JCST Journal of Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7766140

DETECTION OF CYBER ATTACK AND PERPETRATOR PREDICTION USING SUPERVISED MACHINE LEARNING ALGORITHM

Sowmiya Sree C
"M.Tech CFIS, Department of Computer Science and Engineering, Dr. M.G.R Educational and Research Institute, Maduravoyal, Chennai-600 095".

Dr. G.Soniya Priyatharsini
Associate Professor, Department of Computer Science and engineering, Dr. M.G.R Educational and Research Institute, Maduravoyal, Chennai-600 095".

Dr. S. Geetha
"Head of the Department, Department of Computer Science and Engineering, Dr. M.G.R Educational and Research Institute, Maduravoyal, Chennai-600095".

Abstract— One of the major issues facing the globe now is cyberattacks. Every day, they lead individuals and nations to suffer severe financial losses. Cybercrime is also on the rise along with cyberattacks. The ability to recognize cybercriminals and comprehend their methods of attack is essential in the fight against crime & criminals. Cyber-attacks may be difficult to recognize and prevent. However, scholars have recently developed security systems and made predictions using AI (Artificial Intelligence) techniques to resolve the issues. There are several crime prediction techniques listed in the literature. However, they struggle to forecast the strategies used in cybercrime and cyberattacks. The solution to this issue is to use actual data to identify an assault and its perpetrator. The information includes the sort of crime, the perpetrator's gender, the damage, and the assault techniques. Applications made by people who were the target of cyberattacks may provide the forensic units with the necessary data. In this study, we use Machine Learning (ML) to assess two alternative cybercrime models and project the impact of the stated variables on the identification of the cyberattack technique and the perpetrator. In our methodology, we used eight ML techniques and found that their accuracy rates were comparable. With an accuracy rate of 95.02 percent, the Support Vector Machine (SVM) was determined to be the most effective cyberattack technique. With a high degree of accuracy, the initial model allowed us to forecast the sorts of assaults that the victims were most likely to experience. The most accurate approach for finding attackers was the Logistic Regression, which had a 65.42 percent accuracy rate. In the 2nd model, we forecasted whether a comparison of the perpetrators' traits would allow for their identification. Our findings indicate that the likelihood of a cyberattack diminishes with the victim's level of education and income. We anticipate cybercrime units will implement the recommended approach. It would also ease the identification of cyberattacks and make combating them simpler and more efficient.

Keywords: Cyber-attack-crimes, Machine learning, Artificial intelligence, Data analysis, Crime prediction, Security and privacy

1. INTRODUCTION:

Cyberattacks are expanding quickly, making present detection techniques ineffective and increasing the need to develop more accurate prediction models and methods. This challenge is still unresolved since present attack prediction algorithms are unable to keep up with the vast number and diversity of attacks. Researchers have recently shown a great deal of interest in ML methodologies, particularly DL (Deep Learning) techniques, due to their unequaled high performance in various prediction-based domains. Countries are attempting to adapt to this field and safeguard cyberspace security. Nations need to protect their essential infrastructure. These include the chemical, financial, health, and energy industries, as well as nuclear power facilities in certain nations. Financial losses significantly rise daily as a result of the millions of cyberattacks. Data stolen from the Airbus Company's data system were sold on the dark web in 2020. Due to assaults on several cities, medical information for millions of individuals has been taken, and even an emergency has been proclaimed. Availability, nonrepudiation, authorization, authentication, confidentiality, and integrity, are the key components in guaranteeing cyber security. New solutions are needed as the workforce becomes insufficient to combat cyber incidents every day. Solutions such as intrusion detection systems, smart cyber security assistant designs, and autonomous cyber defence systems are being considered to combat cyber-attacks and crimes. ML techniques are used by scholars to avoid Internet of Things (IoT) vulnerabilities and identify power disruptions brought on by cyberattacks. Other applications include identifying spam and network attacks, spotting banking sector phishing attempts, and lowering sexual crimes on social media. These techniques were used in areas including pattern recognition, stock prediction, criminal identity identification, and crime prevention. Our research has three primary goals. The first is to anticipate a cybercrime technique using real cybercrime data as input and compare the accuracy outcomes. The 2nd objective is to determine whether cybercriminals could be anticipated using the existing data. The 3rd purpose is to study the influence of cyber-attack victim profiles. This work used actual 5-year Turkish data on cybercrime. The attack's methodology was anticipated using machine learning techniques, and the attacker was located. Employment, marital status, education, income, gender, age, and the extent of the crime's damage were among the factors used in the detection process. We were able to forecast the types of attack tactics and the victims of these assaults by focusing on certain characteristics like age, gender, etc. The police forces tasked with combating cybercrimes will apply these findings in attack detection and cybercrime modeling.

Significant contributions of the suggested method include:

- Giving the cybercrime department benefits by enabling the use of actual data;
- Enabling the prediction of potential assaults on victims;
- Enabling comparison of ML algorithms to determine the optimum performance

This research investigates the use of ML approaches for anticipating cyberattacks and their perpetrators in this context.

Preparing the Dataset: The 1802 records of extracted features in this dataset were divided into two classes and eight categories.

- Attack approach (Comprises six Categories)
- The perpetrator (Comprises two Categories)

2. LITERATURE SURVEY:

ML algorithms may be trained and applied to examine if a cyberattack has occurred. When a threat is detected, security engineers or users may get an email notification. Any classification approach could be utilized to examine if an attack is a DDoS/DoS or not. Classification algorithms, like SVM, are a kind of supervised learning technique used to examine data and identify patterns. Our best hope at the moment is early detection, which would assist limit the risk of irreversible harm such attacks might do, since we could not predict when, when, or how an attack might come our way and total protection against them could not yet be assured. Organizations may use current technologies or develop their own to identify cyber assaults at a very early stage to mitigate their effects. The ideal system would need the least amount of human involvement.

2.1 Enhancing Cyber Security Assurance Model:

Every time a group of auditors participates in an IT, compliance audit, or data security, there are recurring phases such as designing, defining objectives & scope, clarifying terms of engagements, performing the audit, obtaining supporting documentation, assessing risks, reporting the audit outcomes, and scheduling follow-up tasks. Designing an audit is similar to designing an audit of any other sort. However, given the high level of several cyber security areas, this will need a lot of work. However, the scope of the internal audits does not include reviewing most cyber capabilities. This particular system contains compliance/risk management, security program, development life cycle, 3rd party management, asset/information management, access management, threat/vulnerability management, and the need for assurance which will be attained by management reviews, information management & protection, cyber risk assessments, crisis management, risk analytics, and response. Additionally, Deloitte's system is in line with industry frameworks like COSO ("Committee of Sponsoring Organizations of the Treadway Commission"), ITIL ("Information Technology Infrastructure Library"), NIST ("National Institute of Standards and Technology"), and ISO ("International Organization for Standardization") Equations.

