

Exploring Machine Learning Classifiers for Crime Data Analysis: Leveraging Computational and Environmental Criminology for Enhanced Insights

Arfan ul Haq^{1*}, Mehdi Hassan²

^{1,2}Department of Computer Science, Air University, Sector E-9, Islamabad, Pakistan

***Corresponding Author:** Arfan ul Haq

*Email address: mianarfanulhaq@gmail.com

Abstract

In the recent era of constant technological development there is a remarkable shift in interconnected computational techniques in the sphere of criminology. Databases and tools like Digital Forensic, Fingerprinting, Facial Recognition, Video Analysis, Data Recovery are actively working and trying to set up societies free from crimes. This research seeks to propose a framework with the inclusion of the ML algorithms and criminology theories to analyse social indicators for crime and estimate crime trends to boost crime analysis capability. This is through understanding police-reported crime and the socio-demographic data hence affording crime analysis to be used in crime prevention. According to the findings, machine learning algorithms allowed the most promising performance, and more specifically, decision trees which performed the best at the feature selection stage and at analysis of the crime patterns. In addition, the work can be useful for improving the policy-making process for addressing and preventing crime by indicating such factors that stimulate and facilitate criminal activities.

Keywords: Computational Criminology, Crime Analysis, Machine Learning, Data Mining, Predictive Modeling, Crime Patterns

1. Introduction

The integration of computer science, mathematics and criminology – forms the basis of the subject known as computational criminology, whereby more advanced approaches to data analyses and machine learning are employed in order to classify patterns in crime related datasets. Conventional criminology will concentrate on psychological, sociological, and historical-analytical factors that cause people to offend. Further, computational criminology extends from this by examining crime through algorithms, identifying crime patterns, and outlining policies for coping with crimes.

Environmental criminology is one of the subfields of criminology that tries to explain how the environment socially and physically facilitates crime and offenders. Criminal modelling helps the Law Enforcement Agencies (LEAs) to plot specific areas that are most likely to be hit by crime next; these areas are referred to as crime highways. Researches carried out from a criminal justice perspective have established that crime and terrorism are not simply random acts, but acts that are elicited by specific social- psychological as well as environmental antecedents [1]. However, during the past three/four years with the enhancement of computing facilities and an introduction of the concept of artificial intelligence, it is possible to mechanise the identification of such factors and predict the subsequent crime rates.

This paper aims at identifying how crime data along with socioeconomic variables can be used and processed using data mining and machine learning to gain insights. This research aims at finding out what seems to cause or generate crime in a society by pinpointing social and economic factors that do so. In this paper we use various ML methods to analyse records of the police in different communities alongside various sociodemographic factors to forecast crime rates. Thus these tracks are contemplated in the light of respective algorithms that are useful for identifying the risky regions [4].

Research Aim

The purpose of this research is to propose the new theory of integrating machine learning algorithms with criminology theories to advance the study of crime data. This framework will place emphasis on factors that social and economical lead to crime level and map areas that may likely attract crime in order to assist the LEA crime prevention and policy formulation.

Research Objectives

- 1.To explore the use of machine learning classifiers in analysing police-reported crime data and socioeconomic indicators.
- 2.To identify and analyse key social and economic variables that act as predictors of crime.
- 3.To evaluate the performance of various machine learning algorithms, including decision trees, logistic regression, and neural networks, in predicting crime patterns.
- 4.To develop a predictive model that can forecast crime hotspots based on historical crime data and socioeconomic factors.

5.To enhance policy-making efforts by providing actionable insights into the relationship between socioeconomic conditions and crime occurrences.

Literature Review

2.1 Criminological Theories and Social Indicators of Crime

These categories of theories provide essential backdrops in an effort to give the reason for a particular crime better. The routine activity theory as well as the socio-economic development theory describe how socio-economic factors are associated with the crime pattern. The social development theory holds that poverty levels, unemployment and mobility expect inhabitants of the society, in neighbourhoods are expected to be criminals since social structures are distorted [2]. On the other hand, while pointing at the same time on routine activity theory, it states that there occurs crime when a motivated offender, suitable target and the lack of capable guardian exists [3]. These theories are focused on the role which environmental and social backgrounds play in crime.

