

**APPLYING BUSINESS ANALYTICS IN PRACTICE:
DATASET ANALYSIS USING WEKA: “PHISHING WEBSITES DATASET”**

BADAR KHALFAN AL-MAHROUQI

Supervisor: Dr Dmitri Roussinov

**This dissertation was submitted in part fulfilment of requirements for the
degree of MSc Advanced Computer Science**

**DEPT. OF COMPUTER AND INFORMATION SCIENCES
UNIVERSITY OF STRATHCLYDE
SEPTEMBER 2016**

DECLARATION

This dissertation is submitted in part fulfilment of the requirements for the degree of MSc of the University of Strathclyde.

I declare that this dissertation embodies the results of my own work and that it has been composed by myself. Following normal academic conventions, I have made due acknowledgement to the work of others.

I declare that I have sought, and received, ethics approval via the Departmental Ethics Committee as appropriate to my research.

I give permission to the University of Strathclyde, Department of Computer and Information Sciences, to provide copies of the dissertation, at cost, to those who may in the future request a copy of the dissertation for private study or research.

I give permission to the University of Strathclyde, Department of Computer and Information Sciences, to place a copy of the dissertation in a publicly available archive.

(please tick) **Yes** [☒] No [☐]

I declare that the word count for this dissertation (excluding title page, declaration, abstract, acknowledgements, table of contents, list of illustrations, references and appendices is **18,484**

I confirm that I wish this to be assessed as a Type 1 2 **3** 4 5
Dissertation (please circle)

Signature: **Badar**

Date: **01/09/2016**

ABSTRACT

Phishing attacks are one of the sophisticated and most dangerous web threats that Internet users and organisations face today. Phishers utilise social engineering tricks and spoofing techniques in order to impersonate legitimate websites which phishers use to steal personal and private data. The rapid growth of online services and the lack of web security skills held by many Internet users are two main factors that contributed heavily to the continuous success of phishing attacks. Besides stealing personal and private data of users and organisations, phishing attacks continue to impose huge financial damages. They also create a direct negative effect on the reputation of organisations and on the trust of users on the security of online transactions. While different phishing countermeasures such as black lists and anti-phish plugins had been implemented to fight phishing attacks, most of those approaches had limitations and were not very effective. An automatic and on-the-fly evaluation of websites is indeed required to provide users with high protection from phishing attacks. In this dissertation report, I applied business analytics techniques on a dataset that contains different features and characteristics of both legitimate and phishing websites. I used Weka – open source machine learning software – in order to train a classifier and develop a model that is capable of effectively and efficiently distinguishing between legitimate and phishing websites based on their features and characteristics. The results showed that there are some machine learning algorithms that are capable of correctly classifying the websites with an accuracy of over 92%. These models can be utilised to protect Internet users from phishing attacks while they are surfing the Internet.

ACKNOWLEDGEMENTS

I would like to express my great thanks to everyone who contributed – directly and indirectly – to the work described in this dissertation. Without the support of those people, this dissertation report would not exist in its current professional shape. First of all, I would like to thank all my academic instructors who provided me with the guidance and advice whenever I needed. Special thanks to all the academic staff in the department of Computer Science at the University of Strathclyde who were always ready to guide me and clarify things to me.

Second, I would like to thank Dr. David McMenemy and Dr. Martin Halvey who provided me with very useful information regarding the professional ways and the best academic databases to search for information. Through their guidance and insights, I was able to explore the university academic library as well as various academic databases. In addition, they were always ready to provide me with the support on how to use different tools such as the Endnote referencing software for example. I appreciate their insights and support that helped me working on my dissertation.

Third, I would like to thank Dr. Ian Ruthven who was always ready to clarify things to me. I emailed him many times requesting different answers and clarifications on various aspects related to pursuing my dissertation. He was very keen to describe to me the different types of dissertations, and the different qualitative and quantitative research approaches. In addition, he enlightened me with different skills and cleared many doubts for me which made me well prepared to pursue my dissertation.

Finally, I would like to thank my government for giving me the full-covered scholarship to study and acquire the Masters' degree. I would also like to thank my family members who supported me all the time – especially my Dad and my wife. My Dad was always encouraging me to study hard. His best wishes and prayers were inspiring me to perform better. My wife prepared the best studying environment for me at home. She was always ready to sacrifice her time whenever I was busy studying. She took care of the children and prepared food for us. With her great support, I had enough time dedicated for my study.

Table of Contents

1. Introduction	7
1.1 The Growth of Online Services.....	7
1.2 The Internet is Not a Safe Place	7
1.3 The Solution: Business Analytics Can Help	8
2. Literature Review	9
2.1 The Phishing Problem.....	9
2.2 Some Statistics and Facts about Phishing	10
2.3 Phishing Tricks and Techniques	10
2.4 Phishing Countermeasure Approaches	11
2.5 The Suggested Solution: Applying Business Analytics	12
3. The Increasing Adoption of Business Analytics/Business Intelligence Techniques.....	14
3.1 The Interest on Data Mining and Business Analytics/Business Intelligence	14
3.2 Data Sources and Data Collection.....	14
3.3 The Concept of Big Data and Its Relationship to Business Analytics	15
3.4 Business Analytics Infrastructure	16
4. The Dataset	18
4.1 Dataset Name: Phishing Websites Dataset	18
4.2 Description of the Dataset	18
4.3 How the Dataset was Created?.....	18
4.4 Dataset Attributes	19
4.5 Description of the Dataset Attributes	20
5. Weka.....	31
5.1 Weka History	31
5.2 Weka Algorithms and Functionality	32
5.3 Weka Interfaces	32
5.4 Weka: A Platform Independent.....	33
5.5 Downloading Weka	34
6. Loading the Dataset into Weka.....	35
6.1 The Dataset File.....	35
6.2 Loading the Dataset File.....	36
6.3 Reading Dataset Information from the Explorer Window	38

7. Running the Algorithms before Dataset Pre-Processing	40
8. Dataset Pre-Processing Stage	52
8.1 The Importance of Dataset Pre-Processing	52
8.2 Running Some Attribute Evaluators.....	52
8.3 Running the Algorithms with Attributes Selected by the CFS Subset Evaluator	55
8.4 Running the Algorithms with Attributes Selected by the Consistency Subset Evaluator ...	56
8.5 Running the Algorithms with Attributes Ranked by the OneR Attribute Evaluator	58
8.6 Running the Algorithms with Attributes Belonging to the Different Categories of the Features.....	63
8.7 Running the Algorithms with Different Combinations of the Top 5 Attributes	68
8.8 Running the Algorithms with the Top Ranked 5 and 15 Attributes Using Different Parameters' Values.....	69
9. Discussion of the Testing Results and Performance of the Algorithms	77
9.1 The Output and the Results	77
9.2 Performance and Functionality of the Algorithms.....	79
Conclusion.....	82
References	84
Appendix A: List of Figures.....	87
Appendix B: List of Tables	89
Appendix C: Applications of Business Analytics/Business Intelligence	91
C.1 Applications of BA/BI in Commercial Enterprises	91
C.1.1 Business Analytics: the Main Tool to Success.....	91
C.1.2 Shared Characteristics of the Competing Enterprise.....	92
C.2 Applications of BA/BI in the Healthcare Sector	93
C.3 Applications of BA/BI in the Supply Chain Sector.....	94
Appendix D: Weka Datasets Websites and Repositories	95

1. Introduction

1.1 The Growth of Online Services

Today, many companies, banks, governmental institutions and other different public and private organisations are utilising computer and information technology to conduct business. Many banks and companies nowadays have transformed their banking and business services to online web and electronic services in order to make it easier and more convenient for their customers to do business with them (Brynjolfsson and Hitt, 2000). Using the Internet, customers can perform many different tasks electronically from their homes by browsing the intended website using their personal computers. The 24-by-7 availability of most online services contributed heavily on attracting more customers to perform online business. This high availability of services enabled customers to choose the best time appropriate for them to conduct business since they are not restricted to the opening hours of shops and malls (Bellman et al., 1999). On the educational side, most universities worldwide are utilising one or more of the e-learning environments in order to provide learning materials and other services to their students online. Different studies have shown that there is an increase in the adoption of online learning (Anderson, 2008).

1.2 The Internet is Not a Safe Place

Most people worldwide are conducting some online transactions in a way or another. We sometimes do online shopping and pay online by providing details of our identity and our bank cards. Many of us have mobile banking applications installed on our smart phones, and conduct some online banking through those applications. We perform many online transactions in our everyday life more than we were in the past few years. While the authors in (Brynjolfsson and Hitt, 2000) believe that conducting business online has made our life easier and more convenient, we actually should be very careful when doing so. We should keep in mind that the Internet is not a safe place. In fact, different web threats have grown rapidly since the 1990s as the Internet has gained more and more popularity (Mohammad et al., 2015). There are many

malicious users out there who always try to target and deceive other innocent Internet users. Some of those malicious users have advanced computer experience and use different tools and techniques to deceive people. Many Internet users are on the other hand not aware about computer and web security; hence they can be vulnerable to web threats (Alsharnouby et al., 2015, Dahamija et al., 2006).

1.3 The Solution: Business Analytics Can Help

Given the fact that many malicious users are targeting others on the Internet, and due to the fact that most of those targeted people lack the knowledge about computer and Internet security; this research was aimed to help to protect those people while they are surfing the Internet. In this research, I had applied business analytics and machine learning techniques on a sample dataset that contains data regarding multiple features and characteristics of a mixture of legitimate and phishing websites. The goal was to develop an accurate and efficient model that is capable of identifying phishing and legitimate websites with high accuracy based on the websites' features and characteristics. This model can then be utilised and used to protect normal Internet users from phishing attacks by allowing the model to check the website automatically for the users and to notify those users if that website is a legitimate or a phishing website. There are different ways in which this model can be utilised in an automated way. For example, it can be embedded in web browsers as a plugin, or it can be installed as a small program on the users' computers (Kausar et al., 2014, Kirda and Kruegel, 2005). The main goal is to provide an automated, accurate and highly efficient protection for Internet users from phishing attacks with a high accuracy classification decision.

2. Literature Review

2.1 The Phishing Problem

Among the different web threats that Internet users are targeted by, phishing is considered to be one of the most sophisticated and dangerous threats (Mohammad et al., 2013). There are different definitions for phishing out there, but all of them point to the same general idea. According to the Anti-Phishing Working Group (APWG), “Phishing is a criminal mechanism employing both social engineering and technical subterfuge to steal consumers’ personal identity data and financial account credentials” (Anti-Phishing Working Group, 2014b). The APWG is “A non-profit corporation established in 2003 focuses on reducing the frauds resulting from phishing, crime-ware and email deceiving” (Mohammad et al., 2015). Another good definition for phishing is “Creating a fake online company to impersonate a legitimate organization; and asking for personal information from unwary consumers depending on social skills and website deceiving methods to trick victims into disclosure of their personal information which is usually used in an illegal transaction” (Mohammad et al., 2015). Therefore, phishing attacks are designed to steal private personal and financial data which is used by phishers in illegal transactions. This data can also be sold on the black market as well (Mohammad et al., 2015).

There are two main reasons that make phishing attacks very dangerous and very sophisticated. The first reason is that phishing attacks use different tricks and techniques to fool innocent users who are usually not aware of these kinds of attacks. In fact, the nature and style of these attacks are constantly changing which makes it very difficult for these attacks to be detected or prevented; especially during the first few hours of their launch. Therefore, shielding users against these kinds of attacks is a very sophisticated task. The second reason is that phishers are usually targeting sensitive private data such as personal login credentials, bank account details and credit card information. This stolen data is generally used by phishers in illegitimate transactions by impersonating the legitimate users. Phishing can be harmful to both the targeted victims and the organisations in which the related stolen data can be used by phishers to

access their website and perform illegitimate transactions as they are the legitimate users (Alsharnouby et al., 2015, Dahamija et al., 2006, Purkait, 2012).

2.2 Some Statistics and Facts about Phishing

Different studies and agencies have revealed some facts and statistics regarding the negative effects imposed by phishing attacks to individuals, organisations and countries. In fact, with the increase of online investments and online trading, phishing is getting increased as well. A report distributed by the Anti-Phishing Working Group showed that around 128,387 phishing websites were recorded in the second quarter of the year 2014 (Anti-Phishing Working Group, 2014a). This is really a very big and frightening number. In addition, in 2003, an estimated loss of about \$1.2 billion was directly caused for U.S banks and card issuers due to information been given away by around two million users to phishing websites (Dahamija et al., 2006). Other findings of a survey disseminated by Gartner revealed that phishing websites are still increasing and costing the financial sector in the U.S an annual estimated loss of about \$3.2 billion. The same survey reported that around 3.6 million people were victims of these attacks (Mohammad et al., 2015). It can clearly be noticed how the number of victims has increased, as well as the total amount of loss. This is a very clear sign that phishing attacks are still threatening online customers and organisations, and there must be something done to prevent such attacks.

2.3 Phishing Tricks and Techniques

Phishers succeed in fooling people by utilising different technical and social engineering skills and techniques in which they could convince their victims about the credibility of the email or the website they send to their victims. One of the main tricks used by phishers is to send the phishing website within an email which is usually a forged email that appears to be from a friend of the targeted users. It is found that users are more likely to trust emails from their friends and hence they are more likely to click on the links included in the body of that email (Dahamija et al., 2006). Another

technique used by phishers is to send a forged email to the targeted victims urging them to update or validate their account information for an organisation they have an account for. Phishers design those emails carefully and make them appear as they were sent from that organisation itself. The phishing website included in the email body also looks exactly similar to the legitimate website of that organisation. In addition, phishers use some convincing tricks and wording of sentences such as “Update your account details now to avoid account lockout” in order to urge users to proceed on and reveal their personal data by submitting the data on the fraudulent website (Dahamija et al., 2006).

Another aspect that makes the phishing issue very threatening and frightening is the fact that many Internet users lack the security knowledge and skills which can help them protect themselves from phishing attacks. Different academic research and studies found out that many ordinary Internet users do not even know what phishing is, and many of those users would not expect that such kinds of attacks do exist (Dahamija et al., 2006). In fact, even though some users had the web security background that would protect them from such attacks, they still got fooled due to their bounded attention which is usually directed to the primary task while surfing the Internet; and not on the security side which is considered as a secondary task for them. Also, phishing websites with high quality design, especially from the graphical and visual side, are more likely to fool even those users with some security experience (Alsharnouby et al., 2015, Dahamija et al., 2006).

2.4 Phishing Countermeasure Approaches

Until today, there have been many different approaches implemented to countermeasure phishing attacks. Some of those approaches require manual intervention, while others work automatically. All of them can be classified into three main categories; legal, educational and technical approaches. Legal approaches have been implemented by different countries like the U.S and the U.K in which convicted criminal phishers are brought to justice and get prosecuted (Mohammad et al., 2015). Educational approaches can be very effective if Internet users are trained well and primed to always direct their attention to the security stuff when surfing the Internet.

However, it was found that many users are ignorant and can still be fooled (Mohammad et al., 2015, Purkait, 2012). Technical approaches have been used widely since the beginning of phishing attacks. They are designed and implemented in different ways. The most common forms of technical countermeasures are black lists, user polling, heuristics, server-side security alarms and indicators, client-side browser plugins and toolbars and many other different forms (Kausar et al., 2014, Kirda and Kruegel, 2005, Mohammad et al., 2015, Shekokar et al., 2015). Each of those technical countermeasures has limitations and some of them were found not to be very effective; mainly because phishing attacks are constantly changing in style and techniques and they usually live out there for only few hours or days. Phishers on the other side are constantly designing new phishing tricks and are targeting the weakest point in the chain, the users (Purkait, 2012). Therefore, the use of one single approach alone to prevent the phishing problem is not sufficient. To overcome those limitations, there should be an intelligent automated solution that is capable of analysing a website on the fly and act right away against phishing attacks to protect users.

2.5 The Suggested Solution: Applying Business Analytics

While the most main objective of the use of business analytics nowadays is to extract knowledge and value from data in order for organisations to achieve competitive advantages, business analytics can be used for different purposes since its main goal is to help in making better and faster decisions. Different studies have shown that utilising business analytics in making decisions has improved the competitive performance of many companies and organisations (Chen et al., 2012). In fact, the main factor for the successful utilisation of business analytics is data. Recently, data has been described as “the new oil” (Acito and Khatri, 2014, Mithas et al., 2013). Business analytics algorithms can be applied on any data in order to extract knowledge and value out of that data. Unlike before, utilising business analytics in recent years has become much easier and less expensive as the cost of computing power and storage devices have decreased dramatically (Chaudhuri et al., 2011).

Therefore, the aim of this research is to utilise business analytics algorithms and techniques in order to develop an automated, intelligent and highly efficient and accurate model or solution for the phishing problem. The general idea is to use business analytics algorithms to analyse the data extracted from websites in order to decide whether a specific website is a phishing or a legitimate website. The extracted data is a mixture of different features and characteristics of the websites (Shekokar et al., 2015, Zhang et al., 2014). Applying business analytics in this field can be very beneficial from different aspects. First, the developed model can overcome some of the above mentioned limitations of the different countermeasures discussed previously. Second, it can protect the normal Internet users who have limited or no security knowledge from falling victims to phishing attacks. Third, it can work automatically on the fly by analysing the websites as users are browsing the Internet. It should notify users whether the website to be visited found to be a suspicious phishing website or a legitimate one. The goal is to automatically protect Internet users from phishing attacks with high accuracy classification of websites while users are surfing the Internet.

3. The Increasing Adoption of Business Analytics/Business Intelligence Techniques

3.1 The Interest on Data Mining and Business Analytics/Business Intelligence

Over the past 20 years, the interest towards business analytics and data mining field has grown rapidly. Many enterprises, organisations and governmental institutions have adopted different tools and techniques related to data mining, business analytics and business intelligence. The main driver for this great interest was to extract knowledge out of row data in order to help managers and executives in making faster, accurate and wise decisions. Such faster and accurate decisions helped many enterprises to gain competitive advantages and become one of the market leaders (Chaudhuri et al., 2011).

