International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 04 | April - 2023 Impact Factor: 8.176 ISSN: 2582-3930

CRIME DATA ANALYSIS OF TAMIL NADU USING MACHINE LEARNING TECHNIQUES

J.R.KISHOR[1], S.ARUNACHALAM[2] ,J.AATHIKESAVAN[3] , ARUL SELVAN G[4] , Dr. GANESAN T[5]
1, 2, 3 Final Year Student, 4 Assistant Professor, 5 Professor
Department of Computer Science and Engineering
E.G.S. Pillay Engineering College (Autonomous), Nagapattinam

Abstract: ***The rate of crime in Tamil Nadu is alarmingly rising. Tamil Nadu had the fifth-highest crime rate among Indian states in 2021. The purpose of this research is to analyse the crime data of Tamil Nadu using machine learning techniques. The result of the analysis is categorised by geographic area in order to alert individuals to the types of crimes that are likely to occur in their areas and when, in simpler words, it examines the features of crime that occurs in every region of the state. The most common crimes in a given location and the time period of a day when such crimes are most likely to occur are included in the analysed output data, which proves effective in decreasing crime activities in that location.***

Keywords: Crime Analysis, Machine Learning, K-NN, K-Means Clustering, Regression, Naïve Bayes Classifier

1. INTRODUCTION

Crime poses a serious threat to humanity. From small towns and villages to large metropolitan cities, crimes occur. Many types of crimes exist, including robbery, murder, rape, assault, false imprisonment, kidnapping, and homicide. These crimes often make people, question the existence of justice.

There is a need to resolve cases, much more quickly because crime rate is rising. The police department must manage and lessen the crime activities, which have increased at a faster rate. Given the vast amount of crime data available, crime prediction and its avoidance, are the two biggest issues of the police department.

According to the Tamil Nadu Police Department's Compendium 2021, the total number of IPC and SLL cases filed in Tamil Nadu was 756,753 (only in 2021). [14] The objective of this research is to enlighten people, with the crime that occur in any locality, and the time when the location is most vulnerable to the crime.

Crime analysis also assists the police department in reducing crime by allowing them to tighten security, during the peak time of crime. We can study the kind of crimes that occur more frequently in a specific location and during which time periods, using machine learning algorithms with Python as the core programming language.

2. LITERATURE REVIEW

The challenges surrounding crime prevention have been addressed by numerous researchers, who have also suggested various crime analysis systems. The researchers of [1], combines the machine learning techniques with computer vision (a branch of artificial intelligence), to analyse characteristics of a crime. Despite greater accuracy and comprehensive research, there are challenges in implementation of this concept, due to lack of advanced technologies to handle enormous data, in less time period.

The regression technique is used in [2] to analyse various crime data pertaining to Tamil Nadu. Their paper's objective is to forecast future crime statistics for each crime category they dealt with. With two separate dataset techniques, Vancouver crime data from the previous 15 years were used in [3].Crime analysis accuracy was from 39% to 44% using machine learning predictive models - KNN and decision tree. Although this model's prediction accuracy is poor, it offers a basic framework for additional research.

In [4], the researcher reviews the effectiveness of data mining techniques that might be utilised to analyse the data collected regarding past convictions. The scholars of [5] employed

machine learning and data science techniques to forecast crimes using a set of crime data pertaining Chicago. In their analysis of K-NN algorithm, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Bayesian methods, revealed that K-NN algorithm has the greatest accuracy of all these algorithms.

3.APPLICATION OF MACHINE LEARNING IN CRIME ANALYSIS

The field of criminology is dedicated to the scientific investigation of criminal activity and its behaviour. This is the important sector, where implementing machine learning techniques can produce remarkable outcomes. Crime analysis is a branch of criminology that looks into and learns about crime and how it relates to criminals.

The goal of law enforcement is to pinpoint the traits of crime. The very first stage of developing detailed analysis is investigating crime characteristics. Criminology is a suitable sector for the application of machine learning techniques due to the evaluation of large volume of crime data and the intricacy of the relationships between them.

Machine learning utilizes algorithms, which enables it to learn from data and generate precise predictions. The Police Department and other judicial institutions have the chance to learn about crime trends, how, and why crimes are committed through data analysis. Both Machine Learning and Data Analysis contribute to enhance criminal analysis and, also assist in crime prevention.

