Second International Conference on Smart Computing and Communication for Sustainable Convergence (ISCCSC 2025)

Crime Pattern Analysis and Prediction Using Machine Learning

Pradeep Nazerath[1], Lathesh Kumar S R[2], Sanketh Patil[3], Shivamani M Nayak[4] and Tejashwini Shailendra Murdeshwar[5]

Department of Artificial Intelligence and Machine Learning, Alvas Institute of Engineering & Technology, Moodbidri, Mangalore, Karnataka, India

Emails: pradeepn@aiet.org.in, latheshkumar06@gmail.com, sanketpatilsp360@gmail.com, shivamaninayak5757@gmail.com, murdeshwartejashwini@gmail.com

Abstract.

Urban crime rate changes that are influenced by tiny spatiotemporal factors are typically outside the realm of a usual police investigation. This paper reflects the authors' hierarchical machine learning workflow to understand crime trends through the" Crime in India" open dataset from 2001 to 2020. Our approach very detailed stages a preprocessing, moves the Synthetic Minority Over-sampling Technique (SMOTE) to solve the class imbalance problem, and has multilayer feature extraction to capture the spatial and temporal aspects so that the classifier's prediction quality can be brought up to standard again. The authors tried different algorithms to have a quantitative comparison of their performances and measured accuracy, precision, recall, and F1-score. The algorithms were K-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, and Naive Bayes. Random Forest was the most accurate classifier as it achieved the predictive accuracy of 89 % and being very stable, it varied almost evenly across all the evaluation metrics.

Keywords. Crime Prediction, Machine Learning, Spatiotemporal Analysis, Data Imbalance, Ensemble Learning, Smart Policing, ST-CrimeNet

1. INTRODUCTION

The fast-paced urbanization, higher population density, and disparities in income and living standards have a direct impact on the occurrence and nature of crimes in today's world. Over the years, the increasing size of cities has made it even harder for the police to come up with and implement the right strategies for detecting and fighting crime, given that they are still using the old-fashioned techniques for analyzing criminal activities. Furthermore, traditional methods mostly depend on laborious investigations and the preparation of historical reports which are not always able to reflect the intricate geographical and temporal interrelationships that large crime datasets contain. Recently, the accessibility of extensive digital crime records and the progress in machine learning have together opened up new vistas for crime analysis based on data. The methods of machine learning make it possible to automate the uncovering of hidden patterns, predict trends, and classify crimes with an even higher level of accuracy. These models, by making use of past crime data, help in examining the variations of crime across the cities, different time periods, and various environmental conditions; thus, they are working in the support of proactive police strategies. Nevertheless, there are drawbacks associated with the deployment of machine learning for crime forecasting. The crime datasets are usually very much unbalanced, wherein, the occurrences of some crime types are much more than the others. This skewness very often results in the formation of biased models that yield poor results for rare but significant crimes.

2. LITERATURE REVIEW

[1] Chawla et al. put forward the Synthetic Minority Over-sampling Technique (SMOTE) as a solution to the problem of class imbalance in datasets. Their research revealed that the performance of classifiers is enhanced by the minority class oversampling especially in the highly imbalanced datasets like crime records which are notoriously difficult to classify.

[2] Moher et al. put forward spatiotemporal point process models for crime analysis and pointed out the significance of place and time in forecasting the areas with most criminal activities. Their method was efficient for spotting hotspots, but it was not able to deal with the interactions between complex features flexibly.

[3] Breiman developed the Random Forest algorithm and has since been applied in crime prediction mainly due to its strength and non-susceptibility to overfitting. Random Forest was reported in various studies to be able to improve the accuracy and stability of crime datasets.

[4] Wang et al. used geographic clustering methods to introduce spatial characteristics into the crime prediction models. Their findings allowed for the most accurate selection of regions with a high incidence of crime, but they still did not dive deep into the issue of temporal dependencies.

[5] Huang et al. made use of spatiotemporal modeling techniques to study the variation of crimes and highlighted the significance of patterns that depend on time such as daily and yearly variations. However, their method required a lot of computer power.

[6] Liu et al. experimented with deep learning models for predicting crimes and proved better performance in pattern recognition. But even with higher accuracy, models were not very interpretable and required more computations, resulting in increased computational cost.