3.2 Data Sources and Data Collection

Row data is one of the main components for the successful adoption of data mining and business intelligence techniques. Enterprises and organisations collect data from different sources in order to make decisions. For example, many commercial enterprises gather data about their customers online through the use of web user accounts. Big enterprises such as Amazon and eBay collect huge volumes of data about their customers online and keep this data on databases and data warehouses. Users of Facebook and Twitter generate millions of tweets and upload millions of images every day (Vossen, 2013). Other companies and retail shops use loyalty cards as their data collection technique. Each loyalty card is usually registered for one customer. Whenever this customer goes for shopping and uses his/her loyalty card, all the shopping information gets recorded in the retailer systems (Graeff and Harmon, 2002). In fact, the loyalty cards are not just used for shopping and retail business, but are also used in casinos, coffee shops, sport clubs, pharmacies and so on. Banks on the other hand store a lot of information about their customers. Such information include the personal details, the home address, the number and type of bank accounts and other different details that can be used to retrieve knowledge (Davenport, 2006). In

addition, many commercial enterprises and governmental organisations employed new technologies such as RFID devices, sensors and imaging devices in order to make it easier to capture, read and store information. These devices collect huge volumes of data every day (Bryant et al., 2008). Other sources of data nowadays include product reviews, online forums and social media websites and mobile applications such as Facebook and Twitter (Chaudhuri et al., 2011). Recently, new small fitness and personal tracking devices have become major sources of data. Some of the very well-known devices include Fitbit and Jawbone Up (Barrett et al., 2013). These devices collect a lot of data about the physical activities of people.

3.3 The Concept of Big Data and Its Relationship to Business Analytics

Big Data is a new concept that has been used as a term to describe the current digital data available worldwide. The term has been used to illustrate the nature and size of the data produced nowadays. In fact, Big Data is always described by the 3 Vs (volume, variety and velocity) or the 4 Vs adding (veracity) to the previous 3 (Vossen, 2013). Many Tera bytes or Peta bytes of digital data are generated every day (Bryant et al., 2008). This is indeed a very huge volume of data. Not only the size, but this data is produced in different forms. There is text data as well as images, audio files, videos and many other formats of data. While some of this data is structured in tables and databases, other data is out there unstructured as in many websites, blogs and forums. This is a very clear indication on how various formats of data is available today. The velocity in which this huge volume of data is generated is not like before. A lot of data is transferred worldwide in seconds, especially after the invention of the web. In addition to all of this, the authenticity and reliability of the data available today is another common issue. These are the main common characteristics of the data available today (Vossen, 2013).

All of those characteristics of the Big Data have created a big challenge for enterprises and organisations on how to deal with such enormous and various types of data. Unlike today, most of the data few years ago was mainly stored in local databases in very low volumes. It was very easy to manage and maintain such data. It was also

easier to run some data mining or business analytics algorithms on that data in order to extract knowledge. In some cases, even the local databases' built-in knowledge discovery tools could do the task. For example, Oracle and IBM BD2 databases come with built-in data mining tools. Oracle has its ODM (Oracle Data Miner), and IBM DB2 has its Intelligent Miner tool (Aggarwal et al., 2012). These built-in data mining tools were suitable and did the work for such small databases. However, they actually cannot handle the huge volumes and enormous formats of the data available today. The Big Data has caused many enterprises and organisations to renovate their data-handling technologies. This huge data requires special technologies and infrastructure in order for enterprises and organisations to be able to extract knowledge out of it in a reasonable amount of time (Bryant et al., 2008, Chaudhuri et al., 2011).

Even though Big Data has challenged enterprises and organisations with new requirements, it has also – on the other hand – brought them many valuable advantages. Big Data enables enterprises and organisations to gain a deeper insight about their customers. In the entertainment sector for example, Disney Parks and Resorts could provide their customers with up-to-date information about their existing offers through their website. More than that, the Disney characters within the parks could greet their customers by their names (Vossen, 2013). This is just to imagine how fast and continuously live data is being gathered and processed. Other recommender systems from Amazon and MoveiLens for example process huge volumes of data and run very sophisticated queries in order to recommend to their customers the products that are more likely to be accepted and purchased. Such data include information about their customers like age, income, gender and marital status , and information about their products as well (Adomavicius and Tuzhilin, 2005). Big Data can bring enterprises and organisations very competitive advantages if managed and utilised properly and wisely.

3.4 Business Analytics Infrastructure

The nature of the data available today requires special infrastructure and technologies in order to get the best advantages of business analytics in extracting knowledge. This of course includes software and hardware technologies as well as

human skills and experience. Unlike before, the huge amounts of data generated nowadays cannot be managed and processed by the normal traditional databases and application tools (Sahay and Ranjan, 2008). This data requires high performance computing infrastructure; which includes both software and hardware. For the software side, applications and tools that support sophisticated queries and dashboards should be in place. Specialised data mining and business analytics algorithms are required as well. On the hardware side, servers with high computing power, storages with huge disk space and networking devices that support high rates of bandwidth are necessary to handle and process Big Data (Chaudhuri et al., 2011, Bryant et al., 2008). For example, data warehouses are being used heavily nowadays to store millions of Pita bytes of data. In fact, within the servers' infrastructure only, there have been groups of servers designated for specific tasks. For example, there are clusters – multiple servers grouped together – that are designed for preparing the data for the different processes of business analytics. These servers are called the Extract-Transform-Load (ETL) engines. Other servers are designated for the Online Analytic Processing (OLAP) process; where multidimensional views are generated and made ready for special analytical tasks such as filtration and aggregation (Chaudhuri et al., 2011). All of these and other new technologies are indeed necessary to handle Big Data, and to be able to extract the best knowledge out of this row data.

4. The Dataset

4.1 Dataset Name: Phishing Websites Dataset

The name of the dataset is the “Phishing Websites Dataset”. It has data about mixture of legitimate and phishing websites. The attributes of the dataset represent the features of the websites; which are used to distinguish between legitimate and phishing websites. This dataset is obtained from the UCI Machine Learning Repository.

Link of the dataset: <https://archive.ics.uci.edu/ml/datasets/Phishing+Websites>

4.2 Description of the Dataset

The Phishing Websites Dataset has been created after many researchers have struggled in finding a reliable dataset in this area as the authors of the dataset described (Mohammad et al., 2012, Mohammad et al., 2014a). It is published on the UCI Machine Learning Repository to allow researchers and any interested people to explore the features of the phishing websites, and to do some tests on it. Even though there has been no absolute agreement on the features of the phishing websites in literature, the owners of this dataset tried to include the most important features when they created the dataset.

4.3 How the Dataset was Created?

The Phishing Websites Dataset is one of the very few publicly available datasets that are related to the phishing issue. To create this dataset, the authors (Mohammad et al., 2012, Mohammad et al., 2014a) gathered thousands of legitimate and phishing websites from the PhishTank archive; which is a public anti-phishing website (“<https://www.phishtank.com>”) that keeps records of phishing information, data and websites. In order to create the instances of the dataset, the authors created and used automatic feature-extraction tools. For example, they used JavaScript programs and PHP scripts to extract many features from the collected websites. They also created other scripts that connected to remote databases such as Alexa (“www.alexa.com”) and

WHOIS (“https://who.is/”) databases in order to extract other features. In some cases, they had to automatically extract other features using the source code of the gathered websites (Mohammad et al., 2012). Each feature was then assigned a weight based on its frequency compared to the overall frequency in the data collection. Based on the weights of all the features of each instance or record in the dataset, a final result was recorded for each instance indicating whether this instance is a legitimate or phishing website.

4.4 Dataset Attributes

The dataset has 31 different attributes. The first 30 attributes represent the features of the websites; which are actually the attributes in the dataset. The last attribute is the result; which tells whether the website is a phishing or a legitimate website. All the attributes have nominal values as 0, 1 and -1. Table 1 below provides the details of the dataset attributes.

No.	Attribute Name	Possible Attribute Values
1	having_IP_Address	1 , -1
2	URL_Length	1 , 0 , -1
3	Shortining_Service	1 , -1
4	having_At_Symbol	1 , -1
5	double_slash_redirecting	1 , -1
6	Prefix_Suffix	1 , -1
7	having_Sub_Domain	1 , 0 , -1
8	SSLfinal_State	1 , 0 , -1
9	Domain_registration_length	1 , -1
10	Favicon	1 , -1
11	Port	1 , -1
12	HTTPS_token	1 , -1

13	Request_URL	1 , -1
14	URL_of_Anchor	1 , 0 , -1
15	Links_in_tags	1 , 0 , -1
16	SFH	1 , 0 , -1
17	Submitting_to_email	1 , -1
18	Abnormal_URL	1 , -1
19	Redirect	1 , 0
20	on_mouseover	1 , -1
21	RightClick	1 , -1
22	popUpWidnow	1 , -1
23	Iframe	1 , -1
24	age_of_domain	1 , -1
25	DNSRecord	1 , -1
26	web_traffic	1 , 0 , -1
27	Page_Rank	1 , -1
28	Google_Index	1 , -1
29	Links_pointing_to_page	1 , 0 , -1
30	Statistical_report	1 , -1
31	Result	1 , -1

Table 1: Details of the dataset attributes.

4.5 Description of the Dataset Attributes

Since the emergence of the phishing problem few years ago, phishers had used different tricks and techniques to deceive their victims and allure them to visit the phishing websites which phishers had created. Therefore, different phishing websites have different features and characteristics. This section is intended to explain the dataset attributes and to summarize the most common features of the phishing websites – which are the dataset attributes – so the reader can be aware of such features and

characteristics. These features and characteristics are found most of the time on the phishing websites, but not on the legitimate ones. In fact, different research studies had classified the features of phishing websites into groups or categories (Mohammad et al., 2014a, Alkhozai and Batarfi, 2011, Mohammad et al., 2013). In their experimental study, the authors in (Mohammad et al., 2014a) had classified those features into four main groups; features based on the address bar, features based on abnormality, features based on the HTML and JavaScript techniques and features based on the domain of the website itself.

Group 1: Features based on the address bar

a) Using an IP Address (attribute 1: having_IP_Address)

In most of the cases, if an IP address is used in the address bar of a website instead of the domain name, then this website is more likely to be a phishing website. An example of this type of websites can look like “http://110.22.95.13/index.html” instead of using the domain name such as “http://www.google.com” for example (He et al., 2011). While there are some few domain-registered phishing websites, most of the phishing websites do not get registered to avoid the domain registration cost. In fact, some phishers can even use the hexadecimal version of the IP address to trick the users. Such websites can have an address bar like “http://0x62.0xCA.0xAA.0x52/index.html” (Mohammad et al., 2014a). Internet users should be aware of this kind of IP address tricks used instead of the domain name.

b) Using Long URL (attribute 2: URL_Length)

Sometimes, phishers tend to use very long URLs in order to hide the suspicious part of the phishing website which is embedded within the domain name. For example, phishers can trick users by using long URL links that look like the below link: “http://company.com/3f/aze/ch43e2e369e73902f416dbe79856frdj3a5e/?cmd=_home&file=192334d56gk4d9b4wx53l4f8dc1e7c2e004d58f9u75gee321e7c2e8dd4108mme2@PhishingWebsite.html” (Mohammad et al., 2013). Legitimate URLs normally look like

“http://www.google.com” or “http://www.strah.ac.uk”; and the link itself has meaning and can be read by the user. Usually, normal URL length is between 10 to 50 characters. In one of the studies that examined the phishing and legitimate websites, the results showed that if the length of the URL exceeds 54 characters long, then the website is more likely to be a phishing website (Mohammad et al., 2014a).

c) Using Short URL (attribute 3: Shortening_Service)

While some phishers tend to use very long URLs, other phishers can still go the opposite way and use very small or tiny URLs. This method is called “URL Shortening” in which phishers use very small URLs that eventually lead to other web pages. Phishers can accomplish this trick by utilizing the HTTP Redirect service. Tiny URLs can look like “bit.ly/73XRRdk6” and generally redirect the user to the intended phishing website (Gastellier-Prevost et al., 2011).

d) Using the “@” symbol within the URL (attribute 4: having_At_Symbol)

The functionality of the “@” symbol in web browsers can help phishers to trick Internet users. Using the “@” symbol within the URL makes the browser to ignore the previous part of the URL before the “@” symbol. Usually, phishers use this technique to append the intended phishing website link after the “@” symbol. A highly-used example of this type would look like “http://www.amazon.com@http://www.phishingSite.com” (He et al., 2011). In this case, the browser will take the user to the phishing site instead of amazon.com. In fact, many ordinary Internet users do not pay attention to these kinds of tricks, and some users do not even know about them (Alsharnouby et al., 2015, Dahamija et al., 2006).

e) Using the “//” symbol for redirection (attribute 5: double_slash_redirecting)

Another common technique exercised by phishers is to use the “//” – double back slash – within the URL link in order to redirect the users into another web page; which is usually the phishing website. URL links of this type would usually look like: “http://NormalWebsite.com//http:PhishingWebsite.com”. In this situation, the second “//”

in the URL link will redirect the user to the second website in the URL; which is the intended phishing website (Gastellier-Prevost et al., 2011).

f) Using the “-” symbol with prefix or suffix (attribute 6: Prefix_Suffix)

Most domain names of the legitimate websites usually do not contain the dash “-” symbol within the URL. Phishers use the “-” symbol with a prefix or a suffix and attach it to the original name of the legitimate website, so the domain name of their phishing website will look similar to the original website. Many users do not actually pay attention to the domain names and they easily get tricked. Example of this kind of tricks can look like this link: “http://www.login-hotmail.com” (Mohammad et al., 2014b).

g) Using Multiple Subdomains (attribute 7: having_Sub_Domain)

It is natural to see URL links and domain names with other subdomains. For example, the URL link for Strathclyde University is “http://www.strath.ac.uk”. The domain name here is “strath” which has two subdomains; the “ac” for academic and the “uk” for the country. These two subdomains are separated by one dot. Most legitimate websites and domain names have only one or two subdomains separated by a dot. On the other hand, most phishing websites have more than two subdomains; and hence the number of dots separating the subdomains is more than one. The number of dots separating the subdomains can be used to distinguish between legitimate and phishing websites (He et al., 2011).

h) Using HTTPS (attribute 8: SSLfinal_State)

HTTPS is the Hyper Text Transfer Protocol (HTTP) with Secure Sockets Layer (SSL). This protocol is intended to provide security and encryption for the data transferred through the web. Using HTTPS gives users the impression that the website they dealing with is a secure and legitimate website. However, the existence of HTTPS alone on the domain name does not always guarantee security and legitimacy. A very important component that users should pay attention to, is the SSL certificate which is used by that website. Users should check the certificate and make sure that the

certificate is valid and issued by a trusted Certificate Authority. Top trusted Certificate Authorities include Comodo, GeoTrust, VeriSign, DigiCert and Doster. Another important aspect regarding certificates is the age of the certificate. Usually, the minimum age of a well trusted certificate is two years (Mohammad et al., 2012, Shahriar and Zulkernine, 2012).

i) Period of the Domain Registration (attribute 9: Domain_registration_length)

It was found that most phishing websites live only for few hours or few days (Aburrous et al., 2010). Therefore, the domains of most phishing websites get registered for less than a year, or not registered at all. On the other hand, the domains of most legitimate websites get registered for at least two years (Mohammad et al., 2014a).

j) Using Favicon (attribute 10: Favicon)

Many websites use favicons as graphical images associated with their web pages. These favicons are used in the address bar as a graphical identity and also as a graphical reminder on different web browsers and newsreaders. However, a lot of phishers use favicons of other legitimate websites and link them to their phishing websites in order to deceive users (Herzberg and Gbara, 2004). Internet users should be careful when dealing with favicons. They need to make sure that favicons are associated to their original legitimate websites and not to other external phishing websites.

k) Using Non Standard Ports (attribute 11: port)

As we know, each service running on a computer or a service is usually attached to a specific standard port. For example, the standard port number of the HTTP service is port 80. Therefore, the port of any running service should be made open. However, it is always recommended to change the standard port number of the most common services such as HTTP, SSH, FTP and so on. This is to prevent attacks and intrusions on these services through their standard open ports. Another important point is to make

sure to shut down any un-required service in order to prevent such attacks on it (Gastellier-Prevost et al., 2011).

l) Adding the HTTPS token in the domain (attribute 12: HTTPS_token)

The existence of the HTTPS token gives the users some impression of the security and legitimacy of the website they are dealing with. However, users can be tricked easily on this point. Phishers can add the HTTPS token to the domain part and make the URL link look like using the HTTPS protocol, but actually it is not. URLs of this kind of trick can look like “http://https-www.abc.com” (Mohammad et al., 2013). Users should pay a lot of attention to the domain part of any URL, and be aware of these kinds of tricks.

Group 2: Features based on Abnormality

a) Request URL (attribute 13: Request_URL)

In most legitimate websites, the web page address and most of the contained objects within that website are coming from the same domain. Such objects include images, sounds, videos and other multimedia content. On the other hand, these kinds of objects within the phishing websites do not share the same domain with the web page address. Usually, phishers upload these objects from other external websites (Alkhozae and Batarfi, 2011). Therefore, the request URL feature examines whether these objects are coming from the same domain (in legitimate websites) or from different domain (in phishing websites).

b) URL of Anchor (attribute 14: URL_of_Anchor)

Similar to the Request URL feature, using anchors can trick users as well. Anchors are items or objects defined by the <a> tag in which phishers can embed external objects within the tag. These embedded objects are usually not coming from the same domain as of that of the webpage itself. The contents of the <a> tag and other tags as

well can help in distinguishing legitimate websites from phishing websites (Alkhozai and Batarfi, 2011).

c) Links within Tags (attribute 15: Links_in_tags)

Tags such as the <Meta>, <Script> and <Link> tags are commonly used in most websites. These tags are very useful and provide different functionalities. For example, the <Meta> tags can be used to provide metadata and some information about different documents. The <Link> tags can be used to refer to other web pages or other web resources. In fact, these tags are not used heavily in legitimate websites. In addition, most of the content of these tags within the legitimate websites usually refer to the same domain as that of the website itself. However, the case is different in the phishing websites. These tags get used heavily, and in most cases, their contents refer to other external resources or domains that are different than that of the website itself (Gastellier-Prevost et al., 2011).

d) Using Server Form Handler (attribute 16: SFH)

Server Form Handlers can be used in websites to submit information. However, phishers can exploit this feature to embed their phishing webpages into the server form handlers. Many users do not pay attention or do not know about these kinds of tricks (He et al., 2011).

e) Sending information to an Email (attribute 17: Submitting_to_email)

Many phishing websites use web forms that ask the users to submit their personal information or any other private information, and then redirect this private information to the personal mail box of the phisher. In this technique, phishers use different functions within the web form in order to achieve their goal. Such functions include the server-side mail() function and the client-side mailto() function (Alkhozai and Batarfi, 2011). Users should be careful when submitting their personal information, and make sure their information is not redirected to personal mail boxes.

f) Abnormal URL (attribute 18: Abnormal_URL)

Usually in the legitimate websites, the URL link includes the host name of the website. This is not the case with phishing websites. The URL links of the phishing websites can include any host name; depending where the website is hosted (Mohammad et al., 2012).

