Machine Learning-Based Predictive Modeling Approach for Criminal

Priyanka Jadhav[1], Amruta Ladgaonkar[2], Kaivalya Patil[3], Shivani Dive[4], Prof. Nakul Sharma[5]

[1234](Students, STES'S SINHGAD ACADEMY OF ENGINEERING, KONDHWA, PUNE, India)

[5](Professor, STES'S SINHGAD ACADEMY OF ENGINEERING, KONDHWA, PUNE, India)

Abstract: ***Increasing in Criminal activities adversely affects any country's economic growth including quality of life of the peoples. There has been enormous growth in crime rates in recent years. Recognizing the patterns of criminal activity of a place is paramount in order to prevent it. Many governments are trying to use advanced techniques to tackle such issues.***

The aim of the study is to identify the crime rate and solve crimes faster based on the data collected. The relationships exist among the various crime types and crime Variables. Scrutinizing this data set can offer insight into criminal activities. Here we are using a random forest algorithm to build a prediction model for identifying the various crimes. The comparison of various Supervised Learning Algorithm i. e Support Vector Machine, KNN, Decision tree, and Naïve Bayesian is done against Random forest.

Keywords: ***Crime Rate Analytics, Machine eLearning Approach, Crime Prevention, crime types, prediction model, Data Collection, Dependent Variable Analysis, Word2Vector Formation, and Algorithm Analysis.***

I. INTRODUCTION

Past 1 year, crimes in India have seen a spike. Increase in the misdemeanor, like theft, murder, rapes, and kidnapping can affects the peoples and life badly along with countries financial growth as additional police forces, courts are required for solving the crime issues [1]. The rise in crimes, generate massive data every year by the law enforcement organizations [2]. So it becomes challenging to analyze crime manually and implement decision for avoiding crimes in future. Today, criminals are becoming technologically advanced, so there is a need to use advanced technologies to keep police ahead of them. Reducing the crime rates have turn out to be a very important and yet challenging task. It is significant to recognize different features, existence relations of crimes and thus determining enhanced technique to reduce crimes rates.

According to the research there are certain region where majority of the misdemeanor happens compared to others. Criminals are mostly active and operate in their comfort zones [3]. There is certain relation between underlying pattern of crime and the region or an area's. Predicting such crime patterns is an important to develop more efficient strategies either to prevent crimes or to advance the investigation efforts based on previous data availability like case description, location, date and time.

Here, we can extract valuable information from an unstructured data. By using random forest techniques prediction of the crime and its types with crime locations is done. Such crime data analysts can benefits the Law enforcement officers to accelerate the crimes solving procedure.

II. LITERATURE SURVEY

Papers used to study Data Mining Techniques for crime detection are explained below;

S. Sathyadevan and S. Gangadharan."CRIME ANALYSIS AND PREDICTION USING DATA MINING", 2014 IEEE Conference

Method: - For identifying trends in crime place, clustering crime-prone areas, and visualizing patterns across maps, a number of techniques are used.

Dataset: - The real-time data, which primarily consists of blogs and social media feedback, as well as posts, was stored in Mongo DB.

Techniques used: - Three approaches are used in this crime detection and forecast approach:

- Naive Bayes Algorithm.
- Apriori Algorithm
- Decision Tree Approach.

Description: - The technique is proposed for forecasting or evaluating places with a high probability of violence.

This work also aids in the visualization of crime-prone areas as well as criminal profiling. The proposed framework was able to forecast crime activity on a day-by-day basis as well as crime-prone regions/locations in India.

Accuracy: - Naive Bayes algorithm gives approximately 90% accuracy for their dataset.

Advantage: - Visualization of crime-prone regions- The proposed work's primary asset is the ability to forecast the areas where crime is most likely to occur today.
Disadvantage: - The data was manually gathered from the Libyan police force.

D. Z. Zubi And A. Mahmmud "CRIME DATA ANALYSIS USING DATA MINING TECHNIQUES TO IMPROVE CRIMES PREVENTION" IJCA.
Method: - Data mining algorithms
Dataset: - The dataset was taken manually from Libyan police department.
Techniques used: - DM (Data mining) algorithms used are;

- Association Rule Mining
- K-Means Clustering.

