



IJIRCCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 8, August 2022

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.165



9940 572 462



6381 907 438



ijircce@gmail.com



www.ijircce.com

Crime Rate Detection Using Machine Learning

Das Gupta Girish¹, Deepak Arora², Srivastava Pratima Kumari³

Department of Computer Engineering, Amity University Lucknow Campus India¹

Department of Computer Science Engineering, Amity University Lucknow, India²

Associate Professor, Department of Zoology, Ch.SD.St.Theresa's College for Women, Eluru, Adhi

Kavi Nanaya University, Andhra Pradesh, India³

ABSTRACT: This quantitative research study was conducted to illustrate, "Crime Rate Detection Using Machine Learning". Crime detection and prevention increasingly heavily rely on Machine learning and AI as well. But the goal of this research is to assess Machine learning techniques and their abilities to analyse the information gathered about past crimes. By contrasting them theoretically and practically, I was able to determine which Machine Learning techniques would be most effective in analysing the information gathered from sources that focused on crime prevention. These data are subjected to methods in order to evaluate how well they may be used to analyse and prevent crime. This was accomplished using performance metrics that outperformed alternative approaches in terms of recall, accuracy, and the number of cases properly categorised. I determine that Machine learning techniques aid in making predictions about the likelihood. This investigation looks at the design and implementation of a system based on past crime rate data and analyses the crime percentage in past regions at specific minutes. For this work, we use crucial information that is gathered from people based on their past crime rate issue. Is it really safe or not, that is why we are faced with many unfavourable conditions. The crime percentage figure is a strategy that makes use of a number of formulas to get the crime percentage based on prior information. We need to travel to numerous locations daily for our daily needs, and typically in our daily lives we face a variety of security challenges, such as grabbing, provocation, seizing, and so on. In this paper, we exhibit various types of crime % using various models and tables, generally using data from datasets containing years of crime rate as well, indicating the degree of expected crime rate in various situations portrayed.

KEYWORDS: Machine Learning, Crime Rate Detection, Linear regression algorithm, Random Forest classifier algorithm, K-Nearest Neighbour(K-NN) algorithm, Support Vector Machine.

I. INTRODUCTION

The greatest threat to humanity is crime. Numerous violations occur frequently throughout the course of a day. Perhaps it is growing and dispersing quickly and greatly. From tiny towns to large urban centres, there are violations. Burglary, murder, assault, battery, fake detail, kidnapping, and manslaughter are a few examples of violations. Given the rising crime rates, it is necessary to resolve cases much more quickly. The police department must regulate and reduce the crime rate exercises since they are expanding at an increasingly rapid rate. Given the vast amount of data on crime rates that are available, the police department faces major problems with crime rate expectations and criminal identification. Innovation is required so that cases can be handled more quickly. The aforementioned problem prompted me to take a test to see how easily a crime rate case could be resolved. It was discovered via extensive documentation and cases that information science, in addition to machine learning, may expedite and simplify the process. The purpose of this project is to estimate the crime rate using the dataset's highlights. There is a deletion of the dataset from the authoritative locales. We can predict the type of crime rate that will occur in a certain area with the use of machine learning calculations, using Python as the central language. Training a specific machine learning model is the objective. The preparation would be completed using the informational collection used for it, which would then be approved using the test dataset. Building the model will be completed with better calculations based on accuracy. The dataset is viewed in order to look at potential infractions that may have occurred. Today, criminal intelligence is growing with the aid of technological advancements every year. As a result, we now need to give the government and the police department access to a brand-new, potent tool (a collection of programmes) that will aid them in their efforts to solve crimes. The primary goal of crime forecasting is to identify crimes before they happen, so it is obvious how

crucial it is to use crime forecasting techniques. A victim's life, the anguish they would experience for the rest of their lives, and damage to private property may all be avoided if the crime was predicted. Even possible terrorist actions and activities may be predicted with it. In general, predictive policing and the detection of criminal patterns can both benefit greatly from Machine Learning. The police can instantly attempt to halt crimes if patterns of criminality are automatically discovered.

