IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 7, July 2023

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2023 ||

| DOI:10.15680/IJIRSET.2023.1207144 |

Machine Learning Models for Crime Pattern Recognition and Prediction

Mr. Likhith R[1], Mr. Muthuramalingam B[2]

P. G. Student, Master of Computer Application, Sir M. Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India[1]

Assistant professor, Master of Computer Application, Sir M. Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India[2]

ABSTRACT: In this era of recent times, crime has become an evident way of making people and society under trouble. An increasing crime factor leads to an imbalance in the constituency of a country. In order to analyse and have a response ahead this type of criminal activities, it is necessary to understand the crime patterns. This study imposes one such crime pattern analysis by using crime data obtained from Kaggle open source which in turn used for the prediction of most recently occurring crimes. The major aspect of this project is to estimate which type of crime contributes the most along with time period and location where it has happened. Some machine learning algorithms such as Random Forest Classifier is implied in this work in order to classify among various crime patterns and the accuracy achieved was comparatively high when compared to precomposed works.

KEYWORDS: Crime, Society, Imbalance, Criminal activities, Time period, Location, Machine learning algorithms.

I. INTRODUCTION

This project develops a machine learning system for crime data analysis, prediction, and prevention. It follows a systematic methodology involving data collection, preprocessing, exploratory analysis, model development, crime prediction, user interface creation, and evaluation. The system utilizes diverse crime data, applies machine learning algorithms for classification, and constructs a predictive model for future crime. A user-friendly interface presents visualizations and insights for decision-making. Performance is evaluated using metrics, and the project aims to enhance community safety through accurate crime analysis and proactive measures.

II. RELATED WORK

LITERATURE REVIEW:

1) Crime Analysis Through Machine Learning

AUTHORS: Suhong Kim, Param Joshi, Parminder Singh Kalsi, Pooya Taheri.

Machine-learning-based crime prediction is investigated using Vancouver crime data for the past 15 years. Two data-processing approaches, K-nearest-neighbour and boosted decision tree models, yield a crime prediction accuracy of 39% to 44% for Vancouver.

2) Survey on Crime Analysis and Prediction using Data mining techniques

AUTHORS: Benjamin Fredrick David. H and A. Suruliandi.

Crime's impact is physical and mental, leaving lasting memories. Advancements in technology enable data integration, deep learning, and data mining for predicting crimes, formulating legal strategies, and analysing large data repositories.

PROPOSED:

The collected data is subjected to initial preprocessing using a machine learning filtering and wrapping technique to eliminate irrelevant and duplicate values. This process effectively cleans the data by reducing its dimensionality. Next, the data is split into separate test and training datasets. The model is then trained using both the training and testing datasets. To facilitate easier classification, the crime type, year, month, time, date, and place are mapped to integer values. This mapping enables the analysis of independent relationships between these attributes, initially employing the Random Forest Classifier. The crime features are appropriately labeled, allowing for the