Description: - The Methodology is to extract crime patterns and clustering them to classify crime records. The Technique was able to extract Relation between criminal age and no of crimes is high. For investigation, clusters are created based on location and crime type.
Advantage: - is analyzing crime patterns become easy due to clusters.
Disadvantage: - is cannot predict for multi-modal data or high dimensional data.

Sunil Yadav, et.al. "Crime pattern detection, analysis and prediction" IEEE Conference 2017
Method: - Data mining algorithms and correlation and regression.
Dataset: - The data was obtained from India's online server (2001-2014).
Techniques used: - DM (Data mining) algorithms used are;

- Apriori.
- K-Means Clustering.
- Naïve Bayes.
- Correlation and Regression.

Description: - Creating clusters and mining regular results, grouping and identifying correlation, and regression from a dataset are all part of the technique. Detect violence in numerous states. The methodology was able to establish a 0.98 association between state and crime rate. Regression reveals that only three cases out of ten are found guilty of the charges.
Advantage: - Predictions of crime are made dependent on states and age ranges in comparison to dates.
Disadvantage: - It is impossible to estimate crime hotspots in terms of time because there is inadequate evidence.

<u>And Machine Learning Techniques include same additional making use of images for crime detection and prediction are as shown below.</u>

"Prediction of Crime Occurrence from multi modal data using deep learning" Research Article, 2017.
Method: - Machine Learning Approaches
Dataset: - The dataset used was American Fact-Finder 2014, weather data, Google street view images.
Techniques used: - The Techniques used are;

- DNN
- Pearson Correlation -Coefficient Analysis
- SoftMax classifier

Description: - The technique consists of fusing multi-modal data and using police enforcement reports in specific locations to forecast the likelihood of crimes. Multi-modal data fusion with DNN has an accuracy of 84.25, and it can also efficiently fuse multi-modal data with environmental background information. The paper uses previous illegal activities to predict crime events.
Advantage: - It fits well for multi-model data and high-dimensional data.
Disadvantage: - Weakness is the DNN – The DNN-based crime incidence and forecast cannot be used on inadequate evidence, which is a flaw.

Mahmud-Zinnah "CRIMECAST: A Crime Prediction and Strategy Direction Service" IEEE, 2016.
Dataset: - The dataset is prepared, probability of crime i in location s is calculated, hotspot detection is done, & using ANN crime-cast is implemented.
Description: - Calculating each crime's risk factor based on its position. ANN is used to measure the risk factor of each crime for each venue. The framework proposed CRIMECAST, a crime prediction model with mathematical and ANN implementation.
Advantage: - Strengths is the ANN implementation gives us more precise prediction than mathematical implementation

<u>Text/ NLP-based methods</u>

Ehab Hamdy, et.al, "Criminal Act Detection and Identification Model", *Proceedings of 7th International Conference on Advanced Communication and Networking*

Dataset: - Sensor data, security camera records, social media updates, and messages are among the specifics found in the datasets.
Description: -
Input Used: - Crime records, age, previous convictions, countries visited, birthplace, average ATM use, types of crimes Entrance based on the time of day, crime hotspots, and errors made by victims.
Pre Processing: - Organizing data in the form of some mentioned parameters like Time, Final Movement, Frequency Rate, Video, Images, and Audio
Feature Extraction: - Sliding window resemblance matching for sensory representations semantics of text Lexical processing and Natural Language Processing were used to interpret the text content (NLP).
Classification/ Clustering: - To predict the resemblance of a given input to a suspect object or position, a qualified classification model is used.
Outcome: - Suspicious conduct is divided into three categories: "High," "Medium," and "Low." **Advantages: -** Take into account location feeds and smartphone user data
Disadvantages: - Not providing a good picture of how criminal activity is processed and compared.

