e-ISSN:2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

ONLINE CRIME MANAGEMENT AND REPORTING SYSTEM, AND PREDICTION OF CRIMES USING MACHINE LEARNING AND DATA SCIENCE

Bubai Kumar Bera[*1], **Chetan Dange**[*2], **Saurabh Singh**[*3], **Akshada Nighut**[*4]

[*1,2,3,4]Department Of Computer Engineering ,Sinhgad Institute Of Technology,
Lonavala , Maharashtra, India.

ABSTRACT

Problem was that people get tired by going here and there for getting justice. As we are aware that crime rates are rapidly increasing in our society .capable of registering FIR online, shows investigation update & predict the pattern of crimes etc. The tendency to predict the future crimes based on the area, pattern and time can serve as a important source of knowledge for them from strategic or tactical perspectives. The vast onerous faced by most of the law enforcement and intelligence organizations is efficiency and accurately analyzing the growing volumes of crime related data. The recommended system is a web-based system which comprises of crime analysis techniques such as hotspot detection areas, crime comparison and crime pattern visualization

Keywords: Crime Reporting, Incident-Based, Victims, Criminals.

I. INTRODUCTION

The project titled as “Online Crime Reporting and prediction of crime using machine learning “ is a web-based application. This software can be used for reporting online crimes, complaints, missing persons, show most wanted person details, show snatchers, show unidentified dead bodies, stolen vehicles. Any clients can connect to the server. Firstly user makes their login to sever to show their availability. The server can be any Web Server. Crime can be splited into a many types such as crime against properties (bulgary, theft, and robbery) and crime of aggression (homicides, assaults and rape).There are quite often reasons for crime analysis like to identify general and specific crime trends, patterns, and series in an current manner to maximize the usage of limited law enforcement resources, to access crime problems locally, regionally, nationally within and often law enforcement agencies, to be proactive in detecting and preventing crimes and to meet the law enforcement needs of the changing society. The potential to analyze the amount of crime data without having computational support will put strain on human because human mechanism is incapable of comprehending with millions of data. introduction to Modern society is characterized by increasing levels of global social mobility and uncertainty relating to certain levels of risk posed by internal and external security threats. Within this climate securities driven by various technologies is increasingly being used by governments, corporate bodies and individuals to monitor and reduce risk (Lyon, 2004). There has been an acceptance that the criminal justice system is limited in its capacity to control crime which has led to the exploration of other avenues for tackling crime (Zedner, 2003) and this has provided a market for private companies to push forward the growth of technological security innovations. It is not controlled by any means and perpetrators getaway with their misconducts as most are not likely to be found after committing serious offence. However, various analysts have resolved into using technologies to aid investigation of crime and proper solutions that will not only reduce crime but also monitor individuals with their immediate environments.

II. METHODOLOGY

Existing System:Residents (Citizens) can't get the data and the present status of the culprits of all urban areas Crime information mining is the use of information digging strategies for Crime investigation. Crimes can be isolated into subcategories dependent on various criteria. In eight crime classes are given. They are criminal traffic offenses, sex violations, burglary, misrepresentation, and fire related crime, medicate offenses and vicious crimes [8]. There were numerous endeavors to investigate distinctive sorts of violations utilizing mechanized strategies yet there is no bound together structure portraying how to apply those procedures to various crime types. In, they have utilized a structure which incorporates a connection between the crime information mining procedure and crime type attributes. There are a few existing frameworks which use crime data digging systems for crime examination, for example, territorial crime investigation program, information digging structure for crime design distinguishing proof and opiates organize in Indian police division.

III. MODELING AND ANALYSIS

Simplified UI for Filing :-FIR and Hassle Free Environment Simplified UI for Filing FIR and Hassle Free Environment Machine Learning will be used for the following purposes Hotspot Area Detection Crime Detection Criminal identification and prediction Crime Verification Hotspot Area Detection will be done by using various Supervised Learning Crime Detection and Verification will be done by using various Clustering techniques. Criminal Identification and prediction will also be done using various clustering techniques. Crime verification will also make use of unsupervised techniques and hence, will give the accuracy of our prediction. **Admin Module:-**View and Reply User Complaint. View and Reply User Crimes. Add, Delete and Hide Latest Hot news. View and Delete User's FeedbackAdd, Delete and View Most wanted Persons. Add, Delete and View Missing Persons. Add and View Criminal Registration. Send Message to user Change password.

