IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-8, Aug 2019 E-ISSN: 2347-2693

Identification of Accurate Classification Technique for Crime Investigation Using Ensemble Approach

Sadhna sharma[1*], Sanjiv sharma[2]

[1,2]Dept. of CSE/IT, Madhav Institute of Technology and Science, Gwalior, India

Corresponding Author: Sadhnas745@gmail.com, Mob: +91-8817406942

DOI: https://doi.org/10.26438/ijcse/v7i8.137143 | Available online at: www.ijcseonline.org

Accepted: 08/Aug/2019, Published: 31/Aug/2019

***Abstract*—** Recently, it's observed that the crime is increasing across the world very rapidly and some technique is required for analysis of the crime data. Analysis of the crime data can be done through data mining (DM). DM techniques are applied to crime data for predicting features that affect the high crime rate. Using the method of data mining on previously collected data for predicting the features responsible for the crime in a locality or area, the Police Department and the Crimes Record Bureau Police Department may take the required measures to reduce the likelihood of the crime. In the current work, a new machine learning ensemble algorithm is opted for predicting feature that affects a high crime rate. It helps the police and citizens to take necessary and required action in decreasing the crimes rate. The ensemble algorithm can predict more accurate and significant features with higher accuracy and efficiency.

***Keywords*—** Crime investigation, Crime Prediction, Crime Prediction, Data Mining, Ensemble approach

I. INTRODUCTION

Crime Investigation is an analysis of a crime using facts. It includes studying the crime scene and the proof collected systematically. Many social, temporal, spatial and demographic factors help police assess crime. When a criminal investigation is complete, a lot of information is gathered & this information can be helpful. It can enable police agencies to better respond to such crimes. Criminal activities across the world resulted in a menace in society. The law enforcement organization generate a volume of criminal data each year and it is a major challenge for them to investigate the data for future crime avoidance decisions. The crime data has many features including information about immigrants, race, sex, population, demographics and so on. Analyzing the data not only helps in recognizing features responsible for high crime rate but also helps in taking necessary actions for the prevention of crimes. Data mining provides powerful techniques and algorithms to analyses data and extract important information from it. Data mining generally involve two categories with respect to the data to be mined that includes Description [1] mining and Classification with Prediction [2]. Description mining usually is the mining of association rules, frequent Patterns, clusters, or correlations. Classification and Prediction involve classifying a class label for data using probability equations and predicting any feature using numeric measures accordingly.

The aim is to predict top most features with the accurate predictive model that affects the high crime rate which will eventually help police or law enforcement makers take necessary action. The literature reviews in section II, Methodology in section III, results and discussions in IV, and conclusion in section V.

II. LITERATURE REVIEW

Sharma [3] proposed an idea that depicts zero crime in the general public. To identify suspicious crimes, they have focused on the importance of DM innovation and have structured an active application for that reason. He assumed a device that relates an upgraded decision tree approach to identify doubtful messages regarding crimes. An ID3 improved algorithm is associated through an advanced highlighting technique and asset significance factor to create a faster & better decision tree which is dependent upon data entropy which gets unevenly with the progress of creating an informational index from some sections. . They proposed another calculation which is a mix of improved component determining technology for improved ID3 characterization calculation and improved effectiveness of calculations.

Hamdy et al. [4] depicted a methodology dependent on the general population's collaboration with informal communities and versatile utilization. Additionally, their work presented an identifying model for doubtful conduct, dependent on informal organization feeds & it not just

portrays another strategy utilizing the social connection of individuals, in any case, their work proposes another framework to help crime investigation make quicker and exact choices. The suspicious development of the element can be resolved utilizing the grouping of derivation rules.

Bogahawatte and Adikari [5] proposed a methodology wherein they featured the utilization of information mining procedures, bunching, and grouping for compelling examination of violations and criminal distinguishing proof by building up a framework named ICSIS i.e. Intelligent Crime Investigation System which could recognize a criminal put together up with respect to the proof gathered from the wrongdoing area. The author utilized bunching onto distinguish the wrongdoing designs which are utilized to perpetrate violations knowing the way that every wrongdoing has certain examples.

Agarwal et al. [6] author utilized the rapid miner device for investigating the crimes rates utilizing various information mining systems. Crime investigation through K-Means Clustering calculation. So, principle goal of the crime investigation examination work is to separate the wrongdoing designs, foresee the wrongdoing dependent onto the spatial appropriation of present information and identification of wrongdoing. So, mainly their examination incorporates the following manslaughter wrongdoing rates starting with one year then onto the next.

