja reddy/Documents/Projects/kmeans.csv')
# df

#three clusters _(centroidsspecific)
c1 = df['Dowry_Deaths'].min()
c2 = df['Dowry_Deaths'].max()
c3 = df['Dowry_Deaths'].max()/2
```

Figure 4.3.1: Implementation of K-Means Clusters



```
In [176]: lclust_values = dict( zip(keyss,valuess))
          mclust_values = dict( zip(keyss,valuess))
          hclust_values = dict( zip(keyss,valuess))

In [177]: i=0
          lclust = 0
          hclust = 0
          mclust = 0
          # print(length)
          for i in range(length):
              a = abs(df.Dowry_Deaths[i]-c1)
              b = abs(df.Dowry_Deaths[i]-c2)
              c = abs(df.Dowry_Deaths[i]-c3)
              #the one with the minimum distance will be considered
              val = min(a,b,c)
              if val == a:
                  lclust = df.Dowry_Deaths[i]
                  lclust_values[i] = lclust
              elif val == b:
                  hclust = df.Dowry_Deaths[i]
                  hclust_values[i] = hclust
              else:
                  mclust = df.Dowry_Deaths[i]
                  mclust_values[i] = mclust
```

Figure 4.3.2: Implementation of K-Means Algorithm

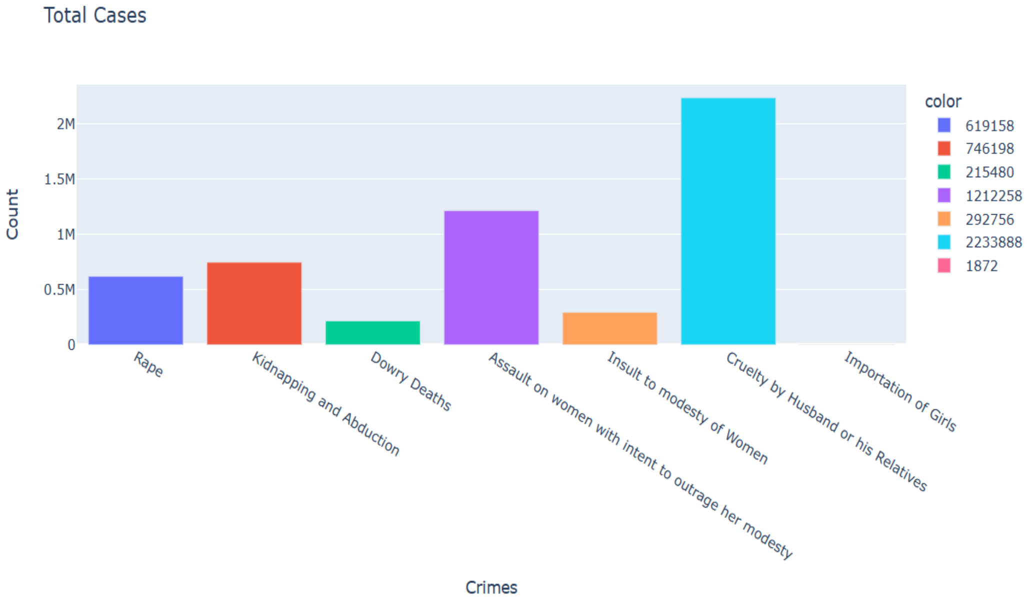


Figure 4.3.3: Data Visualization of Various Crimes

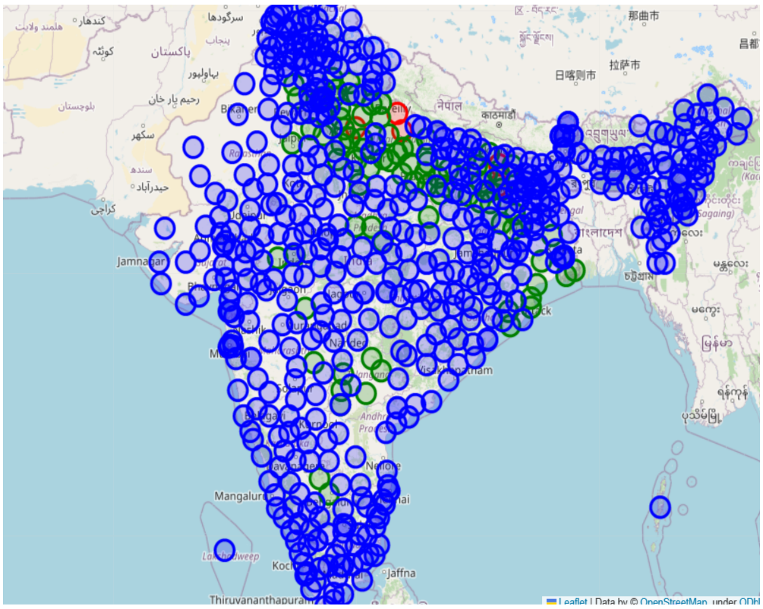


Figure 4.3.4: Crime Analysis Using K-Means Algorithm

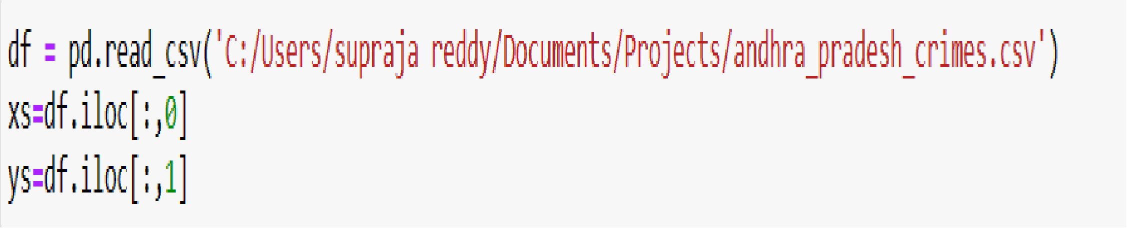
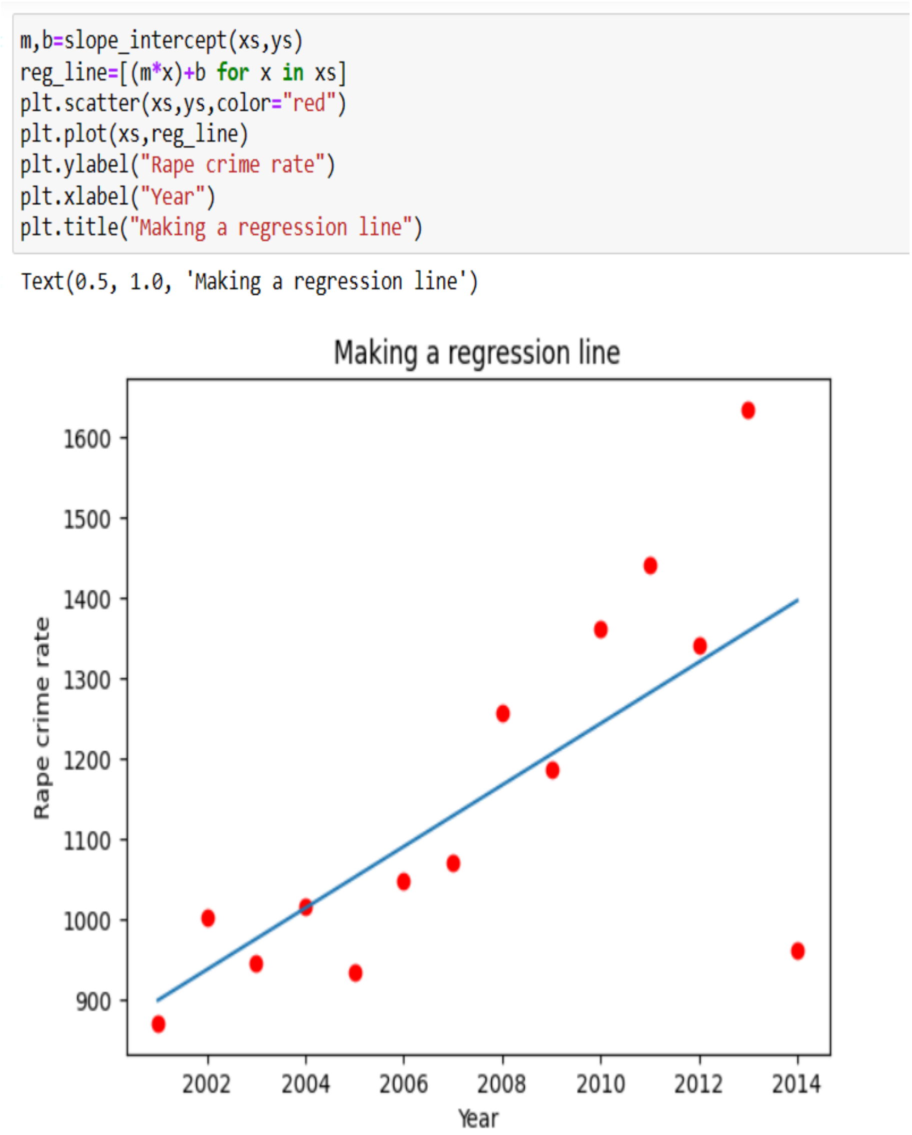