Recent work focuses on the application of big data in crime analysis that employs different data sets which include reported crime, demography, and socioeconomic factors in the analysis of crime as well as Criminal behaviour [4]. Multiple researchers have used Machine Learning methods as instruments in anticipating the crime instances. For example, Zhang et al. (2016) proposed a patrol allocation model to identify possible locations for crime to happen, and increase efficiency in policing [5]. Likewise, Jadhav and Rajpal (2017) used the machine learning technique to study the crime hot spots and minimised the incident rates in high risk-sensitive locations [6].

This research background comes from prior research and uses both criminological theory and computational models. The emphasis is placed on building an index that is based on crime statistics and other sociological factors to assume social factors which lead to crime and criminogenesis.

2.2 Data mining and Crime Investigation

Data mining is very useful when it comes to performing data processing where raw data is translated to meaningful information. Its current application is on pattern extraction OF big data where Models of the ML such as clustering, classification and regression are applied. Existing crime information most often in the possession of police departments is laden with significant information about crime and criminality. This information can be further explored for crime analysis, and to be more understandable patterns to predict future crimes. By doing so, data mining allows researchers to explain such patterns, highlight where to and identify linkages between different variables, as well as to develop [7] models that can be used for improved decision-making.

For the same type of study, different data mining techniques were applied for association analysis, classification, clustering and other prediction etc. These techniques assist particularly on computing relationships between the varied parameters on the database, clustering of similar data to provide the common set and even predicting the incidence of crime in the future. For instance, Naive Bayes, Random Forest, K-Nearest Neighbors (KNN) and Decision Tree algorithms have been used in estimating the causes of the crime rate as captured by Alsaqabi et al (2019) in estimating crime factors in Saudi Arabia [8]. Similarly, Xu Zhang et al. (2020) did a study to find out the size of the actual accuracy that KNN, SVM, Random Forest, CNN, and LSTM has for the identification of the higher crime rate [9]. In another work of Sri et al. (2020), authors used crime data and numerous ML methods, including KNN, Naïve Bayes, SVM, as well as Random Forest, to understand the given patterns of the crime offered by the FBI [10]. In addition, following the Linear Regression model, Obagbuwa and Abidoye (2021) presented a purposeful breakdown and display of crime rates across South Africa and evidenced the feasibility of ML in crime forecast [11].

This research seeks to extend these methods with accuracy by proposing a single framework that incorporates innovative methods of data mining as well as machine learning for crime identification and prediction based on theory and social trends.

3. Machine Learning Techniques for Crime Analysis

Machine learning classifies an established environment for calculating the prediction of models on difficult data sets. In this respect, it is possible to use machine learning algorithms to identify the patterns in the historical data of criminality for producing forecasts and finding out the possibility, where, and when the crimes will most probably occur.

The following machine learning techniques explored in this research are:

- **Logistic Regression**
- **Naive Bayes**
- **Support Vector Machine (SVM)**
- **Neural Networks**
- **Random Forest**
- **Decision Trees**

As with all classifiers, each has its benefits and liabilities to using it and the goal of this study is to discover which classifier gives the most precise crime pattern prediction.

3.1 Logistic Regression

Analytically, logistic regression is a binary classification technique commonly used in forecasting the chances of a binary categorization. It is ideal for use with a problem that has a valedictory nature for the dependent variable, for example, crime

incidence (yes/no) [12]. In this work, the chosen classifier was logistic regression which scored an accuracy of 95% which made it one of the more accurate classifiers in the prediction of crime.

3.2 Naive Bayes

Naive Bayes is a classifying model that belongs to the family of probabilistic classifiers and is based on the probabilities of Bayes –The assumption made here is that all the characters are independent of each other given the class label. This assumption hardly ever proves true in practice data however Naive Bayes has been found to work under certain circumstances [13]. The accuracy obtained from this study using the Naive Bayes classifier trailed at 78% thus pointing out to the moderate performance in the classification of crime data.