Yerpude et al. [7] authors utilized Naïve Bayes, Linear Regression & Random Forest for recognizing factors which affect the high crime rate & than author compared their performance. As a result of the comparison the authors conclude that the among all the approaches Random Forest gives better performance with 83.39% accuracy.

Bruin et al. [8] author utilized an approach that is utilized to decide the criminal's group depends upon criminal careers. Every year profiles of Criminal are removed from the database & calculated profile distance. The distance matrix with frequency value is created to create a cluster employing the clustering naive algorithm.

Chen et al. [9] author introduced an all-purpose structure for crime DM which pulls on experience picked up within Cop connection venture within specialists at Arizona & their work for the most part centers around demonstrating the connections between wrongdoing forms & the connection between the criminal associations. The author utilized an idea space method that would remove criminal after the occurrence rundowns.

Sharma et al. [10] author presented the use of DM crime detection techniques such as association rules, sequential patterns, clustering, and others and it also presented the comparative study of techniques along with its strength and weakness.

III. METHODOLOGY

This section describes the dataset used in the experiments and the boosting ensemble model namely: XGBoost and CatBoost. The aim is to extract top most features that affects the high crime rate which will eventually help police or law enforcement makers take necessary actions. In the previous work, a decision tree named Random forest (ensemble classifier) was used. However, for certain datasets with noisy Classification task Random Forest have been witnessed to overfit. Unlike decision trees, random forest classification are hard to interpret for humans. Random forests are biased towards those features with greater concentration for information. Thus, variable score of importance are not reliable through random forests. So, due to these limitations and issues faced by it. We opted a new machine learning ensemble approach to overcome the deficiencies of the existing classifier. The flow diagram of opted approach is shown in figure1.

A. DATASET

In this paper, for performing crime investigation, the Communities & Crime dataset are from UCI repository has been used which consists of crime data in Chicago. It includes features affecting crime rate like population, race, sex, immigrants etc. The 'Per Capita Violent Crimes' attribute to be predicted which was pre-determined in the data utilizing population & the aggregate of Variables of crime regarded in the United States as violent o ffenses: rape, assault, homicide, theft and assault.

B. DATA PRE-PROCESSING

Pre-processing of data is a DM approach which involves raw data transforming into format which is understandable. Often the data is inconsistent, unstructured, has missing values, and lack in definite behavior that gives many errors. Therefore, it needs to be cleaned, integrated, transformed, and hence reduced. Cleaning fills in the missing values and removes noise. Integration take the data cubes or chunks together using multiple databases. Transformation uses normalization and aggregates the data and Reduction helps in decreasing the volume of data keeping similar analytical results.

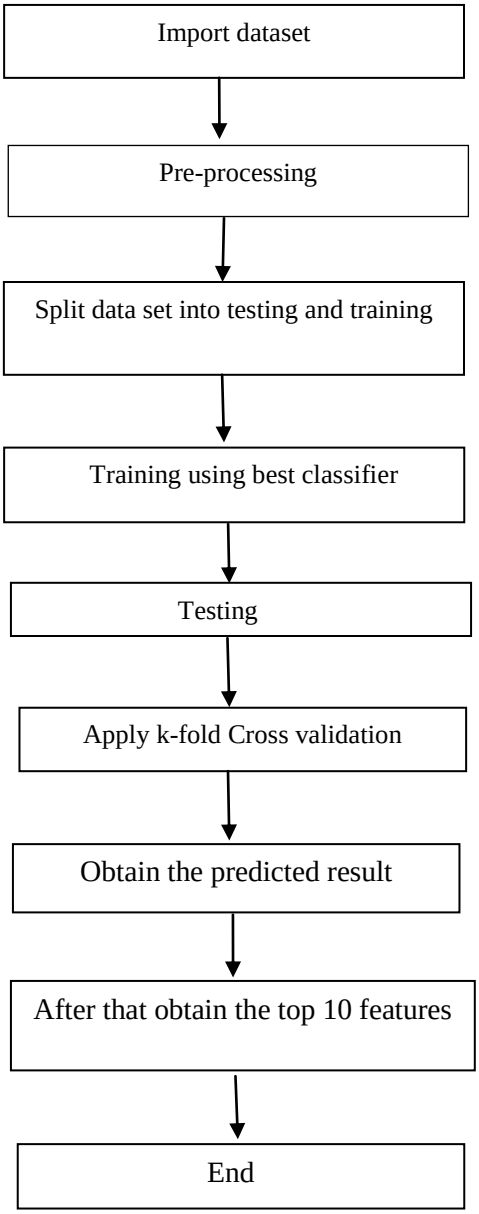

Figure 1: Flow Diagram of current work

.

