ISSN: 2320-2882

IJCRT.ORG



INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

FORECASTING FUTURE CRIME: A PREDICTIVE MODEL FOR CRIME RATE USING MACHINE LEARNING

¹Gaurav Prajapati, ²K Monika, ³Avni Singh, ⁴Vaibhav Kumar Srivastava

¹Student, ²Student, ³Student, ⁴Student

Department of Computer Science and Engineering

SRM Institute of Science and Technology, Modinagar, Ghaziabad, Delhi-NCR

Abstract: The rise of criminal activities worldwide has created a need for law enforcement agencies to implement more efficient methods for preventing crime. Traditional crime-solving techniques have been shown to be inadequate in the face of increasing crime rates. As a solution, the integration of machine learning (ML) and computer vision algorithms and techniques is a promising approach for predicting and preventing crime. The main objective of this study is to illustrate the potential of ML and computer vision in assisting law enforcement agencies in detecting, preventing, and solving crimes with greater speed and accuracy. The study highlights previous successful cases where these techniques have been applied, demonstrating their impact on law enforcement agencies through statistical analysis. The significance of this research lies in the potential to revolutionize law enforcement by significantly improving crime detection and prevention. If implemented successfully, the use of these techniques has the potential to ease the burden on police officers and lead to more effective crime prevention measures. The primary objective of this study is to create a machine learning-based predictive model for crime rates. The dataset utilized in this research encompasses various socio-economic, demographic, and geographic attributes that may influence crime rates. Different machine learning algorithms such as decision trees, random forests, and neural networks are used in this study to build and compare predictive models. The findings reveal that the predictive model based on the random forests algorithm delivers the highest accuracy for predicting crime rates. This research highlights the potential of machine learning techniques as a valuable tool to aid law enforcement agencies and policymakers in making informed decisions to prevent and reduce crime rates.

Index Terms - Machine learning, Computer vision, Crime forecasting, Data Gathering, Gaussian Naïve Bayes.

I. INTRODUCTION

For this study secondary data has been collected. From the website of KSE the monthly stock prices for the sample firms are obtained from Jan 2010 to Dec 2014. And from the website of SBP the data for the macroeconomic variables are collected for the period of five years. The time series monthly data is collected on stock prices for sample firms and relative macroeconomic variables for the period of 5 years. The data collection period is ranging from January 2010 to Dec 2014. Monthly prices of KSE -100 Index is taken from yahoo finance.

Crime rate prediction using machine learning is an emerging field that aims to use machine learning algorithms to predict the likelihood of criminal activities in each area based on historical crime data. This approach can help law enforcement agencies to take a more proactive approach to crime prevention and resource allocation, ultimately improving public safety. Machine learning-based crime rate prediction is a nascent field that seeks to leverage machine learning algorithms to forecast the probability of criminal activities in a particular region based on past crime data. The approach can aid law enforcement agencies in taking a proactive stance towards crime prevention and resource allocation, resulting in better public safety.

One of the primary benefits of using machine learning in crime rate prediction is its ability to identify patterns that may not be visible to human experts. This, in turn, can translate to more precise forecasts and more efficient crime prevention tactics. Furthermore, machine learning can help law enforcement agencies optimize their resource allocation and target their efforts in areas with the highest need.

Traditional crime analysis methods often rely on human expertise and intuition, which can be subject to biases and errors. However, machine learning models can process vast amounts of data and discover patterns that may not be apparent to humans. By considering various factors such as socio-economic, demographic, and geographic data, machine learning models can create more accurate predictions of future crime rates, enabling law enforcement agencies to prioritize their efforts in areas where they are most needed.

One significant challenge in crime rate prediction using machine learning is the quality and completeness of the data. Historical crime data often has gaps and inconsistencies, and some variables that may have an impact on crime rates, such as drug addiction rates or unemployment rates, may not be accurately recorded or available. Therefore, it is crucial to ensure the quality and completeness of data to build robust and accurate predictive models.

In conclusion, crime rate prediction using machine learning has the potential to transform how law enforcement agencies approach crime prevention and resource allocation. By leveraging machine learning algorithms to analyse large amounts of data, law enforcement agencies can improve public safety by focusing their resources on areas with the highest risk of crime.