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2023 ||

| DOI:10.15680/IJIRSET.2023.1207144 |

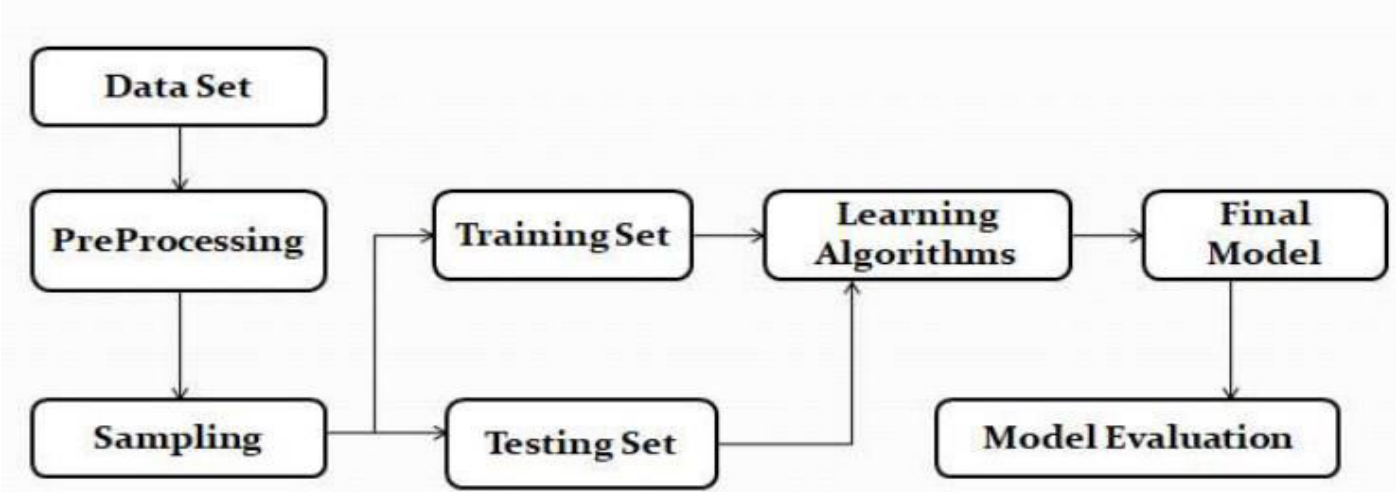
analysis of crime occurrences at specific times and locations. Finally, the most frequently occurring crimes, along with their spatial and temporal information, are determined. The performance of the prediction model is evaluated by calculating the accuracy rate. The prediction model is designed using the Python language and executed using data analysis and machine learning techniques.

- **Advantages of the proposed system:**
- The proposed algorithm is well suited for the crime pattern detection since most of the featured attributes depends on the time and location.
- It also overcomes the problem of analysing independent effect of the attributes.
- The initialization of optimal value is not required since it accounts for real valued, nominal value and concern the region with insufficient information.
- The accuracy has been relatively high when compared to other machine learning prediction model.

III. METHODOLOGY

In the pursuit of advancing public safety, a comprehensive and systematic approach to creating innovative machine learning models for crime pattern recognition and prediction has been devised. This methodology involves meticulous data collection from diverse sources, encompassing historical crime records, demographical data, socio-economic factors, and geographical information, all of which is meticulously anonymized and ethically handled. Preprocessing techniques are applied to address missing values, outliers, and inconsistencies, ensuring the data's suitability for training the machine learning models. Moreover, feature selection methods are employed to identify pertinent attributes, reducing data dimensionality, and promoting model efficiency and interpretability. Model selection encompasses diverse algorithms, both supervised and unsupervised, including decision trees, random forests, support vector machines, and k-means clustering. Rigorous model training, hyperparameter tuning, and evaluation using distinct test datasets are conducted to optimize model performance. Ensuring interpretability, feature importance analysis, and visualization techniques enable comprehension of influential factors driving crime patterns. Ultimately, the most effective model can be deployed in real-world settings for law enforcement assistance, bolstered by ongoing monitoring to maintain reliability and relevance over time.

IV. SYSTEM ARCHITECTURE



V. EXPERIMENTAL RESULTS

IMPLEMENTATION:

- Data Set Reading and Inspection.
- Text Preprocessing.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2023 ||

| DOI:10.15680/IJIRSET.2023.1207144 |

- Analysis.
- Classification.
- Evaluation.

TOOLS AND TECHNOLOGIES USED:

Python: Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.
Flask Framework: Flask is a web application framework written in Python. Armin Ronacher, who leads an international group of Python enthusiasts named Pocco, develops it. Flask is based on Werkzeug WSGI toolkit and Jinja2 template engine. Both are Pocco projects. Http protocol is the foundation of data communication in world wide web. Different methods of data retrieval from specified URL are defined in this protocol. By default, the Flask route responds to the **GET** requests. However, this preference can be altered by providing methods argument to **route()** decorator. In order to demonstrate the use of **POST** method in URL routing, first let us create an HTML form and use the **POST** method to send form data to a URL.

