*sustainability*

MDPI

*Article*

# Computational System to Classify Cyber Crime Offenses using Machine Learning

**Rupa Ch [1], Thippa Reddy Gadekallu [2], Mustufa Haider Abidi [3,*] and Abdulrahman Al-Ahmari [3]**

1 Department of Computer Science, VR Siddhartha Engineering College, Vijayawada 520007, India; rupamtech@gmail.com
2 School of Information Technology and Engineering, Vellore Institute of Technology, Vellore 632014, India; thippareddy.g@vit.ac.in
3 Raytheon Chair for Systems Engineering, Advanced Manufacturing Institute, King Saud University, Riyadh 11421, Saudi Arabia; alahmari@ksu.edu.sa
* Correspondence: mabidi@ksu.edu.sa

Received: 31 March 2020; Accepted: 14 May 2020; Published: 16 May 2020


**Abstract:** Particularly in the last decade, Internet usage has been growing rapidly. However, as the Internet becomes a part of the day to day activities, cybercrime is also on the rise. Cybercrime will cost nearly $6 trillion per annum by 2021 as per the cybersecurity ventures report in 2020. For illegal activities, cybercriminals utilize any network computing devices as a primary means of communication with a victims' devices, so attackers get profit in terms of finance, publicity and others by exploiting the vulnerabilities over the system. Cybercrimes are steadily increasing daily. Evaluating cybercrime attacks and providing protective measures by manual methods using existing technical approaches and also investigations has often failed to control cybercrime attacks. Existing literature in the area of cybercrime offenses suffers from a lack of a computation methods to predict cybercrime, especially on unstructured data. Therefore, this study proposes a flexible computational tool using machine learning techniques to analyze cybercrimes rate at a state wise in a country that helps to classify cybercrimes. Security analytics with the association of data analytic approaches help us for analyzing and classifying offenses from India-based integrated data that may be either structured or unstructured. The main strength of this work is testing analysis reports, which classify the offenses accurately with 99 percent accuracy.

**Keywords:** integrated cybercrimes; security analytics; machine learning approaches; supervised learning; classification; clustering; India

## 1. Introduction

Currently, cybercrime is being used with diverse terminologies such as computer crime, e-crime, Internet crime, etc. [1]. However, the association of chief police officers of England (ACPO) and the U.S Department of Justice (DOJ) define cybercrime as any crime committed by any electronic computing devices [2–4]. Likewise, CERT-In and CERT-US are providing services to the users by prior notifications on the list of vulnerabilities that causes them to become victims of the cyber-attacks. Further, in 2010 [5,6] recommended a cybercrime model for classification that is dependent on the role of the computers' intelligence in cybercrime. Similarly, various taxonomies are proposed by the researchers, scholars and users to categorize the cybercrimes.

The growing rate of cybercrime threats is increasing day by day. Currently, there is no foolproof systematic and reliable tool on reviews of cybercrimes due to a lack of record maintenance at concerned offices because of various reasons such as victims' assumptions on police response, lack of awareness

of the users about IT (information technology) act on cybercrimes and the inability of the victims to be recognized that they have been victimized. Promoting and educating the people on cybercrimes, identification and maintenance of area wise cybercrime rates could also assist in reducing and classifying the cybercrimes.

The main focus of this research work is to find the attacks that take advantage of security vulnerabilities [7–15] and classify these attacks by making use of machine learning techniques [16–19]. The framework developed in this research work is essential for the creation of a model that can support analytics regarding the identification, detection and classification of integrated cybercrime offenses. The proposed tool will provide the essential broad knowledge of cybercrime offenses in the society, enable them to consider the threat landscape of such attacks, and avoid the personification of the cybercrime offenses. This will make it easier to understand the patterns of various cybercrime offenses in a particular area and then concerned authorities can take the necessary measures in order to reduce the incarnation of the frequently occurred cybercrime offenses.

Prediction analysis of the integrated cybercrime offenses are made, to make year-wise analysis and find the occurrences of the various offenses in a particular location and then find the crime rate in a particular year. In this proposed work, we present a framework that will analyze and classify the cybercrime offenses and the datasets (for cybercrime in India) were obtained from Kaggle and CERT-In repositories. The main objectives of this research work are: 1) Analyze and classify the various types of cybercrimes; 2) Group them as different clusters that help to analyze and predict the rate of the incidents easily; 3) Test and analyze the performance of the proposed system.

After discussing the prerequisite of the paper in Section 1 the remaining study is organized as follows: Literature survey related to the cybercrime offenses is presented in Section 2. Section 3 consists of a detailed description of the proposed model methodology and algorithms. Analysis of the results obtained is presented in Section 4. Section 5 presents conclusions deduced from the research work along with future research directions.

## 2. Related Works

To date, several methods were proposed to analyze and evaluate the cybercrime offenses. Several researchers presented methods to analyze the cybercrime offenses, but there are some pros and cons with their works. They are presented in this section.

Ganesan & Mayilvahanan [19] propose a methodology that offered the discovery of unpredicted patterns. They analyzed the cybercrime data from the database, which is a collection of data fields from Internet web pages. The data fields include cyber-bullying, stalking, scams, robbery, identity theft, defamation and harassment. They introduced this model in order to categorize cybercrime offenses such as whether they are violent or non-violent, and further, they can be categorized as various types of cybercrimes such as cyber terrorism, cyberstalking, pornography, cyberbullying, cyber fraud and cyber theft.

Khan, et al. [20] discussed a system that identified the occurrence of the Denial of Service (DoS) attacks in the network by using the data mining approaches such as pattern recognition, Breadth-First Search (BFS). Here the model only concentrated on the log files, not on the records that were communicated through the network.

Nouh et. al [21] designed a multipurpose cybercrime intelligent system framework. The core aim of this framework was to minimize cognitive biases usage during the investigation process. This model provides six major steps, i.e., defining the problem, generating the hypothesis, collecting the information, evaluating the hypothesis, selection of related hypothesis and the information monitoring about incidents continuously. The main drawback of this framework is that it does not provide specificity though it is designed to achieve specific functionalities only. Moreover, it is for specific kinds of analytics only.