4.MACHINE LEARNING TECHNIQUES USED IN THE PROPOSED SYSTEM

4.1 REGRESSION - Regression analysis uses one or more independent variables to describe the relationship between a dependent (target) and independent (predictor) variables. More specifically, regression analysis enables us to comprehend how, while other independent variables remain unaltered, the value of the dependent variable changes in relation to an independent variable. On the basis of the data set gathered for the project, the algorithm employs linear regression techniques. [11]

With the aid of statistical techniques, the linear regression technique aids in forecasting the future behaviour of events. To forecast future behaviour, the algorithm calculates the mean and variance of the dependent variables and applies the formula Y=b0+b1*x.

4.2 K-NEAREST NEIGHBOUR ALGORITHM - One of the simplest machine learning algorithms, based on the supervised learning method, is K-Nearest Neighbour.

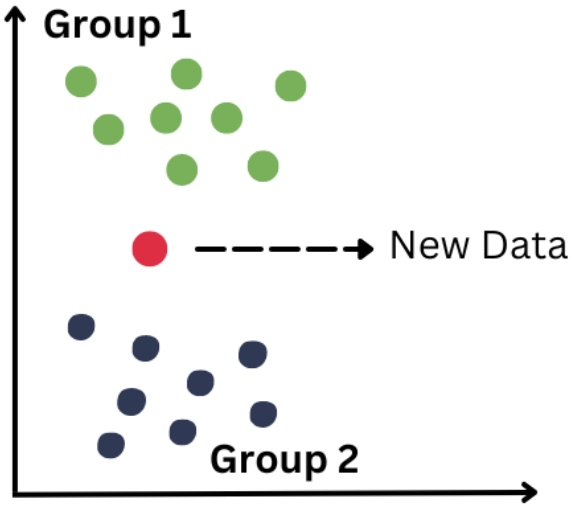

Figure 4.2.1– Before using KNN Algorithm

The K-NN algorithm makes the assumption that the new case and the existing cases are comparable, and it places the new case in the category that is most like the existing categories.[9] A new data point is classified using the K-NN algorithm based on similarity after all the existing data has been stored. This means that utilising the K-NN method, fresh data can be quickly and accurately sorted into a suitable category.

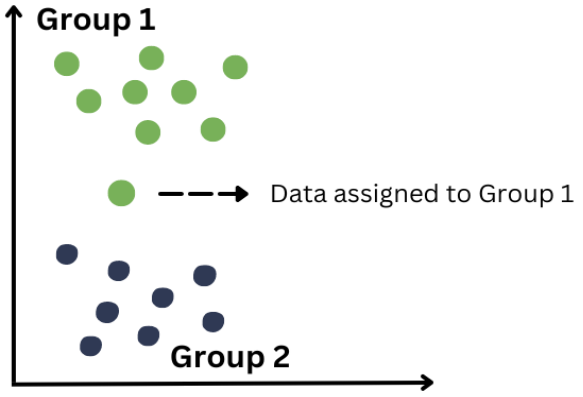

Figure 4.2.2 – After using KNN Algorithm

Although the K-NN approach is most frequently employed for classification problems, it can also be utilised for regression. Since K-NN is a non-parametric technique, it makes no assumptions about the underlying data.

4.3 K-MEANS CLUSTERING - K-Means Clustering divides the unlabelled dataset into various clusters. This enables us to divide the data into various groups and is a practical approach to quickly recognize the various groupings in the unlabelled dataset without the need for any training. [10]

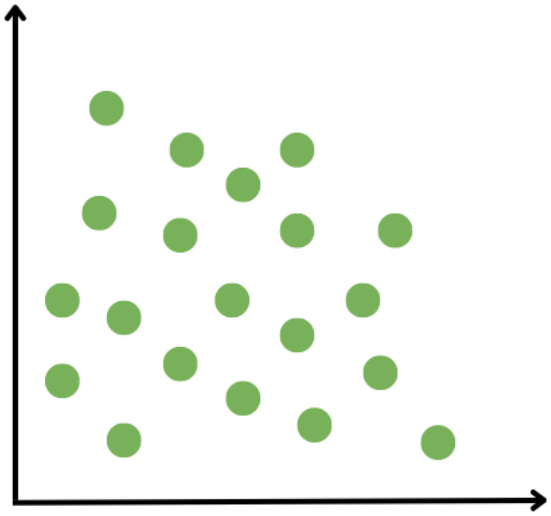

Figure 4.3.1 – Before using K-Means Clustering

Each cluster is assigned a centroid, as this technique is centroids-based. The primary goal of this approach is to reduce the total distances between data points and their respective clusters.