Group 3: Features based on HTML and Java Scripts

a) Website Redirecting (attribute 19: Redirect)

It was found that most of the phishing websites get redirected more than three or four times. This is not the case with legitimate websites where the website gets redirected at most one or two times only (Mohammad et al., 2014a). Phishers use the website redirection feature to trick users in order to convince them about the legitimacy of their phishing websites.

b) Using Events to Change the Status Bar (attribute 20: on_mouseover)

Some phishers can trick users by customizing the status bar. Phishers can use some Java Script code within the status bar to show the users some forged URLs. If users do not pay attention to these fake URLs, they will be victims to the intended phishing websites. In some cases, users may need to check the source code of the web page to find out which website they are redirected to. While going through the source code, users might need to look for some events that phishers may use to make changes on the status bar. One of the popular events used by phishers is the “onMouseOver” event. The actual functionality of these kinds of events can be checked by browsing the source code of any web page (Mohammad et al., 2012).

c) Disabling the Right-Click Functionality (attribute 21: RightClick)

Many phishers may enable or disable different features or functionalities in order to deceive their victims. For example, some phishers can use Java Script code to disable the right-click functionality in their websites in order to prevent users from viewing the

source code. In fact, the source code of any website can be used to retrieve different information (He et al., 2011). Such information includes actual details about the functionality of the website; which phishers intend to hide. However, if the right-click functionality is disabled intentionally, then users might not be able to view the source code; and hence there is more likely something bad is going on behind the scenes in which users should be careful with (Mohammad et al., 2012).

d) Using Pop-up Windows to Submit Information (attribute 22: popUpWidnow)

It is not very common in most legitimate websites to use pop-up windows to ask users to submit their personal information. In fact, this feature is used in some legitimate websites to show users some useful information or warn them about other things. However, phishers may use pop-up windows to ask users to submit their personal information (Alkhozai and Batarfi, 2011). Users should be careful when they encounter websites that use pop-up windows.

e) Using IFrame for Redirection (attribute 23: Iframe)

IFrame is one of the HTML tags that can be used to display other webpages in a website. Phishers abuse this feature by hiding the borders of the IFrame and make the additional external webpage looks as part of the original website. These additional external webpages are usually the intended phishing websites where users get redirected to (Alkhozai and Batarfi, 2011).

Group 4: Features based on the Domain

a) Age of the Domain (attribute 24: age_of_domain)

As I mentioned before, most phishing websites do not live for long time. They are designed to stay up there for short period of time in order to deceive as many users as possible before getting caught. In fact, some phishing websites live only for few hours. On the other hand, legitimate websites usually live for long period of time. Legitimate websites can live from 6 months up to couple of years. Therefore, users should pay

close attention to the age of the domain of any website they deal with (Mohammad et al., 2013).

b) DNS Record Status (attribute 25: DNSRecord)

The DNS records of most legitimate websites get registered in the Whois database. However, the DNS records of the phishing websites usually are not found in this database. They are either empty or not found at all. It is always recommended that users check this database; especially if they have any doubt about the legitimacy of that website (Mohammad et al., 2013).

c) Web Traffic (attribute 26: web_traffic)

Web traffic analysis service is provided by different companies such as Google and Alexa. These companies keep track on websites over the Internet and calculate the web traffic on each of them. Website traffic indicates how much a specific website is browsed and visited by Internet users. Unlike legitimate websites, phishing websites have usually very low web traffic due to their short period of living (Mohammad et al., 2012).

d) Website and Page Rank (attribute 27: Page_Rank)

The rank of any website or any specific web page is determined by the number of visitors to that website or that page, or by the number of other web pages linking to that website or page. Usually, popular websites get more visitors and are linked to by many other websites. These websites are legitimate websites in most of the time. On the other hand, phishing websites do not get as many visitors as the legitimate websites. One clear reason is because that phishing websites do not live for a long period of time (Mohammad et al., 2012).

e) Found on Google Index (attribute 28: Google_Index)

Most legitimate websites get indexed by Google once they are published on the Internet. After any website is indexed by Google, then it gets displayed on the search

results if any of the search keywords is part of that website or its contents. On the other hand, phishing websites usually do not get indexed by Google because they live for a very short period of time. Sometimes, checking Google indexed websites can help in differentiating between legitimate and phishing websites (Whittaker et al., 2010).

f) Links to Website (attribute 29: Links_pointing_to_page)

Similar to the website and page rank, the number of web pages linking to a specific website can indicate whether that website is a legitimate or a phishing website. In most cases, while legitimate websites have a big number of other websites linking to them, phishing websites have a very small number or nothing at all linking to them (Mohammad et al., 2012).

g) Statistics and Reports (attribute 30: Statistical_report)

There are different official organisations like PhishTank and StopBadware that provide periodical statistics and reports about websites on the Internet. These organisations provide different information about the websites. Part of this information is about the legitimacy of the websites. Users can follow the statistics and reports published by these organisations in order to be aware about the legitimacy of the websites they browse (Whittaker et al., 2010).

5. Weka

5.1 Weka History

The Weka (Waikato Environment for Knowledge Analysis) project started in the early 1990s. During that time, it was not easy for researchers to get access to machine learning tools and techniques (Hall et al., 2009). Most of the available tools at that time were not unified and not state of the art tools. Those tools were written in different languages and did not support all the platforms. It was a very tedious task for researchers and businesses to work on those tools. The Weka project was intended to overcome those issues and to help researchers get access to a state of the art machine learning software. Unlike the previously available tools, Weka was designed by the Java programming language; which is a platform independent language. In addition, Weka was published as part of the free and open source software. Weka gained great popularity because of these factors. Weka was not just free to use, but also its source code could be modified and improved by developers. In fact, being part of the open source software has contributed effectively on the great success of Weka since its launch. More than 1.4 million downloads were recorded by the Source Forge – the website hosting Weka – since Weka was hosted there in April 2000 (Frank et al., 2005, Hall et al., 2009).

Weka was sponsored by the New Zealand government when the project was started. The main goals of the Weka project were stated in (Hall et al., 2009) as *“The programme aims to build a state-of-the-art facility for developing techniques of machine learning and investigating their application in key areas of the New Zealand economy. Specifically we will create a workbench for machine learning, determine the factors that contribute towards its successful application in the agricultural industries, and develop new methods of machine learning and ways of assessing their effectiveness”*. The first few versions – versions 2.1 to 2.3 – of Weka were written mainly in C and Prolog programming languages. However, due to some complexities in supporting the application libraries and in managing the various dependencies faced by the developers as well as the installation burden faced by users, it was decided to rewrite the whole

application in Java. It was a very risky decision given the fact that Java was still in its early years. However, Java's platform independence feature was the main driver for this decision. The first Java version of Weka was version 3.0 which was released in 1999. From this version and on, many new algorithms, features and enhancements have been added to Weka. As of July 2016, the stable version of Weka is version 3.6, the latest stable version is version 3.8 and the development version is version 3.9 (Waikato).

5.2 Weka Algorithms and Functionality

Weka is supplied with different comprehensive machine learning algorithms. Weka has many algorithms that can do some classification, regression, clustering and association tasks. The algorithms in Weka enable researchers and any other interested users to work on different datasets. Users can perform data processing and try different machine learning algorithms from the same window. They can test and compare the results of multiple classification and regression methods. Users can even go further and test different values for the parameters of each algorithm in order to see the effects on the output they get. In addition, Weka provides different data pre-processing tools as well as other graphical and visualisation tools (Frank et al., 2005, Hall et al., 2009).

5.3 Weka Interfaces

Weka has different user interfaces that allow users to interact with the system in various flexible ways (Frank et al., 2005). They are designed to make it easy for users to work with Weka and reach its various functionalities. The main interface in Weka is the Explorer Window. Through this window, users can upload the dataset files into Weka. Weka supports different data sources and data files. For example, data can be loaded into Weka from web URLs, databases or physical files. The file formats supported by Weka include the ARFF (Attribute-Relation File Format) files, the CSV (Comma Separated Values) files, the LibSVM (Library for Support Vector Machines) files and the C4.5 files (Hall et al., 2009, Chang and Lin, 2011).

The Explorer Window itself has also multiple panels where there are different available functions on each panel. As of Weka version 3.6, there are 6 panels on the Explorer Window; which are the Preprocess, Classify, Cluster, Associate, Select Attributes and Visualize. The Preprocess panel enables users to perform some data cleaning, data filtering and feature selection tasks. The Classify panel includes many classification and regression algorithms which users can run and test. The test results are shown on the right side as text. However, the results can be viewed graphically as in the case with decision trees and ROC curves for example. In addition, the resulted model can be saved and loaded again at any other time. The Cluster panel provides users with some clustering tools and techniques. The Associate panel includes different association and rules functions. The Select Attributes panel enables users to run different attribute evaluators that can suggest what best attributes to include for testing. The Visualize panel provides some graphical scatterplots based on the statistics and facts from the dataset. These graphical scatterplots and representations provide some insights about the dataset. Within a scatterplot, users can even drill down and view more details about a specific data point in that scatterplot (Frank et al., 2005, Hall et al., 2009).

5.4 Weka: A Platform Independent

Weka was redesigned to be platform independent software. This means that Weka can run on different operating systems such as Windows, UNIX and Linux. The only important thing that should be available on the client's machine is the JVM (Java Virtual Machine) that is compatible with that specific operating system. Another important point to mention here is to reserve a good amount of heap memory in order for Weka to run smoothly without getting stuck in the middle of the testing. This is especially important when working with big datasets or datasets that have many attributes. This is important also for some algorithms that perform a lot of heavy data processing tasks (Frank et al., 2005, Hall et al., 2009).

5.5 Downloading Weka

Weka is available for download from the University of Waikato website as well as from the Source Forge website. The URL links for both websites are included below. The Weka software is available for Windows, Linux and Mac operating systems. For each operating system, the software is available for both the 64-bit and 32-bit platforms. There are also other websites where Weka can be downloaded (Waikato).

Weka Download URLs:

The Waikato University website:

<http://www.cs.waikato.ac.nz/ml/weka/downloading.html>

The Source Forge Website:

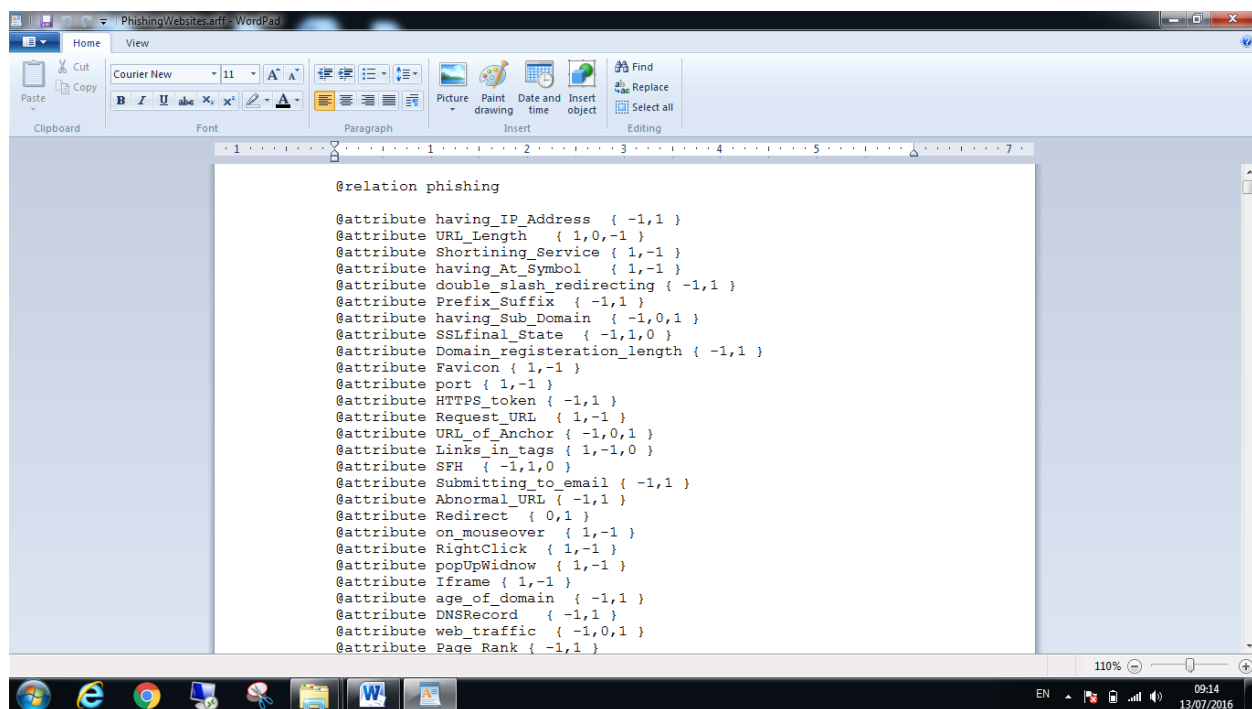
<https://sourceforge.net/projects/weka/>

For this dissertation, I used Weka version (3.6).

6. Loading the Dataset into Weka

6.1 The Dataset File

I downloaded the dataset file from the UCI Machine Learning Repository website. I called the file “PhishingWebsites.arff”. Weka can deal with different types of files; including the ARFF (Attribute-Relation File Format) files. Therefore, the dataset file I have downloaded is ready to be directly loaded into Weka. In fact, the ARFF files should be written in a certain structure that Weka can read and process. Figure 1 illustrates how the dataset attributes and their values are written in the ARFF file. Figure 2 illustrates how the actual attributes’ values of the dataset are written in the ARFF file.



The screenshot shows a WordPad window titled "PhishingWebsites.arff - WordPad". The text inside the window is as follows:

```
@relation phishing

@attribute having_IP_Address { -1,1 }
@attribute URL_Length { 1,0,-1 }
@attribute Shortining_Service { 1,-1 }
@attribute having_At_Symbol { 1,-1 }
@attribute double_slash_redirecting { -1,1 }
@attribute Prefix_Suffix { -1,1 }
@attribute having_Sub_Domain { -1,0,1 }
@attribute SSLfinal_State { -1,1,0 }
@attribute Domain_registration_length { -1,1 }
@attribute Favicon { 1,-1 }
@attribute port { 1,-1 }
@attribute HTTPS_token { -1,1 }
@attribute Request_URL { 1,-1 }
@attribute URL_of_Anchor { -1,0,1 }
@attribute Links_in_tags { 1,-1,0 }
@attribute SFH { -1,1,0 }
@attribute Submitting_to_email { -1,1 }
@attribute Abnormal_URL { -1,1 }
@attribute Redirect { 0,1 }
@attribute on_mouseover { 1,-1 }
@attribute RightClick { 1,-1 }
@attribute popUpWidnow { 1,-1 }
@attribute Iframe { 1,-1 }
@attribute age_of_domain { -1,1 }
@attribute DNSRecord { -1,1 }
@attribute web_traffic { -1,0,1 }
@attribute Page_Rank { -1,1 }
```

Figure 1: The dataset attributes and their values in the ARRF file.

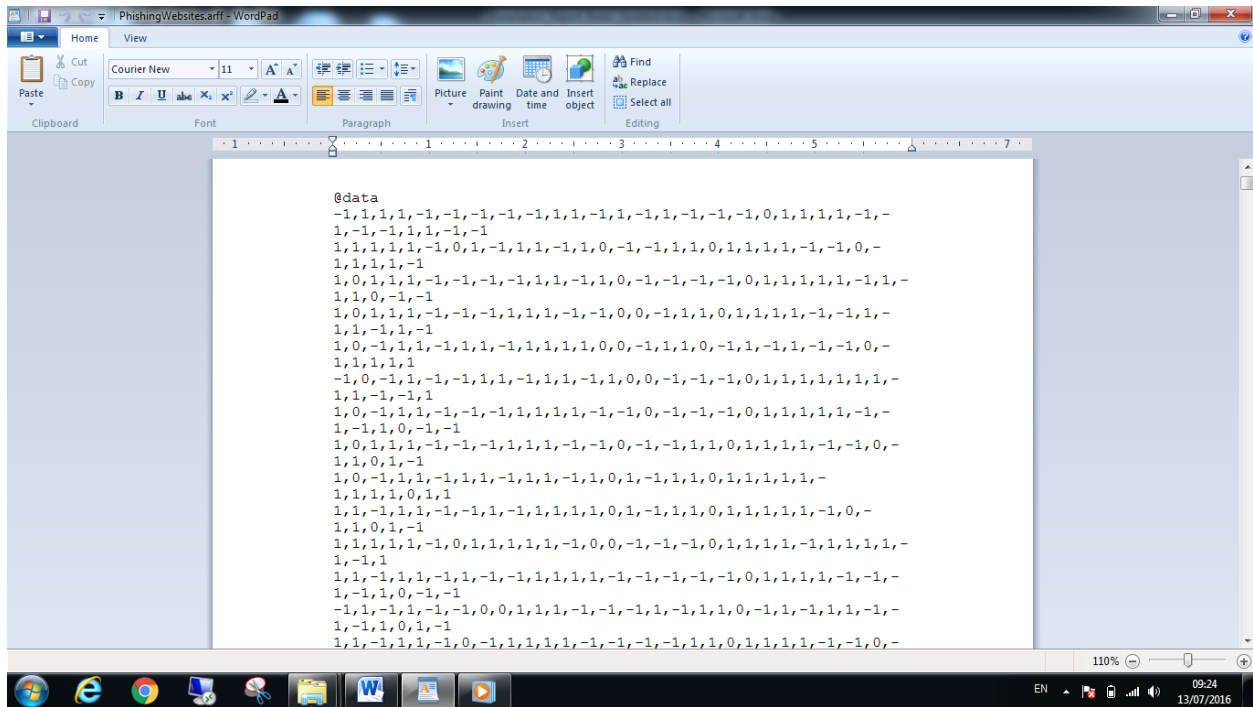


Figure 2: The actual attributes' values of the dataset in the ARRF file.

