Improving Cyber Situational Awareness via Data mining and Predictive Analytic Techniques

Sina Pournouri

A dissertation submitted in partial fulfilment of the requirements of
Sheffield Hallam University for the **Degree of Doctor of Philosophy**

January 2019

Acknowledgment

Many people have earned my gratitude for supporting me in this journey. First, I would like to send my love to my beautiful family. My father, Mansour, you have been always a role model for me and as I always say you are my best friend. My mother, Farahnaz, I want to say you are the best. In this journey you listened to my moans always and you have been very supportive. My sister, Saba, in difficult times, you have been a great support. I hope one day you achieve what you deserve as you just started your path towards a great success.

Also I would like to express my gratitude to my Director of Studies, Professor Babak Akhgar. His support, encouragement, and enthusiasm have been inspirational. I would also like to thank my supervisor Doctor Shahrzad Zargari for her support and advice.

I would like to express my appreciation to Sheffield Hallam University and Cultural Communication and Computing Research Institute (C3RI) for support and the opportunity I was given to learn and develop my skills and follow my dream.

At last I want to say thank you to everyone that I cannot name them here, you all will stay in my heart forever.

Abstract

As cyber-attacks have become more common in everyday life, there is a need for maintaining and improving cyber security standards in any business or industry. Cyber Situational Awareness (CSA) is a broad strategy which can be adopted by any business or government to tackle cyber-attacks and incidents. CSA is based on current and past incidents, elements and actors in any system. Managers and decision makers need to monitor their systems constantly to understand ongoing events and changes which it can lead to predict future incidents. Prediction of future cyber incidents then can guide cyber managers to be prepared against future cyber threats and breaches.

This research aims to improve cyber situational awareness by developing a framework based on data mining techniques specifically classification methods known as predictive approaches and Open Source Intelligence (OSINT). OSINT is another important element in this research because not only it is accessible publicly but also it is cost effective and research friendly.

This research highlights the importance of understanding past and current CSA, which it can lead to more preparation against future cyber threats, and cyber security experts can use the developed framework with other different methods and provide a comprehensive strategy to improve cyber security and safety.

List of Publications

Pournouri, S., Zargari, S. and Akhgar, B., 2018. Predicting the Cyber Attackers; A Comparison of Different Classification Techniques. In *Cyber Criminology* (pp. 169-181). Springer, Cham.

Pournouri, S., Akhgar, B. and Bayerl, P.S., 2017, January. Cyber attacks analysis using decision tree technique for improving cyber situational awareness. In *International Conference on Global Security, Safety, and Sustainability* (pp. 155-172). Springer, Cham.

Pournouri, S. and Akhgar, B., 2015, September. Improving Cyber Situational Awareness Through Data Mining and Predictive Analytic Techniques. In *International Conference on Global Security, Safety, and Sustainability* (pp. 21-34). Springer, Cham.

Table of Contents

LIST OF FIGURES

List of Tables

Chapter 1 Introduction

Since Computers have made life easier, they can be also vulnerable targets for cyber criminals. Cyber-attacks have become a common threat to all different aspect of daily life and they can cause a different level of disruption regarding their target, Type of Threat etc. For instance, Sony Pictures Company was targeted by unknown hackers and some of their products including movies, contracts and market plans were leaked thus analysts predict money loss of the company was around 83 million dollars (Savov, 2014).

Cyber-attacks have a huge negative implication of their targets and the damages caused by them are not always technological. Depends on the victims, the effects can be varied from money loss to reputation damage. For instance, the impact of cyber-attacks to a bank not only can have a negative financial effect but also can damage the reputation of the bank and then that will lead to decreasing the number customers and partners. A cyber-attack to a critical infrastructure of a country can also lead to a major damage like losing confidential and secret information to adversaries.

Existing and new cyber threats to businesses and critical infrastructure have made cyber security experts to plan and implement efficient strategies for prevention and protection to mitigate the risk of cyber-attacks. Managers and authorities are always attempting to find an efficient way to be prepared and secured against future and current cyber-attacks. One of the common methods is applying security standards and policies including cyber security awareness programs. "How to improve cyber security awareness" is a significant challenge for security experts, therefore, they always try to understand the current and past trends of cyber security world (Pournouri and Craven, 2014). Understanding trends of cyber security can be divided into two different levels as follows:

1. Detection of weaknesses and bugs in the system: This step can be taken by security specialists by examining systems using penetration tests in order to find security bugs and w1`eaknesses. By detection of weaknesses in the system, security managers can implement and design effective and solid security standards and procedures. In addition, in order to fill technical bugs and gaps, security patches and equipment will be installed.

2. Identifying cyber hackers and their methods: This level builds on the previous stage and it aids security managers to be aware of cyber-attacks recent methods. The concept of cyber-attack analysis will be highlighted in this stage meaning that by analysing past historical cyber-attacks to cyber firms and finding a relationship between different involved factors, a better landscape will be obtained and let managers to make effective decisions based on recent cyber threats (Pournouri and Craven, 2014).

Cyber Situational Awareness (CSA), as an area of concern of this study, is a term that cyber security experts are using to define a broad strategy to tackle cyber incidents. Cyber Situational Awareness is based on understanding current and past situation within the cyber space and prediction and preparation against future incidents. There are many kinds of research about improving cyber situational awareness and they are explained in the literature review section; however, this study seeks to investigate using data mining techniques and Open Source Intelligence to improve cyber situational awareness through prediction of future cyber-attacks and their different features.

Data mining has been used increasingly in order to meet demands and operate efficiently in many fields and especially those fields which their performance is based on prediction of future as well as interpretation of past and current circumstance. In addition, cyber security can benefit from data mining as an operational way to find a better way to deal with cyber-attacks and enable decision makers to comprehend current and past situational awareness (Ahn et al. 2014).

Another main concept which will be one of the main parts of this study is Open Source Intelligence (OSINT). OSINT refers to those

information and intelligence which is accessible publicly and they are being used by different businesses and even intelligence services. The source of data of this research is OSINT and the full explanation of OSINT has been given in the literature review section.

This study aims to investigate possible ways of CSA improvement through data mining techniques and predictive analytics. In order to fulfill the aim of this research, various steps should be taken including collection of cyber-attacks, pre-processing, analysing and interpretation of the analysed result. The result of this project supports decision makers in cyber firms to take decision more effectively based on experiences and prediction of the future.

1.1 Research Rationale

In this section, the research methodology and philosophy, which has been adopted in this project, will be explained. In the first part, the research question will be described and then based on that in the second part, the aim and objectives will be defined. The third part will discuss the research methodology adopted in this study.

1.1.1 Research Question

Cyber experts adopt different approaches and solutions to tackle cyber incidents and prediction is commonly used by them as Barford et al. (2010) reported the understanding current situation and learning from past incidents are two main principles of cyber countermeasures. Collected data from past and current conditions of cyber space can be subject to knowledge discovery process, which is based on prediction, clustering and discovering a correlation between cyber incidents and their different factors (Skillicorn, 2009). Therefore, the following research question can be concluded:

To what extent can a predictive framework based on classification algorithms and open source Intelligence (OSINT) contribute to improving and a better understanding of Cyber Situational Awareness (CSA)?

1.1.2 Aim and Objectives

This project aims to design a predictive framework using predictive based on classification techniques and past historical data in terms of cyber-attacks to tackle future cyber threats and previous unsolved cyber-attacks incidents. This will contribute to a deeper understanding and improving cyber situational awareness.

The following objectives are identified in order to develop a predictive framework and achieve the aim of this study:

1. Collect data from Opens Source Intelligence. This data includes past cyber attacks to different sectors happened from 2014 to 2017
2. Create a dataset in the form of spreadsheet by cleaning the raw data collected OSINT
3. Train the predictive models based on appropriate classification techniques
4. Evaluate predictive models based on different measurements

Based on aim and objectives of this PhD, the process can be broke down into 5 different steps as they are shown in figure 1-1

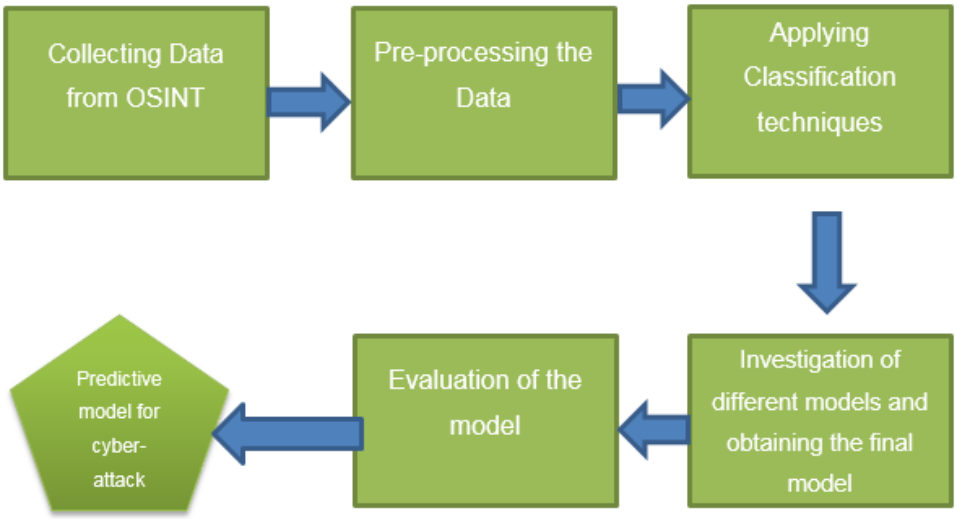

Figure 1-1 The process of the study

1.1.3 Research Methodology

Research methodologies are divided into two main categories; inductive and deductive. Inductive approach tries to reach the new theory based on findings, however, the deductive approach uses the proposed theories in order to new findings. (McNeil, 1990).

This project seeks to use an Inductive approach which means based on examples and findings, a new theory will be obtained. In other words, based on the collected past cyber-attacks incidents, a predictive framework will be obtained which comes in form of a general theory.

Research methods are divided into 4 main categories (Kothari, 2004):

1. Action research: This method is based on pursuing an action and comprehension of the current condition. This method is based on a cycle where the method will be refined based on the interpretation of condition.. Action research is often described as a challenging approach due to the fact that the researchers adopting this method can face uncertainty and unexpected challenges. Action research method has been used widely in educational projects. Another platform which action research is applied to is operational management because it involves with

different unexpected conditions and situations (Coughlan and Coghlan, 2002).

2. Case Study: This method is based on in depth investigation of a single situation and it will lead to generating large amount of subjective. Case study method not only tries to report on the data obtained from a single situation but also attempts to extend the findings and conclude a general interpretation for other situations.
3. Survey: This method is based on questionnaires or interviews. Two main important elements that should be taken into consideration in this method are designing the questionnaire or interview and the sample size and sample variables.
4. Experiment: This research method seeks to investigate casual relationships in a controlled environment, which is set by the researcher. This method usually is adopted in problem solving, development and evaluation projects.

Based on the discussed research methods, this study is based experiment design method. The desired research methodology is experimental methodology but because of limitation of this study, this method will be modified in order to be more feasible within the time scale of this project.

In order to better understanding of the proposed experimental research method, following principals of this PhD are considered:

1. Type of Data: Type of data in any project is the main principle (Walliman, 2011) and this stage is mainly concerned with the form of data to whether it is qualitative or quantitative. The initial format of the data is qualitative and for this research, there is no need to convert them to quantitative as inputs. The outputs of analysis will be quantitative and need to interpret to qualitative format.

2. Data collection method: in this stage, the data will be collected in the form of data sets which are available online and publicly.

Open Source Intelligence (OSINT) is the main source of required datasets due to the fact that they are accessible publicly. The concept of OSINT will be defined in chapter 3 extensively and details of this step will be explained in chapter 4.

3. Authentication and Credibility of the Data: This stage is mainly concerned with the reputation and authentication of the organization supplying the data set. As it was mentioned in the previous stage, this research project will benefit from OSINT for data collection purpose. OSINT includes newspaper, online websites, and companies' announcements and so on. Although there are different organization and companies recording cyber-attacks, they pass their data to other organizations or research centers at different prices.

4. Data analysis: This stage of the research will benefit from data mining techniques and specifically classification techniques which have been used by researchers for building up the predictive framework and more details will be explained and defined in chapter 3. According to Walliman (2011), data mining is a method to find important information from huge databases and gives substantial help to decision makers in order to fulfill their demands and achieve their goals.

5. Result and its interpretation: This study is a combination of qualitative and quantitative research supporting decision makers in order to increase cyber security awareness within their systems. After applying classification algorithms to the dataset, their result will be compared against each other and the desired model will be concluded in a form of five dimensional predictive framework for prediction of future cyber attacks' features including cyber attackers, Type of Threat, and Type of Target, cyber-attack activity, and targeted country.

6. Validation of results: At this stage, the framework will be tested and evaluated through applying an unseen and different data to it and the accuracy will be measured and the success rate of the framework will be discussed.

This study will design a method to address the research question based on the following steps which is inspired by experiment research methodology:

1. Data collection: this research will focus on cyber-attack historical data and the data will be collected from Open Source Intelligence (OSINT). There were contacts with some companies collecting cyber intelligence, however, they do not tend to share their intelligence with students and individuals. Several contacts were made with some companies gathering cyber intelligence and cyber activities including the Recorded Future (www.recordedfuture.com), Norse(www.Norse-corp.com) and Cyber Intelligence Center (Cyber Intelligence Center) but there was no reply from them. The goal of these companies is passing their intelligence to businesses and organizations based on financial purposes and they do not tend to share their intelligence for research purposes. Therefore, it has been decided to use other sources such as news and websites and any other sources which are publicly accessible and their information does not raise any ethical issues. There are some websites recording cyber-attack such as http://hackmageddon.com/ and their data can be used as a raw data set. Cyber incidents which happened from 2013 to 2017 will be collected as the most recent cyber incidents and the data collection process will be explained in section 3.3.

2. Data structure and pre-processing: Operational errors can happen due to the implementation of the system because obtained data from the real world has some errors, contradictions, incompatibility, and missing values. This stage is based on Al-Janabi (2011) method which consists of various tasks to make the data ready for analysis purposes. These tasks include dealing with missing values, removing noises, fixing

incompatibilities and removing outliers. Then the data should be categorized in order to make it more meaningful and operating. Past historical cyber-attacks will be categorized based on the date of the incident, type of attack, Type of Target, hacker, type of cyber breach etc. Further details about this step are described in section 4.4.

3. Data mining techniques: Based on the literature review, in order to find patterns among the data, data mining techniques will be utilized. According to Ahn et al. (2014) using classification techniques can help cyber experts to find current patterns and based on findings try to predict the future patterns. Also based on the type of data which is categorical and discrete, classification techniques will be adopted in this project. According to Kaur et al. (2015) decision trees, Support Vector Machine, Naïve Bayes, K Nearest Neighbour and Multi-Layer Perceptron which is a form Artificial Neural Network can fit the purpose of prediction when the dataset is categorical. There are few examples of applications of classification techniques which have been explained in section 3.6. Al-Janabi (2011) reports that decision trees are easy to use, interpret and make them meaningful to decision makers for predictive analytic and he used it for crime predictions. Bhardwaj and Pal (2011) suggest that naïve Bayes can be used when attributes within the data set are independent and they apply that algorithm to the educational dataset to predict students' future performance. For this stage, R programming as a powerful language for data analysis will be used. R language was initially developed in 1999 and it has many packages which they are highly suitable for analysing and visualizing the data.

4. Expected result and obtaining the final model: after applying classification techniques, the results of each classification technique will be compared with each other and then the most efficient and most accurate model will be concluded.

5. Interpretation of the obtained model: This stage is a crucial stage, as Morris et al. (2011) and Schreiber and Koch (2012) report that making the result meaningful to managers and decision makers is the most significant stage of CSA improvement. For instance, it should be determined which attributes or elements have more effect and cyber-attacks or which factors are weakest or strongest in the CSA.

6. Model evaluation or validation: After the interpretation of the final model, then unseen data will be fed into it to evaluate how accurate it can be for predicting future attributes of a cyber-attack.

1.2 Thesis Outline

This thesis consists of seven different chapters as follows:

1. The first chapter aims to explain a brief introduction to the issue of cyber security, protection, and prevention solutions. It also chapter aims to highlight the research methodology and give a clear view about the research question, aim and objectives of this study.

2. The second chapter named literature review includes 5 sub chapters; Data mining explaining main techniques and how they apply, Open Source Intelligence describing the definition, advantages and disadvantages of using OSINT, Cyber security demonstrating main concepts of cyber security, Cyber Situational Awareness explaining the definition and the main theory and previews and existing approaches to improve CSA and the last subchapter discussing about the application of predictive analytic in other areas.

3. The third chapter aims to discuss the data collection, structure and preprocessing.

4. Chapter four aims to explain data analysis and applying classification techniques in order to train the predictive models.

5. Chapter five will compare the predictive models based on their accuracy rate in order to select the best one in terms of different measurements of accuracy

6. Chapter six will discuss the importance of each factor in the predictive model and also evaluate the result of this research by applying unseen data to the nominated predictive model.

7. Chapter seven summarizes this study and highlight the conclusion and the contribution to knowledge. The limitation of this research and also the future work that can be done will be explained as well.

Chapter 2 Literature review

2.1 Introduction

The principal of this research is the combination of four main elements; Cyber security, data mining, Open Source Intelligence, and Cyber Situational Awareness. We have tried to illustrate how data mining and statistical techniques support cyber security in terms of the improvement of cyber situational awareness. In essence, the principal of this PhDis governed by the notional representation of Cyber Security, Cyber Situational Awareness, Open Source Intelligence, and Data Mining.

2.2 Data mining

As was discussed briefly in Chapter 1, data mining techniques can be applied to cyber security elements in order to have more efficient decisions and actions for Improving CSA. According to Ledoltar (2013), since customers and data have become a strategic goal, data mining has been used in order to improve the performance in serving customers' needs and decision makers' task. The technique of analysing, extracting and discovering meaningful information and pattern, is defined as data mining. Managers are trying to understand the current situation and predict the future trends within their business are also using data mining methods. (Ahlemeyer-Stubbe and Coleman, 2014)

Data mining techniques are divided into two main approaches as follows (Odei Danso, 2006):

1. Supervised: This approach of data mining concludes a pattern of a function from labelled training data.
2. Unsupervised: In this approach, the data mining technique tries to extract hidden patterns from unseen or unlabelled data.

In another categorization, data mining techniques are fallen into the following general methods (Pournouri and Akhgar, 2015 p23):

1. “Regression analysis: This technique tries to establish a function leading to model the data.
2. Association rules: Refers to a technique used to discover interesting relationships among different variables in a data set.
3. Classification: classification techniques are mainly used to classify data set into a different subgroup and the result can be interpreted as a predictive model.
4. Clustering: This technique tries to arrange similar objects in a specific group based on their similarity factors.”

2.2.1 Classification

In this section, classification techniques will be explained in more details as they are being used in this project. Classification techniques include five main knowledge discovery methods; Decision Tree, Naïve Bayes, Support Vector Machine, K Nearest Neighbour and Artificial Neural Networks which will be described in this chapter later. (Han et al., 2011)

2.2.1.1 Decision Tree

Decision trees are also known prediction trees and include a sequence of decisions and their outcomes as consequences. If inputs are a set of variables $x_1, x_2, x_3, \dots, x_i$, the goal is to predict a set of result including $y_1, y_2, y_3, \dots, y_i$. The prediction process can be done through making a decision tree with nodes and their branches. Each node represents a specific input variable and each branch means a decision making process. Leaf node refers to those nodes that they do not have branch and return class labels and in some occasion, they generate probability scores. Decision trees are being used in most data mining application with predictive purposes because they are easy to implement, visualize and present. Input variables can be categorical and continuous.

Decision trees are divided into two types; classification and regression. When output variables are categorical, the decision tree is called classification tree and when the outcome is continuous such as numbers, they are called as regression trees.

Decision tree algorithms are mainly divided into two forms; ID3 and C4.5. ID3 or Iterative Dichotomiser 3 was developed by John Ross Quinlan (1990) and the algorithm is as follows:

ID3 (I,O ,T) # T is training set, I is input variables, O is Output variables

If T ϵ Φ

Return Φ

If all records in T have the same value for O

Return a single node with that O value

If I ϵ Φ

Return a single node with the most frequent values of O in T

Compute information gain for each attribute in I relative to T

Pick attribute D with the largest gain

Let $\{d_1, d_2, d_3, \ldots, d_i\}$ *be the values of attribute D*

Partition T into $\{T_1, T_2, T_3, \ldots, T_i\}$ *according to the values of D*

Return a tree with root D and branches labelled $d_1, d_2, d_3, \ldots, d_i$

Going respectively to trees ID3(A-{D}, O, T_1*) ,*

ID3(A-{D}, O, T_2*),…, ID3(A-{D}, O,* T_i*)*

C4.5 improves some weaknesses in ID3 such as dealing with missing values which makes data analysts able to handle and analyse data more efficiently.

Decision trees always get the best available option to split by using greedy algorithm, however, that option might be the most desirable option at that stage and not in the overall process. Therefore, if a bad split is chosen, it will permeate through the rest of the tree.

There are different ways to investigate whether the obtained decision trees are desirable. These methods are as follows:

1. Checking whether the splits make sense or not by validating them with domain experts by conducting a survey or an interview.

2. Investigating of nodes and depth of the tree. The existence of too many layers and nodes can be an indication of an overfit model. In this case, the model fits training set well but it underperforms on test set. To address the issue of overfitting in decision trees, stopping growth of the tree before it gets to the stage when all training set is classified well or using post prune option to reduce errors, can be considered.

According to Freund and Mason (1999) C4.5, Recursive partitioning and Random forest are 3 main decision tree algorithms:

1. C4.5 is a decision tree algorithm which was developed by Ros Quinlan (1993) and it is an extension of ID3 algorithm based on information entropy. One of the implementation method of C4.5 is using RWeka package includes J48 function to train classifier by C4.5. R Weka Package was developed by Hornik et al. (2009) and includes different way of implementation of classification algorithms.

2. Recursive Partitioning is a method of decision trees based on greedy algorithms to classify members of population correctly based on independent variables. It mainly suits categorical variables and does not perform well on continues variables (Friedman, 1976). One of the way of implementation of RP in R is using Rpart package which was developed by Therneau et al. (2010). They provide implementation of RP by a function called rpart and visualize the decision trees which can be interpreted easier and more comprehensible.

3. Random Forest is another important decision tree algorithm developed by Breiman and Cutler (2007). This algorithm builds decision tree by a process called bagging which means combination of learning trees to increase classification accuracy and Random forest package exists among R packages which was developed by the same authors.

2.2.1.2 Naïve Bayes

Naïve Bayes is one of the most important data mining methods which leads to classification and prediction. One of the most common applications of Naïve Bayes is spam filtering and text categorization. (Murphy, 2006)

Naïve Bayes assumes that all variables in the dataset are independent from each other and calculate the probability of different events based on the variables, however, some authors such as Ray (2015) consider this as a disadvantage because it is almost impossible to obtain data set that its features are completely irrelevant in real life. One of the main advantages of this algorithm is that it only needs a small number of training set in order to build a classifier and according to Murphy (2006), Naïve Bayes is much more efficient, accurate and faster compared to some other algorithms such as decision trees in some classification problems.

Naïve Bayes is based on Bayes theorem and produces conditional probability models, the following formula describes Bayes theorem (Vapnik and Vapnik, 1998):

$$P(c|x) = \frac{P(c)P(x|c)}{P(x)}$$

Equation 2-1 Naïve Bayes theorem

In this formula, P(c|x) is subsequent probability of Class C when attribute x happens, P(c) is prior probability of class, P(x|c) is the chance which is the probability of predictor given class and P(x) is the prior probability of predictor.

Naïve Bayes classifier is among the well-known and most efficient classifiers because not only they are easy to install but also in this approach, there is no need of prior knowledge about the data. This method is being used in different application such as spam filtering and fraud detection (Murphy, 2006).

Naïve Bayes algorithm's advantages and disadvantages can be categorized as follows (Leung, 2007):

1. Naïve Bayes is very simple and easy to implement compared to other classification algorithms.
2. It does not need high level of prior knowledge for training purposes so it can produce reliable models even with small training set.
3. In some cases, because of basic assumption of Naïve Bayes which considers all variable independent the accuracy will get lower.

2.2.1.3 Support Vector Machine

Support Vector Machine (SVM) is a supervised classification technique that is being used for both classification and regression problems. SVM has been used in different application such as text classification and image processing due to high level accuracy compared to other algorithms. Initially, Cortes and Vapnik (1995) developed SVM algorithm and it is based on choosing the optimal hyperplane which separates two or more classes with the maximum distance called Margin between their closest. Support vectors refer to those objects located on the boundaries. Support vectors are the most challenging objects to classify and they play a significant role in defining and identification of the optimal hyperplane. Figure 2-1 Shows support vector machine mechanism in a simple form.

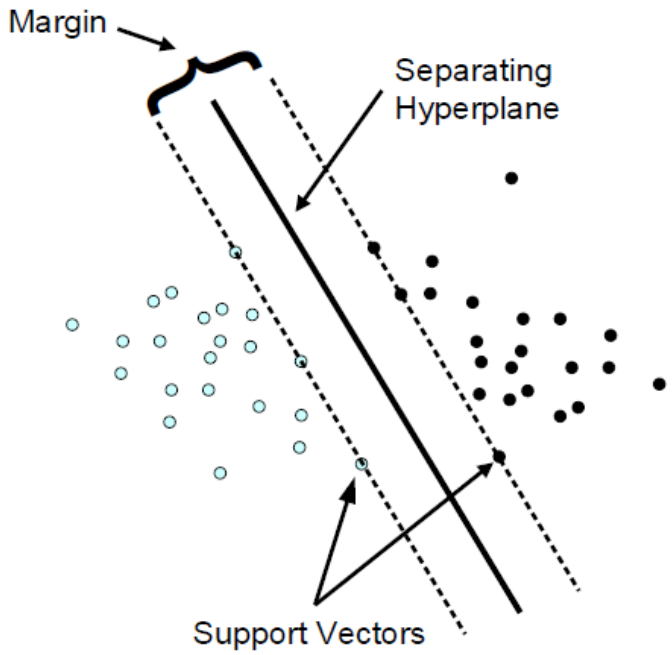

Figure 2-1 Support Vector Machine Hyperplane

SVM has the following advantages (Tong and Koller, 2001):

1. It performs significant analysis in high dimensional spaces.
2. It is memory efficient because it will do analysis on part of training set which is support vectors.
3. It is very useful when number of objects is less than number attributes or dimensions.

On the other hand, SVM cannot deal with large dataset quickly and it is very time consuming. In addition, SVM is highly noise sensitive algorithm meaning that noise in the training set can have negative impact on the final model. (Tong and Koller, 2001)

2.2.1.4 K Nearest Neighbour

According to Larose (2005), K Nearest Neighbour (KNN) is an algorithm used in both classification and regression problems. As Input, the algorithm gets K nearest training example and produces the output as class label. The object will be classified based on majority vote of its neighbour. In the KNN algorithm, the training set is the set of vectors in

multidimensional space assigned to different classes. Then in classification stage, user will define K and the object will be assigned to relevant classed based on the maximum frequency of that class. As an example Figure 2-2 explains how KNN works, two classes are presented as green and blue and the red star is the object needs to be classified. If K is equal to 3, 3 nearest neighbour of the red star will be chosen. Then the next step is counting how many votes each class will be given to red star and in this case, the blue class has the majority and the red star will be classified as blue. The second case scenario is when the K gets 5, 5 nearest neighbour of the red star will be defined and in this case, the red star will be classified as green because the majority of the neighbours are green objects.

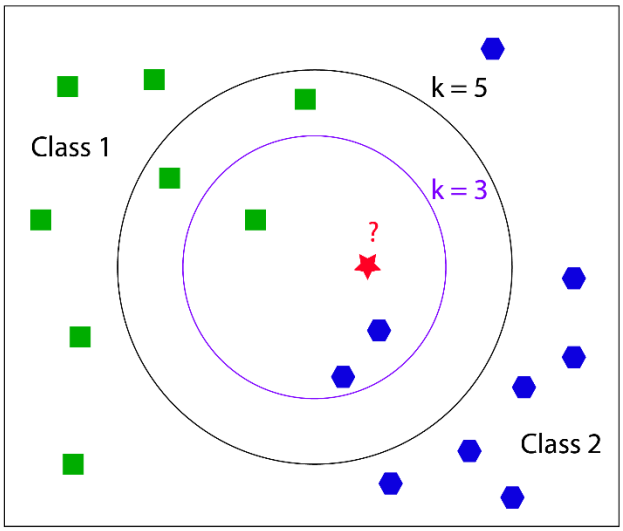

Figure 2-2 KNN example

Like other classification algorithms, KNN has some advantages and disadvantages and they are as below (Beyer et al., 1999):

1. KNN algorithm is very simple and easy to implement.
2. It has significant performance on multiclass problems and cases.
3. It handles the noisy datasets very well.
4. The performance of KNN can be more powerful and effective if the training set is large enough.

5. KNN is known as lazy learner because it does not learn from training set and simply classify the test set based on training set.

6. Operationally KNN is expensive because of computation of K parameter and K computation is based on measuring the distance between the objects.

2.2.1.5 Artificial Neural Networks

Artificial Neural Network (ANN) is another classification algorithm which is designed based on neural system of the brain. In the initial stage, ANN classifies objects then it will compare them with actual classes and the measure the error rate. Then it will repeat the process until the error reaches to its minimum value. Figure 2-3 shows the structure of the ANN algorithm (Agatonovic-Kustrin and Beresford, 2000).

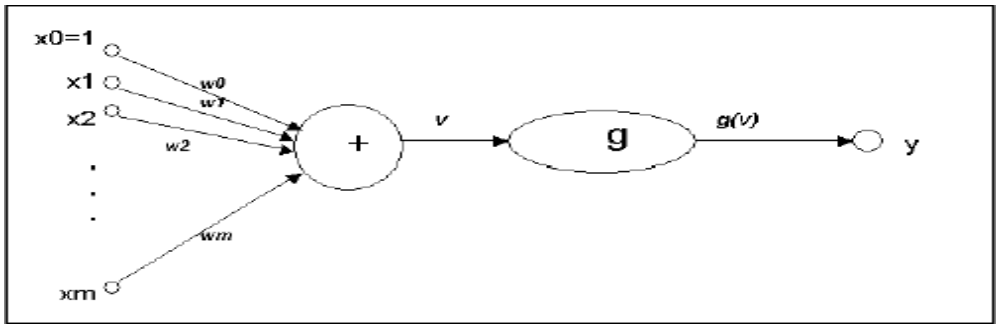

Figure 2-3 structure of Artificial Neural Network

The structure as it is shown in the figure includes Xi and related wright Wi as inputs and g as function which will process the input and produces the result or output as Y.

An ANN algorithm consists of 3 main layers; the first layer is input layer getting inputs, the second layer named hidden layer which can have one or more sublayers doing the process task of the algorithm and the last layer is the output layer produces the result (Agatonovic-Kustrin and Beresford, 2000).

According to Singh (2011), ANN is a supervised algorithm so the classes for each record are defined before the process and the output nodes need to be given to the correct classes. Therefore 1 for the

correct node and 0 for other nodes will be considered. This assumption gives the possibility of calculation of success in terms of correct classification. Based on the error rate, the iterative learning process will be carried out until the error rate gets the minimum value. Measuring error rate in each time of learning process will define the modification of weights associated with inputs.

Table 2-1 demonstrates the main advantages and disadvantages of ANN (Singh, 2011).

Advantages	**Disadvantages**
Requires less statistical techniques	Black box performance
Significant performance on non-linear data	Requires large training data set for better models
Strong and robust detection of the relationship between predictors	Can produce overfitting

Table 2-1 Advantages and Disadvantages of ANN

2.2.2 Clustering

Clustering is an unsupervised learning referring to grouping set of entities based on their similarities. Clustering is a bottom-up approach which its goal is that entities in one group are similar to each other and different from other groups' object. Clustering methods are being used in many different applications such as pattern recognition, image processing and medical imaging and so on. Clustering methods are divided into 6 main categories as follows (Berkhin, 2006):

1. Partitioning method: This method applies when the dataset has n objects and method will build k partitions of the data. Each partition includes one object at least and each object must be allocated to a group or partition. Partitioning method benefits from iterative technique for allocation of objects to the groups by moving them from one to other groups in order to improve the partitioning.

2. Density-based method: This approach is based on density which means the given cluster will keep growing until it exceeds some threshold for each data point. A minimum number of data points are required in order to determine the radius of a given cluster.

3. Model-based method: This method is based on a hypothesis describing a model for each cluster and finding the optimum data to be allocated to the given model. By clustering function which presents spatial distribution of the data points, the cluster will be determined. This method is considered as an effective clustering method because it deals with outliers and noise of the data and statistically allocates the number of clusters.

4. Constraint-based method: This method refers to the clustering based on user or applications' preference and expectation of the clustering process. This method includes must-link constraint defining that two entities with a must link relation should belong to the same cluster and a cannot-link constraint specifies that two entities should not belong to the same cluster if they have cannot-link relation. The constrained based model is a semi-supervised approach and some constrained methods will stop processing if they cannot meet the requirements and some others try violating less constraint to find an efficient clustering result.

5. Hierarchical method: This method is divided into 2 main methods; Agglomerative and Divisive. Agglomerative is a bottom-up approach which begins with grouping each object and then groups and objects will get merged until it reaches its defined conditions of clustering result. On the other hand, the divisive method starts the process with

the objects in the same cluster and with iterative technique, it splits up the clusters into smaller ones until it reaches to the defined expectations of the clustering process.

6. Grid-based method: Grid-based method is similar to the density-based method because they are both space driven approaches rather than data-driven therefore in terms of processing speed they are both very fast algorithms. This method makes partitions within the embedded space by make the space objects into a finite number of cells and the clustering process takes place in the created cells.

2.2.3 Regression

Regression is a data mining approach which is used for prediction of numeric and continuous values. This approach is very similar to classification, where both are used for prediction but classification predicts categorical or nominal values. Regression algorithms are applied to different sections such as financial forecasting, trend analysis etc. Regression algorithms are categorized into 5 types as follows (Hand, 2007):

1. Linear regression: it is the most basic regression which mainly used for prediction of relationship between 2 variables. By using best fit straight line, linear regression tries to specify the relationship between dependent and independent variables. The basic form of linear regression is formed into; Y=a + b*x+ e, which x is independent variable, Y is dependent variable, *a* is intercept, *b* is slop of the line and *e* refers to error. There are two types of linear regression; Simple linear regression where there are one independent variable and multiple linear regression there is more than one independent variable. The main disadvantage of linear regression is its poor performance when the data has outliers and noise.

2. Logistic regression: This type of regression is used in classification problems and when the dependant variable is in binary format. Unlike linear regression, logistic does not need linear relationship between variables because it has non-linear log transformation. Logistic regression can perform better with large sample dataset and make better and accurate result.

3. Polynomial regression: It refers to the type of regression when the independent variable has the power more than one. The equation of this regression is:

$$Y= a + bx^2.$$

Equation 2-2 Polynominal regression

4. Ridge Regression: This type of regression is used when independents variables are highly correlated.

5. Stepwise regression: This type of regression is applied when there are multiple independent variables. Independent variables are chosen automatically without any human interference. This selection will be done based on statistical values such as R-square and t-test. Stepwise regression operates very well when data has many dimensions. The task of stepwise regression includes:

    a. In each step, predictors can be added or eliminated.

    b. In each step, variables will be added and the most significant predictors will be selected in forward selection.

    c. Backward elimination includes removing irrelevant variables and all predictors will be selected.

2.2.4 Association rule

Association rule is a data mining technique, which tries to find a relationship in form of rule between objects in a large dataset. The aim of association rule is find a strong rule to find patterns in a large dataset. This technique has 4 main concepts as follows (Mining, 2006):

1. Support: it refers to indication of how frequently the variable appears in the data set and it is formulated as: , t: transaction T: proportion of transaction.

$$supp(x) = \frac{|\{t \in T; xCt\}|}{|T|}$$

Equation 2-3 Support formula

2. Confidence: it refers to number of time when a rule has been found and its equation is

$$Conf(x \to y) = \frac{supp(x \cup y)}{supp(x)}$$

Equation 2-4 Confidence formula

3. Lift : it is formulated as

$$lift(x \to y) = \frac{supp(x \cup y)}{supp(x) * supp(y)}$$

Equation 2-5 Lift equation

4. Conviction: it refers to following equation:

$$Conv(x \to y) = \frac{1 - supp(y)}{1 - conf(x \to y)}$$

Equation 2-6 Conviction formula

Association rule has 3 main algorithms as follows (Mining, 2006):

1. Aprior: this approach benefits from breadth-first search method to measure the support of itemsets and it feat the descending closure property of support by employing candidate generation function.

2. Eclat (Equivalence Class Transformation): this approach is based on a depth-first search algorithm and it can be executed parallel or sequential.

3. FP-growth (Frequent pattern algorithm): This algorithm initially counts the appurtenance of items and save them into the header table then it inserts the instances and makes the fp-tree structure and the last step is the elimination of least frequent items.

2.3 Open Source Intelligence

Open Source Intelligence (OSINT) is obtained from publicly available sources such as newspapers, journal, radio and TV (Stalder and Hirsh, 2002). OSINT not only is being used by businesses but also intelligent services and governments along with their classified information are utilizing it in order to support their decision and strategy making process. According to (Steele, 2007) OSINT gives the ability to have access to different sources of information and its combination with classified intelligence produces a structured information for operative purposes. There are four separated blocks of open and public information (Steele, 2007):

1. Open Source Data (OSD): Open source data is raw materials such as photographs, broadcasts etc. coming from a primary source.

2. Open Source Information (OSIF): OSIF is processed of data which can be extracted by applying some filters and validation stages. OSIF includes newspapers, books etc.

3. Open Source Intelligence (OSINT): OSINT is that type of information which has been discovered, processed and targeted a specific group of audience. OSINT is the result of applying process of Intelligence creation to wide range of information for addressing specific issues and questions.

4. Validated OSINT (OSINT-V): It refers to a type of OSINT, which its accuracy reaches very high level. OSINT-V either can be obtained by combining classified information with OSINT can result from guaranteed source such as satellite images.

OSINT has advantages and disadvantages, which can be explained as below (Stalder, 2002):

1. The main advantage of OSINT is its accessibility and sharing feature, the Information because of its nature can be easily accessible and shared with anybody or organization without any legal and ethical problems.

2. Cost-effectiveness of OSINT is a significant feature makes researchers and organizations with a low and tight budget, able to do their research and investigation for free or very low price.

3. The main disadvantage of OSINT is, it sometimes needs a heavy analytical process to make it ready to use. Verification, validation and filtering noise and inaccuracy from OSINT need high amount of analytical work and consume big window of time.

2.4 Cyber Security

The usage of computers in all businesses not only makes their job more efficient but also changes the Type of Threats that they face. Cyber-attacks are a new and dynamic phenomenon threatening

companies and governments and cyber security experts need to learn from the past cyber-attacks and design a comprehensive strategy preparing companies and governments against current threats and new threats in future. (Das et al. 2013)

Sony as an entertainment company in May 2011 (Aspan and Soh, 2011), Citi Bank as a financial and banking service in June 2011 (Levik, 2011) and FBI as a law enforcement agency in FBI and Department of Justice in January 2012, Sony in May 2011 (Aspan and Soh, 2011) and Citi Bank in June 2011 (Levik, 2011) were victims of cyber-attacks and that shows that any organizations and businesses regardless of their size can be a potential target for cyber attackers. (Das et al. 2013)

2.4.1 Cyber threats

Different researchers have various views about categorization of cyber-attacks; however, they mainly define cyber-attacks as follows: (Ker et al. 2010) (Nikshin, 2004):

1. Virus: a piece of code which can penetrate the computer system and infect different parts of the system.
2. Worm: a malicious piece of code which causes an interruption in a process of a computer.
3. Trojan horse: a piece of malicious program, which initially shows it, is harmless to the system but then it will cause disruption in the system.
4. Logic bomb: a piece of code, which is designed for destruction and damaging, a computer system at a certain date and time.
5. Key logger: a piece of program, which can collect and save key strokes in a computer system. Key loggers are being used by both cyber criminals to steal sensitive information and companies to monitor their employees.
6. DOS: It refers to a type of attack targeting a server in victims' side by sending too many constant request, which leads to take down the server.

7. SQL injection: a piece of malicious code programmed to target the database of a victim and steal stored information.
8. Zero-day threat: It refers to an undiscovered security bug in an application or a computer system which has not been identified and addressed by security experts and taken advantage by cyber criminals to target the victim through that.
9. Phishing: it refers to deceiving the victim by a fake email pretending to be legitimate and steal sensitive information.

Cyber-attacks have different and sophisticated methods behind every cyber-attack can be a thought and motivation behind it. In next section Type of cyber attackers and their motivations will be discussed.

2.4.2 Cyber attackers and their motivations

Computers also make significant changes to crime, motivations, and thoughts behind them. In order to deeper understanding and efficient analysis of cyber-attacks leading to improvement of CSA, there is a need to categorize cyber criminals and their motivations and goals.

According to Awan and Blakemore (2012), cyber attackers are divided into the following groups:

1. “White hat: This group also is defined as cyber security experts where they are hired by organizations and businesses to test their security standards and demands. The task of white hackers is identifying current and potential weaknesses of the computer and network systems through various approaches such as black box test or white box test. After identification of those weaknesses, they present the efficient solution in order to address them.

2. Black hat: This group refers to those hackers who use their abilities to attack systems and obtain unauthorized and sensitive information. Although Black hat hackers’ motivations do not

always focus on stealing information, sabotage and damaging to the systems are other motivations (Jaishankar, 2011). Cyber terrorists can be a subgroup of black hats whereas their motivations are illegitimate. According to Lewis (2002) those cyber attackers targeting critical infrastructure such as power, government operations in order to make public fear, are defined as cyber terrorists. There are some disagreements and dissimilarities between sociologists' definition of cyber terrorism term. Some authors like Aviksoo (2008) and Pollit (1998) define cyber terrorists as type of cyber attackers who follow political and social interests and carry cyber-attacks to achieve them. On the other hand, some authors such as Cox (2015) the term of cyber terrorism makes sense when human casualties are the main risk because of cyber-attacks. However, all of them agree on the type weapon for this act which is a computer and Type of Target which is critical infrastructures.

3. Grey Hat: grey hat hackers can be categorized in both previous groups. In other words, they are judged based on the result of their performance whether is peaceful and leading to improvement of security standards or harmful and leading to security breaches"(Pournouri and Akhgar, 2015, p24).

2.4.3 Cyber Activities

Cyber activities are divided into 4 main levels which will be explained as follows (Saini et al., 2012):

1. Cyber Espionage: It refers to an attempt to gain access to confidential and secret information held by organizations or governments without their knowledge and cyber attackers use sophisticated techniques such as malware to spy on victims. For instance, in 2011 Mitsubishi was targeted by cyber espionage attack claimed that their data on missiles and submarines were spying for unknown amount of time through a malware detected in their computer systems. (Nitta,2013)

2. Cyber Warfare: It refers to a state-sponsored attack using computer devices and techniques to target other states or governments' organization and properties, however, in some cases the cyber-attack is not carried out by a government directly or they do not take credit of that attack. Most of these type of cyber-attacks target critical infrastructure such as Stuxnet in 2010 which was designed to target Iranian nuclear facilities using a malware to penetrate to SCADA systems performing in nuclear sites. Although there is no evidence that which country or group was preparing Stuxnet, US and Israel were blamed for the cyber-attack (Kelley, 2013).

3. Hacktivism: These type of cyber-attacks are motivated by social or political goals and hackers (Hacktivists) use attacking techniques to penetrate into victims' system and send their political or social message either to public or victims themselves. One of the notable example of hacktivism was the cyber-attack carried out by Anonymous against child pornography hidden websites in October 2014. Anonymous group managed to take down 40 child pornography websites which were hidden from search engines. (Goode, 2015)

4. Cyber Crime: it refers to those type of crimes when a computer device is being used as a tool or as a target. Cybercrime is more general term using for description of cyber-attacks on a smaller scale. (Saini et al., 2012)

2.5 Cyber Situational Awareness

According to Dua and Du (2011) in order to fill the present gaps in cyber security and deal with recent threats, an effective collaboration between cyber specialists and agencies is needed. These days cyber security researchers intend to design a solid and efficient framework maintaining confidentiality (the effort of keeping information secret between eligible and authorized parties and protecting it from unauthorized parties), Integrity (the ability of compatibility and accuracy

of information) and Availability (Accessibility to cyber infrastructure and information) to protect computer and network systems (Dua and Du,2011). Cyber situational awareness is one of the frameworks designed by cyber security experts in order to preserve cyber security's interest and prevent any form of security breaches.

The definition of Situational Awareness should be taken into consideration. Situational Awareness is often described as an understating different factors in an environment which leads to predict and precept the near future events and trends for decision makers (Antonik, 2007). Akhgar (2015) defines **Situational Awareness** in security and policing context; *"as a capability to identify (e.g. people, places, locations, details of an incident etc.), contextualise, visualise, process, and comprehend the critical elements of intelligence about particular area of concern. Area of concern can be anything from an investigation to management of major cyber-attack".* For our research, we are adapting Akhgar's definition of SA and will contextualise it for Cyber Situational Awareness.

In other definitions from other authors such as Tada and Salerno (2010) and Harrison et al (2012) the factor of time plays crucial roles, in other words the time in situational awareness is coming with past information and learning from failures in order to analyse and extract any possible relations among them for a deeper and clearer understanding of future condition and situation. Situational Awareness (SA) is mainly divided to 2 different aspects as follows:

1. Cognitive aspect: from cognitive point of view, SA is mainly concerned with human perception. Endsley (1995) suggests that SA comes down to three main criteria: Basic perception of important data, Interpretation, and conversion of the data to knowledge and capability of using found knowledge for prediction of near future.
2. Technical aspect: In terms of technical according to Bryrielsson (2006) and Arnborg et al. (2000) SA is a combination of three

main factors; arrange, analyse and integrate information as Arnborg et al. (2000) concentrates on arrange meaning collecting the data which suits the main demands and Bryrielsson (2006) reports that analyse and integration are two significant criteria in SA.

Bryielson and Frank (2014) and also Weick et al.(2005) suggest that Cyber Situational Awareness (CSA) can be a sub group of SA where the environment is cyber space and also in order to CSA, Data from IT equipment will be gathered and converted to suitable format for processing stage and that leads to better decision making. According to the study conducted by Weick et al. (2005) cyber sensors play prominent role gathering data for CSA improvement purposes in a deeper and detailed condition such as logs and data recorded by Intrusion Detection Systems.

Barford et al. (2010) categorizes existing methodologies for improving CSA into two main categories:

1. Low level: Low level of improvement of CSA includes factors that are more technical rather than other factors including human factors. Vulnerabilities assessment, damage assessment and alert correlation are significant factors in the low level. For example, security experts can correlate alert extracted from IDS and vulnerabilities of their system in order to better understanding of the current situation and predicting future issues occurring in the network.
2. High level: High level of CSA is more general than low level, in other words, it is the combination of human elements and technical factors. Human elements include human resources and human interference.

Based on the definition of SA by Akhgar and taxonomical classification suggested by of Tada and Salerno (2010) and Harrison et al (2012), we seek to enhance the CSA conceptual understanding, in order to develop a comprehensive framework of understanding for cyber experts to manage their security issues by using data mining and predictive analytic techniques.

2.5.1 Practical examples of Cyber Situational Awareness

This section aims to give some hypothetical examples and scenarios based on some sectors and firms cyber security strategies in order to make this study approach and aim more understandable as it is mentioned in aim and objectives in chapter 2, this research aims to design a framework using data mining analytic techniques to contribute to understanding of CSA based on past cyber incidents:

1. Financial sectors: Financial sectors such as banks and stock change market are always hit by different cyber-attacks. In order to mitigate the risk of cyber breaches in financial organization cyber situational awareness can be considered as general strategy. Over the past few years, cyber-attacks have become more complicated and sophisticated and damages caused by them can have significant negative influence on business and economy in larger scale. 2007 FDIC Technology Incident reports that there is an increasing trend in cyber threat against financial organizations: unknown unauthorized access, online bill applications, and spear phishing are the top three threats to financial sectors. Security reports such as FIDC report highlight security issues in financial sectors, however, security professionals can address these issues by having a program for improving CSA as a wider strategy for mitigation of cyber threats. By getting more detail about these threats not only they can increase the level of security inside their institution, but also they can feed their clients and customers with cyber security instruction. CSA in financial sector can improve knowledge of managers about ongoing security and help them to prioritize the security needs by identification of needs, vulnerabilities, and weaknesses. IT experts can use regular security reports including merging threats against financial sectors and by analysing them conclude a wider and more extensive CSA program to increase level of security in both institution side and client side. For instance in bank side by allocation of suitable and effective IT resources they can prevent or decrease the damage of cyber-attacks and in client side, also they can improve the knowledge of customers about cyber threats such as spear phishing and different malware in order to decrease the risk and the probability of cyber-attacks.

Health care Industry: health care industries also need CSA program as a security strategy. For instance, their database can contain sensitive and classified information about their customers and patients and that makes the more responsible and accountable to adopt effective CSA improvement program. The Health Information Trust Alliance (2014) announces that their analysis demonstrates the fact that although some healthcare sectors utilize threat intelligence to understand the ongoing situation in terms of cyber, they are not aware of countermeasures and solution to merging threats. CSA programs and tools can come together and help healthcare industries to improve the level of security and protect their patients against any sensitive information leakage. An extensive improving program applied to CSA should have following competences:

a) Identification of vulnerabilities and weaknesses within the computer systems in health care institution.
b) Monitoring merging cyber threats
c) Classification of merging cyber threats based on their impact on the health care industry.
d) Updating security countermeasures regularly in order to be prepared for new and unknown cyber attacks
e) Holding educational program for staff in the health care industry about cyber security issues.

2. Critical Infrastructure (CI): critical infrastructures are mainly important and sensitive to governments and nations. A cyber-attack to CI can damage heavily and lead to unprecedented loss to victims. US army highlights CSA in CI, therefore we conclude that CSA program in CI should have following capabilities (Gould,2015):

a) Monitoring cyber activities: this process should be done continuously and intelligence should be gathered in terms of cyber activities to address new issues timely.

b) Identification of vulnerabilities: this task should be continuously carried out by the CSA program in order to resolve and patch security bugs.

c) Profiling cyber adversary behaviour: an extensive CSA should use past intelligence about past cyber-attacks in order to profile cyber attackers for adopting effective security countermeasures against known enemies such as state sponsored cyber attackers.

2.5.2 Existing approaches

As it mentioned previously there are two main approaches to improve Cyber situational awareness; High level and Low level. In this section we reviewed both high- and low-level existing methods.

2.5.2.1 High Level methods

This section discusses existing methods for improvement of CSA from high level point of view and high level methods focus on roles, tasks, actors and human resources in cyber space.

One of the high level methods for improvement of CSA was proposed by Morris et al. (2011) based on intelligence gathering and tackle cyber threats through a mission-based hierarchy. By understanding the current situation in this approach, a clear view will be given to decision makers to plan a strategy to improve CSA. Figure 2-4 shows their proposed method and they present their method through 8 levels as follows:

1. Mission tools: based on type of cyber-attack happened in the system, tools and general solution to overcome the situation will be identified.

2. Mission need: based on the output of the previous level, needs and requirements of the mission to tack the cyber-attacks will be identified.

3. Mission Question: various questions will be asked in order to discover the weaknesses and vulnerabilities of the system.

4. Mission area: this stage generates outcome based on the mission questions. The outcome is monitoring, counter intelligence and indication and warnings.

5. Mission activity: the task and activity against cyber-attacks will be defined in more details at this stage.

6. Mission capability: the ability to do the defined tasks from the previous stage will be an area of concern of mission capability stage.

7. Mission resources: the type of resources which make a backbone for the mission capabilities will be discussed at this stage.

8. Mission sources: the raw and unprocessed data coming from different sources is the area of concern at this basic stage.

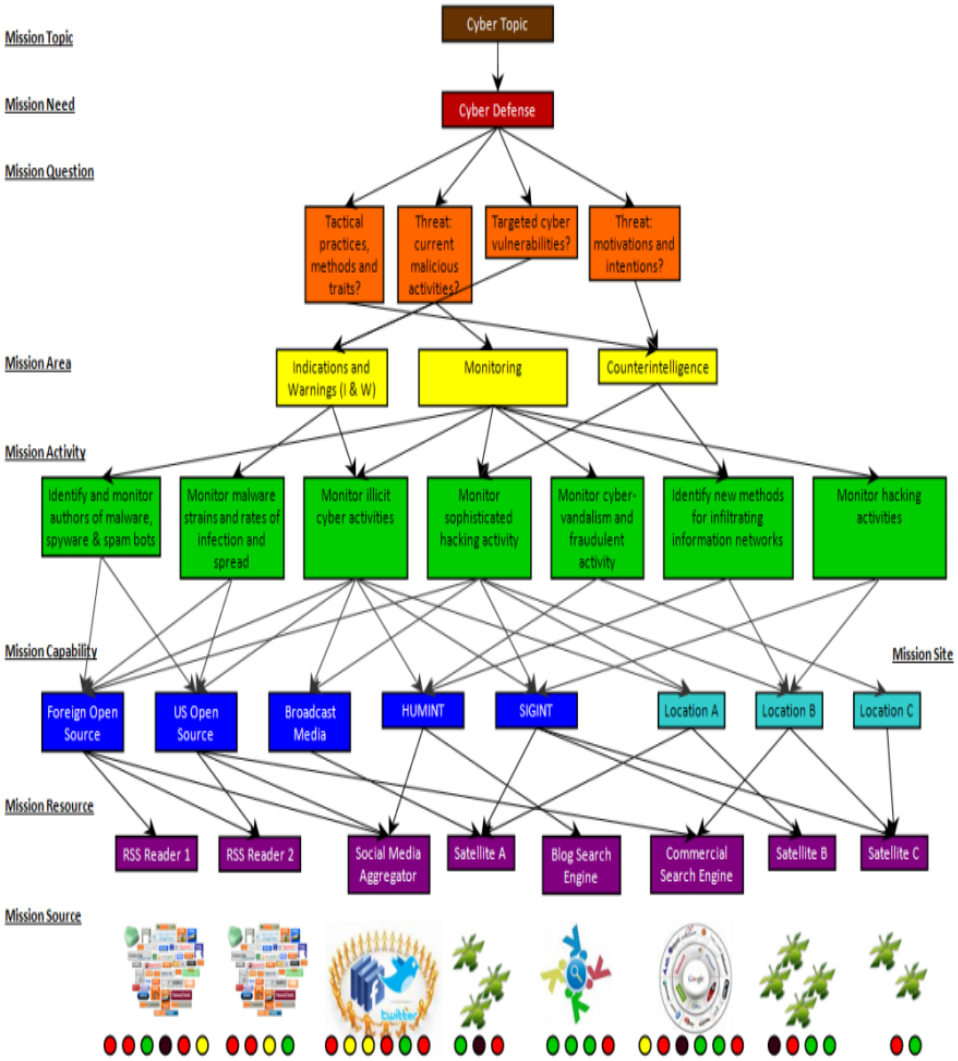

Figure 2-4 Mission based approach for improving CSA (Morris et al., 2011)

Morris et al. (2011)'s framework is mission-based approach in other words regarding current needs, missions and different levels will be defined. Although their method can be effective in decision making, their framework is designed for collecting information and gathering intelligence rather than applying analytical algorithms to obtained intelligence. The positive point about Morris et al. (2011)'s approach is it covers both high and low levels models. For instance, it includes monitoring cyber space in order to learn about malicious behavior from cyber hackers which is in the high-level category and also includes gathering information from IDS in order to find out about new attack signatures which is in the low-level category of CSA.

Another method has been presented by Huang et al. (2016) which includes the usage of Fuzzy based system to help cyber analysts to share their preferences and reach a general strategy to tackle cyber threats based on CSA. Huang et al. (2016) emphasize teamwork to improve CSA in a way that each cyber analyst detects an ongoing attack by finding abnormal network traffic or through scanners and IDS. Then the proposed fuzzy system will be activated and accepted following inputs:

1. For an ongoing attack, different solutions and their alternatives and success rate in form of probabilities will be suggested by cyber analysts.

2. Selection criteria of solutions by cyber analysts will be accepted as inputs in the fuzzy based system.

3. Each cyber analyst will be judged in terms of decision-making weight by their skills, ability, and experience and a weight will be assigned to each of them.
4. Based on the above parameters a belief matrix will be constructed.

5. By calculating, all the solutions and their parameters will be evaluated and a general approach will be generated which will be agreed by the cyber analysts.

Huang et al. (2016) proposed the fuzzy-based approach to improve CSA through a teamwork and generating a comprehensive strategy to tackle cyber-attacks. This method covers high level of CSA which emphasizes on roles and tasks.

Instant Based Learning Theory (IBLT) was used in Dut et al. (2013)'s approach. IBLT is an algorithm for prediction of human interactions and it contains of a storage including three blocks:

1. Situation: the ability to describe an attack.
2. Decision: an action or a strategy against an attack.
3. Utility: the measurement of the desired result of strategy against an attack.

Dut et al. (2013) suggest that their IBLT based approach takes defender behaviour, adversary behaviour and tolerance level into consideration and by analysing these actors' behaviour and also the level of tolerance of human against a cyber-attack, a broad and comprehensive plan in form of a cognitive model will be generated in order to increase defenders ability to deal with cyber attacks. The weakness of their framework is the validation of result and they are not able to compare it with real world scenario due to a different aspect of human perception. This model also can be developed through combination with low level approach which will give an extensive and more efficient approach for improving CSA. According to Akhgar (2015) actors in the cyber can be a factor in intelligence gathering for improving CSA, however, Dut et al. (2015) suggest that their proposed method only focuses on the human level of CSA based on the response against cyber attackers.

I-Hope (IT control, Human Resource control, Organization Control, P&E control, External Control) framework is another approach presented by Das et al the framework aims to investigate effective factors in the deterrence of cyber breaches by increasing CSA. Their initial data was obtained by the CSI/FBI questionnaire. CSI/FBI questionnaire includes data collected from security firms in the United States in anonymously which is conducted from those firms from 1997 to 2010. Das et al. (2013) purpose following stages in their methodology section:

1. Data pre-processing: After obtaining their initial data from the CSI/FBI survey, they combine it with ISO/IEC 2007. ISO/IEC is a standard for Information security incident management includes 5 categories: (I) Organizational Controls, (ii) External Controls, (iii) Human Resources Controls, (IV) Physical and Environmental controls and (v) Information Technology controls. Das et al. (2013) determined 4 questions from the process of combination of CSI/FBI survey and ISO/IEC 2007. Table 2-2 shows those questions.

	Year → / Questions ↓	1999	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
ORGANIZATIONAL CONTROLS	**Security Policy**														
	A. IS policy												√	√	√
	B. Role of Executive and management priorities													√	√
	C. Importance of contracts with business partners													√	√
	Organization of IS														
	IT Expenditure Decision														
	D.% of IT budget spent on awareness training											√	√	√	√
	E.% of IT budget spent on security								√	√	√	√	√	√	√
	F. Investment in security operating expenses										√				
	G.% of IT budget on regulatory compliance													√	√
	H.% of IT budget spent on forensics services														√
	Make or Buy Decisions														
	I.% of Security Function Outsourced								√	√	√	√	√	√	√
	J. Use of Cloud computing														√
	Monitoring Decisions														
	H. Techniques to evaluate effectiveness of IS										√	√	√	√	√
	I. Techniques to evaluate effectiveness of AT											√	√	√	√
	J. Security audits								√	√					
	External Information														
	K. Importance of news reports of other attacks													√	√
	Asset Management														
	L. Data retention/destruction policy												√	√	
	Information Security Incident Management														
	M. Action taken after an attack	√	√	√	√	√	√	√	√	√	√	√	√	√	√
	N. Importance of security breach notifications													√	√
	O. Analyzing points and sources of attack	√	√	√	√	√	√	√							
	Business Continuity Management														
	P. Data retention/destruction policy												√	√	
	Q. Do you hire reformed hackers as consultants	√	√	√	√	√	√	√							
	R. Importance of previous attacks on your firm													√	√
	S. External insurance policy								√	√	√	√		√	
EC	**Compliance**														
	T. Impact of regulatory compliance on overall IS													√	√
	U. Laws applying to organization (HIPAA, SOX)													√	√
	V. Importance of Industry standards(ISO, NIST)													√	√
	W. Importance of sector specific													√	√
	X. Impact of SOX								√	√		√			
	Y. Impact of SOX on security program's focus to											√			
HRC	**HR Security**														
	Z. Satisfaction with security technology													√	√
	AA. Why not report attacks to law enforcement	√	√	√	√	√	√	√	√	√	√	√	√	√	√
	AB. Importance of security awareness training								√	√	√	√			
PC	**P&E Security**														
	AC. Unauthorized use of computers	√	√	√	√	√	√	√	√	√	√	√	√	√	√
ITC	**Access Control**														
	AD. Unauthorized use of computers	√	√	√	√	√	√	√	√	√	√	√	√	√	√
	ISADM														
	AE. Security technologies used	√	√	√	√	√	√	√	√	√	√	√	√	√	√
	AF. Is Software Development process secure?												√	√	√
	Communications and operations management														
	AE. Security technologies used	√	√	√	√	√	√	√	√	√	√	√	√	√	√
	AH. Importance of vulnerability notification													√	√

AT = Awareness Training; EC = External Controls; HRC = HR Controls; PC = P&E Controls; ITC = IT Controls; HIPAA = Health Insurance Portability and Accountability Act , SOX = Sarbanes-Oxley Act,

Table 2-2 Mapping of CSI/FBI questions with ISO/IEC 20071 (Das et al., 2013)

2. Hypothesis formulation: They formulate their hypothesis into 4 formulates as follows:

    a) Increase in usage of technological defense will reduce the probability of a security breach.

    b) Reporting a cyber-attack to law enforcement will prevent and reduce the chance of further cyber security breaches.

    c) Spending more budgets on IT security will reduce the possibility of a cyber-attack.

    d) Increase in percentage of IT security outsourcing will lead to less probability of cyber-attacks.

3. Data analysis: Generalized Linear Model is used as a mathematical model to predict uncertainty due to the fact that Das et al. (2013) assume all of the mentioned cyber-attacks follow a binominal distribution. Below is the GLM equation where n_i is the total number of respondents for year i, y_i is totathe l number of respondent who suffered from the attack happening in year i, p_i is the probability of the attack happening in year i.

$$p_i(y_i) = {}^{n_i}C_{y_i}\, p_i^{y_i}\, (1-p_i)^{n_i-y_i}$$

Equation 2-7 GLM equation (Das et al.,2013)

4. Result: by analysing the model, the hypothesis questions will be answered as follows:

    a) By installing a specific type of security equipment, some type of attacks can be prevented.

    b) Reporting the first cyber-attack to Law Enforcement Agency does not stop further attacks.

    c) By installing ITO and increasing IT budget, the chance of a security breach will be decreased

Managers can benefit from this study because it shows how they should allocate budget and resources in their company, however, this study has one main weakness which is using CSI/FBI report as it is

sparse and not extensive. In addition, the calculation and equations can be time consuming and they do not suit a real time algorithm for improving CSA. The area of concern in i-Hope framework concentrates on the high level and can lead managers to discover the flexibility of their organizations in case of cyber incidents.

2.5.2.2 Low Level Methods

Ahn et al. (2014) report that improvement of CSA is possible with benefiting from Machine Learning and Artificial Intelligence. They suggest usage of classification for understanding the current and past CSA, regression to find any possible pattern between incidents and findings and Association rules for finding relationship between collected data and detecting any abnormal behaviour. As it is shown in figure 2-5 Ahn et al. (2014) suggest a framework with different stages. The first step deals with collecting the raw information from network devices and Intrusion Detection Systems and in the second step is categorizing the data and place them in appropriate groups and appliances for processing. The processing happens in the thirds stage which classification, regression and association techniques will be applied to the data and in the fourth step, the result of the analysis will be interpreted and presented to managers and decision makers.

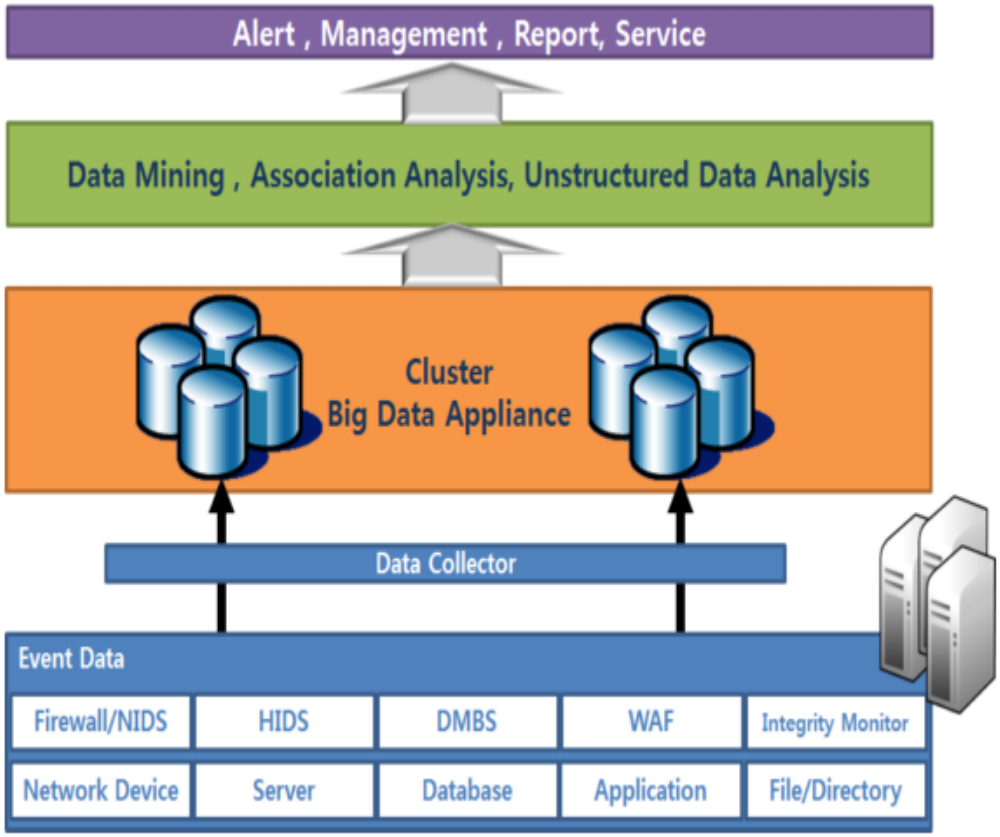

Figure 2-5 Big Data analysis system architecture (Ahn et al., 2014)

Although Ahn et al. (2013) do not purpose a real time algorithm, they introduce a significant framework for obtaining valuable information raw data collected from network sensors and devices in order to improve CSA. According to Akhgar (2015), the area of concern is important in CSA and because Ahn et al. (2015) focus on technical issues within the computer systems, the area of their concern categorised as the low level concern. In addition, the project will benefit from classification techniques mentioned in Ahn et al. (2013)'s proposed model and it will be explained in detail in methodology part.

The limitations and weaknesses of Ahn et al. (2014) are concluded as follows:

1. There is no evaluation for the performance of their framework and they did not mention how the obtained result can be projected as a suitable output for managers.
2. Their framework is not real time response to the problems and it does not give the ability to decision makers to confront with cyber issues in appropriate time.
3. This approach can be improved and developed by analyzing users' behavior and it can be combined with high level CSA in order to get a better outcome.

Wu et al. (2013) suggest the usage of the Bayesian network to improve CSA through prediction and prevention of cyber-attacks. This approach focuses on environmental and internal criteria and they are as follows:

1. Identify vulnerabilities and weaknesses: This step will be done by using powerful scanners such as Nessus.
2. The usage situation of the network: this criterion will focus on traffic load and number of requests and services which each node should deal with.
3. The value of asset in the network: This criterion is concerned with the type of data and type of task that each node has.
4. Attack history: based on previous cyber-attacks, this criterion will specify the likelihood of further attacks on different nodes.

Wu et al. (2013) correlate all of the above criteria and then apply Bayesian network to them. Although these criteria mentioned by Wu et al. (2013) will not be used in this research, the Bayesian network might be used because it is a significant technique to analyse uncertain knowledge. Not only, real time response is an issue for Wu et al. (2013) proposed framework, but also, they did not propose any approach for making obtained result meaningful for decision makers and there is lack interpretation of results and transfer the knowledge to decision makers in their approach. Based on Akhgar (2015) suggestion about CSA, Wu et al. (2015) try to address the area concern of the

computer networks and their assets and they propose an investigative method for identifying vulnerabilities in a system based on their value and the attack history.

Feasel and Ramos (2013) proposed a defence model for organizations against ongoing or plausible cyber threats. OSINT is the main source of their raw data. Recorded future and Thermopylea Sciences are two data solution companies providing past historical data in terms of cyber-attacks. Feasel and Ramos (2013) do not explain what type of data analytic techniques will be applied to these data. Their proposed model explains a business plan toward a real time protection rather than a technical solution to improve cyber situational awareness. Figure 2-6 explains the main blocks in their method, various inputs from RSS feeds, Web pages etc. will be fed into predictive engines and then the result will go to Correlation and Collaboration block when all the outcomes will get together and the result will be provided to decision makers.

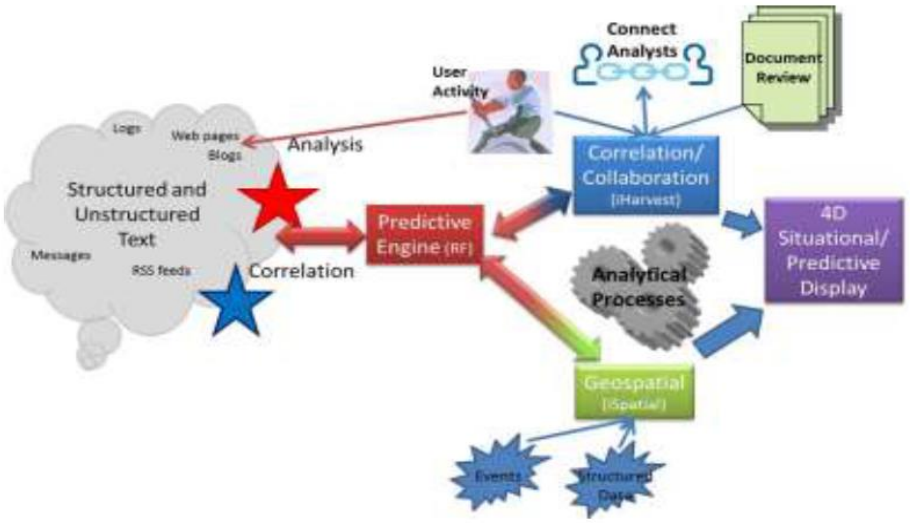

Figure 2-6 Feasel and Ramos (2013) approach for improving CSA

Musliner et al. (2011) suggest an approach based on fuzzy logic which an artificial intelligence technique. As it is shown in figure 2-7, the

approach is divided into 2 parts; proactive and reactive. Proactive block deals with the identification of vulnerabilities in real time and reactive black is concerned with possible prevention solution against different cyber-attacks. The outcome of these will be combined and encoded to the fuzzy system in form of facts and rules. A cyber shield will be constructed against cyber-attacks. Although the fuzzy logic method is not a concern of this research, generating an automatic cyber shield is an interesting subject that can be discussed.

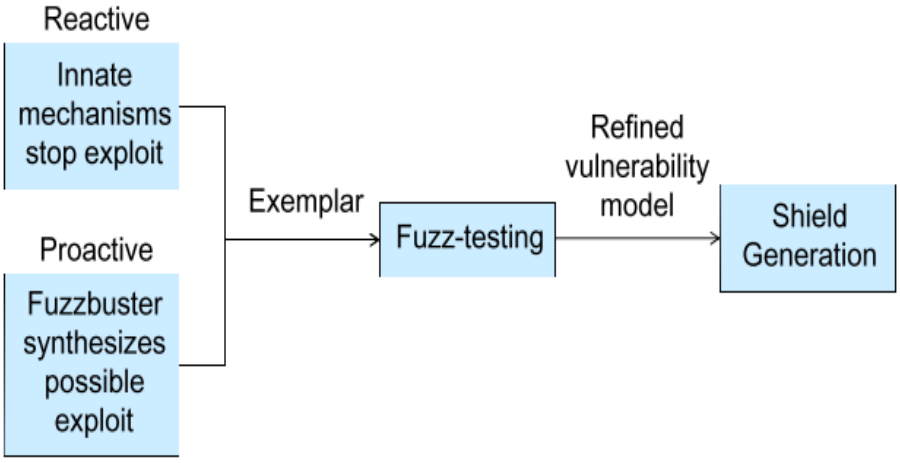

Figure 2-7 Fuzzy defense shield overview (Musliner et al., 2011)

Musliner et al. (2011) proposed a framework based on fuzzy logic for improving CSA which includes generation of cyber shield against future cyber-attacks. Fuzzy logic is a field of Artificial Intelligence and it has few disadvantages which can have a negative effect on performance and accuracy of the framework. The first one is fuzzy logic is not flexible and in the field of cyber situational awareness, there is a need for a dynamic approach for improving current level (Antonik, 2007). In fuzzy logic methods, it is difficult to include all of the detail for the system, therefore, the accuracy of Musliner et al. (2011) will be low and thus the performance does not meet complete and effective improvement of CSA.

Schreiber-ehle and Koch (2012) propose an approach using a data fusion method called JDL for increasing CSA. Data fusion refers to the

process of considering and mixing different data from various resources and forms them in a way to better comprehension and projection for solving issues. In other words, data fusion means recording, storing, filtering, analysing and presenting of the result of the analysis. JDL model of data fusion is a cohesive method presenting each block of data fusion to managers for a better understanding of the process. Figure 2-8 shows proposed JDL model of their methodology which has 6 levels as follows:

1. “Level 0: is a component accepting input from monitoring sensors within the system. IDS is one of those sensors which monitors activities in the system and in case of suspicious one, it will notify the system administrator.

2. Level 1: level 1 aims to refine inputs from level 0, in other words, those raw data extracted by level 0 needs to be pre-processed and allocated to specific objects on the system. For instance, log files and IP addresses recorded by IDS need to map to relevant nodes or objects in the network.

3. Level 2: level 2 investigates for the current relationship between cyber entities through different types of analytic algorithms such as classification, clustering and so on.

4. Level 3: level 3 is the stage of prediction of the future situation based on current and past condition in terms of enemies, threats, vulnerabilities, weaknesses and possible future operations to combat them. Information at this level is extracted from known attacks, signatures, and templates and so on.

5. Level 4: level 4 operation includes observation of overall data fusion to maintain the system performance and try to improve it if it is feasible. Sensor and resource management is the main task at this level.

6. Level 5: level 5 provides Human-computer interaction where the process can be modified and refined by human experts.” (Pournouri and Akhgar, 2015, p.25)

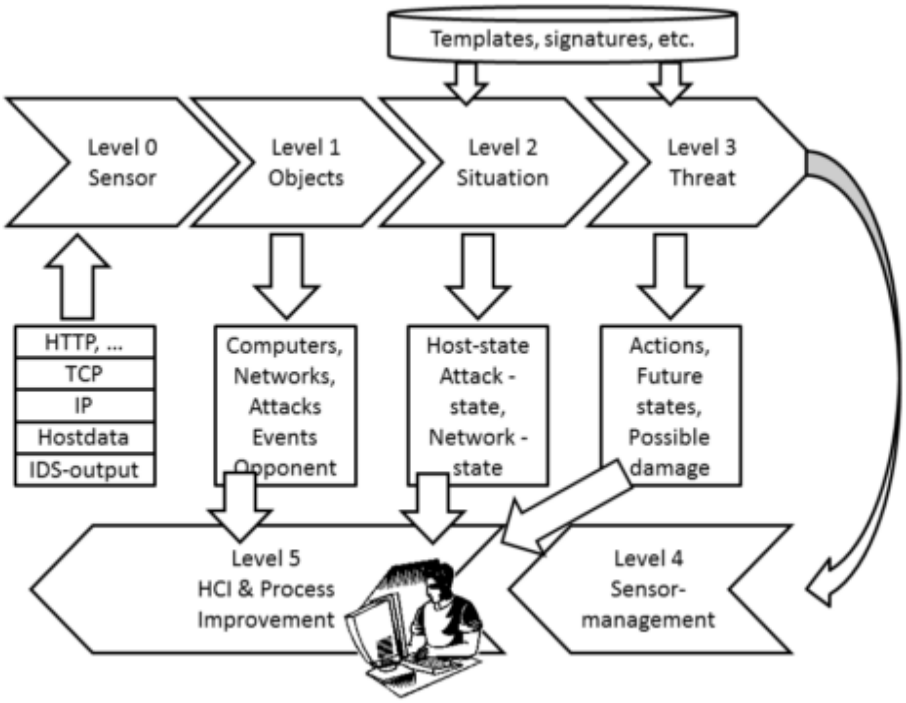

Figure 2-8 JDL fusion system (Schreiber-ehle and Koch, 2012)

This project focus on improvement CSA based on predictive and classification analytics which will be achieved by threat analysis and this aim highlights level 3 of Schreiber-ehle and Koch (2012)'s proposed JDL model. Schreiber-ehle and Koch (2012) draw a general model of improvement of CSA and it can be suitable for projects in big scale. Improving CSA by Scheriber and Koch (2012) based on JDL data fusion has some gaps and disadvantages. As they proposed, their framework has different levels with different inputs but the main gap is combining all of the levels' result together. The more important gap is the projection of result to the managers which was not mentioned in their framework.

Pourboire and Akhgar report that Fayyad and Meinel (2013) design a methodology for prediction of new attack scenarios leading to improve CSA. Figure 2-9 shows their methodology. Fayyad and Meinel (2013) use three main resources of data; IDS database, Attack graphs, and vulnerability database. IDS records all of the alerts and by applying clustering and aggregation algorithms, they will be formed into a suitable format for the correlation process. After correlating alerts, they

will be stored in another built data set. Now it is time for processing attack graphs data. Fayyad and Meinel (2013) define an attack graph as “is a directed graph has two types of vertices, exploit and condition. An exploit is a triple (hs, hd, v), where hs and hd represent two connected hosts and v a vulnerability on the destination host”. By this definition attacks model will be initiated and then in another stage, they will combine with correlated alerts. The next stage is processing vulnerability database and trying to find a relationship between them, attack graphs and correlated alerts. The result of this model provides a real time protection. The advantage of this methodology is staging each part of attack scenario which means defence indicators can have a specific plan for each stage of ongoing attack when it happens.”(Pournouri and Akhar, 2015, p.26)

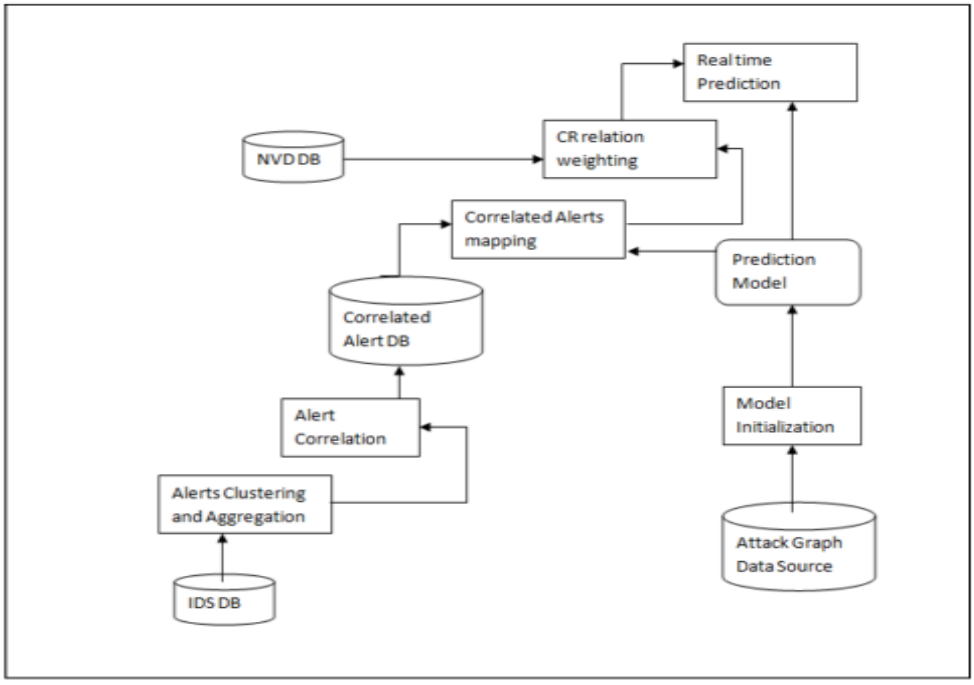

Figure 2-9 Fayyad and Meinel (2013) method for prediction of new cyber-attack scenarios

.

Paxton et al. (2015) suggest that because malware and cyber-attacks are becoming more complicated and sophisticated day by day, conducting current methods of cyber incident analysis is time

consuming and even are not able to help to improve CSA. They proposed a framework including 5 different modules as follows:

1. Traffic and log collection module: The task of this module is capturing and monitoring of network traffic. Network sensors installed on honey-nets will carry out this process. This module has two modes; a) capturing live traffic as packets b) loading captured traffic. The result of this part will be transferred to the second module.

2. Traffic processing module: this module aims to pre-process the obtained result from the previous module and prepare it for flow correlation stage. Firstly, it divided the traffic into 10 minutes time windows in order to investigate behavioural changes over time. The reason for choosing 10 minutes is that it is default time for connection interval. Secondly, some information such as source IP (SrcIP), destination IP (DstIP), source port and destination port will be extracted from heard of the captured packets.

3. Flow correlation module: this module aims to generate communities based on information gained from the second module by utilizing the clique-based method. Clique-based method generates communities by infiltration of fully connected adjacent subgraphs. Primary nodes will be set of IP address and the secondary nodes will be equal to message size. All of the nodes were connected to each other based on links including a tuple of {Time, SrcIP, DstIP, Msglen}. Once the community will be established, Paxton et al. (2015) will use the tuple of {SrcIP, DstIP} in order to identify the role of each primary node. The reason for considering message size as secondary nodes is that they will be able to figure which communication is relevant and meaningful. Thus, the community graphs will be generated in 10 minutes time interval.

4. Botnet behaviour monitoring module: This module aims to monitor behavioural changes of the nodes during the time by comparing the current statues with their previous ones.

5. Bot masters and C&C server (command and control) detection module: based on the obtained result from the fourth module and observing of changes during the time Bot masters and C&C server in a cyber-attack can be detected.

Based on Akhgar's (2015) definition of CSA, Paxton et al.(2015) proposed an investigation method to find out about actors within the cyber in order to which community of computers represents as a malicious group such as bots and how cyber managers can deter their attack to the organization. The advantage of Paxton et al.(2015)'s framework is the ability of identification of bot basters and block them before causing more damages, however, identification of nodes and communication can be time consuming and complicated and sometimes if botnet has a lot of bots it is difficult to use this framework. This approach also focuses on the low level of CSA and it is more technical, although this study attempts to cover, both high and low level of CSA in its proposed approach.

Angelini and Santucci (2017) suggest a visual analytic approach for promoting CSA from the network level to high management level. Their proposed model will be applied to a critical infrastructure in Italy which provides water and electricity purification for cities in central Italy. Angelini and Santucci (2017)'s approach is defined on two layers; the first layer demonstrates the network nodes which makes security specialists able to identify targeted nodes in a cyber-attack incident and the second layer aims to merge the compromised nodes and evaluate their influence in the organization. The mothed handles both geographical and topological view of the network and all the information will be combined with risk factors and will be prepared for decision makers. Figure 2-10 shows an overview of Anglini and Santucci (2017)' approach presenting the critical infrastructure in Italy.

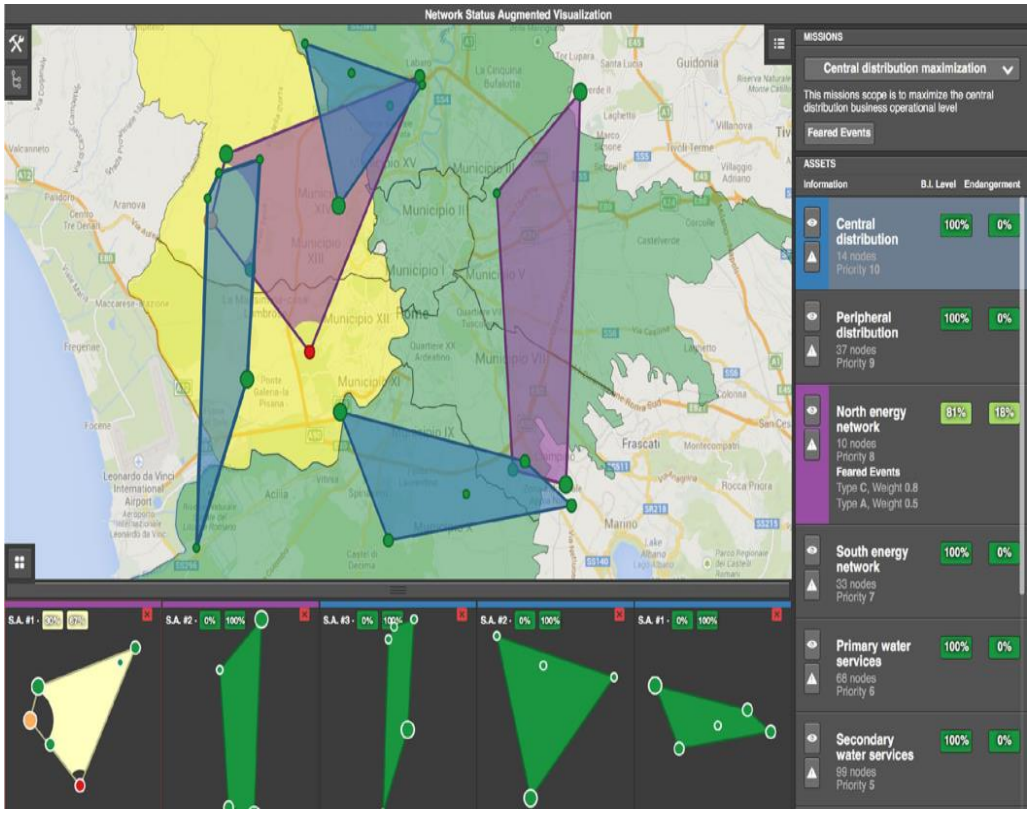

Figure 2-10 visual analytic of critical infrastructure by Angelini and Santucci (2017)

The process of the network appears on the right showing different assets with the number of nodes they have, the blue fragment shows that part of the network are fully functioning and the purple ones show they are compromised. The red node shows this node has been compromised. Angelini and Santucci suggest a method for increasing high level of CSA, however, this method is based on gathering information from technical assets and network nodes to risk analysis and the result will be presented to decision makers. It can be mentioned that the main weakness of this method is lacking any response before cyber-attacks happen so it means mitigation strategies will be in place after any cyber breaches.

Unwubiko (2016) proposes a framework for preserving the security of online web services through enhancing CSA. The proposed framework is based on intelligence gathering through web analytic tools. Web analytic refers to a tool or an approach to collect, analyse and interpretation of web usage data. Figure 2-11 shows intelligence centric web analytic proposed by Unwubiko (2016).

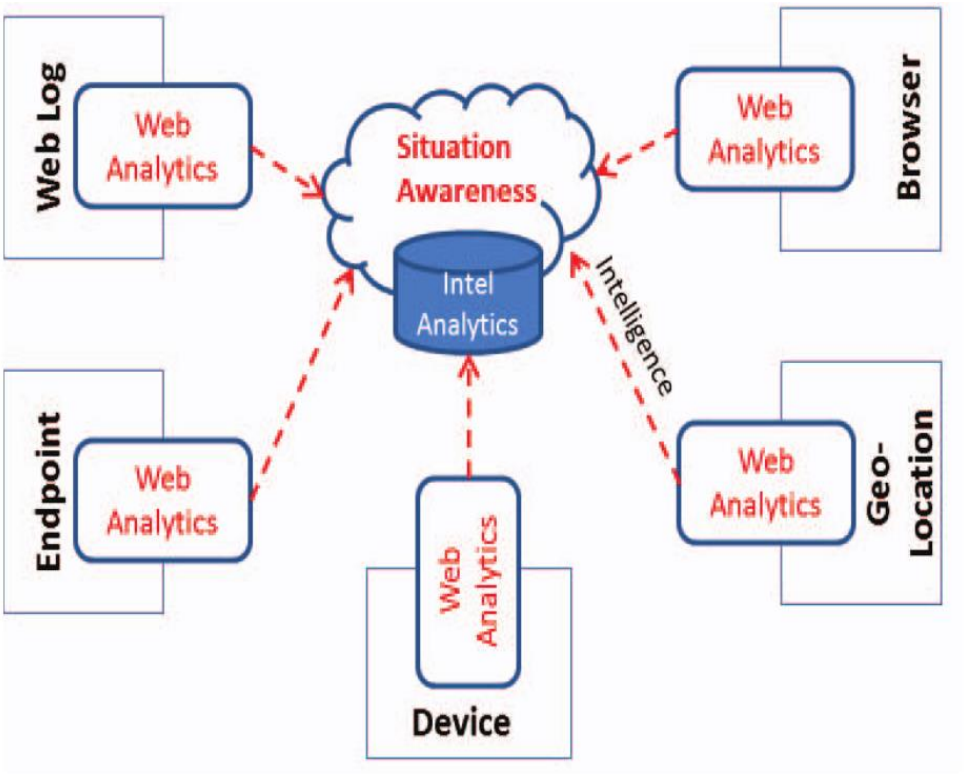

Figure 2-11 Intelligence web-centric approach by Unwubiko (2016

According to figure 2-11, the intelligence framework has 5 different blocks as follows:

1. Web log: when any visit happens in websites, a log will be generated, this log can contain useful information about the location and IP resource of the visitor.
2. Browser: this block contains the information about which browser was used during the visiting the website.
3. Geo-location: This segment includes geographical information about the visitor such as City, Country etc.
4. Device: this block has information about what type of device was used during the visiting the website such as PC, tablet etc.
5. Endpoint: it refers to a block, which contains findings of Operation systems, keyboard and so on from the visitor.

The information in the framework the will be used for profiling of transactions. This profiling is based on 11 key features; date/time, source IP, browser, OS, origin, device, keyboard, endpoint, new/repeat robotics.

The next step is analysing which is based on the comparison of transactions' fingerprinting. After recording transactions, they will be compared against each other and if anything suspicious is seen, it will be flagged. For instance, a visitor can make a transaction with a different type of device and different geo-locations, however, if the geolocations are far from each other and they happen in very little time, the transaction will be flagged as suspicious.

The method proposed by Unwubiko (2016) is focusing on low level of CSA which is more related to logs and transaction gathered by web analytic tools and the main strength of it, is its real time feature, however, there are some disadvantages to this method which can happen sometimes is false alarms and because there is no human interfere in the method sometimes a normal activity can be considered a suspicious or abnormal.

Al shamisi et al. (2016) report that a defensive approach is not enough for tackling cyber-attacks and can make cyber attackers more motivated and innovative to carry out further and more complicated attacks. Therefore, Al shamisi et al. (2016) propose a CSA model which is based on offensive approach towards cyber attackers and it has 2 main feature:

1. It will be based on interaction with cyber attackers which means the active CSA strategy tries to penetrate to cyber criminals' system and influence and sabotage their domain.
2. It will be based on attack deception meaning when an attack happens, it will be forwarded to a deception server.

Figure 2-12 shows their theoretical framework for CSA to maintain the security of cyber space.

Figure 2-12 Detailed overview of the proposed framework by Al shamisi et al. (2016)

comprehension and projection as 3 main blocks in Situational Awareness depends on different factors. Al shamisi et al. (2016)

suggests perception depends on the quality of intelligence, comprehension relies on skills and projection is influenced by the aim and the anticipated result.

Al Shamisi et al. (2016)'s proposed framework has to be applied through 3 main steps:

1. Through the passive, once a cyber-attack happens, the relevant information should be obtained such as motivation, location, attackers' identity etc.
2. After obtaining relevant information about the attackers, their domain will be targeted and compromised actively and some other information will be discovered such as OS, opened ports and running services.
3. The last stage is to do attack deception and carry out counter attack. Protective measures should be considered and the cyber-attack will be redirected to the deception server, meanwhile, a counter attack will be carried out against cyber attackers' domain based on information obtained in previous steps.

The method suggested by Al shamisi et al. (2016) highly relies on the quality of intelligence gathered from cyber attackers. To measure the quality of intelligence, accuracy, completeness, timeliness, and reliability of it should be taken into consideration. This framework is to maintain both high level and low level of CSA and unlike most of other existing methods, it adopts an active and offensive strategy to tackle cyber-attacks.

2.5.3 Summary

To sum up, the following can be distillate from a critical review of the literatures within context of this research:

1- Framework: The governing framework is highly important in context of CSA. Our project seeks to develop an effective framework which it not only fulfils the purpose of this project but also delivers an operational / practical method dealing with improvement of CSA. By reviewing the relevant literatures as

indicated in section 4.4, it can be argued that all of them, same principle in designing their framework where they have same blocks in their framework with different names and same operations, however, not all of them look for same result and performance. Huang et al. (2016) Ahn et al. (2014) and Wu et al. (2013), Feasel and Ramos (2013) and Schreiber-ehle and Koch (2012) do not propose a real time algorithm and their system mainly was designed for decision makers to improve CSA and protect their network through those solutions. Hence, it may not be an effective solution or baseline framework for CSA. Furthermore, the study by Das et al. (2013) follows same path but in different way. In other words, they combine decision-making process with financial issues in their approach of improvement of CSA. On the other hand Morris et al.(2011), Musliner (2011) and Fayyad and Meinel (2013) try to present real time framework of combating cyber breaches by probing technical elements in cyber security. This study can be a combination of both types of proposed models where by analysing past cyber breaches incidents in terms of not only type of attacks and methodology but also cyber attackers motivations and behaviours, it aims to present deeper understating of current situation of cyber environments and its players and predict future conditions based on current and past state.

2- Type of data and its collection: The data collection will be based on Open Source Intelligence because they are publicly accessible and do not raise any ethical issues and this research benefits from Open source intelligence.

3- Type of analysis: Statistical and data mining techniques will be used to improve CSA through data mining methods which will be explained in detail in section 5.

Table 2-3 shows a table as distillation of literature review from type of analysis, type of data and its collection and their framework.

	Framework	**Type of data**	**Type of**	**Type**

		and collection	**analysis**	**of CSA**
Ahn et al. (2014)	Designed for decision making process.	Log, IDS alarms and so on.	Not applicable.	Low level
Das et al. (2013) (i-HOPE)	Designed for resource allocation as a result of CSA understanding	Estimating the cost of cyber-attacks to different businesses through CSI/ FBI annual report	Not applicable.	High level
Wue et al. (2013)	Designed for decision making process.	Past historical attacks, usage of network services, value of assets in the network and vulnerabilities of the system	Bayesian network.	Low level
Fayyad and Meinel (2013)	Designed for protection against cyber-attacks based on past information	IDS data base, NVD data base and attack graphs	Correlation algorithms.	Low level
Feasel and Ramos (2013)	Designed for decision making process.	OSINT and blogs	.Not applicable.	High level
Schreiberehl and Koch (2012)	Designed for decision making	Attack signatures and methods.	JDL data fusion method.	Low level

	process.			
Dut et al.(2012)	Designed for decision making process based on internal behaviours	Attacker and Defender behaviours	Instance Based Learning Theory (IBLT).	High level
Morris et al. (2011)	Designed for task management based on different missions of different levels within the network system.	Each level task	Not applicable.	High level
Musliner et al. (2011)	Designed for identifying technical elements in CSA.	Past cyber-attack signature and common countermeasures against them.	Fuzzy Logic algorithm.	Low level
Paxton et al. (2015)	Designed for cyber breaches deterrence by improving CSA	Using honeynets for detecting cyber bots	Graph-analysis	Low level
Huang e al. (2016)	Designed for generating a comprehensive cyber strategy based on CSA	Using different solutions by cyber analysts to tackle ongoing cyber attacks	Fuzzy based System	High level

	using cyber analysts' solutions			
Al shamisi et al. (2016)	Designed a CSA improvement system to combat cyber-attacks with offensive approach	Intelligence from attackers such as their domains, opened ports and etc.	Theoretical Framework	Low level
Unwubiko (2016)	Designed for Website security by improving CSA	Logs and info which gathered by Web analytic tools	Not Applicable	Low level
Angelini and Santucci (2017)	Designed for improving CSA within critical infrastructures.	Network information such as nodes and compromised nodes.	Visualization and visual analytics.	Low level

Table 2-3 distillation of literature review

2.6 Application of predictive analytic in similar fields

Al-Janabi (2011) proposed a framework for crime data analysis based on classification and clustering algorithms. For classification purposes, Decision trees and for clustering Simple K Means were used in his approach. He used OSINT for source of the data set which was Austin Police Department crime data set. He applied the following stages to the initial data set:

1. Removing outliers: some parts and records of the data are unusable and cannot be included analysing process because of some reason such as incompatibility with the type of analysis, incompleteness and so on.
2. Dealing with missing and unknown values: some of these values can be filled or deleted based on analysis requirements.
3. Data reduction: this stage can be done through normalization and aggregation techniques.

Figure 2-13 shows Al-Janabi (2011) proposed framework including different blocks.

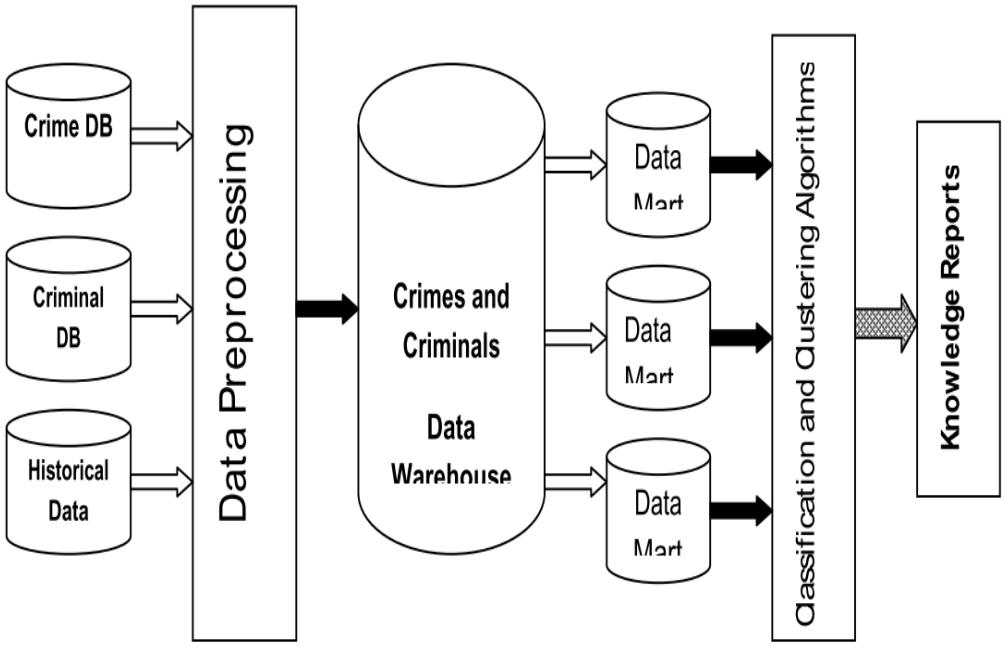

Figure 2-13 crime data analysis framework by Al-Janabi (2011)

First blocks are data sources and by applying pre-processing techniques to data as it is discussed before, they will be stored in data fountain named data warehouse. Data mart blocks are optional blocks designed for performance or data ownership issues. Next block is classification and clustering algorithms and Al-Janabi (2011) explained Decision tree algorithm was chosen for classification because they are easy to be interpreted and used in order to explain dependency between different attributes in a dataset.

Also for clustering purpose, Al-Janabi (2011) benefit from Simple K Means and he reported that this can help to better identification of similar criminal behaviours based on their group. For instance figure 2-14 shows the dependency between the type of offence and sex, income and marital status.

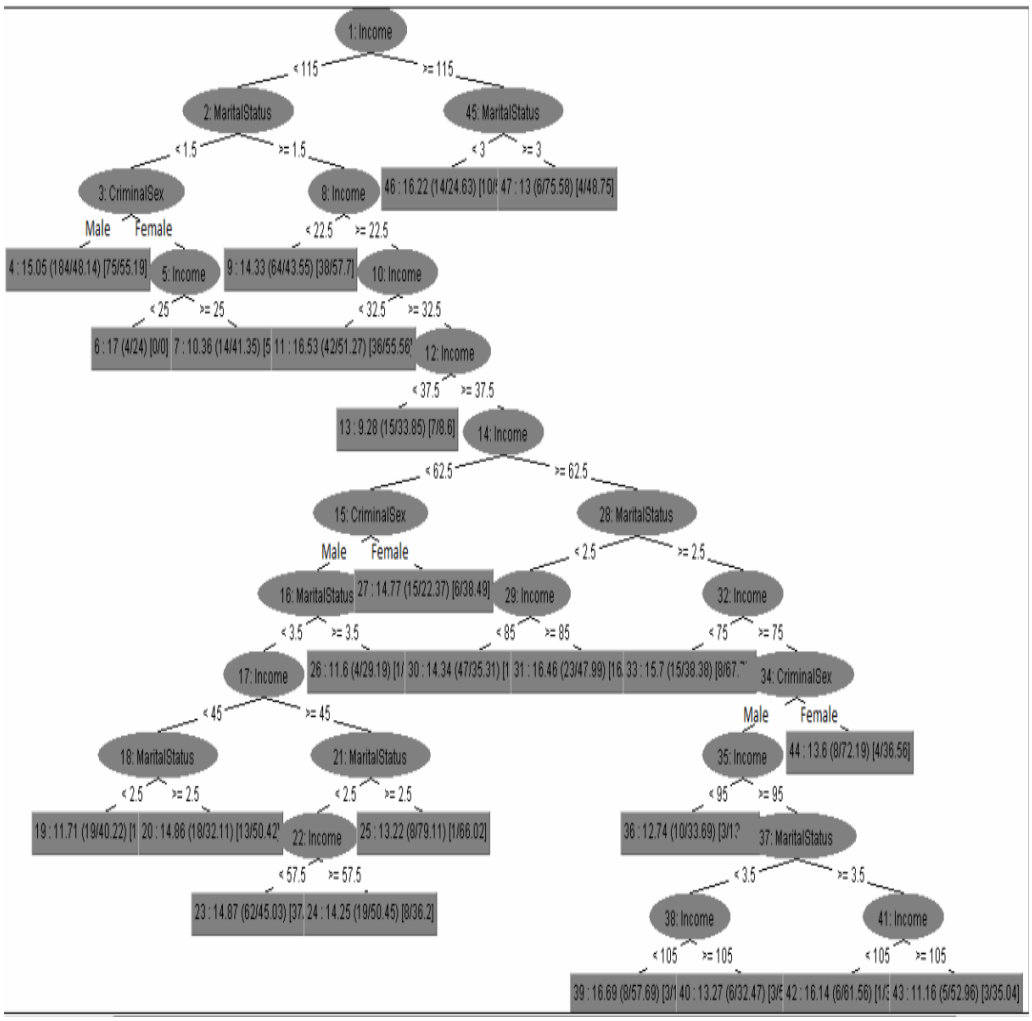

Figure 2-14 Al-Janabi (2011)'s generated decision tree

In another study by Bhardwaj and Pal (2011), Bayesian classification is used for prediction of student performance based on different attributes. They provide a dataset from a college in India showing different attributes of the students. Table 2-4 shows all of the attributes, Description and possible values for each attribute. Bhardwaj and Pal (2011) report that classification is one of the most frequent technique for predictive analytic where the value of categorical attribute can be predicted by analysing others. Bayes classification is also called naïve

because it analyses the data with the assumption of independency between attributes within the data set. Following formula is Bayes theory:

$$P(h_i|x_i) = \frac{P(x_i|h_i)P(h_i)}{P(x_i|h_i) + P(x_i|h_2)P(h_2)}$$

Equation 2-8 Bayes theory (Bhardwaj and Pal, 2011)

Bhardwaj and Pal (2011) justify the usage of naïve Bayes classifier with the following advantages:

1. It is easy to use and only one scan of the training data is needed.
2. Only means and variances of variables are essential for classification purposes.

Bhardwaj and Pal (2011) benefit from Matlab software to implement and apply the methodology to their data set. The final formula is as follow:

$$P(t_i|c_j) = \prod_{k=1}^{p}(x_{ij}|c_j)$$

Equation 2-9 Matlab processed formula of Bayes theory (Bhardwaj and Pal, 2011)

In addition, they define above formulas as “To calculate P (ti) we can estimate the likelihood that ti is in each class. The probability that ti is in a class is the product of the conditional probabilities for each attribute value”.

Based on the analysis Bhardwaj and Pal (2011) conclude the following result showed in Figure 12. It presents the probability of having the effect of each attribute on students’ performance.

Variable	Description	Probability
GSS	Students grade in Senior Secondary education	.8642
LLoc	Living Location	.7862
Med	Medium of Teaching	.7225
MQual	Mother's Qualification	.6788
SOH	Students other habit	.6653
FAIn	Family annual income status	.5672
FStat	Students family status	.5225

Table 2-4 Variable effects on students' performance (Bhardwaj and Pal, 2011)

This project can also use naïve Bayes in order to classify cyber-attack data set like Bhardwaj and Pal (2011) because of different attributes in the cyber-attack data set and their independency and naïve Bayes provides a suitable prediction based on past data.

Nath (2006) uses clustering techniques in order to detect patterns in crime data and he suggests that this technique will draw a deep overview of crimes and help them to solve related issues. Nath (2006)'s approach is a combination of Clustering algorithm and semi-supervised machine learning and it can make prediction more accurate. Clustering algorithms are used for identification of records and allocation of each of them to same groups based on their similarities. The most common clustering algorithm is K-mean dealing with big data sets. In the first stage, Nath (2006) applies pre-processing techniques to raw data set obtained from Sharriff's office under a non-disclosure agreement. This stage includes resolving missing values, removing outliers and so on. Because of the nature of data which is categorical

and nominal Nath (2006) needs to weigh the most important attributes for clustering requirements. Although he suggests that the weighing stage can be done through classification techniques, he benefits from semi-supervised learning which is asking a criminology expert to identify the most significant attributes in the data set. After applying the K-mean algorithm, he mapped the result into a plot and In order to validate the result, he compares it with court disposition. Figure 2-15 shows the geo spatial plot generated by Nath (2006):

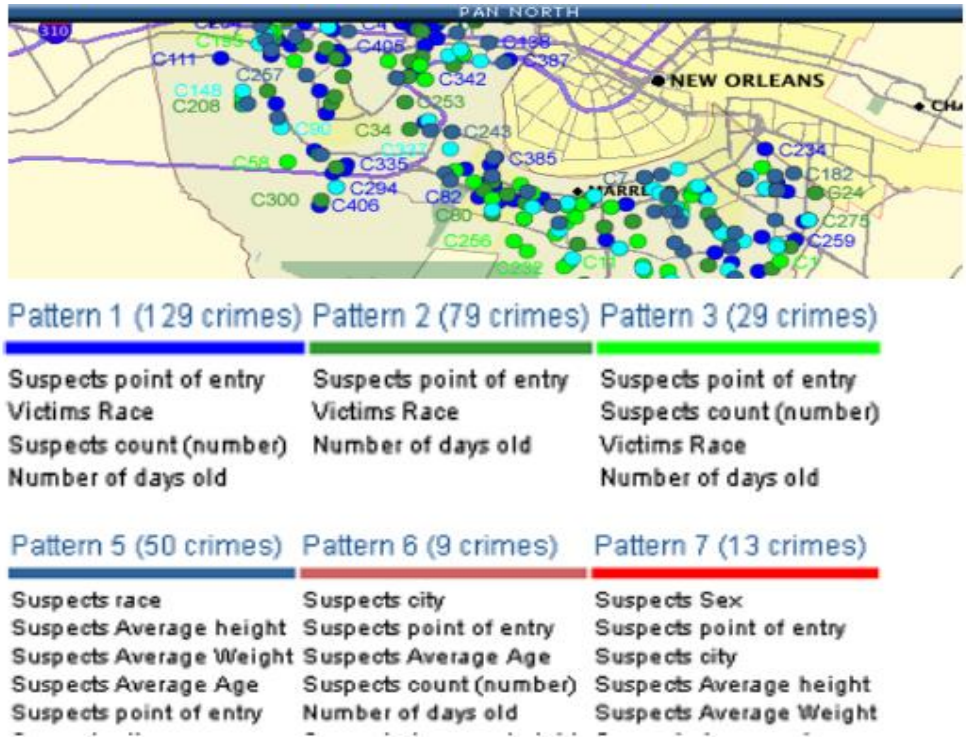

Figure 2-15 geo spatial cluster plot (Nath, 2006)

The reason of mentioning Nath (2006) is the usage of clustering algorithms and the type of the data used in his approach, however, the pre-processing stage was done through classification techniques, which highlight the importance of each factor, by assigning weight to them.

2.7 Conclusion

In this chapter, 4 main concepts have been investigated; Cyber security, Open Source Intelligence, Data mining techniques and Cyber

Situational awareness. Cyber Security basics and components based on aim and objectives of this research have been explained such as cyber threats, different type of cyber activities and cyber attackers and their motivations. The Open Source Intelligence, different types of them have been explained and benefits and disadvantages of them have been discusses. Data mining definition and different techniques have also been described. Specifically regarding to the aim of this study classification techniques have been investigated and demonstrated that why they are suitable for prediction. The main block of this research is Cyber Situational Awareness and the definition and different approaches to improve them have been investigated in this chapter. As it was mentioned, the general approaches for improving them are divided into 2 states; high and low. Based on the literature review, aim, and objectives of this study it has been decided to adopt both high level and low level approaches. In the last section of this chapter, different applications of predictive analytic techniques have been explained in form of a couple of relevant examples to the aim of this project.

The next chapter will discuss and explain the method which has been adopted by this research.

Chapter 3 Data Structure and Pre-processing

3.1 Introduction

This research aims to design a framework applying data mining and predictive analytic techniques applied to past historical data in terms of cyber-attacks to predict future cyber incidents which will help to a deeper understanding of cyber situational awareness. Based on the literature review the research question will be addressed in this project. This chapter aims to demonstrate the data type, structure and pre-processing stage. Different tools which have been used, will be explained. Pre-processing in any data mining project is the most critical step.

3.2 Tools and platforms

3.2.1 Open Refine

In this research, it has been decided to employ a very powerful tool which is developed by Google initially in Java language called Open Refine. Open refine includes different functions for data cleansing and transformation of data. This web-based tool supports the following functions (Verborgh and De Wilde, 2013):

1. Supporting different formats of datasets
2. Support basic and advanced cell transformation within the dataset.
3. Dealing with multiple and missed valuables
4. Connecting different datasets together
5. Supporting regular expression to filter and divide a dataset
6. Supporting the General Refine Expression Language to run different and advanced data operations.

3.2.2 R

R programming language has been adopted in this study. R is a programming language supporting a wide range of statistical, graphical

data mining techniques. These techniques include classic statistical operations, classification, clustering, linear and nonlinear regressions etc. R is an extension of S programming language which was developed by John Chambers and his colleagues. The R language adds an important feature to S programming language which enables developers to actively contribute to developing new functions. In the other words, R provides an open source way by supporting developers' packages including a wide range of functions. R is supporting the following features (Team, 2000):

1. Significant and various ways to store and handle different formats of data.

2. A wide range of basic and advanced tools for data analysis.

3. Dealing and running arrays and matrices

4. Supporting basic components of any programming language such as loops and conditioning.

5. Facilitating graphical functions and outputs.

For the implementation of the R language, either its own environment can be used or other platforms. R studio has been used in this study as a platform, which provides an environment for programming and executing of R (Studio, 2012). Figure 3-1 Shows R studio environment.

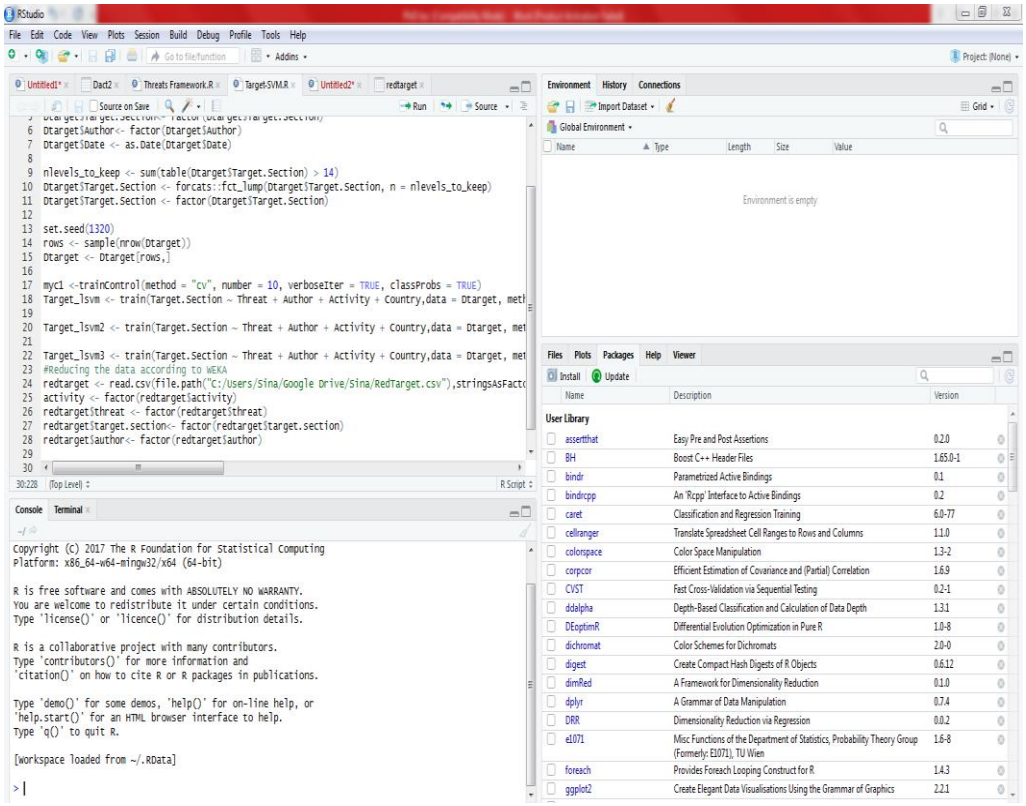

Figure 3-1 R Studio environment

As it is shown R studio has 4 windows, the top on the left side is an area which programmes will be written and down left is the execution and console area. On the right side, at the top the instances and workspace will be loaded and in the below different packages are shown which can be loaded based on the type of operations and needs.

3.2.3 WEKA

Weka is another tool which has been used to verify the accuracy of the models and also it will be used in the evaluation stage of this research. Weka was developed initially in 1997 by University of Waikato and it is based on Java programming language. Weka supports different data mining tasks such as clustering, classification and so on. It also provides pre-processing techniques such as normalization, dealing with missing values and etc. Weka is a significant platform for data analysts for knowledge discovery purposes. It is free and open source

and can be run on any basic computer. It also helps data analysts to visualize their result for presenting to decision makers and managers.

3.3 Data Collection

This subchapter aims to demonstrate the data collection resource and process. The main resource of the data in this research is OSINT as it is impossible to get access to information gathered by governments and official agencies. In addition, there are different companies such as Recorded Future that they collect information about cyber-attacks but they want to sell them for commercial purposes. As it was mentioned previously using OSINT enables researchers to get access to information more cost-effectively and without any ethical concern, however, the data gained from OSINT comes with inaccuracy, incompleteness and lots of noises sometimes and it is data analysts' job to address those issues.

There are a couple of blogs, websites and Twitter accounts, which broadcast cyber breaches and incidents happening all around the world in any business or any sector. Websites and blogs like hackread.com, International Business Time DarkRead.com and etc. are resources of cyber attacks' data. In addition, there is a blog called hackmagedon.com, which gathers hacking and cyber-attacks incidents from different resources and forms them into timeline format month by month. It has been decided to use hackmagedon blog as an initial data resource and collect cyber-attacks happened from 2013 to end of 2015 as the most recent cyber breaches for training the predictive models and we employ the data from 2016 to end of March 2017 for testing and validation of our final model.

The initial dataset includes 3851 records of cyber-attacks incidents, however, as it is said previously the dataset will be divided for training and validation process so the training set has 2694 records and the validation set consists of 1157 records.

In next section, the data structure and data pre-processing will be explained in order to prepare data for next steps and further analysis.

3.4 Data Categorization

In this stage, the initial data set needs to be probed in more detail. Firstly, the data structure needs to be described. Based on the obtained data, each incident of cyber-attack comes with 9 different features:

1. Date of incident: The time of cyber-attack incident.

2. Cyber attacker: who was behind the cyber-attack? For example, Anonymous is one of the cyber attacker groups.

3. Cyber threat: it refers to a Type of Threat that the cyber-attack poses to its victim such Denial of Service and etc.

4. Type of Target: it describes the nature of the target(s) of the cyber-attack such as the University of Maryland as an Educational organization etc.

5. Target: it describes the name of the target, for example, Mitsubishi

6. Targeted Country: it refers to the origin of the target(s) of the cyber-attack such as Mitsubishi in Japan.

7. Type of Cyber-attack: This feature explains the type of cyber-attack in terms of the type of activity such as Hacktivism and etc.

8. Description: This column describes how the cyber-attack happened based on Open source intelligence available for it.

9. OSINT resource: It refers to the reference of OSINT related to the cyber-attacks.

A sample of dataset is shown in the appendix chapter 9.

As this project aims to train models based on the classification algorithm and Time-series Analysis is not the area of this study, Date column will be eliminated. In addition because Targets specifically are not area of interest, the Target column will also be removed from data set, however, the Type of Target column will still remain. In addition, the description column and OSINT resource will be eliminated because they are not involved in the predictive model.

After restructuring the dataset and forming the attributes into 5 different columns, preparation and pre-processing the values will begin. There are different tools for data cleansing such as Excel, R and etc., however, as it mentioned in section 4.2.1 Open refine will be used as a data cleansing tool. The data will be fed into Open refine and pre-processing method will be divided into the following steps:

1. Removing Doubles and ambiguous records: This stage will be done through facet text, as the values in our dataset are nominal values. Facet feature in open refine gives a general and extensive overview of values in one column or in the whole dataset. Doubles will be removed easily by apply text facet to description column, so any record with more than one description will be removed. In addition, those ones with no description or no reference will be removed. The cells in Cyber attacker and Type of Threat with no value, they will be named “Unknown”.

2. Removing irrelevant records: This stage will be done manually. Although it is highly time-consuming because it needs to go through the description column and probing if an attack actually happened or it was just a claim.

3. Dealing with capital and small letters: This stage also needs to be done in order to make the data ready for analysis. Another significant feature of Open Refine is clustering the texts that can help to unify small and capital letters automatically. This will help to integrate the dataset for further analysis.

4. Integration of values: Values need to be integrated in order to make them ready and categorize them well for further analysis:

    a) The first step is categorizing the Type of Targets. Targets based on their nature of business will be categorized. Type of targets have been shown in appendix chapter 9.

    b) The next step is the categorization of the Type of Threats posed by cyber-attacks. Type of threat categories in chapter 9 in appendix

    c) The third step is naming the Target countries by their acronym in order to make their name shorter and easier for analysing purposes. A full list of countries and their acronym is available on Nation Online Project (2017).

5. Based on the data analysis requirement, the dataset needs to be divided into two datasets, one for training the predictive model and one for evaluation of it. The training set includes 2694 records and it starts from the beginning of 2013 to the end of 2015 and the testing set consists of cyber-attack incidents from 2016 to end of March 2017 with 1157, this split point was considered based on the timeline of this project and the most recent cyber-attacks.

6. Removing irrelevant columns: as it was mentioned previously, the dataset has 9 columns, so due to the data analysis requirement 3 columns need to be eliminated from the data that are; Date, Description and OSINT resource.

After data pre-processing stages, the next step is data analysing step which is extensively explained in the next chapter of this thesis.

3.5 Summary

This chapter discussed about data collection and different methods and techniques employed in order to prepare the data for further analysis and training purposes. The next chapter will investigate applying different classification techniques to the data in order to prepare

Chapter 4 Data analysis and applying classification techniques

4.1 Introduction

As it was mentioned in section 1.1.3 based on the dataset and prediction purpose 5 main classification algorithms will be used in this research; Decision Tree, K nearest neighbour, Naïve Bayes, Support Vector Machine and Multilayer Perceptron.

In this chapter, the models will be trained according to the training set and classification algorithms. There are 2 elements that need to be highlighted in the training process. The first one is Train Control which is the type of validation of the accuracy of each model. K fold cross validation has been chosen for this purpose which according to Yin et al. (2011) Is more efficient and accurate compared to dividing the dataset into training and test sets. The reason is in k fold cross-validation, in the training process, the dataset will be partitioned into k equal size sets, then 1 set will act like a training set and k-1 set will be considered as training set and the training set repeats k times. This method presents more accurate and solid result due to repetitive cycle and in this research k is considered 10 so the training process will repeat 10 times.

The second training element, which is different in each classification algorithm, is Tune Grid. Tune Grid allows the training process to be tuned in order to get more accurate and concrete result and helps to the better understanding of training models when they have different reactions to different changes.

One of the vital and useful packages which is being employed in this section is the Caret package. The Caret package was developed by Max Kuhn (2008) which includes functions to train predictive models based on classification and regression algorithms. The goals of developing Caret package are as follows:

1. Integration of different functions for building a predictive model.
2. Creating fast and semi-automated methods for tuning the models in order to optimize them.

Caret also makes data analyst able to compare different models and help them to choose the best model in terms of accuracy and efficiency. In this research, Caret package will be the main package using for the training process which enables the automatic way of train control and tune grid application.

4.2 Decision Tree Analysis

After Categorization and pre-process of the initial data, in this stage, it has been decided to analyse the data by decision tree methods which is described extensively in section 2.1.1 and R has been used as an analytical programming language.

As it was discussed in section 2.2.1.1, Freund and Mason (1999) divide decision tree techniques into 3 main categories; C4.5, Random Forest and Recursive partitioning so it has been decided to analyse the obtain data set by these algorithms and then choose the best one among them. Package RWeka developed by Hornik et al. (2017) for C4.5, Package Rpart by Therneau et al. (2017) for Recursive Partitioning and Random Forest package fo Liaw and Wiener (2015) will be used for implementation of different decision tree algorithms.

4.2.1 Prediction of Type of Threat by Decision tree

Among 2694 records in our dataset there are 2210 cyber-attacks which the type threat known to security experts. Therefore, they will be used as the training set to make the predictive model. It has been decided to convert those threats appear less than 9 times to other which can make the more extensive and accurate.

Now C4.5 algorithm will be used in order to train our model. The training control is set to 10 fold cross-validation and variation of

confidence level from 0.1 to 0.5. Figure 4-1 shows the trained model for the Type of Threat prediction by C 4.5.

```
Threat_j48 <- train(Threat ~ Target.Section + Author + Activity + Country, data
= CThreat, method = "J48",trControl = trainControl(method = "cv", number = 10,
returnResamp = "all", search = "random"), tuneGrid = data.frame(C = (1:5)/10))

Print(Threat_j48)

C4.5-like Trees



2210 samples

   5 predictor

  11 classes: 'AH', 'CSS', 'DF', 'DH', 'DS', 'MWV', 'SQ', 'TA', 'UA', 'ZD',
'Other'

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 1990, 1990, 1988, 1988, 1989, 1991, ...

Resampling results across tuning parameters:



  C    Accuracy   Kappa

  0.1  0.5932992  0.5096207

  0.2  0.5960080  0.5133991

  0.3  0.5887434  0.5049543

  0.4  0.5864645  0.5014032

  0.5  0.5864543  0.5013415

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was C = 0.2.
```

Figure 4-1 Training process of Type of Threats' predictive model based on C4.5

Figure 4-2 shows the plot of our model based on confidence and accuracy level. As it is shown in the plot the maximum accuracy of the model which is 59.60 % happens when confidence rate is 0.2 and it can be interpreted when the confidence rate is more than 0.2 there

should not be more pruning and if the size of the tree is getting larger the accuracy will be decreasing.

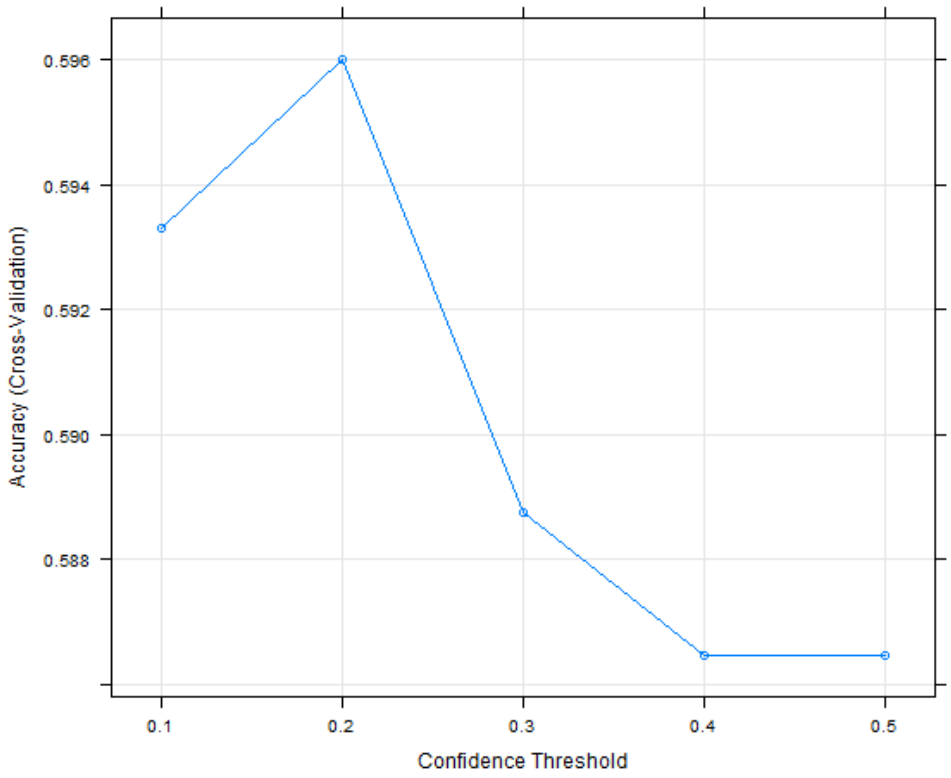

Figure 4-2 Prediction of the Type of Threat Accuracy vs Confidence Threshold in C4.5

Recursive portioning will be applied to the training set to make a predictive model in the second stage. Same stages like C4.5 will be applied and 10 fold cross validation will be used for train control parameters. The following codes shown in figure 4-3 describe our model.

```
Threat_rpart <- train(Threat ~ Target.Section + Author + Activity + Country,
data = CThreat, method = "ctree",trControl = trainControl(method = "cv", number
= 10, returnResamp = "all", search = "random"))

> Threat_rpart

CART



2210 samples

   5 predictor

  11 classes: 'AH', 'CSS', 'DF', 'DH', 'DS', 'MWV', 'SQ', 'TA', 'UA', 'ZD',
'Other'



No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 1988, 1988, 1989, 1989, 1987, 1989, ...

Resampling results across tuning parameters:


  cp           Accuracy   Kappa

  0.001872440  0.5508298  0.4575610

  0.005266238  0.5295657  0.4323073

  0.057050907  0.3727746  0.2258066


Accuracy was used to select the optimal model using  the largest value.

The final value used for the model was cp = 0.00187244.
```

Figure 4-3 Type of Threat prediction by Recursive Partitioning

Figure 4-4 shows the plot of our model based on criterion and accuracy level and it illustrates that the model reaches its maximum accuracy in criterion of 0.0018 and it is 55.08% accurate and reliable and the complexity parameter indicates that increasing the size of the tree will lead to less accurate model and prediction as the result.

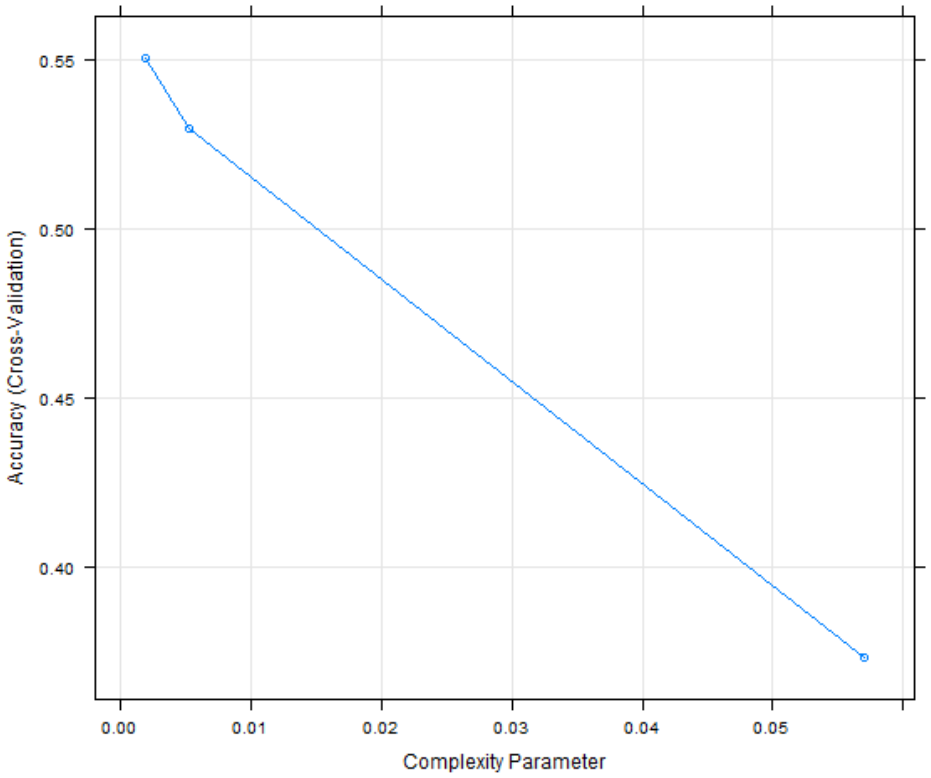

Figure 4-4 Prediction of Type of Threat Accuracy vs Criterion in RP

Random forest is the third algorithm which will be applied to make the predictive model for threat prediction. The train control is the same as C4.5 and Recursive partitioning. Figure 4-5 describes the training process of the model.

```
Threat_rf <- train(Threat ~ Activity + Target.Section + Country + AuthorID, data
= CThreat, method = "rf",trControl = myCont, importance = TRUE)

> Threat_rf

Random Forest

2210 samples

   5 predictor

  11 classes: 'AH', 'CSS', 'DF', 'DH', 'DS', 'MWV', 'SQ', 'TA', 'UA', 'ZD',
'Other'

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 1988, 1990, 1989, 1987, 1988, 1991, ...

Resampling results across tuning parameters:

mtry  Accuracy   Kappa

    2   0.2267059  0.0000000

   34   0.5923942  0.5089259

  595   0.5905269  0.5068367


Accuracy was used to select the optimal model using  the largest value.

The final value used for the model was mtry = 34.
```

Figure 4-5 Type of Threat prediction by Random Forest

Figure 4-6 shows the plot of the model based on accuracy level and number of randomly selected variables. It illustrates that the most accurate model with 59.23 % will be reached when 34 predictors randomly selected by the random forest algorithm

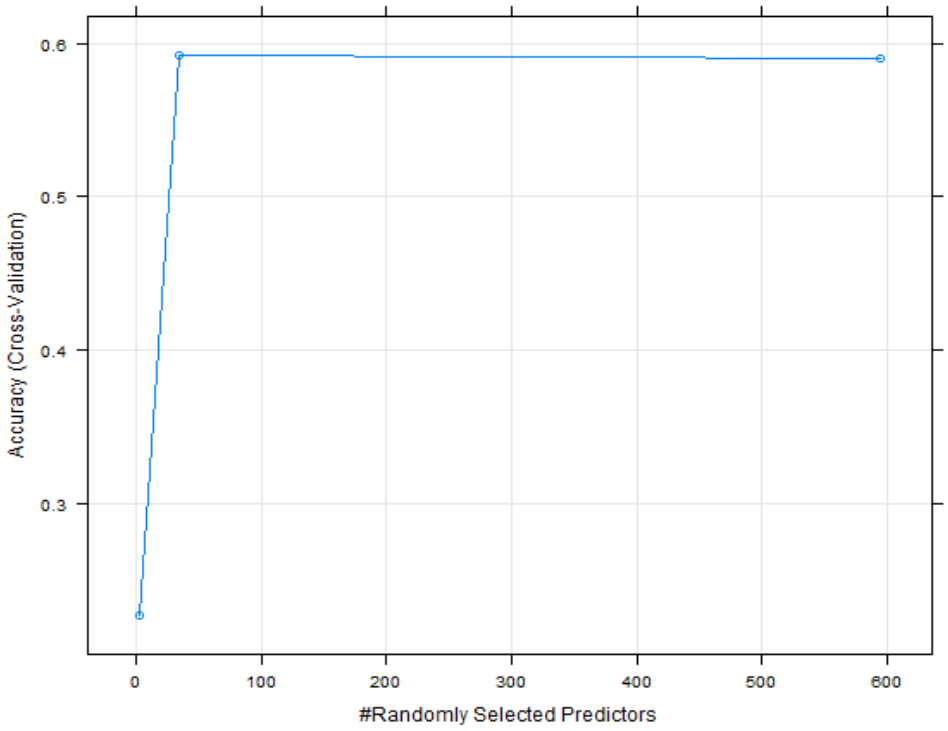

Figure 4-6 Prediction of Type of Threat Accuracy vs Randomly Selected Predictors in Random forest

In order to compare models and determine the most accurate and reliable one, Caret package has another function called Resample. This function compares all models and gives a comprehensive overview of all of the obtained models and help decision makers to choose the desirable model. Figure 4-7 shows applying Resample function to the decision tree models for Type of Threat prediction:

```
threat_model_list <- list(threatc45 = Threat_j48, threatrandomf = Threat_rf,
threatrecp = Threat_CT )

resamples_threat <- resamples(threat_model_list)

summary(resamples_threat)

Call:

summary.resamples(object = resamples_threat1)


Models: threatc45, threatrandomf, threatrecp

Number of resamples: 10









Accuracy

                Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's

threatc45     0.5381  0.5727 0.5946 0.5960  0.6043 0.6606    0

threatrandomf 0.5541  0.5590 0.5960 0.5924  0.6200 0.6393    0

threatrecp    0.4887  0.5174 0.5573 0.5508  0.5747 0.6147    0
```

Figure 4-7 Comparison of Decision Tree models by Resample function

Figure 4-8 visualizes the comparison of the models in order to make it more understandable.

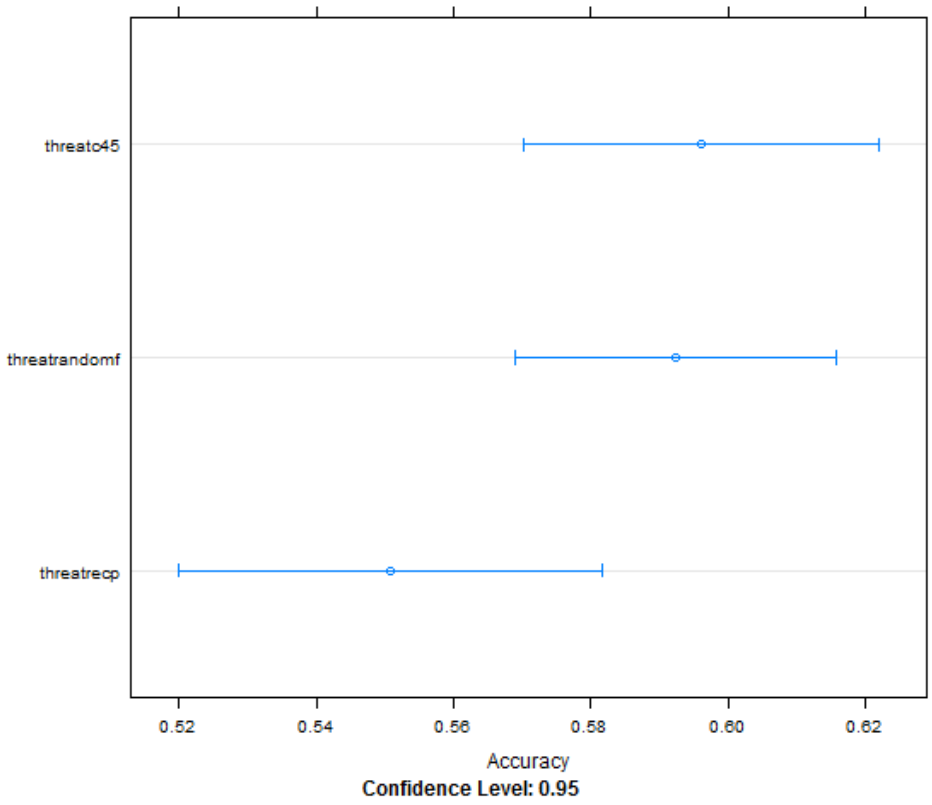

Figure 4-8 Comparison of the decision trees for prediction of Type of Threat

As it is shown in the plot and scripts, the model obtained by C4.5 algorithm has better average accuracy than others have and can be the first option between decision trees in order to predict future cyber threats. Random Forest also performs slightly less accurate than C4.5, however, they are both very close with accuracy level of 59.24% and 59.60% respectively and the Recursive partitioning tree for cyber threat prediction can be the last option with accuracy level of 55.08%.

4.2.2 Prediction of Cyber Attackers by Decision tree

Among 2694 cyber-attacks happened from 2012 to 2015, 1432 cyber-attacks were linked to known attackers or the attackers took credit for them. Therefore, they are included in the training set. It has been decided to convert the level those attackers who carried out attack less than 5 times to Other to get more accurate and extensive result. The following process will explain applying decision tree algorithms to them:

C4.5 is applied to our training set with train control adjusted to 10 fold cross-validation and confidence level varies between 0.1 and 0.5. Figure 4-9 shows the training process of cyber attackers model by C4.5

```
Author_j48 <- train(Author ~ Target.Section + Threat + Activity + Country, data
= Cauthor, method = "J48",trControl = myCont1, tuneGrid = data.frame(C =
(1:5)/10))


C4.5-like Trees


1432 samples

   5 predictor

  32 classes: '@smitt3nz', '@th3inf1d3l', 'Ag3nt47', 'AnonGhost', 'Anonymous',
'Armada Collective', 'Chinese hacker', 'Cyber Islamic State', 'CyberBerkut',
'DarkWeb Goons', 'DERP', 'Dr.SHA6H', 'Guccifer', 'HAXOR', 'Iranian Hackers',
'Izz ad-Din al-Qassam Cyber Fighters', 'JokerCracker', 'KelvinSecTeam',
'LizardSquad', 'LulzSec', 'Maxney', 'NetPirates', 'NullCrew', 'RedHack', 'Rex
Mundi', 'Syrian Electronic Army', 'TEAM MADLEETS', 'TeamBerserk', 'Tunisian
Cyber Army', 'Turkish Ajan', 'XTnR3v0lT', 'Other'


No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 1290, 1288, 1290, 1288, 1293, 1290, ...

Resampling results across tuning parameters:


  C    Accuracy   Kappa

  0.1  0.6068296  0.4251467

  0.2  0.6074943  0.4400457

  0.3  0.6047473  0.4368234

  0.4  0.6054129  0.4368838

  0.5  0.6012011  0.4322257


Accuracy was used to select the optimal model using the largest value.

The final value used for the model was C = 0.2.
```

Figure 4-9 Cyber attackers prediction by C4.5

Figure 4-10 shows the plot of changes of accuracy based on confidence level. As the script and plot show the accuracy rate will be maximum with a value of 60.74% on confidence level 0.2 meaning that the tree does not need more pruning, otherwise the accuracy will be reduced.

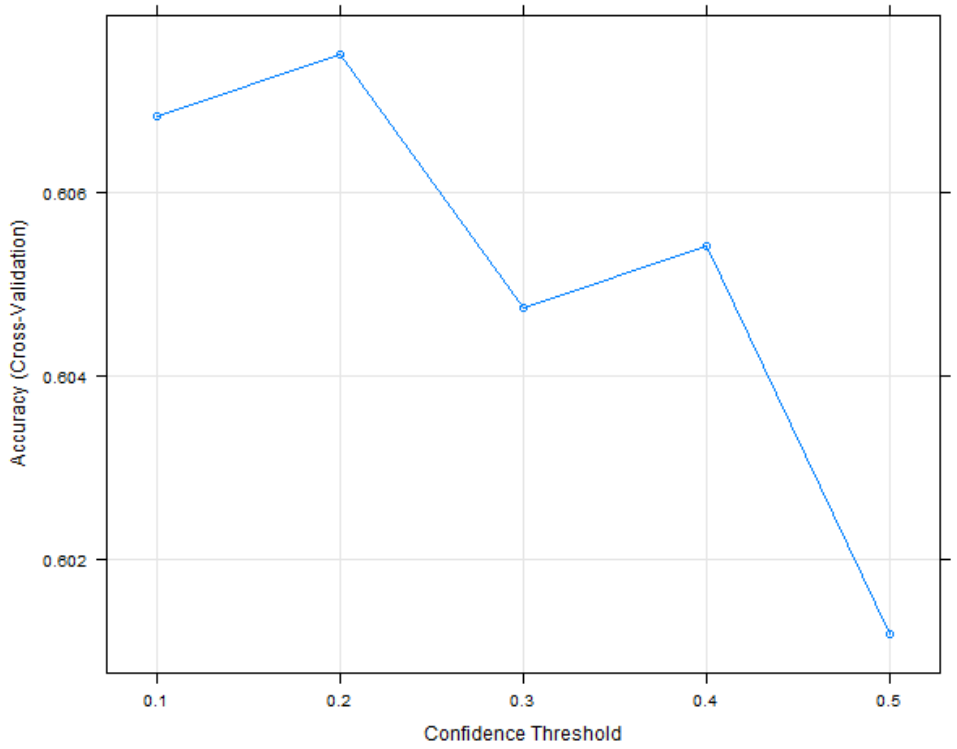

Figure 4-10 Prediction of Attackers Accuracy vs Confidence Threshold in C4.5

In the next stage, Recursive partitioning will be applied to the training data set with train control set to 10 fold cross-validation and figure 4-11 shows the training process of cyber attackers' predictive model based on Recursive Partitioning.

```
Author_CT <- train(Author ~ Target.Section + Threat + Activity + Country, data
= Cauthor, method = "ctree",trControl = myCont1)

> Author_rpart

CART



1432 samples

   5 predictor

  33 classes: '@smitt3nz', '@th3inf1d3l', 'Ag3nt47', 'AnonGhost', 'Anonymous',
'Armada Collective', 'Chinese hacker', 'Chinese hackers', 'Cyber Islamic State',
'CyberBerkut', 'DarkWeb Goons', 'DERP', 'Dr.SHA6H', 'Guccifer', 'HAXOR',
'Iranian Hackers', 'Izz ad-Din al-Qassam Cyber Fighters', 'JokerCracker',
'KelvinSecTeam', 'LizardSquad', 'LulzSec', 'Maxney', 'NetPirates', 'NullCrew',
'RedHack', 'Rex Mundi', 'Syrian Electronic Army', 'TEAM MADLEETS', 'TeamBerserk'
'Tunisian Cyber Army', 'Turkish Ajan', 'XTnR3vOlT', 'Other'



No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 1286, 1290, 1286, 1293, 1292, 1289, ...

Resampling results across tuning parameters:


  cp           Accuracy   Kappa

  0.002538071  0.5986191  0.3924139

  0.015228426  0.5744204  0.3525648

  0.057741117  0.5009787  0.1566863


Accuracy was used to select the optimal model using the largest value.

The final value used for the model was cp = 0.002538071.
```

Figure 4-11 Training process of Cyber attackers' predictive model based on RP

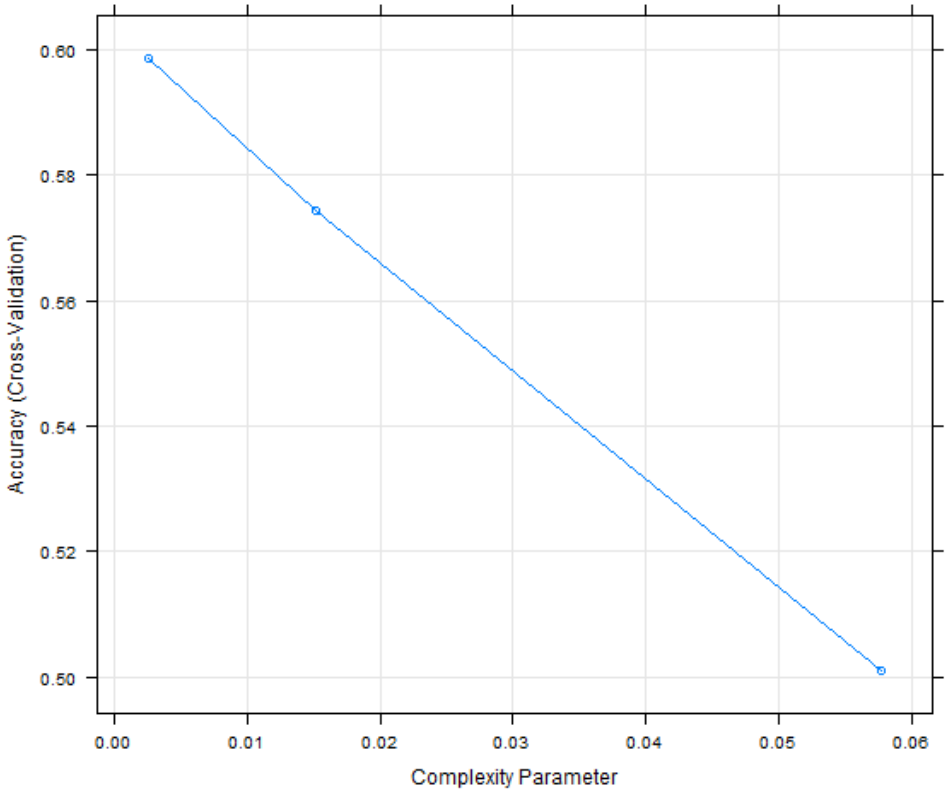

Figure 4-12 Prediction of Attackers Accuracy vs Complexity Parameter in RP

Figure 4-12 shows the changing trend of accuracy based on the complexity parameter in recursive partitioning tree. The maximum accuracy of RP tree which is 59.86 % in the prediction of attackers happened when the complexity parameter is minimum and it is 0.0025 meaning that a smaller tree will be more accurate.

Random forest is the third algorithm applying to the training set. 10 cross fold validation is used as train control and figure 4-13 demonstrates the training process of cyber attackers' predictive model based on Random Forest algorithm.

```
Author_rf <- train(Author ~ Activity + Target.Section + Country + Threat, data
= Cauthor, method = "rf",trControl = myCont,importance = TRUE)

Random Forest



1432 samples

   5 predictor

  32 classes: '@smitt3nz', '@th3inf1d3l', 'Ag3nt47', 'AnonGhost', 'Anonymous',
'Armada Collective', 'Chinese hacker', 'Cyber Islamic State', 'CyberBerkut',
'DarkWeb Goons', 'DERP', 'Dr.SHA6H', 'Guccifer', 'HAXOR', 'Iranian Hackers',
'Izz ad-Din al-Qassam Cyber Fighters', 'JokerCracker', 'KelvinSecTeam',
'LizardSquad', 'LulzSec', 'Maxney', 'NetPirates', 'NullCrew', 'RedHack', 'Rex
Mundi', 'Syrian Electronic Army', 'TEAM MADLEETS', 'TeamBerserk', 'Tunisian
Cyber Army', 'Turkish Ajan', 'XTnR3vOlT', 'Other'



No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 1291, 1293, 1288, 1291, 1290, 1286, ...

Resampling results across tuning parameters:



  mtry  Accuracy   Kappa

    2   0.4498520  0.0000000

   77   0.5958310  0.4348090

  152   0.5798063  0.4222175



Accuracy was used to select the optimal model using the largest value.

The final value used for the model was mtry = 77.
```

Figure 4-13 Training process of Cyber attackers' predictive model based on

Figure 4-14 shows the plot presenting trend of accuracy change over randomly selected variable. The accuracy will reach to its peak level when 77 predictors will be selected randomly, so the accuracy of the attacker predictor model based on random forest is 59.58%.

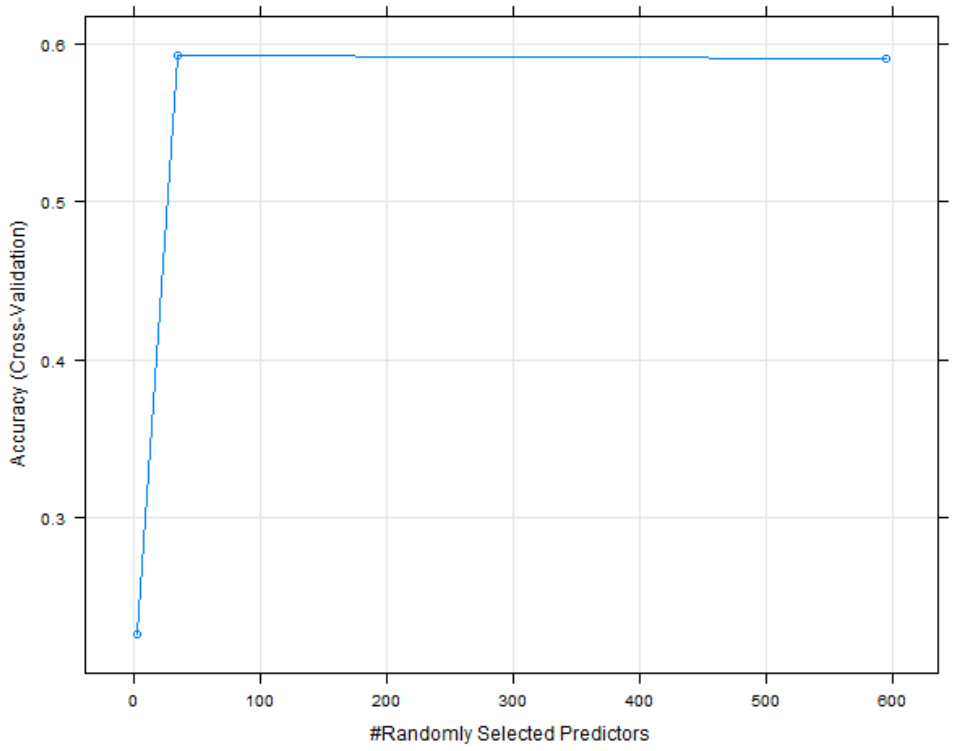

Figure 4-14 Prediction of Cyber Attackers Accuracy vs randomly selected predictors in Random Forest

In the last stage, 3 models need to be compared by Caret in order to see which model is more optimal and accurate. The codes shown in figure 4-15 demonstrates the comparison process by Resample function

```
Author_model_list <- list(AttackerC45 = Author_j48, Attackerrandomf = Author_rf,
Attackerrecp = Author_CT)

resamples_author <- resamples(Author_model_list)




Call:

summary.resamples(object = resamples_author1)


Models: AttackerC45, Attackerrandomf, Attackerrecp

Number of resamples: 10


Accuracy

                  Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's

AttackerC45     0.5694  0.5877 0.6076 0.6075  0.6252 0.6549    0

Attackerrandomf 0.5694  0.5758 0.5925 0.5958  0.6088 0.6454    0

Attackerrecp    0.5646  0.5836 0.5908 0.5986  0.6058 0.6620    0
```

Figure 4-15 Comparison of C4.5, Random Forest and RP in terms of accuracy of cyber attackers' predictive model

Figure 4-16 shows the comparison of three models in a dot pot plot.

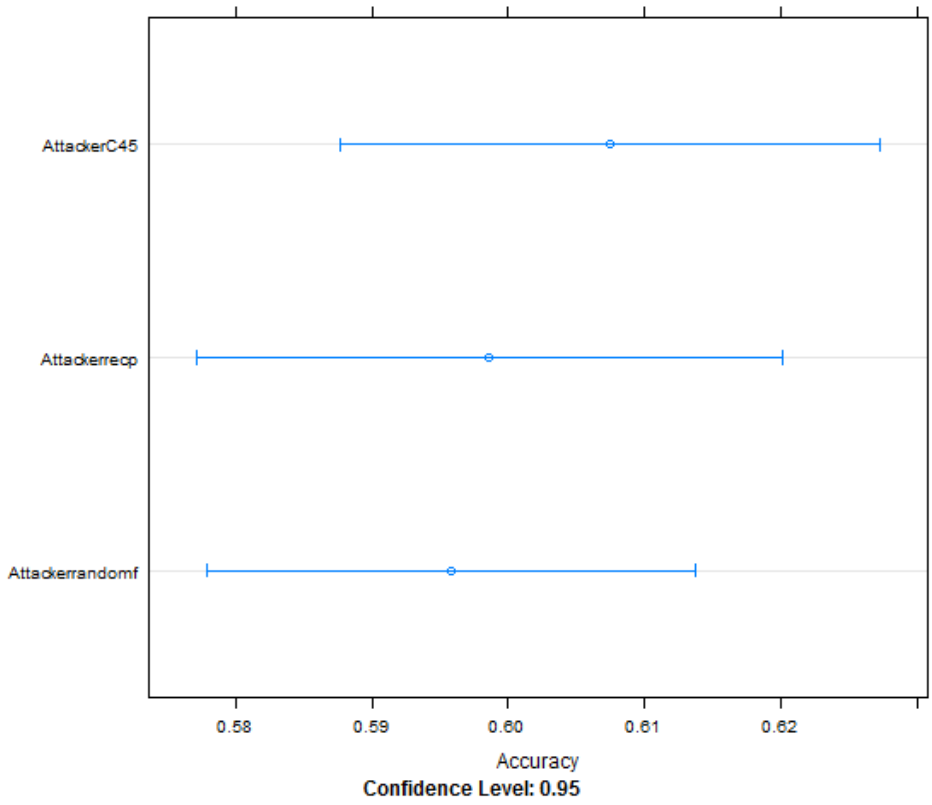

Figure 4-16 Comparing Decision Tree of Prediction of targeted country

According to the plot and Resample process, C4.5 algorithm provides the best accuracy among decision tree algorithms with average accuracy of 60.75% and Recursive partitioning and Random Forest has been placed in next ranks with 59.86% and 59.58 % respectively. Therefore C4.5 will be chosen as the first option for prediction of cyber attackers.

4.2.3 Prediction of Targeted Country by Decision tree

In this part, the models will be built to predict targeted or victim countries against cyber-attacks and all of the 2694 records will be used as a training set to train the model by decision tree algorithms. In order to decrease the level of countries and make the models more accurate and extensive, the name of those countries which they are targeted less than 1 percent of a number of cyber-attacks will be turned into Others. Prediction of the targeted country can help decision makers

and strategist to see which country is more vulnerable to a specific type of cyber-attacks.

In the first stage of the experiment, C4.5 will be used to train the model. 10 fold cross validation is being used as train control parameter and confidence threshold will be changed from 0.1 to 0.5. Figure 4-17 shows the training process of targeted countries' predictive model based on C4.5.

```
Country_J48 <- train(Country ~ Target.Section + Threat + Activity + AuthorID,
data = CCountry, method = "J48",trControl = myCont1, tuneGrid = data.frame(C =
(1:5)/10))


C4.5-like Trees


2694 samples

   6 predictor

  21 classes: 'AU', 'BR', 'CA', 'CN', 'CZ', 'DE', 'FR', 'IL', 'IN', 'INT', 'IT',
'JP', 'KR', 'PH', 'PK', 'RU', 'SA', 'TR', 'UK', 'US', 'Other'


No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 2429, 2422, 2425, 2426, 2421, 2423, ...

Resampling results across tuning parameters:


  C    Accuracy   Kappa

  0.1  0.4770271  0.2163178

  0.2  0.4733545  0.2240196

  0.3  0.4696728  0.2295342

  0.4  0.4729828  0.2401866

  0.5  0.4685245  0.2352127

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was C = 0.1.
```

Figure 4-17 Training process of targeted countries' predictive model based on C4.5

Figure 4-18 shows the plot of changing trend of the accuracy over confidence threshold and the most accurate model has 47.70 % accuracy when confidence level is 0.1. This means a smaller tree with less pruning has more reliability and accuracy.

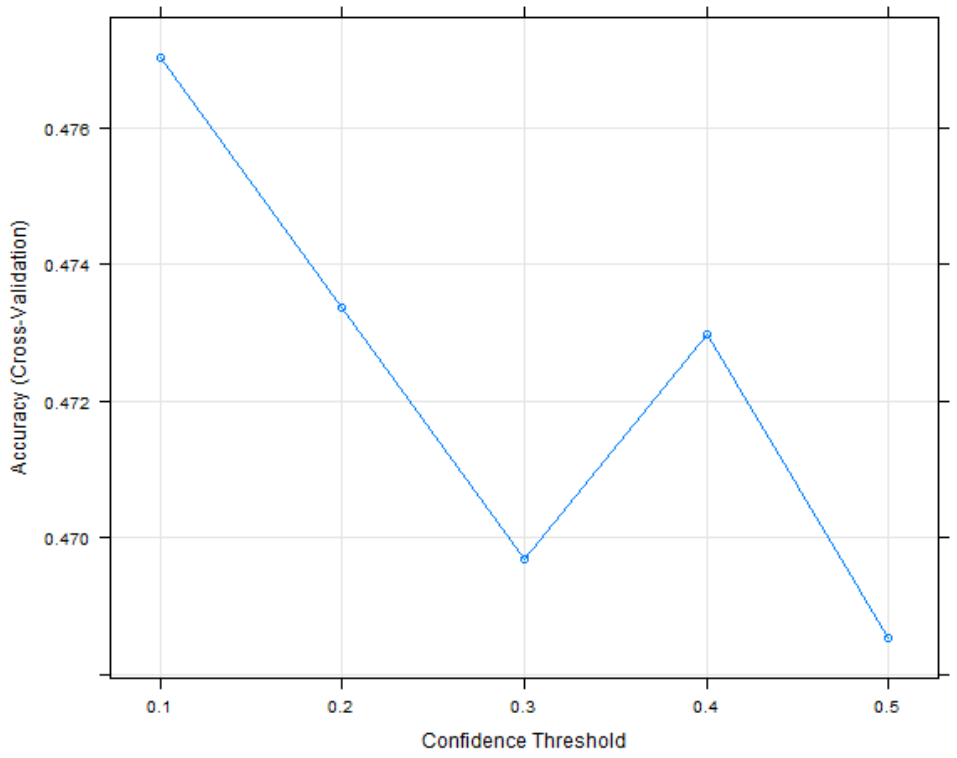

Figure 4-18 Prediction of Targeted Country Accuracy based on Confidence Threshold in C4.5

The second way of training the model is using the Recursive partitioning and the train control will be adjusted to 10 fold cross-validation. The training process of targeted countries' predictive model based on Recursive Partitioning, has been shown in figure 4-19.

```
Country_rpart <- train(Country ~ Target.Section + Threat + Activity + AuthorID,
data = CCountry, method = "rpart",trControl = myCont1)

 CART

2694 samples

   6 predictor

21 classes: 'AU', 'BR', 'CA', 'CN', 'CZ', 'DE', 'FR', 'IL', 'IN', 'INT', 'IT',
'JP', 'KR', 'PH', 'PK', 'RU', 'SA', 'TR', 'UK', 'US', 'Other'

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 2426, 2424, 2426, 2422, 2426, 2425, ...

Resampling results across tuning parameters:


  cp            Accuracy   Kappa

  0.0009609225  0.4629233  0.2023961

  0.0036301516  0.4528843  0.1350404

  0.0057655349  0.4499061  0.1308916


Accuracy was used to select the optimal model using  the largest value.

The final value used for the model was cp = 0.0009609225.
```

Figure 4-19 Training process of targeted countries' predictive model based on RP

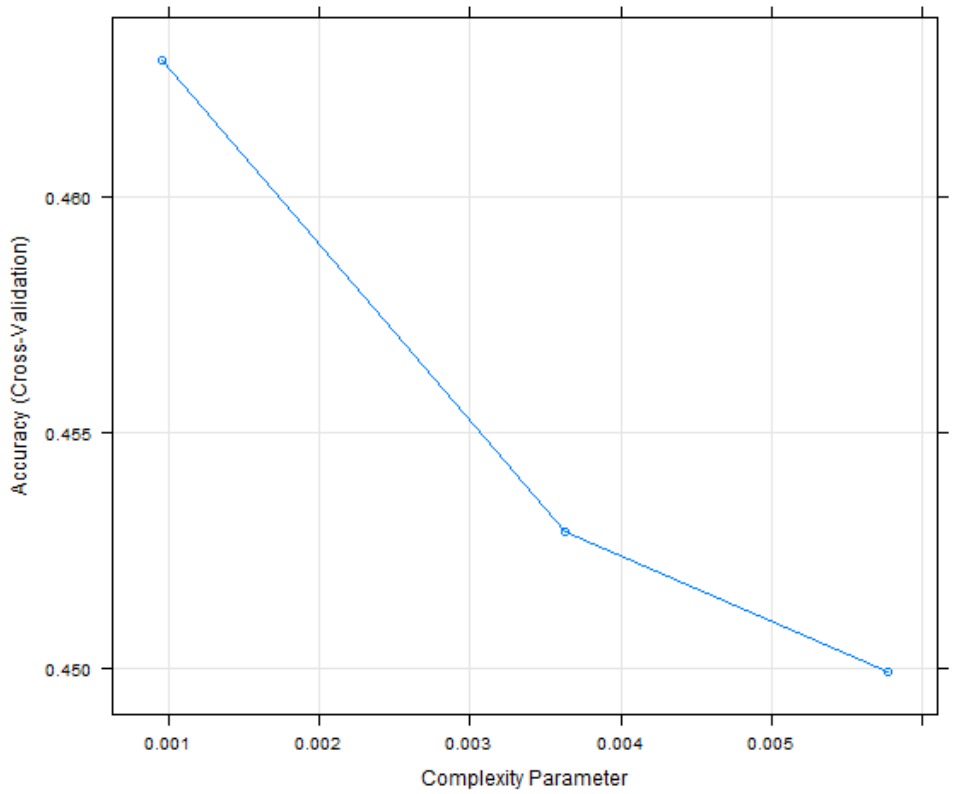

Figure 4-20 Prediction of Targeted country Accuracy vs Complexity Parameter in RP

As it is shown in the codes and figure 4-20, the accuracy is at its maximum level when complexity parameter is at its minimum, which can be interpreted that smaller tree will have better accuracy and larger tree with more nodes can cause less accuracy. Therefore, the Caret process chooses the accuracy of 46.29% with 0.00096 complexity rate.

Random forest is being employed to train the model for prediction of targeted countries and. The scripts shown in figure 4-21 illustrates the process of training the model where 10 fold cross validation is being used for the train control parameter

```
Random Forest


2694 samples

   6 predictor

  21 classes: 'AU', 'BR', 'CA', 'CN', 'CZ', 'DE', 'FR', 'IL', 'IN', 'INT', 'IT',
'JP', 'KR', 'PH', 'PK', 'RU', 'SA', 'TR', 'UK', 'US', 'Other'


No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 2423, 2423, 2422, 2427, 2424, 2424, ...

Resampling results across tuning parameters:


  mtry  Accuracy   Kappa

    2   0.4205784  0.0000000

   32   0.4836506  0.2053243

  533   0.4561586  0.2301219


Accuracy was used to select the optimal model using  the largest value.

The final value used for the model was mtry = 32.
```

Figure 4-21 Training process of Targeted countries' predictive model based on Random Forest

The changing trend of accuracy over randomly selected predictors is shown in figure 4-22. The model is at the most accurate level with 48.36% when 32 predictors are chosen by the process randomly.

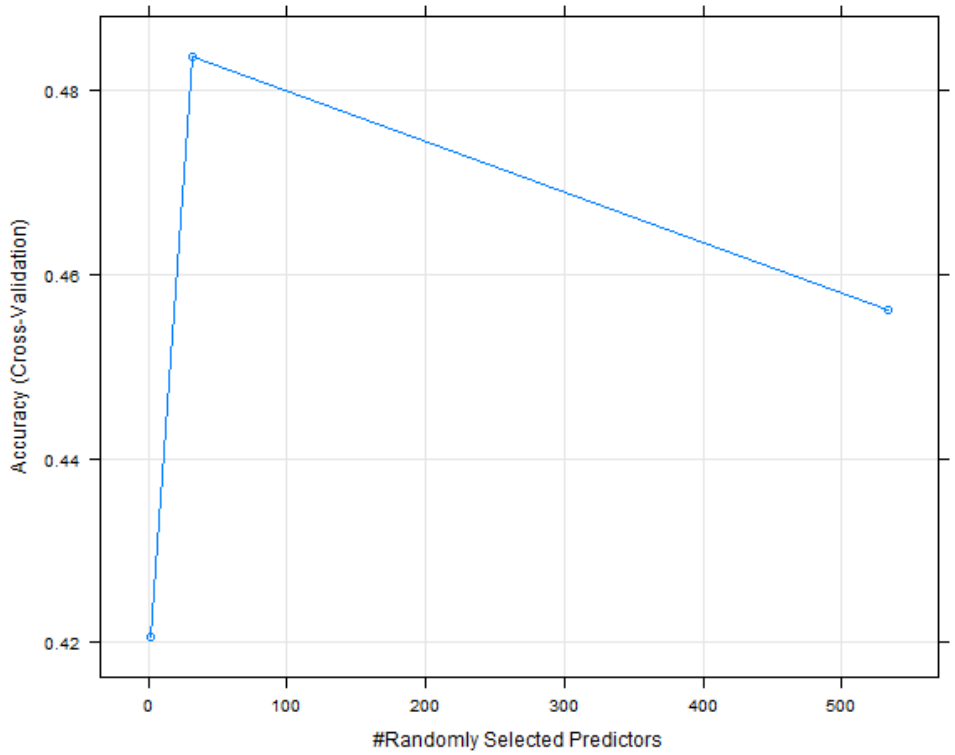

Figure 4-22 Prediction of Targeted country Accuracy vs randomly selected predictors in Random Forest

In the last stage, Resample function needs to be applied to the predictors in order to compare them and chose the most accurate and optimal one. Figure 4-23 shows the comparison process.

```
country_model_list1 <- list(countryc45 = Country_j48, Countryrandomf =
Country_rf, CountryRecP = Country_rpart)

resample_country1 <- resamples(country_model_list1)

summary(resample_country1)

Call:

summary.resamples(object = resample_country1)


Models: countryc45, Countryrandomf, CountryRecP

Number of resamples: 10

Accuracy

                Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's

countryc45     0.4312  0.4713 0.4815 0.4770  0.4865 0.5038    0

Countryrandomf 0.4419  0.4731 0.4879 0.4837  0.4954 0.5093    0

CountryRecP    0.4370  0.4553 0.4584 0.4629  0.4739 0.4907    0
```

Figure 4-23 Comparing process of Decision tree models for prediction of targeted countries

According to the outcome of this function, Random forest does more accurate classification and prediction compared to C4.5 and Recursive Partitioning. The average accuracy of the model obtained by random forest algorithm is 48.37 % and will be chosen as the most optimal decision tree algorithm. Figure 4-24 illustrates the dot plot showing the difference between the models.

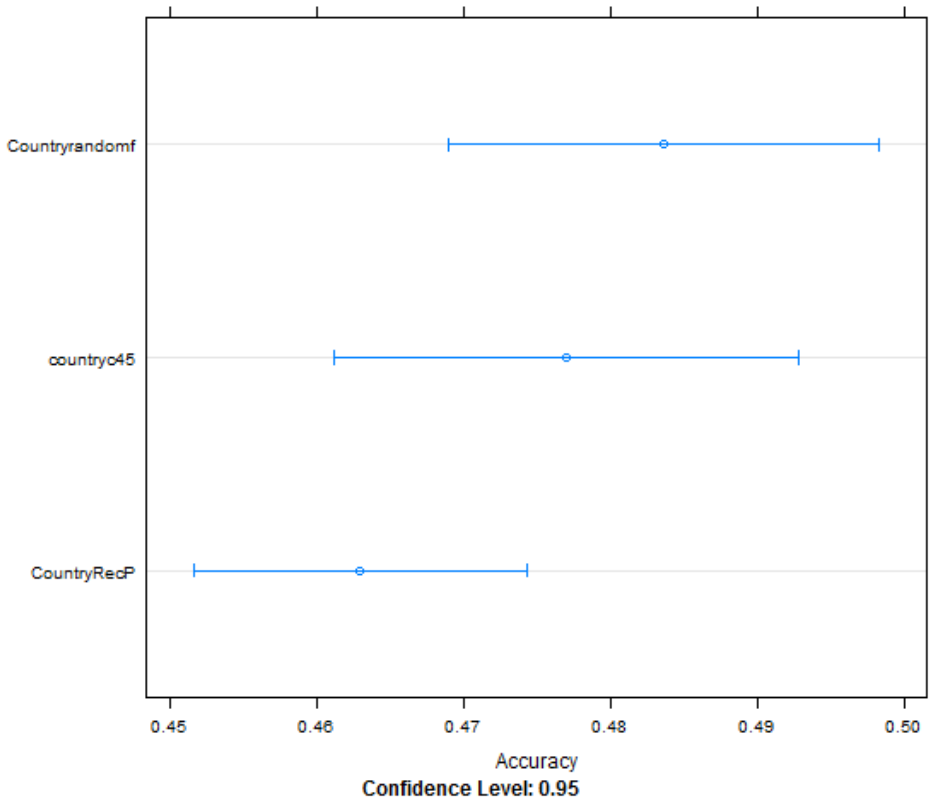

Figure 4-24 Comparison of the decision trees for prediction of targeted countries

Therefore, the most accurate prediction of targeted countries among decision tree algorithms is the model that is trained by Random forest and C4.5 can be the second best option for decision makers with an accuracy level of 47.70. Recursive Partitioning is the last decision tree in terms of accuracy for prediction of targeted countries.

4.2.4 Prediction of Type of Target by Decision tree

In order to make a prediction about the Type of Target of cyber-attacks, decision tree algorithms have been used in this section. The prediction of the Type of Targets will help security experts to find out which section of cyber firms are more vulnerable to different cyber-attacks.

In the first step of the experiment, C4.5 will be used in order to train the model. The train control will be set to 10 fold cross-validation and confidence threshold varied between 0.1 and 0.5. Figure 4-25 shows the training process.

```
Target_J48 <- train(Target.Section ~ Country + Threat + Activity + AuthorID,
data = Ctarget, method = "J48",trControl = myCont1, tuneGrid = data.frame(C =
(1:5)/10))

C4.5-like Trees


2694 samples

   6 predictor

  19 classes: 'BP', 'ED', 'EN', 'ES', 'FB', 'GO', 'HC', 'HT', 'IO', 'MD', 'MU',
'NN', 'RT', 'SI', 'SN', 'TC', 'THS', 'TP', 'Other'


No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 2423, 2425, 2423, 2427, 2426, 2420, ...

Resampling results across tuning parameters:


  C    Accuracy   Kappa

  0.1  0.3894570  0.2775008

  0.2  0.3857753  0.2758925

  0.3  0.3802096  0.2712262

  0.4  0.3750183  0.2667008

  0.5  0.3698176  0.2619773


Accuracy was used to select the optimal model using the largest value.

The final value used for the model was C = 0.1.
```

Figure 4-25 Training process of the Type of Target model based on C4.5

Figure 4-26 plots the changing trend of accuracy based on confidence threshold and as it is shown the process chose the confidence threshold of 0.1 when the accuracy is at its maximum level with 38.94%. This trend shows that if the size of the tree gets larger and it does not get pruned, the result will be less accurate.

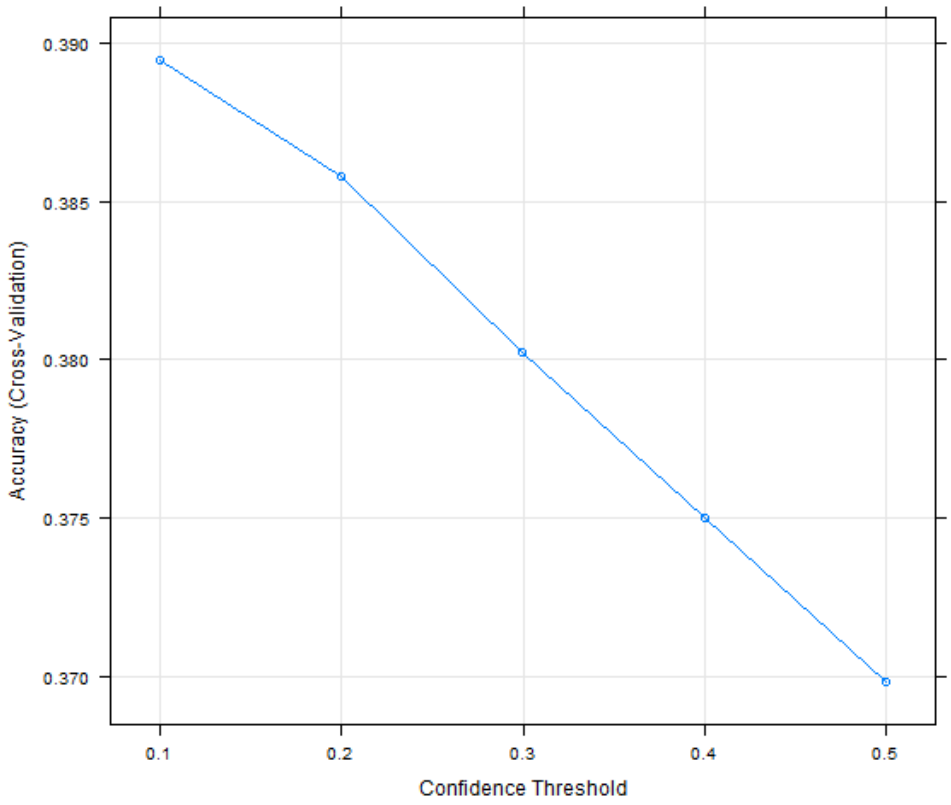

Figure 4-26 Prediction of the Type of Target Accuracy based on Confidence Threshold in C4.5

The second part is to train the model based on Random Forest and in this part again train control will be set to 10 fold cross-validation. The training process is shown in figure 4-27.

```
Target_rf <- train(Target.Section ~ Activity + Threat + Country + AuthorID, data
= Ctarget, method = "rf",trControl = myCont, importance = TRUE)

Random Forest

2694 samples

   6 predictor

  19 classes: 'BP', 'ED', 'EN', 'ES', 'FB', 'GO', 'HC', 'HT', 'IO', 'MD', 'MU',
'NN', 'RT', 'SI', 'SN', 'TC', 'THS', 'TP', 'Other'

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 2423, 2427, 2424, 2427, 2422, 2424, ...

Resampling results across tuning parameters:


  mtry  Accuracy   Kappa

    2   0.2617023  0.0000000

   36   0.3912359  0.2647949

  649   0.3641662  0.2632921


Accuracy was used to select the optimal model using  the largest value.

The final value used for the model was mtry = 36.
```

Figure 4-27 Training process of Type of Target model based on Random Forest

Figure 4-28 demonstrates the plot of changing accuracy over number of selected variable by the process of training and as it is illustrated the accuracy will be maximum with 39.12% accuracy when 36 predictors will be selected randomly.

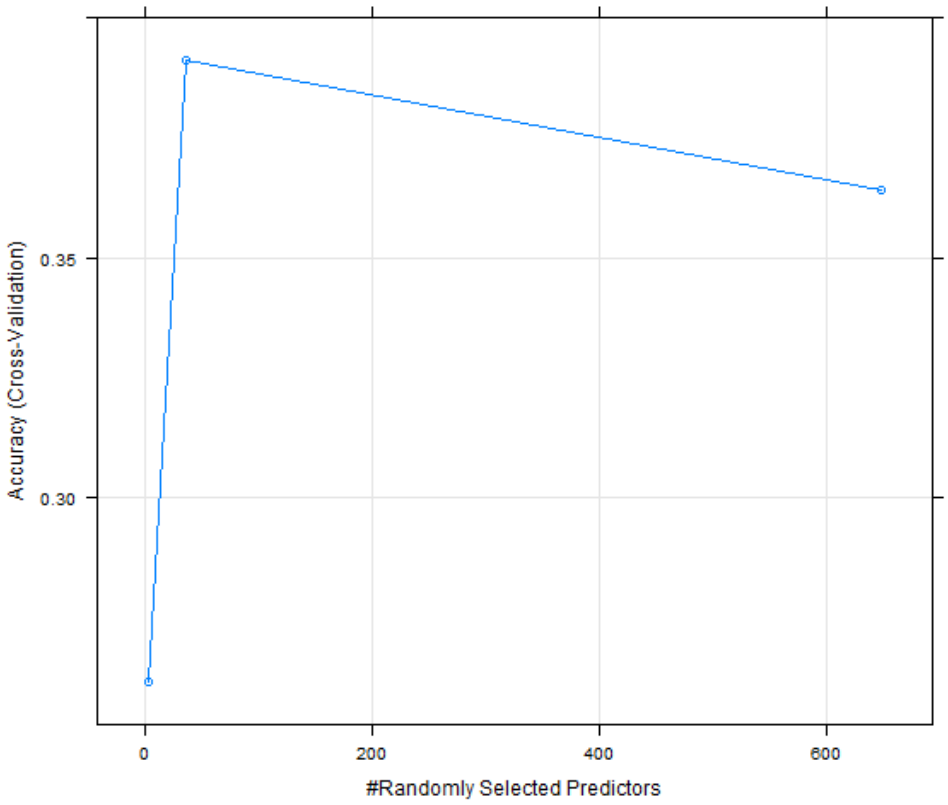

Figure 4-28 Prediction of type of Accuracy vs randomly selected predictors in Random Forest

The last part of the experiment is the making the model by Recursive Partitioning method with 10 fold cross-validation and the training process is shown in figure 4-29.

```
CART

2694 samples

   6 predictor

  19 classes: 'BP', 'ED', 'EN', 'ES', 'FB', 'GO', 'HC', 'HT', 'IO', 'MD', 'MU',
'NN', 'RT', 'SI', 'SN', 'TC', 'THS', 'TP', 'Other'

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 2427, 2425, 2426, 2425, 2425, 2427, ...

Resampling results across tuning parameters:

  cp            Accuracy   Kappa

  0.0002513826  0.3742025  0.2578841

  0.0005027652  0.3753082  0.2576031

  0.0020110608  0.3759801  0.2564984


Accuracy was used to select the optimal model using the largest value.

The final value used for the model was cp = 0.002011061.
```

Figure 4-29 Training process of Type of Target model based RP

The codes represent that the accuracy level is constant over the changing of complexity parameter and the optimal and maximum accuracy is 37.59% when Recursive Partitioning is being used as a method for training the model. Figure 4-30 illustrates the trend of accuracy based on complexity parameter and it describes that they have a direct relationship which means the accuracy reaches its maximum level when complexity parameter is maximum. This is interpreted that despite the C4.5 tree in the prediction of the Type of Target if the tree gets more pruned the result will be less accurate so there is a need for more details for prediction of the Type of Target.

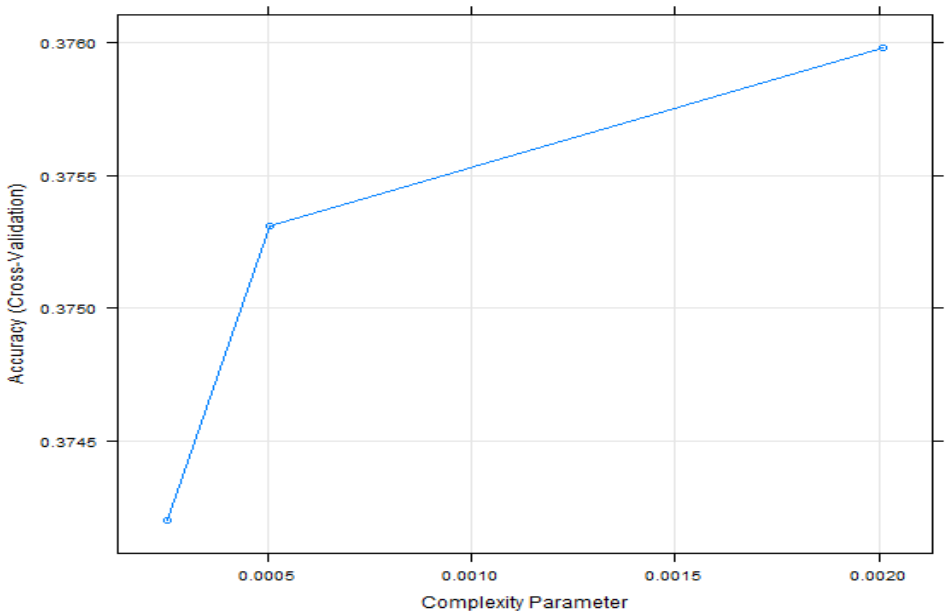

Figure 4-30 Prediction of Type of Target- Accuracy vs Complexity Parameter in RP

The last stage of the experiment is comparing the models with Resample function to find out which algorithm performs better and more accurate in order to make prediction for target section. Figure 4-31 demonstrates the comparison process by Resample function

```
target_model_list1 <- list(Targetc45 = Target_j48, Targetrandomf = Target_rf,
TargetRecp = Target_rpart)

resample_target1 <- resamples(target_model_list1)

> summary(resample_target1)

Call:

summary.resamples(object = resample_target1)

Models: Targetc45, Targetrandomf, TargetRecp

Number of resamples: 10

Accuracy

               Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's

Targetc45     0.3579  0.3728 0.3816 0.3895  0.4021 0.4440    0

Targetrandomf 0.3670  0.3819 0.3919 0.3912  0.4024 0.4154    0

TargetRecp    0.3408  0.3624 0.3752 0.3760  0.3856 0.4164    0
```

Figure 4-31 Comparing Decision tree models for prediction of Type of Target

Figure 4-32 shows the plot of this comparison and Random forest does the better task in terms of accuracy compared to C4.5 and Recursive Partitioning. C4.5 will be placed as the second best option and RP is the third best option when it comes to prediction of Type of Target.

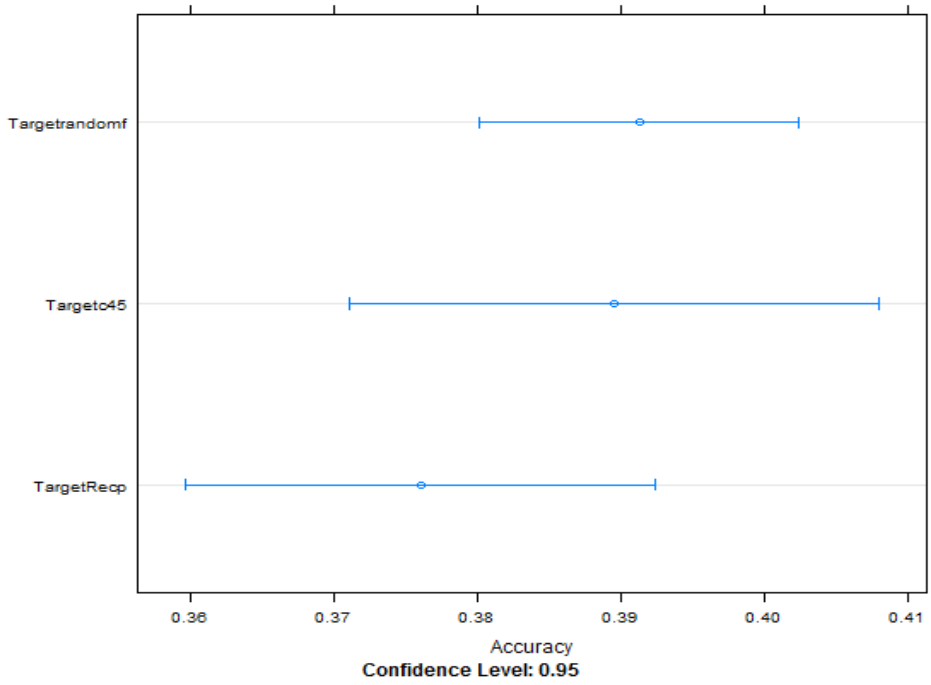

Figure 4-32 Comparing Decision Tree of Prediction of Type of Target

4.2.5 Prediction of Cyber Attack Activity by Decision tree

In the dataset, there are 2694 records and all of the attacks are known in terms of the type of activity where the intention of them was obvious. Prediction of type of activity of cyber-attacks can give the ability of security expert to figure out the intention of future cyber-attacks when they are unknown and ambiguous and profile cyber criminals for law enforcement agencies.

In the first step of the experiment, C4.5 will be used with 10 fold cross-validation and variation of confidence level from 0.1 to 0.5 and figure 4-33 demonstrates the training process.

```
C4.5-like Trees

2694 samples
   6 predictor
   4 classes: 'CC', 'CE', 'CW', 'HA'
No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2424, 2425, 2424, 2424, 2426, 2426, ...
Resampling results across tuning parameters:
  C    Accuracy   Kappa
  0.1  0.8024937  0.6386700
  0.2  0.8054788  0.6463893
  0.3  0.8013785  0.6391037
  0.4  0.8010012  0.6381488
  0.5  0.7950573  0.6272107

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was C = 0.2.
```

Figure 4-33 Training process of cyber-attack activity model based on C 4.5

The plot of the changing the accuracy level based on confidence threshold has been showed in figure 4-34 and the accuracy will get to the maximum level with 80.54 % when the confidence level is 0.2.

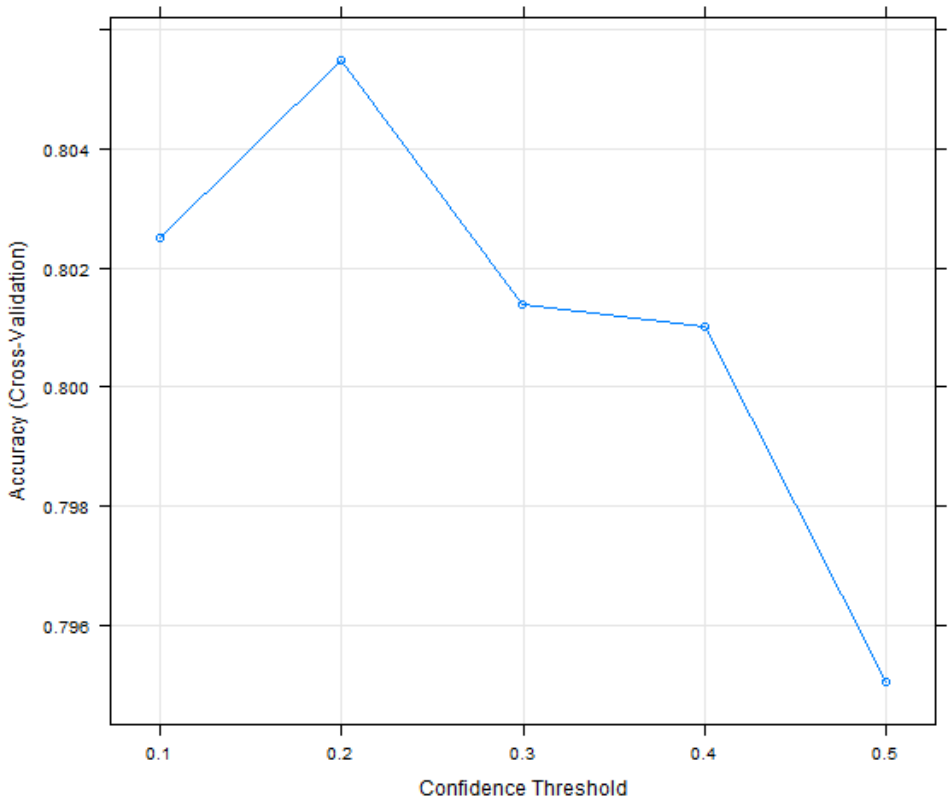

Figure 4-34 Prediction of type of activity- Accuracy vs Confidence Threshold in C4.5

In the second stage, Recursive Partitioning will train the model and the train control will be adjusted to 10 fold cross-validation. Figure 4-35 demonstrates the training process.

```
Act_rpart <- train(Activity ~ Country + Threat + Target.Section + AuthorID, data
= CT2, method = "rpart",trControl = myCont1)

> Act_rp

CART

2694 samples

   6 predictor

   4 classes: 'CC', 'CE', 'CW', 'HA'

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 2425, 2425, 2425, 2426, 2425, 2425, ...

Resampling results across tuning parameters:

  cp            Accuracy   Kappa

  0.0008503401  0.8065922  0.6448783

  0.0048185941  0.8043685  0.6407419

  0.0408163265  0.7453537  0.5314421


Accuracy was used to select the optimal model using  the largest value.

The final value used for the model was cp = 0.0008503401.
```

Figure 4-35 Training process of cyber-attack activity model based on RP

Figure 4-36 shows the process of training and trend of accuracy based on complexity parameter in Recursive partitioning.

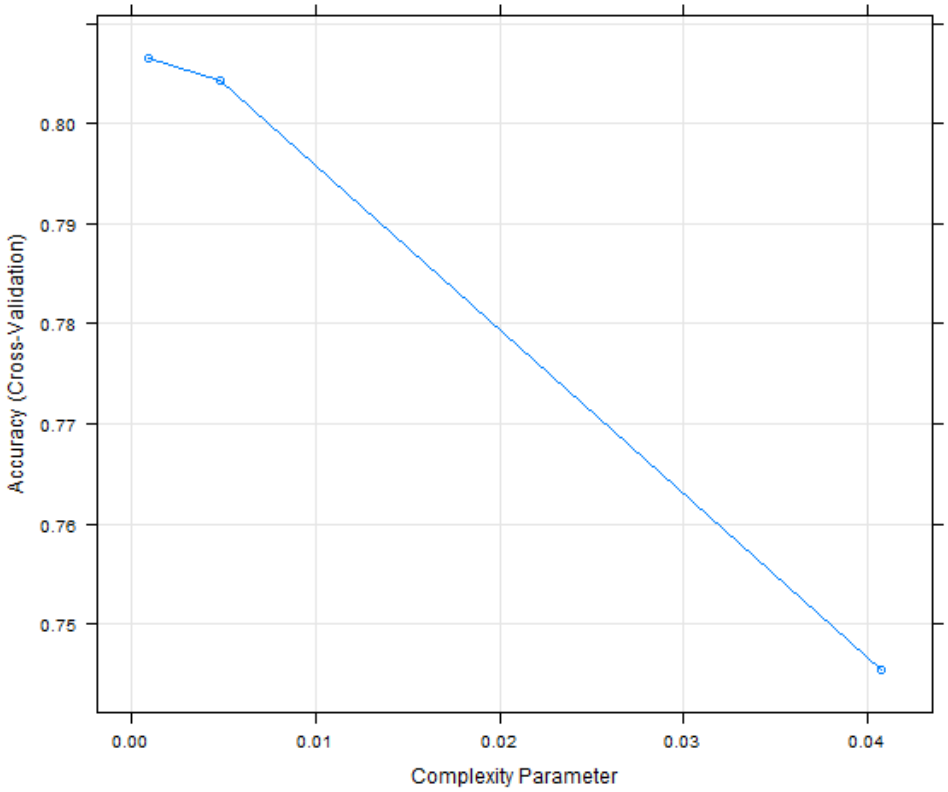

Figure 4-36 Prediction of activity- Accuracy vs Complexity Parameter in RP

As the codes and figure 4-36 are presenting, the accuracy and complexity have the opposite effect on each other and the maximum level accuracy is 80.65% when complexity parameter is at its minimum level of 0.0085.

The third part of the experiment is the making the model based on Random Forest. Train control will be set to 10 fold cross-validation. The training process has been shown in figure 4-37.

```
Act_rf <- train(Activity ~ Country + Threat + Target.Section + Author, data =
CT1, method = "rf",trControl = myCont, importance = TRUE)

Random Forest


2694 samples

   5 predictor

   4 classes: 'CC', 'CE', 'CW', 'HA'


No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 2424, 2425, 2426, 2424, 2424, 2425, ...

Resampling results across tuning parameters:


  mtry  Accuracy   Kappa

    2   0.5634769  0.000000

   36   0.8248042  0.679942

  665   0.8162830  0.668461


Accuracy was used to select the optimal model using the largest value.

The final value used for the model was mtry = 36.
```

Figure 4-37 Training process of cyber-attack activity model based on Random Forest

Figure 4-38 presents the change of accuracy based on the number of predictors which were selected randomly by the random forest algorithm. The process chose 82.48% accuracy when 36 predictors are selected randomly by the training process.

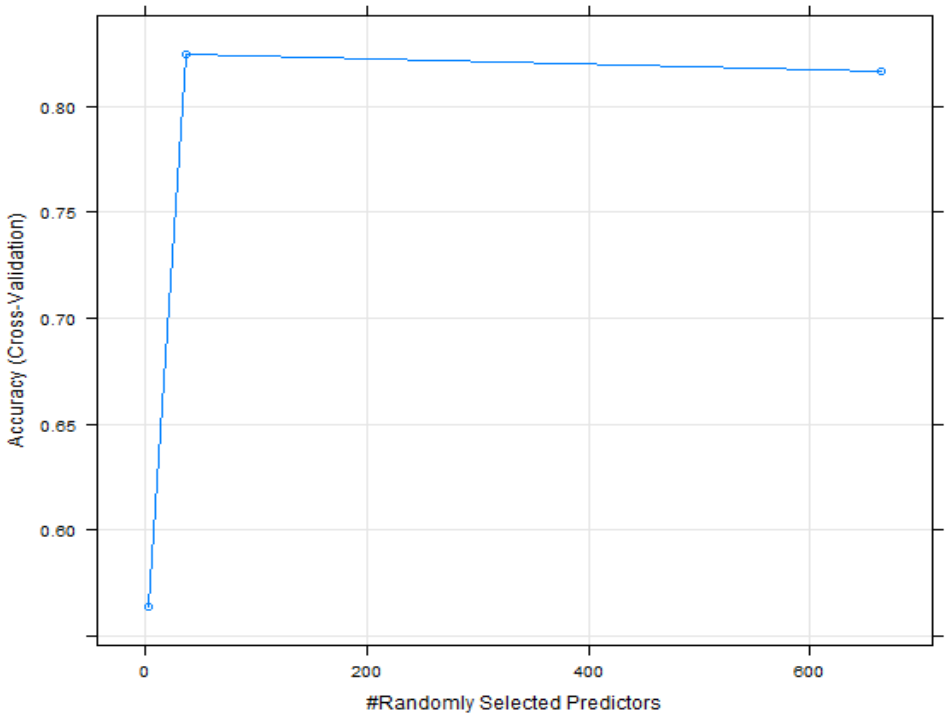

Figure 4-38 Prediction of type of activity- Accuracy vs randomly selected predictors in Random Forest

The last part of the experiment is comparing the models obtained by decision tree algorithms by using Resample function in order to determine which algorithm is more reliable and accurate and figure 4-39 explains the comparison process.

```
Call:

summary.resamples(object = resample_activity1)


Models: ActivityC45, Activityrandomf, ActivityRecp

Number of resamples: 10


Accuracy

                  Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's

ActivityC45     0.7724  0.7859 0.8000 0.8055  0.8306 0.8433    0

Activityrandomf 0.7844  0.8130 0.8249 0.8248  0.8405 0.8513    0

ActivityRecp    0.7704  0.7955 0.8141 0.8111  0.8250 0.8476    0
```

Figure 4-39 Comparison process of Cyber Attack Activity models by Resample function

Figure 4-40 also shows the comparison of the 3 different models for prediction of type of activity. Random Forest method is the most reliable and accurate model among the decision tree methods for prediction of type of activity with 82.48% average accuracy and RP with a minor difference is placed in the second place and C4.5 is in the third place with 80.55% accuracy.

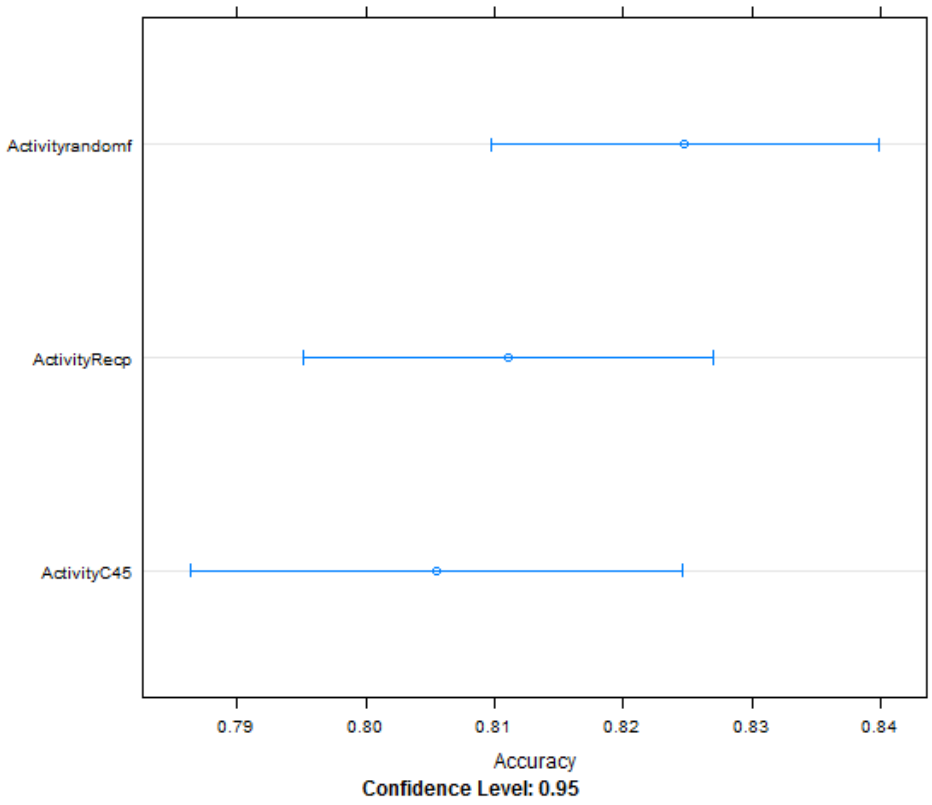

Figure 4-40 Comparison of the decision trees for prediction of activity

4.2.6 Discussion and Interpretation

According to analysis using the main decision tree techniques table 4-1 shows the obtained result:

Prediction	Optimal decision tree	Accuracy
Cyber Attack Activity	Random Forest	82.48%
Cyber Attacker	C4.5	60.75%
Targeted Country	Random Forest	48.37%
Type of Target	Random Forest	39.12%
Type of Threat	C4.5	59.60%

Table 4-1 Optimal Prediction by Decision tree

1- In terms of prediction of activity of cyber-attacks, Random forest has been chosen as the most accurate among the main decision tree algorithms and there is 82.48% reliability in the model to predict the future type of activities, however, other algorithms have a very slim difference with the random forest.
2- Prediction of cyber attackers with the decision tree algorithms, has been more successful and precise with C4.5 and it is 60.75%, however, that might not be completely suitable for cyber experts but it can be used as supportive method along with other methods and tools. It should be mentioned that cyber attackers' profiling can be based on other factors such as the origin of the attack which does not exist in the dataset obtained in this research and according to the limitation of this research OSINT has been used as the main resource of the data and there were no more details on cyber-attacks.

3- Targeted country has more accurate prediction with C4.5 compared to other decision tree techniques. Cyber experts can be 48.37% sure in terms of prediction of vulnerable and targeted countries. Although 48.37% cannot be a reliable accuracy, this method can be employed with other security techniques in order to protect countries against different type of attacks. The accuracy level might be improved if there will be more details on cyber-attacks such as the origin of the attack and technical details.

4- Type of Target: Random forest technique produces a more accurate model compared with C4.5 and RP in the prediction of the Type of Target, however, 39.12% is not the accuracy level

that security experts can rely on. If there were more details, the accuracy level might be improved, as figure 15 shows if the tree gets larger and has more nodes, increasing accuracy level will be highly likely. It should be also noticed that according to the experiments, prediction of the Type of Target by decision trees has been the least accurate among the predictions in this project which can be interpreted that decision trees do not provide high accuracy level for prediction of Type of Target.

5- Type of Threat: Prediction of Type of Threat in cyber-attacks has been more accurate with C4.5 and the accuracy is 59.6% showing that security experts can be used this model as a supportive tool along with other approaches. The accuracy might be improved if there were more details on cyber-attacks such as technical details and etc as Type of Threat can be dependent on many details.

According to the experiments in this chapter, the decision trees has the average accuracy of 58.06 in predictions in this research. In the next chapter, K-nearest Neighbour will be discussed in order to investigate in term of reliably how they perform in predictions.

4.3 K nearest Neighbour Analysis

K nearest neighbour algorithm has been discussed in section 3.2.1.4 and in this section K nearest neighbour will be applied to the training set in order to train a classifier for prediction of different aspects of cyber-attacks. Class is the package in R which provides the library including functions for implementation of KNN. This package which was developed by Ripley and Vanables (2015) and it will be used in combination with Caret package in order to achieve a better and extensive overview of the training process. K is the parameter which will be used in the training process and it represents the number of the nearest neighbours.

4.3.1 Prediction of Type of Threat by KNN

This stage aims to train a KNN classifier for prediction of Type of Threat. The training set includes 2210 cyber-attacks with a known Type of Threat. For training control, 10 fold cross validation will be applied

and K varies from 5 to 11. Figure 4-41 describes the training process of the KNN model for prediction of Type of Threat.

```
k-Nearest Neighbors
Threat_knn <- train(Threat ~ Target.Section + Author + Activity + Country,data
 = Dthreat, method = "knn", tuneLength = 4,trControl = myc1,metric = "Accuracy
")
2210 samples
   5 predictor
  11 classes: 'AH', 'CSS', 'DF', 'DH', 'DS', 'MWV', 'SQ', 'TA', 'UA', 'ZD', 'O
ther'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1769, 1765, 1771, 1768, 1767
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   5  0.5515434  0.4623274
   7  0.5515465  0.4624442
   9  0.5442861  0.4540032
  11  0.5447673  0.4551347

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was k = 7.
```

Figure 4-41 Training process of Type of Threat model based on KNN

As figure 4-41 shows, when K is 7, the accuracy of the model reaches the maximum level; 55.15% so it will be chosen as the most reliable model. Figure 4-42 demonstrates the changing trend of accuracy over K and it can be concluded that there is no robust relationship between the value of the K and the accuracy, however, after K gets the value of 9 it can be seen the accuracy will be decreased.

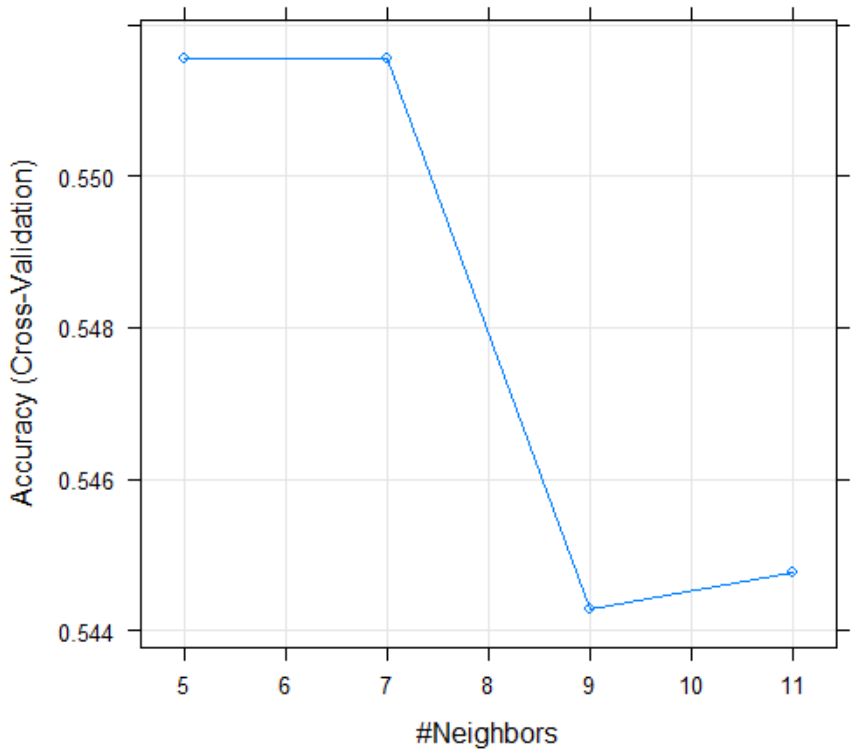

Figure 4-42 Accuracy trend for prediction of type of cyber threat in KNN

4.3.2 Prediction of Cyber attackers by KNN

At this step, KNN will be used to train a model for prediction of potential cyber attackers behind cyber breaches. Among all cyber attacks in the dataset obtained from OSINT, 1432 cyber-attacks were known in terms of perpetrators. Therefore, they will be used training set. 10 fold cross validation will be applied as training control and the amount of K will be set from 5 to 11 and Figure 4-43 demonstrates the training process.

```
Attacker_1knn <- train(Author ~ Target.Section + Country + Activity + Threat,d
ata = DAttacker2, method = "knn", tuneLength = 4,trControl = myc1,metric = "Ac
curacy")

k-Nearest Neighbors

1432 samples
   5 predictor
  33 classes: 'Ag3nt47', 'AnonGhost', 'Anonymous', 'Armada.Collective', 'Chine
se.hacker', 'Chinese.hackers', 'Cyber.Islamic.State', 'CyberBerkut', 'DarkWeb.
Goons', 'DERP', 'Dr.SHA6H', 'Guccifer', 'HAXOR', 'Iranian.Hackers', 'Izz.ad.Di
n.al.Qassam.Cyber.Fighters', 'JokerCracker', 'KelvinSecTeam', 'LizardSquad', '
LulzSec', 'Maxney', 'NetPirates', 'NullCrew', 'Other', 'RedHack', 'Rex.Mundi',
 'Syrian.Electronic.Army', 'TEAM.MADLEETS', 'TeamBerserk', 'Tunisian.Cyber.Arm
y', 'Turkish.Ajan', 'X.smitt3nz', 'X.th3inf1d3l', 'XTnR3vOlT'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1290, 1289, 1292, 1290, 1286, 1287, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   5  0.6029454  0.4262068
   7  0.6071261  0.4264062
   9  0.6078348  0.4264133
  11  0.6127445  0.4327187

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was k = 11.
```

Figure 4-43 Training process of Cyber Attacker model based on KNN

As it is shown in figure 4-43, when k accepts 11, the accuracy will be 61.27% which is the maximum level in the training process. Thus 11 nearest neighbour classifier will be identified as the best predictive model. Figure 4-44 shows the trend of accuracy over k and there is a direct relationship between them as when K gets higher the accuracy will be higher as well and the model will become more accurate.

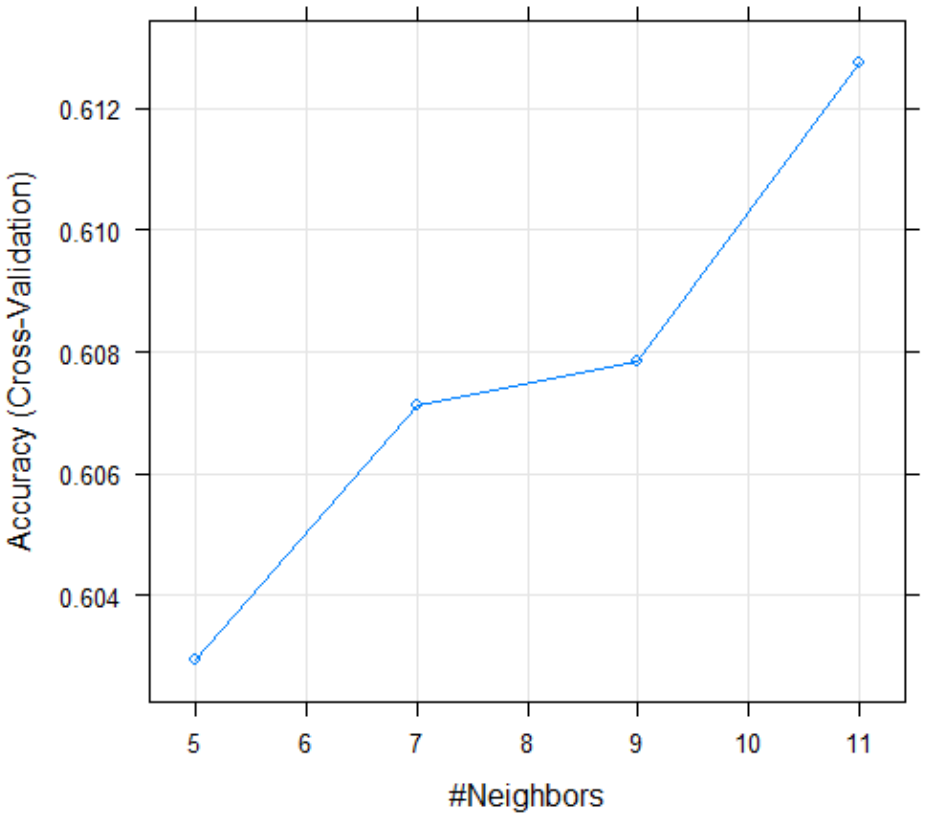

Figure 4-44 Accuracy trend for prediction of cyber attackers in KNN

4.3.3 Prediction of Type of Target by KNN

This part of the experiment demonstrates the training process of a KNN classifier in order to identify and predict vulnerable targets against cyber attacks. 2694 samples exist in the cyberattack dataset obtained from OSINT and the targets are categorized which was described in section 4.4. The training control will be defined as 10 fold cross-validation and K as tuning parameter will be changeable from 5 to 11. The training process is shown in figure 4-45.

```
k-Nearest Neighbors

2694 samples
   5 predictor
  19 classes: 'BP', 'ED', 'EN', 'ES', 'FB', 'GO', 'HC', 'HT', 'IO', 'MD', 'MU'
, 'NN', 'RT', 'SI', 'SN', 'TC', 'THS', 'TP', 'Other'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2155, 2156, 2154, 2153, 2158
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   5  0.3775008  0.2689595
   7  0.3823253  0.2717591
   9  0.3819618  0.2704042
  11  0.3752786  0.2620751

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was k = 7.
```

Figure 4-45 Training process of Type of Target model based on KNN

The training process indicates that highest accuracy in prediction of Type of Target occurs in 7 nearest neighbour classifier with 38.22% reliability. Figure 4-46 demonstrates the training process and there is an abnormal behaviour of accuracy over K. The minimum accuracy in the training process is 37.52% when K is equal to 11.

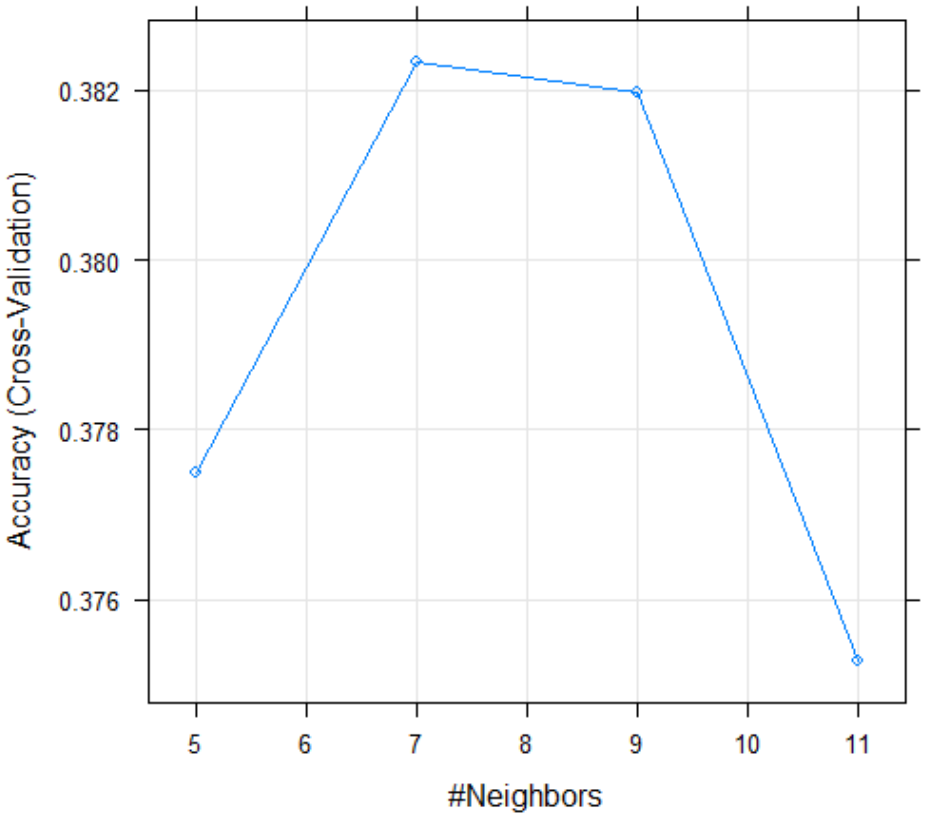

Figure 4-46 Accuracy trend for prediction of Type of Target model in KNN

4.3.4 Prediction of Targeted Country by KNN

This stage of analysis aims to develop a predictive model to identify and forecast vulnerable countries against different cyber attacks based on KNN algorithm. K will be changeable from 5 to 11 as tuning control and 10 fold cross validation is being used as training control for applying to 2694 cyber attacks in the obtained training dataset. The process is described in figure 4-47.

```
k-Nearest Neighbors

2694 samples
   5 predictor
  21 classes: 'AU', 'BR', 'CA', 'CN', 'CZ', 'DE', 'FR', 'IL', 'IN', 'INT', 'IT
', 'JP', 'KR', 'PH', 'PK', 'RU', 'SA', 'TR', 'UK', 'US', 'Other'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2427, 2422, 2423, 2427, 2424, 2425, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   5  0.4710325  0.2135890
   7  0.4706830  0.2056459
   9  0.4661735  0.1943039
  11  0.4709944  0.1961512

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was k = 5.
```

Figure 4-47 Training process of Targeted Country model based on KNN

As the codes are shown in figure 4-47 and figure 4-48 illustrates, the optimal model is 5 nearest neighbour classifier and the accuracy is 47.10% and when k gets the value of 9 the least accuracy happens with 46.61% reliability. The accuracy has a decreasing trend until the k is equal to 9 and then after that, it follows an increasing path.

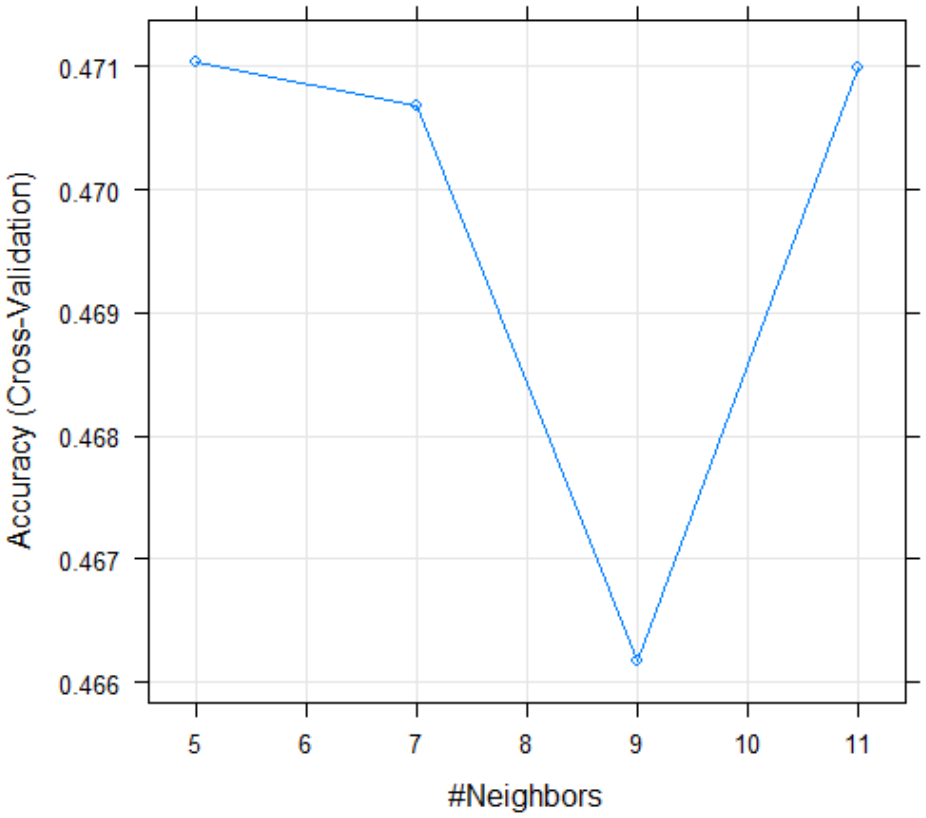

Figure 4-48 Accuracy trend for prediction of targeted country in KNN

4.3.5 Prediction of Cyber Attack Activity by KNN

In order to build a classifier to predict the type of cyber-attack activity, KNN algorithm will be applied to the obtained dataset. In the dataset, cyberattacks have been categorized based on their motivation of them which extensively described in section.

The training control has been set as 10 fold cross-validation and K is set from 50 to 11 as tuning parameter, the code shown in figure 4-49 explains the training process.

```
k-Nearest Neighbors

2694 samples
   5 predictor
   4 classes: 'CC', 'CE', 'CW', 'HA'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2425, 2425, 2424, 2424, 2425, 2425, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   5  0.8084827  0.6502840
   7  0.8044045  0.6424817
   9  0.8088545  0.6473233
  11  0.8070026  0.6416226
Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was k = 9.
```

Figure 4-49 Training process of Cyber Attack Activity model based on KNN

Figure 4-49 and Figure 4-50 indicate that 9 nearest neighbour has been the most reliable model with 80.88% accuracy and also the accuracy does not have a constant trend based on K meaning that there is no interaction and relationship between them.

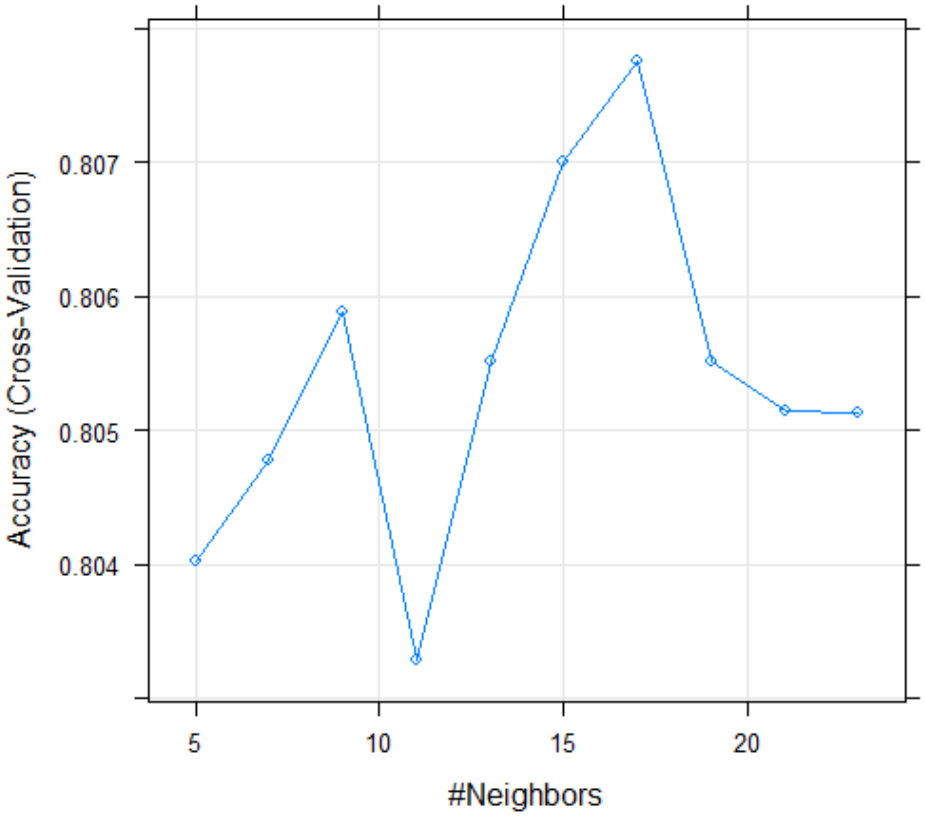

Figure 4-50 Accuracy trend for prediction of type of cyber-attack activity in KNN

4.3.6 Discussion and Interpretation

As it was investigated in the last section, the optimal models by KNN have been achieved and table 4-2 demonstrates them.

Type of prediction	Accuracy level
Type of Threat	55.15%
Cyber attackers	61.27%
Targeted Country	47.10%
Type of Target	38.23%
Type of activity	80.88%

Table 4-2 KNN models' accuracy

Regarding the training process and the accuracy of these KNN models following points can be concluded as interpretations:

1. KNN models for prediction of Type of Target and the countries that they are located in, have performed weak and the accuracy of the is less than average, however, they can be tools for combining with other tools and methods for better and accurate prediction of vulnerable targets and countries.

2. KNN predictive models for Type of Threats and cyber attackers have done a reasonable job with an accuracy of 55.15% and 61.27% respectively. These models can be even more accurate when they mix with other techniques and methods. As it was seen in the training process, the number of neighbours has an effect on the accuracy of both models which in the Type of Threat prediction the number of neighbours has a negative impact and in cyber attackers identification has a positive effect on the accuracy of the models.

3. The accuracy of the KNN predictive model for cyber-attack activity is significantly high and can be a standalone and independent tool for prediction purposes. Cyber experts can be

80.88% confident when they apply 9 nearest neighbour model to their data set for prediction of cyber-attacks activities.

4.4 Naïve Bayes analysis

Naïve Bayes is a classification algorithm which was explained in section 3.2.1.2 and this Stage of analysis will discuss and investigate the usage of Naïve Bayes algorithm in order to build a classifier for prediction of different aspects of cyber-attacks.

Implementation of Naïve Bayes method can be carried out with different packages such as naive Bayes by Majka (2017). This package will be used along with Caret Package in order to obtain an extensive view over the training process and choose the best model. There are features in caret package giving the tuning offer when the models are trained by Naïve Bayes classifier and they are as follows:

1- Laplace Correction: Laplace correction is the solution of naïve Bayes classifier to deal with zero probabilities values. In Conditions that in test set there are some classes that they do not exist in training set the model will give zero probability to that class, however, by applying Laplace correction (estimation) this problem can be solved.

2- Kernel: This feature mainly used for measuring the density of predictors in continues or numeric values. Therefore, because the data set in this project is nominal and categorical, it is not necessary to apply this feature to the training process.

4.4.1 Prediction of Type of Threat by Naïve Bayes Classifier

In this stage, 484 cyber-attacks will be removed from the data set because the Type of Threat was unknown so the rest of cyber-attacks contribute as the training set. In addition, like the previous chapter in order decrease the level of instances, it has been decided to convert those threats, which happened less than 9 times from 2013 to 2015 to “other”. The training control is adjusted to 10 fold cross-validation and

Laplace correction is set from1 to 8 as tuning parameter. Figure 4-51 demonstrates codes generating the model and its construction process.

```
gridtest1                <-                expand.grid(fL=c(0,1,2,3,4,5,6,7,8),
adjust=c(1) ,usekernel=c(FALSE,TRUE))

Threat_nbt1 <- train(Thrx, Thry,data = Dthreat, method = "nb", trControl =
myc1,metric = "Accuracy",tuneGrid = gridtest1)

Naive Bayes

2210 samples

4 predictor

  11 classes: 'AH', 'CSS', 'DF', 'DH', 'DS', 'MWV', 'SQ', 'TA', 'UA', 'ZD',
'Other'

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 1990, 1988, 1988, 1989, 1989, 1987, ...

Resampling results across tuning parameters:

 fL  usekernel  Accuracy   Kappa

  0   FALSE      0.5588208  0.4748552

  0    TRUE      0.5588208  0.4748552

  1   FALSE      0.5547381  0.4654439

  1    TRUE      0.5547381  0.4654439

  2   FALSE      0.5307230  0.4305766

  2    TRUE      0.5307230  0.4305766

  3   FALSE      0.5284707  0.4264311

  3    TRUE      0.5284707  0.4264311

  4   FALSE      0.5248507  0.4211414

  4    TRUE      0.5248507  0.4211414

  5   FALSE      0.5203257  0.4148365

  5    TRUE      0.5203257  0.4148365

  6   FALSE      0.5198753  0.4135546

  6    TRUE      0.5198753  0.4135546

  7   FALSE      0.5176108  0.4102733

  7    TRUE      0.5176108  0.4102733

  8   FALSE      0.5166997  0.4084220

  8    TRUE      0.5166997  0.4084220

Tuning parameter 'adjust' was held constant at a value of 1
Accuracy was used to select the optimal model using the largest value.

The final values used for the model were fL = 0, usekernel = FALSE and adjust
= 1.
```

Figure 4-51 Training process of Type of Threat model based on NB

As it is described in the process, the optimal is chosen when Laplace correction is 0 and kernel value gets false value. Figure 4-52 shows the changing trend of accuracy based on Laplace correction.

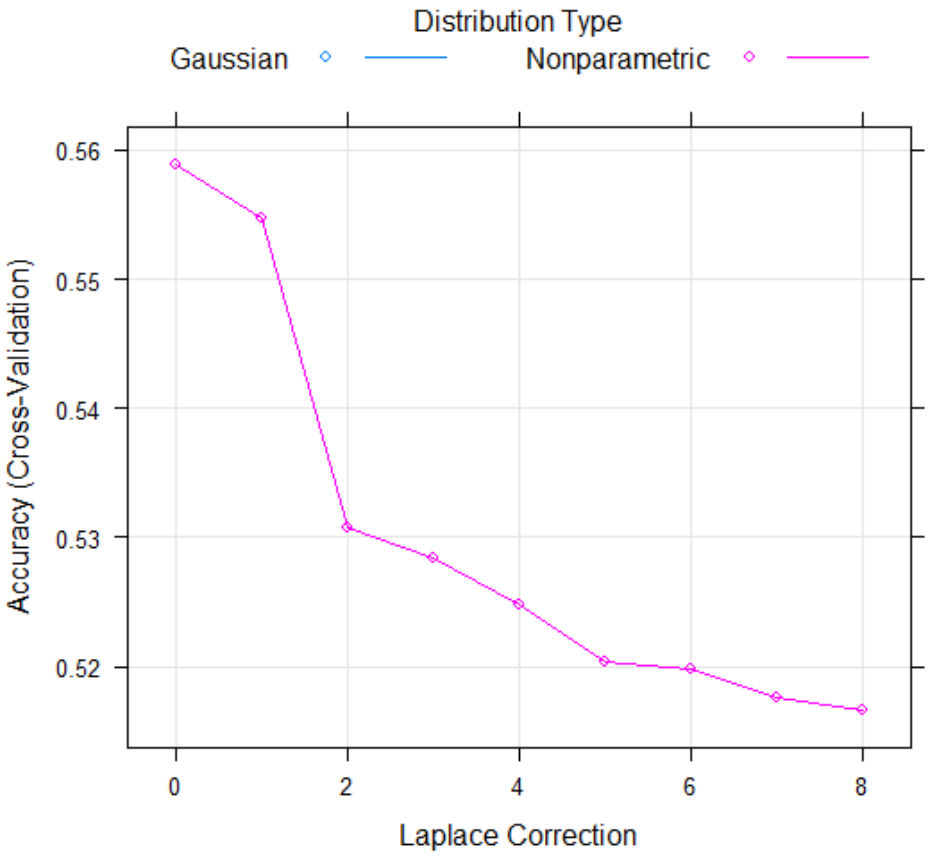

Figure 4-52 Accuracy trend for prediction of type of cyber threat in NB

As the plot shows the maximum level of accuracy happens when Laplace correction is 0 and then after that, the accuracy has a downward trend and it reaches its minimum when Laplace is 8. Optimal model with 55.88% accuracy has been chosen for prediction of type of cyber threat by Naïve Bayes.

4.4.2 Prediction of Cyber attackers by Naïve Bayes Classifier

For making the classifier for prediction of Cyber attackers, those attacks where attackers were known are chosen as training dataset so 1432 cyber-attacks remain. In addition, those less active attackers who committed cyber-attacks less than 5 times will be converted to "others".

Laplace correction will be adjusted from 0 to 8 as tuning grid and also 10 fold cross-validation is set as training control. Figure 4-53 shows the scripts generating the training process of the desired classifier.

```
gridtest1                <-                expand.grid(fL=c(0,1,2,3,4,5,6,7,8),
adjust=c(1) ,usekernel=c(FALSE,TRUE))

Attacker_nbt1 <- train(Attx, Atty,data = DAttacker2, method = "nb", trControl =
myc1,metric = "Accuracy", tuneGrid = gridtest1)

Naive Bayes

1432 samples

   4 predictor

  33  classes:  'Ag3nt47',  'AnonGhost',  'Anonymous',  'Armada.Collective',
'Chinese.hacker',  'Chinese.hackers',  'Cyber.Islamic.State',  'CyberBerkut',
'DarkWeb.Goons',  'DERP',  'Dr.SHA6H',  'Guccifer',  'HAXOR',  'Iranian.Hackers',
'Izz.ad.Din.al.Qassam.Cyber.Fighters',     'JokerCracker',     'KelvinSecTeam',
'LizardSquad', 'LulzSec', 'Maxney', 'NetPirates', 'NullCrew', 'Other', 'RedHack'
'Rex.Mundi',    'Syrian.Electronic.Army',    'TEAM.MADLEETS',    'TeamBerserk',
'Tunisian.Cyber.Army',     'Turkish.Ajan',     'X.smitt3nz',     'X.th3int1d3l',
'XTnR3v0lT'

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 1285, 1290, 1292, 1288, 1288, 1290, ...

Resampling results across tuning parameters:

  fL  usekernel  Accuracy   Kappa

  0   FALSE      0.5816532  0.4380890

  0    TRUE      0.5816532  0.4380890

  1   FALSE      0.5765824  0.3839308

  1    TRUE      0.5765824  0.3839308

  2   FALSE      0.5753258  0.3669369

  2    TRUE      0.5753258  0.3669369

  3   FALSE      0.5732716  0.3569336

  3    TRUE      0.5732716  0.3569336

  4   FALSE      0.5718738  0.3479041

  4    TRUE      0.5718738  0.3479041

  5   FALSE      0.5746812  0.3462661

  5    TRUE      0.5746812  0.3462661

  6   FALSE      0.5775077  0.3418538

  6    TRUE      0.5775077  0.3418538

  7   FALSE      0.5761094  0.3353987

  7    TRUE      0.5761094  0.3353987
  8   FALSE      0.5802863  0.3403856

  8    TRUE      0.5802863  0.3403856

Tuning parameter 'adjust' was held constant at a value of 1

Accuracy was used to select the optimal model using  the largest value.

The final values used for the model were fL = 0,usekernel = FALSE and adjust =1
```

Figure 4-53 Training process of Cyber Attackers model based on NB

According to scripts and training process, the optimal model was chosen where the Laplace correction is 0 and the kernel is equal to

false as the accuracy reaches its maximum level with 58.16% reliability. Figure 4-54 shows the plot of changing accuracy level based on Laplace correction and it highlights the inconstant behaviour of accuracy over Laplace correction where the accuracy has the highest level when Laplace is 0 and then it has decreasing trend and it gets the lowest level when Laplace is 4 then it will get increasing trend again.

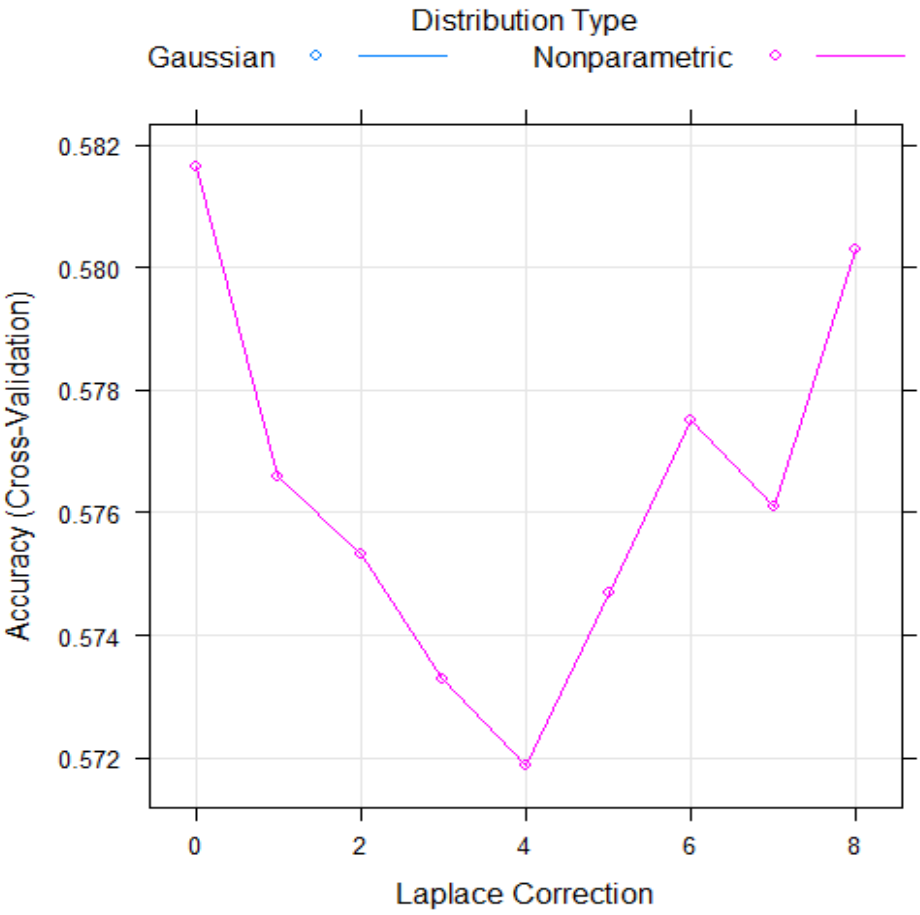

Figure 4-54 Accuracy trend for prediction of cyber attackers in NB

4.4.3 Prediction of Type of Target by Naïve Bayes Classifier

In this stage, a classifier will be trained in order to classify and predict the future potential Type of Targets in cyber-attacks. In this scenario, all of the cyber-attack records including 2694 incidents will be used as training data. Laplace correction is set from

0 to 8 and 10 fold cross validation will be applied as a training control parameter. Figure 4-55 presents the training process of the classifier.

```
Target_nbt1 <- train(Tarx, Tary,data = Dtarget, method = "nb", trControl =
myc1,metric = "Accuracy",importance = TRUE,tuneGrid = gridtest1)

Naive Bayes


2694 samples

   4 predictor

  19 classes: 'BP', 'ED', 'EN', 'ES', 'FB', 'GO', 'HC', 'HT', 'IO', 'MD', 'MU',
'NN', 'RT', 'SI', 'SN', 'TC', 'THS', 'TP', 'Other'


No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 2423, 2426, 2425, 2426, 2425, 2423, ...

Resampling results across tuning parameters:


  fL  usekernel  Accuracy   Kappa

  0   FALSE      0.3652246  0.2721376

  0    TRUE      0.3652246  0.2721376

  1   FALSE      0.3782290  0.2548629

  1    TRUE      0.3782290  0.2548629

  2   FALSE      0.3760164  0.2459761

  2    TRUE      0.3760164  0.2459761

  3   FALSE      0.3771303  0.2443304

  3    TRUE      0.3771303  0.2443304

  4   FALSE      0.3652617  0.2273073

  4    TRUE      0.3652617  0.2273073

  5   FALSE      0.3608090  0.2198033

  5    TRUE      0.3608090  0.2198033

  6   FALSE      0.3567225  0.2132026

  6    TRUE      0.3567225  0.2132026

  7   FALSE      0.3518828  0.2046388

  7    TRUE      0.3518828  0.2046388

  8   FALSE      0.3481859  0.1990067

  8    TRUE      0.3481859  0.1990067


Tuning parameter 'adjust' was held constant at a value of 1

Accuracy was used to select the optimal model using  the largest value.

The final values used for the model were fL = 1, usekernel = FALSE and adjust =
1.
```

Figure 4-55 Training process of Type of Target model based on NB

The maximum accuracy happens when the Laplace correction is 1 and kernel gets False, therefore the chosen model has 37.82% accuracy. Figure 4-56 shows the plot of accuracy changing trend over Laplace

correction and it demonstrates that accuracy will be at its maximum level when Laplace is equal to 1 and then after that, it follows a downward trend and reaches its minimum level when the Laplace is 8.

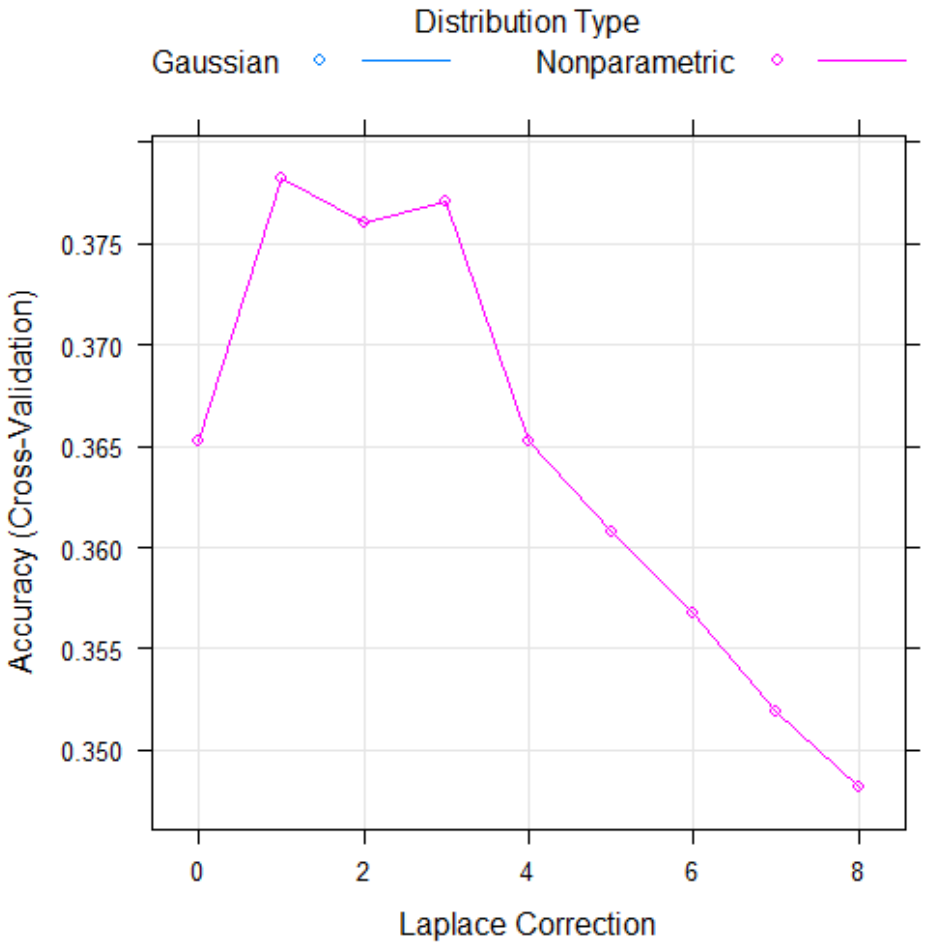

Figure 4-56 Accuracy trend for prediction of Type of Target in NB

4.4.4 Prediction of Targeted country by Naïve Bayes

In this part, the model will be trained to predict targeted countries in cyber-attacks and all of the 2694 records will be used as training set to train the model by Naïve Bayes algorithms. In order to decrease level of countries and make the models more accurate and extensive, the name of those countries which they get targeted less than 1 percent of number of cyber-attacks will be turned into Others. 10 fold cross

validation will be applied to the training process as training control parameters and Laplace correction will vary from 0 to 8 and figure 4-57 shows the process of training the classifier:

```
Country_nbt1 <- train(Cntx, Cnty,data = Dcnt, method = "nb", trControl =
myc1,metric = "Accuracy",importance = TRUE, tuneGrid = gridtest1)

Naive Bayes


2694 samples

   4 predictor

  21 classes: 'AU', 'BR', 'CA', 'CN', 'CZ', 'DE', 'FR', 'IL', 'IN', 'INT', 'IT',
'JP', 'KR', 'PH', 'PK', 'RU', 'SA', 'TR', 'UK', 'US', 'Other'


No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 2423, 2421, 2423, 2425, 2425, 2425, ...

Resampling results across tuning parameters:


  fL  usekernel  Accuracy   Kappa

  0   FALSE      0.4603262  0.2272340

  0    TRUE      0.4603262  0.2272340

  1   FALSE      0.4755642  0.2074231

  1    TRUE      0.4755642  0.2074231

  2   FALSE      0.4711183  0.1914337

  2    TRUE      0.4711183  0.1914337

  3   FALSE      0.4696338  0.1846077

  3    TRUE      0.4696338  0.1846077

  4   FALSE      0.4677832  0.1793312

  4    TRUE      0.4677832  0.1793312
  5   FALSE      0.4666612  0.1756461

  5    TRUE      0.4666612  0.1756461

  6   FALSE      0.4659246  0.1726962

  6    TRUE      0.4659246  0.1726962

  7   FALSE      0.4644527  0.1682928

  7    TRUE      0.4644527  0.1682928

  8   FALSE      0.4637024  0.1649815

  8    TRUE      0.4637024  0.1649815


Tuning parameter 'adjust' was held constant at a value of 1

Accuracy was used to select the optimal model using  the largest value.

The final values used for the model were fL = 1, usekernel = FALSE and adjust =
1.
```

Figure 4-57 Training process of Targeted Country model based on NB

The largest value of accuracy happens when the Laplace is 1 and the kernel is false so the model will be chosen with an accuracy of 47.55%. Figure 4-58 also presents the value of the accuracy based on Laplace correction amount. As it is shown the maximum level happens when

the Laplace is 1 and then after that, the accuracy has decreasing course and it reaches to its minimum when Laplace is equal to 8.

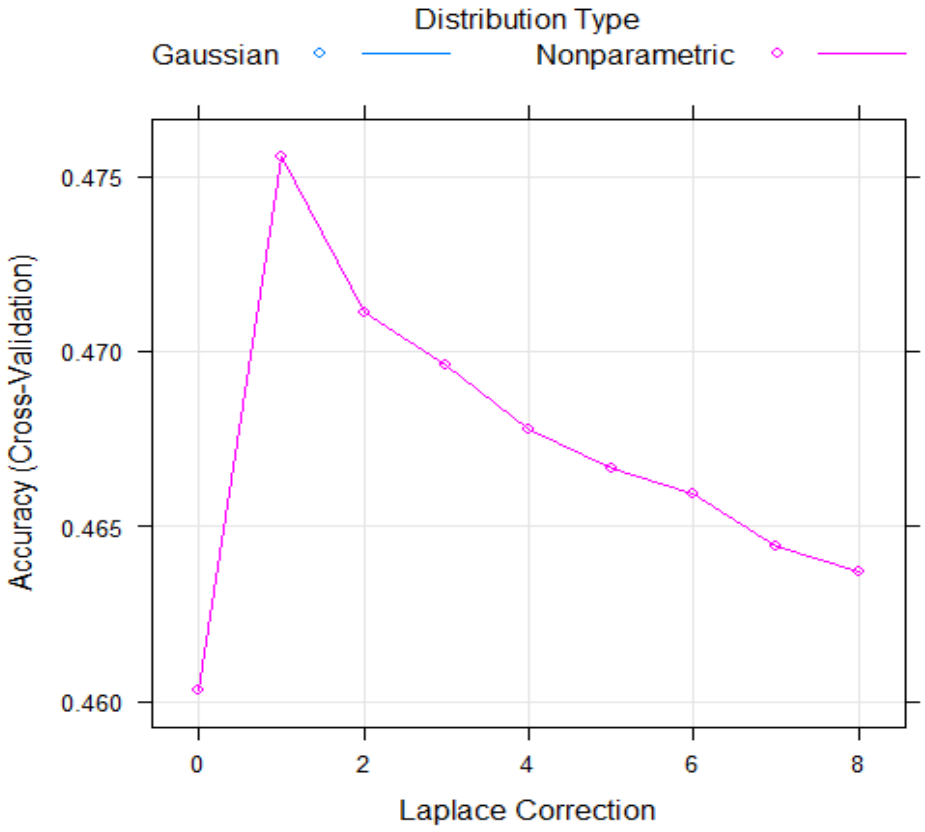

Figure 4-58 Accuracy trend for prediction of targeted country in NB

4.4.5 Prediction of Cyber Attack Activity

In the dataset, there are 2694 records and all of the attacks are recognized in terms of the type of activity where the intention of them was clear. Prediction of the type of activity of cyber-attacks can give the ability to security experts to figure out the intention of future cyber-attacks when they are unknown and ambiguous and profile cyber criminals for law enforcement agencies. Naïve Bayes algorithm will train this classifier and Laplace correction will be varied from 0 to 8 and 10 fold cross validation will be used as training control technique and Figure 4-59 describes the process of the training the model.

```
gridtest1                <-                expand.grid(fL=c(0,1,2,3,4,5,6,7,8),
adjust=c(1) ,usekernel=c(FALSE,TRUE))

Activity_nbt1 <- train(Actx, Acty,data = Dact, method = "nb", trControl =
myc1,metric = "Accuracy",importance = TRUE,tuneGrid = gridtest1)

Naive Bayes


2694 samples

   4 predictor

   4 classes: 'CC', 'CE', 'CW', 'HA'


No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 2424, 2426, 2425, 2423, 2424, 2424, ...

Resampling results across tuning parameters:


  fL  usekernel  Accuracy   Kappa

  0   FALSE      0.8158638  0.6688016

  0    TRUE      0.8158638  0.6688016

  1   FALSE      0.8192302  0.6710477

  1    TRUE      0.8192302  0.6710477

  2   FALSE      0.8107020  0.6507867

  2    TRUE      0.8107020  0.6507867

  3   FALSE      0.8073562  0.6417407

  3    TRUE      0.8073562  0.6417407

  4   FALSE      0.8025207  0.6310110

  4    TRUE      0.8025207  0.6310110

  5   FALSE      0.7988018  0.6225276

  5    TRUE      0.7988018  0.6225276

  6   FALSE      0.7999157  0.6237972

  6    TRUE      0.7999157  0.6237972
  7   FALSE      0.7962065  0.6153845

  7    TRUE      0.7962065  0.6153845

  8   FALSE      0.7947126  0.6115147

  8    TRUE      0.7947126  0.6115147


Tuning parameter 'adjust' was held constant at a value of 1

Accuracy was used to select the optimal model using  the largest value.

The final values used for the model were fL = 1, usekernel = FALSE and adjust =
1.
```

Figure 4-59 Training process of Cyber Attack Activity model based on NB

The most accurate and optimal model will be chosen when the kernel is false, Laplace is equal to 1 and the value of the accuracy is 81.92%.

Figure 4-60 demonstrates that increasing Laplace value has a negative effect on accuracy level and as it is shown the maximum level of accuracy happens in Laplace value of 1 and the minimum accuracy happens when Laplace is 8.

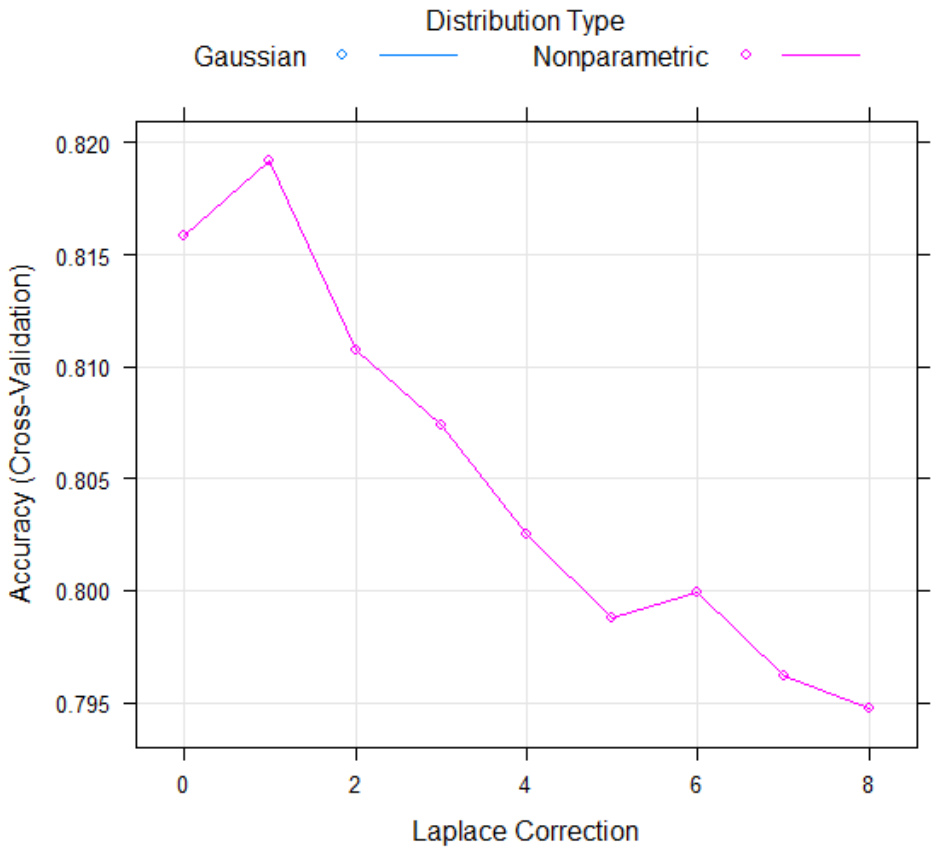

Figure 4-60 Accuracy trend for prediction of type of cyber-attack activity in NB

4.4.6 Discussion and Interpretation

Table 4-3 shows the accuracy of the Naïve Bayes classifier in terms of prediction of different features in cyber-attacks.

Type of	Naïve Bayes accuracy rate

Prediction	
Cyber Attack Activity	81.92%
Cyber Attacker	58.16%
Targeted Country	47.55%
Type of Target	37.82%
Type of Threat	55.88%

Table 4-3 NB accuracy in prediction

According to the result following results can be concluded from the data analysis by Naïve Bayes:

1- In terms of prediction of cyber-attack activity, Naïve Bayes approach had 81.95% accuracy which is significant reliability and this model can be used as a single tool for cyber security experts to predict attackers' motivation in future based on the given information. Naïve Bayes has done the accurate job on classifying cyber-attack activities which can be interpreted that the independence assumption of different variables by this algorithm is a reasonable hypothesis in this project data set.

2- In terms of prediction of cyber attackers and Type of Threat, Naïve Bayes algorithm did an average job with more than 55% accuracy, however, this approach can be a supplementary tool along with other methods for cyber security managers to protect their interests.

3- Prediction of the Type of Target and the targeted country has less than average accuracy showing that Naïve Bayes cannot be trusted among other methods and tools for forecasting future potential targets for cyber attackers. This can be interpreted as a prediction of the Type of Targeted or targeted country can also depend on other factors such as attack origin or attacker country which are not available in the obtained dataset in this project.

4.5 Support Vector Machin Analysis

The Support Vector Machine will be applied to the data set after the pre-processing stage. In this step, e1071 package will be used along with Caret which gives a better overview of the training process. E1071 package was developed by University of Vienna including different

classification functions and also has a built-in library for SVM called libsvm which covers SVM algorithms (Meyer, 2017).

In the training process, there is a parameter called Cost which defines the size of margin in SVM classifier. If the cost gets the high value the size margin will get smaller and that means the hyperplane will try to classify all of the training objects without avoiding them. On the other hand, if C gets small value, the hyperplane will avoid misclassifying objects in the result of increasing margin.

4.5.1 Prediction of Type of Threat by Support Vector Machine

Among 2694 records of cyberattacks happening from 2013 to 2015, 2210 attacks were known to cyber experts or mentioned in our dataset obtained from OSINT thus they will be used as training set for building the SVM classifier. The training control is set to 10 fold cross-validation and the cost parameters will be adjusted from 0.1 to 1. Figure 4-61 is describing the training process.

```
Threat_lsvm <- train(Threat ~ Target.Section + Author + Activity + Cou
ntry,data = Dthreat, method = "svmLinear", trControl = myc1,metric = "
Accuracy",tuneGrid = data.frame(.C = c(.25, .5, 1)),importance = TRUE)

Support Vector Machines with Linear Kernel



2210 samples

   5 predictor

  11 classes: 'AH', 'CSS', 'DF', 'DH', 'DS', 'MWV', 'SQ', 'TA', 'UA',
'ZD', 'Other'



No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 1989, 1989, 1987, 1988, 1990, 1990, ...

Resampling results across tuning parameters:


  cost  Accuracy   Kappa

  0.1   0.5633251  0.4715831

  0.2   0.5814209  0.4941164

  0.3   0.5922849  0.5078894

  0.4   0.5931940  0.5094549

  0.5   0.5950060  0.5117738

  0.6   0.5963716  0.5136368

  0.7   0.6022561  0.5211247

  0.8   0.6035890  0.5231325

  0.9   0.6035849  0.5231657

  1.0   0.6049322  0.5249531


Accuracy was used to select the optimal model using the largest value.

The final value used for the model was cost = 1.
```

Figure 4-61 Training process of Type of Threat model based on SVM

As it is described in the codes, the accuracy reaches its highest amount with 60.49% when the cost is equal to 1 and the lowest level of accuracy is 56.33 when the cost is 0.1. Figure 4-62 shows an

increasing trend of accuracy based on cost parameter, which indicates they have a direct relationship as if the cost increases the model, will be more reliable and accurate.

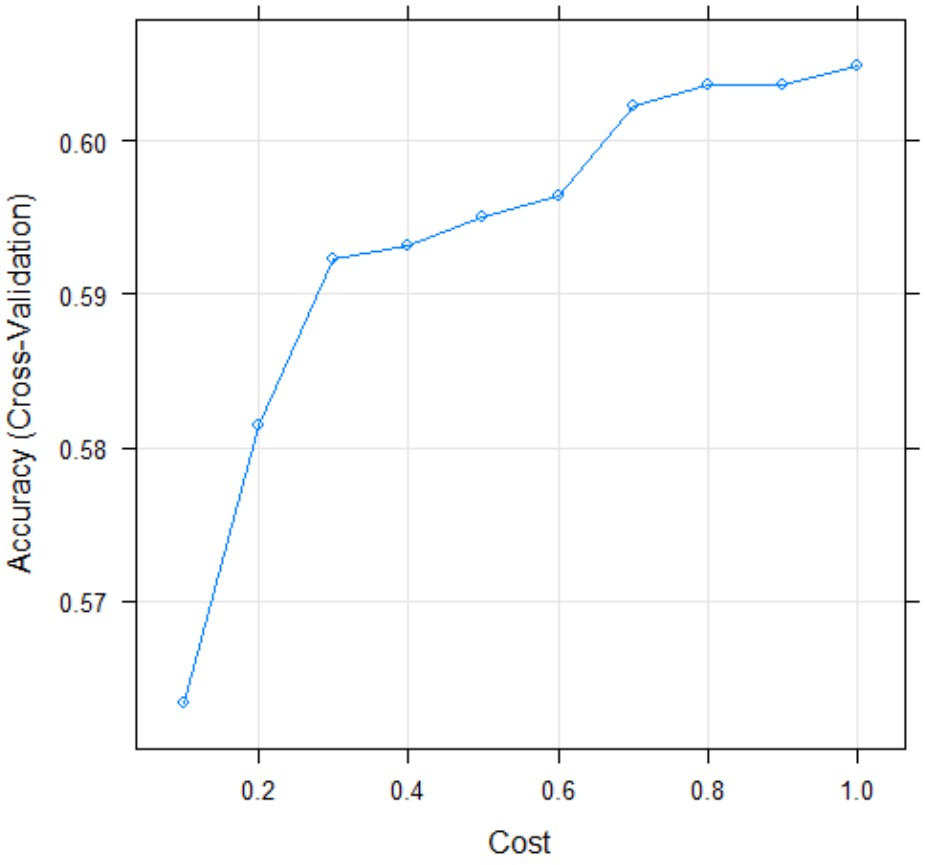

Figure 4-62 Accuracy trend for prediction of type of cyber threat in SVM

4.5.2 Prediction of Cyber Attackers by Support Vector Machine

1432 cyber attackers were identified or they took credit for their attacks in the obtained dataset and their records will be employed as the training set. In addition, it has been decided to convert the name of those attackers who committed cyber-attacks less than 5 times to "others". 10 fold cross validation has been applied as training control and the cost parameter will be varied from 0.1 to 1. Figure 4-63

explains the building of the SVM classifier for prediction of cyber attackers:

```
Support Vector Machines with Linear Kernel

1432 samples
   5 predictor
  33 classes: 'Ag3nt47', 'AnonGhost', 'Anonymous', 'Armada.Collective
', 'Chinese.hacker', 'Chinese.hackers', 'Cyber.Islamic.State', 'Cyber
Berkut', 'DarkWeb.Goons', 'DERP', 'Dr.SHA6H', 'Guccifer', 'HAXOR', 'I
ranian.Hackers', 'Izz.ad.Din.al.Qassam.Cyber.Fighters', 'JokerCracker
', 'KelvinSecTeam', 'LizardSquad', 'LulzSec', 'Maxney', 'NetPirates',
 'NullCrew', 'Other', 'RedHack', 'Rex.Mundi', 'Syrian.Electronic.Army
', 'TEAM.MADLEETS', 'TeamBerserk', 'Tunisian.Cyber.Army', 'Turkish.Aj
an', 'X.smitt3nz', 'X.th3inf1d3l', 'XTnR3vOlT'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1292, 1289, 1288, 1286, 1291, 1290, ...
Resampling results across tuning parameters:

  cost  Accuracy   Kappa
  0.1   0.5974527  0.3827388
  0.2   0.6064288  0.4032040
  0.3   0.6106790  0.4079709
  0.4   0.6099794  0.4109877
  0.5   0.6092807  0.4117749
  0.6   0.6071968  0.4151346
  0.7   0.6113593  0.4236463
  0.8   0.6134189  0.4298831
  0.9   0.6071295  0.4221077
  1.0   0.6085812  0.4285040

Accuracy was used to select the optimal model using  the largest valu
e.
The final value used for the model was cost = 0.8.
```

Figure 4-63 Training process of Cyber Attacker model based on SVM

As it is shown in the codes and figure 4-64 the highest accuracy in the training process happens when the cost is 0.8 and on the cost of 0.1 the lowest accuracy will occur. Therefore the training process chose 61.34% as the most accurate and reliable model. It should be mentioned that the changing trend of accuracy does not have any relationship with the cost and it does not follow a constant behaviour.

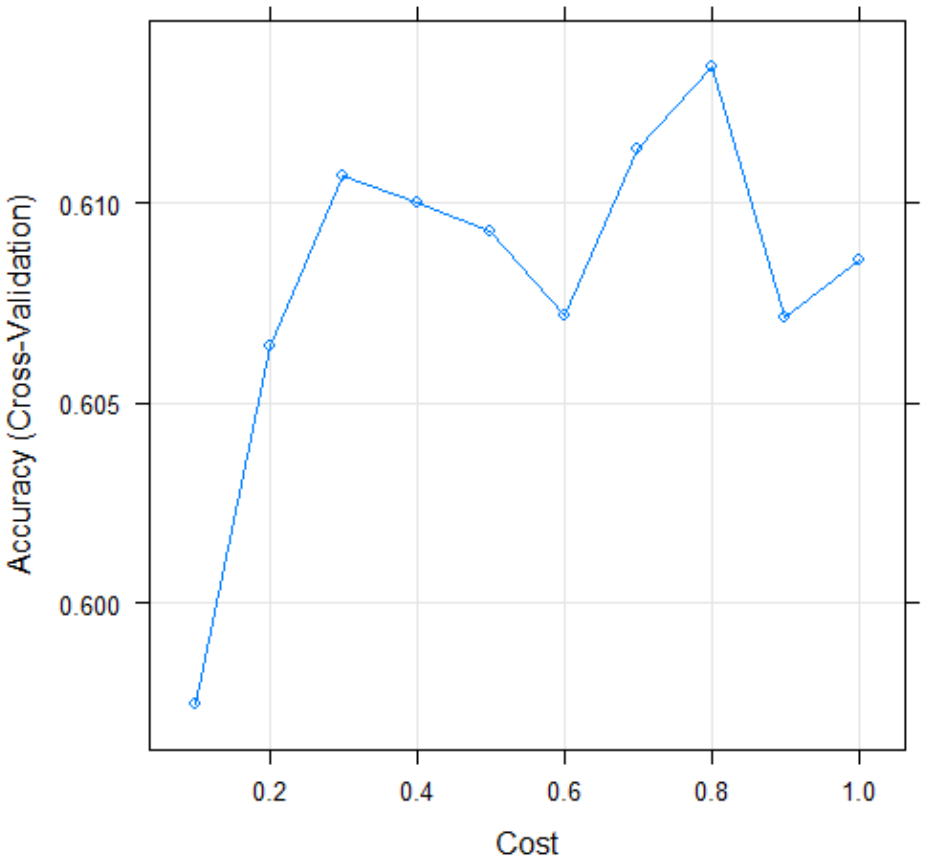

Figure 4-64 Accuracy trend for prediction of type of cyber attackers in SVM

4.5.3 Prediction of Targeted Country by Support Vector Machine

This stage aims to make an SVM classifier to make prediction about the potential burnable countries against cyber-attacks. The training set includes all 2694 records where the country of the victim has been identified and it has been decided to convert classes of those countries which have been targeted less than 22 times to “others”. The training control is defined as 10 fold cross-validation and also the cost will be changeable from 0.1 to 1. The process of training the classifier is shown in figure 4-65.

```
Support Vector Machines with Linear Kernel

2694 samples
   5 predictor
  21 classes: 'AU', 'BR', 'CA', 'CN', 'CZ', 'DE', 'FR', 'IL', 'IN', '
INT', 'IT', 'JP', 'KR', 'PH', 'PK', 'RU', 'SA', 'TR', 'UK', 'US', 'Ot
her'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2425, 2424, 2423, 2425, 2426, 2426, ...
Resampling results across tuning parameters:

  cost  Accuracy   Kappa
  0.1   0.4639358  0.1708123
  0.2   0.4706178  0.1838156
  0.3   0.4858582  0.2082579
  0.4   0.4806730  0.2019414
  0.5   0.4817785  0.2053648
  0.6   0.4880928  0.2164289
  0.7   0.4895840  0.2201068
  0.8   0.4888349  0.2183081
  0.9   0.4862368  0.2166928
  1.0   0.4862478  0.2164701

Accuracy was used to select the optimal model using  the largest valu
e.
The final value used for the model was cost = 0.7.
```

Figure 4-65 Training process of Targeted Country model based on SVM

According to figure 4-66 which demonstrates the changing trend of accuracy over the cost parameter and also considering the training process described in the script, the most accurate predictive model will be achieved when the cost is 0.7 and its accuracy is 48.95%. It is also necessary to consider that accuracy level does not have a constant relationship with the cost and it behaves abnormally.

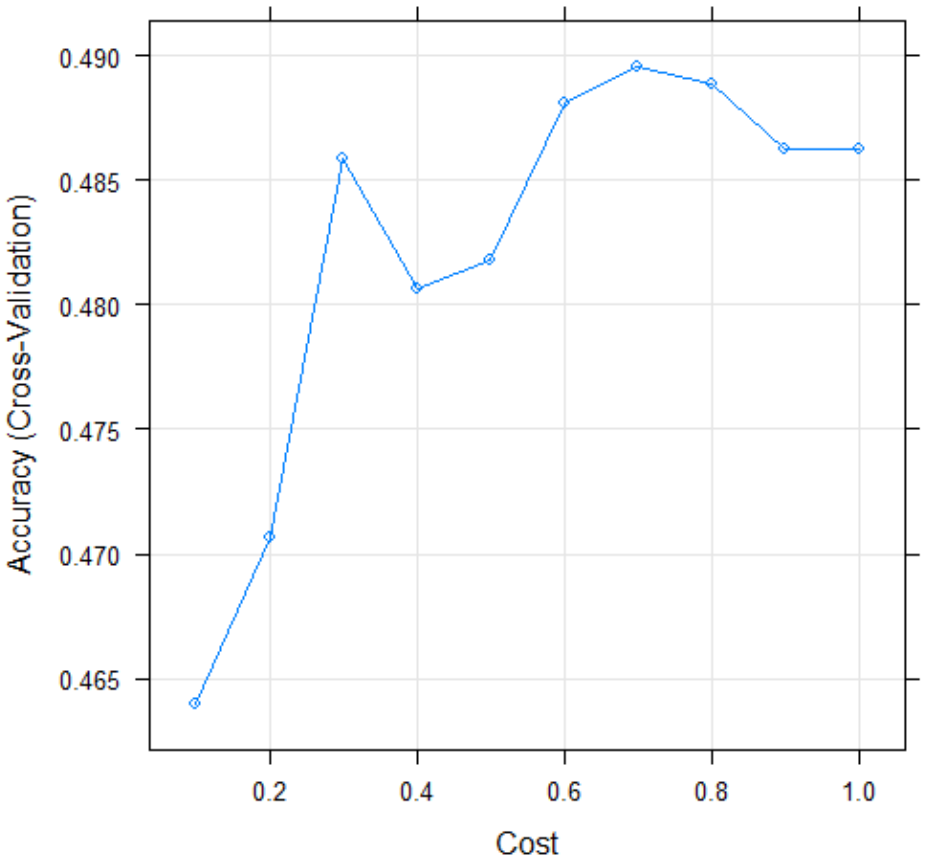

Figure 4-66 Accuracy trend for prediction of targeted country in SVM

4.5.4 Prediction of Type of Target by Support Vector Machine

In this section, SVM classifier will be trained in order to make prediction about Type of Targets in cyber-attacks. All 2694 records will be used for the training set and 10 fold cross validation will be applied as training control and the cost parameter is set from 0.1 to 1. The training process is described in figure 4-67.

```
Support Vector Machines with Linear Kernel

2694 samples
   5 predictor
  19 classes: 'BP', 'ED', 'EN', 'ES', 'FB', 'GO', 'HC', 'HT', 'IO', 'MD', 'MU'
, 'NN', 'RT', 'SI', 'SN', 'TC', 'THS', 'TP', 'Other'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2426, 2421, 2426, 2426, 2425, 2425, ...
Resampling results across tuning parameters:

  cost  Accuracy   Kappa
  0.1   0.3809047  0.2592448
  0.2   0.3920308  0.2757371
  0.3   0.3968884  0.2831498
  0.4   0.3920789  0.2782685
  0.5   0.3905781  0.2774454
  0.6   0.3887083  0.2753842
  0.7   0.3879675  0.2747299
  0.8   0.3886865  0.2758873
  0.9   0.3879361  0.2764398
  1.0   0.3901504  0.2801462

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was cost = 0.3.
```

Figure 4-67 Training process of Type of Target model based on SVM

Figure 4-68 shows an abnormal trend of the accuracy over the cost parameter and when the cost is 0.3, the accuracy will reach its peak and it is 39.68%. The lowest accuracy happens when the cost is 0.1. The most accurate and optimal model will be made with 39.68 % reliability in the prediction of the Type of Target.

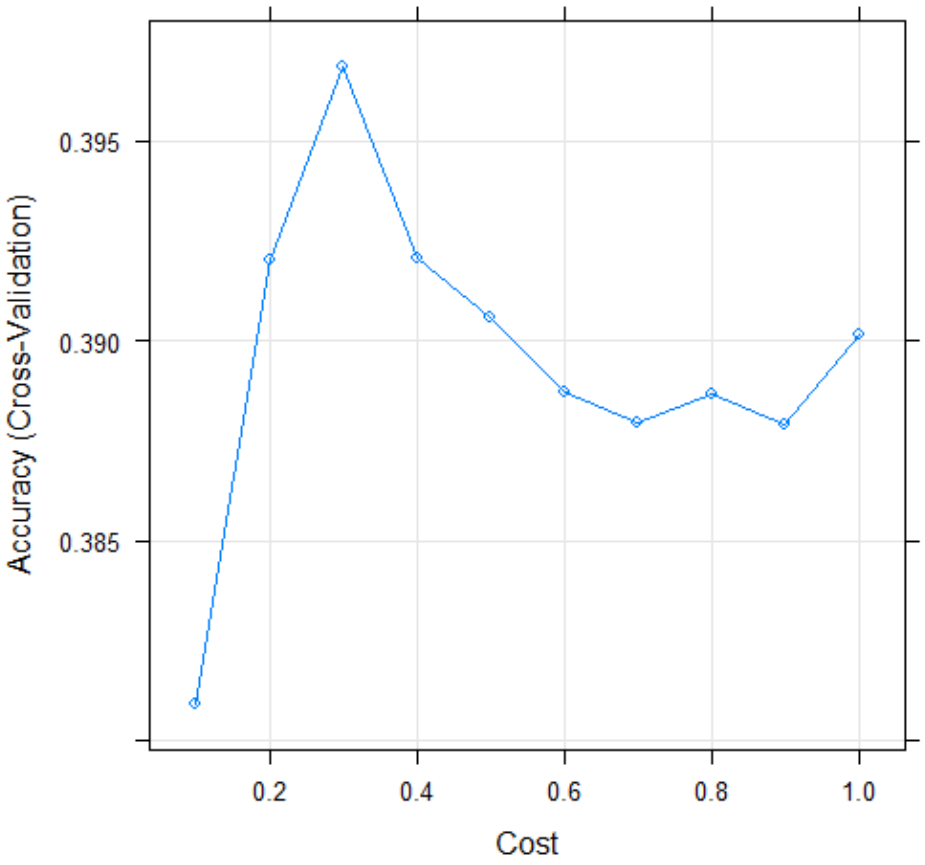

Figure 4-68 Accuracy trend for prediction of Type of Target in SVM

4.5.5 Prediction of Cyber Attack Activity by Support Vector Machine

2694 records existing in the dataset, their motivations have been identified regarding OSINT sources. They will be used as the training set in order to train a classifier making cyber security experts to identify and predict cyber-attacks' type of activity and motivation. Like the previous sections, the same training control and tuning parameter will be applied to the training process and figure 4-69 shows the training process.

```
Support Vector Machines with Linear Kernel

2694 samples
   5 predictor
   4 classes: 'CC', 'CE', 'CW', 'HA'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2425, 2423, 2425, 2424, 2425, 2424, ...
Resampling results across tuning parameters:

  cost  Accuracy   Kappa
  0.1   0.8125544  0.6559966
  0.2   0.8188617  0.6675169
  0.3   0.8214680  0.6721891
  0.4   0.8244379  0.6778967
  0.5   0.8248110  0.6788787
  0.6   0.8255573  0.6806306
  0.7   0.8248069  0.6788444
  0.8   0.8221964  0.6743097
  0.9   0.8214487  0.6733986
  1.0   0.8203376  0.6720097

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was cost = 0.6.
```

Figure 4-69 Training process of Cyber Attack Activity model based on SVM

The changing accuracy level over the cost has been shown in figure 4-70 as it demonstrates the accuracy level has increasing trend until the cost is equal to 0.6 and at that point the accuracy in 82.55% and again after that it has decreasing pattern so the training process will choose the best model with 82.55% reliability with cost of 0.6.

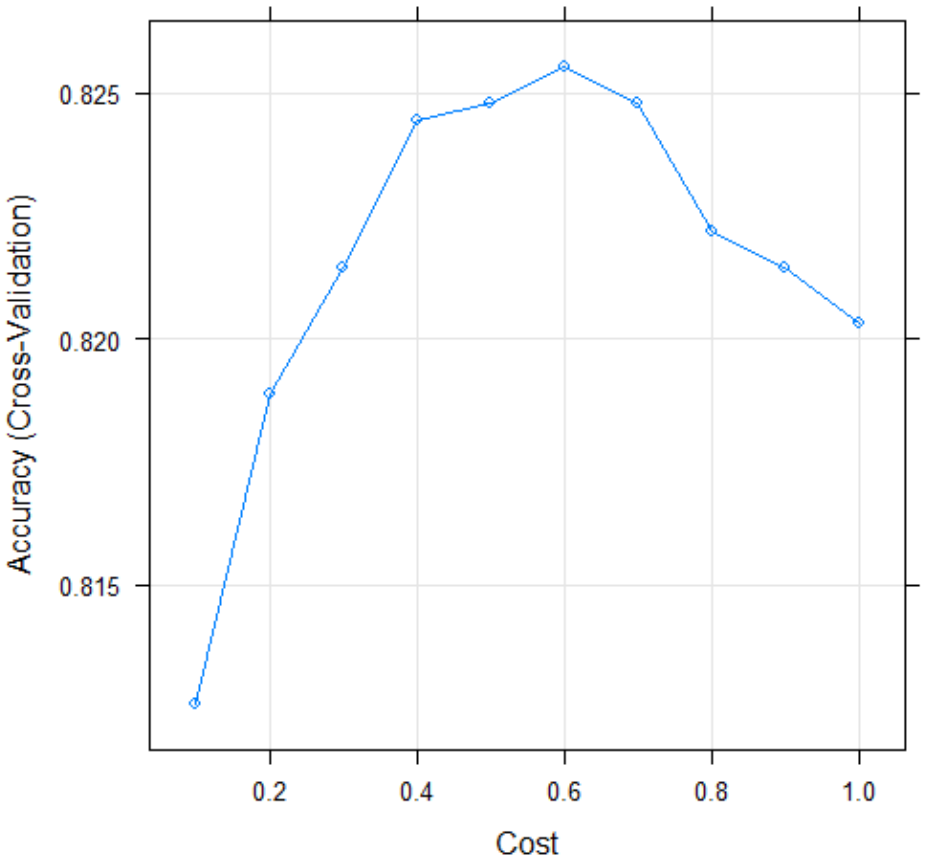

Figure 4-70 Accuracy trend for prediction of type of cyber-attack activity in SVM

4.5.6 Discussion and interpretation

After training the dataset with SVM method, optimal models in terms of accuracy and reliability will be concluded. Table 4-4 shows the accuracy of the obtained model in the prediction of different aspects of cyber-attacks.

Type of	Accuracy level

prediction	
Type of Threat	60.49%
Cyber attackers	61.34%
Targeted Country	48.95%
Type of Target	39.68%
Type of activity	82.55%

Table 4-4 Accuracy prediction in SVM

According to table 4-4, the following points will be extracted:

1. In terms of prediction of type of cyberattacks, SVM has done a reliable job where the accuracy of prediction is 82.55%. This result demonstrates that there is a strong and firm pattern in the dataset among different attributes of cyber-attacks to forecast the type of cyber activity.

2. In terms of prediction of cyber attackers and cyber threat, SVM can predict potential cyber attackers and threats with 61.34% and 60.49% confidence respectively, however, these models are not accurate enough to be relied on. Cyber security experts can use this model as a tool combined with other tools and methods to identify and tackle cyber criminals.

3. The model for prediction of targeted country has 48.95% accuracy indicating that it is not reliable as a method to predict vulnerable countries against cyberattacks. This might be because of other factors and actors playing in cyber-attack incidents. With that mentioned, this predictive model can stand along with other methods and sources to identify vulnerable and victim countries in order to protect them against future cyber threats.

4. In terms of prediction of target type, the accuracy of SVM classifier is not appropriate showing that it is not reliable method to identify the target of cyberattacks, however, compared to other classifying algorithms, SVM has done a slightly better job.

4.6 Neural Network (Multilayer Perceptron) Analysis

This section aims to investigate the implementation of ANN as a training algorithm to build a classifier to predict different features of cyber-attacks. There are different packages in R for implementation of ANN such as NNET developed by (REF) and RSNNS designed by (Bergmeir and Benitez, 2016). It has been decided to analyse the date by RSNNS because it is easy to implement and use and it includes a user-friendly library with different functions on Neural Network Application. RSNNS along with Caret package will be employed in this part of the experiment to have a more comprehensible overview of the training process. Size is the main parameter in RSNNS package which defines the number of hidden layers in the neural network structure and it will be identified as turning parameter in the training process.

4.6.1 Prediction of Type of Threat by NN

In order to build a predictive model for forecasting future and unknown cyber threats, ANN algorithm will be applied to the training set which has 2210 records of known cyber threats. 10 fold cross validation is training control and the size of ANN can vary from one layer to 7 layers and figure 4-71 explains the training process.

```
Multi-Layer Perceptron

2210 samples
   5 predictor
  11 classes: 'AH', 'CSS', 'DF', 'DH', 'DS', 'MWV', 'SQ', 'TA', 'UA', 'ZD', 'O
ther'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1768, 1770, 1767, 1768, 1767,…
Resampling results across tuning parameters:

  size  Accuracy   Kappa
  1     0.3895910  0.2335295
  3     0.5746652  0.4860765
  5     0.5814577  0.4947028
  7     0.5886842  0.5035666

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was size = 7.
```

Figure 4-71 Training process of Type of Threat model based on ANN

The code illustrates that the most accurate ANN model for prediction of type of cyber threat has 7 layers and 58.86% accuracy. Figure 4-72 shows the trend of accuracy based on the size of ANN which can be seen as a direct relationship between them. In other words, if the ANN has more hidden layers the accuracy will be higher as well.

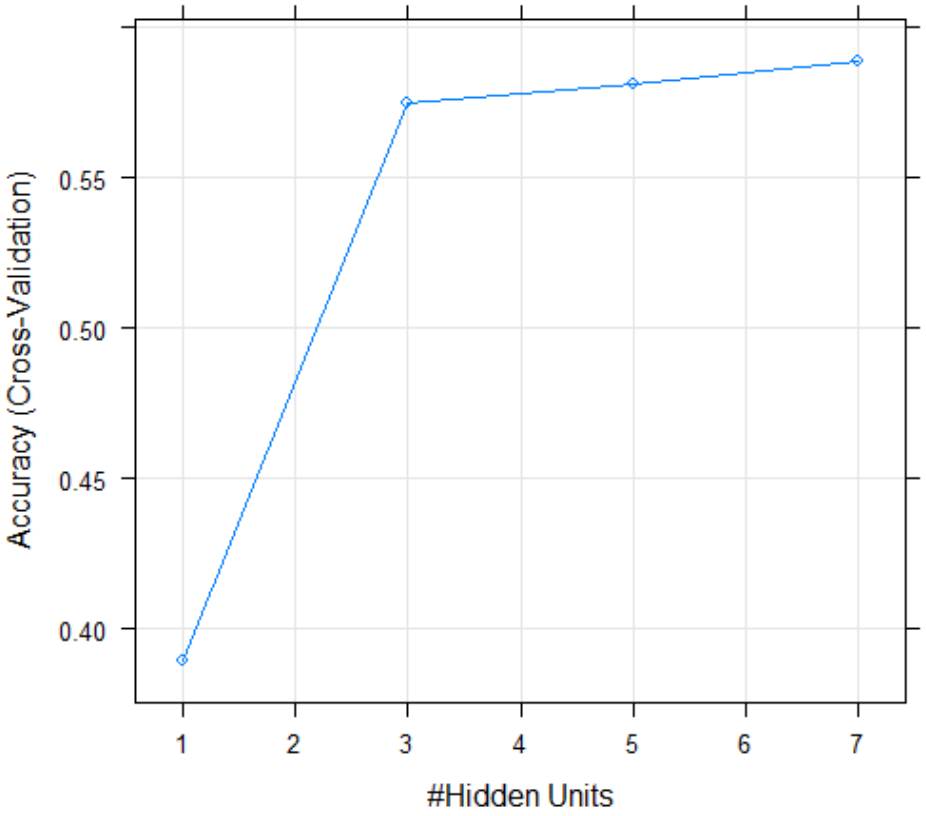

Figure 4-72 Accuracy trend for prediction of type of cyber threat in NN

4.6.2 Prediction of Cyber attacker by NN

In the obtained dataset, 1432 cyber-attacks exist that cyber attackers took responsibility for them and this sample will be used as training set for building an ANN predictive model for cyber attackers identification. The training control will be set as 10 fold cross-validation and the size will be defined from 1 to 7 and the training process is explained in figure 4-73.

```
Multi-Layer Perceptron

1432 samples
   5 predictor
  33 classes: 'Ag3nt47', 'AnonGhost', 'Anonymous', 'Armada.Col
lective', 'Chinese.hacker', 'Chinese.hackers', 'Cyber.Islamic.
State', 'CyberBerkut', 'DarkWeb.Goons', 'DERP', 'Dr.SHA6H', 'G
uccifer', 'HAXOR', 'Iranian.Hackers', 'Izz.ad.Din.al.Qassam.Cy
ber.Fighters', 'JokerCracker', 'KelvinSecTeam', 'LizardSquad',
 'LulzSec', 'Maxney', 'NetPirates', 'NullCrew', 'Other', 'RedH
ack', 'Rex.Mundi', 'Syrian.Electronic.Army', 'TEAM.MADLEETS',
'TeamBerserk', 'Tunisian.Cyber.Army', 'Turkish.Ajan', 'X.smitt
3nz', 'X.th3inf1d3l', 'XTnR3vOlT'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1290, 1290, 1288, 1295, 1286, 1289, .
..
Resampling results across tuning parameters:

  size  Accuracy   Kappa
  1     0.5378261  0.2700091
  3     0.5928264  0.3858660
  5     0.5971274  0.3910163
  7     0.5962974  0.4057647

Accuracy was used to select the optimal model using  the large
st value.
The final value used for the model was size = 5.
```

Figure 4-73 Training process of Cyber Attackers model based on ANN

The most reliable model with 59.71% accuracy has been obtained when the size of ANN model is 5. Figure 4-74 plots the accuracy trend in the prediction of cyber attackers based on number of hidden layers. There is an increasing trend in accuracy when the number of hidden layers rises, however, after 5 layers the trend gets steady.

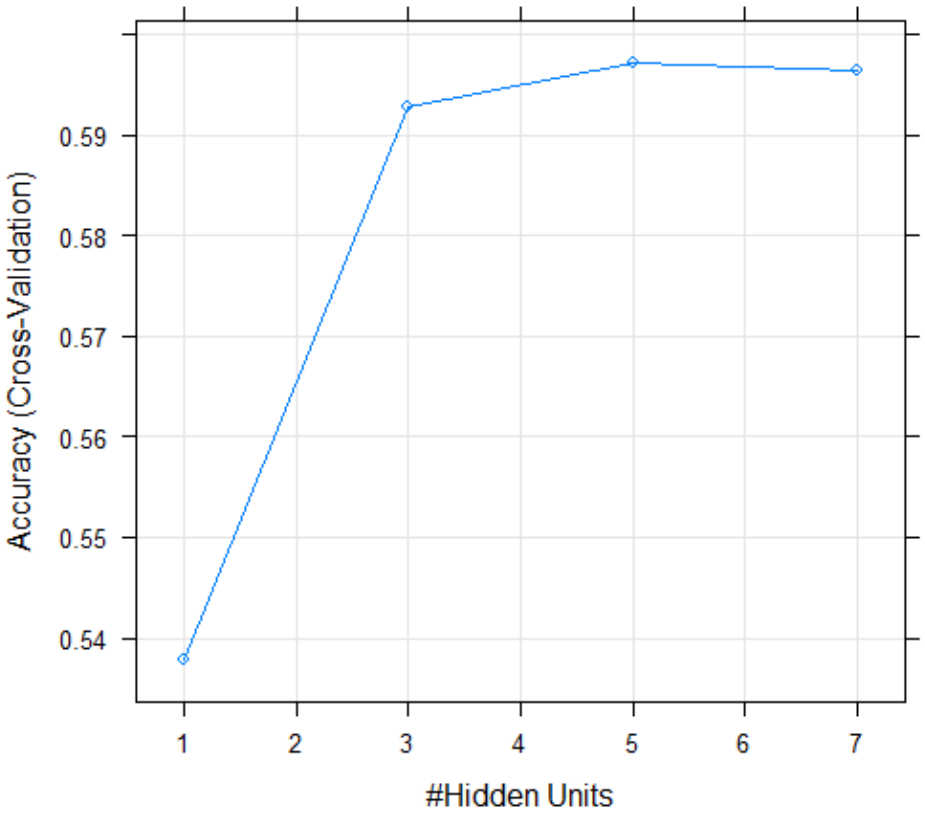

Figure 4-74 Accuracy trend for prediction of type of cyber attackers in NN

4.6.3 Prediction of Type of Target by NN

This stage of analysis aims to train a classifier to predict vulnerable targets against cyber breaches, cyber-attacks' targets have been categorized which was explained in chapter 3. The obtained dataset has 2694 attack where their targets were known and identified, therefore they take apart in the analysis as the training set. 10 fold cross validation as training control and changeable size of ANN from 1 to 7 as tuning control will be implemented in the training process and figure 4-75 demonstrates the training process.

```
Multi-Layer Perceptron

2694 samples
   5 predictor
  19 classes: 'BP', 'ED', 'EN', 'ES', 'FB', 'GO', 'HC', 'HT',
'IO', 'MD', 'MU', 'NN', 'RT', 'SI', 'SN', 'TC', 'THS', 'TP', '
Other'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2153, 2158, 2155, 2155, 2155
Resampling results across tuning parameters:

  size  Accuracy   Kappa
  1     0.3232881  0.1767404
  3     0.3592887  0.2399841
  5     0.3682273  0.2489307
  7     0.3726800  0.2587109

Accuracy was used to select the optimal model using  the large
st value.
The final value used for the model was size = 7.
```

Figure 4-75 Training process of Type of Target model based on ANN

The training process and figure 4-76 explain when the size of ANN model will be larger the accuracy will get higher as well. The most accurate model will be achieved when the NN model has 5 hidden layers and the accuracy is equal to 37.26%.

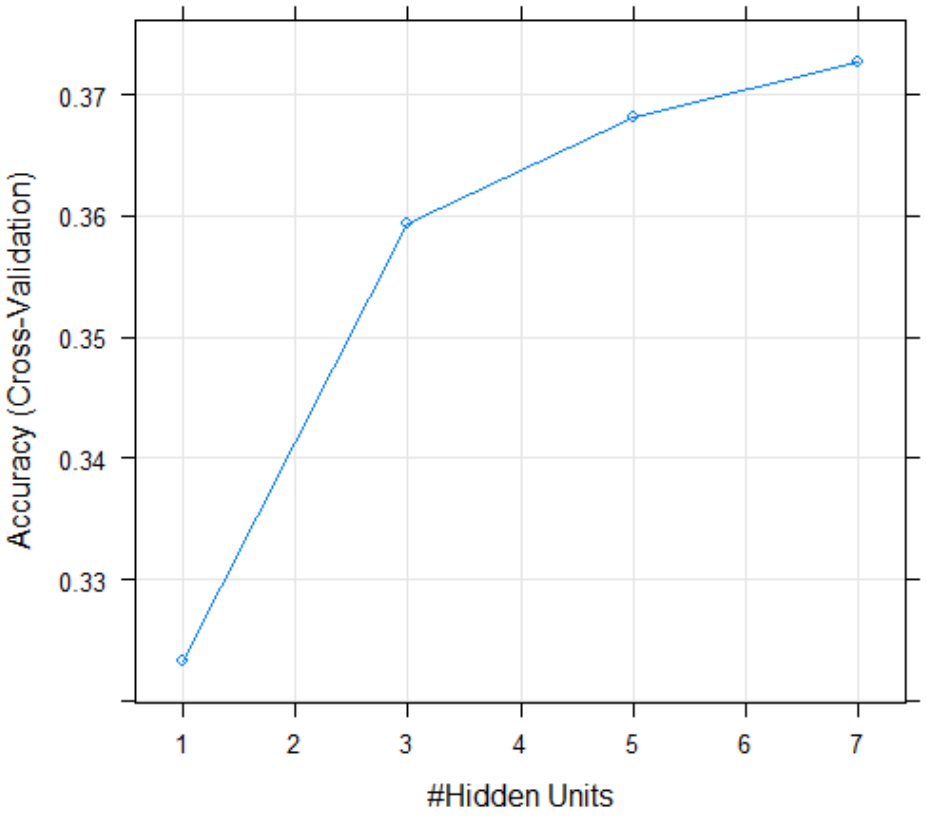

Figure 4-76 Accuracy trend for prediction of Type of Target in ANN

4.6.4 Prediction of Targeted Country by NN

This step of experiment aims to investigate the usage of ANN in the prediction of vulnerable and victim countries against cyber-attacks. The training set has 2694 records and those countries targeted by cybercriminals have been identified. In the training process, 10 fold cross validation will be applied as the training control and the size is changeable from 1 to 7 and figure 4-77 illustrates the codes generating the training process.

```
Multi-Layer Perceptron

2694 samples
   5 predictor
  21 classes: 'AU', 'BR', 'CA', 'CN', 'CZ', 'DE', 'FR', 'IL', 'IN', 'INT', 'IT
', 'JP', 'KR', 'PH', 'PK', 'RU', 'SA', 'TR', 'UK', 'US', 'Other'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2426, 2424, 2425, 2424, 2425, 2422, ...
Resampling results across tuning parameters:

  size  Accuracy   Kappa
  1     0.4643578  0.1714348
  3     0.4681225  0.1875797
  5     0.4718084  0.2118312
  7     0.4662211  0.2073910

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was size = 5.
```

Figure 4-77 Training process of Targeted Country model based on ANN

As training process shows, Caret functions identify the most reliable model with 47.18% accuracy and the ANN model has 5 hidden layers. In addition, figure 4-78 demonstrates the abnormal behaviour of accuracy level over the size parameter indicating there is no constant relationship between them.

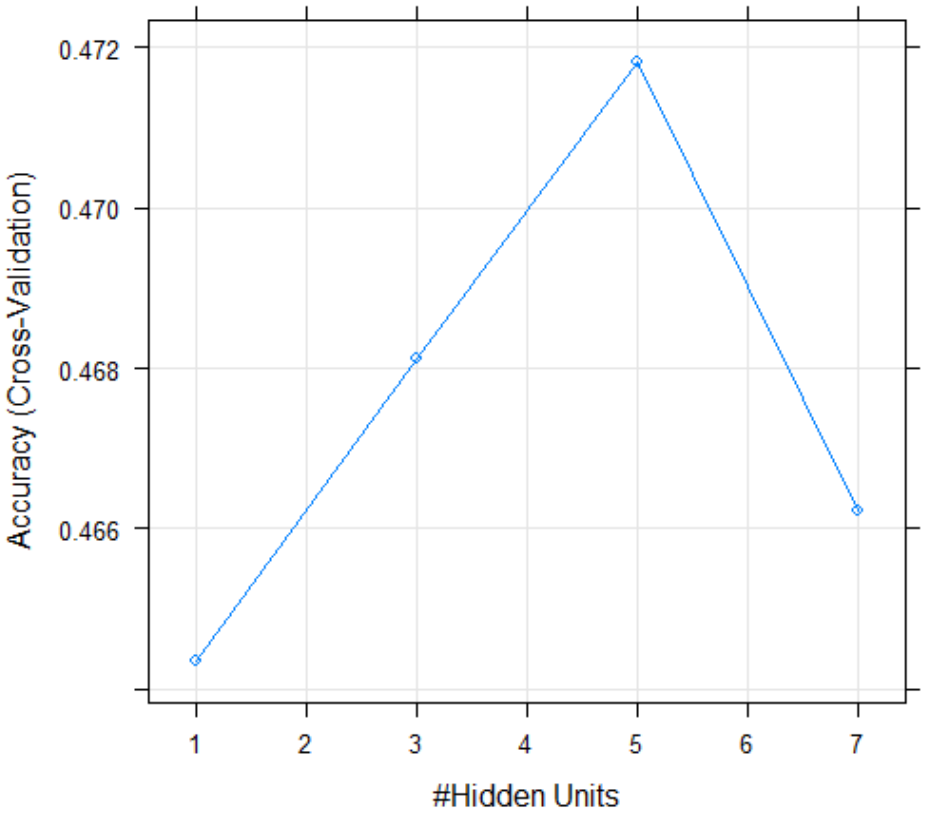

Figure 4-78 Accuracy trend for prediction of targeted country in ANN

4.6.5 Prediction of Cyber Attack Activity by NN

Prediction of cyber-attack activity helping cyber experts to detect the motivation of cyber-attacks is the aim of this stage of analysis. This stage benefits from ANN to classify cyber-attacks based on their type of activity which has been defined in section 3.4.3. The training set has 2694 records, 10 fold cross validation is the key role in training control and the size of ANN model can be changed from 1 to 7 and the training process is shown figure 4-79.

```
Multi-Layer Perceptron

2694 samples
   5 predictor
   4 classes: 'CC', 'CE', 'CW', 'HA'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2425, 2425, 2423, 2425, 2425, 2426, ...
Resampling results across tuning parameters:

  size  Accuracy   Kappa
  1     0.7724481  0.5615696
  3     0.8121510  0.6612874
  5     0.8107053  0.6576691
  7     0.8091963  0.6568456

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was size = 3.
```

Figure 4-79 Training process of Cyber Attack Activity model based on ANN

The best ANN model with 81.21% accuracy has 3 hidden layers which has been shown in the training process. Figure 4-80 also shows the minimum accuracy of the ANN model occurs when it has only 1 hidden layer and the behaviour of accuracy over the size is not constant and abnormal.

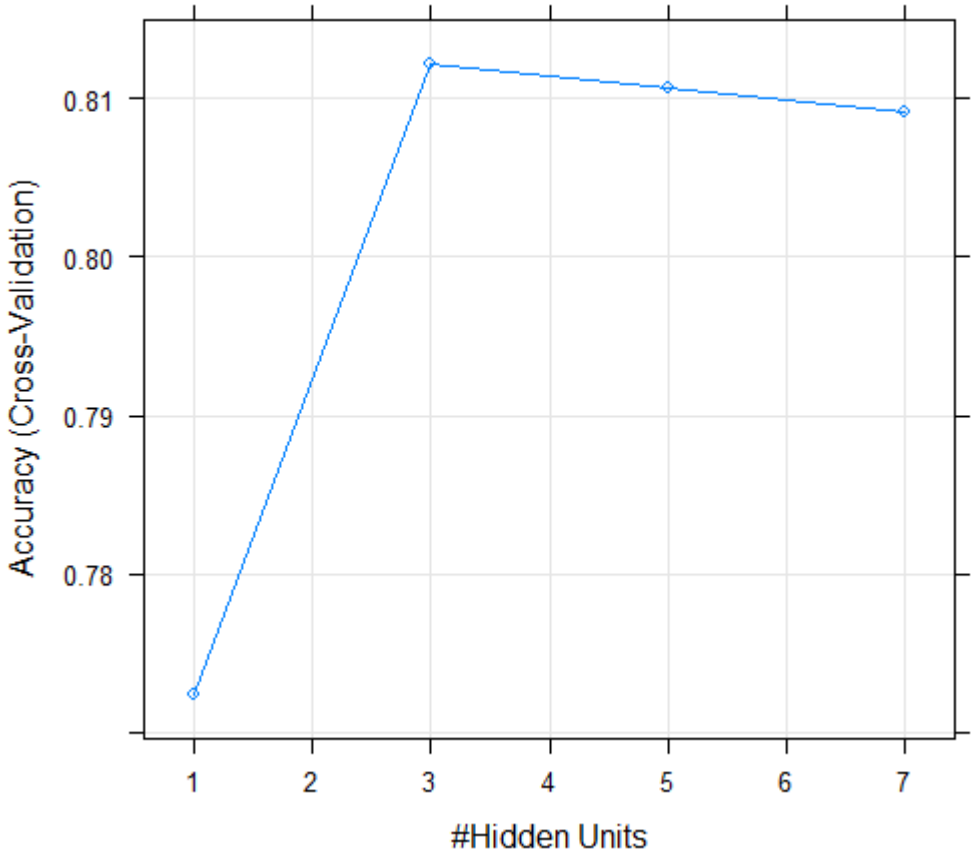

Figure 4-80 Accuracy trend for prediction of type of cyber-attack activity in ANN

4.6.6 Discussion and Interpretation

According to the data analysis by ANN, the optimal models for each feature of cyber-attack have been obtained and the accuracy of them has been shown in table 4-5.

Type of prediction	Accuracy level
Type of Threat	58.86%
Cyber attackers	59.71%
Targeted Country	47.18%
Type of Target	37.26%
Type of activity	81.21%

Table 4-5 Accuracy prediction in ANN

According to the table 4-5 and the data analysis stage the following interpretations and points can be taken into consideration:

1. ANN classifier for prediction of type of cyber-attacks activity is the best and the most consistent among the other models. As it was mentioned before the size of ANN model is 3 which means it has 3 hidden layers and the accuracy and the number of the layers does not directly impact on accuracy of this ANN model. Cyber security experts can be 81.21% assured when they use this model for prediction of cyber-attacks activity which means it can perform as an independent tool.

2. The prediction of the cyber attacker with ANN has 59.71% accuracy and 5 hidden layers. This model has reasonable reliability for cyber experts to identify cyber criminals behind attacks and can be employed along with other approaches for more accurate prediction.

3. Type of Threat prediction by ANN model is the third accurate model compared to others. This model has 7 hidden layers and the process of training indicate the number of hidden layers can have a positive impact on the model accuracy. It is recommended to use this model in combination with other techniques to improve the accuracy of the prediction.

4. The prediction of the Type of Target and victim countries by ANN model has less than 50% accuracy which is not suitable and needs to be improved with other methods or even maybe more factors. The unreliability of these models can be because of two main reasons, firstly there might be a need for more features and factors in the data set to predict Type of Targets and victim countries and secondly, the attackers might choose their targets regardless of their type and the countries that they are located.

4.7 Summary

In this chapter, the methodology of this research has been explained. Data collection, pre-process, structuring and also used tools and platforms have been described. At last data analysis was presented, classification algorithms and how they are applied where the area of concern of this section. For each classification algorithm, a broad interpretation and discussion were carried out. The next chapter will

aim to present the results and compare the model of each classification algorithm to conclude the nominated predictive model for each dimension of a predictive model. For the full scripts and dataset please refer to Appendix.

Chapter 5 Comparison of models and Choosing Optimal framework

5.1 Introduction

This chapter aims to compare all the predictive models trained in chapter 4 and choose the best and most accurate one to forecast and identify future and potential cyber events. The models were built based on main classification techniques and the task was to predict each factor including Type of Threat, Type of Target, targeted country, cyber-attack activity and cyber attacker. This section will compare the models based on two main benchmarks; Accuracy and Kappa. Accuracy is the first benchmark presenting the ability of a classifier model to predict the class label correctly. Kappa is another significant and meaningful benchmark used by data analysts to examine how good the predictive models perform and compare them among themselves. Landis and Koch (1977) suggest that Kappa gives a better insight into a classifier reliability based on observed accuracy and expected accuracy. Observed accuracy is the accuracy obtained from a classifier and expected accuracy is a type of accuracy that any classifier is expected to obtain regarding their confusion matrix. For example, consider the following confusion matrix:

	RED	BLUE
RED	10	7
BLUE	5	8

There are 60 different instances in red and blue colour and the observed accuracy in this confusion matrix is (10+18)/60 = 0.6. For expected accuracy, the marginal frequency of Reds should be multiplied by the number of the instance that machine learning classified as RED and divided by total number of instances which will be (15*7)/30 = 8.5 and the same operation should be done on BLUE and it is equal to (13*15)/30 = 6.5. The final step for expected accuracy is adding these amounts together and dividing them by total number of

instances which will be (6.5+8.5)/30 = 0.5. The kappa equation is as follows (Thompson, 2001):

(OBSERVED ACCURACY- EXPECTED ACCURACY)/ (1-EXPECTED ACCURACY)

According to this formula the Kappa is equal to 0.2. The interpretation of Kappa is the area of different opinions where Fleiss et al. (1969) report kappa more than 0.75 excellent, 0.4 to 0.75 reasonable and less than 0.4 poor and also Landis and Koch (1977) categorize kappa 0 to 0.2 as slight, 0.2 to 0.4 as reasonable, 0.4 to 0.6 average, 0.6 to 0.8 significant and more than 0.8 excellent.

The next sections will investigate which model is suitable for each variable of cyber-attacks by comparing their accuracy and kappa. The Caret package in R gives the ability to compare different models accuracy and kappa with a function named Resample.

5.2 Optimal model for Type of Threat Prediction

7 different models have been trained for prediction of the Type of Threat. These 7 models are; KNN, ANN, Naïve Bayes, Decision tree (including recursive partitioning, c4.5, and random forest) and SVM. After applying resample function to this list of models, the result has been shown in figure 5-1.

```
Accuracy
             Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
KNN        0.5114  0.5175 0.5485 0.5448  0.5670 0.5811    0
ANN        0.5067  0.5842 0.5909 0.5910  0.6099 0.6441    0
Rpart      0.4887  0.5174 0.5573 0.5508  0.5747 0.6147    0
C4.5       0.5381  0.5727 0.5946 0.5960  0.6043 0.6606    0
RandomF    0.5541  0.5590 0.5960 0.5924  0.6200 0.6393    0
SVM        0.5682  0.5973 0.6018 0.6049  0.6144 0.6368    0
Naive_Bayes 0.5135  0.5362 0.5578 0.5588  0.5759 0.6045    0

Kappa
             Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
KNN        0.4164  0.4196 0.4596 0.4543  0.4825 0.4971    0
ANN        0.4043  0.4986 0.5053 0.5070  0.5309 0.5702    0
Rpart      0.3805  0.4191 0.4666 0.4576  0.4846 0.5335    0
C4.5       0.4418  0.4855 0.5112 0.5134  0.5237 0.5912    0
RandomF    0.4630  0.4696 0.5140 0.5089  0.5399 0.5621    0
SVM        0.4843  0.5162 0.5192 0.5250  0.5358 0.5628    0
Naive_Bayes 0.4241  0.4483 0.4719 0.4749  0.4958 0.5295    0
```

Figure 5-1 Comparison of the Type of Threat models in terms of accuracy and kappa

In addition, figure 5-2 demonstrates the dot plot for each model in terms of accuracy and kappa's change.

In terms of accuracy and kappa, SVM has taken the first place with 60.49% accuracy and its kappa is equal to 0.53, which can be interpreted as a reasonably good model for prediction of the Type of Threat. C4.5 as a decision tree model has been occupied the second rank in terms of kappa and accuracy. It has been decided to nominate the SVM model for the most suitable model for the final framework for the prediction of the Type of Threat.

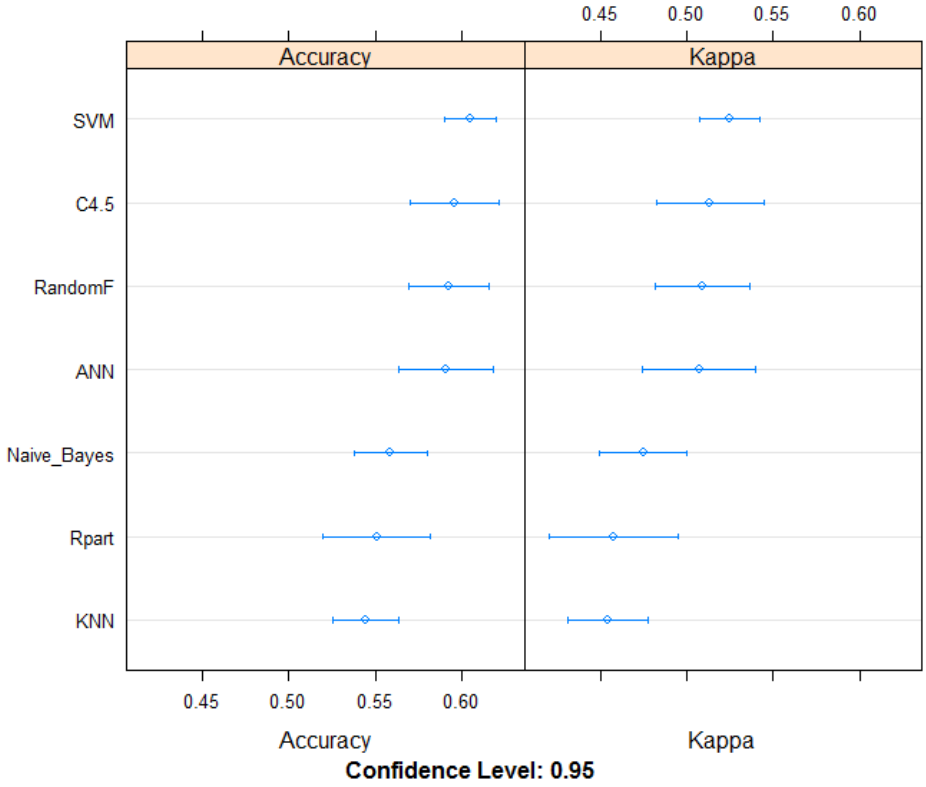

Figure 5-2 Kappa and Accuracy comparison for type of cyber threat prediction

5.3 Optimal model for Prediction of Cyber Attacker

In chapter 4, models were trained by classification techniques for the prediction of future and identification of past cybercriminals in cyber-attacks. In this stage, by comparing the kappa and accuracy of the models, choosing the most significant predictive model will occur. List of the predictive models for cyber attacker will be given to Resample function as input and the result has been created which is shown in figure 5-3.

```
Accuracy
             Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
Naive_Bayes 0.5278  0.5571 0.5804 0.5817  0.6071 0.6197    0
Knearest    0.5694  0.5832 0.6154 0.6127  0.6448 0.6549    0
ANN         0.5532  0.5774 0.6040 0.5971  0.6155 0.6350    0
SVM         0.5548  0.6134 0.6202 0.6134  0.6292 0.6549    0
C45         0.5532  0.5931 0.6211 0.6125  0.6319 0.6503    0
RandomF     0.5448  0.5679 0.5923 0.5951  0.6286 0.6429    0
RPart       0.5524  0.5818 0.6119 0.6048  0.6266 0.6483    0

Kappa
             Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
Naive_Bayes 0.3893  0.4002 0.4355 0.4381  0.4765 0.4818    0
Knearest    0.3741  0.3948 0.4285 0.4327  0.4761 0.4945    0
ANN         0.3380  0.3556 0.3983 0.3910  0.4207 0.4441    0
SVM         0.3288  0.4200 0.4465 0.4299  0.4526 0.4864    0
C45         0.3714  0.4303 0.4542 0.4486  0.4746 0.4923    0
RandomF     0.3652  0.3993 0.4256 0.4330  0.4740 0.5027    0
RPart       0.3357  0.3910 0.4333 0.4205  0.4473 0.4758    0
```

Figure 5-3 Comparison of the Cyber Attacker models in terms of accuracy and kappa

Figure 5-4 shows the comparison of kappa and accuracy of each model. One point that should be mentioned is that accuracy and kappa is not in the same order for models, in other words in terms of accuracy SVM, KNN and C4.5 have been the best models respectively but in terms of kappa C4.5, NB and Random forest models have been the top 3 models. According to this result, it has been decided to nominate 2 preferable models for prediction of cybercriminals; the first one is SVM model with 61.34% accuracy and the kappa of 0.42 and the second of is C4.5 as a decision tree model with 0.44 kappa and the accuracy of 61.25%

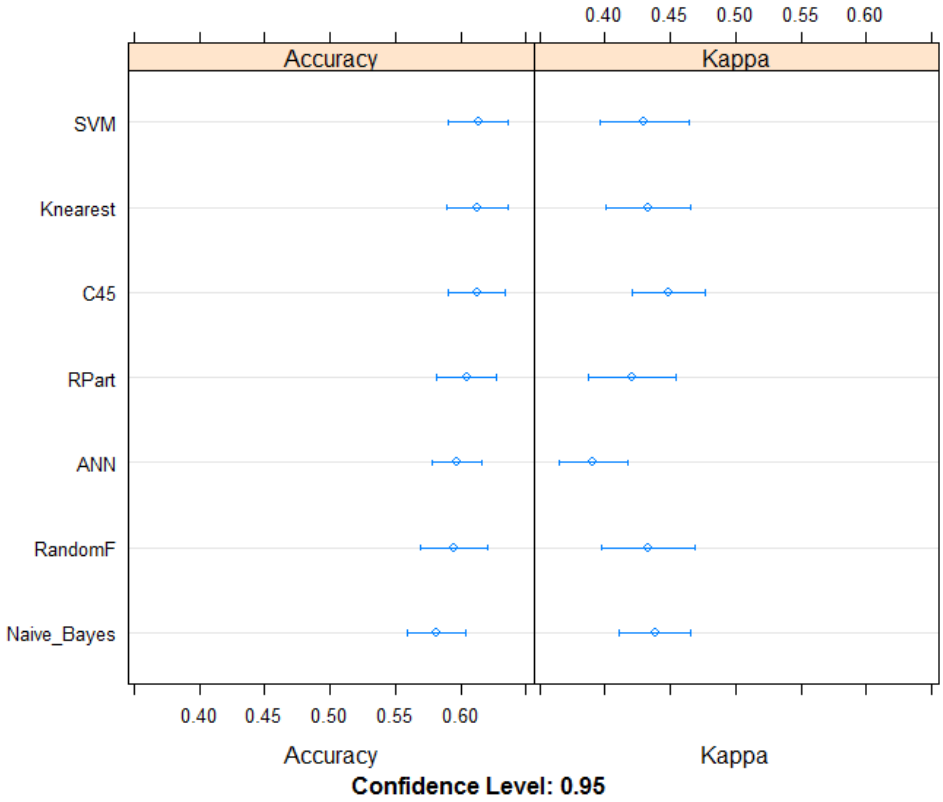

Figure 5-4 Kappa and Accuracy comparison for prediction of cyber attacker

5.4 Optimal model for Prediction of Type of Target

This stage aims to investigate the predictive models for Type of Targets in cyber-attacks in order to find the most accurate and reliable one. Through comparing kappa and accuracy of each model, the investigation will be done and the most desirable model can be identified. Resample function will be applied to those classifier models trained in section 4.5 and figure 5-5 shows the result in terms of average, minimum and maximum of accuracy and kappa of each model.

```
Accuracy
              Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
Naive_Bayes 0.3532  0.3582 0.3796 0.3782  0.3959 0.4059    0
Knearest    0.3259  0.3398 0.3794 0.3739  0.3933 0.4291    0
ANN         0.3569  0.3703 0.3818 0.3793  0.3865 0.3963    0
SVM         0.3680  0.3797 0.3911 0.3969  0.4177 0.4382    0
C45         0.3579  0.3728 0.3816 0.3895  0.4021 0.4440    0
RandomF     0.3670  0.3819 0.3919 0.3912  0.4024 0.4154    0
RPart       0.3408  0.3624 0.3752 0.3760  0.3856 0.4164    0

Kappa
              Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
Naive_Bayes 0.2213  0.2269 0.2595 0.2549  0.2762 0.2905    0
Knearest    0.2098  0.2335 0.2664 0.2639  0.2821 0.3278    0
ANN         0.2410  0.2441 0.2518 0.2544  0.2594 0.2795    0
SVM         0.2486  0.2642 0.2775 0.2831  0.3033 0.3356    0
C45         0.2456  0.2611 0.2729 0.2775  0.2829 0.3379    0
RandomF     0.2331  0.2549 0.2680 0.2648  0.2740 0.2907    0
RPart       0.2183  0.2351 0.2558 0.2565  0.2755 0.2989    0
```

Figure 5-5 Comparison of the Type of Target models in terms of accuracy and kappa

According to the result and figure 5-6 showing the comparison of kappa and accuracy of models, SVM classifier has the highest accuracy with 39.69%, however, kappa of the SVM indicates that the model will be categorized as an insignificant model with 0.28. Although the second best model will be random forest as a decision tree model, its kappa is the third one among other models. After analysing the results, it has been decided to nominate the SVM model for prediction of the Type of Target.

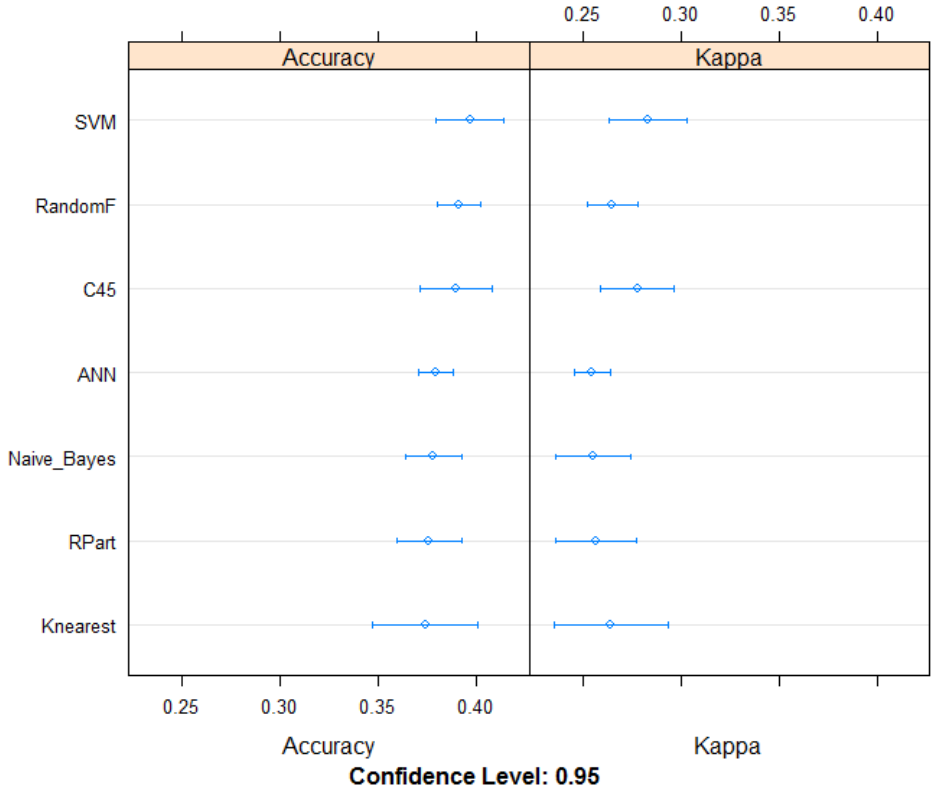

Figure 5-6 Kappa and Accuracy comparison for prediction of Type of Target

5.5 Optimal model for Prediction of Targeted country

Classification algorithms were applied in section 4.5 to train predictive models to identify potential and vulnerable countries against cyber-attacks. This stage aims to identify the best predictive model through comparison of kappa and accuracy of the models. Resample function has been employed and the result has been produced which is shown in figure 5-7.

```
Accuracy
             Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
KNN         0.4607  0.4645 0.4731 0.4710  0.4753 0.4835    0
ANN         0.4440  0.4608 0.4693 0.4718  0.4778 0.5169    0
Rpart       0.4370  0.4553 0.4584 0.4629  0.4739 0.4907    0
C4.5        0.4312  0.4713 0.4815 0.4770  0.4865 0.5038    0
RandomF     0.4419  0.4731 0.4879 0.4837  0.4954 0.5093    0
SVM         0.4627  0.4805 0.4862 0.4896  0.5009 0.5203    0
Naive_Bayes 0.4354  0.4550 0.4668 0.4756  0.5019 0.5243    0

Kappa
             Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
KNN         0.1786  0.2079 0.2166 0.2136  0.2270 0.2328    0
ANN         0.1616  0.1827 0.2183 0.2118  0.2285 0.2900    0
Rpart       0.1679  0.1857 0.2010 0.2024  0.2204 0.2355    0
C4.5        0.1411  0.1991 0.2170 0.2163  0.2401 0.2736    0
RandomF     0.1394  0.1965 0.2090 0.2053  0.2215 0.2435    0
SVM         0.1872  0.2068 0.2147 0.2201  0.2271 0.2728    0
Naive_Bayes 0.1509  0.1837 0.1992 0.2074  0.2415 0.2866    0
```

Figure 5-7 Comparison of the Targeted Countries models in terms of accuracy and kappa

Figure 5-8 shows the comparison of accuracy and kappa of the models. SVM with 48.96% and Random Forest with 48.37% accuracy have been chosen as the first and the second options for prediction of targeted country. Although the accuracy of these models is about to average, kappa indicates they are significantly reliable because the SVM's kappa equals to 0.22 which means it is an unreliable model as a single method for prediction. In conclusion, SVM still needs to be chosen as the best method because compared to other it has better accuracy.

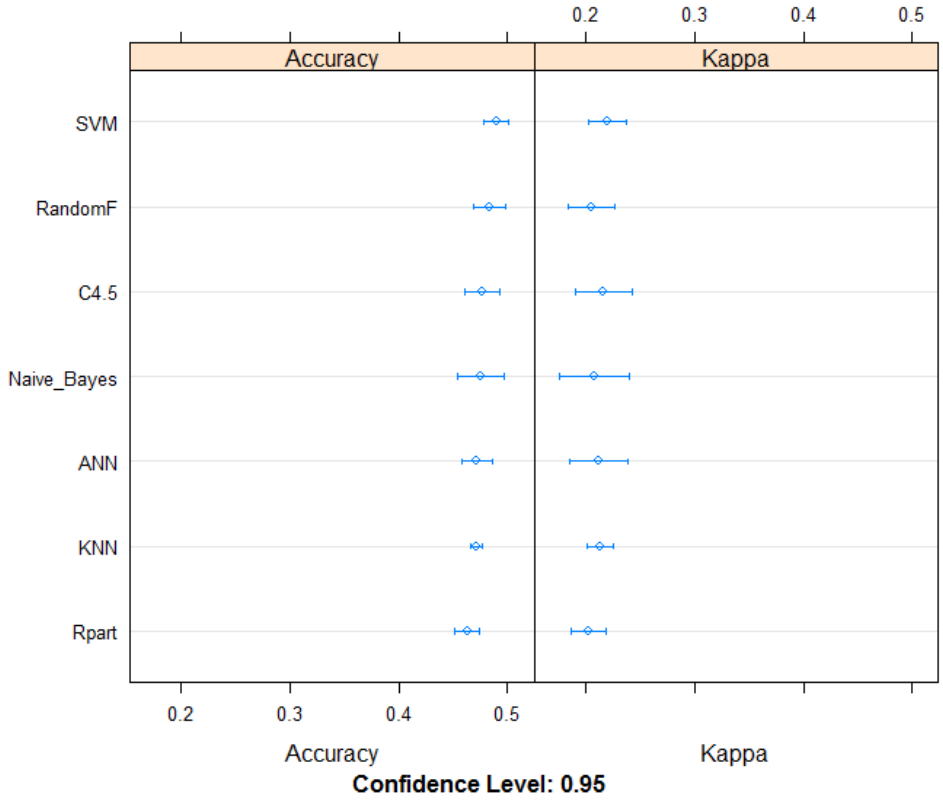

Figure 5-8 Kappa and Accuracy comparison for prediction of targeted country

5.6 Optimal model for Prediction of Cyber Attack Activity

In section 4.5, the classifiers were trained for prediction of cyber-attack activity and in this section models will be compared against each other to nominate the most accurate one in terms of prediction. As it was discussed kappa and accuracy are two main criteria for the comparison, so the list of models will be fed into Resample function to obtain more understandable insight. Figure 5-9 are presenting the result of this function which is in form of maximum, minimum and average of kappa and accuracy for each model.

```
Accuracy
             Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
KNN         0.7704  0.8030 0.8056 0.8089  0.8216 0.8290    0
ANN         0.7815  0.7928 0.8070 0.8122  0.8282 0.8587    0
Rpart       0.7732  0.7892 0.8074 0.8066  0.8155 0.8476    0
C4.5        0.7724  0.7859 0.8000 0.8055  0.8306 0.8433    0
RandomF     0.7844  0.8130 0.8249 0.8248  0.8405 0.8513    0
SVM         0.7955  0.8113 0.8212 0.8256  0.8347 0.8810    0
Naive_Bayes 0.7807  0.7970 0.8182 0.8192  0.8382 0.8662    0

Kappa
             Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
KNN         0.5749  0.6340 0.6490 0.6473  0.6706 0.6868    0
ANN         0.6147  0.6301 0.6465 0.6613  0.6881 0.7455    0
Rpart       0.5845  0.6093 0.6452 0.6449  0.6652 0.7261    0
C4.5        0.5838  0.6094 0.6327 0.6464  0.6914 0.7186    0
RandomF     0.6076  0.6608 0.6806 0.6799  0.7076 0.7309    0
SVM         0.6260  0.6531 0.6714 0.6806  0.6986 0.7812    0
Naive_Bayes 0.6030  0.6311 0.6700 0.6710  0.7037 0.7524    0
```

Figure 5-9 Comparison of the Cyber Attack Activity models in terms of accuracy and kappa

The plot of accuracy and kappa of predictive models for cyber-attack activity has been shown in figure 5-10. As it is presented in the result of Resample function and in figure 5-9, the SVM model has the highest kappa and accuracy so it will be nominated with 82.56% accuracy and 0.68 kappa which means it is a substantial predictive model. Random forest and Naïve Bayes models have minor difference with the SVM model and occupied the second and the third best models respectively.

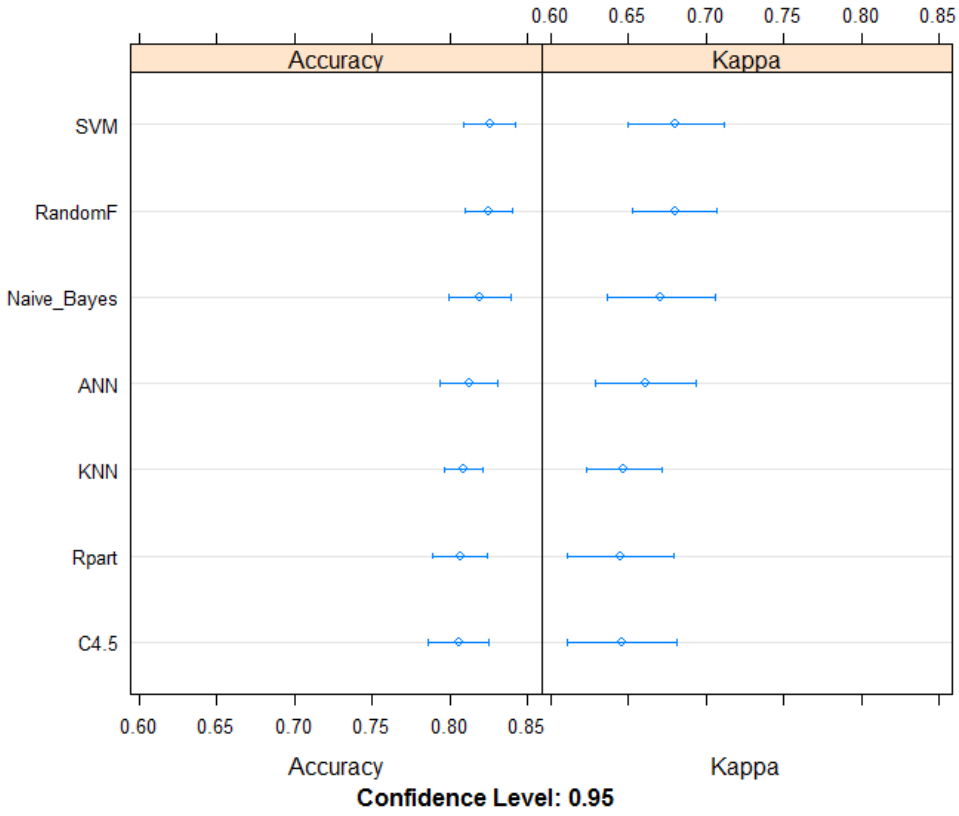

Figure 5-10 Kappa and Accuracy comparison for prediction of cyber-attack activity

5.7 The Final Model

After all the comparison in this chapter, a 5-dimensional model will be concluded which includes 5 different models which can predict 5 different features of cyber-attacks and it can help cyber security experts to evaluate the cyber situational awareness and plan their different strategy for tackling cyber breaches.

By comparison different classification algorithms for prediction purposes, we found out Support Vector Machine has the highest accuracy in prediction of all 5 different features of cyber-attacks. Therefore, it will be chosen as the most desirable algorithm for the 5-dimensional model. Table 5-1 shows the distillation of the accuracy of each classification algorithms in the prediction of each cyber-attack feature.

Algorithm/Dimension	Cyber Attack Activity	Cyber Attacker	Type of Threat	Type of Target	Targeted Country
ANN	81.22%	59.71%	59.10%	37.93%	47.18%
KNN	80.90%	61.27%	54.48%	37.39%	47.10%
NB	81.92%	58.17%	55.88%	37.39%	47.56%
SVM	82.56%	61.34%	60.49%	39.69%	48.96%
Rpart	80.66%	60.48%	55.08%	37.60%	46.29%
Random Forest	82.48%	59.51%	59.24%	39.12%	48.37%
C4.5	80.55%	61.25%	59.60%	38.95	47.70%

Table 5-1 accuracy level for each dimension of cyber attacks

As it is shown in the table Support Vector Machine has done more accurate job in prediction of all features of cyber-attacks, therefore our obtained model will be built based on SVM classification algorithm. In table 5, those cells filled with green, yellow and red colour show the first, second and third best predictive classification technique for each feature. In the next chapter, the model will be evaluated and validated through applying the unseen data of cyber incidents in order to discover its success in prediction of future cyber-attacks.

5.8 Summary

This chapter aimed to present the result by comparing the models' accuracy and obtain the most sufficient and accurate model. Kappa and Accuracy measure was taken into consideration and they were benchmarks for the comparison purpose. The final predictive model is based on Support Vector Machine and in the next chapter, this model will be evaluated by applying unseen data. Also, in the next chapter, variable importance in the final predictive model will be discussed and interpreted. For the full scripts and dataset please refer to Appendix.

Chapter 6 Discussion and Evaluation of the predictive model

6.1 Introduction

In this chapter, an extensive discussion will be taken place about 2 main important subjects. The first part of this chapter, Feature importance will be estimated and discussed in order to investigate which features have more or less contribution to the accuracy of prediction, therefore, the aim of this investigation is to estimate the pattern and relationship between each factor in a cyber-attack. The second part of this chapter aims to evaluate the accuracy of the model when an unseen data will be applied to it. Different benchmarks will be

considered to see how successful the obtained model can perform in terms of prediction of different elements of future cyber-attacks.

6.2 Feature Importance in the prediction

In this section, the variable importance will be investigated to see how much effect each variable had in our predictive framework. This investigation can indicate which factor plays the most and least important role in terms of prediction of different features of cyber-attacks. It is highly important for cyber security experts to find out and discuss which factors have more and less effect on the prediction of different elements of cyber-attacks.

In order to carry out this investigation, four different methods have been used which they are embedded in WEKA for determining the importance of variables in classification. The first method is called Information Gainer which estimates the information gain on the output class by each variable regardless of the classification technique. Every variable will get a value between 0 indicating they have on effect and 1 showing that they have the maximum effect on the output variable. Rajpal et al.(2016) describe Info Gainer Evaluator as a method to measure the informational gain value of each variable regarding the output class. The second one called Correlation Attribute Evaluator which estimates the correlation between input variables and output variable based on the Pearson Correlation Coefficient. Sedgwick (2012) explains that the Pearson Correlation Coefficient is a method to estimate the type of linear relationship between 2 variables and in this research by applying this method each variable will get a value between -1 to 1 so this value indicates if they have negative linear, neutral or positive linear influence on the class output. The third method has been used in this investigation is learner based feature importance called Classifier Attribute Evaluator which is based on the type of classification technique that has been used to build up the predictive model. Like previous methods, each attribute will get a value from -1 to 1 showing their importance in the prediction of the output.

Finally, the last method used in this investigation is used is Wrapper Subset Evaluator (Kohavi and John, 1997). This technique will demonstrate that which variable has the most effect on the accuracy of the prediction and which variables can be considered irrelevant and can be removed.

The nominated predictive model has 5 different dimensions and it is built based on Support Vector Machine. In the next subsections, the importance of each feature on each dimension of this predictive model will be discussed by applying the above methods.

6.2.1 Variable importance in Type of Threats prediction

In order to measure the importance of variables in the prediction of type of cyber threats, 4 methods which explained above will be applied. Table 6-1 shows the result and figure 6-1 demonstrates it a bar chart.

	Activity	Cyber Attacker	Type of Target	Targeted Country
InfoGainer	0.481	1.185	0.465	0.484
Correlation	0.1758	0.0997	0.0829	0.063
ClassifierEval	0.1148	0.1982	0.1602	0.0796
Wrapper	Yes	Yes	Yes	Yes

Table 6-1 variable importance chart in prediction of Type of Threat

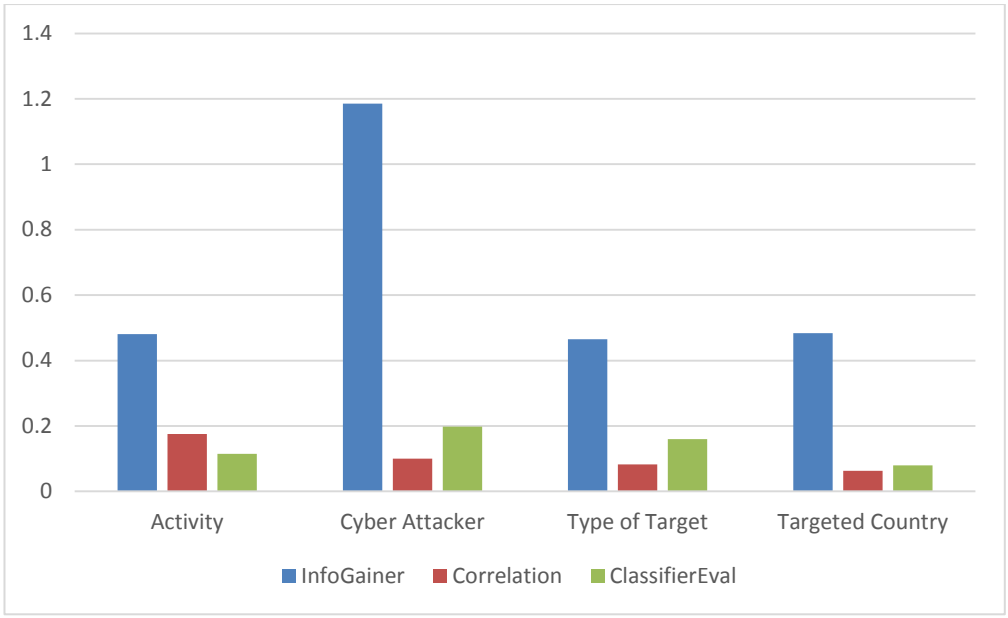

Figure 6-1 variable importance bar chart in prediction of type of cyber threat

Following points can be mentioned regarding to feature importance in prediction of type of cyber threats:

1. Infogain method measures that cyber attacker attribute has the most effect on the class of type of cyber threat. This means changing the value of cyber attacker attribute can change the value of the type of cyber threat. The other attributes have almost equal effect on the type of cyber threat value.

2. By applying CorrelationEval, it has been found out that all of the features have almost a value close to 0 indicating that they have a nonlinear relationship with the type of cyber threat.

3. ClassifierEval method suggested that like Infogain the cyber attacker feature has the most effect on the prediction of type of cyber threat based on the nominated predictive model built by Support Vector Machine. The Type of Target attribute has the second most effect on the result of the prediction of type of cyber threat. In addition, WrapperEval indicate that all the features should remain in the training dataset for SVM classifier.

As the variable importance investigation in the prediction of cyber threats shows cyber attackers have the most influence on this process. This means more information on cyber attackers can lead to better understating of what type of cyber threat they can pose in the future. On the other hand, the targeted countries have the least contribution in

the process of predication, therefore, all countries in the training dataset can be vulnerable to all sort of cyber threats mentioned in the training dataset.

6.2.2 Variable importance in cyber attackers prediction

Variable importance in the prediction of cyber attackers is measured and Table 6-2 and figure 6-2 demonstrate the result of applying 4 main algorithms to the training dataset.

	Activity	Threat	Type of Target	Targeted Country
InfoGainer	0.441	0.763	0.712	0.976
Correlation	0.2167	0.1133	0.0679	0.654
ClassifierEval.	0	0.05537	0.282	0
Wrapper	Yes	Yes	Yes	Yes

Table 6-2 variable importance chart in prediction of cyber attackers

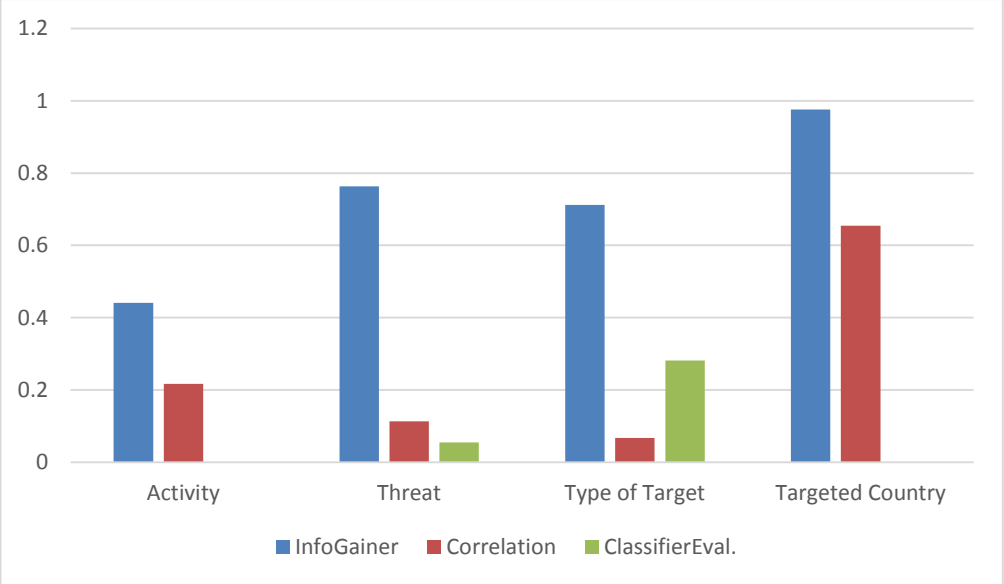

Figure 6-2 variable importance bar chart in prediction of cyber attacker

According to the result, the following points need to be highlighted:

1. InfoGain technique indicates that cyber attacker variable mostly influenced by targeted country which can be interpreted that different countries are targeted by specific cyber attackers. Another important point that can be concluded is that cyber-attack activity has the least effect on cyber attackers' class which means cyber attackers constantly change their motivations and there is no consistency in their intention.

2. Apart from the targeted country feature, all the other variables have almost nonlinear relationship with cyber attacker class according to CorrelationEval method. Type of Target with 0.65 is close to having a linear relationship with cyber attacker class.

3. According to ClassifierEval, considering the nominated predictive model based on SVM, Type of Target and Type of Threat have the most contribution to the predictive model accuracy. In addition, all the features in the training dataset need to be kept according to WrapperEval result.

According to the result of the variable importance for prediction of cyber attackers, although targeted country has the most influence on cyber attacker when SVM algorithm comes to the prediction Type of Threat and Type of Target play crucial role in the process of prediction.

This means Type of Target and type of cyber threat have been considered more important vectors in classifying cyber attackers.

6.2.3 Variable importance in Type of Target prediction

Table 6-3 and figure 6-3 Show the result of the measuring variable importance in prediction of Type of Target.

	Attacker	Country	Type of Threat	Activity
Infogainer	1.162	0.563	0.41	0.263
Correlation	0.0623	0.0577	0.0669	0.1524
ClassifierEval	0.06347	0.00408	0.04139	0.02895
Wrapper	Yes	No	Yes	Yes

Table 6-3 variable importance chart in prediction of Type of Target

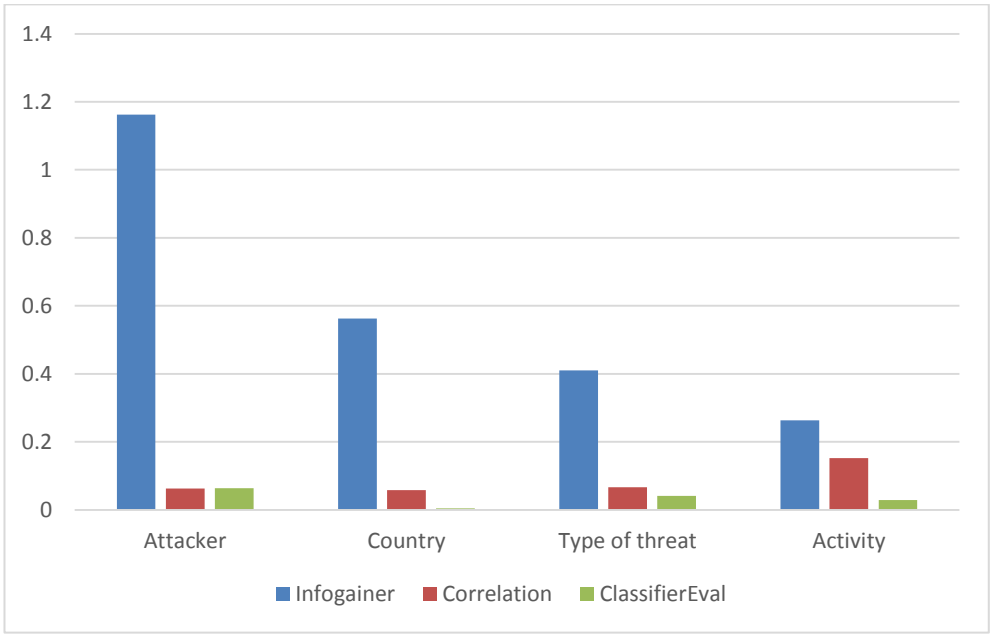

Figure 6-3 variable importance bar chart in prediction of Type of Target

The following points can be mentioned regarding to the table 6-3 and figure 6-3:

1. According to InfoGain, the cyber attacker variable has the most effect on Type of Target, which can be interpreted that changing value of cyber attacker attribute will highly change the value of Type of Target.

2. CorrelationEval indicates that all the features have nonlinear relationship with Type of Target.

3. Cyber attacker and Type of Threat have the main contribution to build the nominated predictive model based on SVM and targeted country has almost zero effect in prediction of Type of Target. WrapperEval suggests that targeted country can be removed from the dataset in order to build more classifier.

As the result of the variable importance, investigation shows the cyber attacker has the most impact on the value of Type of Target without applying any learner algorithm. On the other hand, after applying SVM algorithm to the prediction process, all the factors play equal roles and targeted country can be removed. In order to see how much effect the elimination of targeted country has on the prediction process, this factor will be removed and the predictor will be trained again with the reduced training dataset. The reduced training dataset now has 4 different attributes; Activity, Cyber attacker, Type of Threat and Type of Target. After training the predictive model with SVM algorithm, figure 6-4 shows the training process.

```
Support Vector Machines with Linear Kernel

2694 samples
   3 predictor
  19 classes: 'BP', 'ED', 'EN', 'ES', 'FB', 'GO', 'HC', 'HT', 'IO', 'MD', 'MU',
 'NN', 'RT', 'SI', 'SN', 'TC', 'THS', 'TM', 'TP'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2425, 2424, 2426, 2425, 2424, 2425, ...
Resampling results across tuning parameters:

  cost  Accuracy   Kappa
  0.1   0.3548724  0.2161548
  0.2   0.3797660  0.2554926
  0.3   0.3845891  0.2617549
  0.4   0.3931297  0.2726461
  0.5   0.3905275  0.2709060
  0.6   0.3949844  0.2766345
  0.7   0.3983301  0.2814016
  0.8   0.3998144  0.2833085
  0.9   0.4009227  0.2849955
  1.0   0.4001614  0.2855020

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was cost = 0.9.
```

Figure 6-4 Predictive model for Type of Target after applying Wrapper subset

As it is shown the accuracy of the predictive model has been improved by almost 1 percent. Although it cannot be significant enough, still a slight improvement in some cases can be very important.

6.2.4 Variable importance in Targeted country prediction

4 main methods will be used in this section to estimate variable importance in the prediction of the targeted country. Table 6-4 and figure 6-5 demonstrate the result of the estimation.

	Cyber Attacker	Activity	Type of Threat	Type of Target
Infogainer	1.1691	0.242	0.484	0.563
Correlation	0.0622	0.0916	0.0535	0.0577
ClassifierEval	0.009651	0.000371	0	0.0577
Wrapper	Yes	Yes	Yes	Yes

Table 6-4 variable importance chart in prediction of targeted country

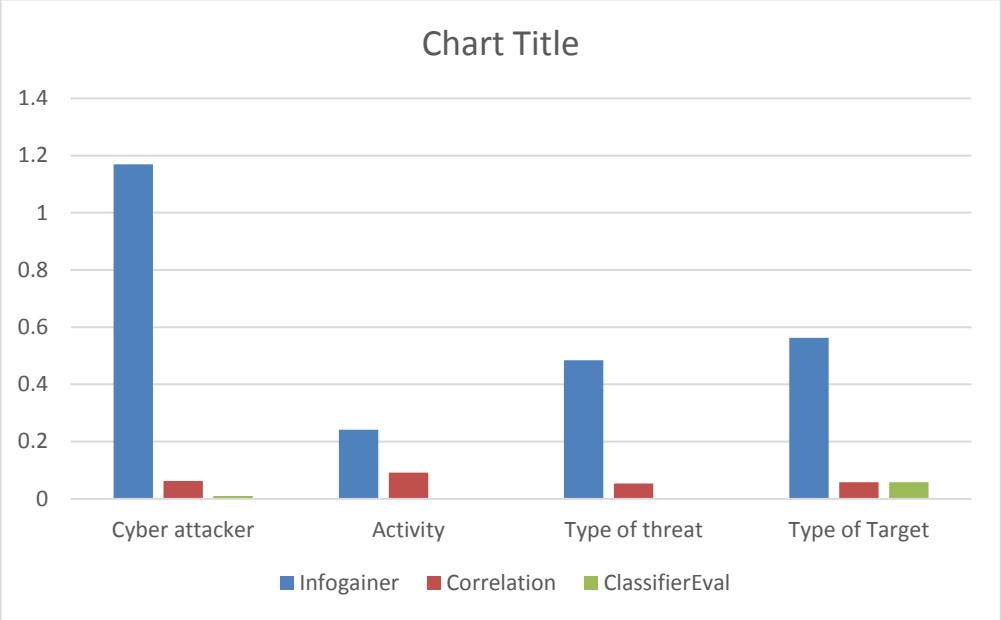

Figure 6-5 variable importance bar chart in prediction of targeted country

The following points are obtained based on the estimation:

1. Targeted country class can be changed highly by cyber attacker value based on InfoGain method, which means different cyber attackers target their victim in specific country. After cyber attacker, Type of Target variable has the most influence on the value of the targeted countries, this can be interpreted that each country such as political, economy and so on is hosting different businesses or targets.

2. There is a nonlinear relationship between different factors according to CorrelationEval technique.

3. In the nominated predictive model, all the features have the equal contribution to building up the classifier based on SVM and as WrapperEval method indicates all the attributes need to remain the training dataset.

According to the investigation of variable importance in prediction of targeted country, cyber attacker has the most influence on the variable of targeted country. This means without applying a learner-based algorithm in prediction of targeted country, cyber attacker variable plays the main role, however, after applying SVM method in prediction all of the variables will contribute equally in prediction of targeted country which means there is no significant factor in the prediction process.

6.2.5 Variable importance in Cyber Attack Activity Prediction

Variable importance in prediction of type of cyber-attack activity is measured by using 4 main methods and the result is shown in table 6-5 and figure 6-6.

	Cyber attacker	Targeted Country	Type of Target	Type of Threat
InfoGainer	0.424	0.242	0.263	0.703
Correlation	0.232	0.0916	0.1524	0.1561
ClassifierEval	0.144	0.046	0.1101	0.195
Wrapper	Yes	Yes	Yes	Yes

Table 6-5 variable importance chart in prediction of cyber-attack activity

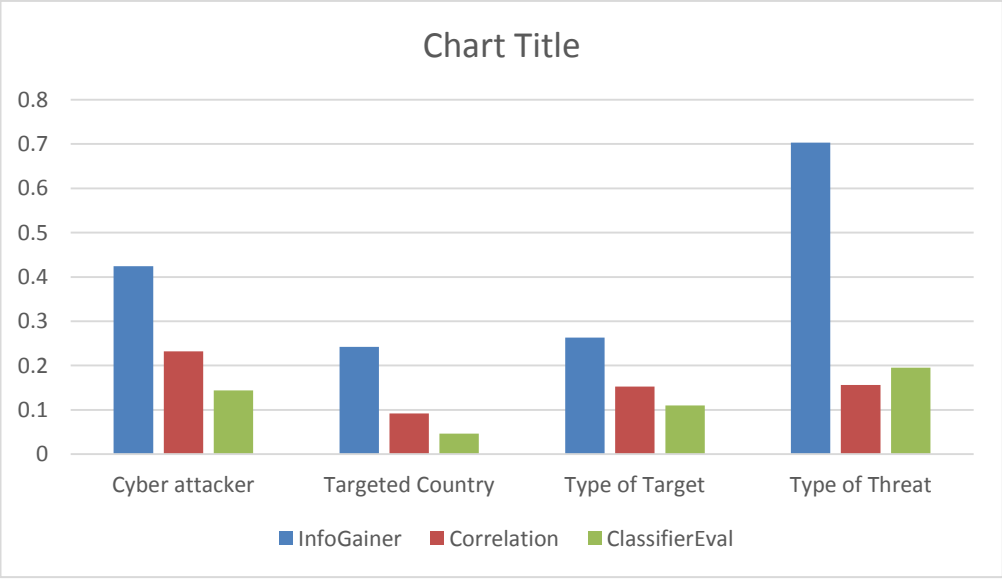

Figure 6-6 variable importance bar chart in prediction of cyber-attack activity

The following conclusions have been made by interpreting from the result:

1. InfoGain indicates, type of cyber-attack activity has been mostly influenced by cyber attacker and Type of Threat attribute, however, all the other attributes contribute to the value of cyber activity fairly equally. This can lead to a conclusion describing a relationship between Type of Threat and cyber-attack activity.

2. According to CorrelationEval technique, there is a nonlinear relationship between all the features and cyber-attack activity.

3. ClassifierEval method indicates that Type of Threat and cyber attacker have the most effect on prediction of cyber-attack activity based on SVM. This means like InfoGain method the type of cyber-attack activity can highly depend on these two features also all the features need to be considered in building in predictive model based on SVM.

Variable importance investigation indicates they cyber-attack activity highly depends on Type of Threat and cyber attacker therefore based who was behind the cyber-attack and what Type of Threat they posed, the type of cyber-attack motivation or activity can be determined. After applying SVM algorithm to the classification process it also shows cyber attacker and Type of Threat play the main roles in the prediction of cyber-attack activity.

6.3 Evaluation of the model

In order to measure the success of the obtained model, a validation dataset will be provided to see how the predictive model will be performed to predict the future cyber attacks' factors and features. The validation data set includes cyber-attacks happened from 2016 to the end of March 2017. The same pre-process method, which was explained in section 4.4, will be applied to the validation data set. After applying the pre-processing stages, 1137 records remain in the data

set in order to apply the predictive models to them for measuring success and accuracy of future attacks prediction.

In the next sections each predictive model will be applied to relevant validation dataset and for measuring success and analyse the accuracy of the following criteria will be considered (Fawcett, 2006):

1. True Positive rate (Recall): It refers to number of instances which are classified correctly in one class, divided by total correctly classified instances and incorrectly declassified which is formulated as: TP rate = TP/(TP+FN)

2. False Positive rate: It is also called false alarm rate, explaining the number incorrectly classified of one class over the total number incorrectly classified instances. The FP rate equation is : FP rate = FP/(FP+TN)

3. Precision: It is also named as Positive Predicted Value which describes the number of instances classified correctly in one class, divided by total number of classified objects in that class. Precision is calculated as : Precision = TP/TP+FP

4. ROC area: Roc explains a two-dimensional graph where X-axis is labelled as False Positive rate and Y-axis is named as True positive rate. The Area under Roc curve is important for measuring the success of a classifier.

5. F-measure: F-measure is another metric that researchers use for evaluating the accuracy of predictive models. F-measure or F-Score is defined based on precision and recall. F- measure is being widely used in information retrieval. The highlight point about this metric is, it does not take True Negative into consideration which in most of cases it is ignored. (Hripcsak and Rothschild, 2005). F-score is formulated as:

$$2 \times \frac{Precision \times Recall}{Precision + Recall}$$

6.3.1 Validation of the Type of Threat predictive model

In this stage, the predictive model for Type of Threat will be examined by applying it to the validation data set. In validation dataset for cyber threats, there are 753 records among 1137 where their cyber threats are identified. The predictive model will be applied and the result has been shown in figure 6-7.

```
Overall Statistics

               Accuracy : 0.1129
                 95% CI : (0.0912, 0.1377)
    No Information Rate : 0.2324
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0607
 Mcnemar's Test P-Value : NA

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC
Area  PRC Area  Class
                 0.056    0.017    0.389      0.056    0.099      0.095    0.71
4     0.263     DS
                 0.012    0.188    0.008      0.012    0.009      -0.148   0.30
0     0.083     SQ
                 0.857    0.551    0.014      0.857    0.028      0.059    0.68
5     0.015     Other
                 0.234    0.022    0.759      0.234    0.358      0.347    0.61
2     0.396     AH
                 0.638    0.143    0.229      0.638    0.337      0.316    0.79
4     0.183     DF
                 0.000    0.000    0.000      0.000    0.000      0.000    0.68
1     0.323     MWV
                 0.000    0.007    0.000      0.000    0.000      -0.007   0.48
6     0.013     DH
                 0.000    0.000    0.000      0.000    0.000      0.000    0.48
2     0.208     TA
                 0.000    0.000    0.000      0.000    0.000      0.000    0.20
9     0.004     CSS
                 0.000    0.000    0.000      0.000    0.000      0.000    0.36
5     0.004     ZD
                 0.000    0.001    0.000      0.000    0.000      -0.003   0.46
6     0.007     UA
Weighted Avg.    0.113    0.043    0.256      0.113    0.122      0.100    0.59
1     0.264
```

Figure 6-7 Validation of Type of Threat Predictive model

As validation result has been shown, the overall accuracy of the model on the validation data set is 11.30%. Also average precision and ROC area get 0.25 and 0.591 respectively. Although these benchmarks do not present an accurate and strong predictive model, by analysing the results deeper, some highlight points can be concluded:

1. As figure 6-8 shows the comparison of TP, FP, and precision for different classes In terms of TP rate (Recall), the Other class has the most portion of the cyber threats which means the predictive model will be more accurate on less frequent and less known attacks with 85.71%. The second and the third rank belong to Defacement attacks and Account Hijacking in terms of recall or TP rate. Considering TP rate by itself does not have a strong indication of significant prediction, therefore, FP rate should be taken into consideration because sometimes the model can have high TP rate and high FP rate at the same time which is not desirable. In terms of precision, the predictive model has done a significant task in AH attacks and then the DS attacks prediction is the second Type of Threat in terms of the success of getting predicted.

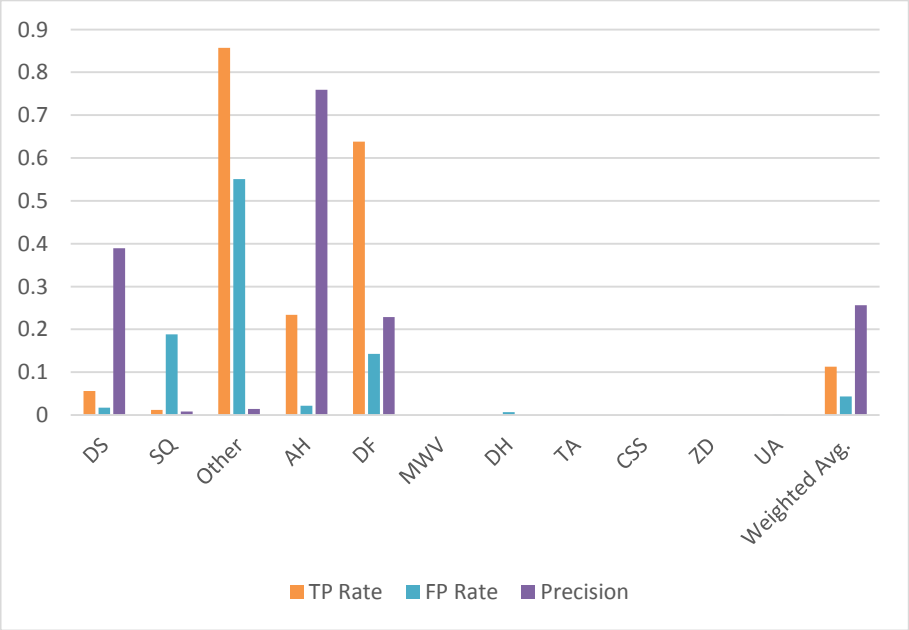

Figure 6-8 TP, FP, and Precision for prediction of type of cyber threat

2. As it was mentioned before, ROC is another benchmark for measuring and comparing of the accuracy of predictive models. As ROC amount was shown in the result and the plot in figure 6-8 representing a bar chart for each class, DF threats have the highest ROC and CSS has the least ROC. The reason for the highest ROC for DF is because the TP rate is the highest compared to its FP rate. As it was mentioned in the previous section about the desirable level of ROC, although the overall

ROC for the model is not significant, for prediction of specific cyber threats the ROC is acceptable such as DF and DS threats with 0.794 and 0.714 respectively.

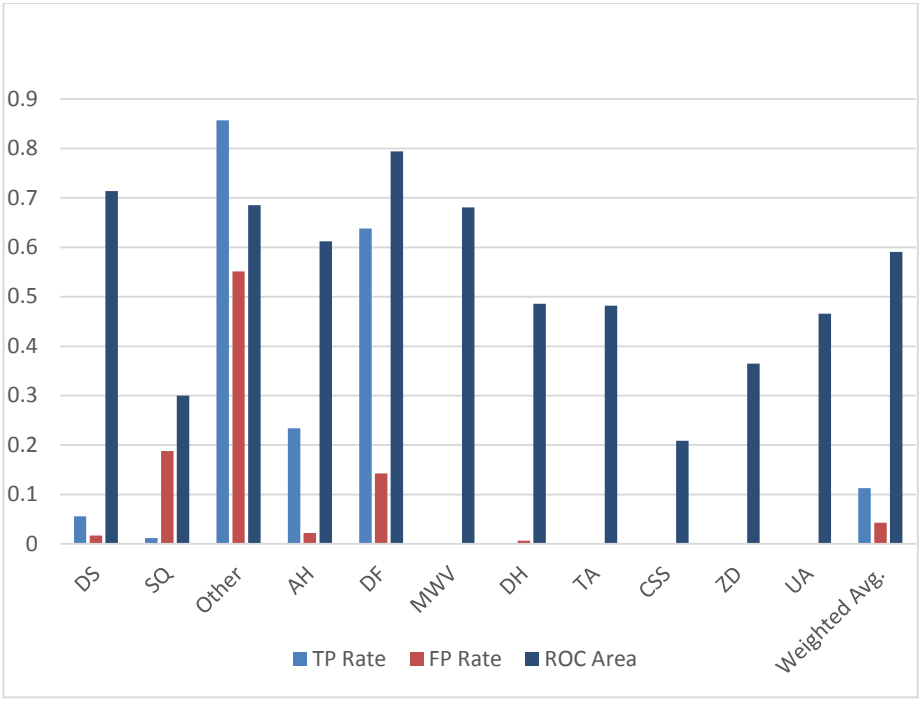

Figure 6-9 TP, FP, and ROC for prediction of the type of cyber threat

3. The third stage of the analysis of validation is comparing the recall, precision, and ROC of the Type of Threat predictive model. The interesting point is DF as the highest ROC has the second most Recall and the third highest precision which indicates the combination of high recall and high precision can be very effective in measuring the success of the predictive model. DS has the second maximum of ROC area with the second highest precision and the fourth highest recall which also indicates that the amount of precision has more effect on ROC. Figure 6-9 shows the comparison of Recall, precision and ROC area for each class.
4. The next step of validation process is dealing with F Score. The overall F-score is 0.122 which is not significant, and it shows it is not reliable as a standalone component for the prediction. The highest F-Score goes to Defacement and Account Hijacking which means the predictive model has performed well in identifying these attacks.

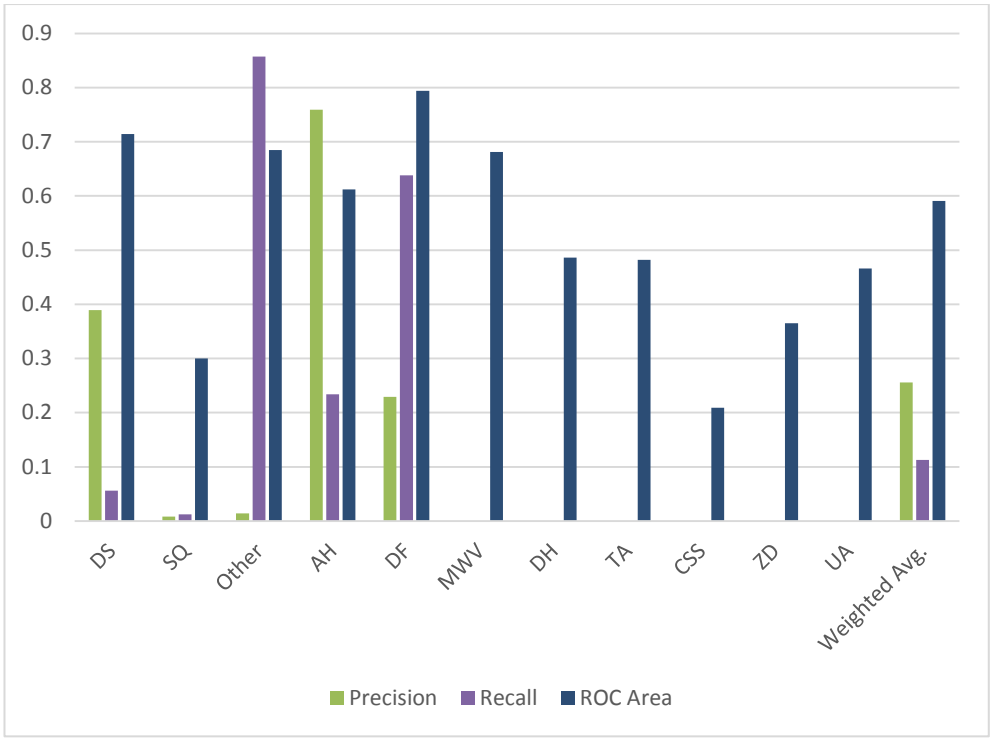

Figure 6-10 Precision, Recall, and ROC for prediction of cyber threat

6.3.2 Validation of the Cyber Attacker predictive model

This stage of validation aims to evaluate the predictive model for cyber attackers. This process will be done by applying the model to the validation data set including 484 cyber-attacks which cyber attackers were known. Figure 6-11 explains the result of the evaluation process:

```
Overall Statistics

               Accuracy : 0.2592
                 95% CI : (0.0758, 0.1316)
    No Information Rate : 0.4855
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1747
 Mcnemar's Test P-Value : NA


    TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Are
a  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    ?
      ?
Dr.SHA6H
                 0.000    0.000    0.000      0.000    0.000      0.000    0.43
8     0.008      Izz.ad.Din.al.Qassam.Cyber.Fighters
                 0.000    0.000    0.000      0.000    0.000      0.000    ?
      ?          DarkWeb.Goons
                 0.000    0.603    0.000      0.000    0.000      0.000    ?
      ?          DERP
                 0.000    0.000    0.000      0.000    0.000      0.000    ?
      ?          CyberBerkut
                 0.538    0.045    0.467      0.538    0.500      0.462    0.84
6     0.126      Chinese.hackers
                 0.627    0.074    0.824      0.627    0.712      0.596    0.89
7     0.464      Anonymous
                 0.000    0.000    0.000      0.000    0.000      0.000    0.55
8     0.032      Guccifer
                 0.000    0.000    0.000      0.000    0.000      0.000    ?
      ?          Maxney
                 0.000    0.000    0.000      0.000    0.000      0.000    ?
      ?          NetPirates
                 0.000    0.000    0.000      0.000    0.000      0.000    0.69
9     0.009      Tunisian.Cyber.Army
                 0.000    0.048    0.000      0.000    0.000     -0.023    0.81
2     0.020      LizardSquad
                 0.000    0.000    0.000      0.000    0.000      0.000    ?
      ?          LulzSec
                 0.000    0.000    0.000      0.000    0.000      0.000    ?
      ?          X.th3inf1d3l
                 0.000    0.000    0.000      0.000    0.000      0.000    0.62
7     0.072      AnonGhost
                 0.000    0.000    0.000      0.000    0.000      0.000    ?
      ?          XTrR3vOlT
                 0.000    0.000    0.000      0.000    0.000      0.000    0.34
2     0.019      HAXOR
                 0.000    0.000    0.000      0.000    0.000      0.000    ?
      ?          TEAM.MADLEETS
                 0.000    0.000    0.000      0.000    0.000      0.000    0.78
9     0.061      Cyber.Islamic.State
                 0.000    0.000    0.000      0.000    0.000      0.000    ?
      ?          Rex.Mundi
                 0.000    0.000    0.000      0.000    0.000      0.000    0.94
5     0.050      Turkish.Ajan
                 0.000    0.000    0.000      0.000    0.000      0.000    ?
      ?          Other
                 0.000    0.000    0.000      0.000    0.000      0.000    ?
      ?          NullCrew
                 0.000    0.000    0.000      0.000    0.000      0.000    0.38
6     0.021      Ag3nt47
                 0.000    0.000    0.000      0.000    0.000      0.000    ?
      ?          JokerCracker
                 0.000    0.000    0.000      0.000    0.000      0.000    0.60
7     0.044      Syrian.Electronic.Army
                 0.000    0.000    0.000      0.000    0.000      0.000    0.46
2     0.012      X.smitt3nz
                 0.000    0.000    0.000      0.000    0.000      0.000    0.56
1     0.010      Iranian.Hackers
                 0.000    0.000    0.000      0.000    0.000      0.000    ?
      ?          KelvinSecTeam
                 0.000    0.000    0.000      0.000    0.000      0.000    ?
      ?          Armada.Collective
                 0.000    0.000    0.000      0.000    0.000      0.000    0.32
3     0.028      RedHack
                 0.000    0.000    0.000      0.000    0.000      0.000    ?
      ?          TeamBerserk
Weighted Avg.    0.259    0.030    0.324      0.259    0.287      0.243    0.68
4     0.195
```

Figure 6-11 Validation of Cyber Attackers Predictive model

The overall accuracy of this model is 25.92% which is not reliable enough for cyber experts as a single predictive method, however, by analysing this result by comparing different benchmarks following points can be highlighted:

1. It should be mentioned that there are some attackers who were not available in validation dataset which means some attackers are not active in recent years or they might have changed their alias names. Figure 6-12 shows the bar plot of the comparison of TP rate, FP rate and precision for those classes that they are available in both the training set and the validation set. The highest TP rate which is 0.627 belongs to the Anonymous hacker group which shows the model despite its insufficient accuracy level, it has done reliable prediction about this cyber attacker group which can be concluded with the contribution of significant level precision which is 0.824. The Chinese hackers are in the second rank in terms of TP rate with the level of 0.538 which also demonstrates accurate task of the predictive model in terms of detection of these cyber attackers. The highest FP rate goes to DERP group which shows in the validation set there are some hackers that they have a similar pattern of feature but they are not in DERP group.

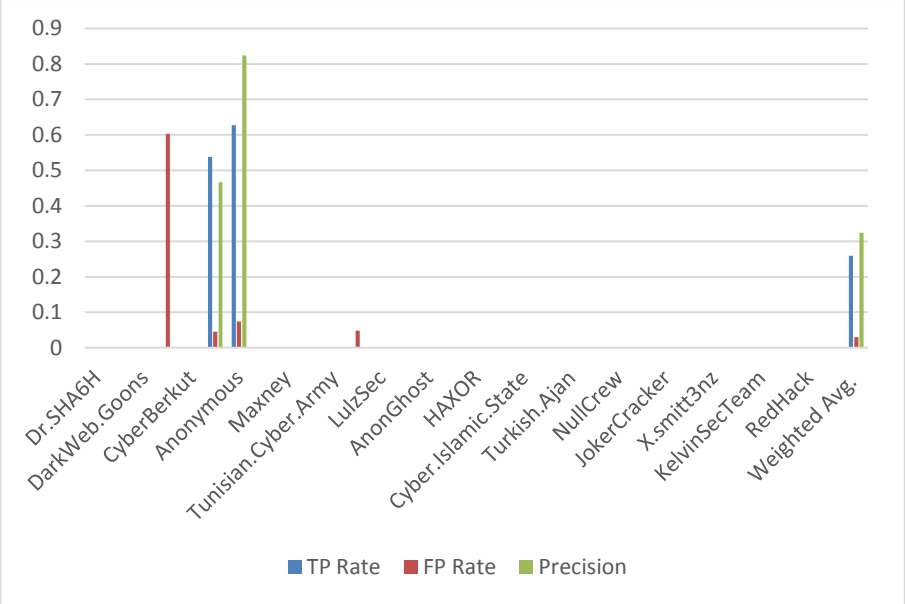

Figure 6-12 TP, FP, and Precision for prediction of cyber attackers

2. ROC is another benchmark to evaluate the predictive model. Figure 6-13 Shows ROC area for each class. The Anonymous group has the highest level of ROC with 0.897 indicating they are will be predicted by the model. Chines hackers are the second in terms of ROC area with the amount of 0.84 which also demonstrates the model is reliable in terms of predicting them.

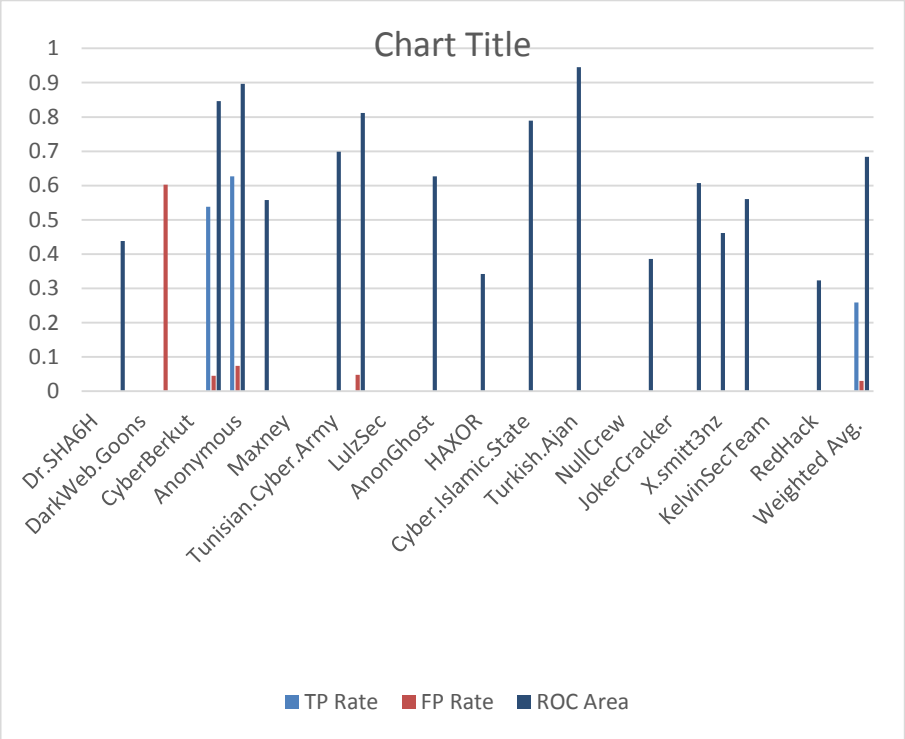

Figure 6-13 TP, FP, and ROC for prediction of cyber attackers

3. The next step of this analysis is comparing recall, precision, and ROC for each class which is shown in figure 6-14. Again this step indicates the Anonymous and the Chinese hackers are the most well predicted among other cyber attackers despite unreliable overall accuracy, precision and ROC with the amount of 0.26, 0.324 and 0.684 respectively.
4. In the final part of validation process, F- measure is considered. F- Score got the value of 0.2592. This predictive model might not be reliable as a single way of prediction but it does an accurate job when it comes to prediction of Anonymous group.

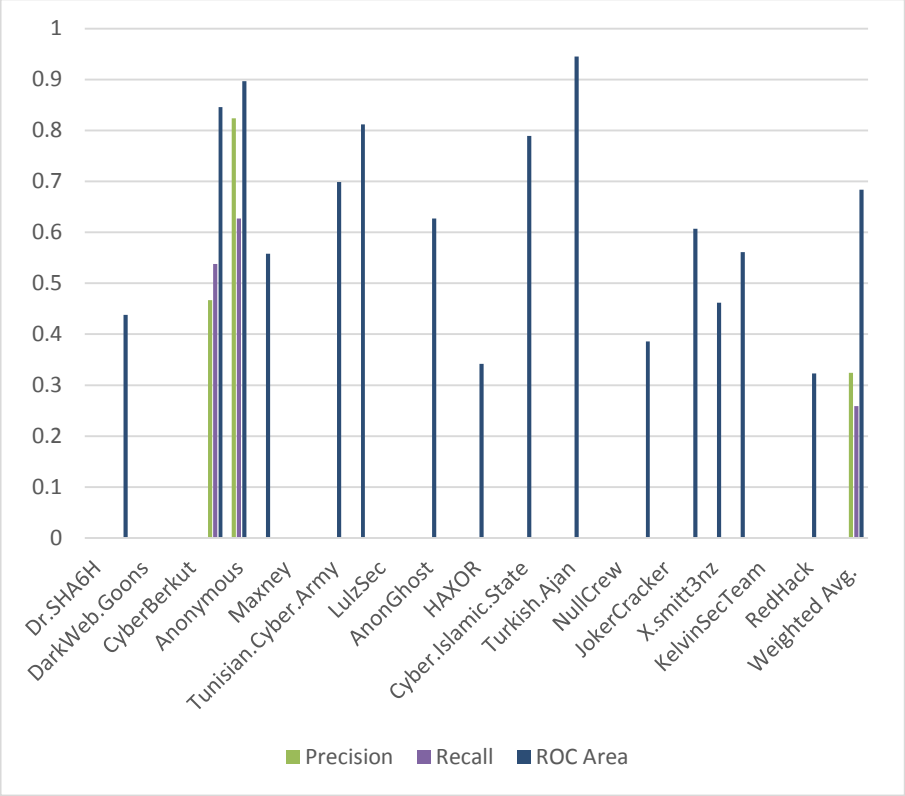

Figure 6-14 Precision, Recall, and ROC for prediction of cyber attackers

6.3.3 Validation of the Type of Target predictive model

This section aims to examine the reliability of the predictive model for the Type of Targets in cyber-attacks. 1137 records of cyber-attacks existing in the validation dataset will be used and fed into the predictive

model obtained in the previous chapter. Figure 6-15 shows the outcome of this validation process for the Type of Targets.

```
Overall Statistics

               Accuracy : 0.1812
                 95% CI : (0.1592, 0.2048)
    No Information Rate : 0.1926
    P-Value [Acc > NIR] : 0.8451

                  Kappa : 0.0382
 Mcnemar's Test P-Value : NA


=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC
Area  PRC Area  Class
                 0.000    0.002    0.000      0.000    0.000      -0.011   0.53
7     0.074      ED
                 0.247    0.315    0.068      0.247    0.107      -0.041   0.52
1     0.089      IO
                 0.521    0.386    0.244      0.521    0.332      0.108    0.61
2     0.258      GO
                 0.188    0.146    0.094      0.188    0.125      0.031    0.60
7     0.099      FB
                 0.000    0.000    0.000      0.000    0.000      0.000    0.45
7     0.006      Other
                 0.124    0.035    0.265      0.124    0.169      0.127    0.59
4     0.160      EN
                 0.000    0.007    0.000      0.000    0.000      -0.011   0.41
8     0.015      RT
                 0.059    0.015    0.056      0.059    0.057      0.042    0.61
4     0.020      TC
                 0.057    0.042    0.042      0.057    0.048      0.013    0.55
2     0.038      BP
                 0.051    0.011    0.200      0.051    0.081      0.077    0.70
2     0.133      MU
                 0.000    0.000    0.000      0.000    0.000      0.000    0.47
0     0.025      SN
                 0.000    0.000    0.000      0.000    0.000      0.000    0.53
3     0.058      HC
                 0.000    0.001    0.000      0.000    0.000      -0.006   0.47
8     0.044      NN
                 0.000    0.001    0.000      0.000    0.000      -0.006   0.61
5     0.054      HT
                 0.000    0.000    0.000      0.000    0.000      0.000    0.78
3     0.088      MD
                 0.000    0.002    0.000      0.000    0.000      -0.010   0.40
0     0.049      THS
                 0.000    0.000    0.000      0.000    0.000      0.000    0.60
2     0.137      SI
                 0.000    0.003    0.000      0.000    0.000      -0.006   0.66
1     0.026      ES
                 0.000    0.000    0.000      0.000    0.000      0.000    0.56
0     0.018      TP
Weighted Avg.    0.152    0.118    0.097      0.181    0.105      0.034    0.57
3     0.120
```

Figure 6-15 Validation of Type of Target Predictive model

According to the above script the overall accuracy of the model is not significant and it is 18.12%, however, by a further and deeper overview of the validation process and considering main criteria, following conclusions can be reached:

1. As figure 6-16 shows the plot of the comparison of TP rate, FP rate, and precision, Government has the highest level of TP rate; 0.521 and the second maximum level of FP rate. Therefore, it should be mentioned the model has done a significant job in terms of predication of cyber-attacks in the government section. The reason for high FP rate in GO can indicate that the attacks' pattern to other types of targets is very similar to governments. The highest precision level belongs to Entertainment section which means not only cyber-attacks to this section is well predicted but also shows its cyber-attack patterns have not been changed over time.

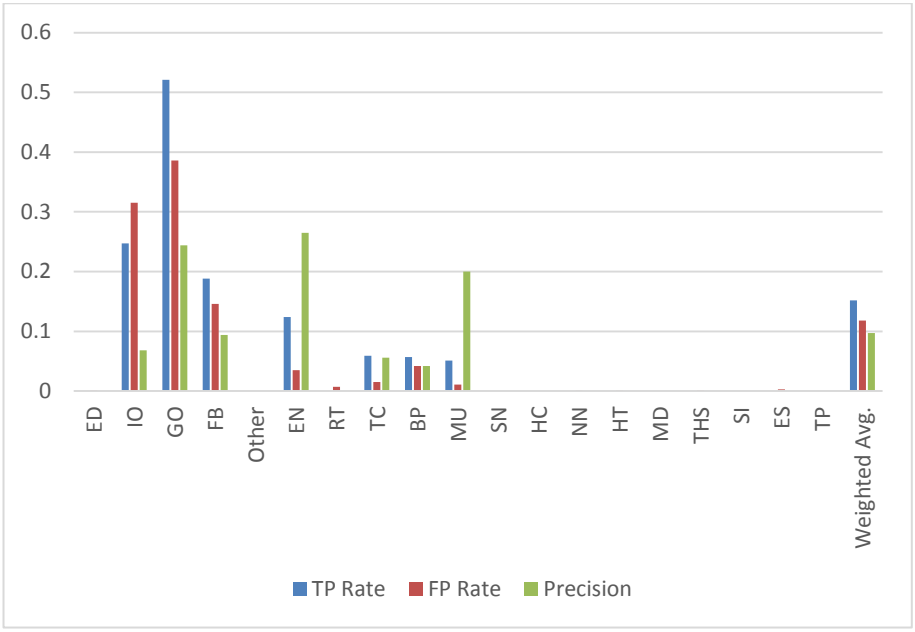

Figure 6-16 TP, FP and Precision for prediction of Type of Target

2. The second outcome of this validation process is ROC area. Figure 6-17 demonstrates the plot of TP, FP and ROC area. In terms of ROC Telecommunication sectors, Governments and Bank and Financial institutions have the highest ROC respectively with the level of 0.614, 0.612 and 0.607. This result indicates that the predictive model for these Type of Targets has worked well despite poor overall accuracy of the model.

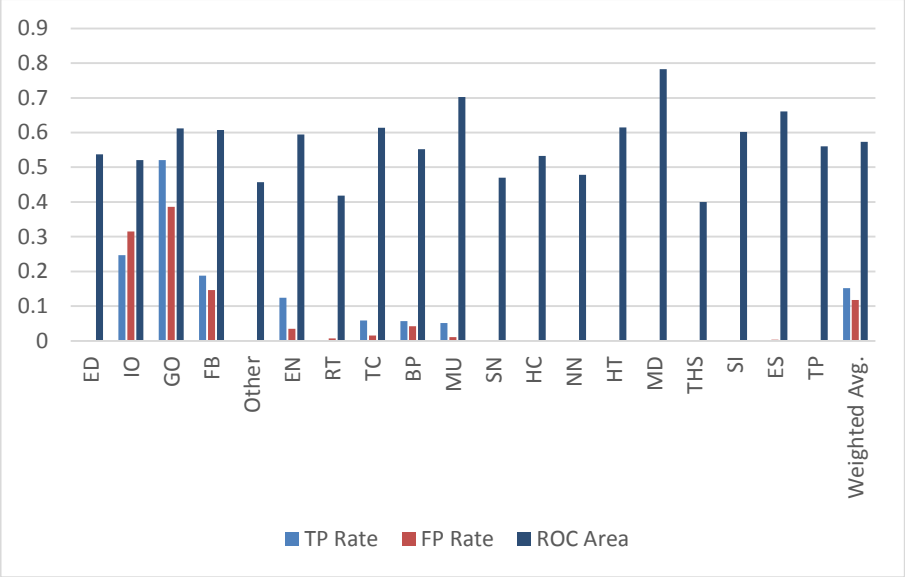

Figure 6-17 TP, FP and ROC for prediction of Type of Target

3. The third step is probing the predictive model by comparing recall, precision, and ROC as it is demonstrated in figure 6-18 by a bar plot. As it is shown, the model predicts attacks to Government with more accuracy because of significant ROC and precision in addition to high level of Recall. The second Type of Target which is predicted reliably is Entertainment section.
4. F-score for prediction of type of target, is 0.105. This result indicates that the predictive model is not accurate enough to be considered reliable and it should be combined with other methods of prediction. Government sector has F-score of 0.33 which not only shows more reliable result in this attribute but also shows the pattern of cyber attacks to governments have not been changed a lot over the years.

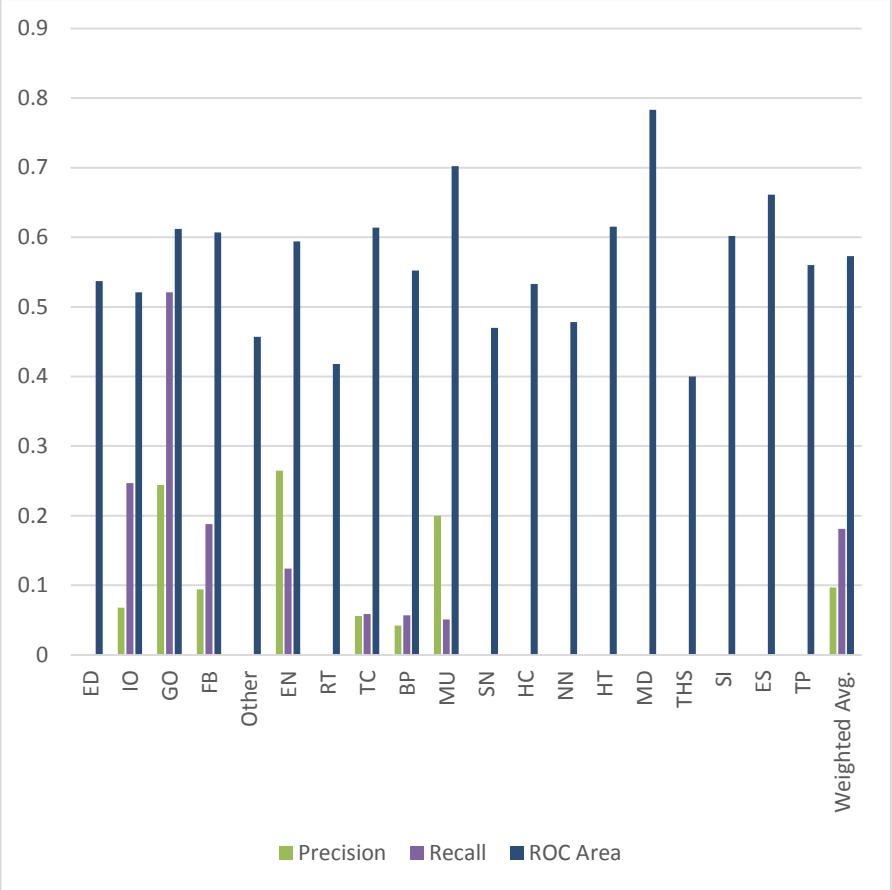

Figure 6-18 Precision, Recall, and ROC for prediction of type target

6.3.4 Validation of the Targeted Country predictive model

This section aims to validate the predictive model for targeted countries in cyber-attacks. This stage will be done by applying the model to the validation dataset which has 1137 records and figure 6-19 shows the results of the validation stage.

```
Overall Statistics

               Accuracy : 0.1428
                 95% CI : (0.1153, 0.1558)
    No Information Rate : 0.4512
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0106
 Mcnemar's Test P-Value : NA
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC
Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    0.28
9     0.009      CN
                 0.781    0.694    0.063      0.781    0.116      0.044    0.55
8     0.063      UK
                 0.000    0.006    0.000      0.000    0.000      -0.007   0.50
9     0.009      TR
                 0.144    0.300    0.284      0.144    0.191      -0.184   0.40
2     0.408      US
                 0.000    0.000    0.000      0.000    0.000      0.000    0.40
0     0.022      IN
                 0.000    0.000    0.000      0.000    0.000      0.000    0.52
6     0.011      IT
                 0.000    0.000    0.000      0.000    0.000      0.000    0.69
5     0.011      BR
                 0.000    0.001    0.000      0.000    0.000      -0.005   0.50
3     0.024      RU
                 0.268    0.029    0.561      0.268    0.363      0.334    0.61
9     0.268      INT
                 0.000    0.000    0.000      0.000    0.000      0.000    0.44
9     0.021      CA
                 0.000    0.000    0.000      0.000    0.000      0.000    0.41
8     0.009      FR
                 0.071    0.001    0.500      0.071    0.125      0.186    0.31
6     0.016      KR
                 0.000    0.001    0.000      0.000    0.000      -0.003   0.62
3     0.017      IL
                 0.000    0.003    0.000      0.000    0.000      -0.006   0.55
3     0.017      AU
                 0.000    0.001    0.000      0.000    0.000      -0.004   0.53
7     0.024      JP
                 0.000    0.000    0.000      0.000    0.000      0.000    0.19
4     0.003      PK
                 0.000    0.000    0.000      0.000    0.000      0.000    0.51
2     0.181      Other
                 0.000    0.000    0.000      0.000    0.000      0.000    0.66
7     0.014      SA
                 0.000    0.000    0.000      0.000    0.000      0.000    0.84
7     0.045      PH
                 0.000    0.000    0.000      0.000    0.000      0.000    0.48
9     0.005      CZ
                 0.000    0.000    0.000      0.000    0.000      0.000    0.63
1     0.023      DE
Weighted Avg.    0.142    0.178    0.206      0.142    0.138      -0.038   0.47
4     0.254
```

Figure 6-19 Validation of Targeted Country Predictive model

The overall accuracy of the model is 14.2% which shows the predictive cannot perform well in terms of detection of future targeted countries in cyber attacks, however, by probing the validation result following points can be mentioned:

1. As the figure 6-20 shows the comparison of TP rate, FP rate and precision between different countries, UK has the highest TP rate which is 0.78, however, the highest level of FP rate also belongs to the UK and that causes a low precision rate. This means attacks to the UK have followed the same pattern over recent years so they can be predicted well. The second maximum TP rate goes to targets located in different countries or they are multinational. This class of targeted countries has very low FP rate which leads to highest precision. Another high precision rate belongs to South Korea because although the TP rate is low, the FP rate is equal to zero.

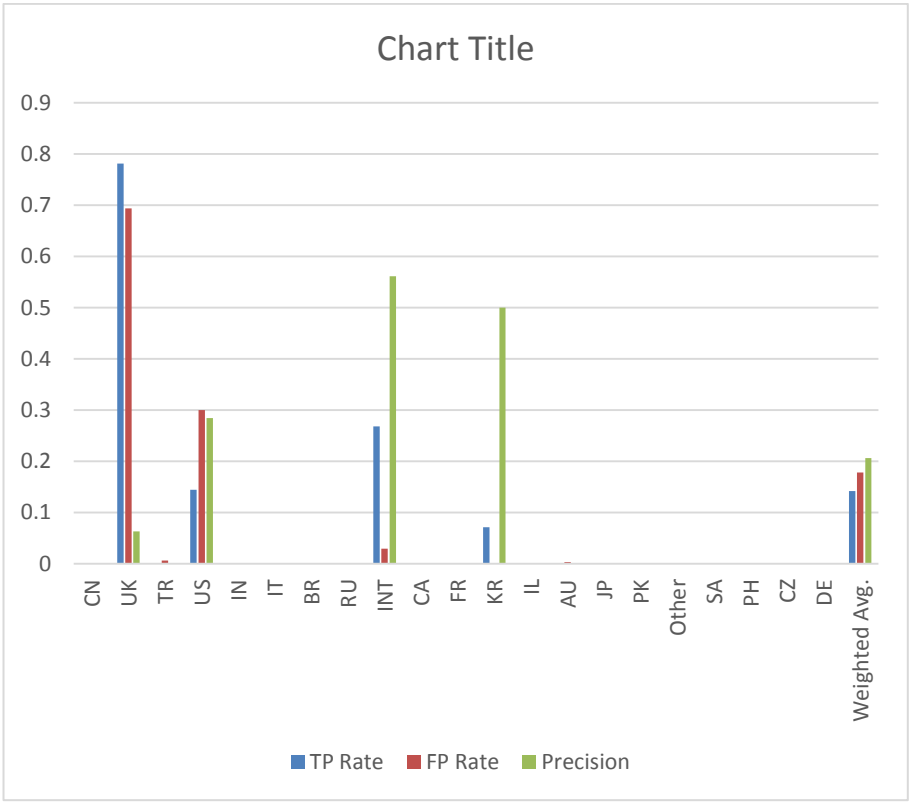

Figure 6-20 TP, FP, and Precision for the Targeted country predictive model

2. The figure 6-21 shows the plot demonstrating the level of TP, FP and ROC area for each class in the targeted countries. As it is shown some classes such as Saudi Arabia, Philippine, Czech Republic and etc. they have zero FP and TP rate but yet they have ROC, the reason is those countries did not exist in the validation dataset. Among those record, multinational targets have the most ROC just like their precision level so the model performs better in terms of prediction of multinational targets. UK and US are in the second and third rank in terms of ROC area.

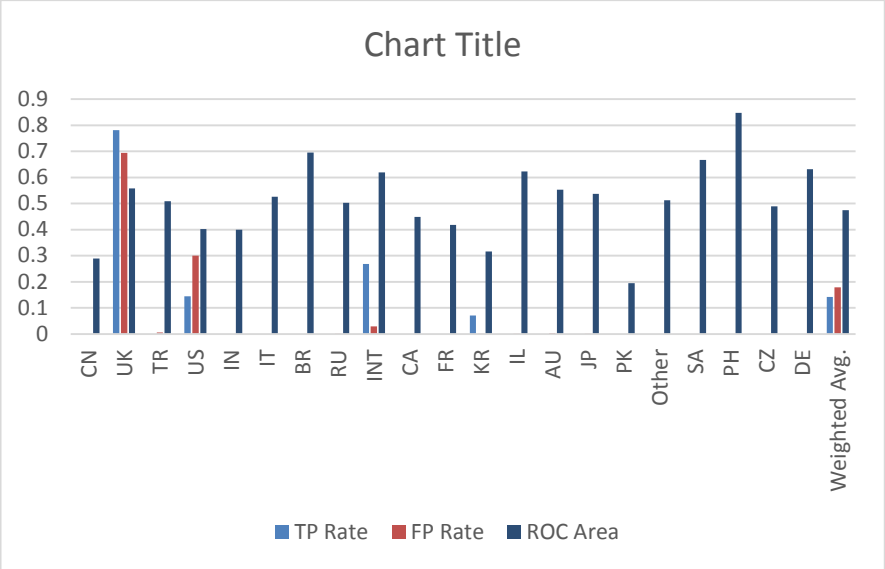

Figure 6-21 TP, FP, and ROC for the targeted country predictive model

3. The third step is probing the predictive model by comparing ROC, Precision, and Recall as it is shown in figure 6-22 by a bar plot. Multinational targets by having the highest precision and ROC are well predicted with the obtained model. Although the second highest precision belongs to South Korea, the insignificant level of recall leads to low ROC. In addition, according to further analysis of validation results, cyber-attacks to the UK and US can be predicted with 0.77 and 0.14 recall rate.
4. The Targeted country predictive model got F-score of 0.138. This indicates that the pattern of cyber attacks to countries can be changed over time, however, because some countries like US, UK and Republic of Korea have more F-score compared to the rest of countries, they have same pattern of cyber attacks.

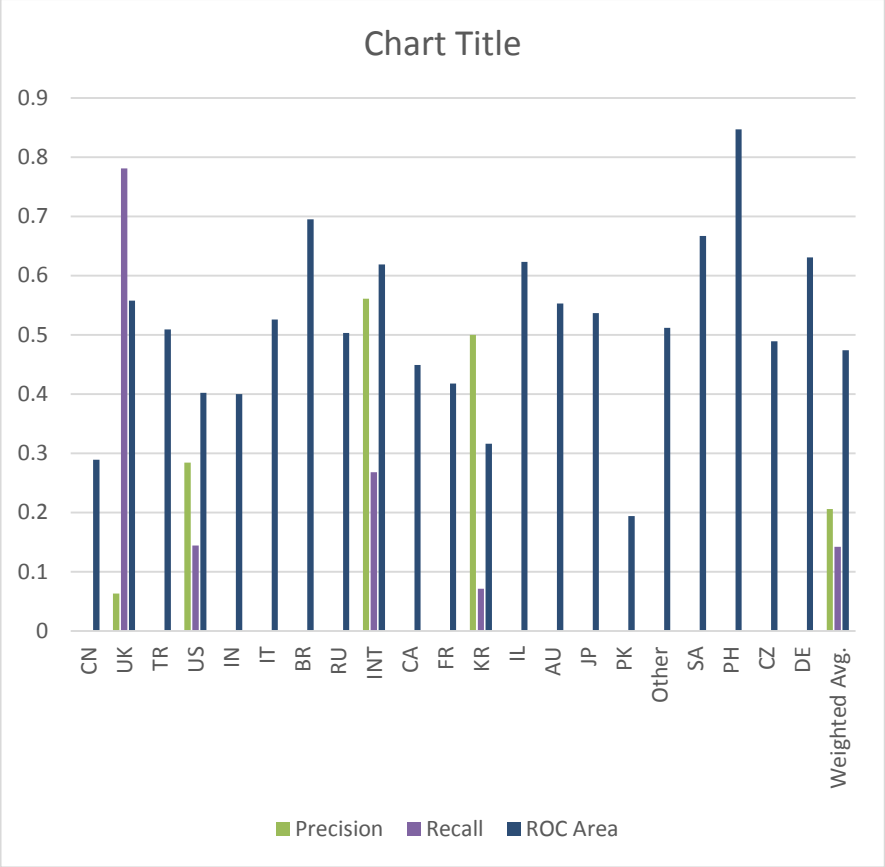

Figure 6-22 Precision, Recall, and ROC for the targeted country

6.3.5 Validation of the Cyber Attack Activity predictive model

In this section, the predictive model for cyber-attack activity will be evaluated by applying it to the validation dataset including 1137 records of cyber attacks from 2016 to the end of March 2017. Figure 6-23 demonstrates the outcome validation stage.

```
Overall Statistics

               Accuracy : 0.8504
                 95% CI : (0.5397, 0.5981)
    No Information Rate : 0.7247
    P-Value [Acc > NIR] : 1

                  Kappa : 0.6537
 Mcnemar's Test P-Value : <2e-16

Statistics by Class:


                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC
Area  PRC Area  Class
                 0.930    0.249    0.908      0.930    0.918      0.695    0.84
7     0.898    CC
                 0.022    0.003    0.250      0.022    0.040      0.063    0.66
9     0.109     CW
                 0.678    0.063    0.615      0.678    0.645      0.591    0.85
0     0.475     HA
                 0.835    0.027    0.789      0.835    0.811      0.788    0.93
9     0.700     CE
Weighted Avg.    0.850    0.192    0.831      0.850    0.836      0.666    0.85
0     0.791
```

Figure 6-23 Validation of Cyber Attack Activity model

The overall accuracy of the model is 85.04% which is very significant and considered as a reliable model. The deeper analysis of the validation stage for this predictive model is explained below:

1. In terms of TP, FP, and precision, Cybercrime has the highest TP rate which is 0.93 showing that the model has the most accurate detection and prediction in Cybercrimes. The lowest TP belongs to Cyber war which indicates that this model does not have a desirable prediction on CW, however, CW is not very common in both validation and training datasets so extracting a specific pattern is almost impossible. The lowest FP rate belongs to Cyber Espionage which again shows the model has done a significant job in the prediction of CE. CC has the maximum precision with the rate of 0.908 indicating that it does a very accurate job in the detection of CC. The figure 6-24 shows the bar plot of TP, FP, and precision for each class in this validation process.

Figure 6-24 TP, FP and Precision for Cyber-attack activity predictive model

2. As figure 6-25 demonstrates in terms of ROC area, all classes apart from CW in cyber attacks' activity have almost same amount of ROC area, CW has the lowest ROC area and the reason is its less frequency in the validation dataset.

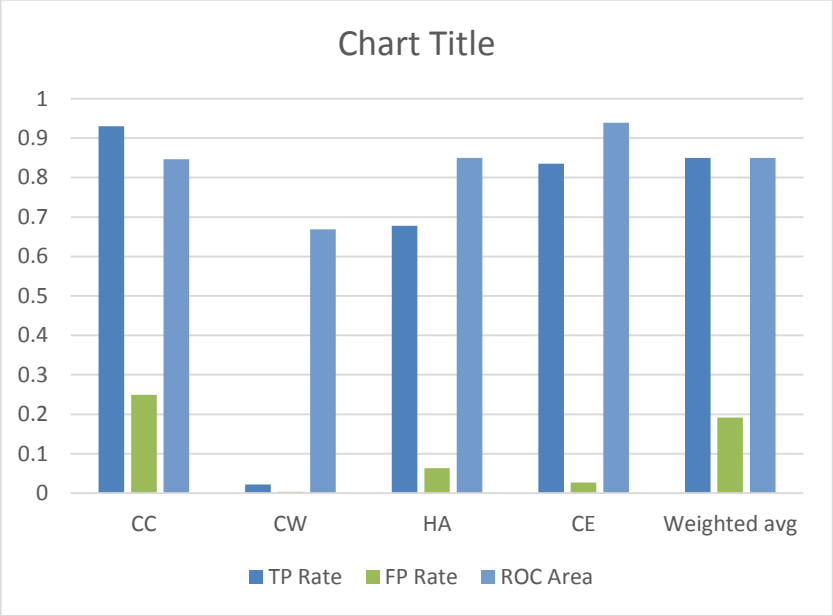

Figure 6-25 TP, FP, and ROC for cyber-attack activity predictive model

3. As figure 6-26 shows in terms of comparing Recall, precision and ROC area, all classes apart from CW have a significant level of these benchmarks.
4. F-score of cyber attack activity is the highest compared to the other models and it got value of 0.836. This indicates this model performed more reliable and can be used for identifying the motivation of any cyber attacks.

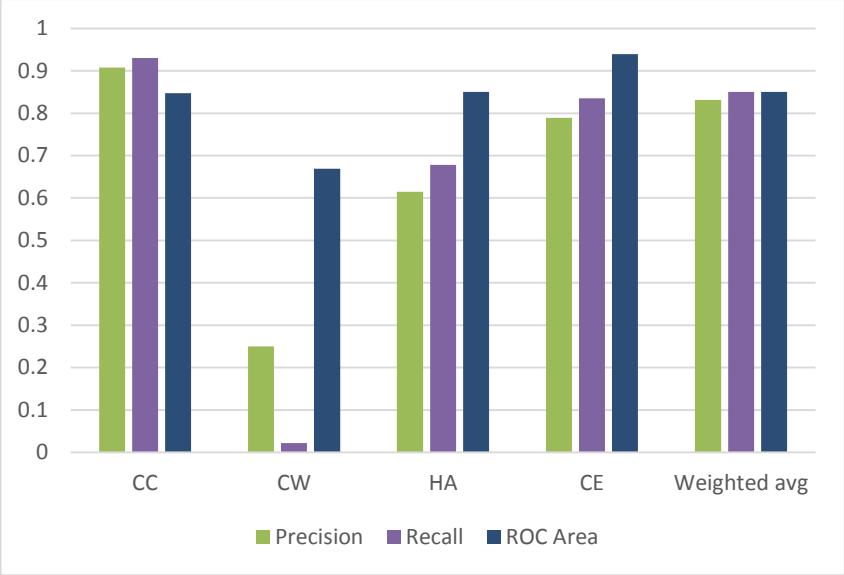

Figure 6-26 Precision, Recall, and ROC for cyber-attack activity predictive model

6.4 Summary

This subchapter aims to summarize the discussion about the proposed predictive model which is obtained in this project. As it was mentioned previously in this chapter in order to evaluate the predictive model, the validation data set which includes unseen records of cyber-attacks taking place from 2016 to end of March 2017 was fed into the model. The obtained model has 5 different dimensions to predict cyber attacks' features including the type of cyber threat, Type of Target, targeted countries, type of cyber-attack activity and cyber attackers. Each of the model's dimension was evaluated and analysed in this chapter and the following points are extracted from the evaluation process:

1. The strongest dimension of the predictive model happens to be cyber-attack activity with 85.04% accuracy and 0.65 Kappa. These factors indicate that the model performs very well in the prediction of the type of cyber-attack activity and also they reflect that the pattern of cyber-attack activity has not been changed by high margin during past years. For more details when the performance of the model on training and validation set is compared, the accuracy on the validation set is slightly higher, whereas the model does a slightly better job in training set by considering Kappa. By analysing the outcome of the validation process, the success of the model is shown in detection of cybercrime and cyber espionage attacks with 0.93

and 0.83 recall level and also by considering ROC area, cyber espionage and hacktivism attacks prediction by the model is highly reliable with recall level of 0.85 and 0.93. Cyberwar attacks have the lowest recall and ROC which can be because of the different reasons; firstly CW attacks are not very common which might show that the victim governments or sectors try to hide their failures from global rivals and secondly the state sponsor cyber-attacks usually are not be taken credit of by any governments.

2. The second most accurate dimension of the predictive model is cyber attacker prediction with 25.92% accuracy and 0.17 kappa which demonstrates using this predictive model as a standalone method is not logical to solve cyber security issues. By deeper probing the validation process, it shows the model can predict and detect Chinese hackers and Anonymous group more accurate with 0.53 and 0.62 recall rate respectively, therefore, it can be decided that these cyber attackers have not changed their attacks' pattern significantly during past 5 years. In addition by comparing the accuracy of the model in training ad validation process, it can be seen that the model has done more reliable job with 61.34% accuracy and 0.43 kappa, whereas the accuracy and kappa in validation process have a significant difference with it. To sum up it can be concluded that cyber attackers changed their cyber-attack methods and patterns over time and sometimes they even change their alias names in order to stay unidentifiable and hide their identity and the more updated data and information cyber experts gain, the chance of detection and prediction of cyber attackers will go higher.

3. The third dimension of the model in terms of accuracy strength is the Type of Target prediction with 18.12% reliability and 0.03 kappa. The kappa and accuracy rate suggest that the model is not a reliable predictive method for potential next Type of Targets of cyber-attacks as a single way. By comparing training and testing process, it is found out that model performs much better on the training set with 39.69% accuracy and 0.28 kappa which indicates that cyber-attacks can happen in any targets and patterns are constantly changing. Low kappa of the validation outcome is an indicator that generally any business or sector should be prepared for different kind of cyber-attacks because past experience cannot be reliable, countable and enough for protection against cyber-attacks. In addition, the validation outcome shows cyber-attacks to government sectors

are highly predictable with 0.52 recall rate and Banks and financial sectors can be well predicted by considering ROC measure.

4. The fourth dimension of the model is targeted country prediction with 14.28% accuracy and 0.01 kappa. The kappa and accuracy rates suggest that the model is not accurate and successful enough for prediction of future targeted countries. The model on the training set gave 48.96% accuracy and 0.22 kappa level in section 6.3.4 which indicates that the model cannot perform well on unseen data, however, UK, US, and multinational-based companies are well predicted with this model regarding their high recall and Roc rate. The reason behind of that is inspired by two main points; firstly the attacks in UK and US are more common in both training and validation dataset in the result of more transparency and announcement of cyber-attacks, secondly, the patterns of cyber-attacks in these countries have been following the same path during past couple years. Multinational companies are also more accurately predicted among other classes in this model because nowadays most of the companies' stocks are shared between different counties and they are more common in both training and test data set.
5. The weakest dimension of the model belongs to the prediction of cyber threats with 11.29% accuracy and 0.06 kappa. This predictive dimension of the model is not reliable enough as a single method for cyber security experts, however, the validation result can be interpreted and more meaningful. Defacement and Account Hijacking are well predicted by the model with 0.63 and 0.23 recall rate which indicates cyber experts can find a usual pattern for these type of attacks to protect different sectors from them because of stability of these attacks during the past couple of years. From another angle, when the validation process is compared with the training process a significant decrease can be seen. In another word the model performs very well on seen data and generally patterns of cyber threats get changed over time and cyber attackers change their way of attacks regularly.
6. Another metric which can be used for measuring the accuracy of the models is F-measure. F- measure will be calculated based on precision and recall. It varies from 0 the worst to 1 the best. In the field of information retrieval, F-measure plays an important role when it comes to document classification

For the full scripts and dataset please refer to Appendix.

Chapter 7 Conclusion

7.1 Contribution to knowledge

Nowadays data mining techniques play a crucial role in daily life including cyber security in other words by analysing current and past cyber breaches, future breaches can be prevented and a meaningful picture can be drawn for managers to improve cyber situational awareness, provide suitable strategy and implement security policies and countermeasures combating future cyber-attacks. In this study classification technique as a data mining approach was used to address a cyber security matter which was Cyber Situational Awareness. Regarding the research question which is to investigate to what extent a predictive framework based on classification techniques and OSINT can contribute to better understanding and improving CSA, this study contribution to knowledge can be divided to the following points:

1. One of the novelties of this research is using Open Source Intelligence for training the predictive models. The dataset which has been used in this study has over 5000 records of cyber-attacks taken place from 2013 to the end of March 2017 and has been gathered from OSINT. The dataset has been cleaned and pre-processed with different kind of tools such as R, Open refine etc. The pre-processing was done automatically and manually. This dataset which has been provided by this study can be used for future researches. The dataset has 7 different attributes; Date, Cyber attacker, Type of Cyber Threat, Type of Target, Targeted country, and cyber-attack activity.

2. The final predictive model was concluded by comparing the models trained based on5 classification techniques. The

Support Vector Machine model has the best result in terms of accuracy for each dimension of the cyber-attacks. In terms of prediction cyber-attack activity, the model has 82.56% accuracy on the seen data but when unseen data applies to the model the accuracy can go higher up to 85% which indicates that the pattern of cyber-attack activity has not been changed generally. The dimension of cyber attacker prediction has 61.34% accuracy, however, if unseen data applies to the model the accuracy goes down to 26%. This will lead to a conclusion that cyber attackers' identity might change over time and their prediction will be difficult apart from Anonymous group and Chinese hackers which they had high recall rate compared to others. The obtained model also does not perform accurate enough to predict the future Type of Targets, however, this can contribute to the fact that various cyber-attack can happen in any Type of Target. In terms of prediction of targeted country, the predictive model predicts UK and US more accurately than other countries which means the pattern of cyber-attacks taken place in these countries still remain the same over past few years. In terms of prediction of cyber threats, although the model does not perform accurate enough, Defacement and Account Hijacking can be predicted very well according to a high recall rate.

3. Another contribution of this research was to investigate how much effect each attribute in a cyber-attack plays in the prediction. In terms of prediction of cyber threats, cyber attackers have the most influence in the accuracy which means often cyber attackers pose same cyber threats to their victims, this can help cyber security experts to plan a broad strategy when they know who is more likely to target their businesses. In terms of prediction of cyber attackers, without applying any knowledge discovery method, the targeted country can be a significant factor to identify cyber attackers, however, the predictive model obtained in this study indicates there are a high influence and dependency from the Type of Target and type of cyber threat on the prediction of cyber attackers. When it comes to prediction of the Type of Target, all the features play an equal role in the obtained predictive model in this research which means all Type of Targets can be vulnerable to any sort of attack. Targeted countries also can be predicted by considering all factors in a cyber-attack in the predictive model, however, without applying any data mining methods it can be seen that cyber attackers can be a significant predictor for targeted

countries. Cyber attackers and Type of Threat that they pose to their victims can also influence the prediction of activity of cyber-attacks or the main motivation of them.

To sum up, this study indicates that Cyber Situational Awareness can be improved by classification techniques and OSINT to some extent but this cannot be a standalone method. It has been tried to cover mainly high level of cyber situational awareness when it comes to planning a broad and extensive strategy to tackle cyber incidents. The 5-dimensional predictive framework which is based on Support Vector Machine can be used by cyber security experts to understand, predict past and future attributes included in cyber-attacks. By prediction of the type of cyber threat, Type of Target and targeted countries, different strategies can be designed in order to prevent future breaches for any businesses or governmental sections. Prediction of cyber attackers in this research also can help law enforcement agencies to collect more evidence on cyber criminals and make their investigation smarter and more time efficient.

7.2 Limitation of Study

Like many other researches, this study has its limitation of study in different sections. The limitations are more around the data element and that is about completeness and different aspects of the data. The data was collected from Open Source Intelligence and as it was mentioned before, any data from OSINT comes with significant amount of noise and irrelevant data. In the section of data pre-processing and data cleansing including removing irrelevant cyber-attack records, the operation was done manually and the only way to validate cyber-attacks was probing each record by reading the resource of information. Therefore, this operation was time consuming and could not be more time efficient and faster. Another limitation of data access was due to sensitivity of cyber security subject, therefore, the accuracy and completeness of data can be challenging as most of the companies and governments try to hidden their cyber security incidents in order to preserve their reputation. Having a complete and accurate dataset can lead to more precise and reliable result in terms of predictive models.

Another limitation of this study was the converting qualitative data to quantitative. This could happen automatically or manually. Doing it manually will be a time consuming task and because of time limitation

in this study, this option was ruled out. In addition applying this process automatically will not have sensible meaning as this should involve more human interference. Many other type of analysis such as clustering and regression analysis, however, based on type of analysis some attributes might need to be assigned a weight which can be done by consulting with cyber security experts through an interview or a survey.

7.3 Future work

The future work that can be done in order to extend and improve this research can be divided into the following areas:

1. Dataset: one of the area in this research which can be improved is the dataset. The cyber-attack dataset was obtained from open source intelligence. The dataset can be extended in terms of number attributes, however, it might need to combine with other data resources. As previously mentioned the dataset in this study has 5 main attributes which was the cyber attacker, type of cyber threat, Type of Target, targeted country and cyber-attack activity. Other attributes can be added if more information is available such as the origin of the attack which can show the IP address or the location of cyber attackers. With more information on cyber-attacks, models that are more accurate can be concluded. In future studies, the dataset can be more in-depth in terms of attributes. The additional attributes can be more technical which then covers a model improving low level of cyber situational awareness or can be more none technical which then leads to a model to cover high level of cyber situational awareness.

2. Data mining techniques: In this study regarding its aim, classification algorithms were used. Another data mining techniques which can be used is Time Series Analysis, however, time series analysis can be highly challenging in cyber security due to the fact the nature of cyber-attacks varies over none technical and technical factors. For instance, due to high usage of mobile devices, cyber threats have become more mobile-based during recent years. None technical factors also play a crucial role and have effect on cyber-attacks, politically or socially motivated attacks can increase or decrease based on the level tension all around the world. Clustering and data

mining techniques can be applied to cyber-attacks data set in future researches. Clustering and other data mining techniques can help cyber security experts for deeper knowledge discovery when it comes to planning a broad strategy to tackle cyber incidents.

3. Real-time protection: For future studies, one of the areas which can be enhanced is real-time protection against cyber-attacks by a dynamic cyber situational awareness improving method. The real-time improving CSA needs live streaming of new data into the cyber-attack dataset. If the new attacks are uploaded in real time, then it will be possible to maintain a real-time protection by applying data mining techniques to the cyber-attack dataset. Live streaming the data can also make a CSA more dynamic and it can be changed based on different conditions.
4. Another future research that can be done in this field is the application of deep learning. Deep learning is often described as a subset of machine learning and data mining with higher complexity. Deep learning has been mainly used for bigger size of data and when the high accuracy is crucial. Examples of deep learning application are image recognition and voice recognition. In the field of Cyber Situational awareness, utilizing deep learning will be useful if the size of the data becomes bigger in terms of size and it has more complexity in terms of attributes and dimensions. One of the modern platforms that can be used of applying deep learning is Tensor flow which has been released in 2015. Tensor flow is a free tool having open source nature and can apply different deep learning algorithms. Tensor flow can help researchers to implement different types of Neural network algorithms to the training data and provide a predictive model. (Abadi et al.,2016)

Chapter 8 References

Aaviksoo, J. 2008, "Cyber-Terrorism", Vital speeches of the day, vol. 74, no. 1, pp. 28

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M. and Kudlur, M., 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)* (pp. 265-283).

Agatonovic-Kustrin, S. and Beresford, R., 2000. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Journal of pharmaceutical and biomedical analysis, 22(5), pp.717-727.

Ahlemeyer-Stubbe, A and Coleman, S 2014, A Practical Guide to Data Mining for Business and Industry.

Ahn, S., Kim, N., & Chung, T. ,2014. Big Data Analysis System Concept for Detecting Unknown Attacks.

Akhgar, B., 2015; National Security and situational awareness, SAP HQ keynote presentation (2015) , Germany .

Akhgar, B., Staniforth, A. and Bosco, F. eds., 2014. Cyber Crime and Cyber Terrorism Investigator's Handbook. Syngress.

Al-janabi, K. B. S. , 2011. A Proposed Framework for Analysing Crime Data Set Using Decision Tree and Simple K-Means Mining Algorithms, 1(3), 8–24.

Al-Shamisi, A., Louvieris, P., Al-Mualla, M. and Mihajlov, M., 2016, May. Towards a theoretical framework for an active cyber situational awareness model. In Systems, Signals and Image Processing (IWSSIP), 2016 International Conference on (pp. 1-6). IEEE.

Angelini, M and Santucci, G 2017, Cyber situational awareness: from geographical alerts to high-level management, Journal of Visualization, 20, (3), Springer Berlin Heidelberg, pp. 453–459.

Antonik, J., 2007. Decision management , In Military Communications Conference (MILCOM ’07), pages 1–5, Orlando, FL, USA, October 2007. IEEE.

Aspan, M., Soh, K., 2011. Citi says 360,000 accounts hacked in May cyber-attack. Reuters.

Awan, I., Blakemore, B. & MyiLibrary 2012, Policing cyber hate, cyber threats and cyber terrorism, Ashgate, Farnham

Barford, P., Dacier, M., Dietterich, T. G., Fredrikson, M., Giffin, J., Jajodia, S., … Wang, C., 2010. Cyber SA : Situational Awareness for Cyber Defense, 3–14.

Barford, P., Dacier, M., Dietterich, T.G., Fredrikson, M., Giffin, J., Jajodia, S., Jha, S., Li, J., Liu, P., Ning, P. and Ou, X., 2010. Cyber SA: Situational awareness for cyber defense. In Cyber situational awareness (pp. 3-13). Springer, Boston, MA.

Berkhin, P., 2006. A survey of clustering data mining techniques. In Grouping multidimensional data (pp. 25-71). Springer, Berlin, Heidelberg.

Beyer, K., Goldstein, J., Ramakrishnan, R. and Shaft, U., 1999, January. When is “nearest neighbor” meaningful?. In International conference on database theory (pp. 217-235). Springer Berlin Heidelberg.

Bhardwaj, B. K., & Pal, S., 2011. Data Mining : A prediction for performance improvement using classification, 9(4).

Breiman, L. and Cutler, A., 2007. Random forests-classification description. Department of Statistics, Berkeley, 2.

Cortes, C. and Vapnik, V., 1995. Support-vector networks. Machine learning, 20(3), pp.273-297.

Coughlan, P. and Coghlan, D., 2002. Action research for operations management. *International journal of operations & production management*, *22*(2), pp.220-240.

Cox, C. 2015, "Cyber Capabilities and Intent of Terrorist Forces", Information Security Journal: A Global Perspective, , pp. 1-8.

Das, S., Mukhopadhyay, A., & Shukla, G. K., 2013. i-HOPE Framework for Predicting Cyber Breaches: A Logit Approach. 2013 46th Hawaii International Conference on System Sciences, 3008–3017.

Dean, J 2014, Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners, Wiley & SAS.

Dua, S., & Du, X., 2011. Data Mining and Machine Learning in Cybersecurity. CRC Press.

Dutt, V., Ahn, Y.-S., & Gonzalez, C., 2012. Cyber Situation Awareness: Modeling Detection of Cyber Attacks With Instance-Based Learning Theory. Human Factors: The Journal of the Human Factors and Ergonomics Society, 55(3), 605–618. doi:10.1177/0018720812464045

Fawcett, T., 2006. An introduction to ROC analysis. Pattern recognition letters, 27(8), pp.861-874.

Fayyad, S., & Meinel, C., 2013. Attack Scenario Prediction Methodology. 2013 10th International Conference on Information Technology: New Generations, 53–59. doi:10.1109/ITNG.2013.16

Feasel, J., & Romas, G., 2013. Visualization, Modeling and Predictive Analysis of Internet Attacks, 8768, 1–6.

Fleiss, J.L., Cohen, J. and Everitt, B.S., 1969. Large sample standard errors of kappa and weighted kappa. Psychological Bulletin, 72(5), p.323.

Franke, U., & Brynielsson, J., 2014. Cyber situational awareness – a systematic review of the literature. Computers & Security, 46, 18–31. doi:10.1016/j.cose.2014.06.008

Freund, Y. and Mason, L., 1999, June. The alternating decision tree learning algorithm. In icml (Vol. 99, pp. 124-133).

Friedman, J.H., 1976. A recursive partitioning decision rule for nonparametric classification. IEEE Trans. Comput., 26(SLAC-PUB-1573-REV), p.404.

Goode, L., 2015. Anonymous and the political ethos of hacktivism. Popular Communication, 13(1), pp.74-86.

Gould, J., 2015. US Army Seeks Leap-Ahead Cyber Defense Tech [WWW Document]. Defense News. URL http://www.defensenews.com/story/defense/policy-budget/cyber/2015/07/01/us-army-seeks-breakthrough-tech-for-cyber-defense/29565733/ (accessed 3.10.16).

Han, J., Pei, J. and Kamber, M., 2011. Data mining: concepts and techniques. Elsevier.

Harrison, L., Laska, J., Spahn, R., Iannacone, M., Downing, E., Ferragut, E. M., & Goodall, J. R., 2012. situ: Situational understanding and discovery for cyber attacks. 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), 307–308. doi:10.1109/VAST.2012.6400503

Hornik, A, Karatzoglou, A, Meyer, D, Buchta, C, Hothorn, T and Zeileis, A 2017, Package ' RWeka '.

Hornik, K., Buchta, C. and Zeileis, A., 2009. Open-source machine learning: R meets Weka. Computational Statistics, 24(2), pp.225-232.

Hripcsak, G. and Rothschild, A.S., 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, *12*(3), pp.296-298.

Huang, Z, Shen, CC, Doshi, S, Thomas, N and Duong, H 2016, Fuzzy sets based team decision-making for Cyber Situation Awareness, Proceedings - IEEE Military Communications Conference MILCOM, pp. 1077–1082.

Jaishankar, K. & Dawsonera., 2011. Cyber criminology: exploring Internet crimes and criminal behaviour, CRC, Boca Raton, Fla; London.

Kelley, M.B., 2013. The Stuxnet attack on Iran's nuclear plant was 'far more dangerous' than previously thought. Business Insider, 20.

Kohavi, R. and John, G.H., 1997. Wrappers for feature subset selection. Artificial intelligence, 97(1-2), pp.273-324.

Kothari, C.R., 2004. *Research methodology: Methods and techniques*. New Age International.

Kuhn, M., 2008. Caret package. Journal of Statistical Software, 28(5), pp.1-26.

Kuhn, M., 2008. Caret package. Journal of Statistical Software, 28(5), pp.1-26.

Landis, J.R. and Koch, G.G., 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics, pp.363-374.

Larose, D.T., 2005. k-Nearest Neighbor Algorithm. Discovering Knowledge in Data: An Introduction to Data Mining, pp.90-106.

Ledolter, J 2013, DATA MINING AND BUSINESS ANALYTICS WITH R, Wiley & SAS.

Leung, K.M., 2007. Naive bayesian classifier. Polytechnic University Department of Computer Science/Finance and Risk Engineering.

Lewis, J. A., 2002. Assessing the Risks of Cyber Terrorism, Cyber War and Other Cyber Threats:, (December), 1–12.

Liaw, A and Wiener, M 2015, Package ‘ randomForest ’.

Majka, M., 2017. High Performance Implementation of the Naive Bayes Algorithm.

McNeill, P. 1990, Research methods, Routledge

Mining, W.I.D., 2006. Data Mining: Concepts and Techniques. Morgan Kaufinann.Hand, D.J., 2007. Principles of data mining. Drug safety, 30(7), pp.621-622.

Morris, I., Mayron, L. M., Smith, W. B., Knepper, M. M., Ita, R., Fox, K. L., & Corp, H., 2011. A perceptually-relevant model-based cyber threat prediction method for enterprise mission assurance, 60–65.

Murphy, K.P., 2006. Naive bayes classifiers. University of British Columbia.

Musliner, D. J., Rye, J. M., Thomsen, D., McDonald, D. D., Burstein, M. H., & Robertson, P., 2011. FUZZBUSTER: Towards Adaptive Immunity from Cyber Threats. 2011 Fifth IEEE Conference on Self-Adaptive and

Self-Organizing Systems Workshops, 137–140. doi:10.1109/SASOW.2011.26

Nath, S., 2006. Crime pattern detection using data mining. Web Intelligence and Intelligent Agent Technology …, 1(954).

nationsonline.org, klaus kästle-, 2017. List of Country Codes :: Nations Online Project [WWW Document]. URL http://www.nationsonline.org/oneworld/country_code_list.htm (accessed 11.1.17).

Nikishin, A. 2004. Malicious software–past, present and future. Information Security Technical Report. 9(2): pp.6-18.

Nitta, Y., 2013. Japan's Approach towards International Strategy on Cyber Security Cooperation. Retrieved September, 13, p.2014.

Odei Danso, S 2006, An Exploration of Classification Prediction Techniques in Data Mining : The insurance domain, Bournmouth University.

Onwubiko, C 2016, Exploring web analytics to enhance cyber situational awareness for the protection of online web services, 2016 International Conference on Cyber Security and Protection of Digital Services, Cyber Security 2016.

Paxton, NC, Jang, D Il, Russell, S, Ahn, GJ, Moskowitz, IS and Hyden, P 2015, Utilizing network science and honeynets for software induced cyber incident analysis, Proceedings of the Annual Hawaii International Conference on System Sciences, 2015-March, pp. 5244–5252.

Pollitt, M.M. 1998, "Cyberterrorism — fact or fancy?", Computer Fraud & Security, vol. 1998, no. 2, pp. 8-10.

Pournouri, S., & Craven, M. ,2014. E-business, recent threats and security countermeasures. International Journal of Electronic Security and Digital Forensics, 6(3), 169-184.

Quinlan, J.R., 1993. C4. 5: Programming for machine learning. Morgan Kauffmann, 38.

Rajpal, R., Kaur, S. and Kaur, R., 2016, July. Improving detection rate using misuse detection and machine learning. In SAI Computing Conference (SAI), 2016 (pp. 1131-1135). IEEE.

Ray, S., 2015. 6 Easy Steps to Learn Naive Bayes Algorithm (with code in Python). Analytics Vidhya.

Saini, H., Rao, Y.S. and Panda, T.C., 2012. Cyber-crimes and their impacts: A review. International Journal of Engineering Research and Applications, 2(2), pp.202-9.

Savov, V., 2014. Sony Pictures hacked: the full story [WWW Document]. The Verge. URL http://www.theverge.com/2014/12/8/7352581/sony-pictures-hacked-storystream (accessed 6.4.15).

Schreiber-Ehle, S., & Koch, W., 2012. The JDL model of data fusion applied to cyber-defence—A review paper. Sensor Data Fusion: Trends, …, (September), 4–6. doi:10.1109/SDF.2012.6327919

Security, C n.d., Cyber-Situational Awareness in the Financial Sector, pp. 1–7.

Sedgwick, P., 2012. Pearson's correlation coefficient. BMJ: British Medical Journal (Online), 345.

Singh, S., 2011. Artificial Neural Network.

Stalder, F. and Hirsh, J., 2002. Open source intelligence. First Monday, 7(6).

Steele, R.D., 2007. Open source intelligence. Handbook of intelligence studies, pp.129-147.

Studio, R., 2012. RStudio: integrated development environment for R. RStudio Inc, Boston, Massachusetts, p.74.

Team, R.C., 2000. R language definition. Vienna, Austria: R foundation for statistical computing.

The Health Information Trust Alliance, 2014. Healthcare Organizations Lack Tools for Cyber Situational Awareness and Threat Assessment [WWW Document]. Dark Reading. URL http://www.darkreading.com/analytics/healthcare-organizations-lack-tools-for-cyber-situational-awareness-and-threat-assessment/d/d-id/1319344 (accessed 3.10.16).

Therneau, T, Ripley, B and Atkinson, B 2017, Package ‘ rpart ’.

Therneau, T.M., Atkinson, B. and Ripley, B., 2010. rpart: Recursive partitioning. R package version, 3, pp.1-46

Thompson, J.R., 2001. Estimating equations for kappa statistics. Statistics in medicine, 20(19), pp.2895-2906.

Tong, S. and Koller, D., 2001. Support vector machine active learning with applications to text classification. Journal of machine learning research, 2(Nov), pp.45-66.

Vapnik, V.N. and Vapnik, V., 1998. Statistical learning theory (Vol. 1). New York: Wiley.

Verborgh, R. and De Wilde, M., 2013. Using OpenRefine. Packt Publishing Ltd.

Walliman, N.S.R. 2011, Research methods: the basics, Routledge, London

Wu, J., Yin, L., & Guo, Y., 2012. Cyber Attacks Prediction Model Based on Bayesian Network. 2012 IEEE 18th International Conference on Parallel and Distributed Systems, 730–731. doi:10.1109/ICPADS.2012.117

Yin, Y., Kaku, I., Tang, J. and Zhu, J., 2011. *Data mining: Concepts, methods and applications in management and engineering design*. Springer Science & Business Media.

Chapter 9 Appendix

9.1 DVD index

Training Dataset…….Data/Datasets

Validation Dataset……Data/ Validation datasets

Decision Tree models……Script and Workspace/ Decision Tree

K nearest neighbour models…….. Script and Workspace/KNN

Naïve Bayes model…… Script and Workspace/Naïve Bayes

SVM model…… Script and Workspace/SVM

ANN model…….Script and Workspace/ANN

9.2 Type of Target

Acronym	Type of Target	Example
BP	Broadcast and Publishing	Including Publisher companies and magazines
ED	Education	Including colleges, schools, and universities
EN	Entertainment	Including music and video game companies and etc.
ES	Energy Section	Companies and sectors operating in Oil, Power and

		etc.
FB	Finance and Banks	Institutions with financing and banking functionality.
GO	Government	Including states and their related departments.
HC	Healthcare	Including health care providers such as hospitals and clinics.
HT	Hospitality and Tourism	Including hotels, restaurants and etc.
IO	Internet and Online Services	Including chat rooms and forums.
MD	Military and Defence Section	Companies operating in military equipment manufacturing.
MU	Multiple	Several targets.
NN	NGO and No Profit	Including non-profit and charity sectors.
RT	Retail	Including retail shops.
SI	Single Individual	Publicly known figures.
SN	Social Network	Including Facebook, Twitter, Instagram

		and etc.
TC	Telecommunication	Sectors providing telecommunication lines such as the internet and telephone.
THS	Technology Hardware and Software	Companies and business providing hardware or software products.
TM	Terrorism	Terrorist groups
TP	Transportation	Including traffic lights and etc.

9.3 Type of Threat

Acronym	**Type of Threat**	D**efinition**
AH	Account Hijacking	Any online account such as email, social media and etc. associated with a person or a company hijacked by a hacker(s).
DF	Defacement	Unauthorized changing a web page by hackers through

		penetration to the web server.
DS	DDOS	Disturbing availability of victims' server by hackers through sending a high volume of requests.
MWV	Malware	A piece of malicious code including virus, worm, Trojan horse and etc. designed by hackers for compromising victims' system.
PH	Phishing	A malicious method tries to steal sensitive information by deceiving victims through an email conversation,
SQ	SQL injection	Attackers' code try to compromise the database
TA	Targeted Attack	Anonymous and untrackable attackers actively are trying to

		penetrate to victims' system
UA	Unauthorized access	Any unauthorized access to computer devices and software by hackers
UN	Unknown Attacks	Those attacks when Type of Threat has not been reported in OSINT resource.
CSS	XSS vulnerability (Cross Site Scripting)	Attackers inject client-side script into a webpage.
ZD	0day	Unresolved Security bugs get exploited by hackers.

9.4 Sample of Dataset

Date	Author	Threat	Target-Section	Activity	Country
01/01/2013	@th3inf1d3l	SQLi	Broadcasting and Publishi	HA	SA
01/01/2013	JokerCracker	SQLi	Education	CC	IN
01/01/2013	JokerCracker	SQLi	Education	CC	IN
01/01/2013	JokerCracker	SQLi	Government	CC	YE
02/01/2013	DarkWeb Goons	Targeted Attack	TechnologyH&S	CC	US
03/01/2013	DarkWeb Goons	SQLi	NGO and Non-Profit	CC	INT
03/01/2013	Anon_Acid	Unk_attack	Government	HA	US
03/01/2013	Izz ad-Din al-Qassam Cyber Fighters	DDoS	Financial and Bank servic	HA	US
03/01/2013	Izz ad-Din al-Qassam Cyber Fighters	DDoS	Financial and Bank servic	HA	US
03/01/2013	Izz ad-Din al-Qassam Cyber Fighters	DDoS	Financial and Bank servic	HA	US
03/01/2013	Izz ad-Din al-Qassam Cyber Fighters	DDoS	Financial and Bank servic	HA	US
03/01/2013	Izz ad-Din al-Qassam Cyber Fighters	DDoS	Financial and Bank servic	HA	US
03/01/2013	Izz ad-Din al-Qassam Cyber Fighters	DDoS	Financial and Bank servic	HA	US
03/01/2013	Unidentified Hackers	Unk_attack	Entertainment	CC	FR
04/01/2013	Anonymous	Unk_attack	Government	HA	DE
04/01/2013	Unidentified Hackers	Defacement	Government	CC	BE
04/01/2013	Chinese hackers	Targeted Attack	Government	CE	JP
04/01/2013	Unidentified Hackers	Unk_attack	Education	CC	US