1.1 Present technologies used in crime rate detection and prediction

The progress in crime detection and prediction technology has advanced significantly in recent years. Among the commonly used technologies is video surveillance, which enables law enforcement agencies to monitor public places and detect criminal activities. Predictive policing utilizes machine learning algorithms to analyse past criminal data and anticipate the possibility of criminal activities occurring in specific areas. DNA analysis is a crucial technology used in crime solving, as DNA samples collected from crime scenes can be examined to identify potential suspects. Facial recognition technology is extensively used in public places to identify individuals from images or videos. Gunshot detection technology uses sensors to identify and locate gunfire in real-time, alerting law enforcement agencies of possible gun violence. Social media analysis involves monitoring social media platforms to detect and prevent criminal activities. Although these technologies have improved the capability of law enforcement agencies in crime detection and prevention, their implementation must be regulated and carefully evaluated due to privacy concerns.

1.2 ML techniques used in crime prediction

Machine learning (ML) techniques have shown great potential in crime prediction, and there are several methods currently being used in this field. Here are some of the ML techniques commonly used in crime prediction:

- 1. Neural Networks: Neural networks imitate the functioning of the human brain and can recognize patterns in large datasets to make predictions. Neural networks are often applied to identify areas with a higher probability of crime and forecast future criminal activities in those locations.
- 2. Decision Trees: Decision trees use a tree-like model to make decisions based on multiple variables. In crime prediction, decision trees can be used to analyse crime data and determine the most significant factors contributing to criminal activities.
- 3. Support Vector Machines (SVMs): SVMs are a type of ML algorithm that helps in classification and regression analysis. SVMs are commonly used in crime prediction to analyse past criminal data and recognize patterns that may result in criminal activities.
- 4. Random Forests: Random forests are an ensemble learning technique that combines multiple decision trees to make predictions. Random forests are employed in crime prediction to improve accuracy and minimize errors in the prediction process.
- 5. Deep Learning: Deep learning is a type of ML technique that employs artificial neural networks to simulate the complex processing of the human brain. In crime prediction, deep learning can be used to analyse and classify crime data, identify patterns, and predict future criminal activities.

These ML techniques, along with others, have shown great promise in crime prediction and have the potential to significantly improve law enforcement agencies' ability to detect and prevent criminal activities. However, their implementation requires careful consideration and regulation to address potential privacy and ethical concerns.

1.3 Comparative study of different forecasting methods

To compare different forecasting methods used in predicting crime rates, researchers conduct a comparative study that involves analysing the performance of various machine learning algorithms on crime datasets. The study aims to determine which algorithm provides the most precise and accurate predictions.

Multiple studies have compared different machine learning algorithms, including decision trees, random forests, support vector machines (SVM), k-nearest neighbours (KNN), Gaussian Naive Bayes (GNB), and gradient boosting. The evaluation of each algorithm's performance is based on different metrics such as accuracy, precision, recall, F1 score, and area under the curve (AUC) of the receiver operating characteristic (ROC) curve.

These metrics provide valuable insights into how well the algorithm performs in predicting crime rates accurately and distinguishing between true positive and false positive cases.

Some studies have also compared the performance of ensemble methods, such as bagging and boosting, with individual algorithms. Bagging is a technique that involves combining multiple models to obtain a more accurate prediction, while boosting is a technique that involves sequentially adding models to the ensemble and adjusting the weights of misclassified instances to improve the prediction accuracy.

The purpose of conducting a comparative study of different machine learning algorithms in crime rate prediction is to understand their advantages and limitations and identify the most suitable algorithm for a particular crime dataset. It's crucial to note that the performance of these algorithms may vary depending on the crime data's characteristics and the quality of the prediction features used. Therefore, conducting a thorough assessment of the algorithms is essential before selecting the most appropriate one for crime rate prediction.

II. LITERATURE SURVEY

Crime rate prediction is an important task in the field of criminology and law enforcement, and machine learning (ML) has shown promising results in this area. In recent years, there has been an increasing interest in developing ML models to predict crime rates, and several studies have been conducted on this topic. In this literature survey, we will review some of the key research papers on crime rate prediction using machine learning:



Table 1: Recent Literature								
Title of the Paper a Published Year	nd	Author's Name		Advantages		Disadvantages		
Crime forecasting: a machine learning and computer vision approach to crime prediction and 		 ah, Innovative Methodology Comprehensive Analysis High Prediction Accuracy Real World Application 		 Complex Methodology Not easily interpretable Limited Generalisability High quality data may not be available in all location 				
Applying learning criminology:machine to semi- parametricPublished By: Roman Marcha Sebastian Haan, Garner Clanc and Sally Cripps PublishedPublished By: Roman Marcha Sebastian Haan, Garner Clanc 		nnt, cey	 Robustness Improved accuracy in predicting crime Better understanding of criminal behaviour Better resource allocation 		 Biased Data used Lack of Transparency Ethical concerns Overreliance on Technology 			
Crime Rate	Published	by: Sakib	•	Scalability	•	Lack of Human Insight		
Prediction Using	Mahmud, Mu	sfika Nuha, and	•	Improved Public Safety	•	False positives and False		
Machine Learning	Abdus Sattar I	Published	•	Improved Investigation		negatives		
(2020)	Place: Springe	.r	•	Better understanding of	•	Privacy concerns		
	r meet springe		•	Early warning systems		law enforcements		
A systematic review on spatial crime forecasting (2020)	Published Kounadi, Alir Araujo Jr. and Published Plac	by: Ourania na Ristea, Adelson Michael Leitner ce: Springer	•	Identification of Best Practices Identification of Knowledge Gaps Potential for Practical Application Improved Decision Making	••••	Limited Original Research Limited Scope Potential for Bias Lack of Context		
Crime Prediction	Published	by: Faisal	•	Novel Dataset	•	Data quality issues		
using Machine	Machine Tareque Shohan, Abu Ubaida		•	Reduced Human Bias	•	Algorithmic bias		
Novel Crime	Mohammad Shafiul Alam		•	Cost savings	•	Potential for misuse		
Dataset (2022)	Published University	Place: Cornell	•	Improved conadoration	•	considerations		
Forecasting Crime with Deep Learning (2018)Published by: Alexander Stec and Diego KlabjanPublished place:		•	Automated analysis	•	High sensitivity to input data			
		•	Integration with other	•	Lack of interpretability			
		•	Flexibility	•	Overfitting			
	Cornell University		•	Data-driven decision making	•	Difficulty in Feature Engineering		
Predicting and Published by: Muzammil Khan.		•	Cost effectiveness	•	Biases in Data			
A Crime A Trans A Li and Vassar A lbark		•	Improved Crime	•	Limited Scope			
Prediction Model		•	Scalability	•	Limited Flexibility			
Using San Francisco Crime Data by Classification	Published Place	lished Place: Wiley Hindawi		Enhances situational awareness				
Techniques (2022)								

III. METHODOLOGY

The methodology of crime rate prediction using machine learning involves the following steps:



Fig: - Workflow of the Model

3.1 Data collection: Crime data is collected from various sources such as law enforcement agencies, government reports, news articles, and social media platforms. In crime rate prediction, the data gathered may include information about the location of the crime, the type of crime committed, the time and date of the crime, demographic information of the area, weather conditions at the time of the crime, and other relevant features. It is important to ensure that the data collected is complete and accurate to avoid any biases or errors in the predictions. Additional data, such as weather, demographic, and geographic information, may also be collected as they can affect crime rates. Initially, we gather a considerable amount of crime news articles from the newspaper, which are then classified into six distinct types of crimes. For each crime incident in all six categories, we extract multiple pieces of information about the crime, including the date, location, time, victim's and criminal's details, and so on. This forms the fundamental dataset.

3.2 Data pre-processing: In crime rate prediction using machine learning, data pre-processing is important. It involves cleaning and transforming raw data into a format suitable for machine learning algorithms. This step ensures data accuracy, completeness, and relevance for the task at hand. It includes data cleaning, integration, transformation, and reduction. Data pre-processing significantly affects the accuracy and effectiveness of the model. Proper techniques should be applied to address inconsistencies, missing values, and errors. In the case of natural language text, data cleaning and feature engineering are crucial, especially when creating datasets from newspaper and news articles, and other records.

3.3 Feature engineering: Feature engineering is the process of selecting and transforming raw data into meaningful features that can be used as input for machine learning models. In crime rate prediction using machine learning, this involves selecting relevant features that can affect crime occurrence and transforming them into a format suitable for machine learning algorithms. It also includes extracting additional features from the data to improve the predictive power of the models. Feature engineering aims to create a high-quality dataset that can enhance the accuracy and effectiveness of machine learning models in predicting crime rates.

3.3.1 Feature Scaling: Feature scaling is a technique for normalizing the numerical valued features in a fixed range. Without scaling features, the learning algorithm may be biased toward features with greater magnitude values. As a result, feature scaling brings all features into the same range, and thus the learning model makes good use of all features. For numeric features, we apply a well-known scaling technique called min-max normalization. This technique brings every numerical attribute in a defined range. The most common ranges used are [0, 1] and [-1, 1].