6.2 Loading the Dataset File

To load the dataset file into Weka, I used the “Open file” button from the explorer window and selected the ARRF file from my hard drive. Figure 3 shows the Weka Explorer Window once I loaded the dataset file. Weka automatically detects the contents of the file and displays them in an organized and readable way in the Explorer Window. As can be seen, the dataset attributes are structured and displayed on the left side in a very neat way. Other textual and graphical information is displayed on the right side of the Explorer Window. In fact, Weka can even detect if the data within the ARRF is not written in the way Weka can read it. For example, when I removed the “@” symbol from the beginning of line 3 where one of the attributes is declared, and tried to load the file again, I got the error shown in figure 4. The error is even indicative and provides information on what is missing, what is expected and which line should be fixed. Weka also enables users to edit and save the ARRF file directly from the Explorer Window through the “Edit” and “Save” buttons respectively.

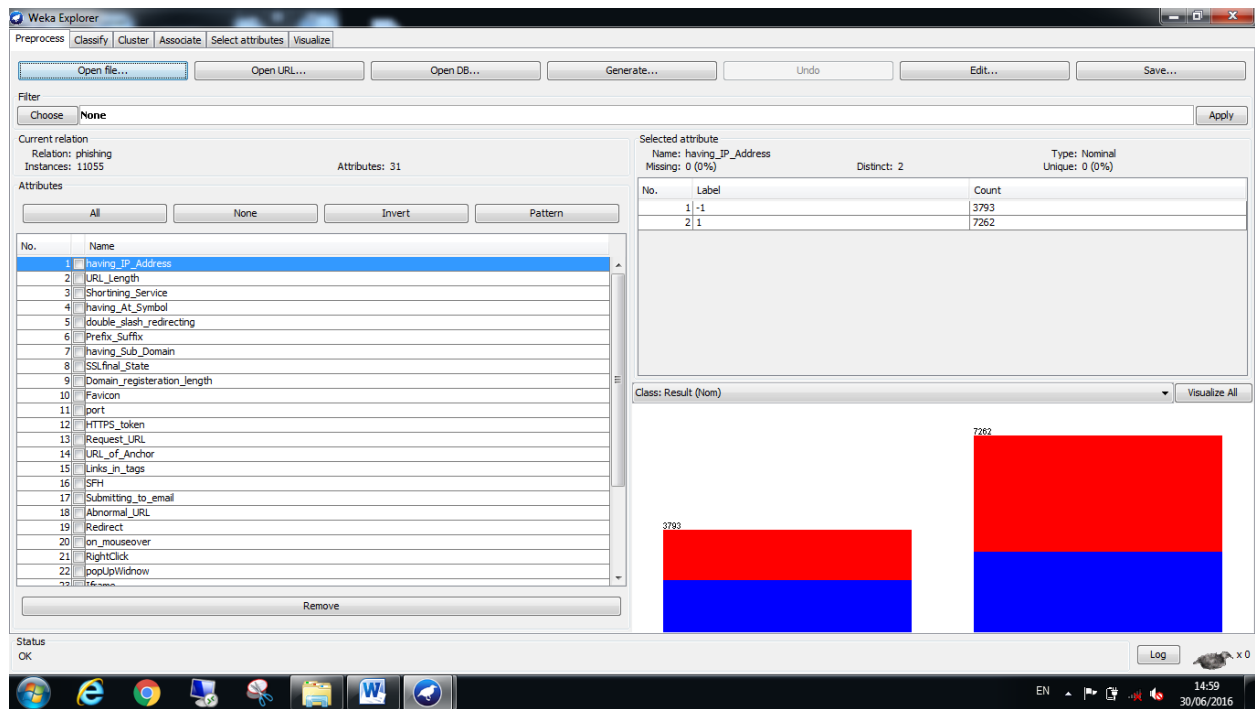


Figure 3: The dataset file once loaded into Weka.

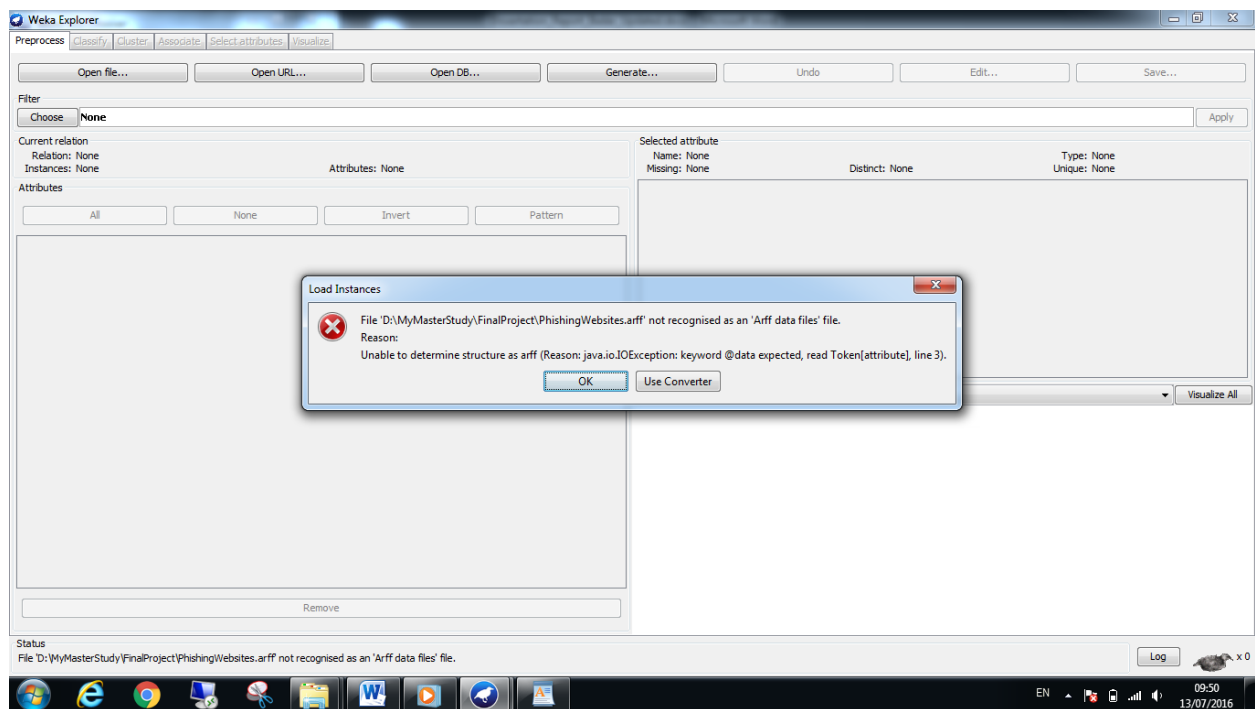


Figure 4: Error displayed by Weka to indicate a missing “@” symbol.

6.3 Reading Dataset Information from the Explorer Window

Different information about the dataset can be read and observed from figure 3. As can be noticed, the dataset has 11055 instances and 30 attributes. The attributes represent the features that can be used to distinguish legitimate websites from phishing websites. The last row in the attributes' list is the Result; which tells whether the website is a phishing website or a legitimate one based on the overall features of the website. In fact, to get some details about a specific attribute, Weka allows users to click on that attribute from the list on the left side and Weka will display the details related to that attribute only. For example, when I clicked on attribute number 22; which is the "Using Pop Up Window" feature to submit personal information, Weka displays some details related to this attribute only. Figure 5 shows the details displayed by Weka for this attribute. We can read from this figure that there are 8918 websites in the dataset that use the pop up window feature. It is also clear that the remaining number of websites which is 2137 do not use this feature. To check another attribute, figure 6 shows the details displayed by Weka once I clicked on attribute number 28; which tells whether the website is indexed by Google or not. As can be noticed from figure 6, there are 9516 out of 11055 websites in this dataset that are indexed by Google. These websites are actually legitimate websites as phishing websites usually do not get indexed by Google because they live for short period of time. The rest of the websites in the dataset are not indexed by Google. These websites constitute 1539 websites as shown in figure 6.

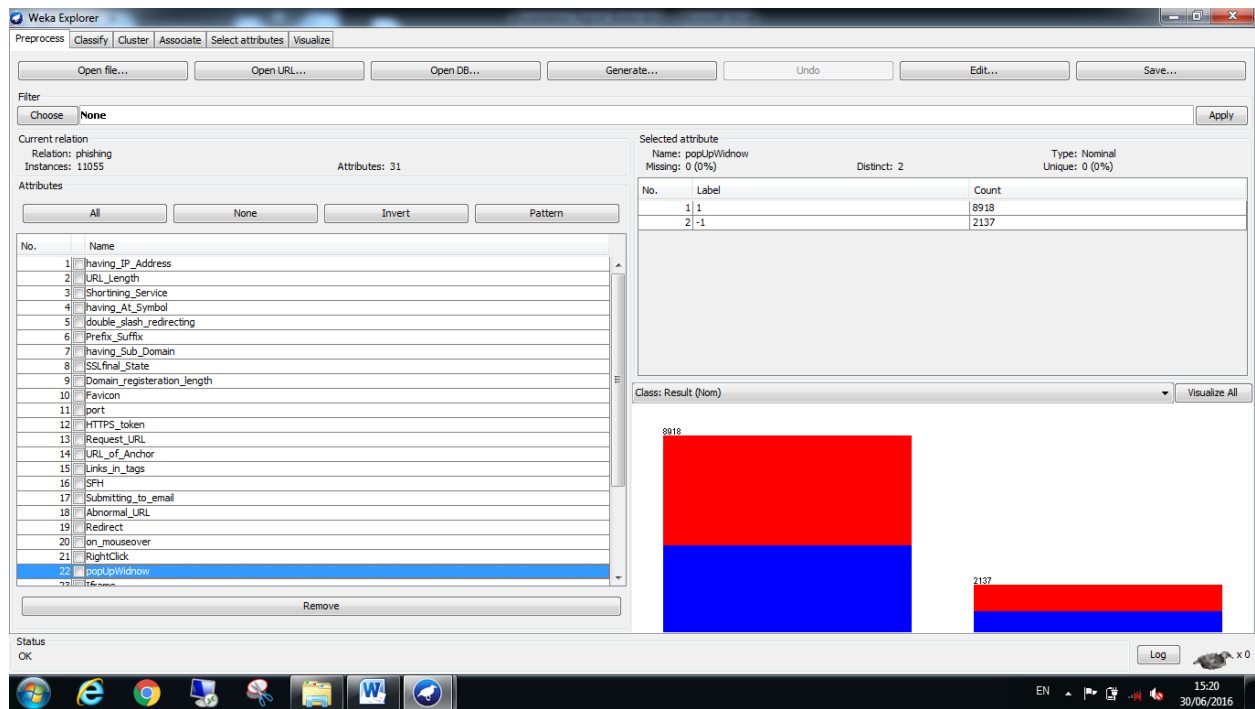


Figure 5: Details displayed for a specific attribute in the list.

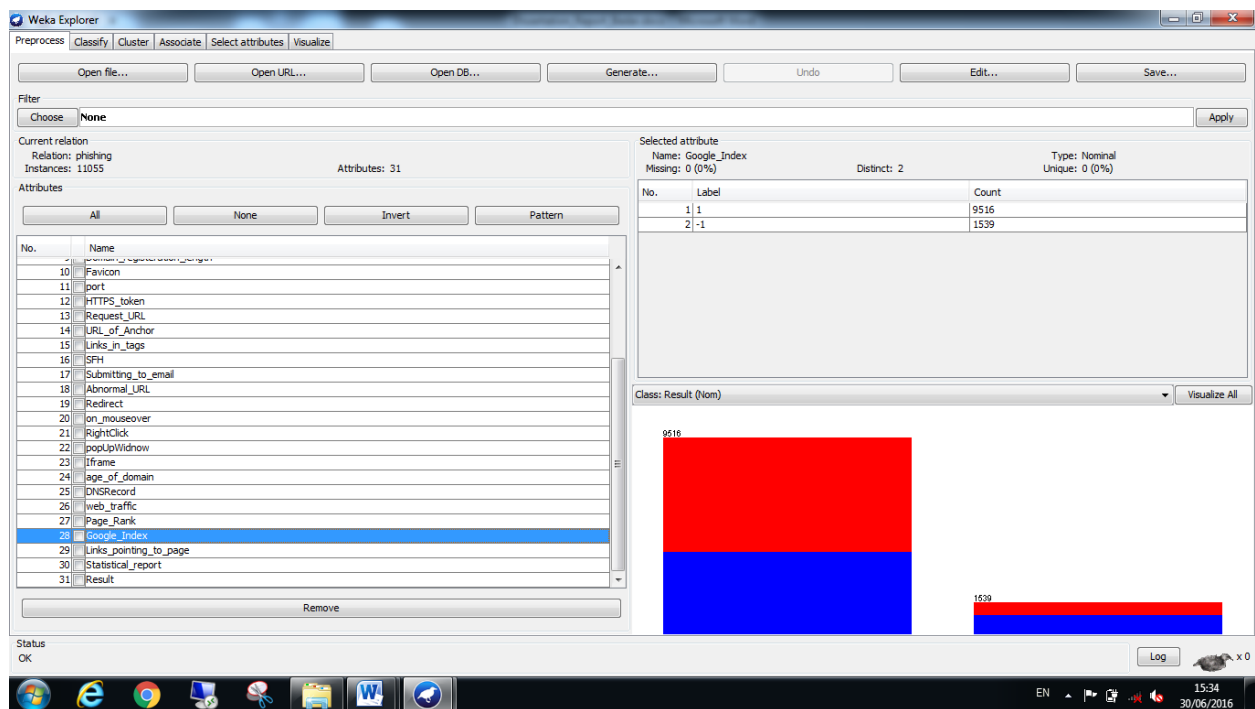


Figure 6: Details displayed for a specific attribute in the list.

7. Running the Algorithms before Dataset Pre-Processing

The pre-processing task is intended to prepare the dataset for running the algorithms and the tests in order to produce an accurate model. This task includes cleaning the dataset, checking for missing data, and selecting the attributes that will contribute effectively when running the tests. In fact, to produce the best model out of the dataset, the dataset should undergo different steps and stages. The first and one of the very early steps is the dataset pre-processing stage.

In this section, I describe the tests and the algorithms run on my dataset without conducting any data cleaning or feature selection. This means that I skipped the pre-processing stage in order to see the effects on the results I receive. The main goal of this task is to run some tests and produce some results that can clearly show the effects of data cleaning and feature selection. This is achieved by performing a comparison on the results I received in this section with the results I obtained later on in the coming sections after I conducted some data cleaning and feature selection.

Since the task required here is to produce the best classification model that can effectively and efficiently distinguish between legitimate and phishing websites based on their features, I need to run different classification algorithms on the dataset. However, we know that Weka includes many different classification algorithms. These algorithms are organised under different categories such as bayes, functions, meta, rules and trees. In fact, it is not practical and not possible to run all of these algorithms in this project. Therefore, I had to choose 13 of the most popular algorithms; algorithms that are more likely to produce good results compared to the other algorithms. I decided to choose 13 algorithms from different categories in order to explore various classification methods. Table 2 presents the 13 different algorithms I decided to run on the dataset. In addition, I used the 10-Folds Cross Validation method as my test option. The cross validation method is a common technique that is used to evaluate classifiers' performance by partitioning the dataset into a number of folds. Most of the partitions are used for training the model, and usually one partition is used for testing. Then, the evaluation process is iterated over the partitions, and the average result is taken

(Moreno-Torres et al., 2012). This technique should be used carefully as it can lead to some kind of classifier bias especially when using big number of iterations for example.

The Selected Algorithms	
Bayes	BayesNet
	HNB
Functions	Logistic
	RBFNetwork
	MultiClassClassifier
	Winnow
Meta	Bagging
	RotationForest
	ClassificationViaClustering
Rules	Ridor
Trees	BFTree
	J48
	REPTree

Table 2: The 13 different algorithms chosen to run on the dataset.

Table 3 provides a summary of the results I obtained for each of the 13 algorithms I tested against the dataset – without dataset pre-processing. Most of the algorithms performed quite well and close to each other. The top two algorithms were the Bagging algorithm and the RotationForest algorithm. These two algorithms showed the best performance among all of the algorithms; with ROC Area values of 0.992 and 0.994 respectively. Some algorithms such as the RBFNetwork and the BFTree came on the second best performing algorithms. They still performed well, but somehow lower than the top two algorithms. On the other hand, there are two algorithms that did not perform quite well. These algorithms are the ClassificationViaClustering algorithm and

the Winnow algorithm. The worst performance was produced by the ClassificatinViaClustering algorithm with ROC Area value of 0.664.

As can be noticed from table 3, the Bagging algorithm and the RotationForest algorithm could correctly classify over 96% of the total instances. They only missed lower than 500 instances out of the 11055 instances. On the other hand, the ClassificatinViaClustering algorithm incorrectly classified 3475 instances; which is about 31.43%. The classification results of the other algorithms are somehow acceptable, but need to be improved.

Even though I received somehow good results for most of the algorithms, this does not mean that those algorithms have already performed up to their best. The algorithms had been tested so far without any dataset cleaning or any feature selection up to this stage. In fact, this dataset has a lot of attributes; 30 attributes in total. There might be some attributes that do not really contribute heavily to the overall performance of the algorithms. Also, there might be some attributes that are crucial and very important to the performance of the algorithms. The next step is to explore these possibilities. I ran the tests again with different combinations of attributes to find out which attributes are really important and which attributes are not. In addition, I explored the different parameters related to each of the algorithms. I tried different values for those parameters and check if changing those parameters' values would produce any better results.

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	ROC Area
BayesNet	10280 92.9896 %	775 7.0104 %	0.981
HNB	10356 93.6771 %	699 6.3229 %	0.986
Logistic	10391 93.9937 %	664 6.0063 %	0.987

RBFNetwork	10071 91.0991 %	984 8.9009 %	0.964
MultiClassClassifier	10391 93.9937 %	664 6.0063 %	0.987
Bagging	10614 96.0109 %	441 3.9891 %	0.992
RotationForest	10700 96.7888 %	355 3.2112 %	0.994
ClassificationViaClustering	7580 68.5663 %	3475 31.4337 %	0.664
Winnow	9510 86.0244 %	1545 13.9756 %	0.86
Ridor	10264 92.8449 %	791 7.1551 %	0.927
BFTree	10579 95.6943 %	476 4.3057 %	0.978
J48	10599 95.8752 %	456 4.1248 %	0.984
REPTree	10539 95.3324 %	516 4.6676 %	0.985

Table 3: Summary of the results of the algorithms (run **without** dataset pre-processing).