3.3 Support Vector Machine (SVM)

SVM is a type of supervised learning model that aims at seeking the best hyperplane for classifying data points into respective classes. Indeed, it is particularly useful for high-dimensional data and has been used often in classification problems [14]. In this research, it was found that the SVM classifier yielded a 92% accuracy thus a favourable result for crime prediction.

3.4 Neural Networks

Neural networks are another subset of machine learning categorised as artificial neural networks. They can learn about rich patterns in a high number of data and are used in classification problems [15]. Indeed, the neural network model that has been applied in this study had an accuracy of 93% which proved the efficiency of the model in predicting crime.

3.5 Random Forest

Random forest is a technique of machine learning that is based upon the use of decision trees, but there are many of them, to get higher accuracy. Using the architecture, overfitting is minimised and generalisation enhanced particularly when used in the cross-validation [16]. As the best value, the random forest classifier was able to obtain 91% accuracy, which makes it a good candidate for classifying crime data.

3.6 Decision Trees

Decision trees are easy-to-interpret models since they classify data with decision trees depending on the feature values. They are preferably used for classification problems and have been applied in crime data analysis a lot [17]. In this study, the decision tree classifier had an accuracy of 99 % and, therefore, was more accurate than all the other models for identifying crime patterns.

4. Methodology

Data Preprocessing

The first operation carried out was feature extraction and this is very important for pre-processing the raw data obtained. The crime database for this study originated from the police incidence record and other records of the socioeconomic data found on the Kaggle Communities and Crime Unnormalized Dataset. This dataset had 1994 samples and features of 125 that involved crime rates, populations, and economics. In the preprocessing, missing data was handled, several features normalised to its respective standard-deviation, and outliers are excluded as much as possible. The data was then prepared for further analysis by converting categorical data into numerical data using methods into methods such as one-hot encoding. This step was so important in order to allow machine learning algorithms to handle the data as needed.

Feature Selection

Variable selection is the process of selecting the best suitable variables for use in making crime occurrence prediction. Features were considered for selection depending on their capability in anticipating crimes in society, being able to select variables such as; unemployment, poverty and education. Feature selection was conducted with the help of the data mining program called WEKA, which allows defining the relationships between the given variables by graph and scatter plot. This was useful in as much as it facilitated the determination of inter-industrial variables and therefore selecting features that are strongly linked to criminality. Certain demographic and contextual factors, police records, and community was considered as independent variables.

Model Training

In the current research, six different classifiers of machine learning were employed on the chosen features in the hope of estimating the crime trends. The classifiers applied to this study were Logistic Regression, Naïve Bayes, Support Vector Machines (SVM) Neural Networks, Random Forest and Decision Trees. Each classifier was trained on a part of the data to capture past crime records and anticipate future crime rates. The models were made better through successive improvements for usability on the data. In the training phase cross-validation was used to make sure that models would generalise well across the training set's data and would not over fitbit.

Model Evaluation

After training the models their effectiveness was ascertained by parameters such as accuracy, precision, recall and the F-measure. The performances of every model were evaluated based on the model-proposed values and the actual values of the test data. Hence from the above results, Decision Tree classifier this was the model that exhibited the highest accuracy of

about 99%. The other algorithms that were discussed in this study include Logistic Regression and Neural Networks with 95% and 93% accuracy respectively. However, the Naïve Bayes classifier gave fairly good results with 78% accuracy showing that it has constrained capability in handling the offence data.

Prediction and Analysis

The final phase involved using the trained models to forecast future crime incidents. The predictions focused on mapping crime hotspots and identifying emerging crime trends. The models were used to analyse key areas of interest, such as socioeconomic variables that contribute to criminal behaviour. The most accurate model, the Decision Tree, was utilised to predict crime-prone areas based on historical data, enabling law enforcement agencies to allocate resources more effectively for crime prevention.