3.3.2 Feature Encoding: Machine learning algorithms can only operate on numerical values. Hence it is necessary to convert the categorical features into numeric ones. There are various methods for encoding categorical features. We use the label encoding method which turns categorical data into machine-readable numeric form by assigning a unique number (beginning with 0) to each class of a particular feature. Different types of information are contained in geo-location data, temporal data, weather data, and demographic data. All of these features are combined in the dataset, thereby resulting in a strong set of features for the prediction task.

3.4 Model selection: To predict crime rates using machine learning, selecting the appropriate model is crucial. This involves considering factors like dataset size and complexity, the type of problem, available computing resources, and desired accuracy level. Various machine learning algorithms are applied to the pre-processed data, and techniques like cross-validation are used to evaluate model performance. The selection process involves experimentation and evaluation using performance metrics like accuracy, precision, recall, and F1 score. The chosen model should also be interpretable and provide insights into the factors contributing to crime rate trends. In our dataset, we have applied algorithms like Random Forest, K-Nearest Neighbour, Naïve Bayes, and Gradient Boost Classifier to select the best algorithm. We also deal with the imbalanced nature of the dataset and conduct experiments on a different setting.

3.5 Model training, evaluation and deployment: After the model selection process, the chosen model is trained on a training set and evaluated on a separate test set using performance metrics like accuracy, precision, recall, and F1 score. Once the model is trained and validated, it can be used for real-time crime rate prediction. This can be done by integrating the model into a software application or dashboard for easy access and visualization of results. In our case, the models used for crime rate prediction on our dataset include Random Forest, Gaussian Naïve Bayes, Decision Tree, Linear SVC, Polynomial SVC, Gradient Boost Classifier, and K-Nearest Neighbour.

3.5.1 To classify or predict outcomes, **Random Forest** is a commonly used machine learning algorithm that creates multiple decision trees during training time and returns the mode of the classes (classification) or mean prediction (regression) of the trees. Each tree is created with a random subset of features and a random subset of the training data. One of the key advantages of Random Forest is that it can manage a large number of input variables without overfitting the model. Furthermore, it is robust to noisy data and missing values. It also provides a way to determine the significance of each feature in predicting the outcome variable.

3.5.2Support Vector Machine Classifier (SVC) is a well-known algorithm used for classification tasks in supervised learning. The algorithm aims to determine the optimal boundary or hyperplane that can best separate the data points of different classes with maximum possible margin. The points closest to the hyperplane, called support vectors, are used to calculate the optimal hyperplane. SVC is flexible and can work with different kernel functions such as linear, polynomial, and radial basis function (RBF). It is also suitable for handling high-dimensional data and can be used for both binary and multi-class classification problems.

3.5.3K-Nearest Neighbours (KNN) is a non-parametric algorithm used for regression and classification that is easy to understand. The classification of a new data point in KNN is based on the majority class of its k-nearest neighbours in the training set. The value of k is a hyperparameter that must be determined before the model is trained. Because KNN does not make any assumptions about the distribution of the data, it is useful for nonlinear and complex decision boundaries. However, for large datasets, it can be computationally expensive because it needs to search through the entire training set to find the nearest neighbours.





Fig: - Results when data is compared on KNN

3.5.4A **Decision Tree** is a supervised machine learning algorithm that is useful for classification and regression tasks. It functions by dividing the data into subsets based on the values of different attributes and then making decisions about which attributes to split on iteratively to achieve the best classification results. This creates a model in the form of a tree, where each internal node represents an attribute, each branch represents a potential value of that attribute, and each leaf node represents a class label or numerical value. The algorithm partitions the data into subsets based on feature values until each subset contains only one class. It selects the feature that results in the greatest information gain, which measures the reduction in entropy or impurity of the data following a split. Decision Trees can handle both categorical and numerical data, and they are easy to interpret. However, they are vulnerable to overfitting and may become complex for large datasets.

3.5.5Polynomial SVC (**Support Vector Machine**) is a type of SVC algorithm that involves using a polynomial kernel function to transform input data into a higher-dimensional feature space. This transformation helps to separate the data into different classes more easily. The degree of the polynomial determines the complexity of the transformation, and a higher degree can lead to overfitting of the model. To avoid overfitting, the degree of the polynomial is typically chosen using cross-validation, which estimates the performance of the model. Polynomial SVC works by finding the hyperplane that separates the classes in the higher-dimensional feature space. The complexity of the decision boundary is influenced by the degree of the polynomial kernel function, with higher degrees producing more complex boundaries. Polynomial SVC is often used for classification tasks where the data is not linearly separable in the original feature space. However, it can be sensitive to the choice of hyperparameters and may be computationally expensive for large datasets.

