Mini Project Report

On

CRIME RATE PREDICTION

Submitted To:

Ms. Preeti Choudhary
Project Coordinator
CSE - V Sem - I

Submitted By:

Rashmi
Roll No.: 2018626
CSE - V Sem - I
Session: 2022 - 23

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GRAPHIC ERA HILL UNIVERSITY, DEHRADUN

TABLE OF CONTENT

INTRODUCTION

A huge number of crimes happen every second in different places, in different patterns and in different times and the number is increasing each growing day. Crimes can be categorized into different types be it murder, theft, rape, kidnap, and burglary so on. Crime is violation of law for which the accused is convicted.

Crime is uncertain and cannot be predicted. Crime prediction is significant to determine increase or decrease in crime rate from preceding years. A good prediction technique provides faster crime data set evolution, helps in predicting correct place of crime and criminal as well as aids in keeping track of resources pertaining to analysis of crime.

Through many documentations and cases, it came out that machine learning and data science can make the work easier and faster. The aim of this project is to make crime prediction using the features present in the dataset. The dataset is extracted from the official sites. With the help of machine learning algorithm, using python as core we can predict the type of crime which will occur in a particular area with crime per capita.

The objective would be to train a model for prediction. The training would be done using Training data set which will be validated using the test dataset. The Multi Linear Regression (MLR) will be used for crime prediction. Visualization of dataset is done to analyze the crimes which may have occurred in a particular year and based on population and number of crimes. This work helps the law enforcement agencies to predict and detect the crime per capita in an area and thus reduces the crime rate.

SOFTWARE / HARDWARE REQUIREMENT

Software Required:

- Operating System: Windows 7 or higher
- Programming: Python 3.9 or higher
- Libraries / Frameworks: Tensorflow, Express.JS, NodeJS, Bootstrap

Hardware Required:

- Processor: Any Processor above 1 GHz.
- Ram: 4 GB
- Hard Disk: 4GB
- Input Devices: Standard Keyboard, Mouse
- Output Devices: VGA and High-Resolution Monitor.

TECHNOLOGY USED / ALGORITHMS

1- Machine Learning

Machine learning is a type of artificial intelligence that utilizes algorithms to learn from data, identify patterns, and make decisions without being explicitly programmed. It is an iterative process where data is used to teach the algorithm to identify patterns and explore further. The goal of machine learning is to enable machines to make decisions or predictions without human intervention. It can be used to automate processes, improve customer experience, and provide insights into data.

Machine Learning is an important field of study due to its ability to enable computers to learn from data, identify patterns, make decisions, and improve over time. Machine learning algorithms are used in a variety of applications, from healthcare to finance, to improve decision making and accuracy. In healthcare, machine learning can be used to help diagnose diseases, predict outcomes, and inform treatment decisions. In finance, machine learning can be used to detect fraud, identify market trends, and manage portfolios. Machine learning can also be used to improve autonomous vehicles, natural language processing, and robotics. With its potential to automate many mundane tasks, machine learning is revolutionizing the way we interact with technology.

2 - Support Vector Machine (SVM)

Support Vector Machines (SVMs) are supervised learning algorithms used for both classification and regression tasks. SVMs are based on the concept of decision planes that define decision boundaries. The algorithm attempts to find a hyperplane in an N-dimensional space (where N is the number of features) that optimally separates the data into two classes. The hyperplane is chosen so that it maximizes the margin between the data points of the two classes. In addition, SVM algorithms use Kernel functions to transform the data and find an optimal hyperplane in a higher dimension.

Support Vector Machines (SVMs) is a powerful machine learning algorithm used for classification and regression problems. SVMs are advantageous over other algorithms because they are incredibly versatile, performing well on nonlinear and high-dimensional data. SVMs are also robust to noise, outliers, and non-linearity, making them a great choice for many real-world problems. Additionally, the algorithm has a strong theoretical foundation, which helps us better understand why it works well.