[7] Rudin underlined the necessity of interpretable models in situations involving decision-making with serious repercussions like police work. The research pointed out that besides being transparent, predictive policing systems should also be understandable user-friendly for their acceptance.

3. METHODOLOGY

The four stages are structured analytically in a highly sequential and integrated manner: it consists of data preparation, feature engineering, model implementation, and evaluation. The stage

3.1 Dataset Collection and Pre-Processing

The research utilizes the open-source dataset "Crime in India", which is publicly accessible on Kaggle and presents a comprehensive account of crimes across the country from the year 2001 to 2020. The data file consists of approximately 1.2 million rows that cover the different crime categories, time and place identifiers, and some demographic variables. The initial preprocessing and scaling the values of the numerical attributes to a range of 0 to 1 using the Min-Max method. Outliers were removed using the Interquartile Range (IQR) method which reduced their influence on the learning phase. A drastic class imbalance where common crimes were several times more than rare ones was the main drawback of the dataset. The Synthetic Minority Oversampling Technique (SMOTE) was used as a solution to this drawback by generating synthetic minority samples and thus promoting classifier generalization. The data set that was processed and distributed equally was considered the starting point for modeling. The focus of feature engineering in this research was to obtain spatial and temporal attributes that would be richer and more informative.

Records that have location coordinates were mapped to localities by K-Means clustering, demonstrating the geographical regions of crime hotspots according to the volume of crime in a given location. The newly created variables provided not just additional information but also enhanced the models' capability to reveal concealed spatiotemporal patterns.

3.2.1 Machine Learning Algorithms:

To identify the top algorithm for multiclass crime classification, five different supervised learning algorithms were tested, and the K-Nearest Neighbors algorithm was applied with the Euclidean distance metric to identify local similarity patterns.

$$d(x, x_i) = \sqrt{\sum_{j-1}^{n} (x_j - x_{ij})^2}$$

KNN, even though it's an effective approach for neighborhood-based reasoning, suffers from the problems of noise and high dimensionality. In accordance with the techniques mentioned in the preceding comparative literature, Support Vector Machines with an RBF kernel were employed to form non-linear class boundaries in the crime data.

$$\hat{y} = \text{mode}\{h_1(x),\, h_2(x),\, \ldots,\, h_T(x)\}$$

The model was able to get good results for categorical variables.

3.2.2 Experimental Setup and Evaluation

All of the experiments were carried out in Python 3.11 using Jupyter Notebook along with the usual standard data science libraries like Scikit-learn, pandas, NumPy, matplotlib, seaborn, plotly, and Folium. The research data was split into two parts, 70% for training and 30% for testing, and 5-fold cross-validation was applied to ensure that the performance estimation was unbiased. Hyperparameter tuning was performed using GridSearchCV.

4. CONCLUSION

The mentioned paper illustrates one of the machine learning techniques which through analyzing and predicting crime patterns by handling a huge volume of historical crime data. The research paper has presented the comparative evaluation as a way to demonstrate the application of different machine learning methods like K-Nearest Neighbor, Support Vector Machine, Decision Tree, Random Forest, and Naive Bayes not just for crime classification but also for getting considerable spatiotemporal patterns discovered. The most important aspect of the methodological framework was data preprocessing, the class imbalance problem which was tackled through SMOTE, and the deliberate feature engineering in order to obtain the most relevant temporal and spatial features.

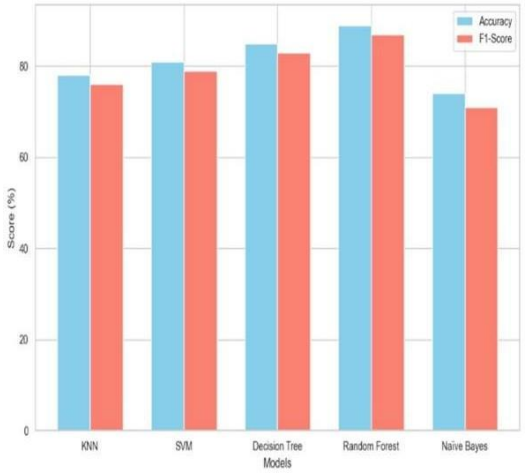

Fig. 1: Model Comparison: Accuracy vs F1-Score

Research on model performance is only a minor segment of the article, which constitutes the larger argument for data-driven crime analytics over traditional evaluation methods. The traditional approaches rely more on proofs done part by part and story-telling insights whereas machine learning algorithms give extensive and data-centric evaluations. This kind of evaluation can be used to spot the specific areas resources are allocated to and those that need to be improved, and at the same time, to make sure that the crime prevention policies created have the

right focus. The real-time data integration, the use of advanced modeling techniques, and the improvement of dataset quality won't just improve these systems, but they will also support the development of more responsive and community-oriented public safety initiatives which will be less reliant on technology.