For clean data, removal of missing values was needed to get an appropriate crime data set. For dirty data, the missing data of a feature were converted into median value of that feature. For predicting feature, 'Per Capita Violent Crimes', a new column called 'High Crime' was created that had a value '1' for Per Capita Violent Crime' greater than 0.1 and '0' otherwise. The threshold of 0.1 was decided upon manual analysis of data by view-through process. All the features had to be predicted using this target feature 'High crime'.

The threshold of 0.1 was decided upon manual analysis of data by view-through process. All the features had to be predicted using this target feature 'High crime'. Clean and dirty data sets were converted into different data frames and the target feature 'High crime' was assigned to a variable 'Target' and the remaining features to 'Features'.

C. PERFORMANCE METRICS

Cross Validation Score: In Cross validation each record is exactly once for testing & used the same number of times for training. For example, in the 2-fold cross verification method, selecting one of the subsets for training & the second to test. Then swap the roles of subset so that the previous training set will be set to the test & vice versa. For analysis, 10-fold cross validation method has been used thereby ruling out the possibility of over fitting the data. Hence, CV Scores of precision, accuracy, recall & F1 Score have been considered for elevating the performance of our model.

Accuracy, Precision, Recall & F1 Score: Performance of a model is measured by reflection of well observed actual events. While training any model, a labelled data set that includes the actual values to be predicted is considered. This introduced the concepts of a confusion matrix. It gives a relation between an actual and predicted class.

True positivity & true negativity are observations that have been correctly predicted, and therefore they are shown in green. The words "false positives" & "false negative" can be confused. False negatives are the comments in which the actual event was positive. The way to think about this is that words refer to observations & not actual events. Therefore, if the word begins with "wrong", then the actual value is the totally opposed of the word which follows it.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

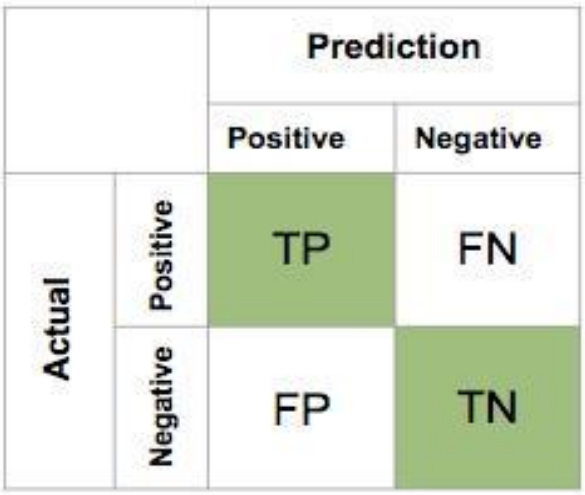

Figure 2: Confusion Matrix

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

XGBoost

XGBoost stands for extreme Gradient Boosting
XGBoost is the application of gradient boosted decision trees for efficiency & velocity. This is the basically the implementation of gradient boosting machines made by Tianqi Chen, with the contribution of many developers. To develop high-performance & fast gradients boosting tree models are develop through XGBoost library. on a range of difficult machine teaching tasks XGBoost is receiving the best performance. XGBoost offers a tree boosting which is parallel (& also called as GBM, GBDT) which resolves many problems faster & more accurately.