```
df = pd.read_csv('C:/Users/supraja reddy/Documents/Projects/andhra_pradesh_crimes.csv')
xs=df.iloc[:,0]
ys=df.iloc[:,1]
```

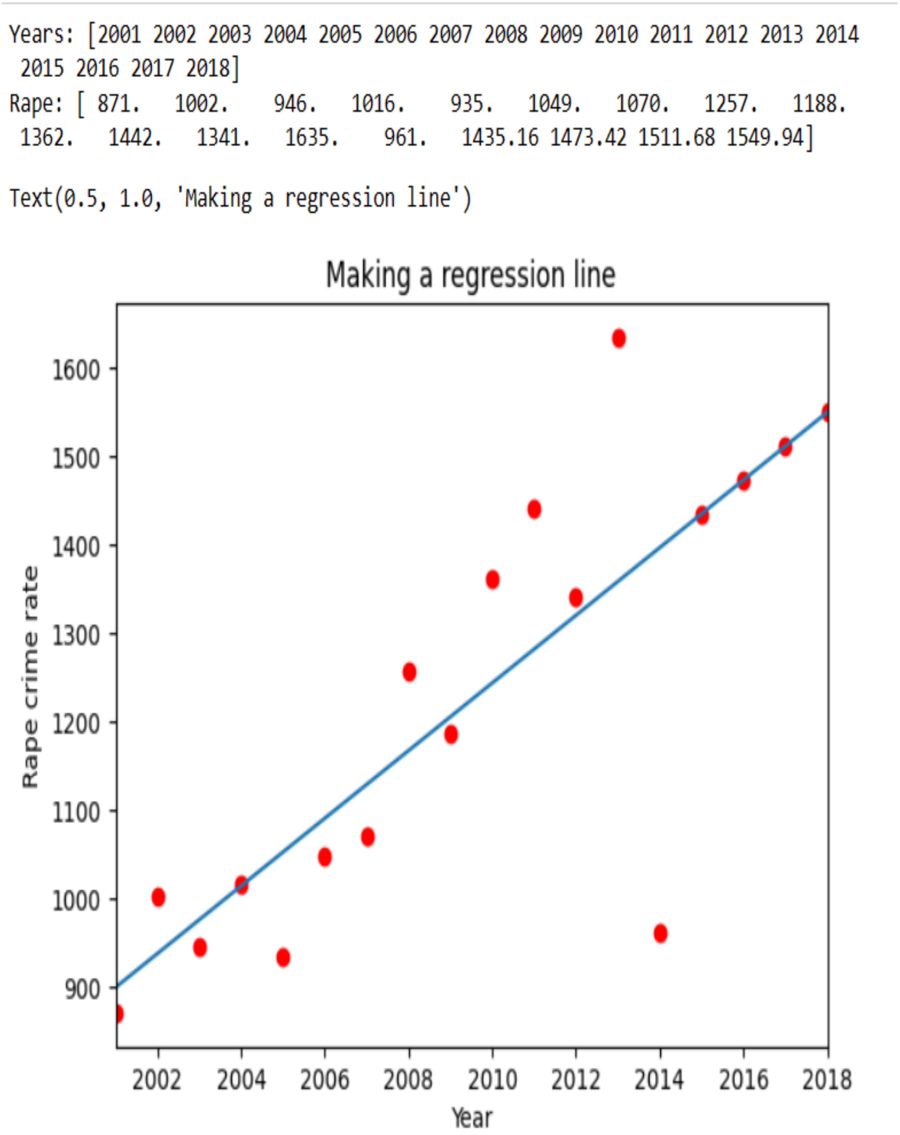
Figure 4.3.5: K-Means Algorithm



