International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Artificial intelligence and machine learning forcrime data analysis

Djidji Etienne Ahikpa
Parul Institute of Computer Application
Parul University
Vadodara, Gujarat, India
etienneahikpa@gmail.com

Vijya Tulsani Associate Professor,
Parul Institute of Computer ApplicationParul Univeristy,
Vadodara, Gujarat, India
Vijya.tulsani42087@paruluniversity.ac.in

Abstract— **The classification of an act as a crime within the realm of criminal law is contingent upon the severity of the offense, encompassing intentional actions causing harm to oneself, others, or property damage. The escalating prevalence and diversity of an unlawful activities underscore the urgency in establishing efficient enforcement agencies. Conventional crime-solving methods, characterized by their tardiness and inefficacy, are no longer proficient in addressing the contemporary surge in criminal activities. Consequently, there is an opportunity to alleviate the workload of police personnel and contribute to crime prevention by developing reliable methods for anticipating crimes before they occur. To achieve this, the implementation of Machine Learning (ML) technologies is suggested. This study presents finding from several cases employing such strategies, simulating further interest in exploration. The shift in crime detection and prevention strategies is fundamentally rooted in observations made by authorities. The primary focus of this research is to demonstrate how machine learning can enhance the speed and accuracy of crime detection, prevention, and resolution. The overarching goal is to enable law enforcement and other authorities to swiftly and accurately detect, prevent, and solve crimes. This research underscores the transformative potential of artificial intelligence and machine learning in reshaping law agencies.**

Keywords—artificial intelligence, machine learning, crime data, decision tree, random forest

I. INTRODUCTION

In recent years, crime has become a major issue in many cities around the world. With the increasing availability of digital data and, artificial intelligence (AI), and machine learning, it has become possible to analyze crime data in more detail than ever before. In this seminar project, we will explore how artificial AI, and machine learning can be used to analyze crime data, with a particular focus on application areas and methodologies.

A. Overview of the project

The project "Artificial intelligence and Machine Learning for Crime Data Analysis" aims to harness advanced technologies, including artificial intelligence and machine learning to analyze crime data in a more efficient and accurate manner. The project will focus on the development and other relevant data sources to identify and predict criminal activity. The project will involve data collection from various sources and pre-processing to ensure data quality and accuracy. Machine learning algorithms like supervised, unsupervised, and deep learning techniques will be implemented to classify, cluster, and predict criminal activity.
The project will provide valuable insights into crime patterns and trends, which can aid in the development of effective crime prevention strategies. It can also help law enforcement agencies to identify potential threats and respond quickly to criminal activity. The project has the potential to make a significant impact on crime prevention and public safety, making it a valuable area of research and development.

B. Problem statement and objectives

The problem statement for this project is that crime data analysis can be efficiently done by using artificial intelligence and ML techniques despites its complexity occurring while using traditional techniques. Traditional crime analysis methods rely on manual data processing and analysis, which can be error-prone and time-consuming.
The objective of this project is to develop and implement machine learning techniques for crime data analysis, with the aim of improving crime prediction, pattern detection, and law enforcement decision-making. Specifically, the project aims to:

- Collect and pre-process crime data from the selected sources, such as surveillance cameras, social media and crime reports, website like Kaggle.
- Apply machine learning algorithms to predict crime and pattern detection, using supervised and unsupervised learning techniques.
- Evaluate the performance of the developed models and compare them with traditional crime analysis methods.
- Provide insights and recommendations to law enforcement agencies on how to use the developed models to improve their crime prevention and investigation strategies.

II.METHODOLOGIES

1) Data collection: Gathering data from the selected sources, such as crime reports, surveillance cameras, and other sources of crime-related data.

Data collection is a crucial step in crime data analysis, as it provides the raw material that machine learning algorithms can analyze.

Some common sources of crime-related data that can be used for data collection are:

Crime Reports: Law enforcement agencies maintain crime reports that contain information about crimes, such as the crime type, its location, the timestamp, and any other relevant details.

Surveillance Cameras: Video footage from surveillance cameras can be used to gather data on criminal activity, such as the movements of suspects and the timing and location of crimes.

Social Media: Twitter, Facebook, and Instagram may provide valuable information into criminal activity.

Public Records: Public records such as court documents, arrest records, and property records can provide valuable data on criminal activity and trends.

Sensors: Sensors can be used to gather data on criminal activity, such as the sound of gunshots or the presence of chemical substances associated with drug use.

Mobile Devices: Mobile devices such as smartphones can provide data on the movements and locations of individuals, which can be useful in tracking suspects and analyzing criminal patterns.