User Module:-Add Online Complaints. Check Complaint Status. Edit Complaints. Add Missing Person Report. Ask Questions. Send Messages. Give Feedbacks change Password. Visitor Module view Hot News. View Missing Persons. View Crime Types.View Faq'S. View Most Wanted. View Help Lines. View Safety Tips. View Stolen.

Dataset Requirements:-As the objective of our project also involve Machine Learning, so, this project will require a dataset to work on. Therefore, we are going to use the Past Crime Record dataset for this purpose, which is made public by the government free of cost, so this objective is also economically viable for our team.Dataset of Crimes in last Decade – by National Crime Records Bureau(NCRB).

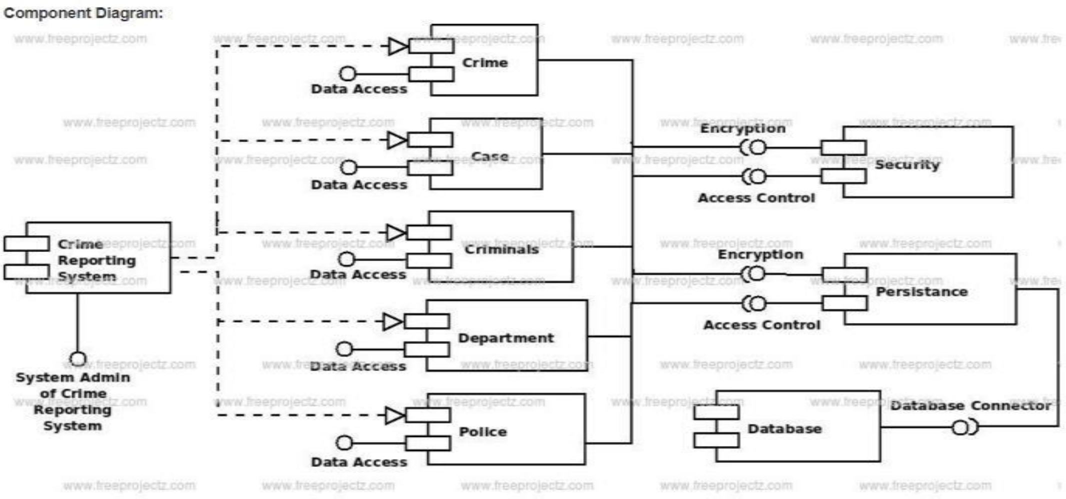
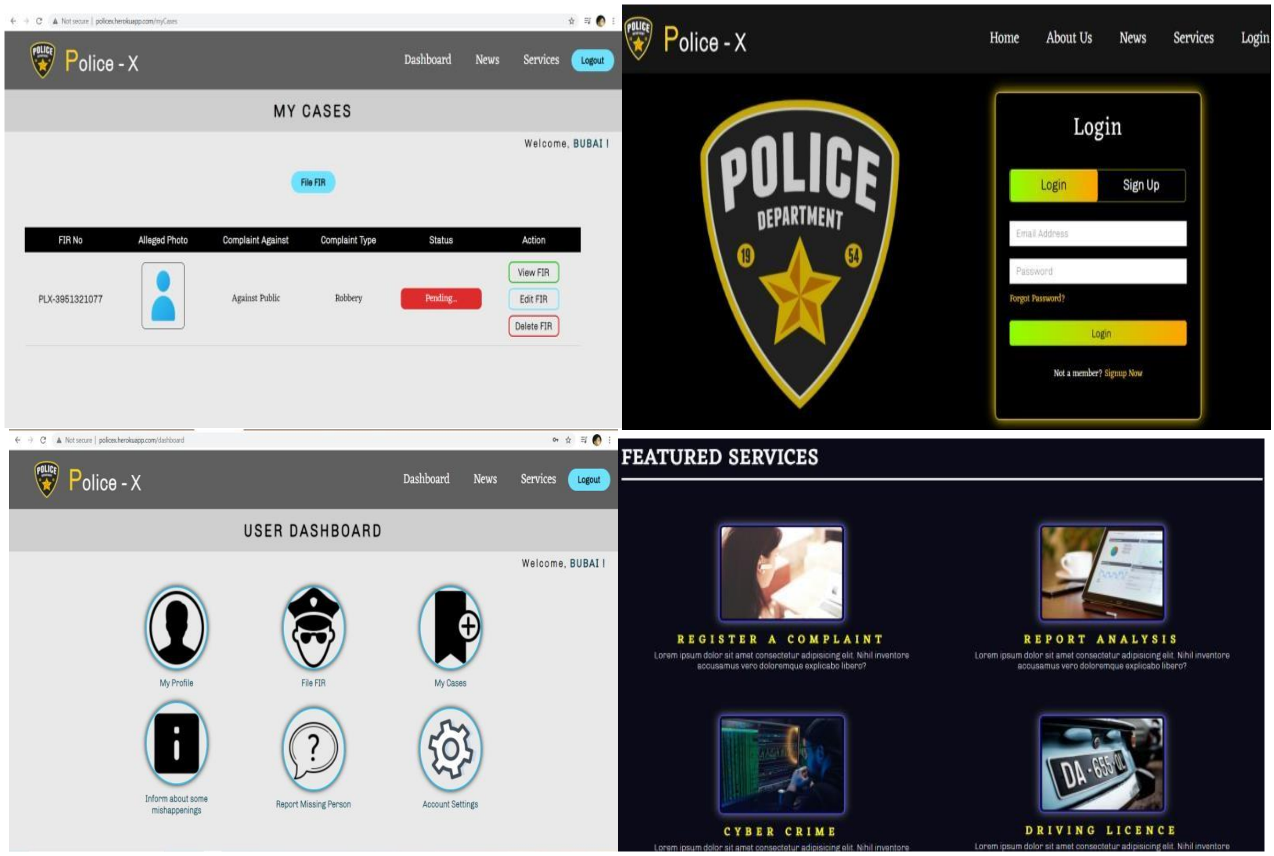