Proposed Algorithm

Step 1. Collect dataset Communities and Crime from UCI Repository:
Step 2. Convert text file into csv format
Step 3. Start
Step 4. Import dataset
Step 5. Clean the dataset by removing missing values
Step 6. If
Violent Crimes per Pop>0.1
High Crime =1
0, Otherwise
Step 7. Dataset split into training (75%) and testing (25%)
Step 8. Training using XGBoost classifier
Step 9. After that apply the cross validation
Step 10. Obtain the predicted result
Step 11. After that obtain the top 10 features
Step 12. End

Table 1: Performance Measures- XGBoost Clean Data

XGBoost classifier-clean data	10-fold Cross-Validation (%)
Accuracy	85.20
Precision	87.63
Recall	87.76
F1 score	88.26

Table 2: Performance Measures- XGBoost dirty Data

XGBoost classifier-dirty data	10-fold Cross-Validation (%)
Accuracy	83.54
Precision	87.40
Recall	85.84
F1 score	87.14

Feature Importance in scikit learn measures the importance of feature in relation to target feature. Therefore, the top 10 most important features that helped in crime investigation, based on this score were fetched. Higher the importance score, more significant is that feature in contribution to crime. The same procedure was applied on both cleaned as well as uncleansed data and the results were as follows:

The top 10 features extracted according to the feature importance scores were:-

PctKids2Par: percentage of kids in family housing with two parents.

racePctWhite: percentage of population that is Caucasian.

Pctllleg: percentage of kids born to never married.

Numllleg: number of kids born to never married.

racePctHisp: percentage of population that is of hispanic heritage.
MalePctDivorce: percentage of males who are divorced.

NumImmig: total number of people known to be foreign born.
PctFam2Par: percentage of families (with kids) that are headed by two parents.

OwnOccLowQuart: owner occupied housing - lower quartile value.
pctWlnvlnc: percentage of households with investment / rent income in 1989.

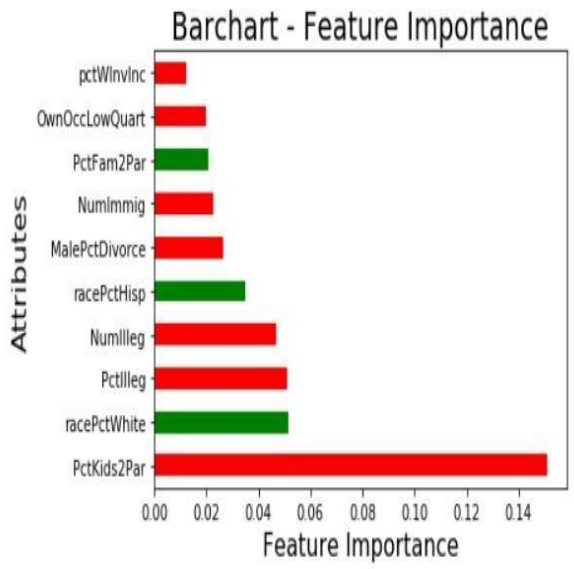

Figure 3: Bar Graph of Feature Importance of XGBoost Classifier

CatBoost

We have used a machine learning (ML) classifier called catbost. This is a newly open sourcing ML algorithm. So, It could work through diverse kind of data which help in solving problems with ordered features while supporting hierarchical features. Category attributes are a variable that can take a certain number of a limited, and usually fixed values (categories). For example, example Animal can be set to 'Feature, Cat', Dog ',' Rat ', etc., which today faces many problems. So, moreover, it provides the good class accuracy. The name "catbost" originates through two words "category" & "boosting".

Step 1. Collect dataset Communities and Crime from UCI Repository:
Step 2. Convert text file into csv format
Step 3. Start
Step 4. Import dataset
Step 5. Clean the dataset by removing missing values
Step 6. Replace “Per Capita Violent Crimes” with a new column “High Crime”
Step 7. If
Violent Crimes per Pop >0.1
High Crime =1
0, Otherwise
Step 8. Dataset is split into training (75%) and testing (25%)
Step 9. Training using CatBoost classifier
Step 10. After that apply the cross validation
Step 11. Obtain the predicted result
Step 12. After that obtain the top 10 features
Step 13. End

Table 3: Performance Measures- CatBoost Clean Data

CatBoost classifier-clean data	10-fold Cross-Validation (%)
Accuracy	84.13
Precision	88.76
Recall	86.59
F1 score	87.10

Table 4: Performance Measures- CatBoost dirty Data

CatBoost classifier-dirty data	10-fold Cross-Validation (%)
Accuracy	81.54
Precision	86.40
Recall	85.84
F1 score	87.26

Feature Importance in scikit learn measures the importance of feature in relation to target feature. Therefore, the top 10 most important features that helped in prediction of crime, based on this score were fetched. Higher the importance score, more significant is that feature in contribution to crime.

The top 10 features extracted according to the feature importance scores were:-

racePctWhite: percentage of population that is Caucasian.

