Project Report

On

Crime Category Prediction

Submitted to D Y Patil International University, Akurdi, Pune in partial fulfilment of full-time degree

Master of Computer Applications

Submitted By:

Name: Miss. Prajakta Bote PRN: 20230804067

Name: Mr. Prit Mahant PRN: 20230804059

Name: Mr. Harshilkumar Munjapara PRN: 20230804004

Under the Guidance of

Ms. Priyanka Karale

School of Computer Science, Engineering and Applications

D Y Patil International University, Akurdi,Pune, INDIA, 411044

[Session 2023-24]

D Y PATIL
INTERNATIONAL
UNIVERSITY
AKURDI PUNE

CERTIFICATE

This report on Crime Category Prediction is submitted for the partial fulfillment of project,which is part of the First Year Master of Computer Applications curriculum, under my supervision and guidance.

14-12-23

Ms. Priyanka Karale
(DYPIU Guide)

1. Prajukta Bote 20230804067
2. Prit Mahant 20230804059
3. Harshil Munjapara 20230804004

DECLARATION

I, hereby declare that the following Project which is being presented in the Project entitled as Crime Category Prediction is an authentic documentation of my own original work to the best of my knowledge. The following Project and its report in part or whole, has not been presented or submitted by me for any purpose in any other institute or organization. Any contribution made to my work, with whom i have worked at D Y Patil International University, Akurdi, Pune, is explicitly acknowledged in the report.

Name: Prajakta Bote	PRN No: 20230804067	Signature :
Name: Prit Mahant	PRN No: 20230804059	Signature :
Name: Harshil M.	PRN No: 20230804004	Signature :

ACKNOWLEDGEMENT

With due respect, we express our deep sense of gratitude to our respected guide (Name of guide), for his/her valuable help and guidance. We are thankful for the encouragement that he/she has given us in completing this Project successfully.

It is imperative for us to mention the fact that the report of project could not have been accomplished without the periodic suggestions and advice of our project supervisor (Name).

We are also grateful to our respected, Dr. Bahubali Shiragapur(Director), Dr. Maheshwari Biradar (HOD BCA & MCA) and Hon'ble Vice Chancellor, DYPIU, Akurdi, Prof. Prabhat Ranjan for permitting us to utilize all the necessary facilities of the college.

We are also thankful to all the other faculty, staff members and laboratory attendants of our department for their kind cooperation and help. Last but certainly not the least; we would like to express our deep appreciation towards our family members and batch mates for providing support and encouragement.

Name: Prajakta Bote PRN: 20230804067
Name: Prit Mahant PRN: 202308059
Name: Harshil Munjapara PRN: 20230804004

Abstract

Crime categorization is a crucial aspect of crime analysis, enabling law enforcement agencies to effectively allocate resources and implement targeted crime prevention strategies. This project aims to develop a machine learning model capable of accurately predicting crime categories based on historical crime data. The project encompasses data preprocessing, feature engineering, model selection, and evaluation.

Data Preparation

The project utilizes a comprehensive dataset of crime incidents, encompassing various attributes such as crime location, time, type, and potential contributing factors. Data preprocessing involves cleaning, handling missing values, and encoding categorical variables to ensure compatibility with machine learning algorithms.

Feature Engineering

Feature engineering plays a critical role in extracting meaningful insights from the raw data. The project employs feature extraction techniques to transform raw data into features that are relevant to crime prediction. This includes creating new features, such as time-based features and location-based features, to enhance the predictive power of the model.

Model Selection and Evaluation

A variety of machine learning algorithms are evaluated for their ability to predict crime categories effectively. The project employs techniques such as k-nearest neighbors (KNN), support vector machines (SVM), and decision trees to identify the most suitable algorithm for the task. Model evaluation involves assessing performance metrics such as accuracy, precision, and recall to determine the best-performing model.

Implications and Future Directions

The successful development of a crime category prediction model has significant implications for law enforcement agencies. By accurately predicting crime categories, agencies can proactively allocate resources to areas of high crime risk, implement targeted prevention strategies, and enhance overall crime prevention efforts.

Future research directions include exploring the integration of additional data sources, such as social media data and demographic data, to further improve the predictive accuracy of the model. Additionally, investigating ensemble methods, which combine multiple machine learning algorithms, could potentially enhance the robustness and generalizability of the model.

TABLE OF CONTENTS

List of Figures

1. INTRODUCTION

1.1. Background

Crime categorization is a fundamental aspect of crime analysis, enabling law enforcement agencies to effectively allocate resources and implement targeted crime prevention strategies. Traditionally, crime categorization has been a manual process, relying on human experts to review crime reports and assign categories based on their judgment. This approach is often time-consuming, subjective, and prone to errors.