5. REFERENCE

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[2] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[3] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita, "Self-exciting point process modeling of crime," Journal of the American Statistical Association, vol. 106, no. 493, pp. 100–108, 2011.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825– 2830, 2011.

[5] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009.

[6] L. Wang, X. Zhang, and Y. Liu, "Predicting crime using spatial features," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2018.

[7] Q. Huang, Y. Guo, and Z. Li, "Spatio-temporal crime prediction using GIS and deep learning," ISPRS International Journal of Geo-Information, vol. 10, no. 9, p. 602, 2021.

[8] C. Rudin, "Why we should use interpretable models instead of black box models for high-stakes decisions," Nature Machine Intelligence, vol. 1, pp. 206–215, 2019.

BIOGRAPHY:

Name: DR. PRADEEP NAZARETH

Designation: Associate Professor (Project Guide)

Email ID: pradeepn@aiet.org.in

Mobile Number: +91 9164525591

Dr. Pradeep Nazareth is an Associate Professor in the Department of Artificial Intelligence and Machine Learning at Alva's Institute of Engineering and Technology (AIET), affiliated with VTU, Belagavi. He holds a B.E. from KVG College of Engineering, an MTech from SJCE Mysore, and a Ph.D. from NITK Surathkal.

Name: LATHESH KUMAR S R

USN: 4AL23AI400

Email ID: latheshkumar06@gmail.com

Mobile Number: +91 7760814609

Areas of Interest: Lathesh Kumar S R is currently pursuing a Bachelor's degree in Artificial Intelligence and Machine Learning at Alva's Institute of Engineering and Technology (AIET), affiliated with Visvesvaraya

Technological University (VTU) in Karnataka. His academic interests span across Generative AI, Web Technologies, Natural Language Processing (NLP), Large Language Models (LLMs), Cloud Computing, Deep Learning, Machine Learning, and Robotics.

Name: SANKET PATIL

USN: 4AL22AI043

Email ID: sanketpatilsp360@gmail.com

Mobile Number: +91 8660966350

Areas of Interest: Sanket Patil is currently pursuing a Bachelor's degree in Artificial Intelligence and Machine Learning at Alva's Institute of Engineering and Technology (AIET), affiliated with Visvesvaraya Technological University (VTU) in Karnataka. His scholarly focus encompasses key areas such as Data Science, Machine Learning, and Artificial Intelligence. I am particularly interested in Generative AI and Cloud Computing Technology.

Name: SHIVAMANI M NAYAK
USN: 4AL22AI051

Email ID: shivamaninayak5757@gmail.com

Mobile Number: +91 98050666408

Areas of Interest: Shivamani M Nayak is currently pursuing a Bachelor's degree in Artificial Intelligence and Machine Learning at Alva's Institute of Engineering and Technology (AIET), affiliated with Visvesvaraya Technological University (VTU) in Karnataka. His scholarly focus encompasses key areas such as Data Science, Machine Learning, and Artificial Intelligence. I am particularly interested in Generative AI and Cloud Computing, and I aspire to explore how these technologies can be integrated to build intelligent, scalable solutions.

Name: TEJASHWINI SHAILENDHRA MURDESHWAR

USN: 4AL22AI060

Email ID: murdeshwartejashwini@gmail.com

Mobile Number: +91 9353999908

Areas of Interest: Tejashwini Shailendhra Murdeshwar is currently pursuing bachelor's degree in Artificial Intelligence and Machine Learning at Alva's Institute of Engineering and Technology (AIET), affiliated with Visvesvaraya Technological University (VTU) in Karnataka. Her academic interests include Generative AI, AI Ethics, Deep Learning, Machine Learning and Data Science, Project Management.