II. LITERATURE SURVEY

Lawrence, Natarajan Meghanathan, and McClendon. "Analyzing crime data with machine learning techniques." Applications of machine learning: Using the same limited set of features, the Linear Regression, Additive Regression, and Decision Stump algorithms have been applied to the Communities and Crime Dataset in An International Journal (MLAIJ) 2.1. Between the three techniques, the linear regression algorithm fared the best overall. This project's goal is to demonstrate how precise and effective machine learning algorithms are. Moreover identification of the areas where there is a high likelihood that a crime will occur. These researchers also depicted places that are prone to crime. They used Naive Bayes classifiers to categorise the data. This algorithm provides the statistical method for classification and is a supervised learning algorithm. This classification has a 90 percent accuracy rate and the study has been undergone by researchers Shiju Sathyadevan, Devan M. S., et al. (2014) [3]. Additionally, On the Communities and Crime Dataset, Lawrence McClendon and Natarajan Meghanathan (2015) [4] employed the Linear Regression, Additive Regression, and Decision Stump algorithms with the identical set of input (features). Comparing the three chosen algorithms, the linear regression algorithm produced the best results overall. Likewise, In Identification, analysis, and prediction of crime patterns The dataset was obtained from an Indian online platform (2001-2014). kmeans clustering, naive Bayes, correlation, and regression are some of the data mining methods employed. Cluster creation, frequent data mining, categorization, correlation detection, and regression analysis are the methods used. Identify crime in various states. The method identified a 0.98 correlation between state crime rates and rates of crime. Regression reveals that only three cases out of ten and are found guilty of the accusations. Strength: Crime statistics are predicted based on states, age groups, and dates. The inability to forecast crime hotspots with regard to time and the challenge of working with little data are weaknesses. This study is undergone by Sunil Yadav's 2017 IEEE Conference Say hello to Nikhilesh Yadav, Timbadia, Ajit Yadav, and Rohit Vishwakarma. Similarly, machine learning techniques leverage photos for crime prediction and detection, as is illustrated below. Deep learning-based crime prediction utilising multimodal data 2017 research article American Fact-Finder 2014, meteorological information, and Google Street View pictures made up the dataset. DNN, Pearson correlation-coefficient analysis, and SoftMax classifier are the techniques employed. The methodology uses criminal activity records in specific areas and a fusion of multimodal data to predict the occurrence of crimes. When adopting multimodal data fusion, accuracy is 84.25. DNN is capable of effectively fusing environmental context information with multimodal data. The study makes crime predictions based on prior criminal activities. Strength: It performs well with multi-model data and high-dimensional data. The DNN-Based crime occurrence and prediction cannot be used on insufficient data, which is a weakness. There is this essay that explains the current method that uses a location's neighbours to forecast the type of crime that will occur there next and demonstrates how the suggested system is superior to the current one. In order to find the most effective machine learning to address this issue, this study compares numerous machine learning models and this study is undergone as follows by Alkesh Bhara and Dr. Sarvanaguru RA. K in International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 09 | September 2018.

IV. PROPOSED ALGORITHM

1. Linear Regression :

Why use Linear Regression for this study? Future predictions can now be made scientifically and with high reliability using linear-regression models. The features of linear-regression models are well understood and can be trained extremely quickly since linear regression is a statistical technique that has been around for a very long time.

A method for presenting the relationship between a scalar ward variable Y and at least one relevant component labelled X is called linear regression. Straightforward direct relapse is an instance of one informative variable. Multivariate refers to multiple variables. This regressor is binded into two types, Simple Linear Regression: A Linear Regression calculation is referred to as Simple Linear Regression if only one free factor is used to predict the value of a mathematical ward variable. Multiple Linear regression: Multiple linear regression is the term for a linear regression computation where more than one independent variable is used to predict the value of a mathematical ward variable. In

simple words itsFinding a relationship between the dependent variables (Victim age) and a group of independent variables using multi-linear regression is sort of a mathematical method. In this project we will be using Linear regression by inputting values and then finding the accuracy score at the end.

