EasyChair Preprint

№ 4376

A Comparative Study of Crimes Against Women Based on Machine Learning using Big Data techniques

Shivani Mishra and Suraj Kumar

October 12, 2020

A comparative study of crimes against women based on Machine Learning using Big Data techniques

Shivani Mishra,Suraj Kumar
shivani1998mishra@gmail.com,
surajbtech762@gmail.com
School Of Computer Engineering, KIIT University
Bhubaneswar, ODISHA

ABSTRACT- Violence against women has become a prominent topic of discussion in India in recent years. Our Government and the Media have placed great focus in this issue due to continuously increasing crime rate. There many crimes committed in different states of India. This project mainly focuses on machine learning in pattern recognition for analysis of the patterns through Indian states for crime against women. Machine learning is a subset of artificial intelligence in the field of computer science that often uses statistical techniques to give computers the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed.This helps in the proper analysis of data related to the crime issues against women and the same could provide an aid to the government in effective policy making for preventive measures.

This project describes how to use machine learning to recognize crimes of different states and display them in a explicitly understandable format which could help in the generation of doable preventive measures. In this project, there would be conception of clustering methodology along with machine learning algorithms like: Cleaning the Dataset, Clustering. The Performance of each algorithm is analyzed and the best algorithm is found out which is having maximum accuracy of crime detection.The basic aim is to provide with an online application that could generate a quick and analyzed view of crime against women scenario in India.

Keywords- Machine Learning, Cleaning, Clustering.

I. INTRODUCTION

India publishes an up-to-date list of all reported crimes against women. A report in June 2017 by London's Reuters media company ranked India as the most dangerous place for women because of its high incidences of sexual violence, lack of access to justice in rape cases, child marriage, female foeticide and human trafficking. India outranked countries such as Syria and Afghanistan that are currently at war.

This has reignited the country's ongoing debate over women's safety. 548 experts for women's issues interviewed for the poll saying India topped the list because its government has done little to protect women. Rapes, female foeticide,sexual assault and harassment has gone unabated.The report noted that reported cases of crimes against women rose 83 percent between 2007 and 2016, where there were four cases of rape every hour.

Apart from sexual violence, India has the most child brides in the world and the government estimated earlier this year that there are 63 million missing women in the country because of girl child abortion, as

well as 21 million unwanted girls taken up by orphanages and NGOs. Although reported rapes in India are on the rise, its rate of rape per lakh people remains far lower compared to some Western countries, including the United States, which experts believe is due to years of fear and under-reporting.

Being able to parse the data presented by these huge data sets is a problem central to understanding the crimes in India. Researchers may be able to glean insight into the underlying causes of crimes, criminal mindset and also may be able to figure out indicators of future crimes to occur, by analyzing the data statistically. All of this categorization and analysis falls under the field of Data Science, a field which analyzes large sets of data using probability and statistics, and makes useful conclusions from the analysis. In this paper, we will attempt to parse India's up-to-date data-set, and try to perform some crime prediction. We will do this by utilizing techniques from Machine Learning.

II. METHODS

In this section we will discuss the methods used to obtain the results.

i. The Data Set

The data set is publicly available through the website https://data.gov.in/

The information presented in this data set is quite comprehensive, including information about the time and location of the crime, type of crime, etc. For the purpose of this paper, we will focus on the place,year and the type of the crime.

ii. Cleaning the Data set

The Data Set has been cleaned using simple Python Coding. Proper plotting of graph requires extra headers to be removed. Extra headers like 'All India','ALL INDIA','TOTAL Crime' section has been cleaned from the Set. Separate Cleaning codes have been written for individual states. After cleaning the Data Set a simple graph plot was obtained. The simple plot was referred for the further methods.

iii. Clustering of Data

Clustering is the assignment of a set of observations into subsets i.e **clusters** so that observation in the same cluster are similar in some sense. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields.

K-means clustering is an algorithm to classify or to group our objects based on attributes/features into K number of group. K is a positive number integer.

The grouping is done by minimizing the sum of squares of distances between the data and the corresponding cluster centroid. Thus the purpose of K-means clustering is to classify the data.

Simple processing of K-means Clustering

The basic step of K-means clustering is simple. In the beginning we determine

number of cluster K and we assume the centroid or center of these clusters. We can take any random objects as the initial centroids or the first K objects in sequence can also serve as the initial centroids.