3. EXISTING SYSTEM:

If a cyberattack has occurred, ML algorithms may be trained to recognize it. An email notification could be sent to security engineers or users as soon as the threat is found. Any classification approach might be applied to examine whether or not an attack is a DoS and/or DDoS. Classification algorithms, like SVM, are a kind of supervised learning technique used to examine data and identify patterns. Given that we could not know when, how, or when an attack might come our way and that complete protection against them cannot yet be guaranteed, early detection is currently our best hope of preventing any irreparable damage that such assaults may inflict.

4. PROPOSED SYSTEM:

The objective of the suggested model is to develop an anomaly detection model using ML. Anomaly detection is a crucial technique for identifying suspicious activity, fraud, network intrusion, and other anomalous occurrences that could be very considerable but are hard to spot. The dependent and independent variables of the ML g model are identified using appropriate data science approaches. The data is then visualized to get insights from the data.

The model is constructed based on a prior dataset from which the algorithm learned data and was trained to make more accurate comparisons. Calculations and comparisons are done for the performance measures. All cybercrime information was analyzed when the dataset was acquired. Different data science techniques were used to eliminate the duplicate sections. Additionally, color groupings are used to separate the specifics of these four variables. Predictions were produced with this data using the Python 3 program and a variety of modules. To show the data, this program's primary libraries, including Matplotlib, Pandas, and Numpy, were employed. The main advantages of using ML methods, according to the article, include the capacity to identify multiple patterns in structured/unstructured data, high success rates in finding shifting criminal tactics, correlations between complex data, and the capacity to generate outcomes that are unexpected to humans.

4.1 MODELING:

4.1.1 Data Pre-processing

4.1.2 Data Visualization Analysis

4.1.3 comparing algorithms with predictions that provide the highest levels of accuracy

4.1.4 Deployment Using Flask

5. SYSTEM DESIGN:

The goal of this investigation is to determine which characteristics are most useful for foretelling network assaults. employed discrete class prediction techniques from machine learning to fit a function that could estimate the incoming input's discrete class.

The repository serves as a learning exercise to:

- Apply the fundamentals of ML to a dataset that is currently accessible. Analyze and interpret my findings and justify my interpretation on the basis of the observed dataset.
- Analyze statistical data and outcomes that are displayed to identify the common trends across all regiments.

Project Goals

- **ANALYSIS OF EXPLORATION DATA FOR VARIABLE IDENTIFICATION**

□ Loading the specified dataset

□ Import the required library packages.

□ Determine the general characteristics

□ Identify duplicate and empty values

□ Examining unique 7-count values

- **UNIVARIATE DATA ANALYSIS**

□ Rename, add, and then remove the data

□ to define the data form

- **DATA ANALYSIS USING BI & MULTIVARIATE EXPLORATION**

□ Bar chart, heatmap, pair plot, along with histogram plot diagram

- **OUTLIER DETECTION APPROACH USING FEATURE ENGINEERING**

□ Pre-processing the specified dataset

□ Splitting the training and test datasets

□ Comparing random forest with the logistic regression model

- **ALGORITHM COMPARISON TO PREDICT THE OUTCOME**
- On the basis of the best accuracy

5.1 SYSTEM ARCHITECTURE

Data Preprocessing

Rescaling features in a normal distribution is referred to as standardization. This must be finished before using machine learning techniques. Numbers from 1 to 10 were provided in accordance with the diversity of data in the columns to make the data appropriate. Data optimization for use in algorithms was done using the Python library's Standard Scaler (). The bidirectional link between kind of crime & damages, as well as assault and attack strategy. 80% of the data have been training data and 20% were test data. In the 1st model, it was attempted to predict the technique of attack by providing information on the perpetrator, assault, education, marital status, income, job, age, gender, and crime. In the second model, attempts were made to forecast the crime's perpetrator by providing information on the crime's characteristics, including income, gender, age, employment, education, marital status, and the attack's purpose and technique.

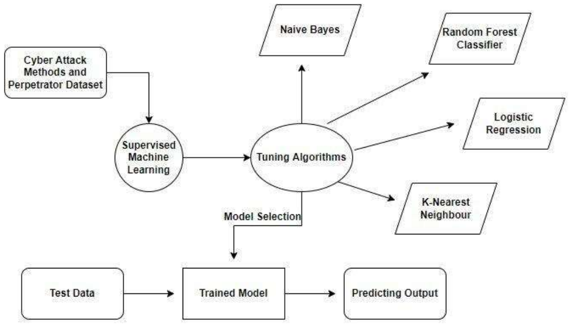


Figure 1: Architecture diagram

1) Logistic regression

It is an eq. that, based on the connection between two or more variables, enables one to anticipate the value of one from the other. The coefficients may be resolved using the least squares approach, supposing that the variance of y is constant. This reduces the error between the regression line and the actual data point. The LR line's eq. is as follows: (1)

$y'=b+w1x1+w1x1+\ldots+wnxny'=b+w1x1+w1x1+\ldots+wnxn$

Here,

y′ indicates desired O/P,
b represents bias value,
x denotes the input property,
w shows the feature's weight

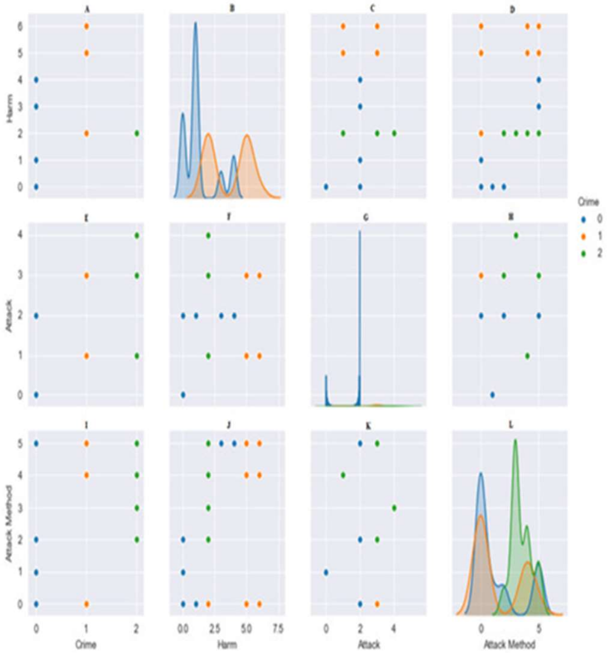


Figure 2: Data Visualization

(1) Harm: Crime. (2) Harm: Harm. (3) Harm: Attack. (4) Harm: Attack Approach. (5) Attack: Crime. (6) Attack: Harm. (7) Attack: Attack. (8) Attack: Attack Approach. (9) Attack Approach: Crime. (10) Attack Approach: Harm. (10) Attack Approach: Attack. (11) Attack Approach: Attack Approach.