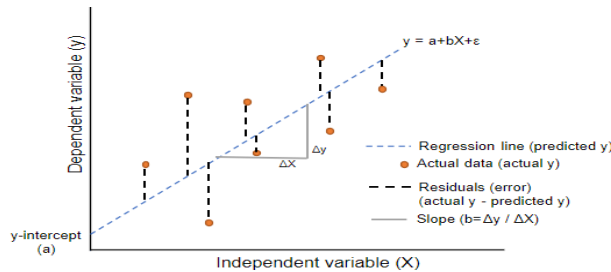


Fig-1: Linear Regression algo predictive analysis.

2. Gaussian Naïve bayes Classifier:

Why use Naïve Bayes Classifier for this study? Naive Bayes classifiers are straight classifiers based on the Bayes hypothesis. It is probabilistic in nature. It measures contingent likelihood, or the probability that something will happen given that something else has actually occurred. For instance, the preceding email is probably spam if the terms "reward" are included.

Because it assumes that all of the dataset's items are free, it is referred to be naïve. In practise, gullible Bayes classifiers frequently outperform expected results despite the freedom postulate, which is frequently ignored. A variation of Naive Bayes that adheres to the Gaussian normal distribution is called Gaussian Naive Bayes.

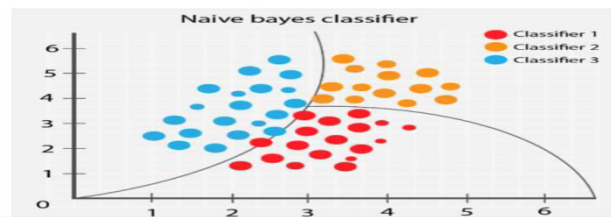


Fig-2: Naïve Bayes Classifier algo predictive analysis.

3. Support Vector machine:

Why use SVM for this study? The hyperplane that produces the greatest degree of class separation for the Wines dataset is chosen via SVM. The great level of precision offered by SVM would be available if input can be divided linearly (Hard Margin). When data cannot be discriminated linearly, all that is required to account for generalisation mistake is to loosen the margin (Soft Margin).

Support Vector Machine a directed Machine Learning algo calculation called is used to focus on portrayal and derive rules from data. It performs admirably well in design affirmation problems. When there are many components and illustrations, it is preferred to use this calculation. Each piece of data is treated as a centre in an n-dimensional space in an SVM model, where n is the number of components, and each component is treated as the value of an orchestrate in the n-dimensional space. Here's how it works:

- (1) The lines or restrictions that successfully request the preparation dataset are the ones that are first examined.
- (2) Next, it selects the particular point that has the best partition from those lines or cut-off points from the closest point or element.

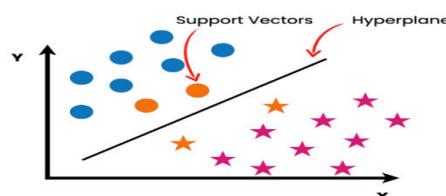


Fig-3: Support Vector machine algo predictive analysis.

4. Random Forest Classifier:

Why use Random Forest Classifier for this study? The aforementioned classification algorithm divides the data into several trees and assists in identifying whether or not a person is at risk for acquiring breast cancer. Additionally, it reduces variance, which enhances the accuracy for the precise prediction about the diagnosis of breast cancer throughout this study. It also dampens the imbalanced datasets and overfitting issue in decision trees.

The random forest algorithm is a directed ML technique that uses a variety of single-choice braids to operate together, as suggested by its name. Every single tree in the Random Woods emits a class assumption, and the class with the greatest number of votes becomes the estimate or forecast of our model. The accuracy score has also been projected as a result of this algorithm's operation on both of the produced datasets. The class with the most votes become the forecast or prediction of our model. Each tree in the Random Forest emits a class expectation. Compared to random forests, decision trees provide a higher level of interpretability. One of the main advantages of random forests is that we don't have to worry as much about choosing suitable hyper-parameter values, but we also have to perform more computations in order to achieve better random forest classifier performance because the ensemble model is very robust and resistant to noise from the individual decision trees and the number of samples in the first training set is used as the sample size for the bootstrap from scikit-learn.

