Jaypee Institute of Information Technology

Department of Computer Science and Engineering

MINOR PROJECT-1

Project Title: Crime Forecasting

Group Number: 41

Project Report till 20th October 2021

Group Members:
Shubham Garg (9919103057)
Pranjal Tiwari (9919103010)
Rahul Sharma (9919103173)

Mentor:
Dr. Shilpa Budhkar

INTRODUCTION

In 1947, India got its independence from the British Colony, with the largest democracy and second most populated country in the world. Crime is still the major concern in the country until now; day by day it keeps on increasing. In 2018, according to the NCRB report, they found that crime rates in India have increased by 1.3%. There had been a major surge in relation to abduction and kidnapping across the country; this increased by 10.3% over the previous years. Therefore, it is important to analyze the major crimes (abduction, kidnapping, rape, theft, etc) happening in every major state and their cities. This can help Law and Enforcement to analyze the trend behind the crime and in the future, they can prevent or stop it from happening. The crimes in urban areas have been rising, national crime records bureau released the data where cites in the northern state's records two times the higher crimes rates than the southern urban agglomerations of India. Crimes against women are the major concern for the government, , violent crimes against women including rape are steadily rising every passing year India's GINI coefficient that has increased from 0.32 to 0.38 in the last two decades. In 2012, the crimes against women reported by official statistics increased by 24.7%, compared to those reported in 2008.

Out of 28 states, 10 states reported more than 10000 cases in 2011. According to the report, 3,78,277 cases of crime against women were reported in the country, up from 3,59,849 in 2017. Uttar Pradesh topped the list with 59,445 cases, followed by Maharashtra (35,497) and West Bengal (30,394). White-collar crime has acquired new dimensions. Political institu-tions have changed very rapidly and cultural norms have not kept pace with them. Hence, there is a ‘cultural lag’ in today’s India.

Power has also become a source of crime for the privileged sections of society. There is a tendency among powerful persons to abuse their influence and authority. Several cases of rape and murder have been reported by wards of influential persons and political heavyweights in the recent past. White-collar crime is a phenomenon found among educated people engaged in trade, professions, and government services.

PROBLEM STATEMENT

The sole intention behind the consideration of this project is to make a Crime Forecasting Model and forecast the crime beforehand. This analysis will help them to understand that which state or city safer for the tourist, and which are unsafe, areas where the government needs to improve the situation.

Top cities will be clustered according to most crimes committed in India this will help to identify the cities within the cluster. The defined cluster will divide the top cities with respect to the count of crimes committed, these subgroups will help the police and the government to take the necessary measure to ensure the safety of the public. Forecasting or prediction of the crime is necessary for every state because it will help the local law enforcement to manage the crimes according to it.

OBJECTIVES:

- To provide aggregate statistics from the dataset such as the highest crime areas.
- To represent key data and findings using a suitable visualization method and tool.
- To cluster applicable data.
- To consider follow on research based on the findings of this work.
- To forecast the trend of crimes for all the states for the next 6 years

ISSUES

a) The total number of crimes and the crime rate has increased. Major reason for the increase in crime due to the population of Indian society has been increased and there is a rapid expansion in industrialization and urbanization as well over the 40 years.

b) One of the oldest civilizations in the world, India is one of the most popular destinations on the planet earth, In June 2019 with over 720 thousand tourists visiting the country. According to the statistics the share of foreigners spending in the year 2017 was around 87% and it is expected to increase by 1% till 2028. Safety of the tourist is a major concern for the government

BACKGROUND STUDY

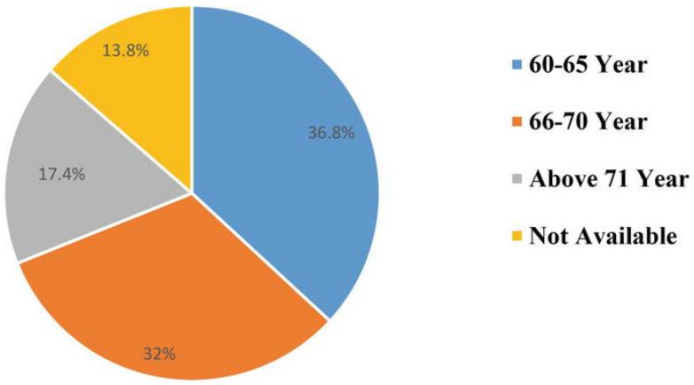
The total number of crimes and the crime rate has increased, Major reason for the increase in crime due to the population of Indian society has been increased and there is a rapid expansion in industrialization and urbanization as well over the 40 years.