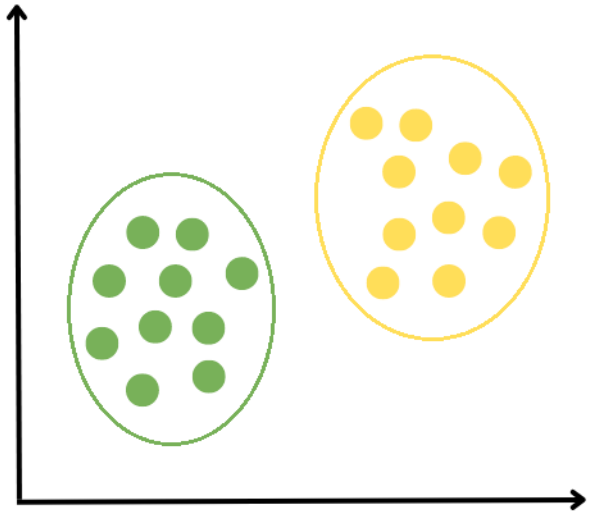

Figure 4.3.2– After using K-Means Clustering

The high and low-frequency crime locations were examined using K-Means Clustering.

4.4 NAÏVE BAYES CLASSIFIER – Naïve Bayes Algorithm is based on Bayes theorem used to solve classification problems. Faster machine learning models that can generate quick predictions can be built using this efficient classification algorithm. Being a probabilistic classifier, it makes predictions based on the probability of an object. [8]

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$ --→ Equation 1

In English terms, this formula is derived as

$$Posterior = \frac{Prior \times Likelihood}{Evidence}$$ --→ Equation 2

5.ARCHITECTURE DIAGRAM OF THE CRIME ANALYSIS

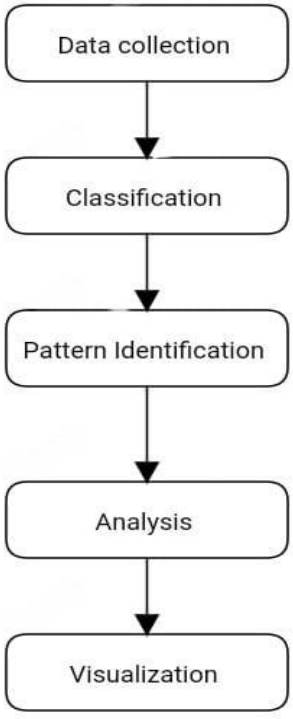

Figure 5.0.1– Architecture diagram of Crime Analysis

5.1 ALGORITHM

1. Analyse various machine learning techniques and choose the one that is most effective.
2. Obtain past crime datasets of Tamil Nadu.
3. Data is classified into various categories.
4. The classified data is analysed using the selected machine learning technique.
5. Then the analysed data is utilized, to examine the characteristics of crime occurrence in every location.
6. The results are helpful in warning the public about crimes that are happening in a location.
7. The results can be used to improve security and lower crime, by predicting when a certain place is most vulnerable to a specific crime.

5.2 DATA COLLECTION - The first significant stage in the construction of a machine learning model is data collection. This is a crucial phase since how well the model performs will be influenced by how much more and better data we can collect. Effective data collection techniques are essential to creating high-performing predictive models. The region wise crime data of Tamil Nadu is obtained from the website "**Kaggle**". Datasets for such Machine Learning projects are typically made available in CSV format (comma-separated values). CSV is a simplistic file format for storing tabular data (number and text) in plain text, like a spreadsheet.

5.3 DATA PRE-PROCESSING - Preparing raw data to be acceptable for a machine learning model is known as data pre-processing. Datasets typically includes corrupt data, missing values, and may be in an undesirable format, which makes it impossible to use directly on machine learning models. Data pre-processing is necessary to clean the data and prepare it for a machine learning model, which also enhances the model's accuracy and effectiveness. The null values in the dataset are removed by dropna() method.

5.4 TRAINING AND TEST – The original dataset is split into two, namely Training data and Test data. The machine learning algorithms are trained how to make predictions for the analysis using the training data. Also, it is the biggest subset of the original dataset. The test dataset examines the model's performance and validates that it generalises successfully to new or unexplored datasets. To split the dataset train_test_split function of scikit-learn is used.[7]

5.5 PREDICTION – Final process in the crime analysis model is prediction. model.predict() function is used to make predictions following model construction after the aforementioned procedure.

5.6 DATA VISUALIZATION – The graphical depiction of information and data in a pictorial or graphical manner is known as data visualization. It is used to discover the trends of the analysed data, and provides us the perspective of the output.