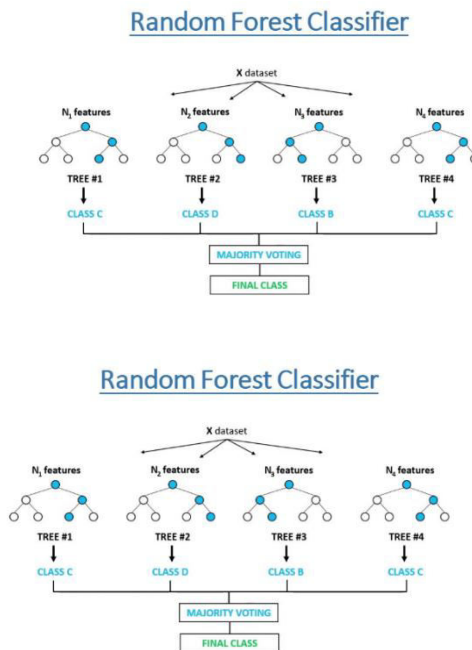


Fig-4: Random Forest Classifier algo predictive analysis.

5. K-Means Clustering:

Why use Random Forest Classifier for this study? To locate groups in the data that have not been explicitly labelled, the K-means clustering algorithm is utilised. This can be used to forecast the level of crime in a certain location and the types of crimes that make up that percentage. Any new data can be quickly allocated to the appropriate group once the algorithm has been performed and the groups have been established.

The computation uses the unlabelled dataset as input, groups it into k groups, then repeats the interaction until it is unable to find the optimal bunches. This calculation should predetermine the value of k. The k-Means algorithm grouping primarily completes two tasks: chooses the best incentive through an iterative cycle for K centre points or centroids. every piece of information and highlights the closest K-Centres. those informational areas that are near the specific specified spot. Since then, each bunch contains datapoints for specific features that it shares and is separate

from other groups. To solve grouping problems in AI or information science, K-Means Clustering is an Unsupervised learning technique. We shall now understand what the K implies bunching calculation is. k-means grouping in Python is implemented together with the bunching calculation, how the computation functions. It is simple to comprehend K-Means.

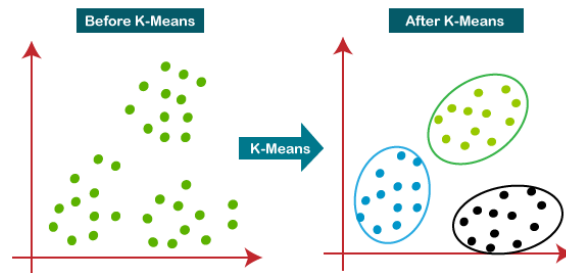


Fig-5:K-Means Clustering algo predictive analysis.

V. EXPERIMENTAL SETUP AND RESULTS

The project's guiding philosophy is to apply machine learning, which is the idea of making predictions about new information based on enormous databases of historical data. It is dependent on the construction of models from impressions known as training data in order to make information-driven decisions on a certain problem or phenomenon. The concept is represented by a flow chart in which the flat portion of the stream is referred to as the assessment or evaluation section and the rising portion of the stream is known as the training component. This study focuses on the model of supervised learning out of several that it contains. According to the supervised learning, the purpose of comparing yields during the information preparation process is to detect accuracy and The capability or model that correctly connects all fed-in outputs to the inputs moved on to the device. In contrast, unsupervised learning does not have a correlation between the input and the outcome. In order to correctly compare input and output, supervised machine learning's main objective is to show how data are fundamentally organised or flow. In order to forecast the outcomes and determine each algorithm's accuracy, I'll be using both supervised and unsupervised machine learning techniques in this research.