LITERATURE SURVEY

Sr	Author	Year	Name	Summary	Limitation	Prediction
1	Muhammad Alkaff, Nurul Fatanah Mustamin, Gusti Aditya Aromatica Firdaus	2022	Prediction of Crime Rate in Banjarmasin City Using RNN-GRU Model	This study proposes a model to predict the crime rate in Banjarmasin City, Indonesia, by using the Recurrent Neural Network (RNN) with the Gated Recurrent Unit (GRU) architecture. The model takes into consideration the inflation rate and discretionary income. Results show that the GRU-RNN model has an R-Squared value of 0.84 and an RMSE value of 2.21.GRU is a modified RNN algorithm that is simpler than the Long-Short Term Memory (LSTM) Neural Network and is more effective in adapting to different timescales and dealing with Vanishing Gradient problems. It consists of two gates, the Update gate (zt) and the Reset gate (rt), and is compatible with data that is not as much as LSTM, achieving optimal results even with less data.	The predicted value is not too close to the original value compared. Due to lower examples in the dataset, the training of the model becomes difficult.	This study applied a RNN-GRU model to predict data. After collecting and normalizing the data, the model produced the best results with the lowest MAE and RMSE values of 1.7368 and 2.21, respectively, and an R-Squared value of 0.84, indicating good model performance.

2	Qifen Dong, Ruihui Ye, Guojun Li	2022	Crime amount prediction based on 2D convolution and long short-term memory neural network	This paper investigates crime amount prediction by employing spatiotemporal correlations and multiple crime-related auxiliary data sources. The main contributions are: analysis of temporal autocorrelation and spatial correlation in crime data, investigation of the relationships between crime and related auxiliary data sources, design of a crime amount prediction model based on 2D convolution and long short-term memory neural networks (2DCONV-LSTM), and experiments on real-world datasets to evaluate the model's prediction accuracy. Results show that the proposed 2DCONV-LSTM model improves prediction performance compared to SVR and LSTM, and that the performance differs for regions with different crime rates.	The full variant of the proposed model, 2DCONV-LSTM-f, did not achieve the best performance, demonstrating that excessive auxiliary data reduces the performance of the prediction model. Founded that too much auxiliary data may contain redundant information, which negatively impacts the performance of the proposed prediction model.	The proposed 2DCONV-LSTM model was evaluated using different combinations of auxiliary data. It was found that data improves prediction performance, and the 2DCONV-LSTM-d variant obtained the best performance with a RMSE value of 9.981. For community areas with medium-high crime rates and above, the proposed 2DCONV-LSTM model had an advantage, but with a decreasing crime rate, the accuracy of crime prediction was reduced. The a RMSE value of the best 2DCONV-LSTM variant was reduced by 0.5% and 0.3% compared with Just-LSTM in regions with a medium and medium-low crime rate, respectively.

3	Wajiha Safat, Sohail Asghar, Saira Andleeb Gillani	2021	Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques	This study aims to analyse crime prediction in the Chicago and Los Angeles datasets by improving the predictive accuracy with the Logistic Regression, SVM, Naïve Bayes, KNN, Decision Tree, MLP, Random Forest, and XGBoost algorithms, time-series analysis with LSTM, exploratory data analysis for visual summary, and crime forecasting for the crime rate and high intensity crime areas for subsequent years with an ARIMA model. The findings of the study will benefit the police by using identified alleged crime areas to allocate additional resources and protection measures against criminals. Sections 2 to 7 cover the literature review, preliminary classification methods, prediction and performance evaluation measures, data and pre-processing, major findings with a comparative analysis, discussions and future directions, and concluding remarks.	Study takes only 5 year data trend which is very less for futuristic predictions. Further, hybrid models should have been used to expand crime prediction accuracy and to enhance the overall performance	This paper investigated the predictive accuracy of eight different algorithms for the Chicago and Los Angeles datasets, with XGBoost performing best with an accuracy of 94% and 88%, respectively. An LSTM model was also implemented, and the performance metrics of RMSE and MAE were used to measure the scale-dependent error. Additionally, future crime density areas were studied by using an ARIMA model, which forecasted that the crime rate in Chicago will continue to increase moderately in the future, followed by a stable decline, while Los Angeles will experience a sharp decline.