Prasanthi et al. [22] proposed a cybercrime prevention and detection model by feature extraction. The features such as incident, type of crime (online or offline) were extracted by using the TFIDF (term

frequency–inverse document frequency) weighted vectors present in the cybercrime data. The main purpose was to get knowledge from the study through the framework that processes the cybercrime offenses classification by feature extraction based on occurrence and severity. The model was evaluated by similar clustering kind of the cybercrime offenses by features extracted so that it can easily identify the repeated cybercrime offenses and can possibly to take precautions on such crimes.

Soomro et al. [23], discussed various cybercrimes related to social media. Here they recommended some techniques to prevent cybercrimes over social media. In current days, social media is using as a tool by the users to convey bulk messages, organize online meetings and work at home and others. In this work, authors have mentioned different types of cybercrimes with corresponding prevention tips and techniques. This study is only helpful for study purposes. It does not detect and predict cybercrimes cases.

Çagrı et al. [24], designed and developed a system that detected the cybersecurity-related accounts on online social networks (OSN) such as Twitter. The authors used machine learning approaches such as SVM, decision tree and Random Forest approaches to detect suspicious accounts automatically on Twitter (OSN) platform. Some behavioral, profile and content features extracted from the tweets later applied a certain approach to identify the anonymous account. However, only twitter-based suspicious account detection concentrated by the authors while implementing this model. This framework is not useful to detect and predict all the cybercrimes on all OSN platforms.

Chen et al. [25], addressed a general framework related to cybercrime mining by showing some examples in the study. This framework showed the relationship between data mining techniques for entity extraction, association, prediction and visualization and crime types. This study addressed an analysis of criminal networks and also criminal groups. The Coplink data were used here to test their methodology.

Prabakaran et al. [26], mentioned different data mining techniques and machine learning techniques in their review work. This work presented various types of crimes like violent crime, traffic violence, sexual assault and cybercrimes. Here they discussed general techniques that help to detect various crimes. In this work, genetic algorithm, hidden Markov model, neural network, kernel destiny algorithm, logistic regression, random forest, and *k*-means were used.

Chauhan et al. [27], presented a review report on crime analysis using data mining techniques. It discussed how the data mining approaches like *k*-means algorithm, random forest, etc. can help to identify the criminals.

Based on the literature review it can be said that machine learning is an efficient tool to detect and classify cybercrimes. However, still, there is a scope of improvements in this regard. Therefore, in this research work a machine learning based tool is proposed to find the attacks that take advantage of security weaknesses and classify these cyber-attacks.

## 3. Proposed Methodology

At present, there is no generalized framework is available to categorize cybercrime offenses by feature extraction of the cases. In the present work, data analysis and machine learning are incorporated to build a cybercrime detection and analytics system. The proposed system's design and implementation utilize classification, clustering and supervised algorithms. Figure 1 depicts the proposed methodology. Here, naïve Bayes is used for classification [28–34] and *k*-means are used for clustering [35]. For feature extraction in the proposed work, the TFIDF or tf–idf vector process is used [36]. This developed methodology is based on 4 phases that are applied to the data, which are reconnaissance, preprocessing, data clustering and classification and prediction analysis.

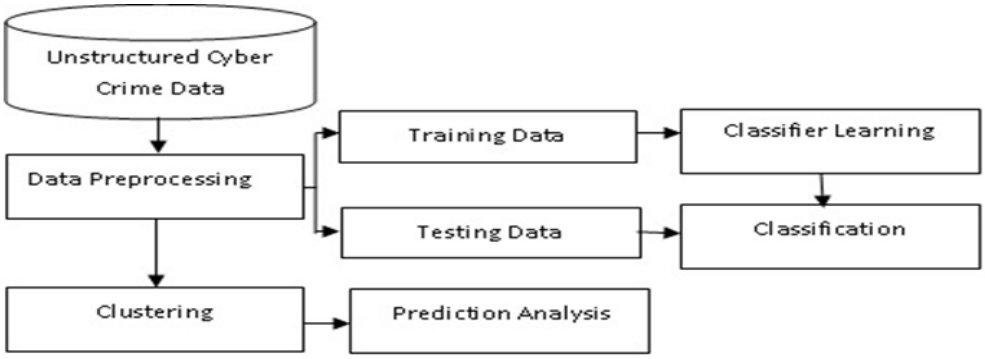


**Figure 1.** Proposed approach to analyze cybercrime incidents.

### *3.1. Information Gathering (Reconnaissance)*

In the reconnaissance phase, the integrated data (structured and unstructured) are collected from Kaggle and CERT-In. The integrated data are stored as raw data in the database. Table 1 depicts sample data of the dataset considered in the proposed model. The next phase of the approach is preprocessing that is used to remove the noisy information from the raw data.

**Table 1.** Sample dataset of the proposed model

| Incident | Offender | Access Violation | Victim | Harm | Year | Location | Age of Offender |
|---|---|---|---|---|---|---|---|
| Illegal downloading | CC | TI | Company | Loss of proprietary | 2013 | Delhi | 27 |
| Pirated textbook | CC | TI | Individual | Loss of copyright | 2012 | Maharashtra | 38 |
| Illegal downloading of application | CC | TI | Company | Loss of proprietary | 2013 | Hyderabad | 26 |
| Pirated software | CC | TI | Company | Loss of intellectual rights | 2012 | Hyderabad | 22 |
| Illegal downloading of music | CC | TI | Industry | Loss of proprietary | 2015 | Hyderabad | 20 |
| Hacking of power plant communication | CH | TI | State | Infrastructure loss | 2013 | Delhi | 34 |
| Hacking of smart phone | CH | TT | Individual | Loss of proprietary | 2014 | Gujrat | 23 |
| Hacking of government website | CH | TI | State | Economic loss | 2015 | Maharashtra | 32 |
| Stealing of credit card information | CC | TI | Individual | Financial loss | 2015 | Banglore | 33 |
| Illegal purchase of goods | CC | TI | Company | Loss of proprietary | 2013 | Banglore | 29 |
| Creating a fake account of reputed person | CC | TI | Individual | Loss of reputation | 2015 | Hyderabad | 21 |
| Siphoned money from a individual account | CT | TI | Individual | Financial loss | 2015 | Tamilnadu | 22 |
| OTP theft | CC | TI | Individual | Financial loss | 2012 | Tamilnadu | 24 |
| Illegal purchase of goods | CC | TI | Company | Loss of proprietary | 2011 | Maharashtra | 26 |
| KYC theft | CT | TI | Individual | Financial loss | 2013 | Maharashtra | 22 |
| Creating a fake ID | CH | TS | Individual | Loss of reputation | 2014 | Bihar | 23 |
| Spoof calling | CC | TT | Individual | Loss of privacy | 2011 | Gujrat | 24 |
| Hacking of password of an account | CH | TI | Individual | Loss of reputation | 2013 | Bihar | 21 |

**Table 1.** *Cont.*

| Incident | Offender | Access Violation | Victim | Harm | Year | Location | Age of Offender |
|---|---|---|---|---|---|---|---|
| Illegal downloading of movie | CH | TI | Industry | Loss of proprietary | 2012 | Maharashtra | 22 |
| Hacking of smart phone | CH | TT | Individual | Loss of security | 2011 | Hyderabad | 24 |
| Illegal access of social account | CC | TI | Individual | Loss of proprietary | 2014 | Hyderabad | 21 |
| OTP theft | CT | TI | Individual | Financial loss | 2012 | Maharashtra | 32 |
| Illegal access of college website | CC | TI | Organisation | Financial loss | 2015 | Delhi | 23 |
| Stealing of bank account details | CC | TI | Individual | Financial loss | 2013 | Maharashtra | 30 |

CC = cyber-criminal; CT = cyber terrorist; CH = cyber hacker; TI = Through Internet; TT = Through telecommunication; TS = Through social network;

### 3.2. *Preprocessing*

In this phase only the feature extraction process takes place. It converts the high dimensional data to low dimensional data. This preprocessed data are helpful for data visualization because a composite data can organize well when that complex data are converted as a less number of dimensions. For feature extraction in the proposed work, the TFIDF or tf–idf vector process is used [36]. It will evaluate the unigrams and bigrams of each and every corresponding cybercrime. One of the best approaches for the feature extraction is the use of a bag of a model, which means a model for each feature in our case finds out the presence of a number of different words that are taken into consideration, but not the order of the words they occur in each of the features.

Table 2 shows the list of content-based features which are considered to make classification of cybercrimes what has listed in Table 1 such as identity theft, copyright attack, hacking and others. It can be done by finding the word frequencies using the tf-idf Vector [37–40]. Text data must be preprocessed before prediction analysis and remove the irrelevant words or noise called tokenizers. In this work, tf–idf was used to identify relevant terms in a document. It consists of two parts: one is Term Frequency and the other one is Inverse Document Frequency.

**Table 2.** List of features.

| Content-Based Features |
|---|
| Incident |
| Offender |
| Harm |
| Access Violation |
| Year |
| Victim |

Term Frequency: This recapitulates how often a word has occurred in the given report and finds out the importance of each word in a document with respect to the whole corpus and is calculated using Equation (1) [41–43].

Inverse Document Frequency: This downscales the given words present in the report that appeared many times in the report [41–43].

The main steps involved in tf–idf are, (1) tokenize the sentence (2) evaluate term frequency (tf) (3) evaluate inverse document frequency (idf) (4) calculate the tf–idf score by multiplying the tf and idf results (5) score the record sentences (6) find the threshold.

Tokenize the sentence: It is an initial step. Here words of the records are tokenized, and weights are assigned to them.

Evaluate Term Frequency: The term frequency (tf) for each term in a record is evaluated based on statistics of its presence in a record. Later, the tf values of each term are stored in a matrix format. That is called as tf matrix [41].

$$\begin{aligned} \text{tf–idf} &= \log \tfrac{N}{1+N_w} \\ \text{tf (term)} &= \tfrac{\text{N}_{\text{term}}(\text{Record})}{\text{T}_{\text{term}}\ (\text{Record})} \end{aligned} \tag{1}$$

where $\text{N}_{\text{term}}$ = No. of times the term appeared in a record $\text{T}_{\text{term}}$ = Total no. of terms in a record. Evaluate Inverse Document Frequency: The inverse document frequency (idf) for each term in a record evaluated here. Later stored those in a matrix referred it as idf matrix.

$$\sum_{i=1}^{N} \chi^2 = \frac{(O-E)^2}{E} \text{idf(term)} = \log \frac{N}{1+N_{term}} \tag{2}$$

where N = Total No. of Records; $\text{N}_{\text{term}}$ = Total No. of terms in Records. Calculation of tf–idf score: tf–idf score evaluated by considering the individual matrices of both the frequencies, i.e., tf and idf.

$$\text{tf–idf } (term) = \text{tf}(term) \times \text{idf (term)} \tag{3}$$

Scoring the record sentences: A sentence of scoring is various in different approaches or algorithms. Here, we have considered tf–idf to allot a score to the terms of a sentence that belongs to a record. The average value of tf–idf of all the terms of a sentence becomes the score of that sentence.

Find the Threshold: Number of approaches existed to evaluate the threshold values. Here we have considered the average score of all the sentences in the record as the threshold. Generally, this value helps to detect the correlated terms in the data.

The chi-squared ($\chi^2$) measure is used to find out the correlation between the two categorical attributes of the incident [25]. It checks whether a relationship between the two variables reflect on the cybercrime dataset or not.

$$\sum_{i=1}^{N} \chi^2 = \frac{(O-E)^2}{E} \tag{4}$$

where O = Observed count of incidents and cybercrimes; E = Expected Count of Incidents and cybercrimes.

For an incident feature, we compute term frequency measure and inverse document frequency measure, which are used to calculate the tf–idf vector for each of the cybercrime incidents.

Python code for Calculation of tf-idf vector – Incident

```
fromsklearn.feature_extraction.text import TfidfVectorizer
tf_idf = TfidfVectorizer(sublinear_tf=True, min_df=5, norm='l2', encoding='latin-1',
ngram_range=(1, 2), stop_words='english')
feat_crime = tfidf.fit_transform(df.Incident).toarray()
features.shape
```

1183 incidents are represented by 148 features to represent tf–idf score. It is considered by different unigrams and bigrams. In the next step, we will find out the most correlated words with each of the cybercrime.

Python code for Calculation of Correlated Words

```
fromsklearn.feature_selection import chi2
import numpy as np
M = 2
for Cyber_crime, c_id in sorted(c_to_id.items()):
feat_crime_chi2 = chi2(feat_crime, labels == c_id)
    indices_crime = np.argsort(feat_crime_chi2[0])
    feat_crime_names = np.array(tfidf.get_feature_names())[indices_crime]
    uni_grams = [j for j in feat_crime_names if len(j.split(' ')) == 1]
    bi_grams = [j for j in feat_crime_names if len(j.split(' ')) == 2]
write (.format(Cyber_crime))
    write (.format('\n. '.join(unigrams[-N:]))
write(.format('\n. '.join(bigrams[-N:]))
```

By using the above code, we can find out the most correlated unigrams and bigrams corresponding to each cybercrime attack.

### *3.3. Clustering and Classification*

Here, naïve Bayes is used for classification [28–34] and *k*-means are used for clustering [35]. The cybercrime offenses are clustered based on the TFIDF weighted vectors obtained from the features. The data has considered by using a 70:30 thumb rule. Where 70% of data were utilized for training and 30% of the data were used for validation and testing purposes. It consists of various features such as incidents, year, location, age, etc. By using these features the cybercrime incidents are categorized [44].

### *3.4. Prediction Analysis*

In the prediction analysis step, the cybercrime data were analyzed and used to predict which crime is occurring more in a particular year at a particular location. Through this analysis, one can predict the cybercrime data and can reduce the incarnation of cybercrime incidents. Therefore, in this step, the prediction of the cybercrime data is classifie

## 4. Results and Analysis

As mentioned in the previous sections, the proposed system is designed and developed by considering the data from sources such as Kaggle and CERT-In. It consists of more than 2000 records with the eight attributes such as incident, offender, victim, harm, year, location, age of the offender and cybercrime. Incidents that occurred in India during 2012–2017 were considered. More than 2000 records are used to construct and test the proposed computational system. Table 3 displays the data after removing missing values in the incident column of the dataset and here, a column is added in which the cybercrime is encoded as an integer.

**Table 3.** Data after removing the missing values.

| | Cybercrime | Incident | Catergory_id |
|---|---|---|---|
| **0** | Identity Theft | Email Id Theft | 0 |
| **1** | Copyright attack | Pirated application | 1 |
| **2** | Identity Theft | Illegal purchase of goods | 0 |
| **3** | Copyright attack | Posting an article without permission | 1 |
| **4** | Copyright attack | Making piracy of an application | 1 |
| **5** | Identity Theft | KYC theft | 0 |
| **6** | Identity Theft | Online shopping fraud | 0 |
| **7** | Hacking | Hacking of Smart phone | 2 |
| **8** | Copyright attack | Illegal downloading of movie | 1 |
| **9** | Identity Theft | Illegal access of bank account | 0 |

Figure 2 shows the data after applying data preprocessing using the TFIDF weighted vector process. Here the dataset consists of 2097 incidents that are represented by 171 features, TFIDF scores for the different unigrams and bigrams obtained, and also results after the chi-squared test is shown in Figure 2.

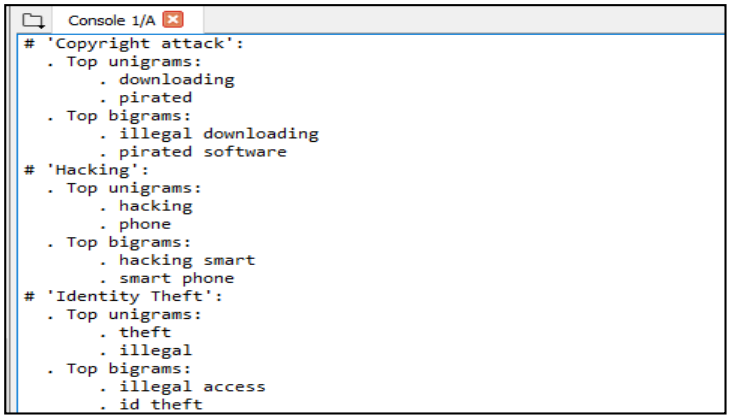


**Figure 2.** Correlated words by applying data preprocessing.

Figure 3 shows the total number of cybercrime incidents clustered under each cluster using the k-means clustering algorithm by considering the word count vectors present in the features provided by the cybercrime dataset. This is done by using the TFIDF vector technique and finding the correlated words corresponding to each incident. The TFIDF parameters are defined upon the case content adequately which facilitates the classification of what all the contents are valuable and unnecessary, respectively and then summarized accordingly. The summarized results of the k-means algorithm are shown in Figure 3.

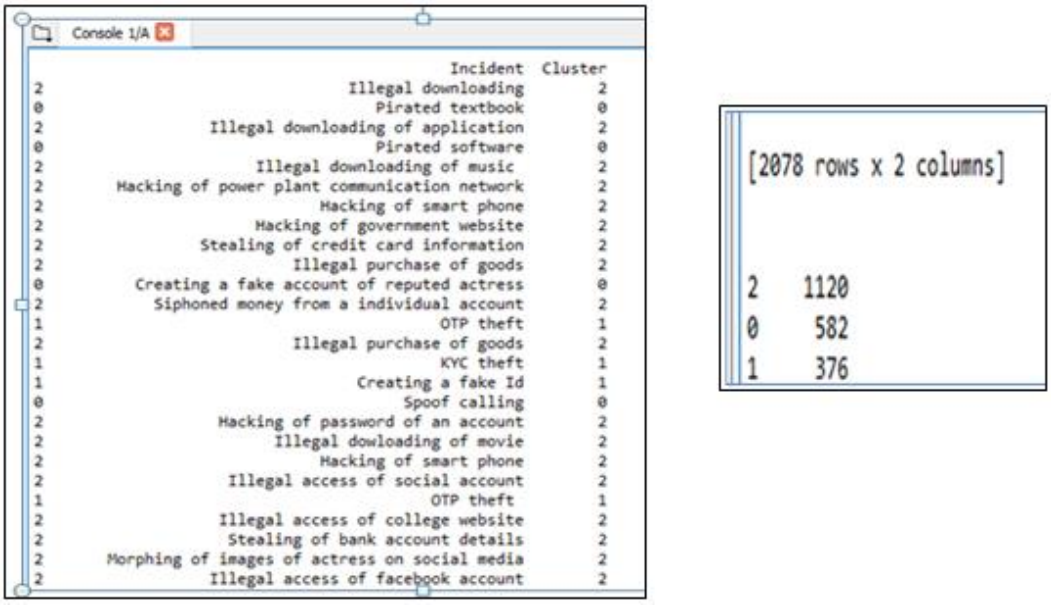


**Figure 3.** Total number of cybercrimes clustered under each cluster.

Figure 4 shows the clustering plot for the cybercrime incidents as per the dataset. This unsupervised learning algorithm makes use of the encoded data of each attribute present in the dataset. The encoded data were obtained by calculating the count vectors and the weighted vectors for the words corresponding to each feature.

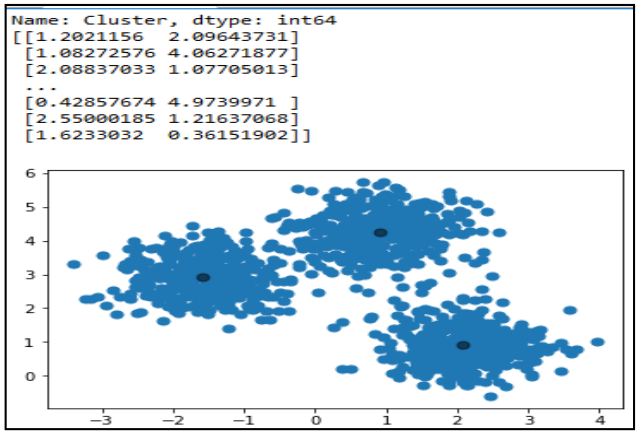


**Figure 4.** Plot for the clustered cybercrime data.

Using the proposed system, it was revealed that the crime rate is more in the Madhya Pradesh state, followed by Haryana. This analysis helps to take countermeasures to mitigate the crime rate state-wise, where the criminal cases registered are more.

Table 4 shows the over-all occurrence of the cybercrime incidents in India during certain specified periods. It demonstrates that the occurrence of the Identity (ID) theft is more when compared to the other two attacks. We can take some countermeasures in order to reduce the existence of the ID theft attack in India. The existence of the copyright attack and hacking is also more.

**Table 4.** Overall occurrence of cybercrime in India.

| | Crime Type | Total Crimes |
|---|---|---|
| **0** | Copyright attack | 466 |
| **1** | Hacking | 282 |
| **2** | ID theft | 868 |
| **3** | Others | 1923 |

Figure 5 demonstrates the accuracy rate of the proposed model and the other models such as logistic regression, random forest, linear *svc*, multinomial *nb*. From this figure, we can analyze that our model classifies data with 99% accuracy, which shows that the classification of the cybercrime offenses is done accurately by using our model. The random forest classifier does not work well for the proposed model.

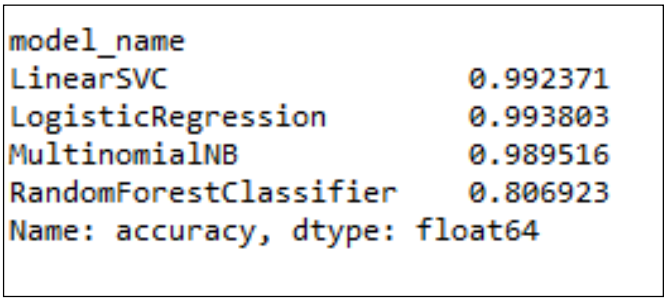


**Figure 5.** The accuracy rate of different classifiers.

Figure 6 demonstrates the precision, recall and f-1 score for our model. These can be obtained by using the confusion matrix obtained in our model and we can get the average accuracy rate for the cybercrimes that are predicted.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Identity Theft | 1.00 | 0.99 | 0.99 | 241 |
| Copyright attack | 0.97 | 1.00 | 0.98 | 95 |
| Hacking | 1.00 | 1.00 | 1.00 | 84 |
| avg / total | 0.99 | 0.99 | 0.99 | 420 |

**Figure 6.** Precision recall and f-1 score for the proposed model.

Precision: It is the measure of truly predicted positive samples to the total number of positively predicted samples. If the precision score is more then it represents that our model is pretty good to classify the samples.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

Recall: It is the measure of truly predicted positive samples of all the samples present in the actual class as yes. It is also termed as the sensitivity of the model.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

F1 score: It is calculated as the weighted average of both precision and recall. Its main components (considerations) are true negatives, true positives, false negatives and false positives. F1 score is preferred more than accuracy in order to know our classifier model performance measure, as shown in equation 7.

$$\text{F1 Score} = 2 \times (\text{precision} \times \text{recall}) \tag{7}$$

Accuracy is the performance measure used to check our model. It is preferred when the number of false positives values and the false negative values are the same. When the false-positive rates and the false negative rates are different then it is not much a good approach to check the performance of our classifier. In this situation, it is better to use f1 score rather than an accuracy measure. It can be calculated using

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{8}$$

Figure 7 depicts the confusion matrix for our model when the training size was 0.8 and the test size was 0.2. By this, we know how many cases are classified correctly and how many are classified incorrectly. It means we can find out the true negatives and true positives and false negatives and false positives classified by using the model.

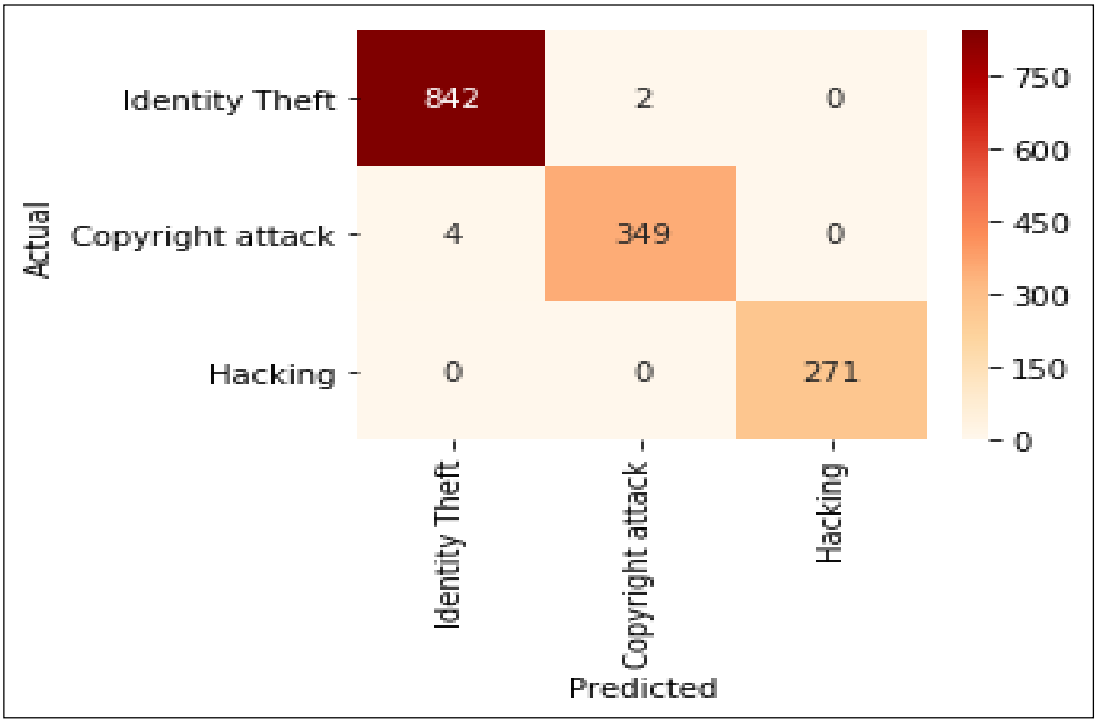


**Figure 7.** Confusion matrix for the proposed model.

*Test Cases*

Figure 8 shows the number of misclassified cybercrime offenses when the test size was 0.33. It demonstrates that the five identity theft cases were predicted as a copyright attack.

```
'Copyright attack' predicted as 'Identity Theft' : 4 examples.
           Cybercrime                                          Incident
1184  Copyright attack  Using the images with the taking permission of...
2096  Copyright attack                        Illegal dowloading of movie
847   Copyright attack                          Illegal access of account
29    Copyright attack  Using the images without the taking permission...

'Identity Theft' predicted as 'Copyright attack' : 2 examples.
         Cybercrime                                 Incident
210   Identity Theft          Illegal downloading of music
1625  Identity Theft  Posting an article without permission
```

**Figure 8.** Incorrect predicted cases when the test size was 0.33.

Figure 9 shows the number of misclassified incidents when the test size was increased to 0.5. It shows that the eight identity theft cases were misclassified as copyright attacks and one copyright attack was misclassified as identity theft and one identity theft case was misclassified as a hacking attack.

```
'Copyright attack' predicted as 'Identity Theft' : 1 examples.
           Cybercrime                    Incident
847  Copyright attack   Illegal access of account

'Identity Theft' predicted as 'Copyright attack' : 8 examples.
          Cybercrime                                        Incident
194   Identity Theft                     Illegal access of music files
1875  Identity Theft                     Illegal access of music files
31    Identity Theft      Infecting the system without owners knowledge
210   Identity Theft                      Illegal downloading of music
1281  Identity Theft  Using the images with the taking permission of...
500   Identity Theft                     Illegal access of music files
185   Identity Theft                     Illegal access of music files
1320  Identity Theft  Using the images with the taking permission of...

'Identity Theft' predicted as 'Hacking' : 1 examples.
       Cybercrime                                Incident
35  Identity Theft  Illegal access of the company website
```

**Figure 9.** Incorrect predicted cases when the test size was 0.5.

Figure 10 demonstrates the total number of misclassified incidents when the test size was increased to 0.6. It shows that the twelve identity theft cases were misclassified as copyright attacks and two copyright attacks were misclassified as identity theft and one hacking attack was misclassified as identity theft attack.

```
'Copyright attack' predicted as 'Identity Theft' : 2 examples.
           Cybercrime                              Incident
847  Copyright attack            Illegal access of account
832  Copyright attack   Illegal access of college website

'Hacking' predicted as 'Identity Theft' : 1 examples.
     Cybercrime           Incident
1397    Hacking   Harassing person

'Identity Theft' predicted as 'Copyright attack' : 12 examples.
          Cybercrime                                                Incident
194   Identity Theft                      Illegal access of music files
1875  Identity Theft                      Illegal access of music files
210   Identity Theft                       Illegal downloading of music
1281  Identity Theft  Using the images with the taking permission of...
500   Identity Theft                      Illegal access of music files
185   Identity Theft                      Illegal access of music files
1320  Identity Theft  Using the images with the taking permission of...
1376  Identity Theft  Using the images with the taking permission of...
1625  Identity Theft              Posting an article without permission
640   Identity Theft              Posting an article without permission
1734  Identity Theft  Using the images with the taking permission of...
59    Identity Theft  Using the images with the taking permission of...
```

**Figure 10.** Cases predicted incorrectly when the test size was 0.6.

Figure 11 shows the number of misclassified incidents when the test size was 0.4. It shows that the four copyright attacks were misclassified as identity theft attacks and one identity theft attack was misclassified as copyright attack.

```
'Copyright attack' predicted as 'Identity Theft' : 3 examples.
          Cybercrime                                          Incident
1184  Copyright attack  Using the images with the taking permission of...
2096  Copyright attack                       Illegal dowloading of movie
847   Copyright attack                         Illegal access of account

'Identity Theft' predicted as 'Copyright attack' : 1 examples.
        Cybercrime                      Incident
210  Identity Theft  Illegal downloading of music
```

**Figure 11.** Cases predicted incorrectly when the test size is 0.4.

## 5. Conclusions and Future Scope

In the present world, cybercrime offenses are happening at an alarming rate. As the use of the Internet is increasing many offenders, make use of this as a means of communication in order to commit a crime. The framework developed in our work is essential to the creation of a model that can support analytics regarding the identification, detection and classification of the integrated cybercrime offenses (structured and unstructured). The main focus of our work is to find the attacks that take advantage of the security vulnerabilities and analyze these attacks by making use of machine learning techniques. The aim is that the developed framework will provide the essential broad knowledge of cybercrime offenses in the society, enable them to consider the threat landscape of such attacks and avoid the incarnation of the cybercrime offenses. From the results, it is evident that the developed framework reduces the time consumption and manual reporting process. It helps to identify the number of filing cases by incident wise and area-wise. This report is useful to predict the cases and to take precautionary steps against filing cybercrime cases on certain hot-spot places identified.

*Future Scope*

The current model works only with the classification and clustering of the cybercrime patterns. A feature extension needs to be considered into account in order to provide some countermeasures and custom actions to the crime agencies in order to reduce the growth of frequently occurred cybercrimes in the specific location. It is used to provide some standard protocols and procedures which can be used to automate a crime management organization. Furthermore, the information processed by our model would be logged as an aggregator to generate some statistics regarding the occurrence of cybercrime offenses, which leads to the optimization of the supervision of cybercrime offenses by the relevant information security agencies. In the future, the features of the frames-work can be enhanced by using deep learning approaches in the prediction of crime cases area-wide.

**Author Contributions:** Conceptualization, R.C. and T.R.G.; methodology, R.C. and T.R.G.; software, R.C. and T.R.G.; formal analysis, R.C., T.R.G. and M.H.A.; resources, T.R.G., M.H.A. and A.A.-A.; data curation, R.C. and M.H.A..; writing—original draft preparation, R.C. and T.R.G.; writing—review and editing, T.R.G., M.H.A. and A.A.-A.; supervision, T.R.G. and A.A.-A.; funding acquisition, M.H.A. and A.A.-A. All authors have read and agreed to the published version of the manuscript.

**Funding:** The authors are grateful to the Raytheon Chair for Systems Engineering for funding.

**Acknowledgments:** The authors are grateful to the Raytheon Chair for Systems Engineering for funding.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Singh, A.K.; Prasad, N.; Narkhede, N.; Mehta, S. Crime: Classification and Pattern Prediction. *Int. Adv. Res. J. Sci. Eng. Technol.* **2016**, *3*, 41–43. [CrossRef]

2. Brar, H.S.; Kumar, G. Cybercrimes: A Proposed Taxonomy and Challenges. *J. Comput. Netw. Commun.* **2018**, *2018*, 1–11. [CrossRef]
3. Pete, I.; Chua, Y.T. An Assessment of the Usability of Cybercrime Datasets. In Proceedings of the CSET @ USENIX Security Symposium, Santa Clara, CA, USA, 12 August 2019.
4. Ngo, F.; Jaishankar, K. Commemorating a Decade in Existence of the International Journal of Cyber Criminology: A Research Agenda to Advance the Scholarship on Cyber Crime. *Int. J. Cyber Criminol.* **2017**, *11*, 1–9.
5. Khusna, A.N.; Agustina, I. Implementation of Information Retrieval Using Tf–idf Weighting Method On Detik.Com's Website. In Proceedings of the 2018 12th International Conference on Telecommunication Systems, Services and Applications (TSSA), Yogyakarta, Indonesia, 4–5 October 2018; pp. 1–4.
6. Zhang, G.Z. Computer Forensics Based on Data Mining. *Appl. Mech. Mater.* **2014**, *536–537*, 371–375. [CrossRef]
7. Numan, M.; Subhan, F.; Khan, W.Z.; Hakak, S.; Haider, S.; Reddy, G.T.; Jolfaei, A.; Alazab, M. A Systematic Review on Clone Node Detection in Static Wireless Sensor Networks. *IEEE Access* **2020**, *8*, 65450–65461. [CrossRef]
8. Iwendi, C.; Jalil, Z.; Javed, A.R.; Gadekallu, T.R.; Kaluri, R.; Srivastava, G.; Jo, O. KeySplitWatermark: Zero Watermarking Algorithm for Software Protection against Cyber-Attacks. *IEEE Access* 2020. [CrossRef]
9. Bhattacharya, S.; Somayaji, S.R.K.; Maddikunta, K.P.; Kaluri, R.; Singh, S.; Gadekallu, R.T.; Alazab, M.; Tariq, U. A Novel PCA-Firefly Based XGBoost Classification Model for Intrusion Detection in Networks Using GPU. *Electronics* **2020**, *9*, 219. [CrossRef]
10. Jia, X.; He, D.; Kumar, N.; Choo, K.-K.R. Authenticated key agreement scheme for fog-driven IoT healthcare system. *Wirel. Netw.* **2019**, *25*, 4737–4750. [CrossRef]
11. Wu, L.; Zhang, Y.; Ma, M.; Kumar, N.; He, D. Certificateless searchable public key authenticated encryption with designated tester for cloud-assisted medical Internet of Things. *Ann. Telecommun.* **2019**, *74*, 423–434. [CrossRef]
12. Aggarwal, S.; Shojafar, M.; Kumar, N.; Conti, M. A New Secure Data Dissemination Model in Internet of Drones. In Proceedings of the ICC 2019–2019 IEEE International Conference on Communications (ICC), Shanghai, China, 20–24 May 2019; pp. 1–6.
13. Wang, T.; Zheng, Z.; Bashir, A.K.; Jolfaei, A.; Xu, Y. FinPrivacy: A Privacy-Preserving Mechanism for Fingerprint Identification. *ACM Trans. Internet Technol.* **2018**, *37*, 111–116.
14. Al Ridhawi, I.; Otoum, S.; Aloqaily, M.; Jararweh, Y.; Baker, T. Providing secure and reliable communication for next generation networks in smart cities. *Sustain. Cities Soc.* **2020**, *56*, 102080. [CrossRef]
15. Alloghani, M.; Baker, T.; Al-Jumeily, D.; Hussain, A.; Mustafina, J.; Aljaaf, A.J. A Systematic Review on Security and Privacy Issues in Mobile Devices and Systems. In *Handbook of Computer Networks and Cyber Security*; Gupta, B., Perez, G., Agrawal, D., Gupta, D., Eds.; Springer: Cham, Germany, 2020; pp. 585–608. [CrossRef]
16. Reddy, G.T.; Swarna Priya, R.M.; Parimala, M.; Chowdhary, C.L.; Reddy, P.K.; Hakak, S.; Khan, W.Z. A deep neural networks based model for uninterrupted marine environment monitoring. *Comput. Commun.* **2020**, *157*, 64–75. [CrossRef]
17. Patel, H.; Singh Rajput, D.; Thippa Reddy, G.; Iwendi, C.; Kashif Bashir, A.; Jo, O. A review on classification of imbalanced data for wireless sensor networks. *Int. J. Distrib. Sens. Netw.* **2020**, *16*, 1550147720916404. [CrossRef]
18. Reddy, G.T.; Reddy, M.P.K.; Lakshmanna, K.; Kaluri, R.; Rajput, D.S.; Srivastava, G.; Baker, T. Analysis of Dimensionality Reduction Techniques on Big Data. *IEEE Access* **2020**, *8*, 54776–54788. [CrossRef]
19. Ganesan, M.; Mayilvahanan, P. Cyber Crime Analysis in Social Media Using Data Mining Technique. *Int. J. Pure Appl. Math.* **2017**, *116*, 413–424.
20. Khan, M.A.; Pradhan, S.K.; Fatima, H. Applying Data Mining techniques in Cyber Crimes. In Proceedings of the 2017 2nd International Conference on Anti-Cyber Crimes (ICACC), Abha, Saudi Arabia, 26–27 March 2017; pp. 213–216.
21. Nouh, M.; Nurse, J.R.C.; Goldsmith, M. Towards Designing a Multipurpose Cybercrime Intelligence Framework. In Proceedings of the 2016 European Intelligence and Security Informatics Conference (EISIC), Uppsala, Sweden, 17–19 August 2016; pp. 60–67.

