SYMBIOSIS INSTITUTE OF
COMPUTER STUDIES AND RESEARCH
Symbiosis International (Deemed University)
(Established under section 3 of the UGC Act, 1956)
Re-accredited by NAAC with 'A' grade (3.58/4) | Awarded Category – I by UGC
Founder: Prof. Dr. S. B. Mujumdar, M. Sc., Ph. D. (Awarded Padma Bhushan and Padma Shri by President of India)

Celebrating 50 Years of Excellenc

DATA WAREHOUSING ARCHITECTURE AND OPERATIONS

MINI PROJECT

Topic: **Crime Analysis in Maharashtra**

Rhea Shree S

PRN-23030141050
MBA-IT (SEMESTER 2)

TABLE OF CONTENTS

Date: 27/02/2024

COMPLETION CERTIFICATE

This is to certify that the project report titled Symbolism in Literature is the bonafide work of Rhea Shree S, PRN-23030141050, MBA-IT Batch 2023-2025 at Symbiosis Institute of Computer Studies and Research (SICSR) who carried out the project work under Dr. Amol Vibhute's supervision. She has completed the project during Start Date: 23/02/2024 to End Date: 27/03/2024

ABSTRACT

Crime analysis helps in understanding patterns or trends of crime in a specific region and to find effective strategies to minimize it. This study focuses on crime analysis specifically in the state of Maharashtra. The dataset is collected from Kaggle website and contains information about 11 categories of crime and all 35 districts in the state of Maharashtra. Data analysis and Data mining techniques such as Linear Regression and Expectation-Maximization (EM) clustering algorithm are applied using the WEKA software. Data Mining provides practical and convenient methods for evaluating extensive and diverse datasets. It enables the discovery of concealed insights within vast criminal records, aiding in the investigation, management, and prevention of crime by organizations and individuals.

INTRODUCTION

In today's world, Crimes are rising in increasing levels due to many factors mainly due to inequality in the socioeconomic area and lack of strict conviction and accountability.

In India, criminal identification mostly relies on traditional methods for crime analysis and determining suspects which makes it challenging for police stations using databases to store and access criminal information.

Crime patterns constantly change and the amount of stored crime data keeps growing, making data management and analysis difficult. Effective crime prevention relies on thorough analysis of this data. Data mining techniques and machine learning algorithms can help extract trends from this data to aid in crime prevention.

To address these challenges, Utilizing WEKA, an open source data mining software tool, aims to improve upon traditional methods by storing criminal data in a specific format, analyzing it, and deriving conclusions. This systematic analysis helps law enforcement agencies identify criminals and analyze crime trends more effectively. Data mining is a technique for extracting hidden data from a generally huge dataset and turning it into useful information for further use

LITERATURE REVIEW

Researchers often employ clustering techniques to classify the data for further analysis and prediction of criminal behavior. They focus on methods for predicting crime by uncovering hidden patterns within existing crime records. Through the use of data mining techniques in WEKA, to study the likelihood of criminal activity in a region, Classification algorithms are considered: SMO, Zero R, and J48 decision trees. The researchers collected over 10,000 crime records from the Indian police department to predict the frequency and behavior of crime. Among the algorithms tested, the Naive algorithm demonstrates reliable predictions in terms of crime frequency.[1]

Ensuring the security and safety of communities is a paramount concern for governments worldwide, prompting extensive research into strategies for reducing crime rates. In recent years, the application of various artificial intelligence (AI) strategies in crime prediction has emerged as a significant area of study. Notably, the supervised learning approach emerges as the most commonly applied method in crime prediction. [2]

The thesis under consideration talks about leveraging predictive analytics to present the predicted output to users in a comprehensible format, possibly through the use of data visualization algorithms like K-means clustering. This approach aims to simplify the interpretation of crime prediction data, enabling stakeholders to make informed decisions and strategic interventions.[3]

By training the linear regression model on historical crime data, the study aims to predict various types of crimes, provide valuable insights for law enforcement agencies by enabling them to anticipate, prevent, and address future criminal activities more effectively. The experimental results demonstrate the effectiveness of the linear regression model in forecasting various types of crimes, with a notable observation that crime rates tend to increase with population growth.[4]