4	Sakib Mahmud, Musfika Nuha, Abdus Sattar	2021	Crime Rate Prediction Using Machine Learning and Data Mining	This paper studied the relationship between crime and different features in the criminology literature. To reduce crime and detect criminal activity, the author used Z-Crime Tools and Advanced ID3 algorithms with data mining technology, K-Means Clustering and deep learning algorithms, random forest and naïve Bayes algorithms, and multi-linear regression. The author also used Apriori algorithm and Naive Bayes algorithm for identification of criminal trends and patterns and prediction of crime rate, respectively. Classification algorithms like Naive Bayes were used to classify each object in a data set into one of the predefined classes or groups	Only last 3 years data has been used to predict the crime rate. The sparsity of crime in many areas complicates the application of the prediction rate area specific modelling.	The accuracy of different algorithms is evaluated, with K-nearest neighbour providing the most precise crime rate forecast system. Linear, Naive Bayes and KNN algorithms had accuracy scores of 73.6%, 69.5% and 76.9% respectively.

5	Gaurav Hajela, Dr. Meenu Chawla, Dr. Akhtar Rasool	2020	A Clustering Based Hotspot Identification Approach For Crime Prediction	The study of crime shows that it can be represented with a spatio-temporal pattern across geographical space. There are many indicators of crime such as urban or census-based indicators, street light and daylight, social media-based indicators, population flow indicators, and climate-based indicators. A crime hotspot is an area with a higher concentration of crime than the rest of the area. This paper proposes a crime prediction model for the dataset of San Francisco, which includes crime hotspot identification, dataset preparation, and crime prediction approach. The performance parameters used to evaluate the model are discussed and the results are presented.	Spitting of the dataset into training, validation and test data should have been done smartly.	The last phase of the crime prediction model is prediction using state of art techniques. Naive Bayes (NB), Naive Bayes with kernel estimator (NB -k), Decision Tree (REPTree) and ensemble approaches are used for classification. The results are compared with and without hotspot analysis and it is found that hotspot analysis improves accuracy. KMeans clustering is used for hotspot identification and X and Y are replaced by the cluster number. Results show that best accuracy is obtained when k=4 and when coupled with hotspot identification. Decision tree approach achieved 83.95 % and outperforms Nave Bayes.

6	Masoomali Fatehkia, Dan O'Brien, Ingmar Weber	2019	Using Facebook interests to improve predictions of crime rates in urban areas	This article discusses the potential for using data from the Facebook Advertising API to gain insight into the distribution of individual-level processes in relation to crime rates across different neighbourhoods. It begins by describing existing theories of carcinogenesis related to factors such as poverty, social disorganization, income inequality, and impulsivity. It then outlines how the API could be used to measure the prevalence of interests among a ZIP code's Facebook population, which can be used to reflect behavioural and attitudinal features of a population. The article concludes by noting that this is an exploratory study, and that the results should be interpreted as predictions of reported crime rates, not necessarily crime itself.	The model have lesser data and can predict only three types of crimes that is assaults, burglaries and robberies.	The results of a series of three regression models for predicting crime rates in three different types of crime: assaults, burglaries and robberies. The models used only demographic factors, only Facebook interests, or both, and controlled for each city's baseline crime rate and the age composition of the neighbourhood. Results showed that the combination of demographic factors and Facebook interests had the greatest predictive strength for all three crime types, both in-sample (using adjusted R2) and out-of-sample prediction (using MAE).