With the advent of machine learning, there is a growing opportunity to automate and improve the process of crime categorization. Machine learning algorithms can analyze large datasets of crime data and identify patterns and relationships that may be difficult for humans to discern. This information can then be used to develop predictive models that can automatically assign crime categories to new incidents.

1.2. Objectives

The objective of this project is to develop a machine learning model that can accurately predict crime categories based on historical crime data.

1.3. Purpose

- Improve the accuracy and efficiency of crime categ
- Deployment of the model to a production environment
- Enhance overall crime prevention efforts

1.4. Scope

- Data collection and preprocessing
- Feature engineering
- Model selection and evaluation
- Deployment of the model to a production environment

1.5. Applicability

The model developed in this project can be applied to crime data from a variety of jurisdictions. The model can be used to predict crime categories for new incidents, as well as to recategorize historical incidents. The model can also be used to identify trends in crime patterns over time.

This project has the potential to make a significant contribution to crime prevention efforts by providing law enforcement agencies with a tool that can help them to better understand and respond to crime.

2. PROJECT PLAN

2.1. Problem Statement

The manual process of crime categorization is time-consuming, subjective, and prone to errors. This can lead to:

- Inefficient allocation of resources: Law enforcement agencies may not be able to allocate resources effectively to areas of high crime risk if they do not have accurate information about crime categories.
- Untargeted crime prevention strategies: Crime prevention strategies may not be targeted to the most relevant crime categories if they are not based on accurate information.
- Decreased effectiveness of crime prevention efforts: Overall crime prevention efforts may be less effective if they are not based on accurate information about crime categories.

2.2. Proposed Solution

Develop a machine learning model that can accurately predict crime categories based on historical crime data. This model would be able to:

- Categorize crime incidents automatically: The model would be able to automatically assign crime categories to new incidents, which would save law enforcement agencies a significant amount of time.
- Identify areas of high crime risk: The model would be able to identify areas of high crime risk, which would allow law enforcement agencies to allocate resources accordingly.
- Develop targeted crime prevention strategies: The model would be able to help law enforcement agencies develop targeted crime prevention strategies by providing information about the most relevant crime categories.

2.3. Benefits of Proposed Solution

The proposed solution would have the following benefits:

- Improved accuracy and efficiency of crime categorization: The model would be able to categorize crime incidents more accurately and efficiently than humans.

- More effective allocation of resources: Law enforcement agencies would be able to allocate resources more effectively to areas of high crime risk.
- More targeted crime prevention strategies: Crime prevention strategies would be more targeted to the most relevant crime categories.
- Enhanced overall crime prevention efforts: Overall crime prevention efforts would be more effective.

2.4. Requirement Specification

A Requirements Specification is a critical document that outlines the functional and non-functional requirements of a system. Here's a sample template for a Requirements Specification document for a crime category prediction project using the K-Nearest Neighbors (KNN) algorithm:

2.4.1. Functional Requirements

The crime category prediction system shall fulfill the following functional requirements:

Data Import and Cleaning: The system shall be able to import crime data from various sources, handle missing values, and clean the data to ensure its integrity and consistency.

Feature Extraction and Transformation: The system shall be able to extract relevant features from the crime data, transform features to improve their representation, and create new features to enhance the predictive power of the model.

Model Training and Evaluation: The system shall be able to train various machine learning models on the prepared data, evaluate their performance using appropriate metrics, and select the best-performing model for crime category prediction.

Prediction and Output: The system shall be able to generate predictions for new crime incidents based on the selected model and provide the output in a clear and understandable format.

2.4.2. Non-Functional Requirements

The crime category prediction system shall adhere to the following non-functional requirements:

Accuracy: The system shall achieve a high level of accuracy in predicting crime categories.

Efficiency: The system shall be able to process and analyze large datasets of crime data efficiently.

Scalability: The system shall be scalable to accommodate future growth in crime data volume.

User-Friendliness: The system shall provide a user-friendly interface that is easy to navigate and understand.

Reliability: The system shall be reliable and maintain high availability to support critical crime analysis tasks.

3. PROPOSED SYSTEM AND METHODOLOGY

The system will look at how to convert crime information into a data-mining problem, so that it will help detectives in solving crimes faster. In terms of crime a cluster is a group of crimes in a geographical region or a hot spot of crime. Whereas, in terms of data mining a cluster is the group of a particular set of objects based on their characteristics of possible crime pattern. Thus relevant clusters or a subset of the cluster will have a one-to-one correspondence to crime patterns. The Proposed system focuses on:

Crime analysis based on available information to extract crime patterns.