```
m,b=slope_intercept(xs,ys)
reg_line=[(m*x)+b for x in xs]
plt.scatter(xs,ys,color="red")
plt.plot(xs,reg_line)
plt.ylabel("Rape crime rate")
plt.xlabel("Year")
plt.title("Making a regression line")
```

```
Text(0.5, 1.0, 'Making a regression line')
```

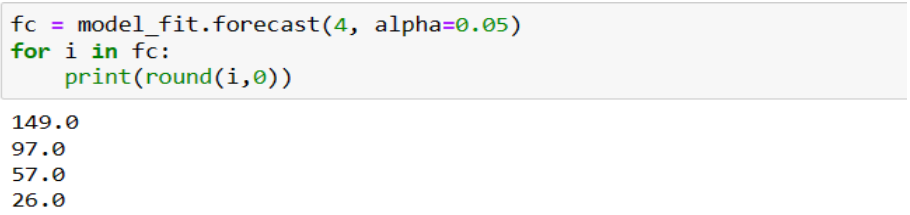
Figure 4.3.6: Rape Crime Before Prediction



```
Years: [2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
 2015 2016 2017 2018]
Rape: [ 871.    1002.     946.    1016.     935.    1049.    1070.    1257.    1188.
 1362.    1442.    1341.    1635.     961.    1435.16 1473.42 1511.68 1549.94]

Text(0.5, 1.0, 'Making a regression line')
```

Figure 4.3.7: Rape Crime After Prediction



```
fc = model_fit.forecast(4, alpha=0.05)
for i in fc:
    print(round(i,0))
```

```
149.0
97.0
57.0
26.0
```

Figure 4.3.8: Crime Prediction using ARIMA model

5. Conclusion and Future Enchancements

Patterns related to the crimes in specific regions can be identified which can help concerned authorities to take preventive measures. Mapping by area in relation to the type of violence helps to prepare better strategies to prevent specific violence. The trends observed will help in better decision making and reducing the future crime against women. Analysis and prediction of crime by clustering and regression will help to identify the regions where specific crimes are more frequent as well as the crime rate, which in turn can be used by concerned authorities to focus on those areas.

The proposed project mainly focuses on emphasizing the crime rates around a particular country (India). So as per the future vision other part of the world can be taken into consideration. Right now it is specifically bound to a single gender (Women). Other genders can be included further. There are certain crimes and cases which are unheard and unregistered around the globe and if they are taken into consideration the accuracy of the crime rates can be improved. Currently the dataset consist of only seven crimes but this can expanded to include many more crimes in the future.

REFERNCES

[1] Keerthi.R, Kirthika.B, Pavithra.S , Dr. V.Gowri, ”Prediction of Crime Rate Analysis using Machine Learning Approach”, IRJET, 2020.

[2] Lavanya, D. Akila, ”Predicting Crimes against Women’s and Criminal Performance in Tamilnadu State”, IEEE, 2020

[3] Priyanka Das, Asit Kumar Das, Janmenjoy Nayak, Danilo Pelusi, ”Framework for crime data analysis using relationship among named entities”, DBLP, 2020.

[4] B.Sivanagaleela, S.Rajesh, Çrime Analysis and Prediction Using Fuzzy C-Means Algorithm”,IRJET, 2019.

[5] Bhajneet Kaur, Laxmi Ahuja, Vinay Kumar, Çrime Against Women:Analysis and Prediction using Data Mining Techniques”, IEEE, 2019.