III. PROPOSED METHODOLOGY

A. Architecure

The proposed system helps to detect the criminal activity in a geographic area to understand the underlying pattern of the crime the area suffers from. The module wise description of the proposed system is given below:

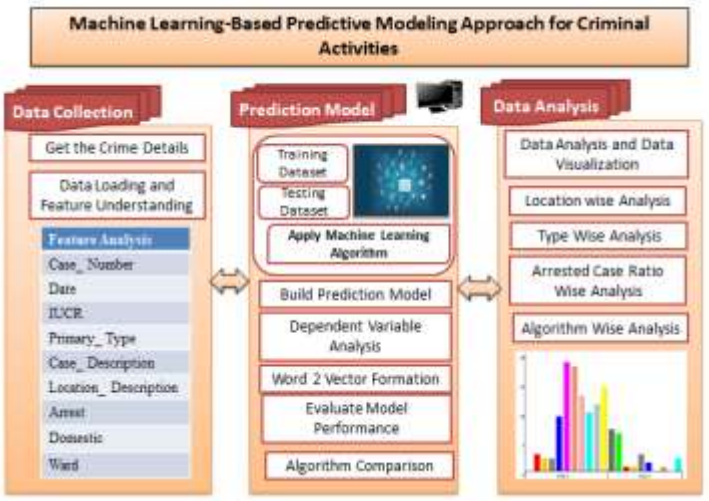

Fig 1: System Architecture

1. **Data Uploading and feature understanding: The** data downloaded from kaggle is preprocessed first so that we can extract important features that are quite natural for predicting the crime. Such as few streets or locations, date and time, areas have a higher concentration of criminal activities compared to others.

This data is used as input to the system to predict and solve crimes at a much faster rate. The main features used for crime rate detection are as follows:

a. Case_ Number
b. Crime occur_ Date
c. Primary_ Type
d. Case_ Description
e. Location_ Description
f. Arrest
g. Domestic
h. Ward

Different studies and researches have shown that significant concentration of crime happens at micro level of a region. Also, an overall study of criminal activity in a geographic area helps to understand the underlying pattern of the crime the area suffers from.

2. **Dependent Variable Analysis:** Dependent variables or Predicted variables are the ones that help to get the factors that mostly dependent on crime-related variables. For example, the Code for the type of crime (iucr) has nothing to do with the crime rate prediction or it is least bother for the prediction. So here by using the dataset, we achieve the terms or the factor that is mostly affecting to make crime prediction.

The analyzed data is visualized for the word to vector formation and on this fine-tuned data we can apply the algorithm to get the final result.

3. **Analytics:** Exploratory Data Analysis is an initial process of analysis, in which you can summarize characteristics of data to can predict numerous crimes and predicting the type of crime, probable location, which may happen in the future depending upon various conditions.
4. **Built Prediction Model:** The system builds a prediction model by using a random forest technique. It is one of the ensembles learning technique which consists of several decision trees rather than a single decision tree for classification. While classifying all the trees in the random forest gives a class to an unknown example and the class having maximum votes will be assigned to the unknown example.

The techniques perform dependent variable analysis and word-formation vector to Predict Important Aspects Of Crime detection and prevention.

The data analysis helps to identify useful features for building a predictive model, Such as Predicting arrests for a given type of crime in a given location and predicting the number of crimes in particular on a given day and time.

B. Algorithm

Random Forest Algorithm:

Random Forest algorithm is a supervised classification algorithm. The decision tree gives a tree like structure to show the possible significances. The decision tree creates some set of rules using targets and features to give the prediction.

The technique mainly has two stages like random forest creation and prediction.

A. Random Forest Creation:

1. Select “K” attributes from overall “m” where k << m, randomly.

2. Between the “K” features, estimate the node “d” using the finest split point

3. Divided the node into daughter nodes by the best splitting.

4. Repeat the 1 to 3 steps till “l” number of nodes has been reached.

5. Create forest by iterating steps 1 to 4 for “n” number times to create “n” number of trees.

B. Prediction Using Random Forest Classifier:

1. Consider the test features and use the rules for every randomly created decision tree to find the result and stores the predicted outcome (target)

2. Analyze the votes for each expected target

3. Consider the high voted predicted target as the final prediction from the random forest algorithm.

IV. RESULT AND DISCUSSION

The proposed system builds a prediction model for forecasting the arrests for a given crime. System accuracy is evaluated using the classification report generated.