2) K-nearest neighbors

The distance function is the foundation of the KNN classifier, and it is used to measure the difference as well as the similarity between 2 or more samples. The "Euclidean distance" *d* (*x*, *y*) in 2 samples is described as:

(2)

d(x,y)=∑k=1n(xk+yk)2d(x,y)=∑k=1n   (xk+yk)2

Here,
Xk & yk indicate the nth element.
n indicate the nth property of the dataset.

The parameter (k) is estimated, and then, based on the existing data, the distance of new data to be added to the dataset is estimated one by one. The nearest neighbor is identified and given to the neighbor class k.

3) SVM

Support vector regression and classification are included in this approach. Based on the idea of decision limitations, SVM can classify data in both binary and more complex ways. Taking into account the training data (D):

(3) D={(xi,yi) |xi∈Rp, yi∈{−1,1}}ni=1D={(xi,yi) | xi∈Rp, yi∈{−1,1}}i=1n

The class to which the point of xi belongs in this case is determined by yi yiis1 or −1. Each xi is a real vector in p-dimensional. The support vector is located nearest to the hyperplane of ideal separation. Locations on the other side of the separation hyperplane and mapping I/P vectors on the other side of the plane are classified into many groups.

4) Naive Bayes

The NB classifier is a straightforward probability classifier that uses the “Bayes' theorem” application together with independence requirements between properties. There is a multi-label classification model-based NB that may be used. Given a dataset labeled {a1a1, a2a2,..., ajaj} and { v1v1, v2v2,..., vjvj},, the following equation is used to predict the outcomes:

(4) vNB = arg max vj∈VP(vj) ∏ iPai | vjvNB = arg max vj ∈ VP(vj) ∏iP ai |vj

5) Decision Tree (DT)

A DT is a classification technique that builds a model like a tree with “decision nodes” and leaf nodes according to classification, feature, along with the target. Every internal node of a classification tree, also recognized as a DT, is labeled with an input attribute. A sub-decision node in a distinct input attribute is reached by arcs from the tagged node that are labeled with all of the target attribute's potential values. By segmenting the resources into subgroups in accordance with an attribute value test, a tree may learn. On each derived subset, this procedure is done recursively in a method known as recursive partitioning. When a node's subset has the whole target variable’s value or the division stops adding value to the predictions, the iteration is finished.

(5) (x,Y) = (x1,x2,x3,…,xk,Y)(x,Y)=(x1,x2,x3,…,xk,Y)

The dependent variable Y is the “target variable” that we are attempting to comprehend, generalize, or categorize. The input variables utilized for this work, like x1,x2,x3x1,x2, and x3 make up the vector *x*.

6) Random Forest (RF)

A classifier is created using the RF method on training data, and the outputs are combined to provide the best estimations on test data. Randomness reduces variation to prevent overlearning from training data.

(6)

“y=argmaxp∈{h(x1)..h(xk)}{∑j=1k(I(h(x|θj)=p))}y=argmaxp∈{h(x1)..h(xk)} {∑j=1k (I(h(x|θj)=p))}”

Here,

h(x|θ) indicates classification tree,

k shows many trees chosen from a pattern random vector.

If D(x, y) presents the training data, every classification tree within the ensemble is built with a distinct subset “Dθ k(x,y) ⊂ D(x,y) Dθ k(x,y) ⊂ D(x,y)” of the training data. Then, every tree performs such as a typical DT. The data is partitioned into segments using a randomly generated value until the maximum depth is achieved or the partitioning is completed.

XGBoost (Extreme Gradient Boosting)

It begins with the 1st prediction (“base score”). This prediction may be any number since it will converge with the operations that will be carried out in the following phases to produce the right outcome. By default, this value is set to 0.

The 1st step is to develop the loss function “L(yi, yi ,*F*(*x*))”. The observed value is yiyi, *F*(*x*) indicates predicted value:

F0(x)=arg min γ∑i=1n(L(yi, γ)F0(x)=arg min γ ∑i=1n (L(yi,γ)

The constant variable is identified in this case. It is represented by the formula's sigma value loss function. γγ (gamma) indicates the predicted value.

rim=−[∂L(yi,F(xi))∂F(xi)]F(x)=Fm−1(x)i=1,….,nrim=−[∂L(yi,F(xi))∂F(xi)]F(x)=Fm−1(x)i=1,….,n

Here,

r indicates residual,

i denotes the observation number,
m represents the established tree's number.
Regression trees serve as the fundamental learning terminal nodes in tree growth. It is written as follows:
(9)γim=argminγ∑xiRij L(yi,Fm−1(xi)+γ)j=1Jmγim=argminγ ∑xiRij L(yi,Fm−1(xi)+γ)j=1…Jm
(10)Fm(x)=Fm−1(x)+∑j=1jmγjmI(x∈Rjm) Fm(x)=Fm−1(x)+∑j=1jm γjmI(x∈Rjm).

Accuracy (Acc), Precision(P), Recall (R), and F1-Score (F1)

The Acc score is an approach applied to assess the performance of the model by comparing the algorithm's predictions with the test data. A number between 0 & 1 is generated based on the ratio of the full projected value for a forecast to match the actual data. To assess the forecast's accuracy:

- TP=Prediction and actual is positive (normal).
- TN=Prediction and actual is negative(abnormal).
- FP=Prediction is positive(normal) as well as actual is negative(abnormal).
- FN=Prediction is negative(abnormal) & actual is positive(normal).

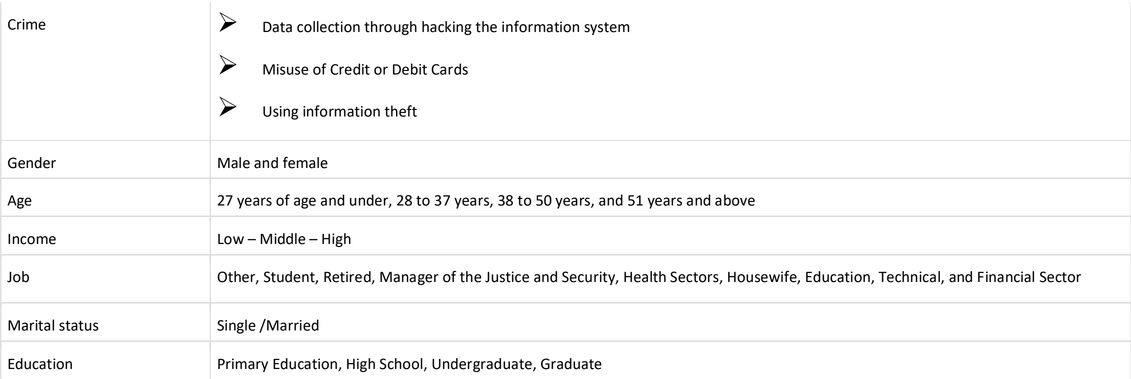
The suggested model's additional assessment criteria include recall (R), precision (P), and *F1*. P is the proportion of cases accurately identified as positive to the entire number of positive cases. R indicates the accuracy with which positive cases are anticipated. *F1* is the weighted mean of P and R.
(11) Acc=TP+TNTP+TN+FP+FNAcc =TP+TNTP+TN+FP+FN
(12) P=TPTP+FPP=TPTP+FP
(13) R=TPTP+FNR=TPTP+FN
(14) F1=2TP2TP+FN+FP.

RESULTS:

The study's objectives include accurate incident data analysis, crime prevention, and the capture of offenders. The major focus of this essay is to make inferences from the data that has been evaluated and, using those inferences, fight crime. These findings will come to light and throw light on any hidden facts as well as the investigations conducted by law enforcement officials. Machine learning techniques may be used to assess whether the same culprit was responsible for the cyberattack based on the victim's information, the cyberattack's methodology and whether the perpetrator has been identified or not. Through a variety of approaches, the losses incurred by the victims of cyber events in the province of Elazığ have been uncovered throughout time. Each victim's financial losses in the dataset were totaled by adding across the years. It is believed that the deterrent provided by the legislation and awareness campaigns is the cause of the decline in such instances, which was particularly evident after 2017.

6. FIGURES AND TABLE:

POSITIONING FIGURES AND TABLES



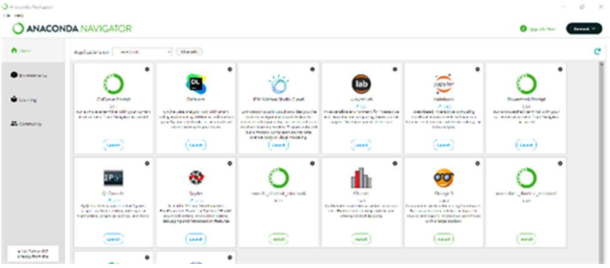
Crime	➢ Data collection through hacking the information system ➢ Misuse of Credit or Debit Cards ➢ Using information theft
Gender	Male and female
Age	27 years of age and under, 28 to 37 years, 38 to 50 years, and 51 years and above
Income	Low – Middle – High
Job	Other, Student, Retired, Manager of the Justice and Security, Health Sectors, Housewife, Education, Technical, and Financial Sector
Marital status	Single /Married
Education	Primary Education, High School, Undergraduate, Graduate

Harm	➢ Obtaining Money by Claiming to Be a Bank Employee ➢ withdrawals made without being informed ➢ Blackmailing/Threatening ➢ Shopping online without knowledge ➢ Using the Internet to Shop by Posing as a Bank Clerk ➢ Do Moral Harm ➢ Fraud
Attack	➢ ATM Cards/Copying Bank ➢ Social Media Accounts Hacking ➢ Data Acquisition and Utilization in a Digital Environment ➢ Making Electronic Bank Accounts Available ➢ Providing Fake products for Sale
Attack technique	➢ Social Engineering ➢ Utilizing social media to get public data ➢ Phishing Attack ➢ Hacking Tools or Malware ➢ Creating a Fake Shopping Site ➢ Generating Devices/Card Copying
Perpetrator	Known/Unknown

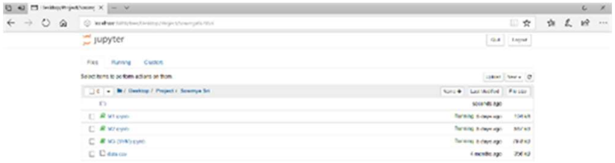
The attributes used for feature selection.

7. OUTPUT SCREENSHOTS:

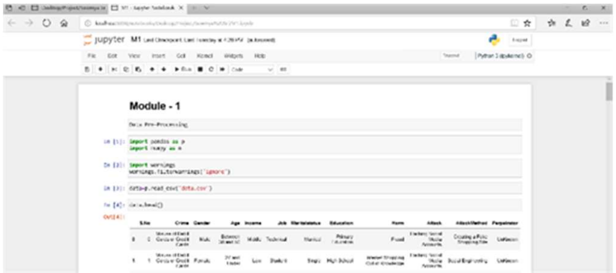
Step 1: Installing and Downloading Anaconda Navigator.



Step 2:In the anaconda navigator, launch jupyter notebook 6.4.5. Desktop>Project>Sowmya sri>M1



Step 3: Module 1 is data-preprocessing.



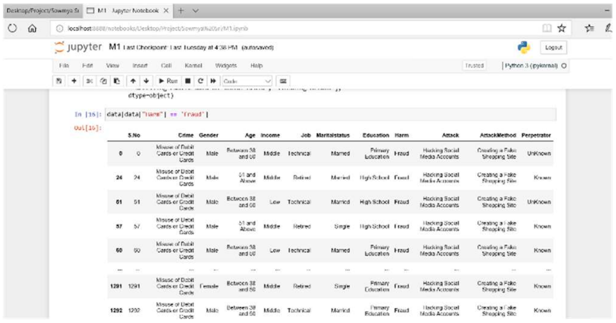
Module - 1

Data Pre-Processing

```
In [1]: import pandas as p
        import numpy as n

In [2]: import warnings
        warnings.filterwarnings('ignore')

In [3]: data=p.read_csv('data.csv')

In [4]: data.head()
Out[4]:
```

	S.No	Crime	Gender	Age	Income	Job	Maritalstatus	Education	Harm	Attack	AttackMethod	Perpetrator
0	0	Misuse of Debit Cards or Credit Cards	Male	Between 38 and 50	Middle	Technical	Married	Primary Education	Fraud	Hacking Social Media Accounts	Creating a Fake Shopping Site	UnKnown
1	1	Misuse of Debit Cards or Credit Cards	Female	[illegible]	Low	Student	Single	High School	Internet Shopping Out of Knowledge	Hacking Social Media Accounts	Social Engineering	UnKnown

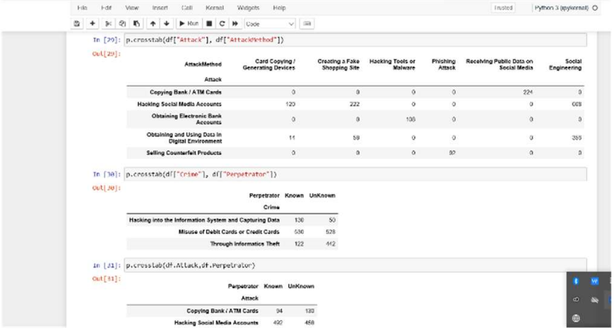
Step 4:It displays the list of attributes used.



```
       dtype=object)

In [16]: data[data["Harm"] == 'Fraud']
Out[16]:
```

	S.No	Crime	Gender	Age	Income	Job	Maritalstatus	Education	Harm	Attack	AttackMethod	Perpetrator
0	0	Misuse of Debit Cards or Credit Cards	Male	Between 38 and 50	Middle	Technical	Married	Primary Education	Fraud	Hacking Social Media Accounts	Creating a Fake Shopping Site	UnKnown
24	24	Misuse of Debit Cards or Credit Cards	Male	51 and Above	Middle	Retired	Married	High School	Fraud	Hacking Social Media Accounts	Creating a Fake Shopping Site	Known
51	51	Misuse of Debit Cards or Credit Cards	Male	Between 38 and 50	Low	Technical	Married	High School	Fraud	Hacking Social Media Accounts	Creating a Fake Shopping Site	UnKnown
57	57	Misuse of Debit Cards or Credit Cards	Male	51 and Above	Middle	Retired	Single	High School	Fraud	Hacking Social Media Accounts	Creating a Fake Shopping Site	Known
60	60	Misuse of Debit Cards or Credit Cards	Male	Between 38 and 50	Low	Technical	Married	Primary Education	Fraud	Hacking Social Media Accounts	Creating a Fake Shopping Site	Known
...	...	...	...	...	...	...	...	...	...	...	...	...
1291	1291	Misuse of Debit Cards or Credit Cards	Female	Between 38 and 50	Middle	Retired	Single	Primary Education	Fraud	Hacking Social Media Accounts	Creating a Fake Shopping Site	Known
1292	1292	Misuse of Debit Cards or Credit Cards	Male	Between 38 and 50	Middle	Technical	Married	Primary Education	Fraud	Hacking Social Media Accounts	Creating a Fake Shopping Site	Known

Step 5: Displays the total number of attack methods and perpetrators found.



```
In [29]: p.crosstab(df["Attack"], df["AttackMethod"])
Out[29]:
```

AttackMethod	Card Copying / Generating Devices	Creating a Fake Shopping Site	Hacking Tools or Malware	Phishing Attack	Receiving Public Data on Social Media	Social Engineering
Attack						
Copying Bank / ATM Cards	0	0	0	0	224	0
Hacking Social Media Accounts	120	222	0	0	0	608
Obtaining Electronic Bank Accounts	0	0	108	0	0	0
Obtaining and Using Data in Digital Environment	11	59	0	0	0	355
Selling Counterfeit Products	0	0	0	92	0	0

```
In [30]: p.crosstab(df["Crime"], df["Perpetrator"])
Out[30]:
```

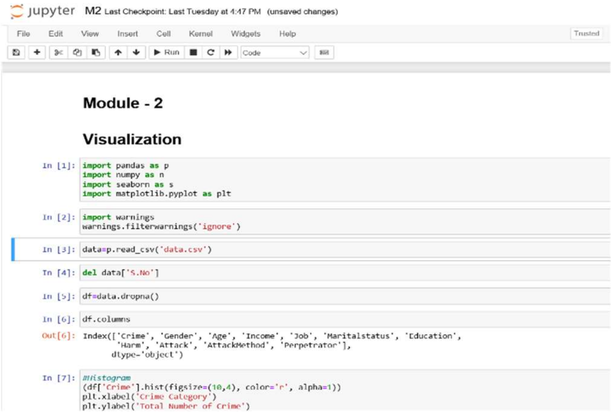
Perpetrator	Known	UnKnown
Crime		
Hacking into the Information System and Capturing Data	136	50
Misuse of Debit Cards or Credit Cards	530	528
Through Informatics Theft	122	442

```
In [31]: p.crosstab(df.Attack,df.Perpetrator)
Out[31]:
```

Perpetrator	Known	UnKnown
Attack		
Copying Bank / ATM Cards	94	130
Hacking Social Media Accounts	492	458

Step 6: In this, the attributes are named in numerical order. For eg, in crime 1) Hacking into the information security and capturing data. 2)Misuse of debit and credit cards. 3)Through informatics theft.

Step 7: Module 2 is about data visualization.

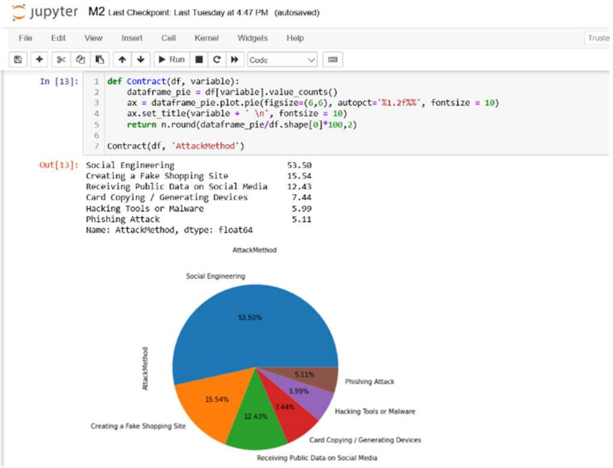


Module - 2

Visualization

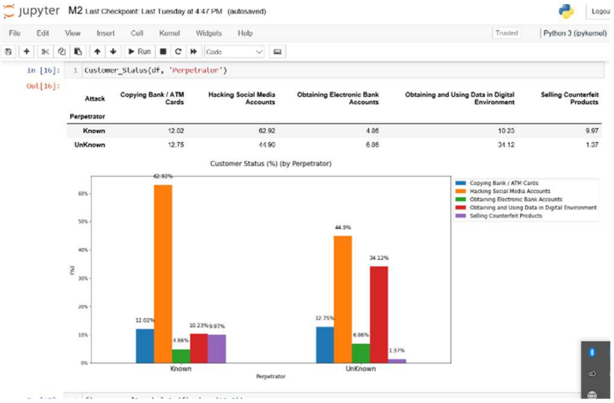
```
In [1]: import pandas as p
        import numpy as n
        import seaborn as s
        import matplotlib.pyplot as plt

In [2]: import warnings
        warnings.filterwarnings('ignore')

In [3]: data=p.read_csv('data.csv')

In [4]: del data['S.No']

In [5]: df=data.dropna()

In [6]: df.columns
Out[6]: Index(['Crime', 'Gender', 'Age', 'Income', 'Job', 'Maritalstatus', 'Education',
               'Harm', 'Attack', 'AttackMethod', 'Perpetrator'],
              dtype='object')

In [7]: #Histogram
        (df['Crime'].hist(figsize=(10,4), color='r', alpha=1))
        plt.xlabel('Crime Category')
        plt.ylabel('Total Number of Crime')
```

Step 8: Displaying the attack data in the form of a pie diagram


jupyter M2 Last Checkpoint: Last Tuesday at 4:47 PM (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Truste
Run
Code
AttackMethod
Social Engineering
53.50%
5.11%
5.99%
7.44%
12.43%
15.54%
Phishing Attack
Hacking Tools or Malware
Card Copying / Generating Devices
Receiving Public Data on Social Media
Creating a Fake Shopping Site
AttackMethod

```
In [13]:  def Contract(df, variable):
              dataframe_pie = df[variable].value_counts()
              ax = dataframe_pie.plot.pie(figsize=(6,6), autopct='%1.2f%%', fontsize = 10)
              ax.set_title(variable + ' \n', fontsize = 10)
              return n.round(dataframe_pie/df.shape[0]*100,2)

          Contract(df, 'AttackMethod')

Out[13]:  Social Engineering                       53.50
          Creating a Fake Shopping Site            15.54
          Receiving Public Data on Social Media    12.43
          Card Copying / Generating Devices         7.44
          Hacking Tools or Malware                  5.99
          Phishing Attack                           5.11
          Name: AttackMethod, dtype: float64
```

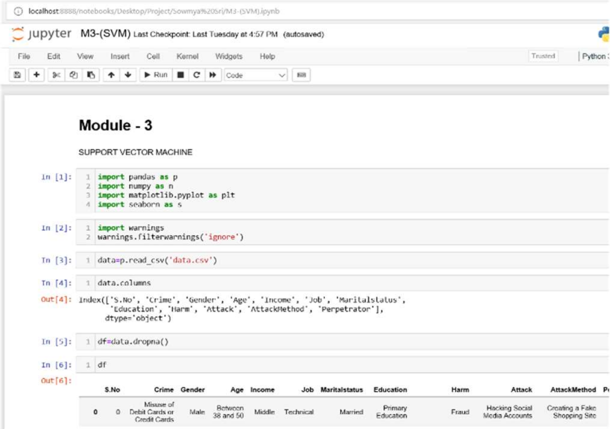
Step 9: Displaying the attack data in the form of a bar diagram.


jupyter M2 Last Checkpoint: Last Tuesday at 4:47 PM (autosaved)
Logou
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
Customer Status (%) (by Perpetrator)
62.92%
12.02%
4.86%
10.23%
9.97%
44.9%
34.12%
12.75%
6.86%
1.37%
60%
50%
40%
30%
20%
10%
0%
(%)
Known
UnKnown
Perpetrator
Copying Bank / ATM Cards
Hacking Social Media Accounts
Obtaining Electronic Bank Accounts
Obtaining and Using Data in Digital Environment
Selling Counterfeit Products

```
In [16]:  Customer_Status(df, 'Perpetrator')
Out[16]:
```

Attack	Copying Bank / ATM Cards	Hacking Social Media Accounts	Obtaining Electronic Bank Accounts	Obtaining and Using Data in Digital Environment	Selling Counterfeit Products
Perpetrator					
Known	12.02	62.92	4.86	10.23	9.97
UnKnown	12.75	44.90	6.86	34.12	1.37

Step 10: Module 3 consists of the display of the Support Vector Machine.


localhost:8888/notebooks/Desktop/Project/Sowmya%20Sri/M3-(SVM).ipynb
jupyter M3-(SVM) Last Checkpoint: Last Tuesday at 4:57 PM (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python :
Run
Code

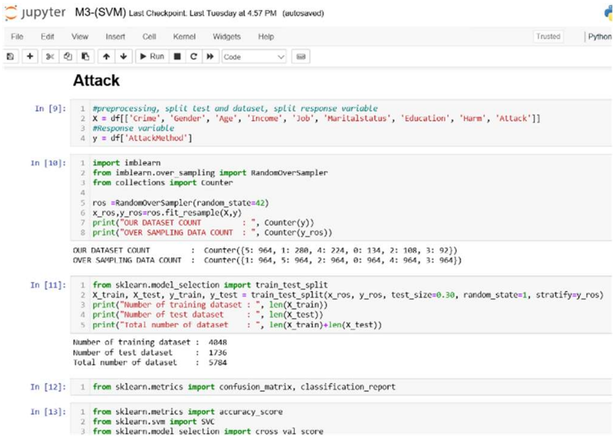
Module - 3

SUPPORT VECTOR MACHINE

```
In [1]:  import pandas as p
         import numpy as n
         import matplotlib.pyplot as plt
         import seaborn as s

In [2]:  import warnings
         warnings.filterwarnings('ignore')

In [3]:  data=p.read_csv('data.csv')

In [4]:  data.columns

Out[4]:  Index(['S.No', 'Crime', 'Gender', 'Age', 'Income', 'Job', 'Maritalstatus',
                'Education', 'Harm', 'Attack', 'AttackMethod', 'Perpetrator'],
               dtype='object')

In [5]:  df=data.dropna()

In [6]:  df
Out[6]:
```

	S.No	Crime	Gender	Age	Income	Job	Maritalstatus	Education	Harm	Attack	AttackMethod	P
0	0	Misuse of Debit Cards or Credit Cards	Male	Between 38 and 50	Middle	Technical	Married	Primary Education	Fraud	Hacking Social Media Accounts	Creating a Fake Shopping Site	

Step 11: Algorithm of the first phase attack method which has training dataset, test dataset, and a total number of the dataset.


jupyter M3-(SVM) Last Checkpoint: Last Tuesday at 4:57 PM (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python
Run
Code

Attack

```
In [9]:  #preprocessing, split test and dataset, split response variable
         X = df[['Crime', 'Gender', 'Age', 'Income', 'Job', 'Maritalstatus', 'Education', 'Harm', 'Attack']]
         #Response variable
         y = df['AttackMethod']
```

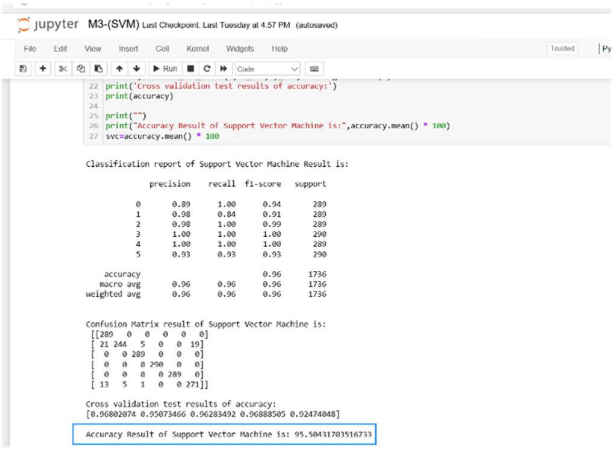
```
In [10]: import imblearn
         from imblearn.over_sampling import RandomOverSampler
         from collections import Counter

         ros =RandomOverSampler(random_state=42)
         x_ros,y_ros=ros.fit_resample(X,y)
         print("OUR DATASET COUNT          : ", Counter(y))
         print("OVER SAMPLING DATA COUNT  : ", Counter(y_ros))

OUR DATASET COUNT         :  Counter({5: 964, 1: 280, 4: 224, 0: 134, 2: 108, 3: 92})
OVER SAMPLING DATA COUNT  :  Counter({1: 964, 5: 964, 2: 964, 0: 964, 4: 964, 3: 964})
```

```
In [11]: from sklearn.model_selection import train_test_split
         X_train, X_test, y_train, y_test = train_test_split(x_ros, y_ros, test_size=0.30, random_state=1, stratify=y_ros)
         print("Number of training dataset : ", len(X_train))
         print("Number of test dataset     : ", len(X_test))
         print("Total number of dataset    : ", len(X_train)+len(X_test))

Number of training dataset :  4048
Number of test dataset     :  1736
Total number of dataset    :  5784
```

```
In [12]: from sklearn.metrics import confusion_matrix, classification_report
```

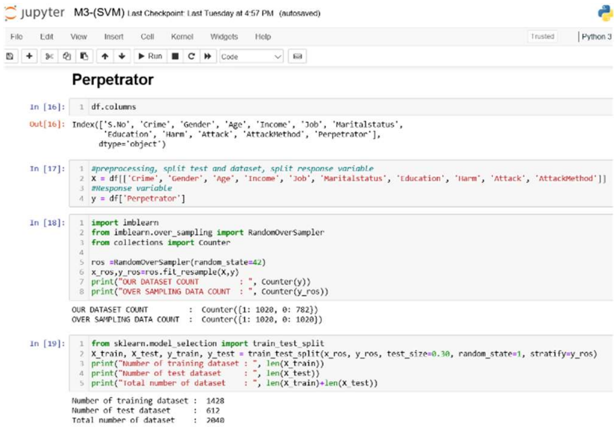
```
In [13]: from sklearn.metrics import accuracy_score
         from sklearn.svm import SVC
         from sklearn.model_selection import cross_val_score
```

Step 12: The accuracy result of the attack method using the support vector machine algorithm is calculated.


jupyter M3-(SVM) Last Checkpoint: Last Tuesday at 4:57 PM (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Trusted
Py
Run
Code

```
print('Cross validation test results of accuracy:')
print(accuracy)

print("")
print("Accuracy Result of Support Vector Machine is:",accuracy.mean() * 100)
svc=accuracy.mean() * 100

Classification report of Support Vector Machine Result is:

              precision    recall  f1-score   support

           0       0.89      1.00      0.94       289
           1       0.98      0.84      0.91       289
           2       0.98      1.00      0.99       289
           3       1.00      1.00      1.00       290
           4       1.00      1.00      1.00       289
           5       0.93      0.93      0.93       290

    accuracy                           0.96      1736
   macro avg       0.96      0.96      0.96      1736
weighted avg       0.96      0.96      0.96      1736


Confusion Matrix result of Support Vector Machine is:
 [[289   0   0   0   0   0]
 [ 21 244   5   0   0  19]
 [  0   0 289   0   0   0]
 [  0   0   0 290   0   0]
 [  0   0   0   0 289   0]
 [ 13   5   1   0   0 271]]

Cross validation test results of accuracy:
[0.96802074 0.95073466 0.96283492 0.96888505 0.92474048]

Accuracy Result of Support Vector Machine is: 95.50431703516733
```

Step 13: Algorithm of the first phase attack method which has training dataset, test dataset, and a total number of the dataset.


jupyter M3-(SVM) Last Checkpoint: Last Tuesday at 4:57 PM (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code

Perpetrator

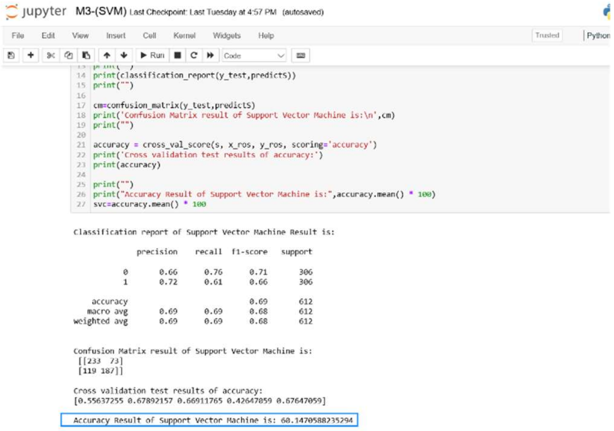
```
In [16]: df.columns

Out[16]: Index(['S.No', 'Crime', 'Gender', 'Age', 'Income', 'Job', 'Maritalstatus',
               'Education', 'Harm', 'Attack', 'AttackMethod', 'Perpetrator'],
              dtype='object')
```

```
In [17]: #preprocessing, split test and dataset, split response variable
         X = df[['Crime', 'Gender', 'Age', 'Income', 'Job', 'Maritalstatus', 'Education', 'Harm', 'Attack', 'AttackMethod']]
         #Response variable
         y = df['Perpetrator']
```

```
In [18]: import imblearn
         from imblearn.over_sampling import RandomOverSampler
         from collections import Counter

         ros =RandomOverSampler(random_state=42)
         x_ros,y_ros=ros.fit_resample(X,y)
         print("OUR DATASET COUNT          : ", Counter(y))
         print("OVER SAMPLING DATA COUNT  : ", Counter(y_ros))

OUR DATASET COUNT         :  Counter({1: 1020, 0: 782})
OVER SAMPLING DATA COUNT  :  Counter({1: 1020, 0: 1020})
```

```
In [19]: from sklearn.model_selection import train_test_split
         X_train, X_test, y_train, y_test = train_test_split(x_ros, y_ros, test_size=0.30, random_state=1, stratify=y_ros)
         print("Number of training dataset : ", len(X_train))
         print("Number of test dataset     : ", len(X_test))
         print("Total number of dataset    : ", len(X_train)+len(X_test))

Number of training dataset :  1428
Number of test dataset     :  612
Total number of dataset    :  2040
```

Step 14: The accuracy result of the perpetrator using the support vector machine algorithm is calculated.



```
print("")
print(classification_report(y_test,predictS))
print("")

cm=confusion_matrix(y_test,predictS)
print('Confusion Matrix result of Support Vector Machine is:\n',cm)
print("")

accuracy = cross_val_score(s, x_ros, y_ros, scoring='accuracy')
print('Cross validation test results of accuracy:')
print(accuracy)

print("")
print("Accuracy Result of Support Vector Machine is:",accuracy.mean() * 100)
svc=accuracy.mean() * 100
```

```
Classification report of Support Vector Machine Result is:

              precision    recall  f1-score   support

           0       0.66      0.76      0.71       306
           1       0.72      0.61      0.66       306

    accuracy                           0.69       612
   macro avg       0.69      0.69      0.68       612
weighted avg       0.69      0.69      0.68       612


Confusion Matrix result of Support Vector Machine is:
 [[233  73]
 [119 187]]

Cross validation test results of accuracy:
[0.55637255 0.67892157 0.66911765 0.42647059 0.67647059]

Accuracy Result of Support Vector Machine is: 60.1470588235294
```

8. CONCLUSION AND FUTURE WORKS:

This study proposes a technique for anticipating and identifying cyberattacks by combining ML algorithms with information from prior cybercrime investigations. The characteristics of those who could be assaulted and the types of attacks they would experience are predicted by the model. ML methods have been shown to be effective enough. The most effective of these techniques is the linear SVMs approach. The program has a 60 percent success rate in identifying the attacker who will launch a cyberattack. There may be ways to attempt to raise this ratio using further artificial intelligence techniques. According to our methodology, it is vital to call attention, in particular, to malware as well as social engineering assaults. It has been shown that the likelihood of a cyberattack decreased with the victim's wealth and education level. This study's main goal is to assist law enforcement organizations in the battle against cybercrime and to give quicker and more efficient ways to identify crime and offenders. As a result of our analytical study's examination of the characteristics of assault victims, new training & warning systems may be developed for persons with comparable traits. Deep learning algorithms may be used to anticipate crime, criminal activity, victim profiling, and cyberattacks in future works, and the outcomes can be compared. Cybercrime statistics from other provinces may also be gathered to utilize for comparison with this research based on conversations with other authorized entities with crime databases. Comparable studies may be used to compare the data from other provinces. To lower crime rates, “intelligent criminal-victim detection” systems that may aid law enforcement in their battle against crime & criminals could be developed.

Acknowledgment

I express my heartfelt thanks to our Head of the Department, Prof. Dr. S. Geetha, who has been actively involved and very influential from the start till the completion of my mini-project. Our sincere thanks to Project guide Dr.G.Soniya Priyatharsini for their continuous guidance and encouragement throughout this work, which has made the project a success.

REFERENCES

1)Arpitha. B, Sharan. R, Brunda. B, Indrakumar. D, Cyber Attack Detection and notifying system using ML Techniques, 2021

2)Arora T, Sharma M, Khatri SK. 2019. Detection of cybercrime on social media using random forest algorithm.

3)Bayuk JL, Healey J, Rohmeyer P, Sachs MH, Schmidt J, Weiss J. 2012. Cyber security policy guidebook. Hoboken: Wiley. 3-4

4) Ben-Asher N, Gonzalez C. 2015. Effects of cyber security knowledge on attack detection.Computers in Human Behavior 48(3):51-61

5)Bharathi ST, Indrani B, Prabakar MA. 2017. A supervised learning approach for criminal identification using similarity measures and K-Medoids clustering.

6) Bharati A, Sarvanaguru RAK. 2018. Crime prediction and analysis using machine learning. International Research Journal of Engineering and Technology 5(9):1037-1042

7)Biju JM, Gopal N, Prakash AJ. 2019. Cyber attacks and their different types.International Research Journal of Engineering and Technology 6(3):4849-4852

8)Biswas AA, Basak S. 2019. Forecasting the trends and patterns of crime in Bangladesh using a machine learning model.

9)Breda F, Barbosa H, Morais T. 2017. Social engineering and cyber security. InternationalTechnology, Education and Development Conference 3(3):106-108

10) Canbek G, Sagiroglu Ş, Temizel TT. 2018. New techniques in profiling big datasets for machine learning with a concise review of android mobile malware datasets.

11)Ch R, Gadekallu TR, Abidi MH, Al-Ahmari A. 2020. Computational system to classify cybercrime offenses using machine learning. Sustainability 12(10):4087

12)Chandrasekar A, Raj AS, Kumar P. 2015. Crime prediction and classification in San Francisco City

13) Dipankar Dasgupta. Immunity-based intrusion detection system: A general framework. In Proceedings of the 22nd National Information Systems Security Conference (NISSC)Arlington, Virginia, USA, 1999.

14)Jonatan Gomez and Dipankar Dasgupta. Evolving fuzzy classifiers for intrusion detection. In Proceedings of the 2002 IEEE Workshop on Information Assurance, West Point,2002