The prime concern of the paper is crime and its prevention by relying on software tools such as Fuzzy System and WEKA for data collection and analysis as well prediction of future crime locations. Fuzzy System is highlighted as one of the most commonly used tools for data collection and filtering in crime analysis. In fuzzy clustering, data elements have the flexibility to be assigned to multiple clusters, unlike traditional clustering methods where each element belongs exclusively to one cluster. The fuzzy system comprises input, processing, and output stages, with fuzzy rules guiding the decision-making process.[5]

The increasing volume of records highlights the importance of a systematic and intelligent approach to crime analysis. The crime analysis process involves the use of patterns, graphs and statistics to identify correlated crimes or future crimes. In this documentation, different types are categorized and various statistical and technical methodologies are described that help for efficient crime investigation and analysis process which helps in minimizing time and error complexity.

METHODOLOGY

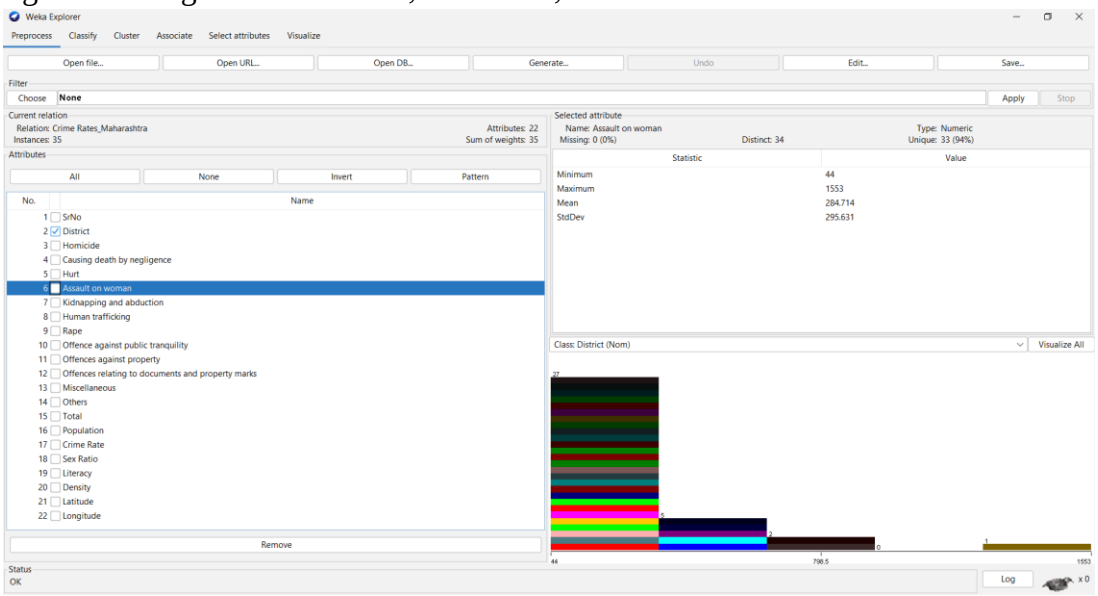
Data Collection:

The dataset related to various types of crime in the state of Maharashtra is collected from Kaggle website. The dataset includes 35 districts and 11 types of crimes namely, Homicide, causing death by negligence, Hurt, Assault on woman, Kidnapping and abduction, Human trafficking, Rape, Offence against public tranquility, Offences against property, Offences relating to documents and property marks, Miscellaneous and Others. The dataset also provides information about the Population, Crime Rate, Sex Ratio, Literacy, Density, Latitude, Longitude of each district.