Cases of Burglary has been decreased by 79.84% over the 53 years, but the murder rate has been increased by 7.39%. Robbery and riots have been declined by 28.85% and 10.58% respectively, but kidnapping has been increased by 47.8%.

The location has played a significant role in crimes in India:

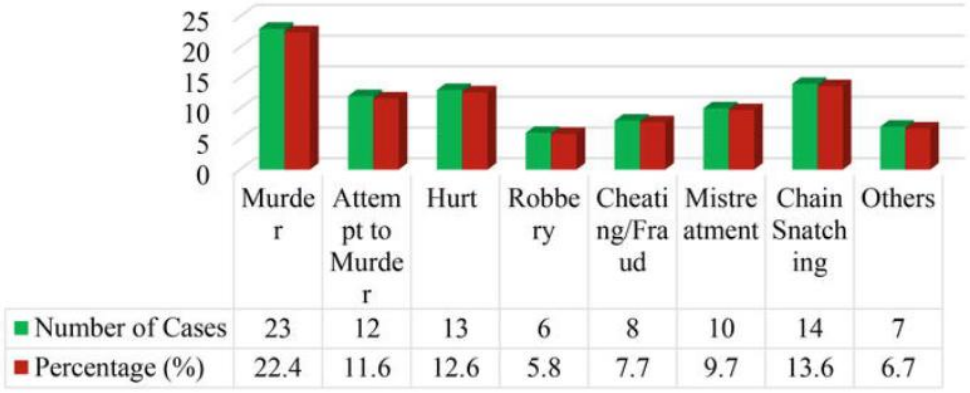
- In 2006, the highest crime rate was reported in Puducherry (447.7%) for crimes under the Indian Penal Code which is 2.7 times the national crime rate of 167.7%.
- Kerala reported the highest crime rate at 312.5% among states.
- Delhi, Mumbai, and Bengaluru have accounted for 16.2%, 9.5%, and 8.1% crime rate 769.1% among the megacities in India followed by Madhya Pradesh (Bhopal) 719.5% and Rajasthan(Jaipur) 597.1%.
- Uttar Pradesh reported the highest incidence of violent crimes accounting for 12.1% of total violent crimes in India (24,851 out of 2, 05,656) followed by Bihar with 11.8% (24,271 out of 2,05,6556).
- Among 35 megacities, Delhi reported 31.2% (533 out of 1,706) of total rape cases. Madhya Pradesh has reported the highest number of rape cases (2,900) accounting for 15.0% of total such cases reported in the country.
- Uttar Pradesh reported 10% (5,480 out of 32,481) of total murder cases in the country and 18.4% (4,997 out of 27,230) total attempts to murder cases.

Today in India, crimes cases against Elderly women is on a rise mainly women are facing problem such as murder, theft, hurt, bag snatching.



Age of Women

Women of age more 60 to 65 have been victimized the most, the chart suggests that 36.8% of women are the most affected.



	Murder	Attempt to Murder	Hurt	Robbery	Cheating/Fraud	Mistreatment	Chain Snatching	Others
Number of Cases	23	12	13	6	8	10	14	7
Percentage (%)	22.4	11.6	12.6	5.8	7.7	9.7	13.6	6.7

Nature of Crime

The nature of the crimes against the women is mostly murder with 22.4% followed by chain snatching and Hurt with 13.6% and 12.6% respectively.

DATASET USED

The data set used for this project is "**District-wise crimes under various sections of Indian Penal Code (IPC) crimes**". This dataset was obtained from the website data gov.in(Ministry of Home Affairs India, 2016). The short description of the variables used in the data set is in Table.

This dataset contains the count of all major crimes that were reported in India from 2001 to 2012. All the crimes that are recorded are for every state and their respective districts.

Feature name	Feature description
STATE/UT	Text Value which denotes all the Name of the state and Union Territories
DISTRICT	Text Value which covers all the urban and rural cities of every state in India
YEAR	Numerical Value denoting the year's when the crimes have occurred from 2001 to 2013
THEFT	Numerical Value that denotes the count of theft that happened in India
ATTEMPT TO MURDER	Numerical Value that denotes the count of attempt to murder that happened in India
CULPABLE HOMICIDE NOT AMOUNTING TO MURDER	Numerical Value that denotes the count of culpable homicide not amounting to murder
RAPE:	Numerical Value that denotes the count of Rape that happened in India
CUSTODIAL RAPE	Numerical Value that denotes the count of Custodial Rape that happened in India
OTHER RAPE:	Numerical Value that denotes the count of Other Rape that happened in India
KIDNAPPING & ABDUCTION	Numerical Value that denotes the count of Kidnapping and Abduction that happened in India
KIDNAPPING AND ABDUCTION OF WOMEN AND GIRLS	Numerical Value that denotes the count of Kidnapping and Abduction of Women and Girls that happened in India

KIDNAPPING AND ABDUCTION OF OTHERS:	Numerical Value that denotes the count of Kidnapping and Abduction of Others that happened in India
DACOITY	Numerical Value that denotes the count of Dacoity that happened in India
PREPARATION AND ASSEMBLY FOR DACOITY	Numerical Value that denotes the count of PREPARATION AND ASSEMBLY FOR DACOITY that happened in India
ROBBERY	Numerical Value that denotes the count of Robbery that happened in India
BURGLARY	Numerical Value that denotes the count of Burglary that happened in India
THEFT	Numerical Value that denotes the count of Theft that happened in India
AUTO THEFT	Numerical Value that denotes the count of Auto theft that happened in India
OTHER THEFT	Numerical Value that denotes the count of Other theft that happened in India
RIOTS	Numerical Value that denotes the count of Riots that happened in India
CRIMINAL BREACH OF TRUST	Numerical Value that denotes the count of Criminal breach of trust that happened in India
CHEATING	Numerical Value that denotes the count of Cheating that happened in India
COUNTERFEITING	Numerical Value that denotes the count of Counterthefting that happened in India
ARSON	Numerical Value that denotes the count of Arson that happened in India
HURT/GRIEVOUS HURT	Numerical Value that denotes the count of Hurt that happened in India
DOWRY DEATHS	Numerical Value that denotes the count of Dowry deaths that happened in India

ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY	Numerical Value that denotes the count of Assault on women with intent to outrage her modesty that happened in India
INSULT TO MODESTY OF WOMEN	Numerical Value that denotes the count of Insult to the modesty of women that happened in India
CRUELTY BY HUSBAND OR HIS RELATIVES	Numerical Value that denotes the count of cruelty by husband or his relatives that happened in India
IMPORTATION OF GIRLS FROM FOREIGN COUNTRIES	Numerical Value that denotes the count of Importation of girls from foreign countries that happened in India
CAUSING DEATH BY NEGLIGENCE	Numerical Value denotes the count of causing death by negligence that happened in India.
OTHER IPC CRIMES	Numerical Value that denotes the count of other IPC crimes reported that happened in India.
TOTAL IPC CRIMES	Numerical Value that denotes the count of all the total crimes combined that happened in India.

TOOLS AND TECHNIQUES

Python Libraries:

The libraries that are used are given below with the description

LIBRARY	DESCRIPTION
Pandas	The Pandas is one of the most important libraries while dealing with data, all the data analysis, and data manipulation can be done with the help of pandas
Numpy	All the mathematical functions can be computed through the NumPy library.
Seaborn	Seaborn library is used for visualization.
Seasonal_decompose	This library was used to find the trend of crimes in every state.
Statsmodels.api	Statsmodels.api is used for modeling purpose, to use the forecasting machine learning algorithms, statsmodels.api provides all the necessary libraries in it.
Sklearn	The Sckitlearning(Sklearn) module contains all the machine learning algorithms eg Linear Regression Classification and clustering, in this project K-Means library was used to derive the cluster
Plotly	Plotly library provides an in-depth visualization of any data
Matplotlib	Matplotlib was used for visualization
Geopandas	Geopandas was used to draw the map of India. All the worst affected states in India were plotted on the map
ARIMA	ARIMA machine learning model was used for forecasting