Figures 7 to 19 illustrate the output results I obtained for each of the algorithms when I ran them against the dataset without any dataset pre-processing activities. Each figure shows the details and results I received for one specific algorithm. Such details include the number of instances correctly classified, the number of instances incorrectly classified, the Kappa statistic, the mean absolute error, the root mean squared error, the relative absolute error, the root relative squared error, the time taken by the algorithm to build the model and the value for the ROC area. In addition, some figures show more

details such as the tree size and the number of its leaves in case there is a tree produced by the algorithm. For example, figures 20 and 21 show the trees produced by the two algorithms; the J48 and the REPTree respectively. As can be noticed, the sizes of the trees are very big, and I could not even fit them clearly on the screen. This is expected because I ran the algorithms without any dataset cleaning and feature selection. Therefore, most of the attributes are included in the tree; and hence the tree size is getting very big. Another thing to mention is that the output of the Ridor algorithm also included the number of rules created by this algorithm. This is the only algorithm from the Rules category. Figure 16 shows that the total number of rules produced by this algorithm for the created model is 39.

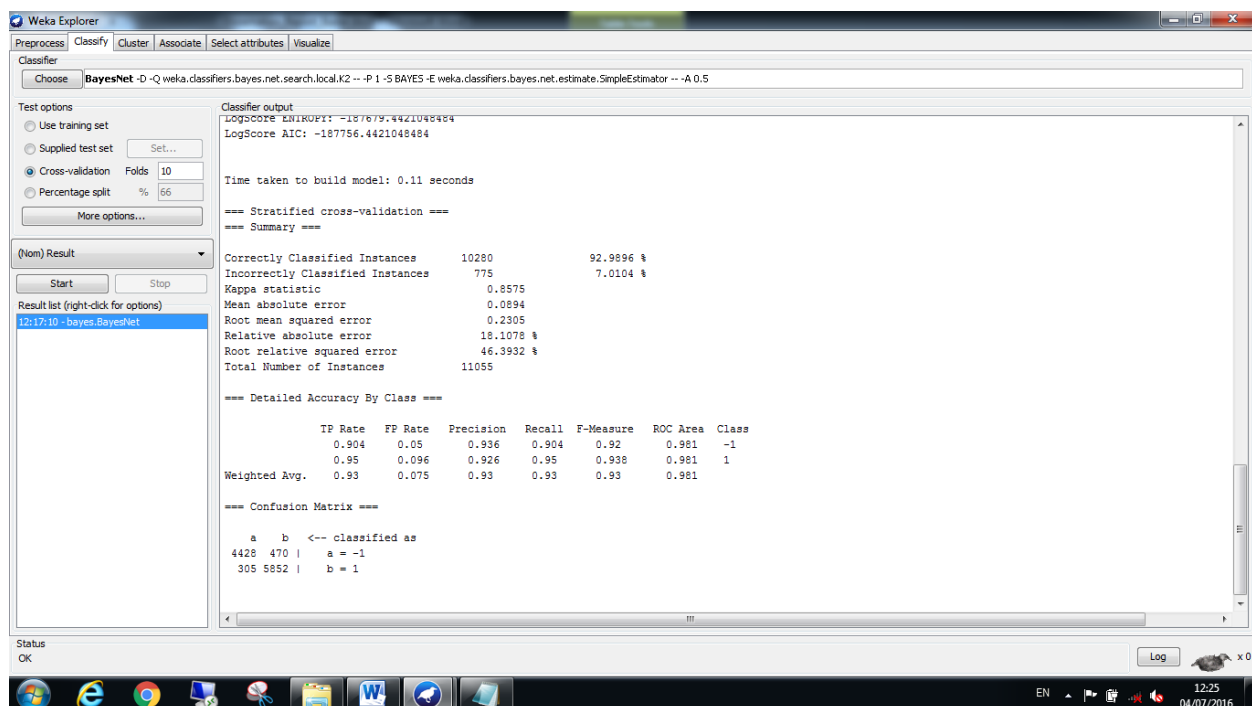


Figure 7: Output of the BayesNet algorithm (without dataset pre-processing).

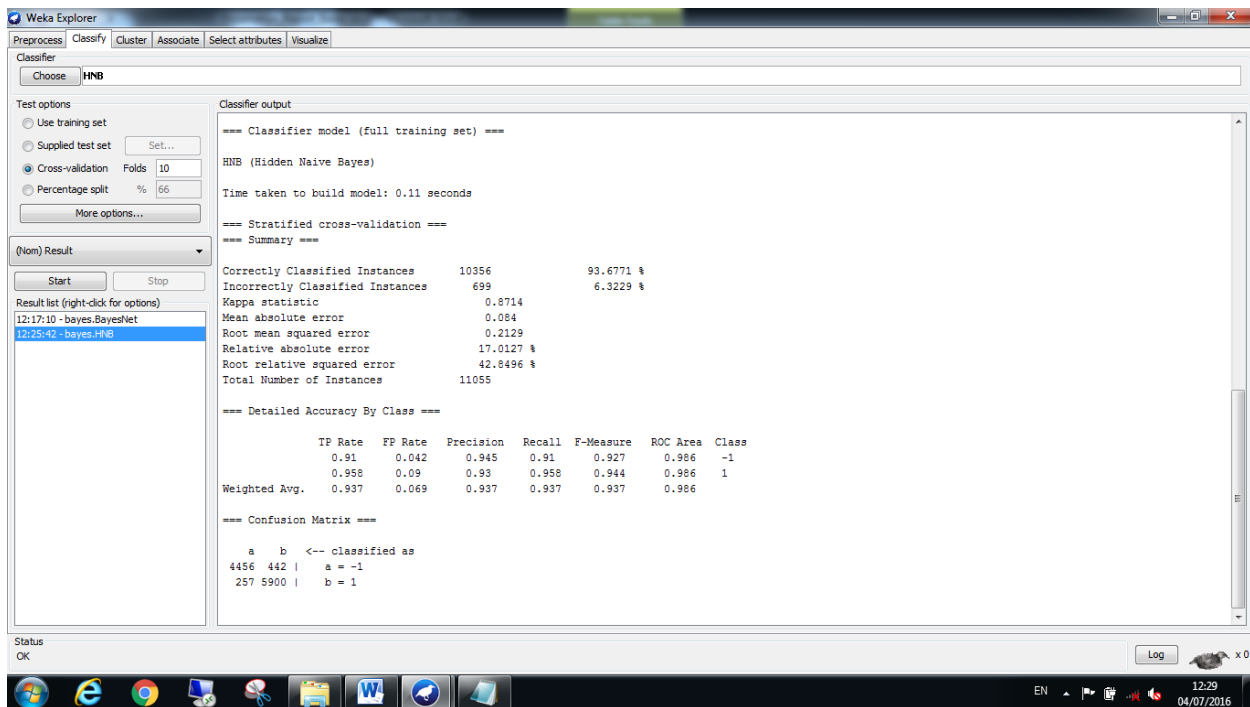


Figure 8: Output of the HNB algorithm (without dataset pre-processing).

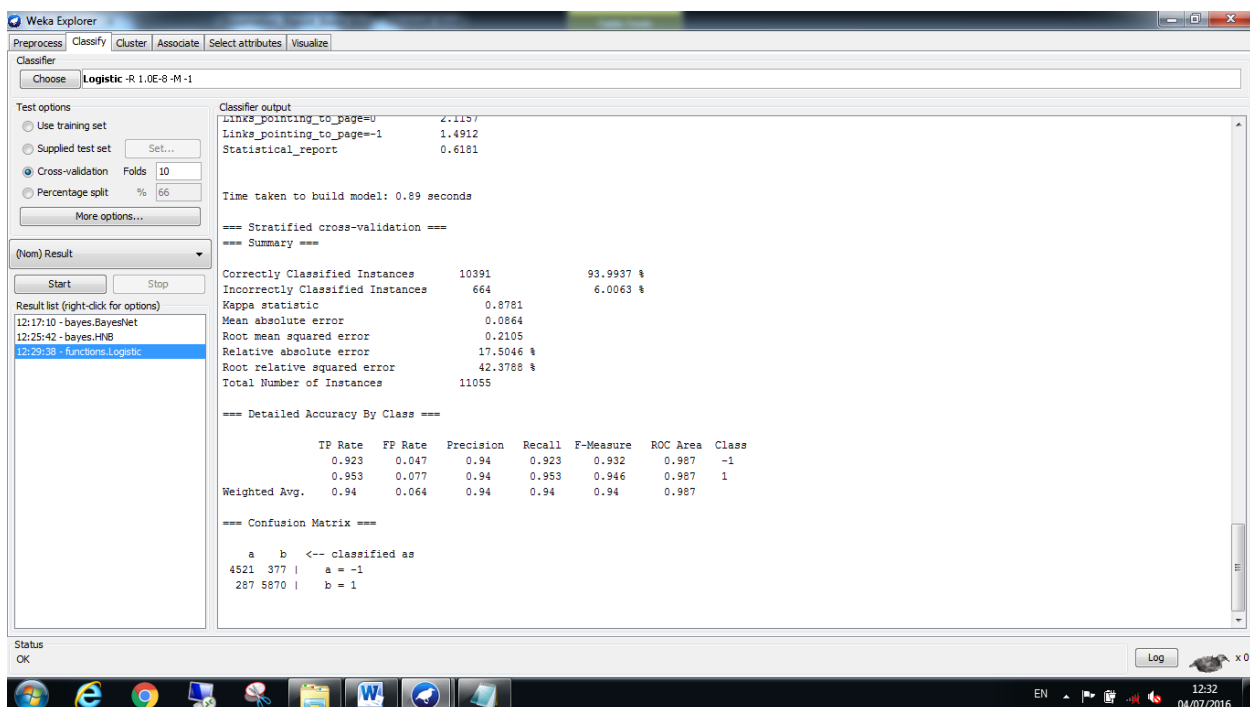


Figure 9: Output of the Logistic algorithm (without dataset pre-processing).

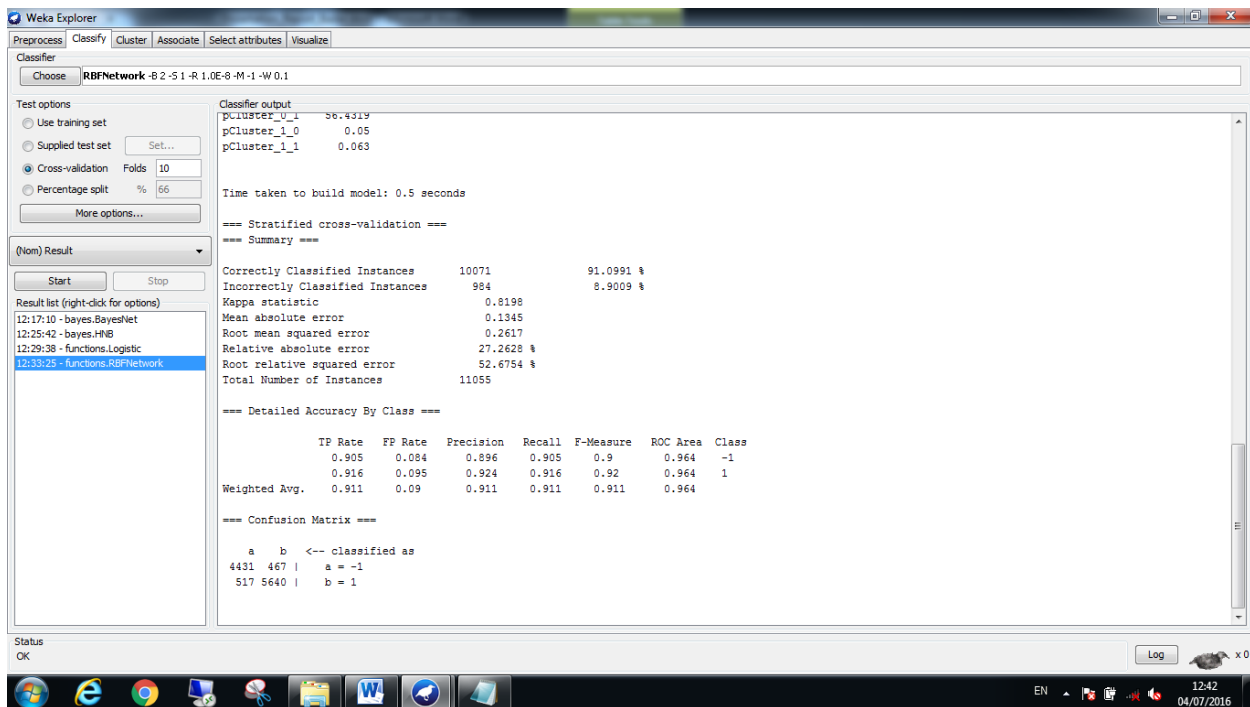


Figure 10: Output of the RBFNetwork algorithm (without dataset pre-processing).

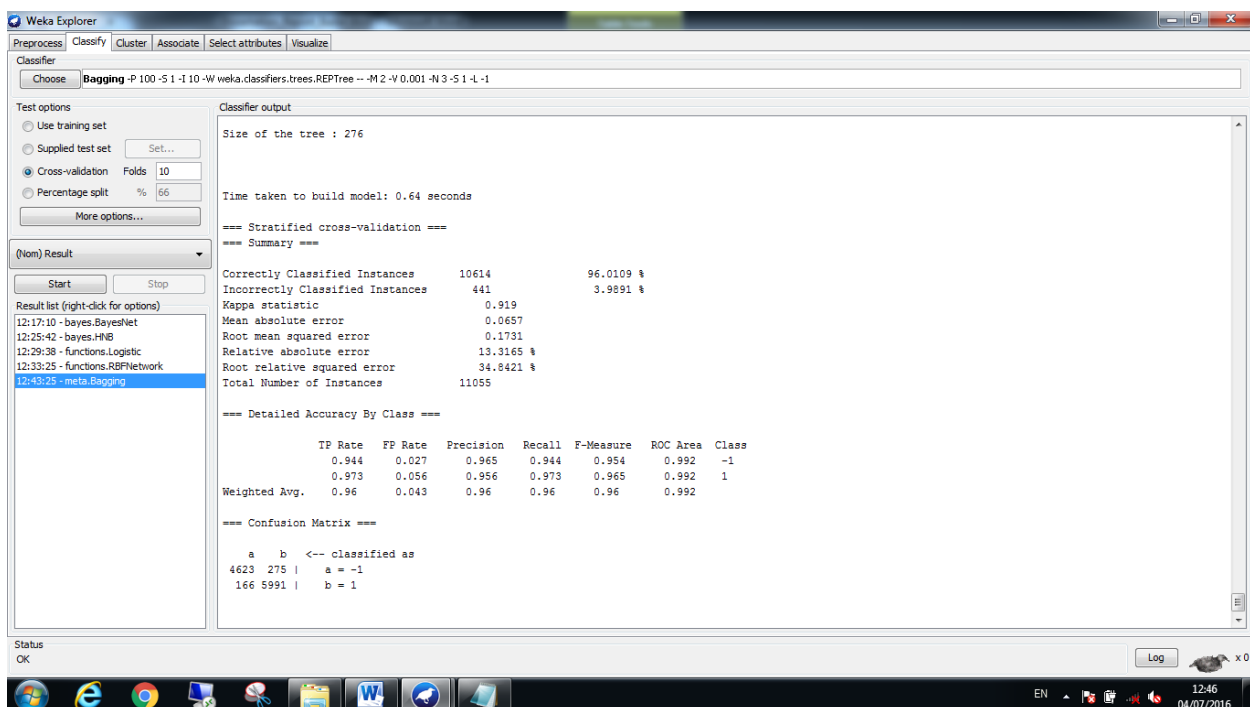


Figure 11: Output of the Bagging algorithm (without dataset pre-processing).

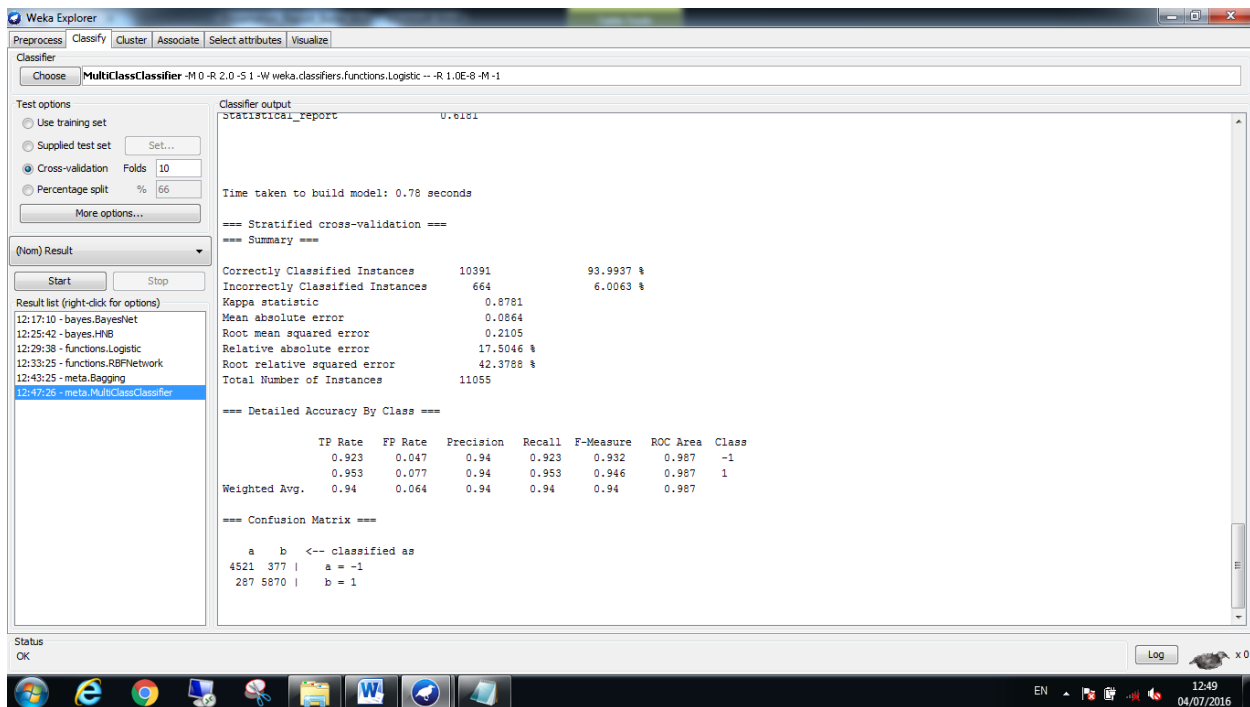


Figure 12: Output of the MultiClassClassifier algorithm (without dataset pre-processing).

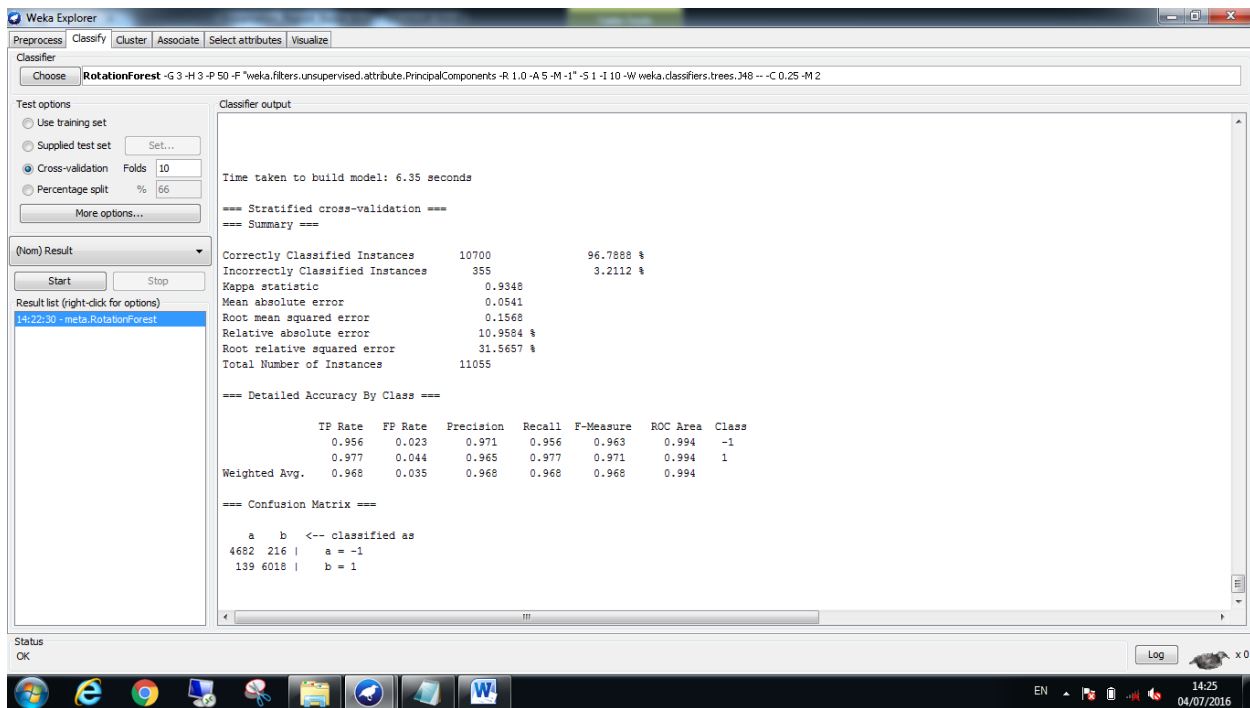


Figure 13: Output of the RotationForest algorithm (without dataset pre-processing).

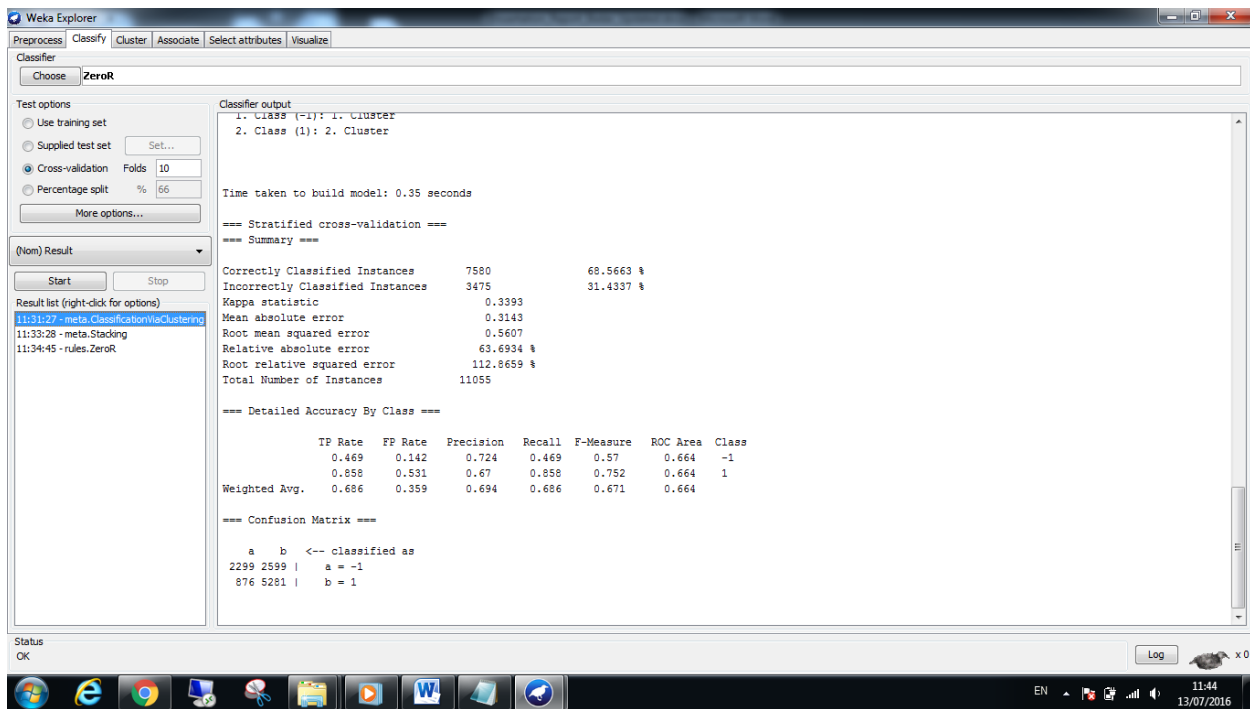


Figure 14: Output of the ClassificationViaClustering algorithm (**without** dataset pre-processing).

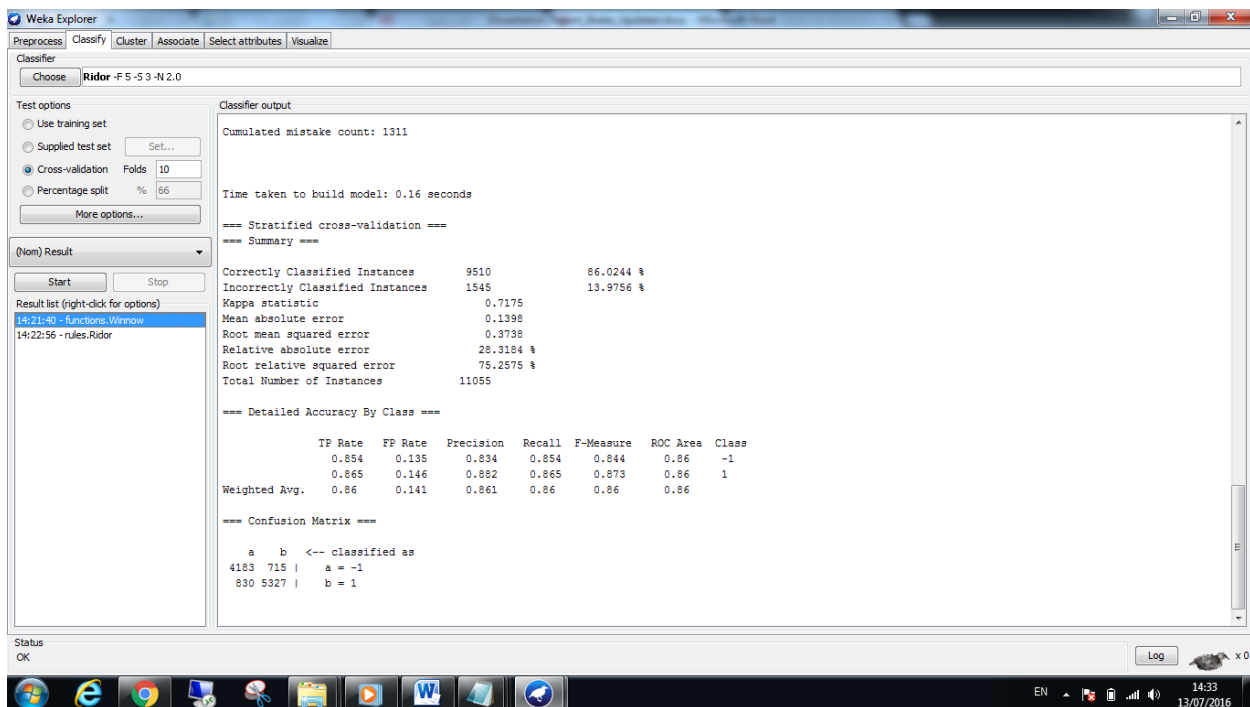


Figure 15: Output of the Winnow algorithm (**without** dataset pre-processing).

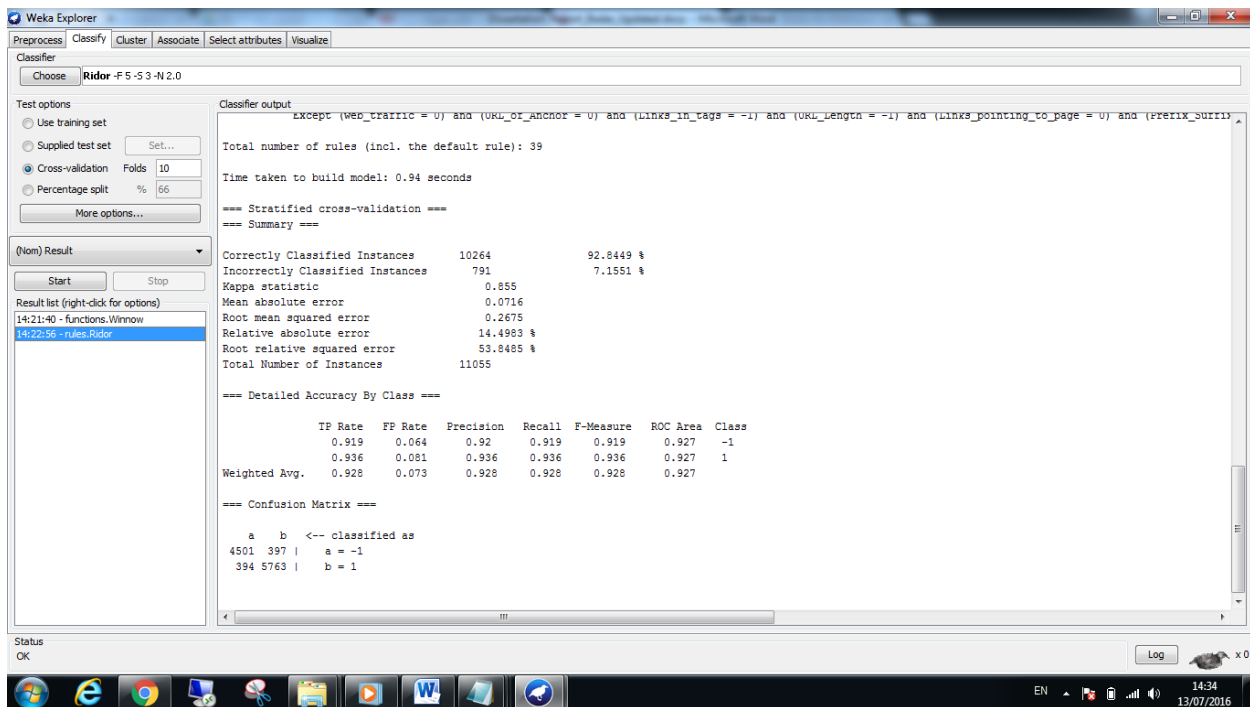


Figure 16: Output of the Ridor algorithm (without dataset pre-processing).

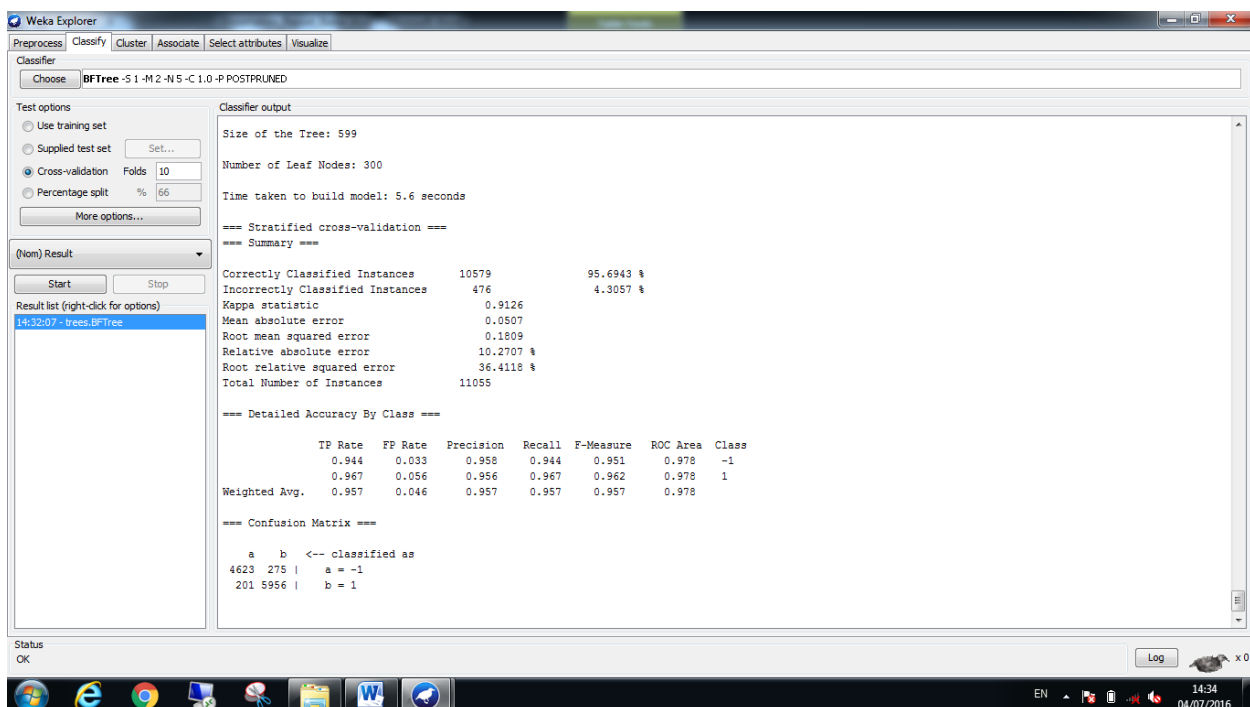


Figure 17: Output of the BFTree algorithm (without dataset pre-processing).

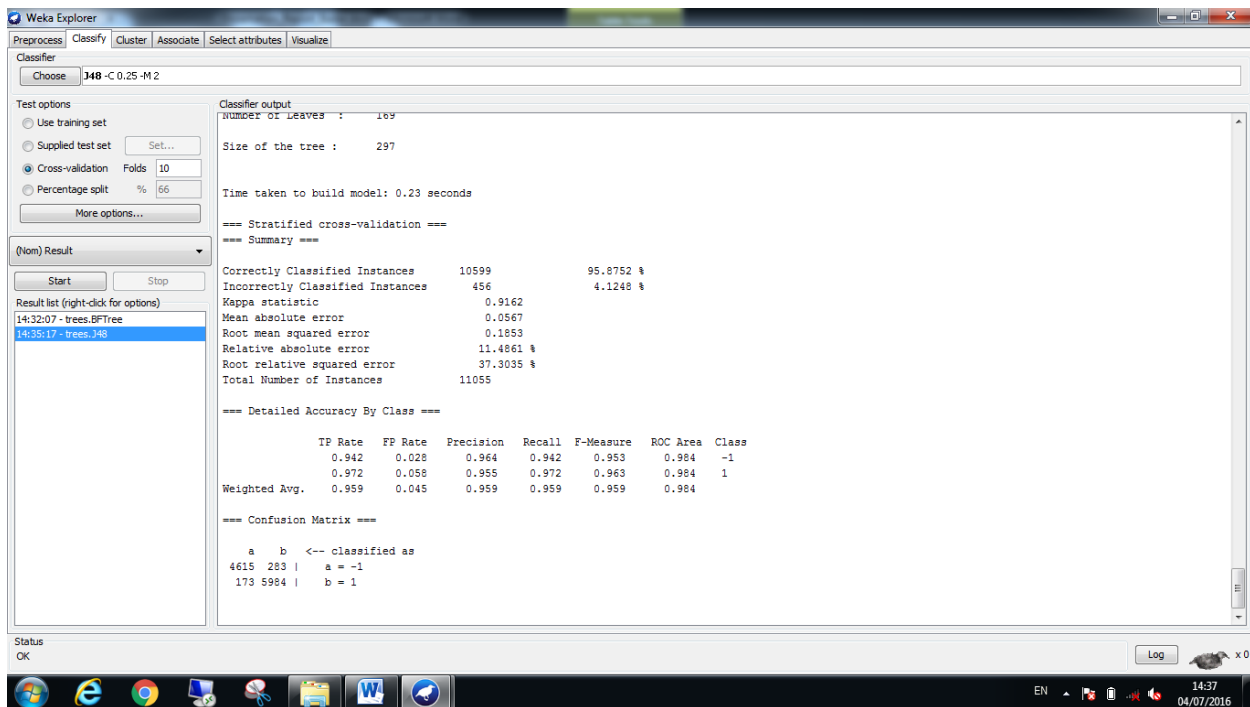


Figure 18: Output of the J48 algorithm (without dataset pre-processing).

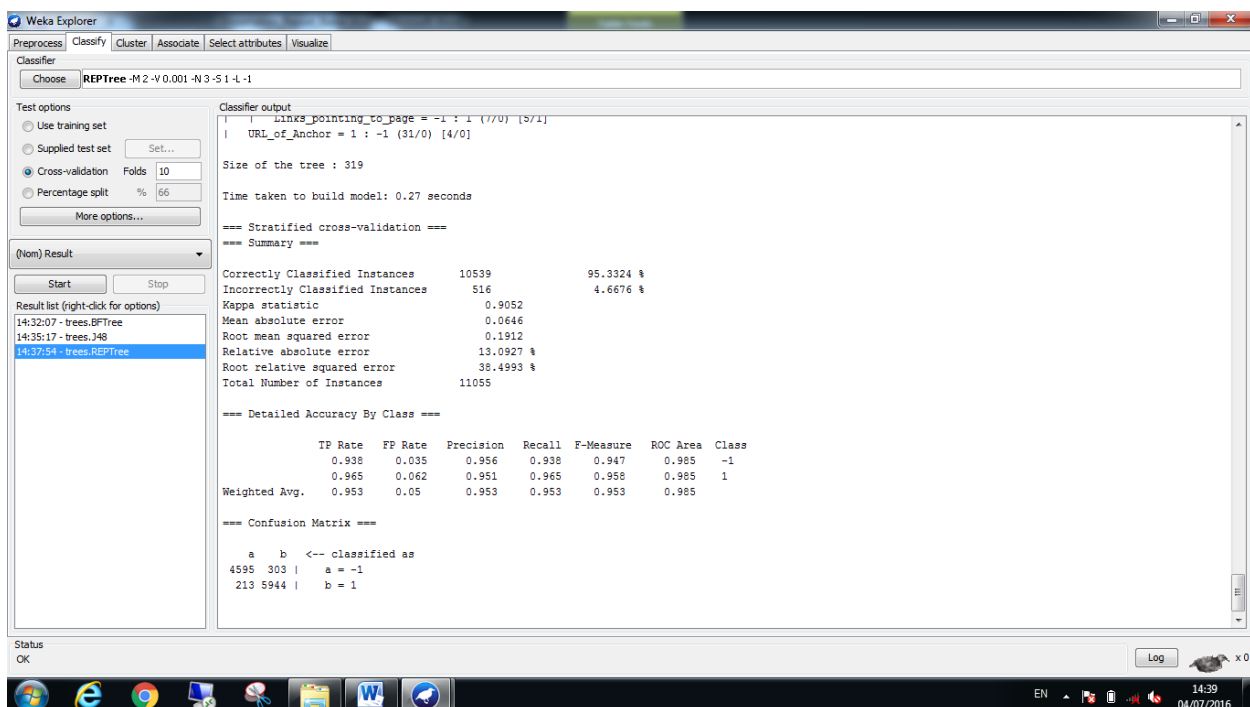
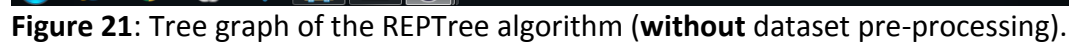
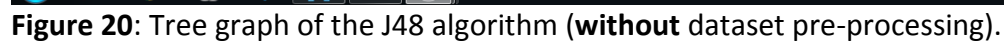


Figure 19: Output of the REPTree algorithm (without dataset pre-processing).



8. Dataset Pre-Processing Stage

8.1 The Importance of Dataset Pre-Processing

As the output results of the previous section have shown, most of the algorithms have performed quite well. Other algorithms performed somehow poorly. In fact, all of the algorithms did not perform up to their best because I ran them without conducting any dataset cleaning or feature selection. The values I obtained for the ROC area for most of the algorithms are not the ideal values. This is because I have not prepared the dataset for analysis yet. Dataset preparation is an important stage that should be conducted before running any analysis algorithms on the dataset. This stage is called dataset pre-processing. In this stage, the dataset is prepared and made ready for the next stages in order to get very effective, efficient and accurate models and results as much as possible. This stage includes checking the dataset for any missing data. It also includes selecting the most effective features or attributes that will help in getting the best output from the analysis algorithms.

8.2 Running Some Attribute Evaluators

Some activities of the dataset pre-processing task can be done manually or automatically. For example, selecting the appropriate attributes for running the algorithms can be done manually or automatically through some tools that are supplied by Weka. Weka has different tools that can help in selecting the best attributes for the analysis algorithms. These tools are called Attribute Evaluators in Weka. For this task, I worked with three different feature selection evaluators; the “CFS Subset Evaluator”, the “Consistency Subset Evaluator” and the “OneR Attribute Evaluator”. I used the default search methods for each of the evaluators. Table 4 illustrates the used evaluators, the used search method for each of the evaluators and the selected attributes produced by each of the evaluators.

Evaluator Name	CFS Subset Evaluator	Consistency Subset Evaluator	OneR Attribute Evaluator
Search Method	Best First	Best First	Ranker
Number of Selected Attributes	9	23	
Selected Attributes	Prefix_Suffix having_Sub_Domain SSLfinal_State Request_URL URL_of_Anchor Links_in_tags SFH web_traffic Google_Index	having_IP_Address URL_Length Shortining_Service having_At_Symbol Prefix_Suffix having_Sub_Domain SSLfinal_State Domain_registration_length HTTPS_token Request_URL URL_of_Anchor Links_in_tags SFH Submitting_to_email Redirect popUpWidnow age_of_domain DNSRecord web_traffic	Ranked attributes: 88.89 , SSLfinal_State 84.73 , URL_of_Anchor 69.78 , web_traffic 66.46 , having_Sub_Domain 63.42 , Request_URL 63.09 , Links_in_tags 62.47 , Domain_registration_length 58.54 , Google_Index 57.55 , Prefix_Suffix 56.85 , Statistical_report 56.37 , age_of_domain 56.22 , having_IP_Address 56.01 , SFH 55.97 , URL_Length 55.69 , Shortining_Service 55.69 , Redirect 55.69 , Page_Rank 55.69 ,

		Page_Rank	double_slash_redirecting
		Google_Index	55.69 , Iframe
		Links_pointing_to_page	55.69 , Favicon
		Statistical_report	55.69 , port
			55.69 , HTTPS_token
			55.69 , Abnormal_URL
			55.69 , Submitting_to_email
			55.69 ,
			Links_pointing_to_page
			55.69 , RightClick
			55.69 , popUpWidnow
			55.43 , having_At_Symbol
			55.36 , on_mouseover
			55.07 , DNSRecord

Table 4: The Output results of the Attributes' Evaluators.

As can be seen from table 4, each evaluator has selected different attributes than the other one. Of course, this depends on the evaluator itself and the search method used by that evaluator. It can also be noticed that the number of the selected attributes by each of the evaluators is different from that of the other one. For example, the first evaluator selected only 9 attributes, while the second evaluator selected 23 attributes. Another point to mention here is that the first two evaluators did not give any ranking for the selected attributes. Only the third evaluator has ranked the attributes based on their importance. The next step is to run the algorithms again using only the attributes selected by each of the evaluators in order to see which attributes produce the best results.

8.3 Running the Algorithms with Attributes Selected by the CFS Subset Evaluator

In this subsection, I ran the tests again but now using only the 9 attributes which are selected by the CFS Subset Evaluator. Table 5 illustrates the results I received for each of the algorithms when I ran them using only the attributes selected by this evaluator. As can be seen from table 5, the performance of most of the algorithms was almost similar to their performance when I used all the dataset attributes. In fact, the performance of the Bagging algorithm and the RotationForest algorithm has decreased a little bit. The values of the ROC area changed from 0.992 to 0.986 for the Bagging algorithm, and from 0.994 to 0.984 for the RotationForest algorithm. The Bagging algorithm performed the best in this round. In addition, the ClassificationViaClustering algorithm and the Winnow algorithm - that performed poorly when using all the attributes – have better performance now. The ROC area values increased from 0.664 to 0.726 for the ClassificationViaClustering algorithm, and from 0.86 to 0.882 for the Winnow algorithm.

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	ROC Area
BayesNet	10241 92.6368 %	814 7.3632 %	0.982
HNB	10352 93.6409 %	703 6.3591 %	0.985
Logistic	10303 93.1976 %	752 6.8024 %	0.984
RBFNetwork	10271 92.9082 %	784 7.0918 %	0.977
MultiClassClassifier	10303 93.1976 %	752 6.8024 %	0.984
Bagging	10424 94.2922 %	631 5.7078 %	0.986
RotationForest	10436	619	0.984

	94.4007 %	5.5993 %	
ClassificationViaClustering	8148 73.7042 %	2907 26.2958 %	0.726
Winnow	9761 88.2949 %	1294 11.7051 %	0.882
Ridor	10274 92.9353 %	781 7.0647 %	0.926
BFTree	10416 94.2198 %	639 5.7802 %	0.984
J48	10426 94.3103 %	629 5.6897 %	0.979
REPTree	10399 94.066 %	656 5.934 %	0.983

Table 5: Summary of the results of the algorithms (using attributes selected by CFS Subset Evaluator).

8.4 Running the Algorithms with Attributes Selected by the Consistency Subset Evaluator

In this subsection, I ran the tests again but now using only the 23 attributes which are selected by the Consistency Subset Evaluator. Table 6 illustrates the results I received for each of the algorithms when I ran them using only the attributes selected by this evaluator. As can be seen from table 6, the performance of all the algorithms is very similar to that performance when I ran them using all the attributes. The values of the ROC area for all the algorithms are almost the same. Also, the best performing algorithm is the RotationForest with ROC area value of 0.994. These results are very important and can indicate something. Because I received almost the same results when using all the 30 attributes as well as when using the 23 attributes selected by this evaluator, this means that I can exclude the 7 attributes which were not selected by this evaluator. These are attributes number 5, 10, 11, 18, 20, 21 and 23. This finding can indicate that those 7 attributes are of less importance to the performance of the algorithms. In fact, none of these 7 attributes was selected by the first evaluator as well.

This means that both evaluators have excluded those 7 attributes. Therefore, I can be assured that those 7 attributes are of less importance and can be excluded. They do not contribute much to the overall performance of the algorithms.

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	ROC Area
BayesNet	10272 92.9172 %	783 7.0828 %	0.981
HNB	10373 93.8308 %	682 6.1692 %	0.987
Logistic	10383 93.9213 %	672 6.0787 %	0.987
RBFNetwork	10272 92.9172 %	783 7.0828 %	0.976
MultiClassClassifier	10383 93.9213 %	672 6.0787 %	0.987
Bagging	10611 95.9837 %	444 4.0163 %	0.992
RotationForest	10685 96.6531 %	370 3.3469 %	0.994
ClassificationViaClustering	7856 71.0629 %	3199 28.9371 %	0.698
Winnow	9552 86.4043 %	1503 13.5957 %	0.866
Ridor	10250 92.7182 %	805 7.2818 %	0.924
BFTree	10575 95.6581 %	480 4.3419 %	0.978
J48	10587 95.7666 %	468 4.2334 %	0.984

REPTree	10530	525	0.985
	95.251 %	4.749 %	

Table 6: Summary of the results of the algorithms (using attributes selected by Consistency Subset Evaluator).

8.5 Running the Algorithms with Attributes Ranked by the OneR Attribute

Evaluator

Unlike the first two attribute evaluators, the OneR Attribute Evaluator did not select a specific set of attributes. Instead, it has evaluated all the attributes and assigned a ranking score for each of them as shown in table 4. In fact, based on the previous two evaluators, I noticed that the results I received when the number of attributes is high are better than the results when the number of attributes is low. This means that the results I received when I used 23 attributes are better than the results I received when I used only 9 attributes. Of course, the different combinations of attributes should affect the results. Therefore, for this subsection, I decided to further explore the attributes and find out what combination of attributes would provide the best results. Since the OneR evaluator had ranked all the attributes, I used this ranking to select the different combinations of attributes. I ran the algorithms again using the top 5 attributes, then the top 10 attributes, then the top 15 attributes and finally with the top 20 attributes.

Tables 7 to 10 present the results I received for the algorithms when I ran them using the top 5, 10, 15 and 20 attributes respectively. As can be noticed from the results, the performance of the algorithms is getting better as the used number of attributes gets higher. However, the algorithms did not show a much better performance when jumping from 15 to 20 attributes. This means that the performance got improved by adding more attributes, but only to some extent. Therefore, using the top ranked 15 attributes – which is half the total number of the attributes – is an ideal combination and gives very good results. The next step is to explore the different parameters for each of the algorithms using only those 15 attributes.

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	ROC Area
BayesNet	10027 90.701 %	1028 9.299 %	0.972
HNB	10192 92.1936 %	863 7.8064 %	0.977
Logistic	10136 91.687 %	919 8.313 %	0.974
RBFNetwork	10081 91.1895 %	974 8.8105 %	0.969
MultiClassClassifier	10136 91.687 %	919 8.313 %	0.974
Bagging	10249 92.7092 %	806 7.2908 %	0.974
RotationForest	10229 92.5283 %	826 7.4717 %	0.974
ClassificationViaClustering	8572 77.5396 %	2483 22.4604 %	0.763
Winnow	9553 86.4134 %	1502 13.5866 %	0.864
Ridor	10090 91.2709 %	965 8.7291 %	0.91
BFTree	10241 92.6368 %	814 7.3632 %	0.978
J48	10235 92.5825 %	820 7.4175 %	0.972
REPTree	10204 92.3021 %	851 7.6979 %	0.973

Table 7: Summary of the results of the algorithms (using the top 5 attributes ranked by the OneR Attribute Evaluator).

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	ROC Area
BayesNet	10233 92.5645 %	822 7.4355 %	0.979
HNB	10314 93.2972 %	741 6.7028 %	0.983
Logistic	10269 92.8901 %	786 7.1099 %	0.983
RBFNetwork	10229 92.5283 %	826 7.4717 %	0.977
MultiClassClassifier	10269 92.8901 %	786 7.1099 %	0.983
Bagging	10405 94.1203 %	650 5.8797 %	0.985
RotationForest	10415 94.2108 %	640 5.7892 %	0.981
ClassificationViaClustering	8116 73.4147 %	2939 26.5853 %	0.724
Winnow	9586 86.7119 %	1469 13.2881 %	0.865
Ridor	10202 92.284 %	853 7.716 %	0.92
BFTree	10384 93.9303 %	671 6.0697 %	0.981
J48	10410 94.1655 %	645 5.8345 %	0.978
REPTree	10352 93.6409 %	703 6.3591 %	0.98

Table 8: Summary of the results of the algorithms (using the top 10 attributes ranked by the OneR Attribute Evaluator).

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	ROC Area
BayesNet	10256 92.7725 %	799 7.2275 %	0.98
HNB	10332 93.46 %	723 6.54 %	0.985
Logistic	10330 93.4419 %	725 6.5581 %	0.985
RBFNetwork	10249 92.7092 %	806 7.2908 %	0.977
MultiClassClassifier	10330 93.4419 %	725 6.5581 %	0.985
Bagging	10556 95.4862 %	499 4.5138 %	0.99
RotationForest	10606 95.9385 %	449 4.0615 %	0.991
ClassificationViaClustering	8179 73.9846 %	2876 26.0154 %	0.731
Winnow	9782 88.4848 %	1273 11.5152 %	0.883
Ridor	10212 92.3745 %	843 7.6255 %	0.922
BFTree	10534 95.2872 %	521 4.7128 %	0.98
J48	10531 95.2601 %	524 4.7399 %	0.984
REPTree	10495 94.9344 %	560 5.0656 %	0.984

Table 9: Summary of the results of the algorithms (using the top 15 attributes ranked by the OneR Attribute Evaluator).

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	ROC Area
BayesNet	10251 92.7273 %	804 7.2727 %	0.98
HNB	10339 93.5233 %	716 6.4767 %	0.986
Logistic	10341 93.5414 %	714 6.4586 %	0.986
RBFNetwork	10243 92.6549 %	812 7.3451 %	0.976
MultiClassClassifier	10341 93.5414 %	714 6.4586 %	0.986
Bagging	10575 95.6581 %	480 4.3419 %	0.991
RotationForest	10624 96.1013 %	431 3.8987 %	0.991
ClassificationViaClustering	8566 77.4853 %	2489 22.5147 %	0.763
Winnow	9577 86.6305 %	1478 13.3695 %	0.864
Ridor	10215 92.4016 %	840 7.5984 %	0.918
BFTree	10545 95.3867 %	510 4.6133 %	0.979
J48	10567 95.5857 %	488 4.4143 %	0.985
REPTree	10509 95.0611 %	546 4.9389 %	0.983

Table 10: Summary of the results of the algorithms (using the top 20 attributes ranked by the OneR Attribute Evaluator).

8.6 Running the Algorithms with Attributes Belonging to the Different Categories of the Features

In this subsection, I ran the algorithms using the attributes from each of the four categories of the websites' features. As described in section 4.5, the dataset attributes – websites' features – had been classified into four main groups; features based on the address bar, features based on abnormality, features based on the HTML and JavaScript techniques and features based on the domain of the website itself. In each testing round, I used the attributes that belong only to one of these four categories. Tables 11 to 14 present the results I received for the algorithms when I used the attributes belonging to each of these categories.

As can be noticed from tables 11 to 14, the results of the algorithms are very interesting. While the algorithms performed quite well when using the attributes of group 1 and group 2, they performed poorly when using the attributes of group 3 and group 4. Using the attributes of the first group (the address bar attributes), the highest ROC value was 0.957 for the bagging algorithm. The BFTree algorithm scored the best ROC value of 0.944 when using the attributes of the second group (the abnormality attributes). The attributes of the third and fourth groups (the HTML and JavaScript group and the domain group) did not contribute much to the performance of the algorithms. For the third group, the highest ROC value was 0.54 for the RotationForest algorithm. The bagging and the BFTree algorithms scored the highest ROC value of 0.8 when using the attributes of the fourth group. These results can indicate that some of the attributes in the first two groups are of more importance to the performance of the algorithms than those attributes in the other two groups. I explored these attributes in details in the coming sections.