7	Charlie Catlett, Eugenio Cesario, Domenico Talia, Andrea Vinci	2019	Spatio-temporal crime predictions in smart cities: A data-driven approach and experiments	This paper presents an approach to detect high-risk crime regions in urban areas and predict crime trends in each region. The algorithm involves several steps, such as discovering crime dense regions through spatial analysis and discovering predictive models from each region. The paper provides an experimental evaluation on two real-world case studies, Chicago and New York City. The results show the effectiveness of the approach, with good accuracy in spatial and temporal crime forecasting. The paper also compares the results with other regression analysis approaches proposed in literature. It presents an approach for predicting crime trends in urban areas by combining spatial analysis and auto-regressive models. It is based on the idea that there are areas of high risk and low risk for crime, and that crime rates can vary with the season	Clusters where splitted with a very large sizes. Also there was difficulty in getting the correct relationship between the trend in crimes and other events of the city.	A comparative analysis was performed of several approaches for crime predictors extraction, including ARIMA models and three classic regression algorithms (Random Forest, REPTree and ZeroR). Results showed that the ARIMA approach generally achieves greater accuracy than other algorithms for one-year-ahead forecasts in the three highest crime dense regions in both Chicago and New York City. This confirms the appropriateness of the autoregressive model for crime prediction

8	Ms. Vrushali Pednekar, Ms. Trupti Mahale, Ms. Pratiksha Gadhave, Prof. Arti Gore	2018	Crime Rate Prediction using KNN	This system looks at how to convert crime information into a data-mining problem to help detectives solve crimes faster. It focuses on crime analysis, extracting target datasets, data pre-processing, data mining, and interpretation and using discovered knowledge. The proposed model of crime analysis and prediction uses a general algorithm which takes raw data of crime from a government repository as input and produces a correlated dimensions model for crime analysis and prediction as output. It also uses various data mining techniques to predict the frequency of occurring crime based on territorial distribution of existing data. It also involves data cleaning and treating missing values to improve the quality of data for mining. Finally, it provides SQL or reports to interpret the discovered patterns and take actions based on the knowledge.	It will be more accurate if a particular state/region have been considered. Also, The system will not predict the time in which the crime is happening.	The proposed system presents a new framework for clustering and predicting crimes based on real data. Considering the methods proposed for crime prediction shows that the parameters such as the effect of outliers in the data mining pre-processing, quality of the training and testing data, and the value of features have not been addressed before. The proposed system predicts crime prone regions in India on a particular day.

PROPOSED METHODOLOGY

The proposed methodology for the crime rate prediction project using the Support Vector Machine (SVM) algorithm involves collecting and analyzing data from various sources such as police reports, census data, and other relevant sources. The data will be pre-processed and then used to build the model. The SVM algorithm will be used to train the model by optimizing the parameters. Once the model is trained, it will be tested on unseen data to evaluate its performance. Finally, the model can be used to predict the crime rate in different areas. The proposed methodology provides an effective way to accurately predict crime rates and can be used to support law enforcement efforts.

- **Data Collection**: Collect crime data such as crime type, time of occurrence, location, and other necessary data from reliable sources.
- **Preprocessing**: Preprocess the collected data to make it suitable for analysis. This includes removing unnecessary data, handling missing values, and converting categorical variables into numerical variables.
- **Feature Selection**: Identify the most important features that can help in predicting the crime rate.
- **Data Splitting**: Split the available data into train and test sets.
- **Model Training**: Train the Support Vector Machine model using the training dataset.
- **Model Evaluation**: Evaluate model performance on the test dataset.
- **Model Deployment**: Deploy the model in the real-world environment to predict the crime rate.

REFERENCES

- https://ijisae.org/index.php/IJISAE/article/view/2130
- https://onlinelibrary.wiley.com/doi/full/10.4218/etrij.2021-0396
- https://ieeexplore.ieee.org/document/9424589
- https://www.researchgate.net/publication/348152019_Crime_Rate_Prediction_Using_Machine_Learning_and_Data_Mining
- https://www.sciencedirect.com/science/article/pii/S1877050920308231
- https://www.researchgate.net/publication/330861216_Correlated_impulses_Using_Facebook_interests_to_improve_predictions_of_crime_rates_in_urban_areas
- https://www.sciencedirect.com/science/article/abs/pii/S157411921830542X
- https://d1wqtxts1xzle7.cloudfront.net/56658735/Crime_Rate_Prediction_using_KNN-with-cover-page-v2.pdf