For our project we chose the "`Crimes_2001_to_Present.csv`" dataset from kaggle (https://www.kaggle.com/code/criptonkaegrey/chicago-crime-analysis)

```
[ ] df = pd.read_csv("/content/Crimes_-_2001_to_Present.csv")

[ ] df.columns

Index(['ID', 'Case Number', 'Date', 'Block', 'IUCR', 'Primary Type',
       'Description', 'Location Description', 'Arrest', 'Domestic', 'Beat',
       'District', 'Ward', 'Community Area', 'FBI Code', 'X Coordinate',
       'Y Coordinate', 'Year', 'Updated On', 'Latitude', 'Longitude',
       'Location'],
      dtype='object')
```

2) Data pre-processing: Cleaning and processing the data to prepare it for analysis.

Pre-process the data is essential in preparing data for analysis, as raw data often contains errors, inconsistencies, and missing values that can negatively impact the accuracy and effectiveness of machine learning algorithms. Here are some common steps involved in data pre-processing for crime data analysis:

Data Cleaning: It includes error detections and identification in the dataset, example missing values, incorrect data types, and formatting errors. This step ensures that the data is accurate and consistent and can be used for analysis.

Data Integration: It includes combining data from different sources into a single dataset. This step ensures that all relevant data is included in the analysis and that there are no inconsistencies or redundancies in the data.

Data Transformation: It consists in converting the data into acceptable format for data analysis. This step may involve scaling or normalizing the data, converting categorical data into numerical data, or creating new variables based on the existing data.

Data Reduction: Data reduction involves reducing the amount of data that needs to be analyzed while preserving the relevant information. This step can be accomplished through techniques such as feature selection, which identifies the most important variables for analysis, or dimensionality reduction, which reduces variables amount in the data while retaining the most necessarily data.

Data Discretization: Data discretization involves converting continuous data into discrete data. This step can be useful in crime data analysis for identifying patterns and trends, such as the frequency of crimes at certain times or in certain locations.

3) Feature extraction: Identifying important features in the data, such as facial features or licence plate numbers.

It consists in select the most important features in the data that can be used for analysis. In machine learning, features are the measurable characteristics of an object or scene that can be used to identify or classify it. Here are some common approaches to feature extraction in crime data analysis:

Facial Features: In facial recognition, the features extracted from a face might include the distance between the eyes, the shape of the nose, and the curvature of the lips. These features can be used to identify individuals and track their movements.

Licence Plate Numbers: Licence plate recognition systems extract features such as the shape and colour of the licence plate, as well as the characters on the plate, to identify vehicles of interest.

Object Features: Object detection systems extract features such as the shape, size, and texture of objects to identify them within an image or video.

Audio Features: In audio analysis, features such as pitch, loudness, and spectral content can be extracted from audio recordings to identify specific sounds or speech patterns.

4) Model training: Using machine learning algorithms to train models for crime detection and prevention.

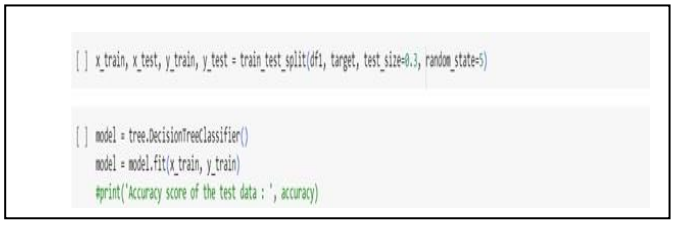
It includes using machine learning algorithms to learn and identify relationships from data to detect and prevent crimes. Some common steps involved in model training for crime detection and prevention:

Data Preparation: Before training a model, the data must be prepared by cleaning, transforming, and reducing it as necessary. This step ensures that the data is suitable for training and that the most important features have been identified.

Selection of Algorithm: A lot of machine learning algorithm are available for crime data analysis, such as decision trees, neural networks, and support vector machines.

The specific analysis task, the size and complexity of the data, and the performance requirements of the system are the requirement of algorithm selection.

Feature Engineering: Feature engineering involves selecting or transforming the most important features in the data for use in the model.
This step can be iterative, with features added or removed based on their impact on model performance.
Model Training: Model training involves feeding the prepared data into the selected algorithm and using it to learn patterns and relationships in the data. The model is adjusted based on its performance on a validation dataset.
Model Evaluation: Once the model has been trained, it must be evaluated on a test dataset to ensure that it is accurate and effective. Evaluation metrics such as accuracy.
Model Deployment: Once the model has been evaluated and validated, it can be deployed in a real-world crime detection or prevention system.



5) Model evaluation: Evaluating the performance of the trained models and making improvements where necessary.

Model evaluation is an important step in the crime data analysis pipeline as it involves assessing the performance of the trained models and making necessary improvements. Here are some common methods for model evaluation in crime data analysis:

Accuracy: Accuracy measures how often the model makes correct predictions. While accuracy is a commonly used metric, it may not be appropriate in all cases, particularly when the data is imbalanced.