PROPOSED METHODOLOGY

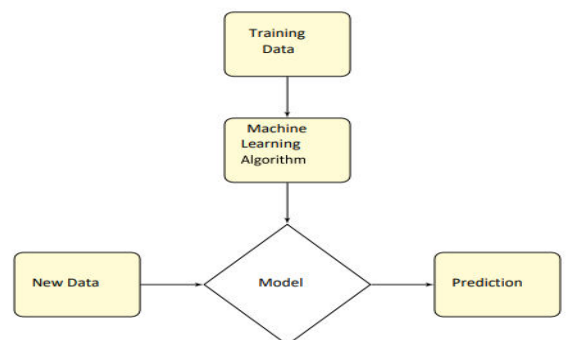
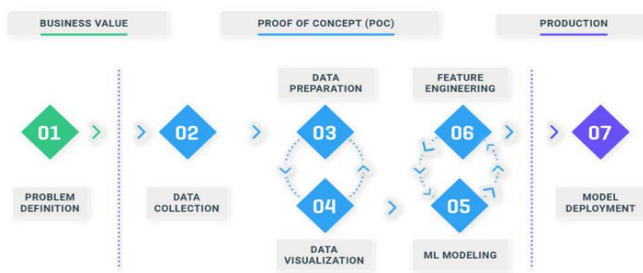


Fig.6(a),6(b):Machine Learning Prediction and analysis process with clear steps.

The four main supervised machine learning techniques I'll be employing for the dataset are as follows::

- Linear regression
- Support Vector Machine (SVM)
- Random Forest Classifier
- Gaussian Naïve bayes Classifier.

The following steps must be taken in order to achieve the desired outcome and prediction:

Data is gathered, filtered, divided into input and output, and tested using training and test variables. The next steps are to normalise (scale) the data, run a classifier or regressor, fit the model, predict the model's output and calculate the accuracy score. These steps through out would be taking the data and filtering it followed by dividing into input and output then main part of this accuracy prediction would be firstly training and testing the test variables secondly

normalizing the data thirdly running a classifier followed by fitting the model finally finding the accuracy score.



Fig-7: Machine Learning prediction accuracy prediction steps involved.

Result for Crime Rate Detection:

Algorithm used: **K-Means Clustering :**

```

In [101]: km_accuracy = cross_val_score(km_clf, X, y, cv=10, scoring='accuracy').mean()
          km_precision = cross_val_score(km_clf, X, y, cv=10, scoring='precision').mean()
          km_recall = cross_val_score(km_clf, X, y, cv=10, scoring='recall').mean()

          print ('Accuracy is for KMeans(Clean data)', km_accuracy)
          print ('Precision is for KMeans(Clean data)', km_precision)
          print ('Recall is for KMeans(Clean data)', km_recall)

Accuracy is for KMeans(Clean data) 0.406989949749
Precision is for KMeans(Clean data) 0.612484172735
Recall is for KMeans(Clean data) 0.513012817885
  
```

Fig.8: K-Nearest Neighbour confusion matrix with accuracy score of 61.2%

Algorithm used: **Decision Tree Classifier :**

```

In [75]: dt_clf = DecisionTreeClassifier(max_depth=3)
          dt_clf.fit(X,y)
          #Predicting
          pred_dt= dt_clf.predict(X)
          dt_accuracy= metrics.accuracy_score(communities_crime_df['highCrime'], pred_dt)
          dt_precision= metrics.precision_score(communities_crime_df['highCrime'], pred_dt)
          dt_recall= metrics.recall_score(communities_crime_df['highCrime'], pred_dt)
          print("Accuracy for DT =",dt_accuracy)
          print("Precision for DT =",dt_precision)
          print("Recall for DT =",dt_precision)

Accuracy for DT = 0.83592574009
Precision for DT = 0.900260190807
Recall for DT = 0.900260190807
  
```

Fig.9: Decision Tree Classifier confusion matrix with accuracy score of 83.5%

Algorithm used:**Gaussian Naïve bayes Classifier:**

```

In [79]: # Using GaussianNB
          gaussian_clf = GaussianNB()
          gaussian_clf.fit(X, y)

          # Applying 10 fold cross validation
          gaussian_accuracy = cross_val_score(gaussian_clf, X, y, cv=10).mean()
          gaussian_precision = cross_val_score(gaussian_clf, X, y, cv=10, scoring='precision').mean()
          gaussian_recall = cross_val_score(gaussian_clf, X, y, cv=10, scoring='recall').mean()
          print("Accuracy for gaussian :", gaussian_accuracy)
          print("Recall for gaussian:", gaussian_recall)
          print("Precision for gaussian:", gaussian_precision)

Accuracy for gaussian : 0.761608040201
Recall for gaussian: 0.692
Precision for gaussian: 0.911799814828
  