6. RESULT AND DISCUSSION

The result of crime data analysis of Tamil Nadu using Machine learning algorithms are obtained. Data of crimes like Thefts, Robbery, Extortion, Blackmailing, Burglary and Vehicle-Theft are used for the analysis.

Machine Learning Techniques	Accuracy Percentage
Linear Regression	72%
K-Nearest Neighbour	73%
K Means Clustering	82%
Naïve Bayes Classifier	78%

Table 6.1 – Listing the accuracy percentage of the used Machine learning techniques

According to the results from the table, even though every machine learning algorithm has a higher accuracy, the algorithm that may be utilised for the research is K-Means Clustering, which has the highest accuracy among the other algorithms with 79%.

6.1 CRIME DATA VISUALIZATION - The analysis of the dataset and its plotting into graphs, are covered in this part.

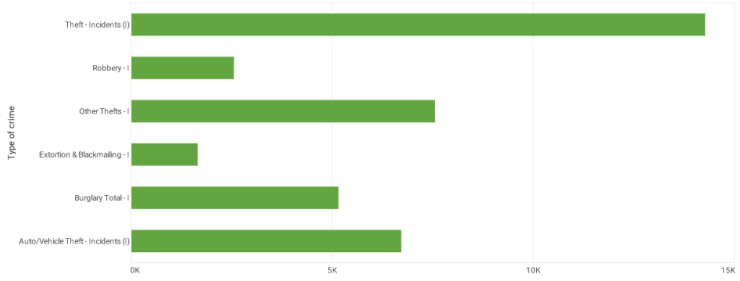

Figure 6.1.1 – Graph of Different Crime Incidents that occurred in Tamil Nadu

The above bar graph shows six types of crimes and number of incidents (of the crimes) occurred in Tamil Nadu (2021 stats alone). The **y** axis of the graph denotes the types of crimes, and **x** axis denotes the number of incidents occurred in the year.

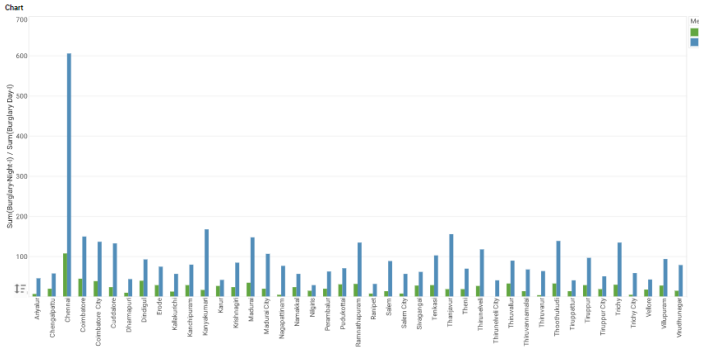

Figure 6.1.2 – Graph of Burglary in the districts of Tamil Nadu.

The data of burglary events in Tamil Nadu are displayed in the double bar graph above, district by district. The graph's y axis indicates the number of burglaries that took place, and the x axis represents the districts of Tamil Nadu. The green bar indicates the incidents of burglary occurred in day time, while blue bar indicates those occurred in night time.

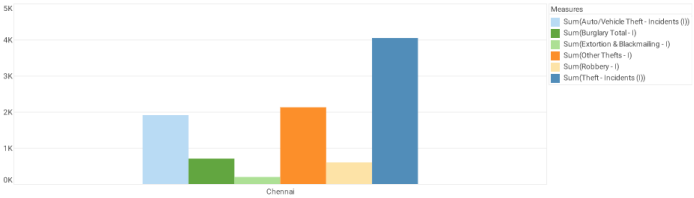

Figure 6.1.3 – Graph of Different Crime Incidents that occurred in Chennai

The above bar graph shows six types of crimes and number of incidents (of the crimes) occurred in the district of Chennai (2021 stats alone). The **y** axis of the graph denotes the types of crimes, and **x** axis denotes the number of incidents occurred in the year.