Analysis using WEKA

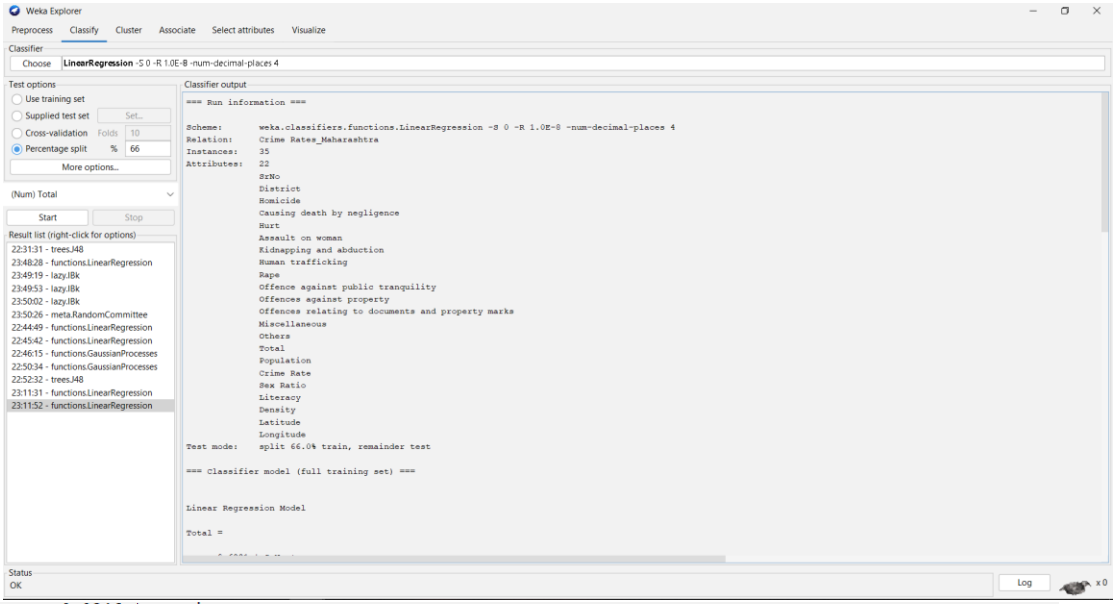
(i) Loading the dataset in Weka

The dataset was in CSV format and is loaded into WEKA for preprocessing.
The Preprocess tab in Weka talks about various attributes in the file as well give figures relating to the minimum, maximum, mean and Standard Deviation of the data.



Preprocess tab in WEKA

(ii) Applying Linear Regression model




```
      0.0346 * Density +
    -30.5521 * Latitude +
     46.9061 * Longitude +
  -5741.4512

Time taken to build model: 0.02 seconds

=== Predictions on test split ===

    inst#     actual  predicted      error
        1   1491       6781.622   5290.622
        2   4862       8698.21    3836.21
        3   5690       9741.912   4051.912
        4   7524      10036.433   2512.433
        5   3622       7199.537   3577.537
        6   6603      10690.484   4087.484
        7   5719      10203.597   4484.597
        8   5146       8619.831   3473.831
        9   6255      10168.574   3913.574
       10   7134       9624.885   2490.885
       11   2190       6549.433   4359.433
       12   2160       5676.785   3516.785

=== Evaluation on test split ===

Time taken to test model on test split: 0.02 seconds

=== Summary ===

Correlation coefficient                  0.9288
Mean absolute error                   3799.6086
Root mean squared error               3873.1545
Relative absolute error                 39.067  %
Root relative squared error             39.039  %
Total Number of Instances               12
```

Screenshots of Linear Regression applied on dataset in WEKA

The model utilized the Linear Regression algorithm with specific parameters.

The dataset contains 35 instances or data points. There are 22 attributes used for prediction, including features like SNo, District, Homicide, Population, Crime Rate, etc. The dataset was split into 66% for training and the remaining data for testing. The coefficients represent the weights assigned to each attribute for predicting the target variable (crime rate).

For example, positive coefficients suggest a positive relationship with the target variable, while negative coefficients suggest a negative relationship.

Predictions on Test Split:

The "Predictions on test split" section presents the actual crime rates alongside the predicted values and the error for each instance in the test dataset.

The "error" column represents the absolute difference between the actual and predicted crime rates for each instance.

Evaluation on Test Split:

Correlation Coefficient: Measures the strength and direction of the linear relationship between actual and predicted crime rates. A value of 0.9288 suggests a strong positive correlation.

Mean Absolute Error (MAE): The average absolute difference between the actual and predicted crime rates is 3799.6086. It represents the average magnitude of errors.

Root Mean Squared Error (RMSE): The square root of the average squared differences between the actual and predicted crime rates is 3873.1545.