Using various data mining techniques, frequency of occurring crime can be predicted based on territorial distribution of existing data.

Crime recognition.

Data mining is needed for crime analysis, because the last is an iterative process of extracting knowledge hidden from large volumes of raw data. To present the proposed model of crime analysis and prediction using data mining, first will begin with a big view of this model explained in the following algorithm:

General Algorithm of Proposal Model

Input : Raw data of crime from Government Repository.
Output : Correlated dimensions model for crime analysis and prediction.

3.1. System Architecture

1.Understanding the crime domain: Goals of the crime prediction and detection includes appropriate prior knowledge.

2.Extracting the target dataset: This is for building a dataset for the three dimensions of the proposed model; crime, criminal and geo-crime. By focusing on a subset of variables, feature selection will be done which is not affected by crime conflictions and geo-crime environment changes.

3.Data pre-processing: For mining it is required to improve actual quality of data. The time

required for mining the preprocessed data is reduced and it also increases mining efficiency. In our proposal we focus on data preprocessing to involve data cleaning and treating missing values.

4.Data mining: To introduce correlated patterns AR is applied on each dimension dataset among the three dimensions to advance the crime analysis.

5.Interpretation and Using discovered knowledge: This includes providing SQL or reports for both separated and correlated dimensions to interpret the discovered patterns. By taking actions based on the knowledge it helps to incorporate this knowledge into the performance system.

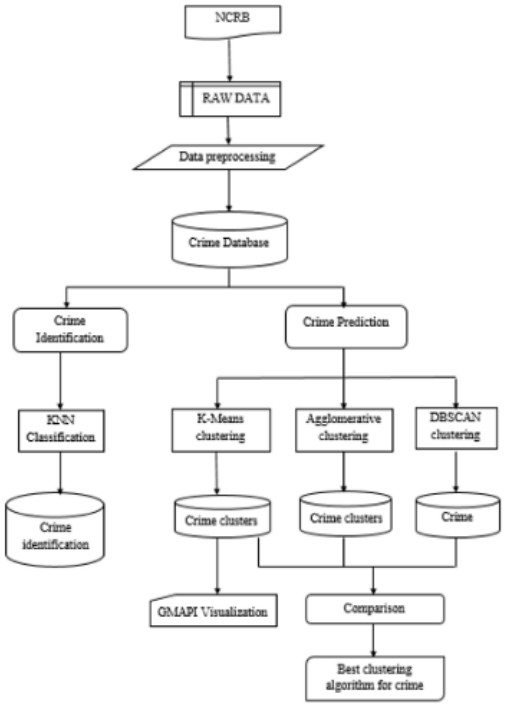

Fig. : System Architecture

3.2. Methodology (Algorithms used)

- KNN

  K-Nearest Neighbors (KNN) is a versatile and intuitive supervised machine learning algorithm designed for classification and regression tasks. At its core, KNN relies on identifying the majority class or average of the k-nearest data points in the feature space to make predictions. The main parameter, 'k,' determines the number of neighbors

considered, influencing the model's flexibility and sensitivity to noise. During the training phase, KNN memorizes the entire training dataset, and predictions are made based on the proximity of new data points. Common distance metrics, such as Euclidean or Manhattan distance, play a crucial role in determining the closeness of neighbors. While KNN is non-parametric and does not assume a specific data distribution, its computational complexity can be a challenge for large datasets due to the need to calculate distances for each new data point against all training samples. The algorithm's performance can also be impacted by the curse of dimensionality, particularly in high-dimensional spaces. Despite these considerations, KNN finds application in various domains, including image recognition, recommendation systems, and medical diagnosis, making it particularly effective for small to medium-sized datasets with nonlinear decision boundaries.

- Random Forest

  Random Forest is a powerful ensemble learning algorithm widely used for both classification and regression tasks in machine learning. Comprising multiple decision trees, it operates by constructing a multitude of trees during training and outputs the mode (classification) or mean prediction (regression) of the individual trees. The randomness element stems from the selection of a random subset of features for each tree and bootstrapped samples from the training data. This diversity helps mitigate overfitting and enhances the model's generalization performance. The aggregation of predictions from various trees often leads to more robust and accurate outcomes compared to individual trees. Random Forest is less prone to the overfitting that single decision trees might exhibit, making it a popular choice in practice. Moreover, it handles missing values well and provides a feature importance measure, aiding in interpretability. While computationally intensive, its parallelization capability makes it suitable for large datasets. The algorithm's versatility extends to applications in areas such as finance, healthcare, and remote sensing.