PctKids2Par: percentage of kids in family housing with two parents.

PctFam2Par: percentage of families (with kids) that are headed by two parents.

Numllleg: number of kids born to never married.

TotalPctDiv: percentage of population who are divorced.

PctYoungKids2Par: percent of kids 4 and under in two parent households.

FemalePctDiv: percentage of females who are divorced.

PctPersDenseHous: percent of persons in dense housing.

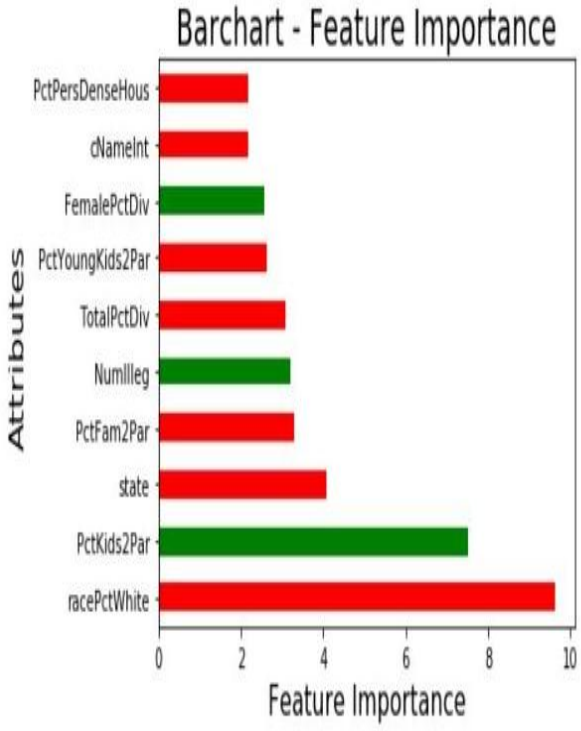

Figure 4: Bar Graph of Feature Importance of CatBoost Classifier

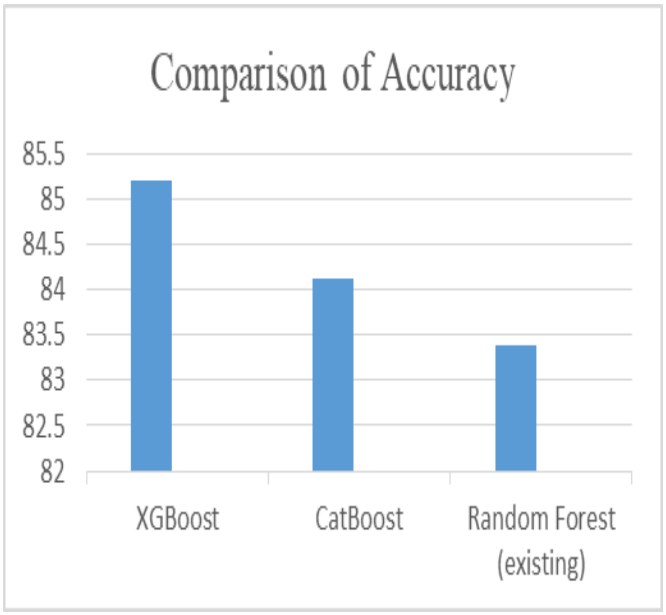

Figure 5: Comparison of Accuracy

The above figure 5 shown the comparison Accuracy graph of the classifiers namely XGBoost, CatBoost and Random Forest. Accuracy of XGBoost is 85.20 %, CatBoost accuracy is 84.16% and Random Forest is 83.39%.

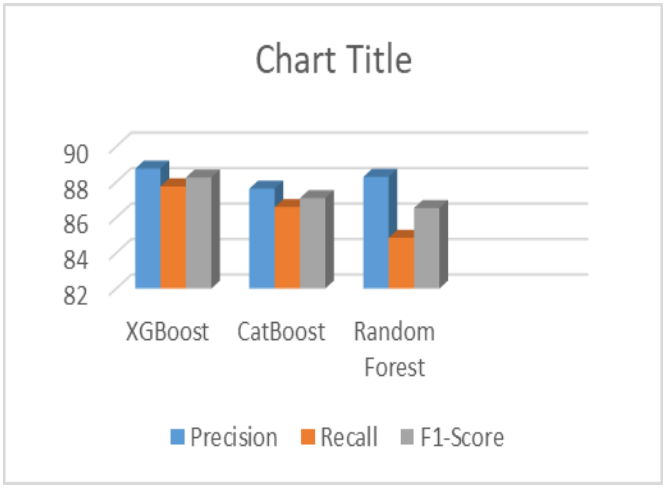

Figure 6: Performance Evaluation

The above figure 6 shown the performance evaluation of current classifier i.e. XGBoost and CatBoost with existing classifier i.e. Random Forest.