Figure 1: Component Diagram

Expected Outcomes:-The manual system has some drawbacks which can be overcome by using the web based software. The following reasons explain by it is needed. Citizens are not needed to go to police stations to see the criminals information. they can directly see information on site. The proposed system consists of five major components. They are classifier, duplicate detector, data base handler, analyzer and graphical user interface. Classified articles will be stored in data base for entity extraction. The main purpose of this module is to identify exact/near duplicates of newspaper articles and remove them from the database. Analyzer module will perform crime analysis operations on processed crime articles. Such as Hot spot detection , Crime comparison, Crime pattern visualization. This module is used to visualize crime statistical details of the previous years. Conclusion The crime analysis is sensitive domain where efficient for prediction and classification to analyze the increasing numbers of crime data. Hence, the crime prediction methods will be evaluated and analyzed by the systematic tool in crime analysis. The biggest challenge facing by many law enforcement is how to efficiently and accurately analyzing the increasing volumes of crime data. This research work focuses on reviewing a crime prediction analysis tool for many scenarios using different crime prediction methods which can help law enforcement to efficiently handle crime incidents. Therefore, a crime analysis should be able to identify the crime patterns as fast as possible and in an effective manner for future crime detection.



FIR No	Alleged Photo	Complaint Against	Complaint Type	Status	Action
PLX-3951321077		Against Public	Robbery	Pending...	View FIR Edit FIR Delete FIR

Lorem ipsum dolor sit amet consectetur adipisicing elit. Nihil inventore accusamus vero doloremque explicabo libero?

Lorem ipsum dolor sit amet consectetur adipisicing elit. Nihil inventore accusamus vero doloremque explicabo libero?