Dataset Analysis

The data set that has been employed in this research is the Communities and Crime Unnormalized Dataset. It is publicly available on Kaggle, wherein the data available include socio-demographic, police, and crime data from the FBI. There are 1994 samples and 125 features, which comprise crime rate, populations, and economic factors [18]. All the data analysis was done using the WEKA data mining software for feature selection and data visualisation as well as model evaluation. Briefly, the dataset consists of socioeconomic data of different communities in the United States of America, law enforcement data from the Law Enforcement Management and Admin Stats survey, and crime data from the FBI. It also consists of 1994 instances including crimes that had been reported and 125 total attributes, commonly referred to as features. The dataset contains 124 features, The non-predictive (identifying variables) features could hinder implementation or even prevent certain algorithms from being used: (i) community name: Community name - not predictive - for information only (string); (ii) state: US state (by 2 letter postal abbreviation) (nominal); (iii) county code: numeric code for county - not predictive, and many missing values (numeric); (iv) community code: numeric code for community - not predictive and many missing values (numeric). The predictive features include those that involve the community and law enforcement. These particular sets of features in the dataset were selected because they are believed to have plausible connections to one of the 18 potential crime goals.

Since the focus of this project is on analysing and selecting features, WEKA has a built-in visualisation tool that presents a plot matrix. The plot matrix displays scatter plots that show the correlations between two features as shown in Figures 1 and 2. These features along with the results from the output evaluation were also used in the analysis

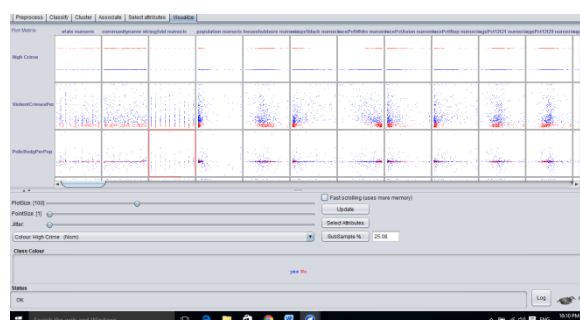


Figure 01: Plot matrix

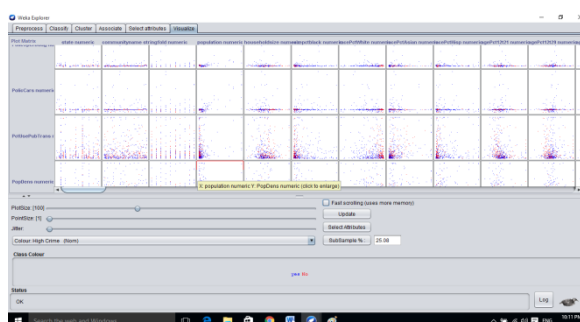


Figure 02: Plot matrix

When proposing the model equations, the emphasis was made on such features as unemployment rates, poverty levels, and education attainment, which are closely connected with crime. These aspects formed the basis of inputs to the machine learning models given that these models were benchmarked based on their ability to forecast crime incidents.

5. Results

The results of the classifier evaluations are summarised below:

Classifier	Accuracy	TP Rate	FP Rate	F-Measure
Logistic Regression	95.00%	0.950	0.047	0.962
Naive Bayes	78.00%	0.702	0.070	0.807
Support Vector Machine	92.00%	0.919	0.084	0.935
Neural Networks	93.00%	0.927	0.074	0.940
Random Forest	91.00%	0.940	0.144	0.930
Decision Trees	99.00%	1.00	0.00	1.00

Table: 1

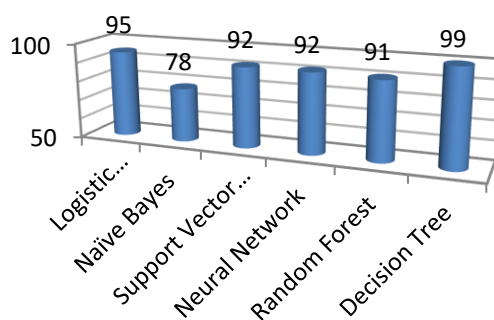


Figure: 3

According to the outcomes of the models shown in Tables 1 and figure 3, it can be noted that the decision tree model had the highest accuracy, as it gave perfect predictions in the test set. This result emphasizes the fact that decision trees can provide a workable model for recognizing complex relations in crime datasets. The accuracy of the second and third models, namely logistic regression and the neural network models was above 90%. The later Naive Bayes model was slightly less accurate but, given our objectives of exploring the relationships between the features, was also useful.