- Dicision Tree

  A Decision Tree is a fundamental machine learning algorithm used for both classification and regression tasks. It works by recursively partitioning the data based on feature values, creating a tree-like structure where each internal node represents a decision based on a feature, and each leaf node corresponds to a predicted outcome. The goal is to create decision rules that lead to accurate predictions. Decision Trees are interpretable and easy to visualize, making them valuable for understanding the decision-making process. However, they can be prone to overfitting, especially with complex trees. Techniques like pruning are employed to mitigate this issue. Decision Trees are well-suited for datasets with a mix of categorical and numerical features. They serve as the foundation for ensemble methods like Random Forests. The algorithm's

simplicity, interpretability, and ability to handle non-linear relationships contribute to its widespread use in fields such as finance, healthcare, and marketing.

3.3. Implementation

1) Data Collection
2) Classification
3) Pattern Identification
4) Prediction
5) Visualization

1) Data Collection:
Large amount of crime data is collected at police records. This data is made available by National Crime Bureau of Records. This data is in the form of number of cases recorded all over the nation throughout the year. The data is in unprocessed form and contains some wrong as well as missing values. Hence preprocessing of data is crucial task in order to bring the data in proper and clean form. Pre-processing of data includes data cleansing and Preprocessing.

2) Classification: The dataset is classified into various groups based on certain characteristics of the data object. Grouping of crimes is done according to states cities. Classification of the crime is done on the basis of different types of crime. K-mean algorithm can be used to group or cluster data with similar characteristics.

3) Pattern Identification In these phase proposed system have to identify trends and patterns in crime. The result of this phase is the crime pattern for a particular place. Here corresponding to each location we take the attributes of that place like weather attributes, area sensitivity, notable event, presence of criminal groups etc. Information regarding patterns helps police officials to facilitate resources in an effective manner.

4) Prediction Corresponding to each place it builds a model. So for getting the crime prone areas we pass current date and current attributes into the prediction software. The result is shown using some visualization mechanisms.

5) Visualization The crime prone areas can be graphically represented using a heat map which indicates level of activity, usually darker colors to indicate low activity and brighter colors to indicate high

3.4. Flow Chart Diagram

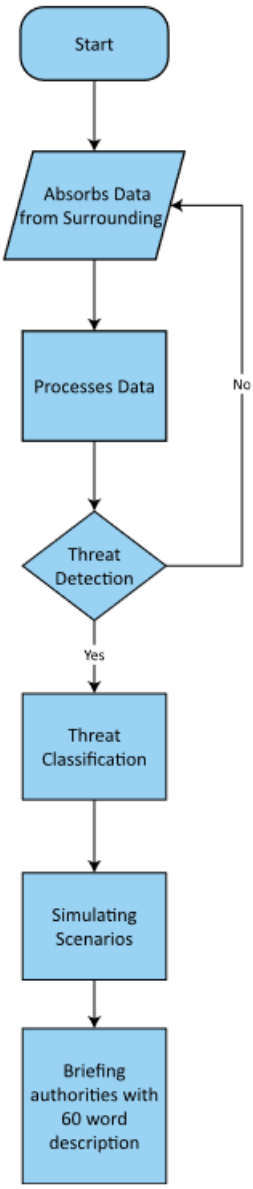

Fig. : Flow Chart

4. RESULTS AND EXPLANATION

4.1. Implementation Approaches

The three implementation approaches for the crime category prediction project are sequential, phased, and parallel. The sequential approach involves completing each step of the project in a linear fashion, while the phased approach divides the project into smaller phases and completes each phase separately. The parallel approach involves working on multiple steps of the project simultaneously. The choice of approach will depend on the specific needs of the project, such as the complexity of the project, the number of team members, and the desired timeline.

4.2. Pseudo Code

Import necessary libraries: sklearn, matplotlib, numpy, pandas

1.Load crime data: Read crime data from CSV or database Convert data into appropriate data structures (e.g., DataFrame)

2.Data Preprocessing: Handle missing values (e.g., imputation, deletion) Encode categorical variables (e.g., one-hot encoding) Normalize numerical variables (e.g., standardization, min-max scaling)

3.Feature Engineering: Identify and extract relevant features from the data Create new features if necessary (e.g., time-based features, location-based features) Transform features to improve their representation (e.g., dimensionality reduction)

4.Model Selection and Training: Split data into training and testing sets Choose and train machine learning algorithms (e.g., KNN, SVM, decision trees) Evaluate model performance on training data using appropriate metrics (e.g., accuracy, precision, recall)

5.Model Evaluation and Selection: Evaluate model performance on testing data using selected metrics Select the best-performing model based on evaluation results

6.Crime Category Prediction: Use the selected model to predict crime categories for new incidents

7.Output and Visualization: Present prediction results in a clear and understandable format Visualize results using charts, graphs, or maps to gain insights

4.3. Analysis (graphs/chart)

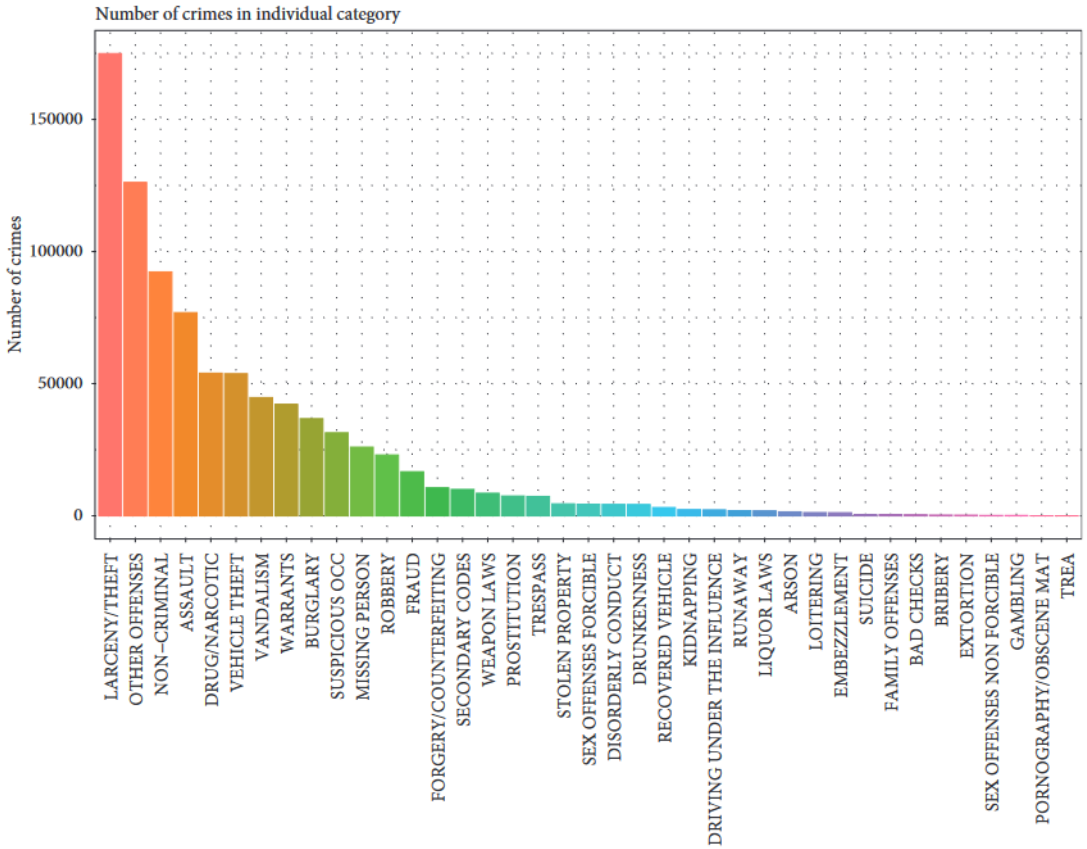

Fig. : Number of crimes in individual category.

5. CONCLUSION

The successful implementation of a crime category prediction model can have a significant impact on law enforcement agencies' ability to understand crime patterns, allocate resources effectively, and enhance crime prevention efforts. The model can automatically categorize crime incidents, identify areas of high crime risk, and inform the development of targeted crime prevention strategies. By leveraging machine learning to analyze historical crime data, law enforcement agencies can gain valuable insights that can help them to reduce crime and improve public safety.

References

[1] Kaggle [Online]. Available: https://www.kaggle.com/

[2] Overleaf [Online]. Available: https://www.overleaf.com/

[3] Google AI Bard [Online]. Available: https://ai.googleblog.com/2022/01/lamda-language-model-for-dialogue.html

[4] ChatGPT [Online]. Available: https://chat.openai.com/beta

[5] YouTube [Online]. Available: https://www.youtube.com/

[6] GitHub [Online]. Available: https://github.com/