Lorem ipsum dolor sit amet consectetur adipisicing elit. Nihil inventore

Lorem ipsum dolor sit amet consectetur adipisicing elit. Nihil inventore

Figure 2: Screenshots of Project

Scope Of The Project:-As a general the scope of our system is as follows:Ensure data's accuracy. anyone can report crime to the crime authority. Reduced manual data entry, greater efficacy and better services. Generating crime reports. Assingning the police officer to the concerned Zones. Action Taken As for any problem there is always a solution. Some solutions for problems are:Constant upgrading of the system as per the user requirements and the issues faced by them. Computer Failure and other such hardware issue, for this our project is capable of multidevice access, in case of hardware failure, the same data can be accessed from other devices. **Future Scope:-** The scope of the project includes that what all future enhancement can be done in this system to make it more feasible to us:In future Users can view the progress of their complaint online. By using more adavance technology user can view the case details and progress of the complaints on their mobile phones. Beteer graphics can be designed to make it more user friendly and understandable. For future works involving crime prediction algorithm in crime analysis field, it is hold that researchers can propose more hybrid methods between crime prediction algorithm in order to get best finding data and results. In addition, the entity extraction module can be improved by incorporating more rules which will improve the accuracy and the comprehensiveness of the entity extraction process.

IV. RESULTS AND DISCUSSION

Existing System:- Many researchers have gone through this problem regarding the criminal cases being unsolved for a long period. They proposed different crime prediction algorithms. In all these models the accuracy will surely vary depending on the data set and the features or attributes we select during data pre-processing. The Crime prediction which are done on the Mississippi crime data set where models like linear regression and Decision stump model are used gave a result of 83%, 88% and 67% respectively. Even though accuray of the predictions may vary accordingly because it is discovered that many machine learning algorithms are implemented on data sets consisting of different places having distinctive features, so predictions are changing in all cases. In the data set used is from kaggle.com and have selected models namely logistic regression, K- Nearest Neighbours (KNN), Decision tree classification, Bayesian methods. Data pre-processing is done by dropping the null values and filling the unknown values [2]. The accuracies obtained by using the above machine learning algorithms are as follows 78.73% for KNN, 78.60% for decision tress classifier 31% for Support Vector Clustering(SVC) and 64.60% for Gaussian Naïve Bayes. The model is trained by cleaning and pre-processing the data and obtaining a best accuracy of 78.73%.

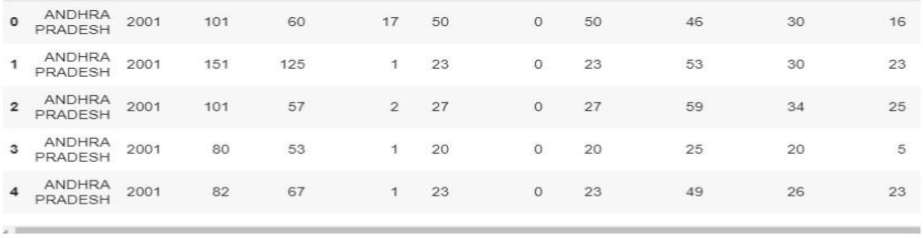
Proposed system: - The proposed system is made on the basis of the research work that is done by going through various such documentations. Majorly all of the crimes are predicted based on the location and the types of crimes that are occurring in those areas. On surveying previous works, Linear Regression, Decision Tree and Random Forest tend to give good accuracy so these models are used in this paper to predict crimes. The dataset used in this paper is from da ta.world.com. The data set contains different types of crimes that being committed in India according to the state and year respectively [4]. This paper takes types of crimes as input and gives the area in which crimes are committed as output. The data pre-processing involves data cleaning, feature selection, dropping null values, data scaling by normalizing and standardizing. After data pre-processing the data is free of null values which m ay alter the accuracy of the model significantly and feature selection is used to select only the required features that won't affect the accuracy of model. After data pre-processing the models chosen i.e., Logistic Regression, Decision Tree and Random Forest are trained by splitting the data into as train and test data. As the expected value required is a categorical value classification models are used here. Here we have use Python language for data prediction.

Data collection: - The data set used is the crimes that are committed in India during the year 2001-2018 which is available In the dataset world. It consists of features like the states of India and the districts of every state where the crimes are committed. It also gives the type of crimes that are being committed such as kidnapping, raping, robbery, theft, criminal breach of trust, etc.

Data Pre-Processing: - The first and major step in data Pre-Processing is done in order to remove the null values and the features or attributes that are unnecessary. Nine thousand entries are present in the dataset that is being used in this. All the null values are removed. To use the data consisting of string values there is a need to convert that string values to float to use the machine learning algorithms efficiently. This conversion of data can be done in mainly two ways one is one hot encoding and the other one is label encoding.

Here we have used label encoding.

Feature Selection: - Feature Selection is the method done in order to avoid the alteration of accuracy or to increase the accuracy by only selecting the required features or attributes in given data. This increases the accuracy of the model by removing unnecessary attributes. The attributes used for the feature selection are the type crime committed being the input and the area in which they are committed being the output. The following figure 4.1 shows the attributes that are selected from the Dataset of crimes in India during year 2001-2018.



0	ANDHRA PRADESH	2001	101	60	17	50	0	50	46	30	16
1	ANDHRA PRADESH	2001	151	125	1	23	0	23	53	30	23
2	ANDHRA PRADESH	2001	101	57	2	27	0	27	59	34	25
3	ANDHRA PRADESH	2001	80	53	1	20	0	20	25	20	5
4	ANDHRA PRADESH	2001	82	67	1	23	0	23	49	26	23

Figure 3: Sample data of dataset containing crime data Year 2001-2018

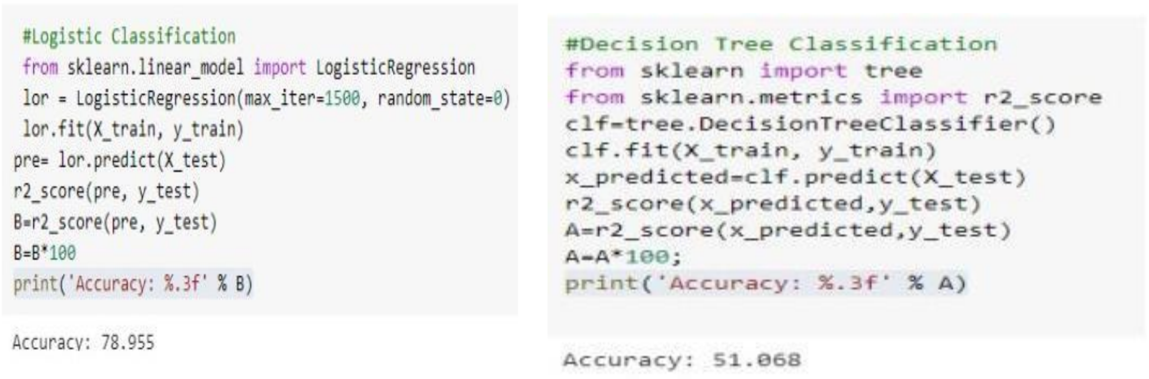
Label Encoding: - Label encoding assigns a numerical value to every categorical value of the data set. By assigning these numerical values the data set is pre-processed and is ready to be used in machine learning models. The one disadvantage is that the model may consider the assigning of numerical values as an order of preference. To avoid such difficulty one hot encoding is used. In the current data set, there will be no difference with an order of preference so label encoding is sufficient to pre-process the data. Sci-kit learn library is used for label encoding which provides us with the code and libraries that are need to be used in order to undergo label encoding. Splitting the data for Training and Testing purpose: - After data Pre-Processing we will split the data for training and testing purpose. Generally, the training data consists of 70-80% of overall data and testing data consists of the remaining 30-20% of the overall data. After the splitting of data into training data and testing data, the data will be ready to be trained using machine learning algorithms which are to be used in this. After splitting, the data standard scaler is used to scale the data and process it.

Machine learning models: - Machine Learning may be a sort of AI that recognizes patterns using data analysis. A computer can learn and make predictions from data through machine learning without being explicitly programmed. Machine learning are often divided into three categories which are Supervised Learning where machine will be told what to be predicted beforehand, Unsupervised Learning where machine is given just multiple inputs and Reinforcement Learning. In this paper, supervised learning methods are wont to predict crime categories. The various classification and regression models used are compared with each other to know which model works best and gives a more accurate prediction of the data given in the dataset. In this crime, type is our input and the areas where this crime is committed mostly is our output, following machine learning algorithms are implemented.

Supervised Learning: - Supervised learning may be a machine learning model which will predict an output from a group of inputs. In supervised learning, the output labels are specifically defined. Input object contains various number of features and typically is represented during a vector form. In the training dataset, each input object is paired with a selected output object. A supervised learning algorithm develops a predictive model using the training data and fits new information i.e. the inputs to the model. Separating train and test data helps supervised learning models to avoid overfitting. The labels of latest information are predicted by the algorithm. Supervised learning models are often implemented on both classification and regression problems. Within the crime dataset, the goal is to predict the category of a criminal incident during a

given time and place. As the crime categories are discontinuous, this is often a supervised classification problem. In this Model Decision Tree, Logistic Regression and Random Forest are used.

Logistic Regression: - A logistic regression model is one of the approaches to linear classifiers. Even though it has a name ending with regression it is a classifier rather than a regressor. Data will be classified into different categories using linear boundaries in Logistic regression. For multiclass dataset, one vs the remainder scheme is employed during this method, logistic regression trains separate binary classifiers for every class. Meaning, each class is assessed against all other classes by assuming that each one other classes is one category.



```
#Logistic Classification
 from sklearn.linear_model import LogisticRegression
 lor = LogisticRegression(max_iter=1500, random_state=0)
 lor.fit(X_train, y_train)
pre= lor.predict(X_test)
r2_score(pre, y_test)
B=r2_score(pre, y_test)
B=B*100
print('Accuracy: %.3f' % B)

Accuracy: 78.955
```

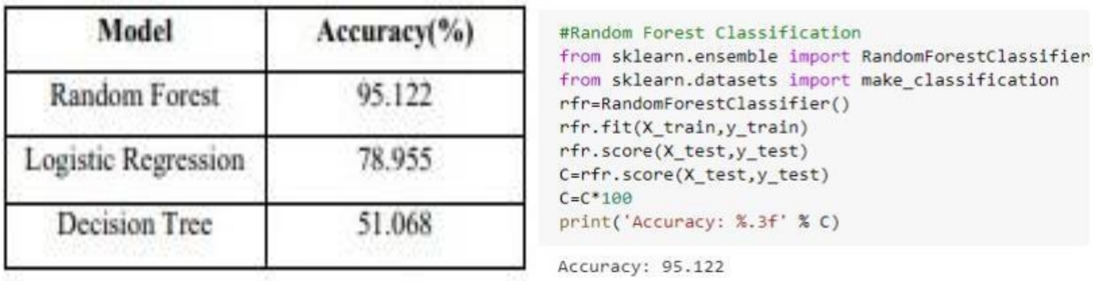
```
#Decision Tree Classification
from sklearn import tree
from sklearn.metrics import r2_score
clf=tree.DecisionTreeClassifier()
clf.fit(X_train, y_train)
x_predicted=clf.predict(X_test)
r2_score(x_predicted,y_test)
A=r2_score(x_predicted,y_test)
A=A*100;
print('Accuracy: %.3f' % A)

Accuracy: 51.068
```

Figure 4: Accuracy for Logistic Regression.

Decision Tree: - Decision tree is the method for finding the approximations of discrete valued targeted functions. In this the learned function is represented by a choice tree. These trees can be represented in a graphical if-then way to make them understandable to Humans. The learning methods are popular for inductive inference algorithms which are successfully applied in broad range tasks. Decision tree is of two types namely Decision Tree Classifier and Decision Tree Regression. The model which we used is the Decision Tree Classifier. This is used when the results required are of type yes/no which means the decision variables are generally discrete or categorical. Decision Tree uses the technique of divide and conqueror. The data is split into various subsets and these subsets are split into even more smaller subsets. Choose a root node and split the data based on information gain(IG) value. This is process is repeated until all the sub-nodes are of the same class i.e. basically the Decision trees classify instances by sorting them down the tree on the basis level to leaf node, which result us by providing the classification of the instance. For the remaining sub trees at new node this processed will be repeated. label is picked randomly consistent with the distribution during a branch. Gini Impurity is computed by summing the probability p times the probability of mistaking while categorizing an item. While building the tree, Information Gain helps to make a decision which feature to separate next at each step. Information Gain are often calculated using entropy, which may be a function to calculate arithmetic mean at each step. The major disadvantage in using this algorithm is there can be overfitting due to the splitting of data into subsets this problem can be solved by setting a depth limit to the tree. Another solution for this problem is to use a random forest algorithm.

Random Forest: - Random forest is formed by combining many Decision trees. It combines all these decision trees consequently trains each one. After the process is done for all the decision trees the final prediction will be the average of the predictions of each tree. It consists of ensemble; an ensemble is nothing but a large number of decision tress. Random forest majorly uses a most important technique called Bootstrap aggregation or it can be simply called as Bagging. It is a simple as well as very powerful ensemble method. An ensemble method is a process that combines the predictions and results from various machine learning algorithms i.e. from various decision tress all together to form more accurate predictions than a person single model. When highly flexible data it memorizes and processes the training data consequently it fits into it perfectly. From above it is clear that the model is not only learning the actual values but also any noise present in it. To avoid this problem, Random forest classifier is used. The decision tree is very sensitive to use. It varies very instantly depending upon the input changes.



Model	Accuracy(%)
Random Forest	95.122
Logistic Regression	78.955
Decision Tree	51.068

```
#Random Forest Classification
from sklearn.ensemble import RandomForestClassifier
from sklearn.datasets import make_classification
rfr=RandomForestClassifier()
rfr.fit(X_train,y_train)
rfr.score(X_test,y_test)
C=rfr.score(X_test,y_test)
C=C*100
print('Accuracy: %.3f' % C)

Accuracy: 95.122
```

Figure 5: Accuracy for Random Forest Classification

Results: - The Following python codes shows the accuracies of the respective algorithms used in this paper to predict the crime data. The above lines of code give the accuracy obtained by using Random

Forest Classification which is about 95.122 The above lines of code give the accuracy obtained by using Logistic Regression which is about 78.955%. The above lines of code give the accuracy obtained by using Decision Tree Classification which is about 51.068%. Following is the table which summarizes the accuracies of all the models used to predict the results. By observing the above results It is clear that Random Forest Classifier method has giving the best accuracy among all the methods used and can be chosen or taken as the best model to predict the data for the given dataset with an accuracy of 95.122%.