22. Prasanthi, M.S.; Ishwarya, T.A.S.K. Cyber Crime Prevention & Detection. *Int. J. Adv. Res. Comput. Commun. Eng.* **2015**, *4*, 45–48. [CrossRef]
23. Soomro, T.R.; Mumtaz, H. Social Media-Related Cybercrimes and Techniques for Their Prevention. *Appl. Comput. Syst.* **2019**, *24*, 9–17. [CrossRef]
24. Çağrı, B.A.; Sağlam, R.B.; Li, S. Automatic Detection of Cyber Security Related Accounts on Online Social Networks: Twitter as an example. In Proceedings of the 9th International Conference on Social Media and Society, Copenhagen, Denmark, 18–20 July 2018; pp. 236–240.
25. Chen, H.; Chung, W.; Xu, J.J.; Wang, G.; Qin, Y.; Chau, M. Crime data mining: A general framework and some examples. *Computer* **2004**, *37*, 50–56. [CrossRef]
26. Prabakaran, S.; Mitra, S. Survey of Analysis of Crime Detection Techniques Using Data Mining and Machine Learning. *J. Phys. Conf. Ser.* **2018**, *1000*, 012046. [CrossRef]
27. Chauhan, C.; Sehgal, S. A review: Crime analysis using data mining techniques and algorithms. In Proceedings of the 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 5–6 May 2017; pp. 21–25.
28. An, J.; Kim, H. A Data Analytics Approach to the Cybercrime Underground Economy. *IEEE Access* **2018**, *6*, 26636–26652. [CrossRef]
29. Tsakalidis, G.; Vergidis, K. A Systematic Approach Toward Description and Classification of Cybercrime Incidents. *IEEE Trans. Syst. Man Cybern. Syst.* **2019**, *49*, 710–729. [CrossRef]
30. Tsakalidis, G.; Vergidis, K.; Madas, M. Cybercrime Offenses: Identification, Classification and Adaptive Response. In Proceedings of the 2018 5th International Conference on Control, Decision and Information Technologies (CoDIT), Thessaloniki, Greece, 10–13 April 2018; pp. 470–475.
31. Gangavane, H.N.; Nikose, M.C. A Survey on Document Clutering for identifying Criminal. *Int. J. Adv. Res. Artif. Intell.* **2015**, *2*, 459–463. [CrossRef]
32. Zubi, Z.S.; Mahmmud, A.A. Crime Data Analysis using Data mining Techniques to Improve Crimes Prevention. *Int. J. Comput.* **2014**, *8*, 39–45. [CrossRef]
33. Sudha, T.S.; Rupa, C. Analysis and Evaluation of Integrated Cyber Crime Offenses. In Proceedings of the 2019 Innovations in Power and Advanced Computing Technologies (i-PACT), Vellore, India, 22–23 March 2019; pp. 1–6.
34. Reddy, G.T.; Sudheer, K.; Rajesh, K.; Lakshmanna, K. Employing data mining on highly secured private clouds for implementing a security-asa-service framework. *J. Theor. Appl. Inf. Technol.* **2014**, *59*, 317–326.
35. Kigerl, A. Cyber Crime Nation Typologies: K-Means Clustering of Countries Based on Cyber Crime Rates. *Int. J. Cyber Criminol.* **2016**, *10*, 147–169. [CrossRef]
36. Wu, H.; Yuan, N. An Improved TF–IDF algorithm based on word frequency distribution information and category distribution information. In Proceedings of the 3rd International Conference on Intelligent Information Processing, Guilin, Chin, 4–6 May 2018; pp. 211–215.
37. Zheng, M.; Robbins, H.; Chai, Z.; Thapa, P.; Moore, T. Cybersecurity Research Datasets: Taxonomy and Empirical Analysis. In Proceedings of the International Conference on Cyber Security Experimentation and Test, Baltimore, MD, USA, 13 August 2018.
38. Wang, C.; Yang, B.; Luo, J. Identity Theft Detection in Mobile Social Networks Using Behavioral Semantics. In Proceedings of the 2017 IEEE International Conference on Smart Computing (SMARTCOMP), Hong Kong, China, 29–31 May 2017; pp. 1–3.
39. Zhijun, L.; Ning, W. A Cyber Crime Investigation Model Based on Case Characteristics. In Proceedings of the 2017 4th International Conference on Information Science and Control Engineering (ICISCE), Changsha, China, 21–23 July 2017; pp. 11–15.
40. Delamaire, L.; Abdou, H.; Pointon, J. Credit Card fraud and Detection techniques: A review. *Banks Bank Syst.* **2009**, *4*, 57–68.
41. Roul, R.K.; Sahoo, J.K.; Arora, K. Modified TF–IDF Term Weighting Strategies for Text Categorization. In Proceedings of the 2017 14th IEEE India Council International Conference (INDICON), Roorkee, India, 15–17 December 2017; pp. 1–6.
42. Sonawane, T.R.; Al-Shaikh, S.; Shinde, R.; Shaikh, S.; Sayyad, A.G. Crime Pattern Analysis Visualization and Prediction using Data Mining. *Int. J. Adv. Res. Innov. Ideas Educ.* **2015**, *1*, 681–686.

43. Williams, M.L.; Burnap, P.; Sloan, L. Crime Sensing With Big Data: The Affordances and Limitations of Using Open-source Communications to Estimate Crime Patterns. *Br. J. Criminol.* **2016**, *57*, 320–340. [CrossRef]
44. Agarwal, A.; Chougule, D.; Agarwal, A.; Chimote, D. Application for Analysis and Prediction of Crime data using Data mining. *Int. J. Adv. Comput. Eng. Netw. (IJACEN)* **2016**, *4*, 9–12.