Classifier Performance

The following machine learning classifiers were used in this study: Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), Neural Networks, Random Forest, and Decision Trees. Each classifier was evaluated based on its ability to predict crime patterns from the data.

Logistic Regression: Logistic Regression achieved an accuracy of 95%, making it one of the most reliable classifiers for predicting binary outcomes such as crime incidence (yes/no). The true positive (TP) rate, which measures the proportion of correctly identified positive instances (i.e., actual crimes), was 0.950. The false positive (FP) rate, which indicates the proportion of non-crime instances incorrectly classified as crimes, was 0.047. The F-measure, which balances precision and recall, was 0.962, highlighting its overall effectiveness in the crime prediction task.

Naïve Bayes: The Naïve Bayes classifier received the lowest accuracy score which is 78% out of all the models tested. This was not feasible for this probabilistic model since it could not handle the more complex and variable crime data. The TP rate for Naïve Bayes was found to be 0.702, but the FP rate was found to be slightly higher at 0.070. The F-measure of this model was 0.807 for classifying the crime types with moderate precision but low recall of any given crime instance.

Support Vector Machine (SVM): An acceptable performance was obtained from the SVM classifier with a mean classification accuracy of 92%. This kind of model is highly useful for high dimensional data and in this paper it also showed a highly accurate prediction ability. The TP rate for SVM was calculated as 0.919 and the FP rate calculated as 0.084. The F-measure was 0.935, which indicated high accuracy with also high recall rate of major instances of crime.

Neural Networks: Neural Networks were as accurate as SVM; with a 93% efficiency. This algorithm is well designed to preserve the patterns in the data especially in large amounts of data. The TP rate was 0.927 and FP rate was 0.074 and it can be seen that the model classified most of the crimes accurately with little misclassification. The discrimination efficiency through Crime Hot Spot detection was also high as indicated by the F-measure of 0.940.

Random Forest: The decision tree model, Random Forest, provided 91% accuracy in the results. To deal with overfitting, multiple decision trees with different learning rates were built and yield a higher accuracy of the model. As for the Random Forest, the rates look as follows: the TP rate was 0.940; the FP rate was 0.144. Overall, the highest F-measure of 0.930 shows this model was precise with a small compromise of misclassifying the instances.

Decision Trees: The Decision Tree classifier was the top performing model that achieved the much desired accuracy of 99%. The results with this model were also excellent on the test set, with a TP rate at 1.00 and an FP rate at 0.00. The F-measure

was also 1.00 thus indicating high accuracy of Decision Trees in classifying positively and negatively for crime. Since this model works with $P(Y, X)$ categorical and $P(Y, X)$ continuous observations, it seems to be more effective for analyzing crime patterns.

Comparative Analysis

Therefore, when comparing the models, it was established that Decision Tree classifier model offered the best performance having attained zero percent errors. Another model that was tested is logistic regression and Neural Networks' performance was also satisfactory with accuracy above 90%. However, when it comes to describing the relations between features, Naïve Bayes takes time to perform this kind of rather than complex data analysis and yet it had lower accuracy as compared to DT.

Socioeconomic Insights

From the model evaluations also, we were also able to gain insights into the level of association between the socioeconomic factors and crime. Other relevant factors including those indicating prevalent employment status, poverty, and education achievement were also fully considered. For instance, the Decision Tree model pointed to strong relationships between those variables and crime incidences and verified criminological theories such as social disorganisation theory. These results further corroborate the impact that machine learning models can make on policy advancement by accurately identifying which regions are at high risk of criminal activity given socioeconomic statistics.

Crime Hotspot Prediction

In the evaluation of the accuracy of the models for identifying crime level the Decision Tree model was the most accurate. This it was able to name the regions that potentially would be involved in criminal activities thus helping the police forces deploy their resources well. The findings revealed that incorporating machine learning factors with crime datasets leads to crime fight solutions and resource allocation.