Here we have used a total of 100000 records for experimentation.

Total Records --> 100000

Training Records --> 75000

Testing Records --> 250001.

The classification report for the above-mentioned algorithm is given below, where,
Class a represents --> No Arrest
Class b represents --> (Yes) Arrest

TABLE I**:** Classification Report for Random Forest

	a	**b**
a	21150	4059
b	1089	23936

TABLE II**:** Classification Report for SVM

	a	**b**
a	19077	6132
b	1975	23050

For above mentioned classes i.e. for class a & class b the accuracy and precision is calculated based on TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative).

TABLE III: Confusion Matrix for Random Forest

	TP	TN	FP	FN
a	21150	23936	1089	4059
b	23936	21150	4059	1089

TABLE IV: Confusion Matrix for SVM

	TP	TN	FP	FN
a	19077	23050	1975	6132
b	23050	19077	6132	1975

Depending on the above TP, TN, FP & FN values the Accuracy, Precision are Calculated by using the below formula:

Precision = TP / (TP+TN)

Accuracy = (TP+TN) / (TP+TN+FP+FN)

F-score = 2*TP/ (2*TP+FP+FN)

The algorithm comparison in terms of Accuracy, Precision is given below:

TABLE V: Algorithm Comparison

Algorithm	Accuracy	Precision	F1-Score
SVM	73.86	65.68	71.48
Random Forest	85.65.75	81.1	85.15

The graphical representation of the above table is given in figure below:

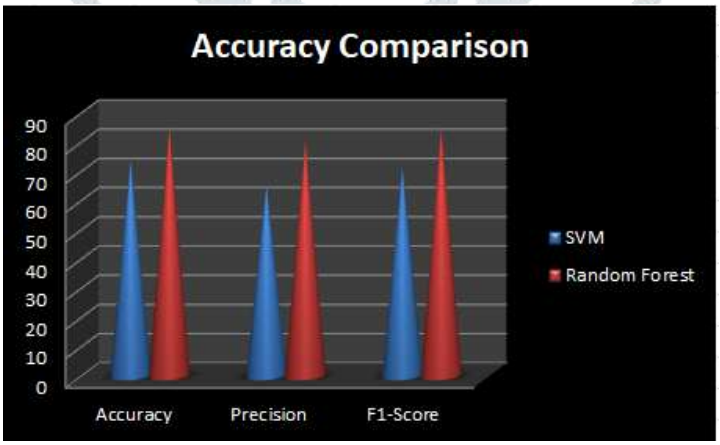

Fig. 2 Graphical Representation of Algorithm Comparison

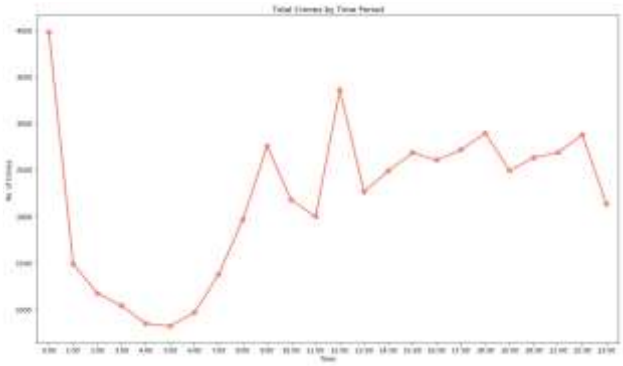

Fig. 3 Time Series Prediction for Crime Data

The graph tell us that according to time featured there are more crime at midnight then it is decreased and it is again increased at morning. The X coordinate denotes the time and Y coordinate denotes number of crime.