TECHNIQUES

FORECASTING

Forecasting is a technique that can help one to predict the future outcome with the given historical data. Currently, the machine learning models that are used for forecasting are ARIMA, SARIMA, SARIMAX etc. Different regression models can give a better result as compared to forecasting models. The main assumption from the regression model is that the patterns in the past data will be repeated in the future. While performing forecasting techniques the obtained observation may be biased on the validation set concerning the original values. To minimize the biasing of the model generalization of the machine learning model must be done so that proper accuracy can be obtained from it. The generalization of the model helps developers to get a more accurate result which is more resilient to the noise.

CLUSTERING

One of the most frequently used supervised machine learning algorithms is Clustering. Clustering creates a homogenous group of entities for better management of it. To derive clusters, the clustering algorithm uses distance formula to create similarity of dissimilarity between entities. Types of distance metrics currently are EUCLIDEAN, MINKOWSKI, JACCARD SIMILARITY COEFFICIENT, COSINE, and GOWER's.

Assigning centroid is the most important step to obtain the equal distance between the entities. K-Means clustering can also be used pattern recognition, data mining, and cloud computing. The K-means is a centroid-based clustering algorithm. K-means calculation works in 2 stages, in the initial step, all information focuses are doled out to the cluster with the closest centroid. In the subsequent advance, all clusters recalculate and refresh the centroids area location on the mean of all information directs doled out toward their clusters.

IMPLEMENTATION

DATA EXPLORATION

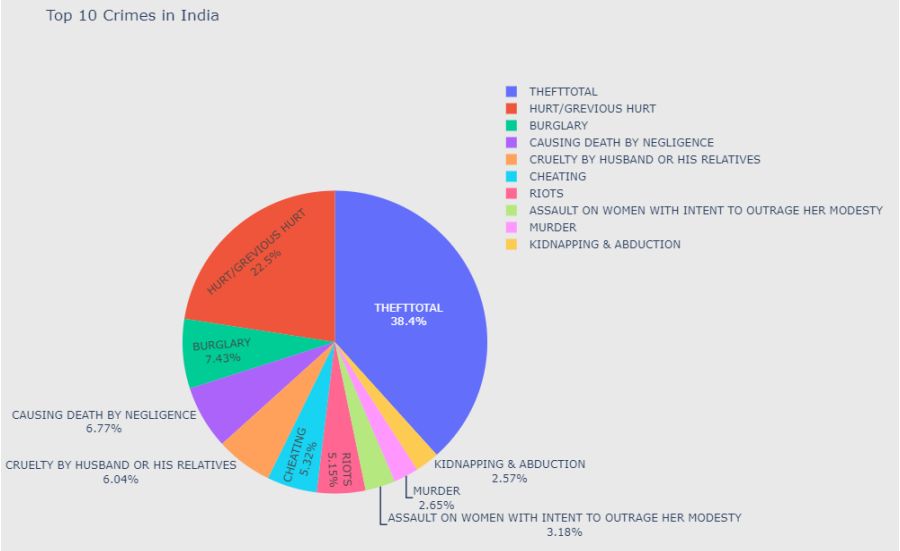
- Importing the data, from data.goi website, as a CSV file.
- Importing all the necessary libraries for the thesis
- Checking for all the information such as the number of columns and rows present in this dataset there are 33 columns present.
- Types of variables: 2 categorical and 30 numerical.
- Checking the count of all the crimes and states
- Checking for the NULL or NAN value in the dataset

DATA VISUALISATION

Data visualization is the demonstration of taking (data) and setting it into a visual setting, for example, a guide or chart. Data visualizations make of all shapes and size data simpler for the human mind to understand, and visualization additionally makes it simpler to detect patterns, trends, and outliers in gatherings of data. Data visualization techniques will help to find in-depth insights from the data such as the worst affected state in India, the count of crimes that are being committed in India by criminals, and which is the most frequent crime, etc.