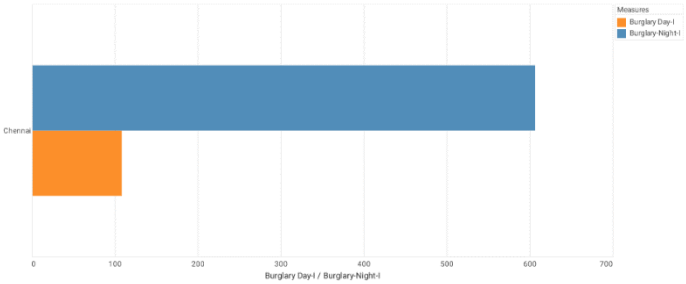

Figure 6.1.4 – Graph of Burglary Incidents that occurred in Chennai

The data of burglary incidences in the Chennai district is displayed in the double bar graph above. Burglary incidents that happened at night are represented by the blue bar, while those that happened during the day are represented by the orange bar.

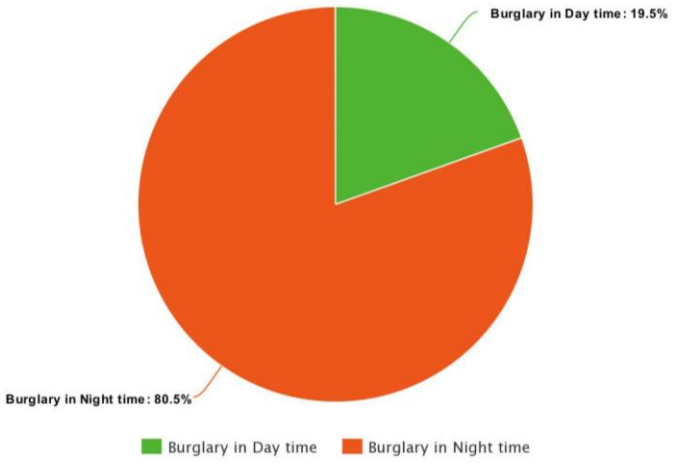

Figure 6.1.5 – Graph of Burglary Incidents that occurred in Chennai

The pie chart above illustrates the time period during which burglary incidents occurred. This chart evidently shows that the certain crime primarily occurs in night time.

7. CONCLUSION

Finding relation and trends among varied crime data has become much simpler with the use of machine learning techniques. The approach of this project mostly consists of utilising these techniques to analyse Tamil Nadu's crime data. The analysis's findings are categorised by geographic area to inform people about the different crimes that are likely to happen there and at which time period. The model created utilising the K-Means Clustering technique benefits in crime analysis with a 79% accuracy rate. Data Visualization is used

to discover the trends of the analysed data, and provides us the perspective of the output. At overall, this outcome would be advantageous since the law enforcement department would strengthen protection during the vulnerable moment, which would reduce crime.

8. REFERENCES

1. Neil Shah,Nandish Bhagat, Manan Shah(2021) A Machine Learning and computer vision approach to crime prediction and prevention. Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention | Visual Computing for Industry, Biomedicine, and Art | Full Text (springeropen.com)

2. Marimuthu Muthuvel, Keerthika.V ,R Krishika, V Kuzhal (2018) Analysis of Crime Data in Tamil Nadu using Regression. (PDF) ANALYSIS OF CRIME DATA IN TAMILNADU USING REGRESSION (researchgate.net)

3. SuhongKim,ParamJoshi, Parminder Singh Kalsi,Pooya Taheri (2018) Crime Anlaysis Through Machine Learning. (PDF) Crime Analysis Through Machine Learning (researchgate.net)

4. OltaLlaha (2020) Crime Analysis and Prediction using Machine Learning(2020) (PDF) Crime Analysis and Prediction using Machine Learning (researchgate.net)

5. Alkesh Bharati, Dr SarvanaguruRA.K (2018) Crime Prediction and Analysis Using Machine Learning IRJET-V5I9192.pdf

6. Train and Test datasets in Machine Learning https://www.javatpoint.com/train-and-test-datasets-in-machine-learning

7. Naïve Bayes Classifier Algorithm https://www.javatpoint.com/machine-learning-naive-bayes-classifier

8. K-Nearest Neighbor (KNN) Algorithm for Machine Learninghttps://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

9. K Means Clustering Algorithm for Machine Learning https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

10. Linear Regression in Machine Learning https://www.javatpoint.com/linear-regression-in-machine-learning

11. Tamil Nadu Crime Review Compendium 2021 https://eservices.tnpolice.gov.in/content/crime_review/tn_cr_compendium_2021.pdf

12. Tamil Nadu Crime Review Statistics 2021 https://eservices.tnpolice.gov.in/content/crime_review/tn_cr_statistics_2021.pdf

13. Crime in India 2021 Statistics (Pg. 45) https://ncrb.gov.in/sites/default/files/CII2021/CII_2021Volume%201.pdf