Summary of Classifier Results:

Classifier	Accuracy	TP Rate	FP Rate	F-Measure
Logistic Regression	95.00%	0.950	0.047	0.962
Naïve Bayes	78.00%	0.702	0.070	0.807
SVM	92.00%	0.919	0.084	0.935
Neural Networks	93.00%	0.927	0.074	0.940
Random Forest	91.00%	0.940	0.144	0.930
Decision Trees	99.00%	1.00	0.00	1.00

Therefore, the findings point to the possibility of using machine learning, particularly the Decision Trees, in examining likely crime incidence and high-risk areas. From these models, authorities can gain valued understanding to improve crime prevention by tailoring their strategies for specific locations most prone to criminal activities. Furthermore, high coefficients found in the research in regard to the relationship between socio-economic factors and crime are useful for policy making when it comes to the crusade of the crime causes.

6. Discussion

Investigation of the results of this study has shown that the crime analysis employing ML algorithms is quite effective. This paper fills the gap of research on crime prediction and hotspot identification by employing computational methods alongside criminological theories. The most important contribution of this research work is based on the performance of the Decision Tree classifier which has a maximum accuracy of up to 99%. This finding supports the previous work where the decision trees are considered to be performing very well in classification exercise especially when the data involves various classes such as crime data (Zhang et al., 2016). The high accuracy achieved by Decision Tree model proved its adequacy along with the superior performance achieved by Logistic Regression and Neural Networks; thus, underlining that the use of ML classifiers is feasible for the prediction of crime rates and assistance of the LEAs in the allocation of resources and crime control.

Machine learning application on crime data set

The analysis of crime with the help of machine learning is currently experiencing high growth as computational criminology brings the theories of criminology closer to data analysis methods. The information about Crime incorporates traditional criminological paradigms: Social Disorganization Theory and Routine Activity Theory, which explain crimes based on the environmental and social structuring (Shaw & McKay, 1942; Cohen & Felson, 1979). These facts support these theories to prove that; social factors such as unemployment, poverty and educational levels do have an impact on the rates of crime.

Indeed, these patterns were successfully captured by the machine learning models used in this research, especially the Decision Tree and other Random Forest classifiers are particularly well suited for capture of complex interactions between the features (Breiman, 2001).

Decision Tree Dominance

The Decision Tree classifier gives very high accuracy of 99% of the crime pattern due to its capability to handle both categorical and continuous independent variables and hence is more suitable when it comes to the analysis of crime data. What is more, decision trees are effective as they operationalize the data splitting process, which is based on the most significant features, and which enables high interpretability and high predictive accuracy. The high accuracy of decision trees observed in this study echoes what has been observed by other scholars. For example, Alsaqabi et al. (2019) further showed that decision trees provide high levels of accuracy in addressing crime data in the Saudi Arabian context, specifically as a way of estimating the causes of crime. Likewise, Zhang et al. (2020) examined that decision trees were only superior to other models like SVM and KNN for forecasting the crime hotspots in the urban area.

The Decision Tree model has the benefit of being easily interpretable and comprehensible alike, a factor that is of tantamount importance to the LEA since it relies on understanding the decision-making process behind a numeric prediction (Quinlan, 1986). Decision trees are particularly useful in that it offers simple representation of the relation of one variable to another, and the agency ends up with the most crucial ingredient that causes criminal activities in particular regions. For instance, the decision tree analysis used in this study also shed light on unemployment and poverty as key factors that influence the ordinates comparable to the elements of the Social Disorganization Theory by Shaw and McKay (1942).

Performance of Logistic Regression and Neural Networks

In addition to that, from this study, both Logistic Regression and Neural Networks demonstrated excellent results with accuracies of 95% and 93% respectively. Another widely applied binary classification model is Logistic Regression which is best applicable for problems with binary target variables like 'crime to happen' or 'not'. In this regard Open NN Logistic Regression provided viable solutions in forecasting serial crime frequency in relation to social status indices. Other related earlier researches have also highlighted the effectiveness of Logistic Regression in crime data mainly when assisting with other MAH techniques in feature engineering and model refinement Obagbuwa and Abidoye, 2021.