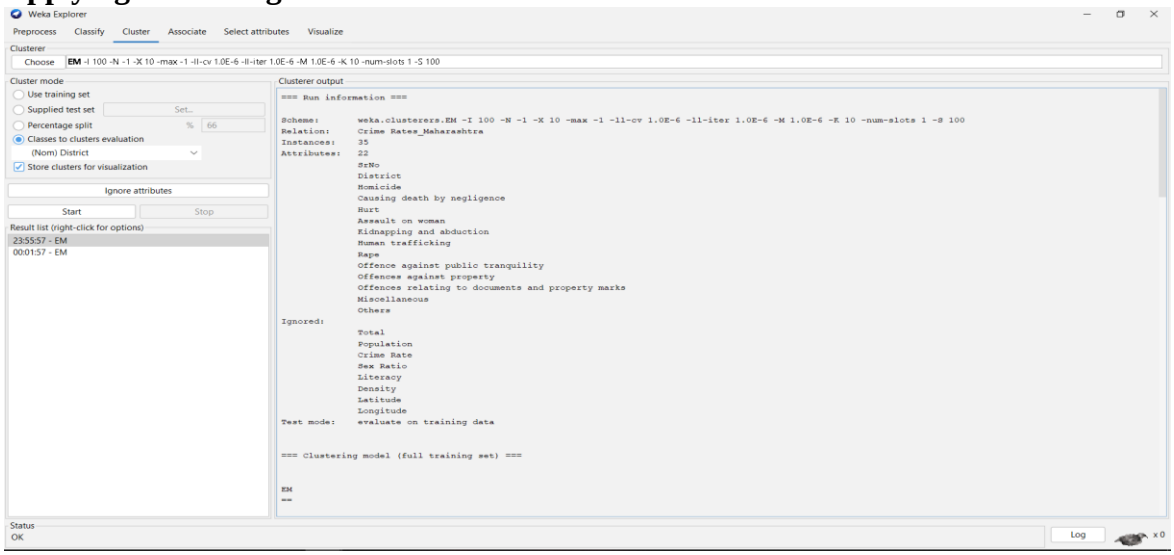
Relative Absolute Error: The MAE relative to the mean of the actual crime rates is 39.067%.

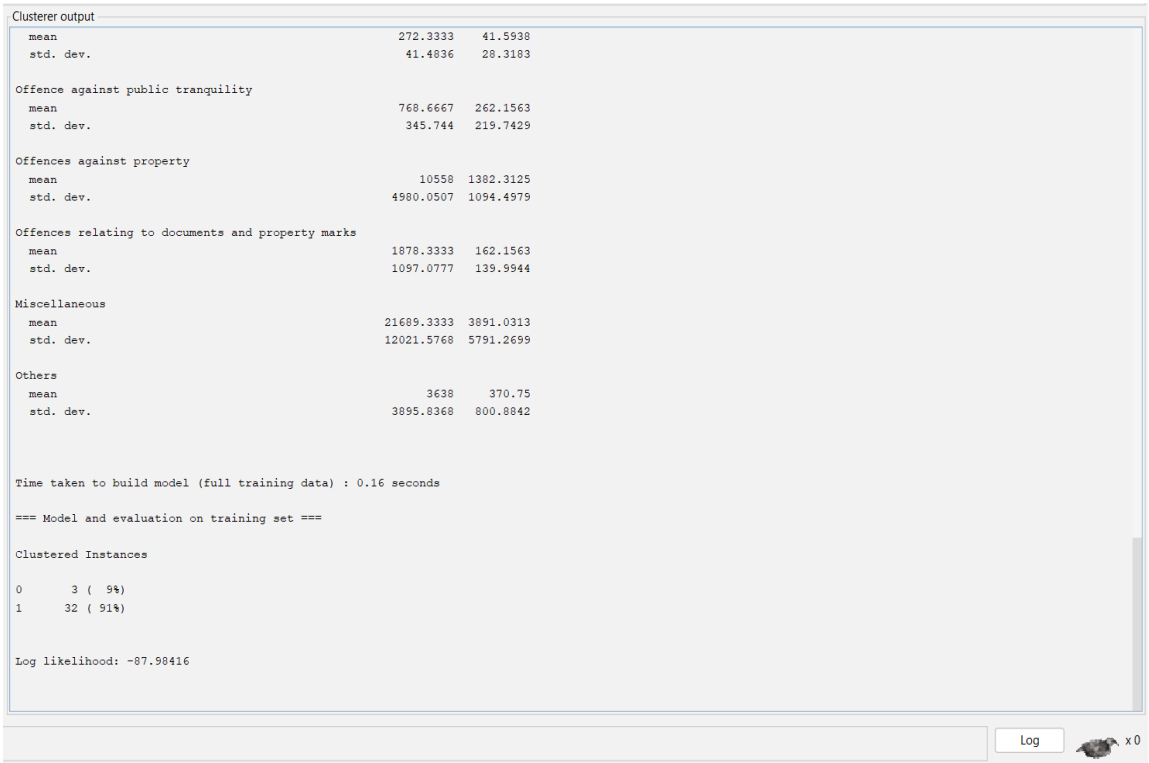
Root Relative Squared Error: The RMSE relative to the standard deviation of the actual crime rates is 39.039%.