3.5.6 Gaussian Naive Bayes (GNB) is a machine learning algorithm used for classification tasks based on Bayes' theorem and an assumption that the features are independent of each other. It is widely used in natural language processing applications such as spam filtering and text classification. The algorithm assumes that each feature follows a Gaussian distribution, and estimates the mean and variance of each feature for each class in the training data. Then, it calculates the likelihood of the test data belonging to each class based on these parameters, taking into account the prior probability of each class occurring in the training data. The class with the highest posterior probability is predicted as the output class for the test data. Compared to other classification algorithms, GNB is simple and computationally efficient. However, it may not work well when the features are correlated or when the assumption of independence among features is not met. Moreover, GNB may suffer from the problem of zero frequency, which can lead to a probability of zero and cause the model to break down.

3.5.7The Gradient Boosting Classifier (GBC) is a popular algorithm for classification tasks that combines the predictions of several individual models to produce a more accurate and robust prediction. It does this by building a series of decision trees, where each tree is trained on the residuals of the previous tree. This process continues until a set number of trees have been built or until the desired accuracy level has been achieved. GBC has several advantages over other classification algorithms, including its ability to handle non-linear relationships between features, its robustness to overfitting, and its ability to perform well on imbalanced datasets and handle missing data. However, GBC also has some drawbacks, such as its computational expense, sensitivity to hyperparameters, and poor performance on high-dimensional data due to the curse of dimensionality.

3.6 Model maintenance and updates: As new data becomes available; the model may need to be updated or retrained to maintain its accuracy and effectiveness over time. Model maintenance and updates are important in crime rate prediction using machine learning as the underlying data distribution may change over time, and the predictive performance of the model may degrade. To maintain the model, regular monitoring and evaluation of the model's performance are necessary. If the model's performance drops below a certain threshold, it may be necessary to update the model with new data or to retrain the model with a different algorithm or hyperparameters. Updating the model with new data can be done by adding new instances to the existing dataset and retraining the model. It can also involve adding new features or modifying existing ones based on the insights obtained from the monitoring process. Feature selection methods can be used to identify the most relevant features that contribute to the model's accuracy. Retraining the model with a different algorithm or hyperparameters may also be necessary if the model's performance deteriorates. This involves selecting a different algorithm or tuning the hyperparameters to improve the model's accuracy. Regular maintenance and updates help ensure that the model remains effective and reliable over time.

Models	Accuracy	
Decision tree	83.5926	
svc	80.8779	1.2
KNN	77.8224	
Random Forest	78.9188	
Polynomial SVC	76.2136	
Gaussian NB	80.56	
Gradient boosting Classifier	79.9219	

Fig: - Accuracy when compared different algorithms

IV SYSTEM SPECIFICATION

Table 2: System Specification

SPECIFICATION	DESCRIPTION		
CPU	RYZEN 5 3550H		
GPU	12GB VRAM		
RAM	8 GB		
SSD Version	SSD MOBILENE VERSION 1 FPN 640-640		
Augmentation Data	90 degrees Clockwise, Anti-Clockwise, 50 Percent Positive Brightness, 50 Percent Negative Brightness.		

REFERENCES

[1] Jenga, K., Catal, C. & Kar, G. Machine learning in crime prediction. J Ambient Intell Human Comput, (2023).

[2] Tayebi, M. A., Frank, R., & Glässer, U. (2012). Understanding the link between social and spatial distance in the crime world. Proceedings of the 20th International Conference on Advances in Geographic Information Systems - SIGSPATIAL '12.

[3] Awal, M. A., Rabbi, J., Hossain, S. I., & Hashem, M. M. A. (2016). Using linear regression to forecast future trends in crime of Bangladesh. 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV).

[4] Nath, S. V. (2006). Crime Pattern Detection Using Data Mining. 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops.

[5] Parvez, M. R., Mosharraf, T., & Ali, M. E. (2016). A Novel Approach to Identify Spatio-Temporal Crime Pattern in Dhaka City. Proceedings of the Eighth International Conference on Information and Communication Technologies and Development - ICTD '16.

[6] Gandhi, M., Kamdar, J., & Shah, M. (2019). Pre-processing of Non-symmetrical Images for Edge Detection. Augmented Human Research, 5(1).

[7] Biswas, A. A., & Basak, S. (2019). Forecasting the Trends and Patterns of Crime in Bangladesh using Machine Learning Model. 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT).