```

Fig.10:Gaussian Naïve bayes classifier confusionmatrix with accuracy score of 76%

As we have already applied the above mentioned algorithms to the input data, let's try adding some more algorithms and working with them to finalize a standard usable algorithm that could really cross the levels and help us get the desired accuracy for the crime prediction of a given set of data as I did.

Algorithm used: **Random Forest Classifier:**

```
In [105]: clf = RandomForestClassifier(random_state=100, max_depth=3)

rf_accuracy = cross_val_score(clf, X_d, y_d, cv=10, scoring='accuracy').mean()
rf_precision = cross_val_score(clf, X_d, y_d, cv=10, scoring='precision').mean()
rf_recall = cross_val_score(clf, X_d, y_d, cv=10, scoring='recall').mean()

print ('Accuracy for RandomForestClassifier is', rf_accuracy)
print ('Precision for RandomForestClassifier is', rf_precision)
print ('Recall for RandomForestClassifier is', rf_recall)

Accuracy for RandomForestClassifier is 0.817963874097
Precision for RandomForestClassifier is 0.843663428685
Recall for RandomForestClassifier is 0.872107936508
```

Fig.11: Random Forest Classifier confusion matrix with accuracy score of 87.2%.

Result for Crime Rate Detection and Accuracy Score:

Functioning with the Crime Rate dataset in this project and using machine learning calculations, specifically Linear Regression, Random Forest Classifier, Support Vector Machine, K-Means algos, and Gaussian Nave Bayes, we can reach a conclusion by determining the precision score's accuracy using the Matrix network found in the measurements module from SK. Learn what the precision score is by seeing the table below.

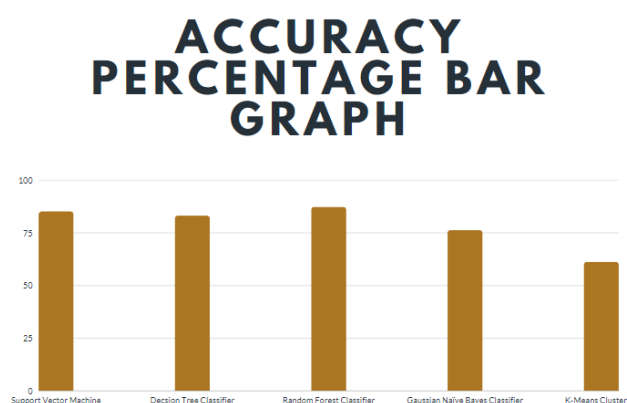


Fig12: Accuracy percentage of algorithms deployed.

The discovery of crime percentages and crime prediction will benefit greatly by computations using advanced Machine Learning algorithms. According to the given input dataset and its high proficiency of accuracy in terms of metrics, the Random Forest classifier is the preferred method or algorithm for the prediction of crime according to my research.

V. RESULT AND DISCUSSION

With the aid of machine learning innovation, completing this project flawlessly, it has become simple to find connections and examples among various pieces of information. The focus of this project's work is predicting the types of wrongdoing that might happen on the off chance that we are aware of their location. We have put together a model using prepared informative index that have undergone information cleaning and alteration using the concept of machine learning. The model's accuracy in predicting the type of wrongdoing with a Random Forest classifier is 87.2. Investigating an informational index is aided by information representation. The diagrams include bar, pie, line, and scatter charts, each with their unique advantages. We created numerous charts and discovered amazing measurements that helped get Chicago violations datasets, which can be used to identify characteristics that can be used to protect