Total Number of Instances: The number of instances used for evaluation.

Overall, the model seems to perform well in predicting crime rates in Maharashtra based on the provided attributes, as evidenced by the high correlation coefficient and relatively low error metrics. However, further analysis and validation may be necessary to assess model's strength and applicability across different scenarios.

(iii) Applying Clustering methods





Screenshots of Clustering Algorithm applied on dataset in WEKA

The clustering algorithm used is Expectation-Maximization (EM), which is the most commonly used approach to find the maximum likelihood estimates of variables that are sometime observed and sometimes not [6]. In this model, it includes the number of iterations, maximum number of clusters, convergence criteria, and random seed.

The dataset contains 35 instances or data points and 22 attributes used for clustering. The EM algorithm determined that the optimal number of clusters is 2 and performed two iterations during the clustering process.

Each numerical attribute's mean and standard deviation values are provided for both clusters (Cluster 0 and Cluster 1). For categorical attributes like District, the frequencies of each category within each cluster are provided.

Cluster 0 contains 3 instances (9%) It has specific characteristics across various attributes, including lower mean values for attributes like Homicide, Causing death by negligence, Hurt, Assault on woman, etc. Whereas Cluster 1 contains 32 instances (91%), the attributes here have a higher mean value

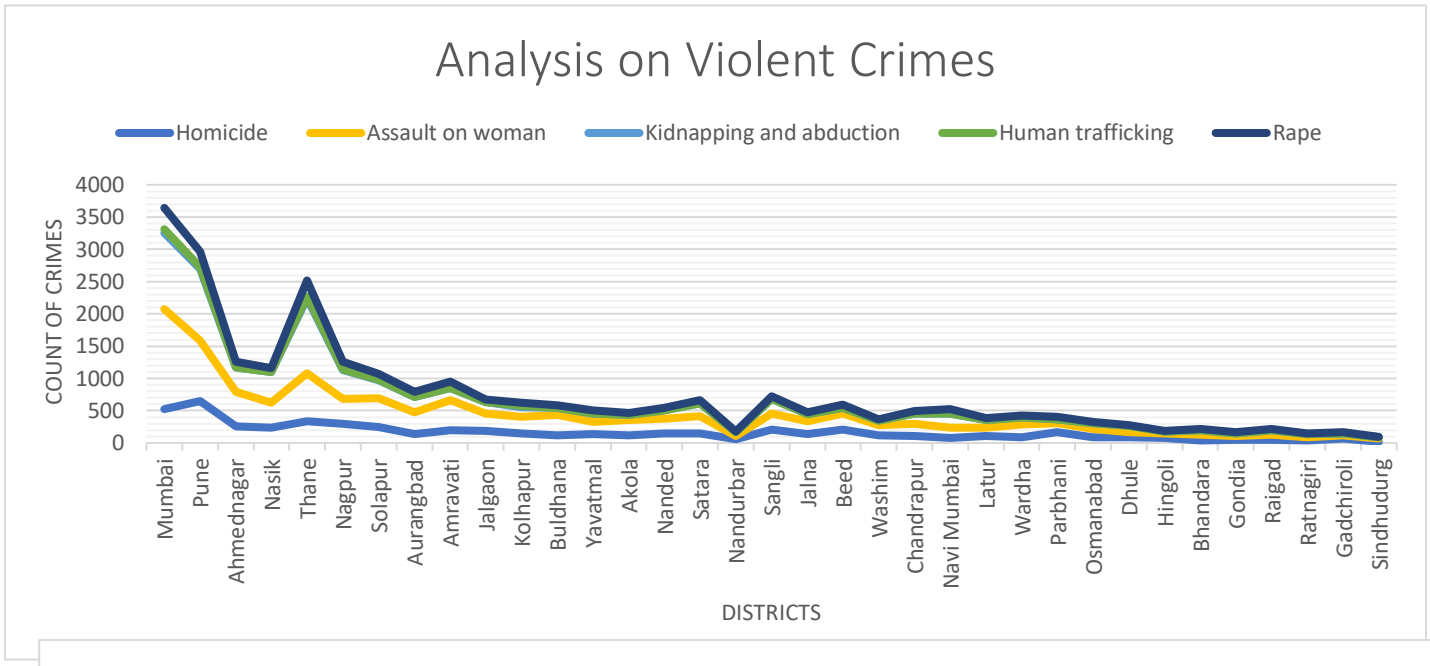
The log likelihood value (-87.98416) indicates the overall goodness of fit of the model to the data. Higher values represent better fit.