Precision and Recall: It evaluates the proportion of true positives among all positive predictions. These metrics are particularly useful in cases where false positives or false negatives are costly.

Confusion Matrix: A confusion matrix is a table that summarizes the performance of a classification model by comparing the actual labels to the predicted labels. The matrix shows the number of true positives, true negatives, false positives, and false negatives.

Area Under the Curve (AUC): AUC is a metric that summarizes the performance of a model across all possible threshold values. It provides a single number that represents the overall quality of the model.

Cross-Validation: Cross-validation is a technique that involve dividing the dataset into several folds and training the model on each fold while assessing its performance on the remaining folds. This approach aids in detecting overfitting and ensures the model's effectivity and accuracy.

III. ALGORITHMS

A. *Decision tree*

The decision tree is a widely supervised machine learning method applicable to both classification and regression tasks. Its structure resembles a flowchart, where each internal node represents a feature, each branch signifies a decision rule, and each leaf node denotes the outcome or class label. The primary objective is to construct a productive model for estimating the value of a target based on input features.
The decision tree algorithm learns from the data by iteratively partitioning the feature space into subsets according to the values of the input features. This portioning the target variable in classification or have minimal prediction error in regression. The learning process typically follows a top-down, greedy approach known as recursive binary splitting. Beginning with the entire dataset at the root node, the algorithm selects the best feature to split the dataset at the root node, the algorithm selects the best feature to split the data based on criteria like information gain Gini impurity, or entropy. Subsequently, the data is partitioned into subsets based on the chosen feature, and the process repeats recursively for each subset until a stopping condition is met, such as reaching a subset until a stopping condition is met, such as reaching a maximum tree depth or having a minimum number of samples in each leaf node.
Once constructed, the decision tree serves to predict new or unseen data by traversing the tree from the root node to a leaf node based on the values of the input features. The predicted class label in classification or the predicted value in regression is then associated with the leaf node.
Decision trees offer advantages, including interpretability, as the tree structure is visualized comprehension. They can handle both numerical and categorical features and capture intricate interactions between variables. However, decision trees are susceptible to overfitting, particularly when they become deep and complex. Mitigation strategies such as pruning.

B. *Random forest*

Random Forest is an ensemble learning method algorithm that combines several individual decision trees to construct a more accurate and robust predictive model. This versatile method is employed for both classification and regression tasks.
How does Random Forest work in machine learning?
Dataset: Start with a labelled dataset consisting of input features and corresponding target variables. For classification tasks, the target variable represents class labels, while for regression tasks, it represents numerical values.
Random Sampling: Randomly generate subsets of the original dataset through a process called bootstrapping, where samples are selected with the replacement. These subsets, referred to as bootstrap samples, contribute to the diversity of the individual decision trees within the ensemble. Each bootstrap sample is typically of the same size as the original dataset but may contain duplicate instances.
Tree Construction: Construct an individual decision tree using each bootstrap sample.
At every node of the tree, a random subset of features is considered for splitting. Typically, the square root of the total

number of features is selected in this random feature selection process.

This approach enhances the diversity of trees and mitigates overfitting.

Decision Tree Growth: Expand each decision tree either to its maximum depth or until a defined stopping criterion is met. The tree grows recursively, with data splitting at each node based on a chosen feature and a specific splitting criterion (e.g., information gain, Gini impurity). The expansion continues until the tree reaches the stopping criterion, like achieving a maximum depth or reaching a minimum number of samples in a leaf node.

Ensemble Creation: After constructing all decision trees, predictions are generated by aggregating the individual predictions of each tree. In classification tasks, the final prediction is determined by the class with the majority of votes among the trees. In regression tasks, the final prediction is the average of the predicted values from all trees.

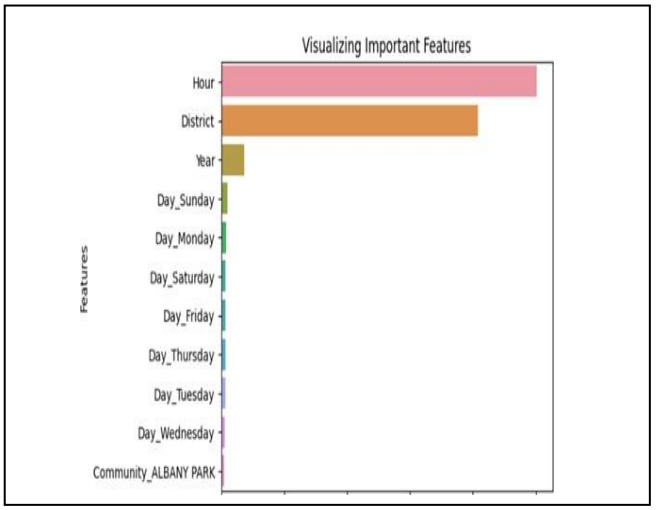
Key Advantages of Random Forest in Machine Learning:

1. Robustness: Random Forest exhibits lower susceptibility to overfitting compared to a single decision tree, owing to the amalgamation of multiple trees and the introduction of randomness during the training process.
2. Generalization: The ensemble nature of Random Forest enhances generalization performance, enabling it to handle a diverse range of input data and capture intricate patterns.
3. Feature Importance: Random Forest offers a measure of feature importance, facilitating the identification of the most influential features in the prediction process.

Random Forests also offer some degree of interpretability, as you can analyze the structure of individual decision trees within the forest.

```
# view the feature scores

feature_scores = pd.Series(clf.feature_importances_, index=X_train_RandomForest.columns).sort_values(ascending=False)

feature_scores

# Creating a seaborn bar plot
import seaborn as sns # statistical data visualization
import matplotlib.pyplot as plt # data visualization

sns.barplot(x=feature_scores, y=feature_scores.index)


# Add labels to the graph

plt.xlabel('Feature Importance Score')

plt.ylabel('Features')


# Add title to the graph

plt.title("Visualizing Important Features")


# Visualize the graph

plt.show()
```

For our project, we use the random forest to know the feature importance.



IV. CURRENT/LATEST R&D WORKS

- Crime Prediction and Analysis by Gayatri Sanjay Kolte, Nupur Rajesh Patel, Prof. Sonali Bodekar[1]

Gayatri Sanjay Kolte, Nupur Rajesh Patel, Prof. Sonali Bodekar developed a system to prevent crimes with the help of machine learning algorithms. The aim is to coach a model for prediction.

The goal of this project is to give a concept of how artificial intelligence specially machine learning will be utilised by enforcement agencies to detect, predict and solve crimes at a way quicker rate and therefore reducing the crime rate. However even for local people in addition to grasp their area, the neighbourhood in terms of crime and security. For this, the data is stored in databases.

- Crime prediction by B.Sivanagaleela, S.Rajesh, B.Sivanagaleela, S.Rajesh**[2]**

B. Sivanagaleela and S. Rajesh conducted a planned crime analysis and prediction using the fuzzy c-means algorithm. They introduced a model that enhances the prediction and analysis of crimes through the refinement of the Fuzzy-c means algorithm. By analyzing crime patterns, they successfully identified and prevented potential criminal activities. Additionally, their study aimed to comprehend the crime patterns to determine the areas where crimes are likely to occur frequently.

- Crime prediction by Sunil Yadav, Meet Timbadia, Ajit Yadav, Rohit Vishwakarma, Nikhilesh Yadav**[3]**

Sunil Yadav, Meet Timbadia, Ajit Yadav, Rohit Vishwakarma, and Nikhilesh Yadav initiated a project focusing on crime pattern detection, analysis, and prediction. Their work explores how societal development might contribute to crime prevention.

The team employed various techniques, including Association mining (Apriori), k-means, Naïve Bayes, correlation, and regression.

These methods were implemented on the dataset using both the rail tool and R tool for comprehensive analysis and prediction in the context of crime prevention.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to all those who have contributed to the successful completion of our research paper.

First and foremost, we would like to thank our research paper supervisor prof. Vijya Tulsani for her invaluable guidance, expertise, and support throughout the entire process. His insightful feedback and constructive criticism have been instrumental in shaping the content and quality of this seminar report.

We would like to extend our appreciation to the staff and faculty members of the Department of Master of Computer, Faculty of Information Technology & Computer Science, Parul University, Vadodara for providing us with the resources and facilities necessary for conducting research and accomplishing it.

Once again, we extend my heartfelt thanks to everyone who has contributed to the completion of this research project.

REFERENCES

[1] Gayatri Sanjay Kolte, Nupur Rajesh Patel, Prof. Sonali Bodekar, Crime Prediction and Analysis, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 09 Issue: 04 | Apr 2022.
https://www.irjet.net/archives/V9/i4/IRJET-V9I4234.pdf

[2] B.Sivanagaleela,S.Rajesh, Crime analysis and prediction using fuzzy c-means algorithm.In 3rd International conference on trends in Electronics and Informatics(ICOEI),April 2019,IEEE

[3] Sunil Yadav,Meet Timbadia ,Ajit yadav, Rohit Viswakarma,Nikhilesh Yadav, Crime pattern detection ,analysis and prediction. In international conference of Electronics, Communication and Aerospace Technology (ICECA), April2020, IEEE.
https://ieeexplore.ieee.org/document/8203676