society. In addition to what was already said, if we have a smaller amount of information that is clean and named identically, the beneficiary will use supervised adaptation; but, if we don't have any named information that is surprisingly present in any way, the beneficiary will use unsupervised learning. In addition to the aforementioned, there are countless other machine learning techniques that are truly a part of the pack, such as neural networks, Deep learning that has already been discussed, reinforcement learning, and NLP (Natural Language Processing), which is essentially MACHINE LEARNING (Artificial intelligence) and is on the verge of becoming the future of a significant portion of technology. Every single object will have its own way of getting things done and won't need someone to be able to monitor the specific circumstance, moment, work, or forecast in first, as we all know that machine learning paired with AI and DL is the upcoming change to be presented to the human generation. We may anticipate that its significance and areas of application will continue to grow, and that the more advanced the innovation becomes, the wider its adoption will be. It is now widely used in facial recognition and image recognition in general, including in content idea for websites like Instagram, YouTube, Netflix, and Pinterest calculations. It is also used in record handling, nude channels, and observation. Frameworks for simulated intelligence are currently and As they are given datasets to analyse and learn from, simulated intelligence frameworks are now developing novel anti-toxins that are designed to treat specified illnesses. In the future, it very likely may be possible to integrate AI into medical services much more, using it to assess patients and provide treatment for their specific ailment. When we go toward fully automating our transportation, we will inevitably see the rise of self-driving cars as ML frameworks get sophisticated enough to handle complicated decision-making much like people do. ML is currently on track, generating in many areas of our lives, and will only continue to coordinate farther and deeper as the technology of ML advances. There are so many uncountable There are countless additional sectors, such as custom design, quantum computing, and others. Because of all the calculated activities, engineering by advances and equipment, or analogy-based applications that are currently open swiftly, ML, if or when used, might result in a result that may exceed our expectations. As a result, the undertakings may also align their vested party. Similarly, we can assert that machine learning represents the future.

REFERENCES

- [1]. Yu-Yueh Huang, Cheng-Te Li and Shyh-Kang Jeng, "Mining Location-based Social Networks for Criminal Activity Prediction", Proceedings of 24th IEEE International Conference on Wireless and Optical Communication, pp. 185-190, 2015.
- [2]. Rasoul Kiani, Siamak Mahdavi and Amin Keshavarzi, "Analysis and Prediction of Crimes by Clustering and Classification", International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No. 8, pp. 11-17, 2015.
- [3]. Chung-Hsien Yu, Max W. Ward, Melissa Morabito and Wei Ding, "Crime Forecasting using Data Mining Techniques", Proceedings of 11th IEEE International Conference on Data Mining Workshops, pp. 779-786, 2011.
- [4]. Arunima S. Kumar and Raju K. Gopal, "Data Mining based Crime Investigation Systems: Taxonomy and Relevance", Proceedings of Global Conference on IEEE Communication Technologies, pp. 850-853, 2015.
- [5]. Arunima S. Kumar and Raju K. Gopal, "Data Mining based Crime Investigation Systems: Taxonomy and Relevance", Proceedings of Global Conference on IEEE Communication Technologies, pp. 850-853, 2015.
- [6]. Kevin Sheehy et al., "Evidence-based Analysis of Mentally 111 Individuals in the Criminal Justice System", Proceedings of IEEE Systems and Information Engineering Design Symposium, pp. 250-254, 2016.
- [7]. 10 Emerging Technologies That Will Change Your World, Available at: http://www.rle.mit.edu/thz/documents/10_emerging_tech.pdf.
- [8]. Marchant, R., Hana, S., Clancey, G., Cripps, S.: Applying machine learning to criminology: semi parametric spatial demographic Bayesian regression. Security Inform. 7(1) (2018).
- [9]. E. W. T. Ngai, L. Xiu, and D. C. K. Chau. "Application of Data Mining Techniques in Customer Relationship Management: A Literature Review and Classification," Expert Systems with Applications, pp. 2592–2602, 2008.
- [10].] R. Bermudez, B. Gerardo, J. Manalang and B. Tanguilig, III. "Predicting Faculty Performance Using Regression Model in Data Mining," Proceedings of the 9th International Conference on Software Engineering Research, Management and Applications, pp. 68-72, 2011.



INNO  SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

 **doi**[®]
cross **ref**

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details