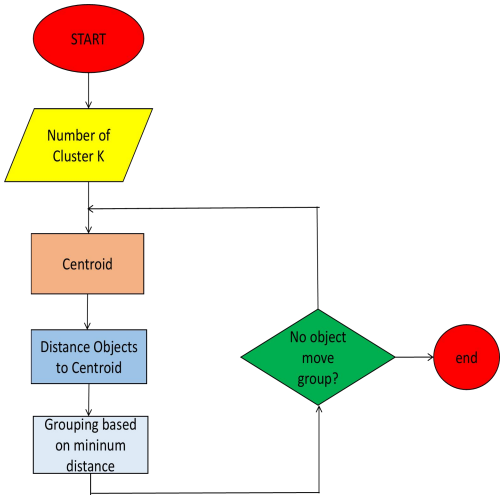

Figure 1: Data flow diagram for clustering algorithm

Step 1- Begin with the decision on the value of K (where, K is number of cluster)

Step 2- Put any initial partition that classifies the data into K clusters. You may assign the training samples randomly, or systematically as the following:

1) Take the first K training sample as single-element clusters.

2) Assign each of the remaining (N-K) training sample to the cluster with the nearest centroid. After each assignment, recomputed the centroid of the gaining cluster.

Step 3- Take each sample in sequence and compute its distance from the centroid of each of the clusters. If a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing the sample.

Step 4- Repeat step 3 until convergence is achieved, that is until a pass through the training sample causes no new assignments.

If the number of data is less than the number of cluster then we assign each data as the centroid of the cluster. Each centroid will have a cluster number. If the number of data is bigger than the number of cluster, for each data, we calculate the distance to all centroid and get the minimum distance. This data is said belong to the cluster that has minimum distance from this data. Iterate until stable (= no object move group):

1) Determine the centroid coordinate.

2) Determine the distance of each object to the centroids.

3) Group the object based on the minimum distance.

III. IMPLEMENTATION

I. Python

Python is an interpreted high-level programming language for general-purpose programming. Created by Guido Van Rossum and first released in 1991. Python is a programming language that lets you work quickly and integrate systems more effectively. Python is an interpreted, object-oriented, dynamic data

type of high-level programming languages.
Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.The programming language style is simple, clear and it also contains powerful different kinds of classes. Python can easily combine other programming languages, such as C or C++ or Java. Python is completely free as it is an open source software. Users can easy to install Python on their own computer and use the standard and extend library.

II. SciKit-learn

SciKit–learn is an open source machine learning library for the Python programming language. It features various classification, regression, and clustering algorithms and is designed to inter operate with the Python numerical libraries NumPy and SciPy . In this project using SciKit-learn we used clustering algorithms and Euclidean formula based on Python and figured out how to implement these algorithms in programming.

from sklearn.cluster import KMeans

III. NumPy and Matplotlib

In Python, there is no data type called array. In order to implement the data type of array with python, NumPy is the essential library for analyzing and calculating data. Its an open source library. NumPy is mainly used for the matrix calculation. In Python programming, it can be used with a simple command:
>>> import numpy

Then Python will call the methods from numpy.

Matplotlib is a famous library for plotting in Python. It provides a series of API and it is suitable for making interactive mapping. In this case, we need to use it to find the best result visually.
>>> import matplotlib.pyplot as plt
command is use to import and implement matplotlib.

IV. Pandas

Pandas is Python Data Analysis Library, pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools.

Primary object types:

- **DataFrame:** rows and columns (like a spreadsheet)
- **Series:** a single column

We use Pandas to load data into Python and perform your data analysis tasks. It is perfect for working with tabular data like data from a relational database or data from a spreadsheet.

import pandas as pd *#imports pandas library to python code*

from pandas.plotting import scatter_matrix *#Uses pandas for graph plotting*

dataset = pandas.read_csv(url, names=names) *#Imports the particular dataset file given by the url name.*

V. Machine learning system design

In general, the principles of machine learning system design should follow two basic requirements : the model selection and creation and the learning algorithm selection and design. In addition, different models can have different learning systems.The accuracy and complexity of different algorithms would be the most important factor of the learning system. If the chosen algorithm is not very adaptive to the learning system, then the efficiency and result of the learning system would be reduced. The selection of training data set can have an influence on learning performance and feature selection.

VI. Using Python to implement the program

For good implementation and good compatibility, Python version 3.6.5 will be in use. The Integrated Development Environment in this case will be IDLE. By using the Scikit-learn software package, there is no need to write a program to implement each algorithm.

We have to import the following modules:

from sklearn.cluster import KMeans

import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

Snippet of code for the computation of total crime in India year wise.

```
from  sklearn.cluster import KMeans
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
df = pd.read_csv('crimedata.csv')
print(df)
s=df[df["CRIME HEAD"]=="TOTAL CRIMES AGAINST WOMEN"]
l=s[s["STATE/UT"]=="TOTAL Crime against Women"]
print(l)
l=l.drop(["STATE/UT"],axis=1)
l=l.drop(["CRIME HEAD"],axis=1)
print(l)
m=[]
for i in l:
  m.append(i)
print(l[m[0]])
z=[]
for i in range(0,len(m)):
  plt.scatter(m[i],l[m[i]],color="red")
plt.xlabel("Year")
plt.ylabel("Number of crimes against women")
plt.title("Total crimes year wise")
plt.show()
l=l.reset_index()
```

```
print(l)
kmeans=KMeans(n_clusters=3)
q=[]
for i in m:
    print(i)
    q.append(list(l[i].values))
print(q)
x=[]
for i in q:
    for j in i:
        x.append(j)
print(x)
x=np.array(x)
x=x.reshape(-1,1)
kmeans.fit(x)
print(kmeans.labels_)
colors=["red","green","yellow"]
for  i in range(0,len(x)):
    plt.scatter(m[i],l[m[i]],color=colors[kmeans.labels_[
plt.xlabel("Year")
plt.ylabel("Number of crimes against women")
plt.title("Total crimes year wise")
plt.show()
```

Snippet of code for the computation of total crime in India state wise.

```
from  sklearn.cluster import KMeans
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
df = pd.read_csv('crimedata.csv')
print(df)
s=df[df["CRIME HEAD"]=="TOTAL CRIMES AGAINS
m=s[s["STATE/UT"]!="TOTAL Crime against Wom
print(m)
m=m.reset_index()
df=m
df['TOTALCRIME']=m['2001']+m['2002']+m['200
+m['2007']+m['2008']+m['2009']+m['2010']+m[
print(df)
s=df[["TOTALCRIME","STATE/UT"]]
print(s)
m=df["TOTALCRIME"].values
print(m)
plt.scatter(df["STATE/UT"],df["TOTALCRIME"]
plt.xticks(rotation="vertical")
plt.xlabel("States")
plt.ylabel("Number of crimes statewise")
plt.title("Total crimes state wise")
plt.show()
```

```
kmeans=KMeans(n_clusters=3)
m=np.array(m)
m=m.reshape(-1,1)
kmeans.fit(m)
colors=["red","green","brown"]
for i in range(0,len(df)):
    plt.bar(df["STATE/UT"][i],df["TOTALC
plt.xticks(rotation="vertical")
plt.xlabel("States")
plt.ylabel("Number of crimes statewise")
plt.title("Total crimes state wise")
plt.show()
```

IV. EVALUATING RESULT

For every machine learning algorithm, exceptions will always exist. In order to find the best result, result analyzing is necessary. The analysis of crimes against women in India yielded some alarming results from every domain. From Figure 3 it can be observed that the total rate of Crimes in India has brought us to high risk zone.The results obtained from K-means clustering imply high vulnerability of crimes in states of Andhra pradesh, Uttar Pradesh, Madhya Pradesh, Rajasthan and Maharashtra. In Figure 4, we see the women in state of Madhya Pradesh are more vulnerable to rapes. The major reasons for such crimes are lack of public safety, not enough police general, blaming provocative clothing, acceptance of domestic violence, Low status of women and the sluggish court system of India.

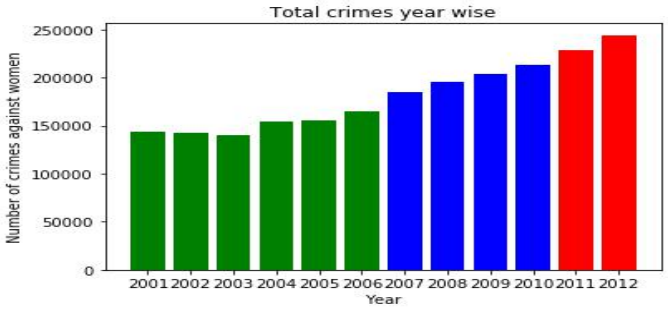

Figure 2: Cluster on Total Crimes year wise

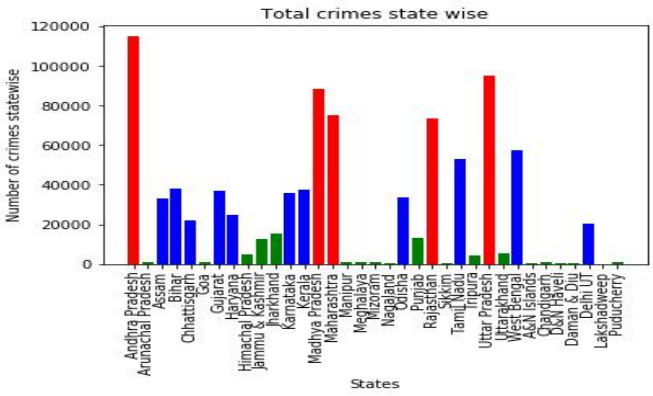

Figure 3: Cluster on Total Crimes state wise

Considering the above graph we see that the state of Andhra Pradesh has highest rate of crimes and the Union Territory, Lakshadweep has the lowest rate of Total Crimes.

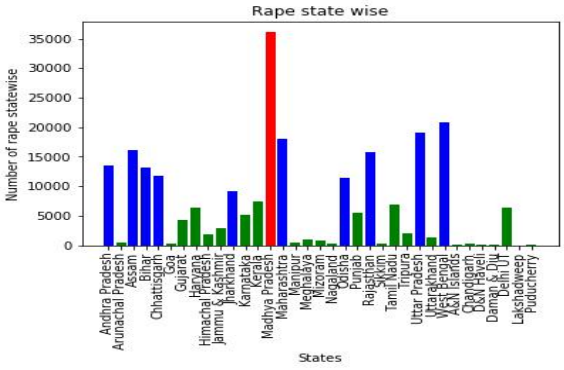

Figure 4: Cluster on number of Rapes state wise

In the above graph, the red bar shows highest number of Rapes in the state of Madhya Pradesh and Lakshadweep has the lowest rate of Rapes.

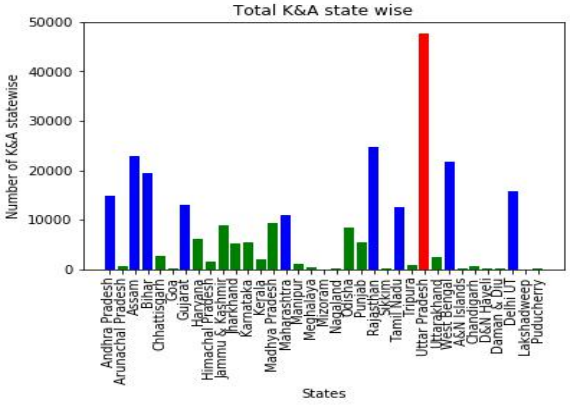

Figure 5: Cluster on total kidnapping and abduction state wise

In the above graph, the red bar shows highest number of kidnapping and abduction in the state of Madhya Pradesh and Lakshadweep has the lowest rate of Kidnapping and abduction.

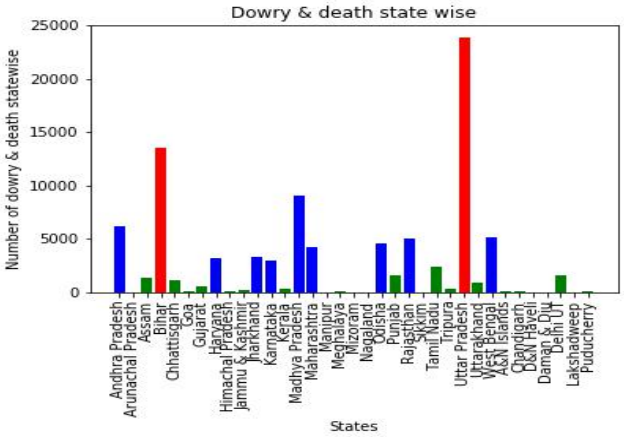

Figure 6: Cluster on number of Dowry deaths state wise

In the above graph, the red bar shows highest number of deaths due to Dowry issues in the state of Uttar Pradesh and Lakshadweep has the lowest rate of Dowry deaths.

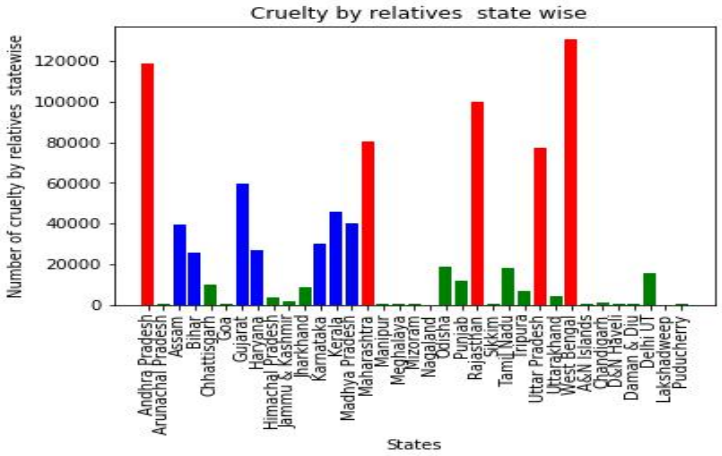

Figure 7: Cluster on number of cruelty by relatives state wise

In the above graph, the red bar shows highest number of Rapes in the state of West Bengal and the blue bar shows Lakshadweep having the lowest rate of deaths due to cruelty by relatives.

V. CONCLUSION

With the rapid development of technology, Machine learning is the most fundamental approach to achieve AI. This project describes the work principle of machine learning using python. In addition, a case study of Women Empowerment to introduce the work flow of machine learning in data analysis is shown. The work also shows how to use SciKit-learn software to learn machine learning.

Many crimes are not being reported and no one has the record about it. To reduce such crimes awareness campaigns should be done. Although punishments have been provided by the law, 3 crore cases have been pending.

VI. FUTURE PROSPECTS

The Crime rate case study shows that the Machine Learning algorithm works well in the data analysis of different crimes state-wise and year-wise. Crimes against women is a problem increasing rapidly and women are becoming more and more prone to rapes, kidnapping, assault and molestation. This is a problem which can be prevented if actions are taken at the right time. The law and order for states having highest rate of crimes should be upgraded. Social Awareness steps should be taken by our government for reducing these crimes, educating people to respect the dignity of women. States having higher Rate of Crimes. Machine Learning used in this case can be applied in many other algorithms for seeking information regarding crimes in different areas. Women can seek help from online protocol systems and can be more aware of the crime conditions in their area.

VII. ACKNOWLEDGEMENT

We would like to express our profound gratitude to the Dean of School Of Computer Engineering KIIT, Dr. Samaresh Mishra for allowing us to proceed with the report and for giving us full freedom to access the lab facilities. Our heartfelt thanks to Dr. Siddharth Swarup Rautaray & Dr. Manjusha Pandey for taking time and helping us through our work. They have been a constant source of encouragement without which the work might not have been completed on time. Their ideas and thoughts have been of great importance.

VIII. REFERENCES

1. Veena Talwar Oldenburg. “Dowry Murder: The Imperial Origins of a Cultural Crime”

2. Ignatius, Arun. "Sexual Violence In India." *The International Journal Of Indian* (2013).

3. Krishnan, Kavita. "Rape Culture and Sexism in Globalising India." *SUR-Int'l J. on Hum Rts.* 22 (2015): 255.

4. Simon-Kumar, Rachel. "Sexual violence in India: The discourses of rape and the discourses of justice." *Indian Journal of Gender Studies* 21.3 (2014): 451-460.

5. East, Patricia, and Joyce Adams. "Sexual assertiveness and adolescents'

sexual rights." *Perspectives on Sexual and Reproductive Health* 34 (2002): 212-213.

6. Singh, Mannat Mohanjeet, Shradha S. Parsekar, and Sreekumaran N. Nair. "An epidemiological overview of child sexual abuse." *Journal of family medicine and primary care*3.4 (2014): 430.

7. Behere, P. B., TS Sathyanarayana Rao, and Akshata N. Mulmule. "Sexual abuse in women with special reference to children: Barriers, boundaries and beyond." *Indian journal of psychiatry* 55.4 (2013): 316.

8. Belur, Jyoti, et al. "The social construction of 'dowry deaths'." *Social Science & Medicine* 119 (2014): 1-9.

9. Bhate-Deosthali, Padma, and Nobhojit Roy. "The invisible face of burns in India." *Current Medicine Research and Practice*5.2 (2015): 53-54.

10. Bhate-Deosthali, Padma, and Lakshmi Lingam. "Gendered pattern of burn injuries in India: a neglected health issue." *Reproductive health matters* 24.47 (2016): 96-103.