I. CONCLUSION

The proposed system helps the enforcement agencies to have a prior assumption of the type of the criminality and it would give them strategic advantages to resolve cases faster. The system detect the crime rate and improves the investigation speed

depending on the collected data. Random forest algorithm is used to build a prediction model for recognizing the various crimes. Dependent variable analysis and word-formation vector helps to detect important aspects of crime and prevention. As compared to popular algorithm like Support Vector, Random forest is gives higher accuracy for crime prediction.

The proposed model helps us to analyze the time series crime data. Time Series Analysis helps in understanding the patterns in the data. As the crime is also time dependent, Forecasting of future crime events can be performed on such data which is dependent on Time. We can find the maximum crime has happened previously in certain time period, year-wise crime using the available crime data. We are more likely to use ARIMA model for time series prediction of crime.

REFERENCES

[1] Hitesh Kumar Reddy Toppi Reddya,, Bhavna Sainia, Ginika Mahajana , "Crime Prediction & Monitoring Framework Based on Spatial Analysis ",International Conference on Computational Intelligence and Data Science (ICCIDS 2018).

[2] Prajakta Yerpude1 and Vaishnavi Gudur,"PREDICTIVE MODELLING OF CRIME DATASET USING DATA MINING",International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.7, No.4, July 2017.

[3] Hitesh Kumar Reddy ToppiReddya, Bhavna Sainia , Ginika Mahajan"Crime Prediction & Monitoring Framework Based on Spatial Analysis",International Conference on Computational Intelligence and Data Science (ICCIDS 2018).

[4] S. Sathyadevan and S. Gangadharan, "Crime Analysis and Prediction Using Data Mining," First Int. Conf. Netw. Soft Comput., 2014.

[5] D. Z. ZUBI and A. MAHMMUD, "Crime Data Analysis Using Data Mining Techniques to Improve Crimes Prevention," Int. J. Comput., vol. 8, 2014.

[6] Sunil Yadav, Meet Timbadia, Ajit Yadav, Rohit Vishwakarma and Nikhilesh Yadav, "Crime Pattern Detection, Analysis &Prediction", in 2017International Conference on Electronics, Communication and Aerospace TechnologyICECA 2017

[7] "Prediction of crime occurrence from multi-modal data using deep learning." http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0176244.

[8] "Crimecast: A crime prediction and strategy direction service -Semantic Scholar."Available:/paper/Crimecast%3A-A-crime-prediction-and-strategy-service-Mahmud-Zinnah/89f4dc20e2c1e713c03f2eac9855411d57fc9b4b.

[9] Ehab Hamdy, Ammar Adl, Aboul Ella Hassanien, Osman Hegazy and Tai-Hoon Kim, "Criminal Act Detection and Identification Model", *Proceedings of 7th International Conference on Advanced Communication and Networking*, pp. 79-83, 2015.

[10] Jazeem Azeez ,D. John Aravindhar ,"Hybrid Approach to Crime Prediction using Deep learning ",2015 IEEE.

[11] Hitesh Kumar Reddy Toppi Reddya,, Bhavna Sainia, Ginika Mahajana ,"Crime Prediction & Monitoring Framework Based on Spatial Analysis ",International Conference on Computational Intelligence and Data Science (ICCIDS 2018).

[12] H. Benjamin Fredrick David1and A. Suruliandi2 ,"SURVEY ON CRIME ANALYSIS AND PREDICTION USING DATA MINING TECHNIQUES", ICTACT JOURNAL ON SOFT COMPUTING, APRIL 2017, VOLUME: 07, ISSUE: 03.

[13] Suhong Kim ; Param Joshi ; Parminder Singh Kalsi ; Pooya Taheri , "Crime Analysis Through Machine Learning", 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON).

[14] Lawrence McClendon and Natarajan Meghanathan*,"USING MACHINE LEARNING ALGORITHMS TO ANALYZE CRIME DATA",An International Journal (MLAIJ) Vol.2, No.1, March 2015.

[15] J. Kiran ; K Kaishveen. ,"Prediction Analysis of Crime in India Using a Hybrid Clustering Approach", 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I- SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference.