Vol 12, Issue 07, July/2021
ISSN NO:0377-9254

Analysis of Crime Data Using Machine Leaning Algorithm

PILLI VINAY KUMAR [1], G.RAJESH [2]

[1]PG Scholar, Dept of MCA, Audisankara College of Engineering and **Technology (AUTONOMOUS), Gudur, AP, India.**

[2]Assistant Professor, Dept of MCA, Audisankara College of Engineering and Technology (AUTONOMOUS), Gudur, AP, India.

Abstract_ The crook instances in India are growing swiftly due to which variety of instances pending are additionally piling up. This non-stop expand in the crook instances is proving to be tough to be labeled and to be solved. Recognizing the crook exercise patterns of a location is vital in order to forestall it from happening. The crime fixing companies can do a higher work if they have a right thought of the sample of crook things to do that are occurring in a specific area. This can be accomplished by way of the usage of desktop studying by way of using exclusive algorithms to locate the patterns of the crook things to do in a unique area. This paper makes use of crime statistics set and predicts the sorts of crimes in a specific vicinity which helps in dashing up the classification of crook instances and proceed accordingly. This paper makes use of the statistics of previous 18 years that is gathered from more than a few depended on sources. Data pre-processing is as vital as remaining prediction, this paper used function selection, disposing of null values and label encoding to easy and nourish the data. This lookup gives an environment friendly computing device leaning mannequin for predicting the subsequent crook case.

1. INTRODUCTION

At present, the crook instances that are pending in India are hastily growing with the wide variety of crimes dedicated are increasing. To remedy a case based totally upon a specific records there need to be a thorough investigation and evaluation that is to be performed internally [1]. With the quantity of crime statistics that is current in India presently the evaluation and selection making of these crook instances is too hard for the officials. Identifying this a predominant hassle this paper concentrates on developing a answer for the selection making of crime that is committed. Machine Learning is the branch of science the place computer systems determine except human intervention. In current

instances Machine Learning is being used in more than a few domains one of the examples of such instances is computerized or self-driving cars. By Machine Learning algorithms there is a way the place we can predict sure outcomes based totally upon our inputs given and furnish a answer to fixing crime instances in India. The two frequent kinds of prediction methods are classification and regression. This crime statistics prediction is a area the place classification is applied. Classification is a supervised prediction approach and it has been used in a range of domains like forecasting stock, medicinal area, etc. [2].The most important intention of this paper is to reflect onconsideration on some algorithms which can be used to predict and analyse the crime facts and enhance the accuracy of these fashions via information processing in order to reap higher results. The cause is to teach the required mannequin to predict the records the use of the coaching statistics set through validation of the check facts set [3]. The fashions which are being used right here are Logistic Regression, Decision Tree classification, Random Forest classification.

2.LITERAURE SURVEY

2.1 McClendon, Lawrence, and Natarajan Meghanathan. "Using laptop getting to know algorithms to analyze crime data." Machine Learning and Applications: An International Journal (MLAIJ) 2.1

Data mining and desktop studying have emerge as a critical phase of crime detection and prevention. In this research, we use WEKA, an open supply records mining software, to behavior a comparative find out about between the violent crime patterns from the Communities and Crime Unnormalized Dataset furnished through the University of California-Irvine repository and genuine crime statistical statistics for the country of Mississippi that has been furnished through neighborhoodscout.com. We applied the Linear Regression, Additive Regression, and Decision Stump algorithms the use of the equal finite set of features, on the Communities and Crime Dataset. Overall, the linear regression algorithm carried out the quality amongst the three chosen algorithms. The scope of this assignment is to show how high-quality and correct the computer gaining knowledge of algorithms used in information mining evaluation can be at predicting violent crime patterns.

2.2 Alkesh Bharati, Dr Sarvanaguru RA. K," Crime Prediction and Analysis Using Machine Learning" in International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 09 | September 2018.

The crook instances in India are growing swiftly due to which variety of instances pending are additionally piling up. This non-stop amplify in the crook instances is proving to be tough to be categorised and to be solved. Recognizing the crook pastime patterns of a vicinity is essential in order to stop it from happening. The crime fixing companies can do a higher work if they have a correct concept of the sample of crook things to do that are taking place in a specific area. This can be carried out by using the usage of computer gaining knowledge of via using specific algorithms to locate the patterns of the crook things to do in a unique area. This paper makes use of crime facts set and predicts the kinds of crimes in a unique vicinity which helps in rushing up the classification of crook instances and proceed accordingly. This paper makes use of the facts of previous 18 years that is accumulated from a number relied on sources. Data pre-processing is as vital as last prediction, this paper used characteristic selection, casting off null values and label encoding to smooth and nourish the data. This lookup offers an environment friendly desktop leaning mannequin for predicting the subsequent crook case. Various Machine getting to know fashions such as Logistic Regression, Decision Tree Classification, and Random Forest Classification had been used to discover the most environment friendly mannequin to predict the kind of crime at a unique region This paper discusses the about current device which makes use of Knearest neighbour to predict subsequent kind of crime at a precise location, and additionally indicates how the proposed machine is higher than the current present system. This paper compares many computing device studying fashions amongst themselves to discover most environment friendly computing device gaining knowledge of to address this problem. Keywords: crook cases, Machine learning, Crime data, Algorithms, Data pre-processing, Logistic regression, Decision tree classification, Random woodland classification. 1. Introduction At present, the crook instances that are pending in India are unexpectedly growing with the range of crimes committed are increasing. To resolve a case based totally upon a specific statistics there have to be a thorough investigation and evaluation that is to be executed internally [1]. With the quantity of crime information that is existing in India presently the evaluation and selection making of these crook instances is too tough for the officials. Identifying this a predominant hassle this paper concentrates on developing a answer for the selection

making of crime that is committed. Machine Learning is the department of science the place computer systems figure out besides human intervention. In current instances Machine Learning is being used in a range of domains one of the examples of such instances is computerized or self-driving cars. By Machine Learning algorithms there is a way the place we can predict positive outcomes based totally upon our inputs given and furnish a answer to fixing crime instances in India. The two frequent sorts of prediction methods are classification and regression. This crime records prediction is a area the place classification is applied. Classification is a supervised prediction approach and it has been used in a range of domains like forecasting stock, medicinal area, etc. [2].The most important intention of this paper is to reflect onconsideration on some algorithms which can be used to predict and analyse the crime statistics and enhance the accuracy of these fashions through records processing in order to acquire higher results. The reason is to educate the required mannequin to predict the information the use of the coaching information set by means of validation of the take a look at records set [3]. The fashions which are being used right here are Logistic Regression, Decision Tree classification, Random Forest classification

2.3 Wang, Tong, et al. "Learning to observe patterns of crime." Joint European convention on laptop studying and know-how discovery in databases. Springer, Berlin, Heidelberg, 2013.

We introduce a novel, sturdy data-driven regularization method referred to as Adaptive Regularized Boosting (AR-Boost), encouraged via a want to limit overfitting. We change AdaBoost's challenging margin with a regularized gentle margin that trades-off between a large margin, at the price of misclassification errors. Minimizing this regularized exponential loss consequences in a boosting algorithm that relaxes the vulnerable getting to know assumption further: it can use classifiers with error higher than 1 two . This allows a herbal extension to multiclass boosting, and in addition reduces overfitting in each the binary and multiclass cases. We derive bounds for education and generalization errors, and relate them to AdaBoost. Finally, we exhibit empirical outcomes on benchmark information that set up the robustness of our method and accelerated overall performance overall. 1 Introduction Boosting is a famous approach for enhancing the accuracy of a classifier. In particular, AdaBoost [1] is regarded the most famous shape of boosting and it has been proven to enhance the overall performance of base rookies each theoretically and empirically. The key thinking in the back of AdaBoost is that it constructs a robust

classifier the usage of a set of susceptible classifiers [2,3]. While AdaBoost is pretty powerful, there are two important limitations: (1) if the base classifier has a misclassification error of increased than 0.5, generalization decreases, and (2) it suffers from overfitting with noisy records [4,5]. The first hassle can emerge as extreme in multiclass classification, the place the error price of random guessing is C−1 C , the place C is the range of training [6]. AdaBoost requires vulnerable classifiers to attain an error fee much less than 0.5, which can be challenging in multiclass classification. The 2nd trouble of overfitting takes place basically due to the fact susceptible classifiers are unable to seize "correct" patterns interior noisy data.

3. PROPOSED SYSTEM

The proposed device is made on the groundwork of the lookup work that is accomplished by means of going via a number of such documentations. Nearly all of the crimes are predicting primarily based on the vicinity and the kinds of crimes that are taking place in these areas.On surveying preceding works, Linear Regression, Decision Tree and Random Forest have a tendency to provide exact accuracy so these fashions are used in this paper to predict crimes. The dataset used in this paper is from data.world.com. The statistics set includes one-of-a-kind sorts of crimes that being dedicated in India in accordance to the kingdom and 12 months respectively [4]. This paper takes sorts of crimes as enter and offers the vicinity in which crimes are dedicated as output. The facts pre-processing entails information cleaning, function selection, losing null values, records scaling by means of normalizing and standardizing. After facts pre-processing the information is free of null values which m ay alter the accuracy of the mannequin appreciably and characteristic choice is used to choose solely the required elements that won't have an effect on the accuracy of model.After information pre-processing the fashions chosen i.e., Logistic Regression, Decision Tree and Random Forest are skilled by means of splitting the records into as teach and take a look at data.As the output required is a specific cost classification fashions are used here. Python language is used for the statistics prediction.

3.1 IMPLEMENTATIONS

3.1.1 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. [6].

3.1.2 Data PreProcessing

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 [8]. 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. The one which is used here is label encoding.

3.1.3 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.

3.1.4 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 this 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. Scikit 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.

3.1.5 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

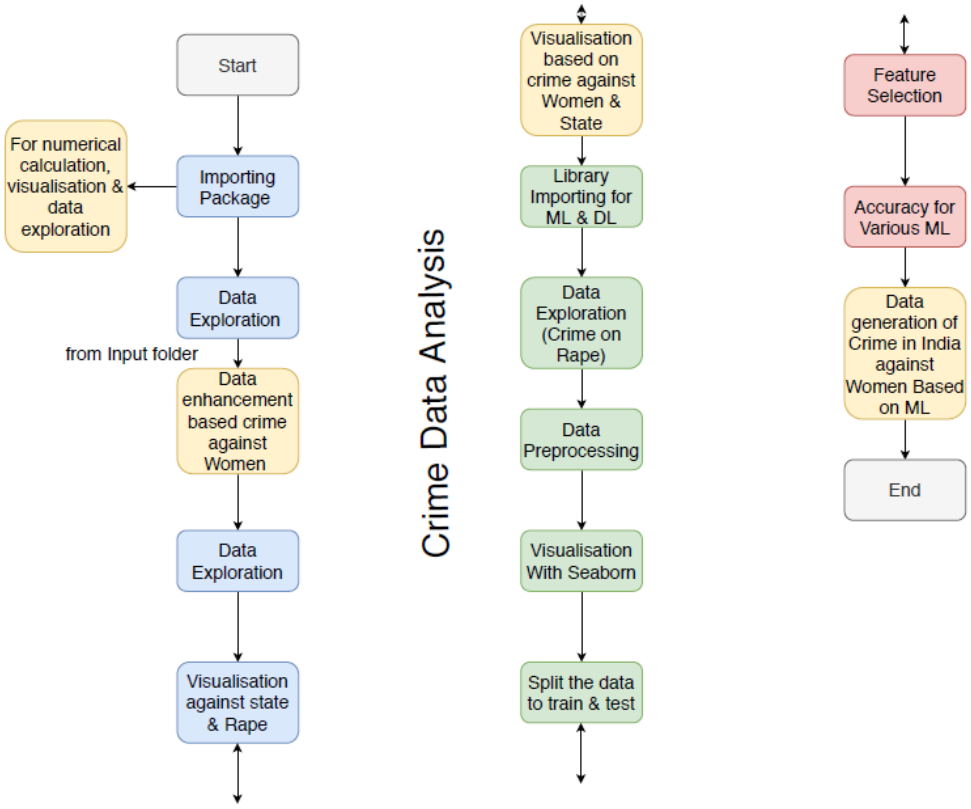


Fig 3.1:Flow Chart

Vol 12, Issue 07, July/2021
ISSN NO:0377-9254

4.RESULTS AND DISCUSSIONS

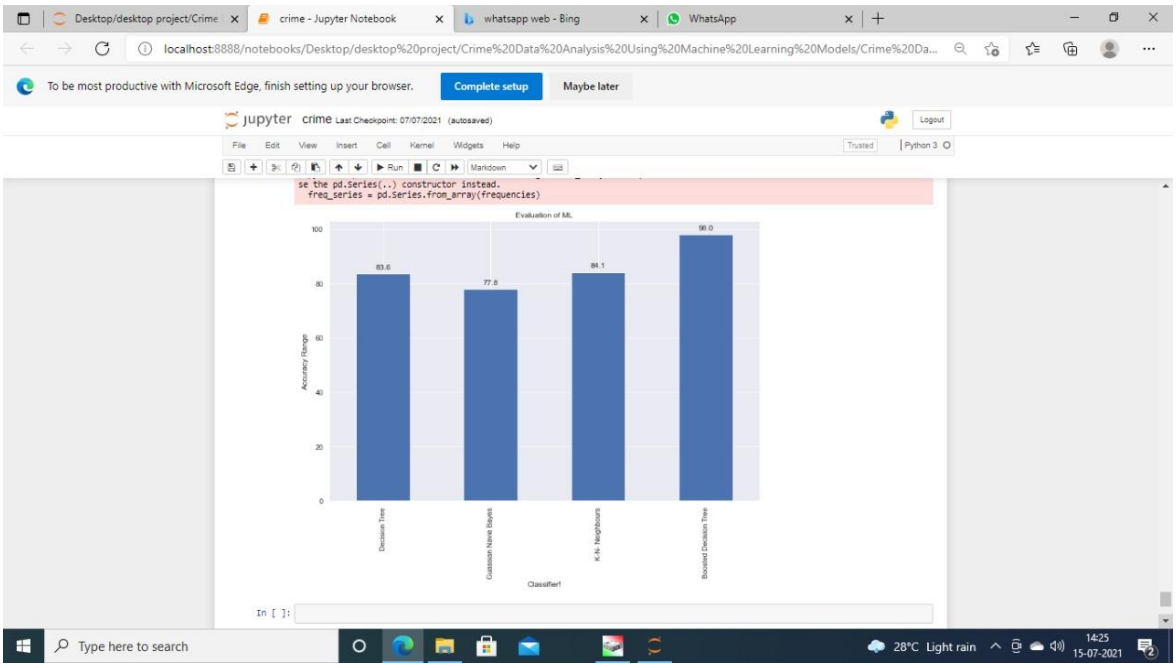
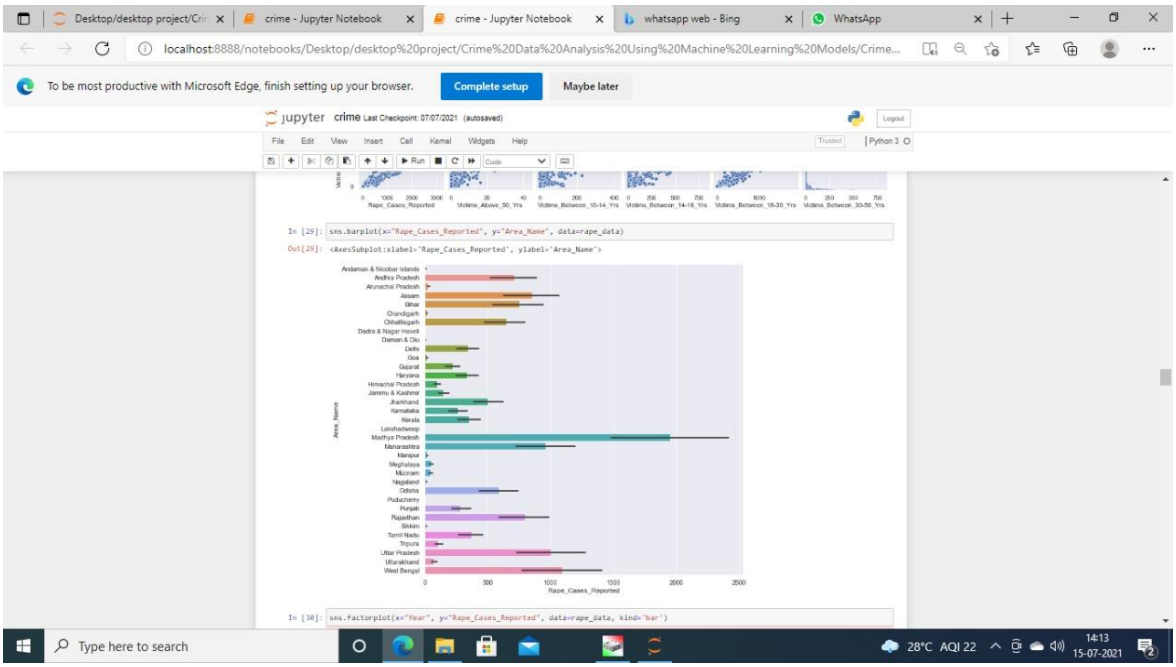
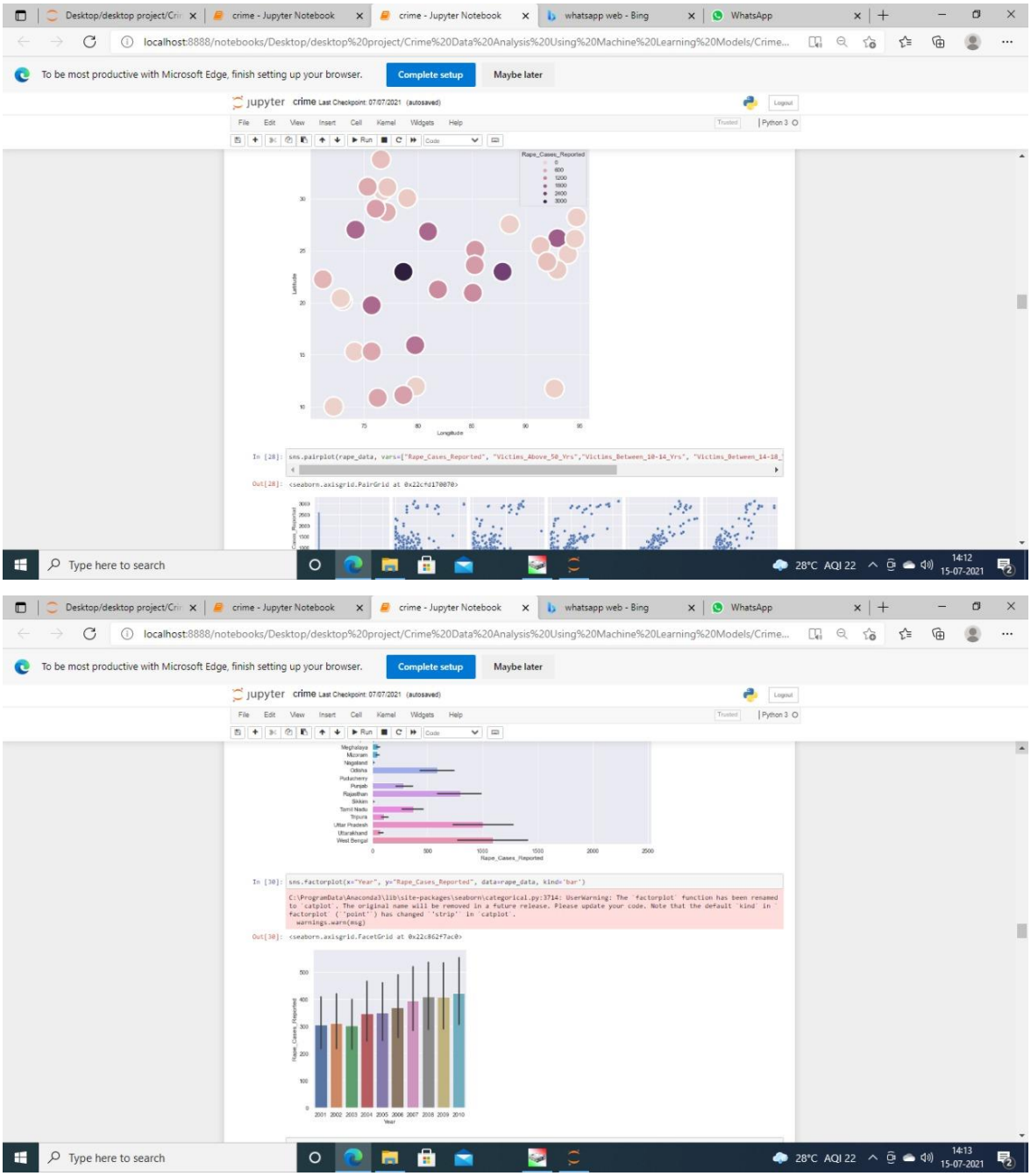
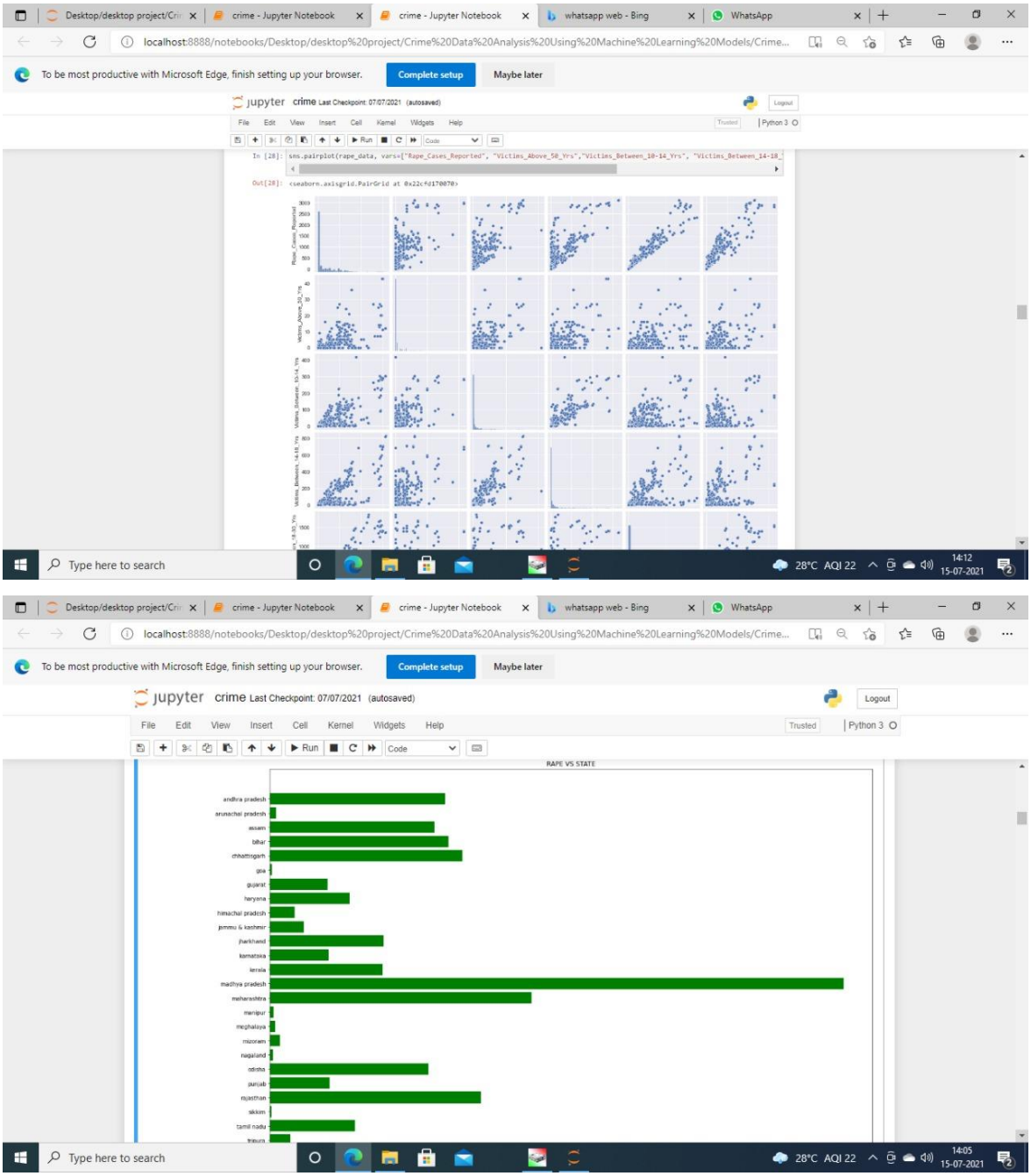


Fig 4.1 Graph



Vol 12, Issue 07, July/2021
ISSN NO:0377-9254





5. CONCLUSION

It is clear that fundamental important points of crook things to do in a neighbourhood comprise indications that will be employed with the aid of laptop getting to know dealers to classify a crook undertaking given a area and date. The coaching agent suffers from imbalanced classes of the dataset, it had been geared up to overcome the hassle via oversampling and under-sampling the dataset. This paper provides a crime records prediction through taking the sorts of crimes as enter and giving are in which these crimes are dedicated as output the use of Colab pocket book having python as a core language and

python grant built in libraries such as Pandas and Numpy thru which the work will be achieved quicker and Scikit gives all the procedures of how to use distinctive libraries offering through the python. Results of prediction are distinct for one of a kind algorithms and the accuracy of Random Forest Classifier observed to be desirable with the accuracy of 95.122%.

6.REFERENCES

1. 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.

2. Alkesh Bharati, Dr Sarvanaguru RA. K," Crime Prediction and Analysis Using Machine Learning" in International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 09 | September 2018

3. McClendon, L., & Meghanathan, N. (2015). Using machine learning algorithms to analyze crime data. Machine Learning and Applications: An International Journal (MLAIJ), 2(1), 1-12.

4. Chen, Hsinchun, et al. "Crime data mining: a general framework and some examples." computer 37.4 (2004): 50-56.

Author's Profile:

PILLI VINAY KUMAR has Pursuing his MCA from Audisankara College of Engineering and Technology (AUTONOMOUS), Gudur, affiliated to JNTUA in 2021. Andhra Pradesh, India.

G. RAJESH has received him M.Tech degree in CSE from Rayalaseema University in 2014, at Audisankara College of Engineering and Technology, Gudur, Nellore Dt, Andhra Pradesh, India. He has 10 years of experience in the field of teaching. He is a research scholar in Rayalaseema University. Kurnool. He did is M.Tech in JNTU, Hyderabad. His areas of interests are Data warehousing and Data Mining and Computer Network.