Table 5: Comparative Result of Classifiers

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
XGBoost	85.20	88.76	87.76	88.26
CatBoost	84.13	87.63	86.59	87.10
Random Forest	83.39	88.30	84.86	86.54

The above table 5 shows the comparative result of classifiers namely XGBoost, CatBoost and Random Forest (existing) in terms of accuracy, precision, recall and F1 score. Accuracy measures the performance of each model that gives percentage of features that are predicted correctly among total number of features, Precision which is defined as number of positive features classified by the model that are positive and Recall that gives number of positive features classified correctly by the model. Also, F1 Score is a harmonic mean of precision and recall for balancing out both has been taken as a measure of performance.

V. CONCLUSION

The paper concludes that XGBoost Classifier is able to predict more balanced results with respect to accuracy, precision, recall and F1 score out for prediction of 'High Crime' feature compared to that of Random Forest. features having high importance scores that proved to be highly predictive of 'High crime' features are "PctKids2Par", "racePctWhite", "Pctllleg" , "Numllleg" , "racePctHisp" , "MalePctDivorce" , "NumImmig" , "PctFam2Par", "OwnOccLowQuart" , "pctWlnvlnc" using XGBoost

Classifier model.This will reduce the complexity of the crime dataset and provide benefit to the citizen to utilize their resources efficiently & take appropriate actions to reduce criminal activities in the society.

REFERENCES

[1] OdedMaimon, LiorRokach, "The Data Mining and Knowledge Discovery Handbook", Springer 2005, Page 6

[2] Han, Jiawei et.al "Data Mining", Second Edition, Page 285

[3] Mugdha Sharma, "Z-Crime: A Data Mining Tool for the Detection of Suspicious Criminal Activities based on the Decision Tree", International Conference on Data Mining and Intelligent Computing, pp. 1-6, 2014

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

[5] Kaumalee Bogahawatte and Shalinda Adikari, "Intelligent Criminal Identification System", Proceedings of 8th IEEE International Conference on Computer Science and Education, pp. 633-638, 2013.

[6] Jyoti Agarwal, Renuka Nagpal and Rajni Sehgal, "Crime Analysis using K-Means Clustering", International Journal of Computer Applications, Vol. 83, No. 4, pp. 1-4, 2013.

[7] Prajakta Yerpude, Vaishnavi Gudur. "Predictive modelling of crime dataset using data mining". In international journal of data mining & knowledge management process, vol.7, pp.43-58, 2017.

[8] Jeroen S. De Bruin, Tim K. Cocx, Walter A. Kosters, Jeroen F. J. Laros and Joost N. Kok, "Data Mining Approaches to Criminal Career Analysis", Proceedings of 6 th IEEE International Conference on Data Mining, pp. 1-7, 2006.

[9] H. Chen, W. Chung, J.J. Xu, G. Wang, Y. Qin and M. Chau, "Crime Data Mining: a General Framework and Some Examples", Computer, Vol. 37, No. 4, pp. 50-56, 2004.

[10] Sadhna shrama, sanjiv sharma, " a compartive study of crime investigation using data mining approaches", International Journal for Research in Applied Science & Engineering Technology,Vol.7,pp. 2073-2079,2019.

Authors Profile

Sadhna sharma, she received B.E (Computer science and engineering) from Maharana Pratap College of Technology, Gwalior in 2016, pursuing M.Tech (Computer science and engineering) from Madhav institute of technology and science, Gwalior.

Sanjiv sharma, work as an assistant professor in department of computer science engineering and information technology in Madhav institute of technology and science, Gwalior. He has 12 years of teaching and research experience. He received his B.E from Pt. Ravishankar University, Raipur and M.Tech from RGPV, Bhopal and PhD from Bansthali VidhyaPeeth, Jaipur. His current research interests include Social Network Analysis, Data Mining, Network Security and Ad hoc Network and Mobile Computing and their interdependency.