ANALYSIS:
The analysis involves utilizing Matplotlib libraries to create visual representations of various values in order to analyse the properties of the dataset. Graphical analysis provides valuable insights and helps guide predictions. To facilitate this analysis, the dataset is typically split into training and testing sets. A common practice is to allocate approximately 70% of the dataset for training and 30% for testing. However, another commonly used ratio is 80:20, where 80% of the dataset is used for training and 20% for testing. These ratios determine the proportion of data used for training machine learning models and evaluating their performance.

CLASSIFICATION:

KNN:
The k-nearest neighbors algorithm (k-nn) is a versatile technique used for both classification and regression in pattern recognition. It determines the output based on the k closest training examples in the feature space. In classification, it assigns an object to the class most common among its k nearest neighbors. In regression, it estimates the property value by averaging the values of the k nearest neighbors.

Random forest classifier:
A random forest is a powerful ensemble learning method that combines multiple decision tree classifiers. It improves predictive accuracy and reduces overfitting by fitting these decision trees on different subsets of the dataset and using averaging to generate the final predictions. The size of the subsets, known as sub-samples, can be controlled using the max_samples parameter. By default, if bootstrap=true, random forest uses sub-samples, but if bootstrap=false, it utilizes the entire dataset to build each tree.

XGboost:
Xgboost is a highly effective ensemble machine learning algorithm that is based on decision trees and utilizes a gradient boosting framework. While artificial neural networks excel in prediction problems with unstructured data like images and text, decision tree-based algorithms are currently regarded as the top performers for small-to-medium structured or tabular data.

Evaluation:
After creating a model, it is crucial to validate it using real-time data values. This process is commonly referred to as validation, where the predicted value of the model is compared with the actual output value. By assessing the agreement between the predicted and actual values, the model's performance and accuracy can be evaluated.

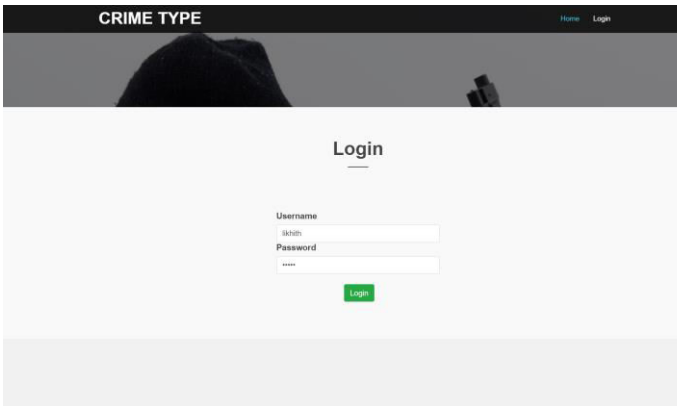
IMPLEMENTATIONS:
This phase is very important and crucial part in any project as it gives the functionality to the design of the project. Each screenshots represents the part of project which is there in our project.

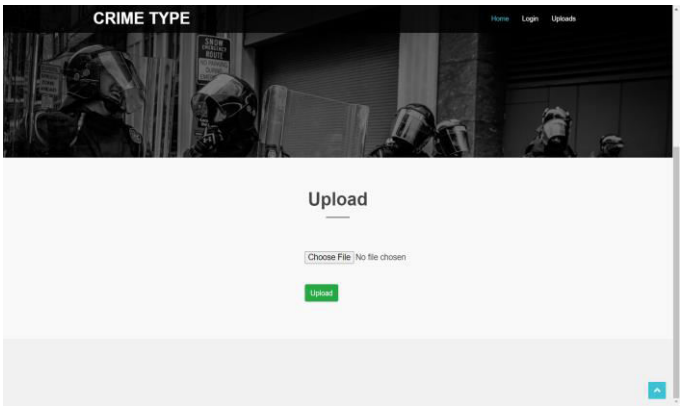
|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