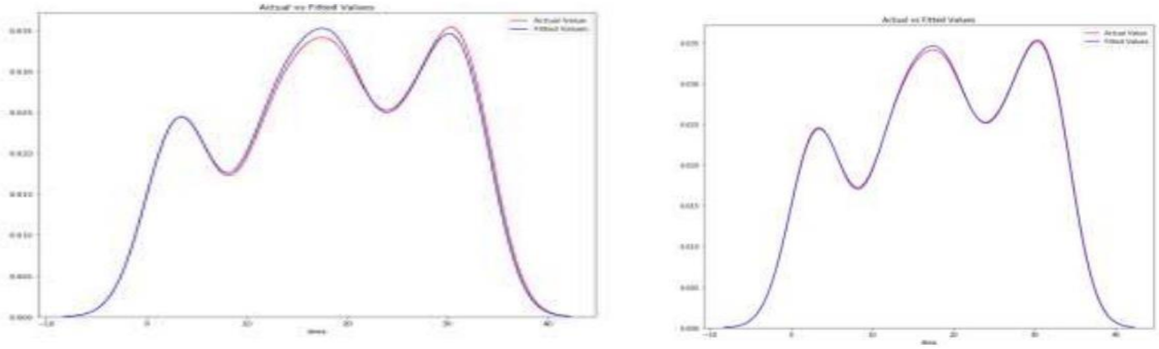


Figure 6: Actual values and fitted values for Decision Tree Classification

Above figure 2.5 depicts the closeness of actual to the predicted values and how are they changing accordingly. The above figure depicts the closeness of actual to the predicted values when we use Decision Tree Classification. Above figure 2.6 depicts the closeness of actual to the predicted values and how are they changing accordingly. The above figure depicts the closeness of actual to the predicted values when we use Decision Tree Classification. The experiment is conducted on Colab notebook using python as core language. Important libraries are provided by Scikit some of them are Pandas, Numpy, Seaborn and Matplotlib. Data Pre-Processing and label encoding is done using Scikit learning tool. Among all the methods we used the best model is found to be Random Forest Classifier with an accuracy of 95.122%.

V. CONCLUSION

The prediction analysis done in this paper provides patterns of crime in a particular area by analyzing and using certain machine learning algorithms. In this paper Random Forest Classification is proven to be a better model for predicting the results which can be observed by comparing the accuracies with the other prediction algorithms. The relatively poor algorithm is Decision Tree classification which is having a low accuracy of 51.068%. As seen in the model fitting curve in the above section. In the future more data can be collected and there will be enhancement on the computer capabilities so, more efficient models can be developed. Throughout the research It is clear that basic details of criminal activities in a neighborhood contain indicators that will be employed by machine learning agents to classify a criminal activity given a location and date. The training agent suffers from imbalanced categories of the dataset, it had been ready to overcome the problem by oversampling and under-sampling the dataset. This paper presents a crime data prediction by taking the types of crimes as input and giving are in which these crimes are committed as output using Colab notebook having python as a core language and python provide inbuilt libraries such as Pandas and Numpy through which the work will be completed faster and Scikit provides all the processes of how to use different libraries providing by the python. Results of prediction are different for different algorithms and the accuracy of Random Forest Classifier found to be good with the accuracy of 95.122%.

ACKNOWLEDGEMENTS

I would like to express my appreciation to all those who provided me the possibility to finish this report. A special gratitude to all our guide , whose contribution to stimulating suggestions and encouragement, helped me to coordinate my project especially in writing this report. Furthermore, I would also like to acknowledge with much appreciation the crucial role of the Faculties of Computer Department, who permitted me to use all required resources and the necessary materials to complete the task "Online Crime Management and Reporting System, and Prediction of Crimes using Machine Learning and Data Science" Also, I appreciate the guidance given by other supervisors as well as the guides especially in our project presentation that has improved our presentation skills thanks to their comments and advice. At last, I would like to say that, it would not have been possible without the help of these people, so thank you to all those who contributed to this project.

VI. REFERENCES

[1] An overview on crime prediction methods nurul hazwani mohd shamsuddin1 , nor azizah ali2 , razana alwee3 1,2,3faculty of computing, universiti teknologi malaysia, johor, malaysia. hazwani.shamsuddin@gmail.com1 , nzah@utm.my2 , razana@utm.my3 .

[2] Crime analytics: analysis of crimes through newspaper articles isuru jayaweera, chamath sajeewa, sampath liyanage, tharindu wijewardane, indika perera department of computer science and engineering, faculty of engineering university of moratuwa sri lanka {jayaweera.10, chamaths.10, sampath.10, tharinduwije.10, indika}@cse.mrt.ac.lk adeesha wijayasiri department of computer and information science and engineering university of florida Gainesville, Skogan (1984), "Reporting Crimes To The Police: The Status Of World Research", Journal Of Research In Crime And Delinquency, Vol. 21, Pp. 113-37. 5. Robert D. Crutchfield, George S. Bridges, And Susan R. Pitchford, "Analytical And Aggregation Biases In Analyses Of Imprisonment: Reconciling Discrepancies In Studies Of Racial Disparity," .

[3] McClendon, Lawrence, and Natarajan Meghanathan. "Using machine learning algorithms to analyze crime data." Machine Learning and Applications: An International Journal (MLAIJ) 2.1 (2015): 1-12.