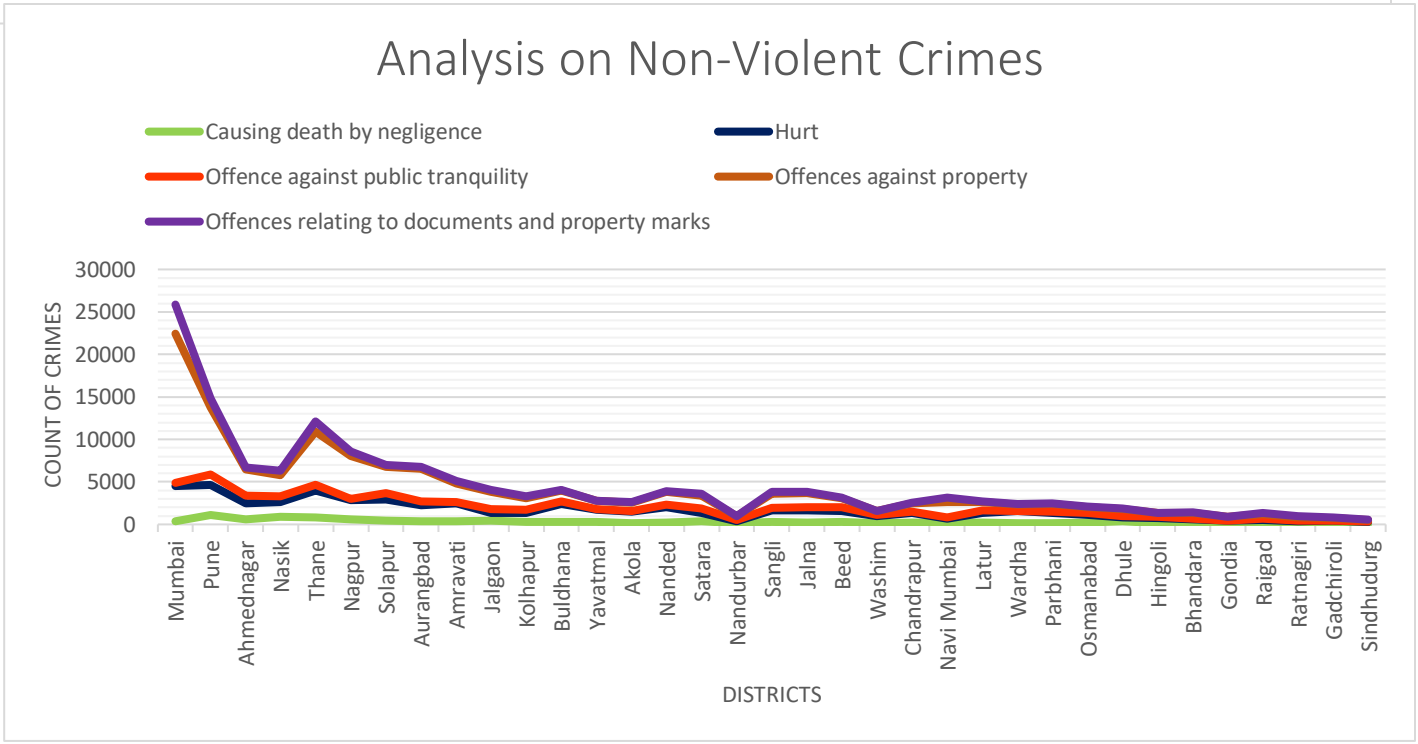
The clustering results indicate that the instances in the dataset can be effectively grouped into two distinct clusters based on the specified attributes and observed patterns. Cluster 1 is significantly larger than Cluster 0, indicating that the majority of instances in the dataset share similar characteristics across various crime rate attributes.