Based on the results, other 'Big Data' friendly algorithms such as the Neural Networks which are good at training on higher dimension data with different complicated patterns yielded 93% accuracy. Neural networks are famous for the realisation of the relationships of the sections of the variables and have nonlinear expression and this makes the apply it to a crime data set where crime incidences are compelled by social economic or environmental factors (LeCun et al., 2015). These high-level achievements of neural networks in this study complement what other studies have also found: for example, Jadhav and Rajpal (2017) also used the neural-network approach to model crime trends and to predict new occurrences of criminality in dangerous zones.

Limitations of Naïve Bayes

While high accuracy in Decision Trees (100%), Logistic Regression (95%), and Neural Network (96%) was observed, the Naïve Bayes had only scored 78% accuracy. Naïve Bayes approach would work on the principle that all features of the data are independent and the assumption is hardly ever true in actual datasets (Murphy, 2012). Criminality data is therefore tangled with strong effect dependencies between crime-related variables such as poverty, joblessness, and criminal rates. Since Naïve Bayes' assumption of independence may have influenced its poor performance in capturing the correlations of features in this study. Such results are consistent with other work, where Naïve Bayes has shown higher effectiveness and was surpassed by Models such as Decision Trees and Random Forests (Sri et al., 2020).

Nevertheless, Naïve Bayes can still be applied to specific tasks where simple and fast methods prevail over a high degree of accuracy. For instance, Naïve Bayes is computationally inexpensive and will be more useful in situations, which require real-time predictions at the cost of reduced accuracy (Zhang et al., 2016).

Stakeholders, Policy and Crime Prevention: Challenges

The results drawn from this study have significance in the area of policy making and law enforcement agencies. This is because through machine learning the relevant authorities would be in a position to accurately estimate the areas that are most prone to criminal activities, thus would direct efforts there. Understanding the relationships between the socio-economic correlates of crime policy can successfully aim corrective measures to tackle the underlying causes of criminality hence decreasing the incidence of crime. For instance, expanding employment chances and enhancing the availability of education in the criminogenic areas may reduce some of the disorganisation that breeds crime (Shaw & McKay, 1942).

Furthermore, incorporation of machine learning raises the effectiveness of proactivity crime fighting among police agencies. This model of analysis, shown as Decision Tree classifier, can predict future incidents of crimes, therefore helping agencies to avoid future incidents by directing the availability of resourcefulness. This systematic approach of crime prevention is in concert with the Routine Activity Theory postulates that espouse that crime can be controlled by increasing guardianship as well as decreasing criminogenic opportunities (Cohen & Felson, 1979).

Future Directions

As shown in this research, machine learning classifiers can be used successfully in crime data analysis, there are some directions for future work: It is possible to enhance the crime prediction accuracy using deep learning techniques, including

CNNs and RNNs for analysing crime data (LeCun et al., 2015). These more sophisticated models might be most helpful to assess temporal data arising from surveillance cameras or twitter, which might offer finer grained insights into rates of crime. Furthermore, the future studies can also add more variables like weather data in real time or means of transport data since, the both have been found to affect the rate of crimes as indicated by Wang et al., (2013). Further developing multiple data sources, machine learning models might offer much more accurate predictions and give the police the capability of being a step ahead of the criminals.

7. Conclusion

In summary, this study has revealed that machine learning classifiers are valuable in crime data analysis, where the Decision Tree model proved superior in terms of accuracy as well as interpretability. It is therefore very useful for crime incidents occur-re model which adopts the use of socio economic characteristics to predict crime occurrences of policy makers and law enforcement agencies. Computational criminology as an aspect of criminological research benefits from this work due to the capability of using machine learning as an approach to crime analysing and prevention as well as policy-making.