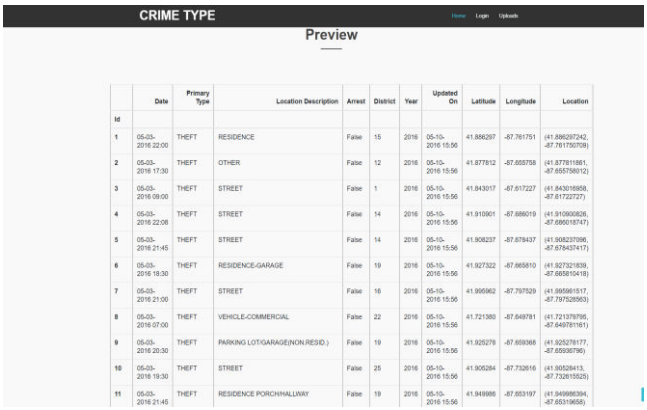
|| Volume 12, Issue 7, July 2023 ||

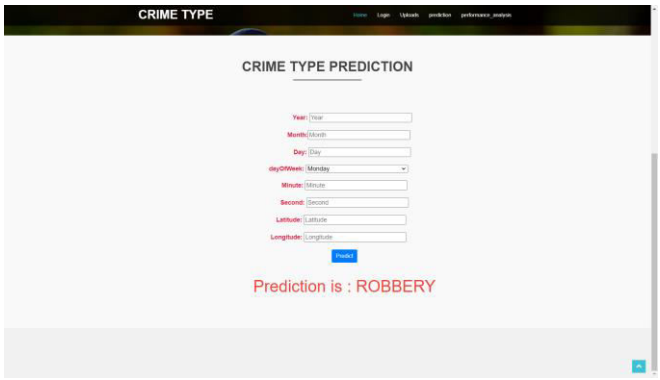
| DOI:10.15680/IJIRSET.2023.1207144 |

SCREENSHOTS:









VI. CONCLUSION

This system effectively handles nominal distribution and real-valued attributes using Multinomial NB and Gaussian NB classifiers. Real-time predictions with minimal training time are achieved, and the system successfully handles continuous target variables. Random Forest Classification enables the identification of the most frequent crimes. Standard metrics like precision, recall, F1 score, and accuracy assess the algorithm's performance. Incorporating machine learning algorithms significantly improves accuracy in this context.

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2023 ||

| DOI:10.15680/IJIRSET.2023.1207001 |

VII. FUTURE ENHANCEMENT

1. Absence of class labels: Address the limitation of relying on class labels for accurate estimation. Explore methods to handle situations without provided class labels.
2. More classification models: Incorporate additional machine learning classification models to enhance crime prediction accuracy. Utilize diverse models with unique strengths and weaknesses for more reliable predictions.
3. Enhancing performance: Improve the system's overall performance by leveraging multiple classification models. Achieve higher accuracy rates and enhance effectiveness in identifying and predicting criminal activities.
4. Considering income information: Incorporate income information into the analysis to uncover relationships between income levels and crime rates. Develop more accurate predictive models by considering this additional feature.
5. Future study and improvement: Analyze income information alongside crime data for further improvements. Gain insights into the relationship between income levels and crime rates to inform crime prevention strategies. Contribute to a comprehensive understanding of crime patterns and guide policies to reduce crime rates.

REFERENCES

Chen, Ling, and Xu Lai. "Comparison between ARIMA and ANN models used in short-term wind speed forecasting." Power and Energy Engineering Conference (APPEEC), 2011 Asia- Pacific. IEEE, 2011.
[2]. Agarwal, Jyoti, Renuka Nagpal, and Rajni Sehgal. "Crime analysis using K-means clustering." International Journal of Computer Applications 83.4 (2013).
[3]. Sathyadevan, Shiju, and Surya Gangadharan. "Crime analysis and prediction using data mining." Networks & Soft Computing (ICNSC), 2014 First International Conference on. IEEE, 2014.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.423

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details