The attributes' means and standard deviations within each cluster provide insights into the characteristics of instances within those clusters. Further analysis and interpretation of these clusters would help to understand the underlying patterns or factors contributing to the variation in crime rates across districts in Maharashtra.

RESULTS

The types of crimes are categorized into Violent and Non-Violent crimes. The violent crimes are: Homicide, Assault on Women, Kidnapping and abduction, Human Trafficking and Rape. The nonviolent crimes are: Offence against public tranquility, Offences against property, Offences relating to documents and property marks, causing death by negligence, and hurt.
From both graphs, We can see that the maximum number of crimes is in Mumbai and lowest count of crimes is in Sindhudurg.





DISCUSSION

These results are important especially to Law and Police authorities to gain insights into the underlying patterns and factors influencing crime rates across districts in Maharashtra.

The clustering results highlight the presence of distinct groups of districts with varying crime rate attributes. This segmentation empowers authorities to prioritize resource allocation and intervention strategies according to the specific characteristics and needs of each cluster or group. The strong positive correlation between actual and predicted crime rates suggests that the model is effective in capturing the relationships between different attributes and predicting crime rates accurately.

Overall, these results offer valuable insights that can inform evidence-based decision-making processes aimed at enhancing public safety, reducing crime, and improving the overall quality of life for residents in Maharashtra.

CONCLUSION

Law enforcement officials and investigators gather data from diverse sources such as telephone records, social networks, police records, and transaction records to aid in the investigative process. Despite being time-intensive, advancements in technology offer opportunities for law enforcement to more effectively identify criminals.

There can be various data mining tools and techniques that can be applied that aims at identifying specific crime patterns occurring in particular location. The findings suggest that crime data mining has the potential to enhance national security and intelligence operations by increasing productivity and efficiency. Moreover, there remain numerous avenues for future research in this field that are still in the early stages of exploration.

Date: 27/02/2024

ACKNOWLEDGEMENT

This project was made possible through ongoing efforts with the determination and guidance of a wide range of faculty and staff of the Institute. I would like to extend my gratitude to Dr. Amol Vibhute, Project guide in-charge, for your help and training and for helping me complete my project successfully. Their constant guidance, valuable supervision, and selfless-assistance help have been proved helpful during this period and completing this project with excellent results.

Name -Rhea Shree S

BBA(IT) – 2023-2025 batch 2nd Semester

PRN - 23030141050

REFERENCES

1. *Kshatri, S. S., & Narain, B. (2020). Analytical study of some selected classification algorithms and crime prediction. Int J Eng Adv Technol, 9(6), 241-247.*

2. *Dakalbab, F., Talib, M. A., Waraga, O. A., Nassif, A. B., Abbas, S., & Nasir, Q. (2022). Artificial intelligence & crime prediction: A systematic literature review. Social Sciences & Humanities Open, 6(1), 100342.*

3. *Pande, V., Samant, V., & Nair, S. (2016). Crime detection using data mining. International Journal of Engineering Research & Technology (IJERT), 5(01), 2.*

4. *Awal, M. A., Rabbi, J., Hossain, S. I., & Hashem, M. M. A. (2016, May). Using linear regression to forecast future trends in crime of Bangladesh. In 2016 5th international conference on informatics, electronics and vision (ICIEV) (pp. 333-338). IEEE.*

5. *Kaur, N. (2016). Data Mining Techniques used in Crime Analysis:-A Review. International Research Journal of Engineering and Technology (IRJET), 3(08), 1981-1984.*

6. *https://www.javatpoint.com/em-algorithm-in-machine-learning*