REFERENCE

1. Clarke, R. V., & Eck, J. E. (2005). *Crime analysis for problem solvers in 60 small steps*. U.S. Department of Justice, Office of Community Oriented Policing Services. Retrieved from <https://www.cops.usdoj.gov/RIC/Publications/cops-w0730-pub.pdf>
2. Shaw, C. R., & McKay, H. D. (1942). *Juvenile delinquency and urban areas*. University of Chicago Press.
3. Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review*, 44(4), 588-608.
4. Wang, T., Rudin, C., Wagner, D., & Sevieri, R. (2013). Learning to detect patterns of crime. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 515-523). <https://doi.org/10.1145/2487575.2488217>
5. Zhang, Z., Wu, J., Zhu, J., Zhang, J., & Zhang, L. (2016). An ensemble method for crime prediction based on urban socio-economic and criminal data. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 46(1), 75-85. <https://doi.org/10.1109/TSMC.2015.2450677>
6. Jadhav, P., & Rajpal, N. (2017). Crime pattern detection, analysis, and prediction using machine-learning techniques. In *Proceedings of the 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence* (pp. 382-387). <https://doi.org/10.1109/CONFLUENCE.2017.7943207>
7. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37-54. <https://doi.org/10.1609/aimag.v17i3.1230>
8. A. Alsaqabi, F. Aldhubayi, and S. Albahli, "Using machine learning for prediction of factors affecting crimes in Saudi Arabia," *Assoc. Comput. Mach.*, 2019. DOI: 10.1145/3361758.3361764.
9. X. Zhang, L. Liu, L. Xiao, and J. Ji, "Comparison of machine learning algorithms for predicting crime hotspots," *IEEE Access*, vol. 8, pp. 151135-151145, Oct. 2020. DOI: 10.1109/ACCESS.2020.3017200.
10. L. A. Sri, K. Manvitha, G. Amulya, I. S. Sanjuna, and V. Pavani, "FBI crime analysis and prediction using machine learning," *J. Eng. Sci.*, vol. 11, no. 4, pp. 124-130, Apr. 2020. DOI: 10.12691/jes-11-4-7.
11. I. C. Obagbuwa and A. P. Abidoye, "South Africa crime visualization, trends analysis, and prediction using machine learning linear regression technique," *Appl. Comput. Intell. Soft Comput.*, vol. 2021, pp. 1-12, 2021. DOI: 10.1155/2021/6658720.
12. Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). Wiley. <https://doi.org/10.1002/0471722146>
13. Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press. <https://mitpress.mit.edu/9780262018029/machine-learning/>
14. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297. <https://doi.org/10.1007/BF00994018>
15. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436-444. <https://doi.org/10.1038/nature14539>
16. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
17. Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81-106. <https://doi.org/10.1007/BF00116251>
18. [UCI Machine Learning Repository. (n.d.). *Communities and Crime Unnormalized Data Set*. Retrieved from <https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized>
19. Alsaqabi, A., Aldhubayi, F., & Albahli, S. (2019). Using machine learning for prediction of factors affecting crimes in Saudi Arabia. *Association for Computing Machinery*. <https://doi.org/10.1145/3361758.3361764>
20. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
21. Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review*, 44(4), 588-608.
22. Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). Wiley. <https://doi.org/10.1002/0471722146>
23. Jadhav, P., & Rajpal, N. (2017). Crime pattern detection analysis and prediction using machine-learning techniques. In *Proceedings of the 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence* (pp. 382-387). <https://doi.org/10.1109/CONFLUENCE.2017.7943207>
24. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444. <https://doi.org/10.1038/nature14539>
25. 25 Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

26. Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81-106. <https://doi.org/10.1007/BF00116251>
27. Shaw, C. R., & McKay, H. D. (1942). *Juvenile delinquency and urban areas*. University of Chicago Press.
28. Sri, L. A., Manvitha, K., Amulya, G., Sanjuna, I. S., & Pavani, V. (2020). FBI crime analysis and prediction using machine learning. *Journal of Engineering Science*, 11(4), 124-130. <https://doi.org/10.12691/jes-11-4-7>
29. Wang, T., Rudin, C., Wagner, D., & Sevieri, R. (2013). Learning to detect patterns of crime. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 515-523). <https://doi.org/10.1145/2487575.2488217>
30. Zhang, Z., Wu, J., Zhu, J., Zhang, J., & Zhang, L. (2016). An ensemble method for crime prediction based on urban socio-economic and criminal data. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 46(1), 75-85. <https://doi.org/10.1109/TSMC.2015.2450677>