A pie chart is a method of showing information where a circle is partitioned into sections (or cuts") that mirror the relative size or recurrence of the classifications. Now, we're trying to identify the top crimes that are being committed in India by criminals. This help us to classify the count or frequency of each crime so that it can be further analyzed which district or which state is worst affected concerning the top crimes in India.

TOP 10 CRIMES OF INDIA



Top 10 Crimes of India

Identifying crimes plays an important role in keeping the country safe. In India the top 10 crimes that were committed between 2001 and 2012 were:

MURDER, KIDNAPPING & ABDUCTION, BURGLARY, RIOTS, CHEATING, HURT/GRIEVOUS HURT, ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY, CAUSING DEATH BY NEGLIGENCE, THEFTTOTAL, CRUELTY BY HUSBAND OR HIS RELATIVES.

The author combined theft, auto theft, and other theft as the Theft total. It can be observed from the pie chart that the most committed crime in India was **THEFT** with 38.4% followed by **HURT** with 22.5%. The ratio of these two crimes was the most from 2001 to 2013. Crimes whose percentage was less than 2% were **MURDER, KIDNAPPING & ABDUCTION** with 2.65% and 2.57% respectively. **BURGLARY** with 7.43% and **CAUSING DEATH BY NEGLIGENCE** with 6.77%, and were third and fourth respectively. At fifth position recoded crimes was **CRUELTY BY HUSBAND OR HIS RELATIVES** with 6.04%. Crimes that were below 6% was **CHEATING, RIOTS**, and **ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY** with 5.32%, 5.15%, and 3.18% respectively.

MODELING

ARIMA Model

An ARIMA model is a class of statistical models for analyzing and forecasting time series data. It explicitly caters to a suite of standard structures in time-series data, and as such provides a simple yet powerful method for making skillful time-series forecasts. ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. It is a generalization of the simpler AutoRegressive Moving Average and adds the notion of integration. This acronym is descriptive, capturing the key aspects of the model itself. Briefly, they are:

- **AR**: AUTOREGRESSION. A model that utilizes the dependent relationship between perception and some number of lagged perceptions.
- **I**: INTEGRATED. The utilization of differencing of crude perceptions (for example deducting an observation from the previous perception at the past time step) to make the time series stationary.
- **MA**: MOVING AVERAGE. A model that utilizes the reliance between a perception and a remaining mistake from a moving normal model applied to slack perceptions.

Standard notation is used of ARIMA(p,d,q) where the parameters are substituted with integer values to quickly implies the specific ARIMA machine learning model being used. Each of these parts of the model are explicitly specified in the model as a parameter.

- AR is a component of p lags
- Integration component(d)
- MA with q lags, MA(q)
- The parameters of the ARIMA model are defined as follows:

Parameter Name	Default value	Parameter Description
P	1	The number of lag observations included in the model also called the lag order.
d	0	The number of times that the raw observations are differenced also called the degree of difference
q	0	The size of the moving average window also called the order of the moving average

Description Parameters used

A linear regression model is constructed including the specified number and type of terms, and the data is prepared by a degree of differencing to make it stationary, i.e. to remove trend and seasonal structures that negatively affect the regression model. A value of 0 can be used for a parameter, which indicates to not use that element of the model. This way, the ARIMA model can be configured to perform the function of an ARMA model, and even a simple AR, I, or MA mode. In this thesis, the author divided the dataset into two parts training and testing. The training data set was trained from the year 2001 to 2010 and was further validated from 2011 to 2012.

REFERENCES:

https://indianexpress.com/article/opinion/columns/crime-data-in-india-6035032/

https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

https://www.lawctopus.com/academike/crime-crime-rates/

https://data.gov.in/catalog/district-wise-crimes-under-various-sections-indian-penal-code-ipc-crimes?filters%5Bfield_catalog_reference%5D=87615&format=json&offset=0&limit=6&sort%5Bcreated%5D=desc

https://chartio.com/learn/charts/line-chart-complete-guide/

https://vciba.springeropen.com/articles/10.1186/s42492-021-00075-z