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	ROC Area
BayesNet	9927 89.7965 %	1128 10.2035 %	0.945
HNB	9937 89.8869 %	1118 10.1131 %	0.949
Logistic	9923 89.7603 %	1132 10.2397 %	0.947
RBFNetwork	9830 88.919 %	1225 11.081 %	0.938
MultiClassClassifier	9923 89.7603 %	1132 10.2397 %	0.947
Bagging	9995 90.4116 %	1060 9.5884 %	0.957
RotationForest	10003 90.4839 %	1052 9.5161 %	0.951
ClassificationViaClustering	7198 65.1108 %	3857 34.8892 %	0.634
Winnow	9419 85.2013 %	1636 14.7987 %	0.851
Ridor	9915 89.6879 %	1140 10.3121 %	0.894
BFTree	9996 90.4206 %	1059 9.5794 %	0.958
J48	9984 90.3121 %	1071 9.6879 %	0.945
REPTree	9962 90.1131 %	1093 9.8869 %	0.949

Table 11: Summary of the results of the algorithms (using attributes number 1 to 12 that belong to the Address Bar Category).

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	ROC Area
BayesNet	9620 87.0194 %	1435 12.9806 %	0.934
HNB	9621 87.0285 %	1434 12.9715 %	0.937
Logistic	9638 87.1823 %	1417 12.8177 %	0.938
RBFNetwork	9541 86.3048 %	1514 13.6952 %	0.932
MultiClassClassifier	9638 87.1823 %	1417 12.8177 %	0.938
Bagging	9666 87.4355 %	1389 12.5645 %	0.931
RotationForest	9648 87.2727 %	1407 12.7273 %	0.934
ClassificationViaClustering	6327 57.232 %	4728 42.768 %	0.569
Winnow	8903 80.5337 %	2152 19.4663 %	0.805
Ridor	9611 86.938 %	1444 13.062 %	0.86
BFTree	9639 87.1913 %	1416 12.8087 %	0.944
J48	9654 87.327 %	1401 12.673 %	0.919
REPTree	9668 87.4536 %	1387 12.5464 %	0.922

Table 12: Summary of the results of the algorithms (using attributes number 13 to 18 that belong to the Abnormality Category).

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	ROC Area
BayesNet	6198 56.0651 %	4857 43.9349 %	0.505
HNB	6277 56.7797 %	4778 43.2203 %	0.532
Logistic	6206 56.1375 %	4849 43.8625 %	0.51
RBFNetwork	6183 55.9294 %	4872 44.0706 %	0.517
MultiClassClassifier	6206 56.1375 %	4849 43.8625 %	0.51
Bagging	6321 57.1777 %	4734 42.8223 %	0.537
RotationForest	6327 57.232 %	4728 42.768 %	0.54
ClassificationViaClustering	6018 54.4369 %	5037 45.5631 %	0.502
Winnow	5008 45.3008 %	6047 54.6992 %	0.494
Ridor	6298 56.9697 %	4757 43.0303 %	0.516
BFTree	6324 57.2049 %	4731 42.7951 %	0.539
J48	6323 57.1958 %	4732 42.8042 %	0.526
REPTree	6318 57.1506 %	4737 42.8494 %	0.527

Table 13: Summary of the results of the algorithms (using attributes number 19 to 23 that belong to the HTML and JavaScript Category).

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	ROC Area
BayesNet	7836 70.882 %	3219 29.118 %	0.752
HNB	8002 72.3835 %	3053 27.6165 %	0.777
Logistic	7918 71.6237 %	3137 28.3763 %	0.755
RBFNetwork	7725 69.8779 %	3330 30.1221 %	0.747
MultiClassClassifier	7918 71.6237 %	3137 28.3763 %	0.755
Bagging	8245 74.5816 %	2810 25.4184 %	0.8
RotationForest	8213 74.2922 %	2842 25.7078 %	0.792
ClassificationViaClustering	6196 56.047 %	4859 43.953 %	0.545
Winnow	6843 61.8996 %	4212 38.1004 %	0.614
Ridor	8087 73.1524 %	2968 26.8476 %	0.729
BFTree	8235 74.4912 %	2820 25.5088 %	0.8
J48	8200 74.1746 %	2855 25.8254 %	0.772
REPTree	8203 74.2017 %	2852 25.7983 %	0.789

Table 14: Summary of the results of the algorithms (using attributes number 24 to 30 that belong to the Domain Category).

8.7 Running the Algorithms with Different Combinations of the Top 5 Attributes

In this subsection, I tried to explore the importance of the attributes to the performance of the algorithms. I picked the top 5 attributes ranked by the OneR attribute evaluator. Then, I ran the tests using different combinations of those 5 attributes. I noticed that when running the algorithms using only the top one or two attributes which are the SSLfinal_State and URL_of_Anchor, I received the same results for all the algorithms. Table 15 presents the results I received for all the algorithms. I also received similar phenomenon – same results for all the algorithms – even when I used some different combinations of 3 of those attributes. Even when I used four attributes, the results of the algorithms are very close to each other. The results started to clearly change and varied for each algorithm when I used more than 4 attributes. This indicates that at least 5 attributes are required to clearly distinguish between legitimate and phishing websites.

The top 5 attributes ranked by the OneR attribute evaluator are all needed to produce good classification results. They are all important and can be used as the discriminative attributes or features to distinguish between legitimate and phishing websites. These attributes are the SSLfinal_State, URL_of_Anchor, web_traffic, having_Sub_Domain and Request_URL. In fact, two of these attributes belong to the first category of attributes, and another two attributes belong to the second category. One attribute belongs to the fourth category. These findings explain why I received good results when I used the attributes of the first and second groups in section 8.6. They also prove that these 5 attributes are all important and contribute heavily to the performance of the algorithms.

Used Attributes	Correctly Classified Instances	Incorrectly Classified Instances	ROC Area
SSLfinal_State	10251 92.7273 %	804 7.2727 %	0.98
SSLfinal_State and URL_of_Anchor	10089 91.2619 %	966 8.7381 %	0.951

Table 15: The similar results of all algorithms (when using only top one or two attributes).

8.8 Running the Algorithms with the Top Ranked 5 and 15 Attributes Using Different Parameters' Values

In this subsection, I explored the effects of changing the values of some parameters for each of the algorithms on the overall results and performance of the algorithms. I ran the algorithms again using the top ranked 5 and 15 attributes. I used 5 and 15 attributes to check whether there is any big difference in the results. This is to explore the appropriate minimum number of the most discriminative attributes. Table 16 presents the top ranked 5 and 15 attributes. Figure 22 illustrates the attributes' names, their numbers and their ranking score.

The Top 5 Ranked Attributes		
SSLfinal_State	URL_of_Anchor	web_traffic
having_Sub_Domain	Request_URL	
The Top 15 Ranked Attributes		
SSLfinal_State	URL_of_Anchor	web_traffic
having_Sub_Domain	Request_URL	Links_in_tags
Domain_registration_length	Google_Index	Prefix_Suffix
Statistical_report	age_of_domain	having_IP_Address
SFH	URL_Length	Shortining_Service

Table 16: The top 5 and 15 ranked attributes used for the final testing.

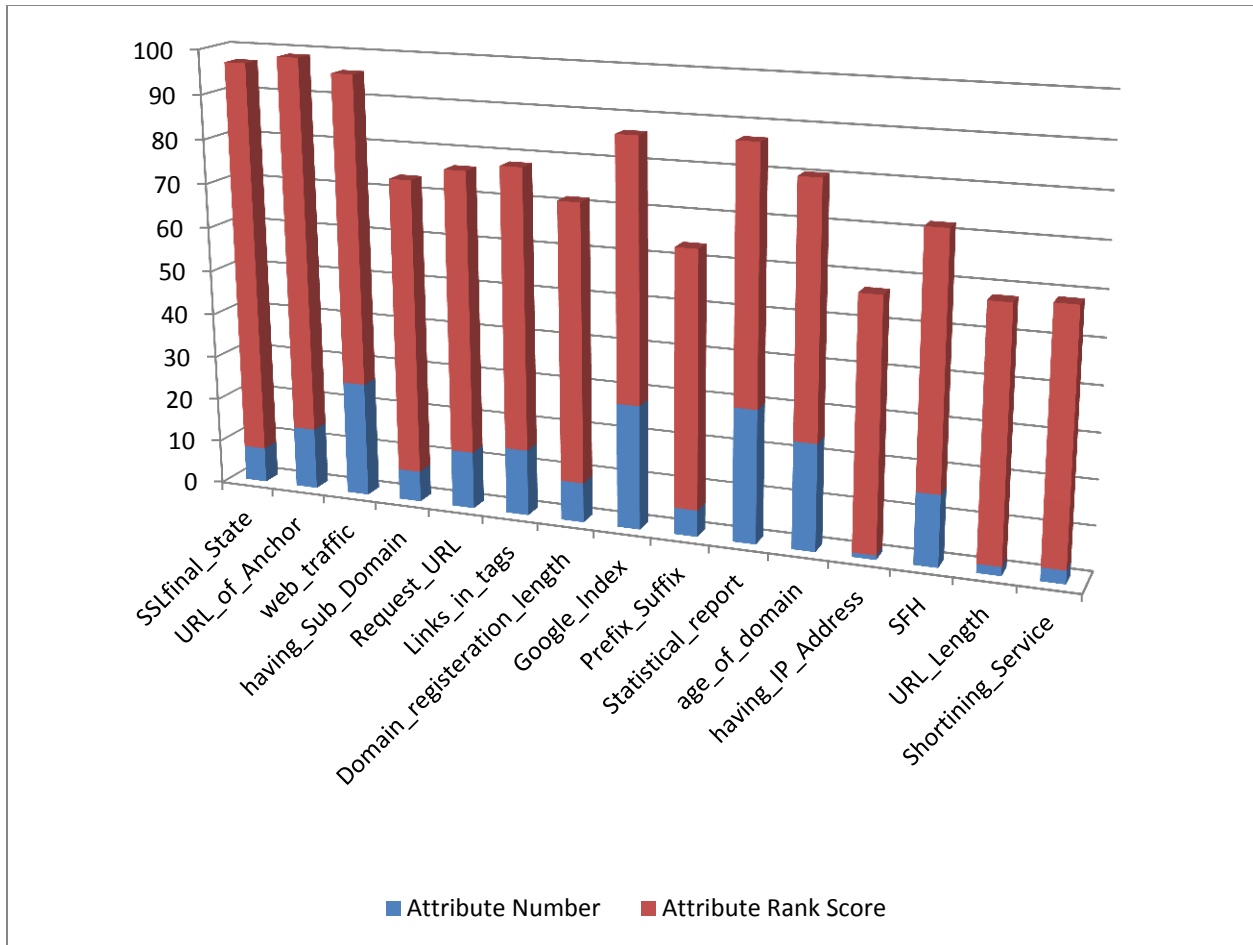


Figure 22: The top 5 and 15 ranked attributes, their number and their ranking score.

This subsection is intended to explore the different parameters of the algorithms, and to find out if changing their values would produce any better results. I explored these parameters using the top ranked 5 and 15 attributes to further check the effect on the results. However, I worked only with 6 of the best performing algorithms. I chose the algorithms from different groups to explore different classification methods again. Table 17 presents the 6 selected algorithms and the parameters I selected to explore for each of the algorithms. It also illustrates what different values I used for each of the selected parameters. Tables 18 and 19 present the best results I received for each of the algorithms when I used 5 and 15 attributes respectively.

Algorithm	Explored Parameters	Parameter Values Tested
RBFNetwork	numClusters	2, 10, 20, 50, 70 and 100
MultiClassClassifier	randomWidthFactor	2, 5 and 10
Bagging	numIterations	10, 15, 20, 30 and 50
RotationForest	numIterations	10, 15, 20, 30 and 50
	confidenceFactor	0.25 , 0.3, 0.4 and 0.5
J48	confidenceFactor	0.25 , 0.3, 0.4 and 0.5
REPTree	numFolds	3, 5, 10 and 20

Table 17: The different parameters tested for each algorithm.

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	ROC Area	Best Performance with Parameter / Value
RBFNetwork	10250 92.7182 %	805 7.2818 %	0.978	numClusters / 50
MultiClassClassifier	10136 91.687 %	919 8.313 %	0.974	randomWidthFactor / All
Bagging	10251 92.7273 %	804 7.2727 %	0.975	numIterations / 20
RotationForest	10244 92.664 %	811 7.336 %	0.977	numIterations / 15 confidenceFactor / 0.5
J48	10236 92.5916 %	819 7.4084 %	0.974	confidenceFactor / 0.5
REPTree	10221 92.4559 %	834 7.5441 %	0.973	numFolds / 5

Table 18: Summary of the best results of the 6 algorithms along with the parameters' values (using top 5 attributes).

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	ROC Area	Best Performance with Parameter / Value
RBFNetwork	10511 95.0791 %	544 4.9209 %	0.989	numClusters / 100
MultiClassClassifier	10330 93.4419 %	725 6.5581 %	0.985	randomWidthFactor / All
Bagging	10566 95.5767 %	489 4.4233 %	0.991	numIterations / 30
RotationForest	10633 96.1827 %	422 3.8173 %	0.993	numIterations / 50 confidenceFactor / 0.5
J48	10564 95.5586 %	491 4.4414 %	0.984	confidenceFactor / 0.5
REPTree	10495 94.9344 %	560 5.0656 %	0.984	numFolds / 3

Table 19: Summary of the best results of the 6 algorithms along with the parameters' values (using top 15 attributes).

As can be noticed from tables 18 and 19, all of the algorithms performed very well and close to each other. The RBFNetwork algorithm showed slight improvement in the performance as I increased the value for the number of clusters. It obtained its best performance when the value of this parameter was 50 when using 5 attributes, and 100 when using 15 attributes. It scored the best performance among the other algorithms when using 5 attributes with an ROC area value of 0.978. Figure 23 shows the best output of this algorithm along with the parameter's value used to achieve this result. The MultiClassClassifier algorithm did not show any better performance when I changed the value of the randomWidthFactor parameter. It always gave the same results with ROC area value of 0.974 when I used 5 attributes and 0.985 when I used 15 attributes. The Bagging algorithm showed little improvement in its performance as I increased the value of the numIterations parameter. It scored its highest performance when the value of this parameter was 20 with 5 attributes and 30 with 15 attributes. While the RBFNetwork algorithm showed the best performance when I used 5 attributes, the RotationForest algorithm showed the best performance when I used 15 attributes. It showed an improvement in its performance as I increased the values of the two parameters; the

numIterations and the confidenceFactor. It scored its best performance when I used 15 attributes with an ROC area value of 0.993 when the values of the two parameters were 50 and 0.5 respectively. Figure 24 shows the best output of this algorithm along with the parameters' values used to achieve this result. The J48 and the REPTree algorithms scored almost the same performance in all the tests. Both algorithms did not show much improvement as I changed the values of their parameters. Both algorithms scored ROC values close to each other. Figures 25 and 26 illustrate the overall performance of the 6 best performing algorithms when I used 5 and 15 attributes respectively. Figure 27 shows how the ROC value improved for these 6 algorithms as the number of selected attributes increased. It can also be noticed that the ROC value did not improve much when going from 5 to 20 attributes. It got even lower for some algorithms.

Overall, the performance of all the algorithms was poor when I used the category-based attributes, especially for the third and fourth groups. Using the ranked attributes, I noticed that there was not much improvement in the performance of the algorithms when using 5, 10, 15 or 20 attributes. The performance was slightly improving. Therefore, the top 5 attributes can be used as the most discriminative attributes to distinguish between legitimate and phishing websites with correct classification and accuracy over 92%.

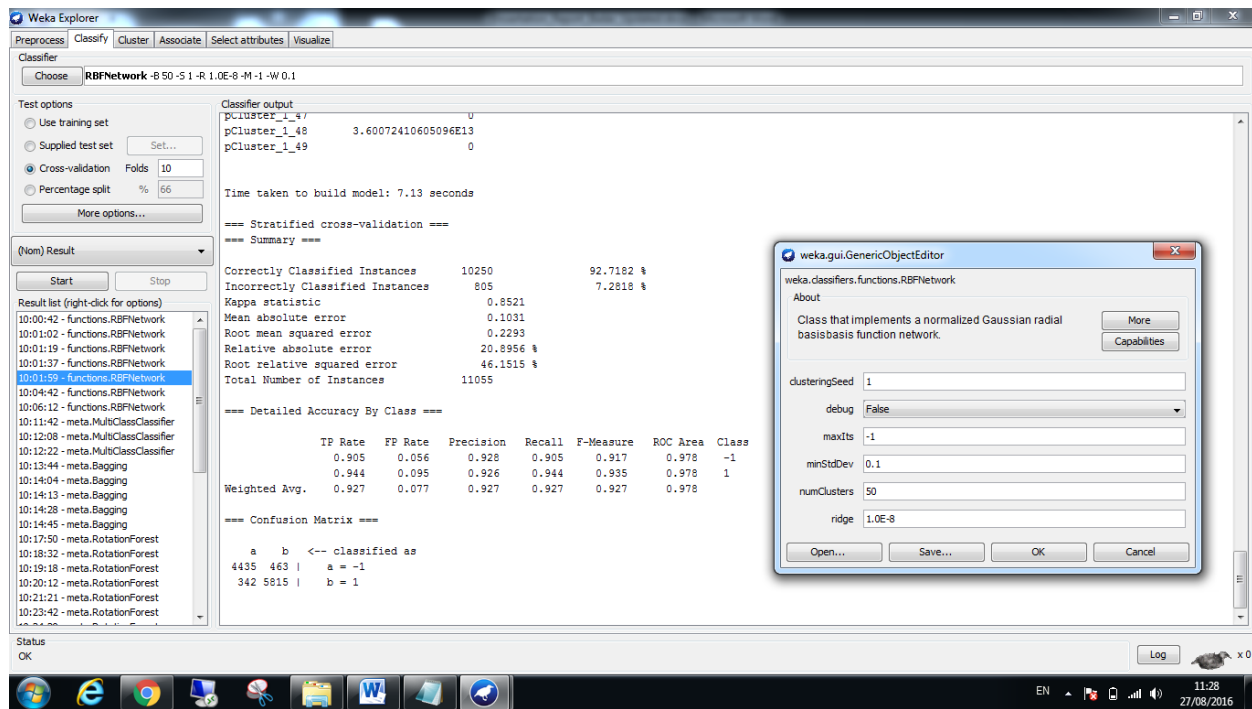


Figure 23: The best performing algorithm using 5 attributes “RBFNetwork” (with the used parameters’ values).

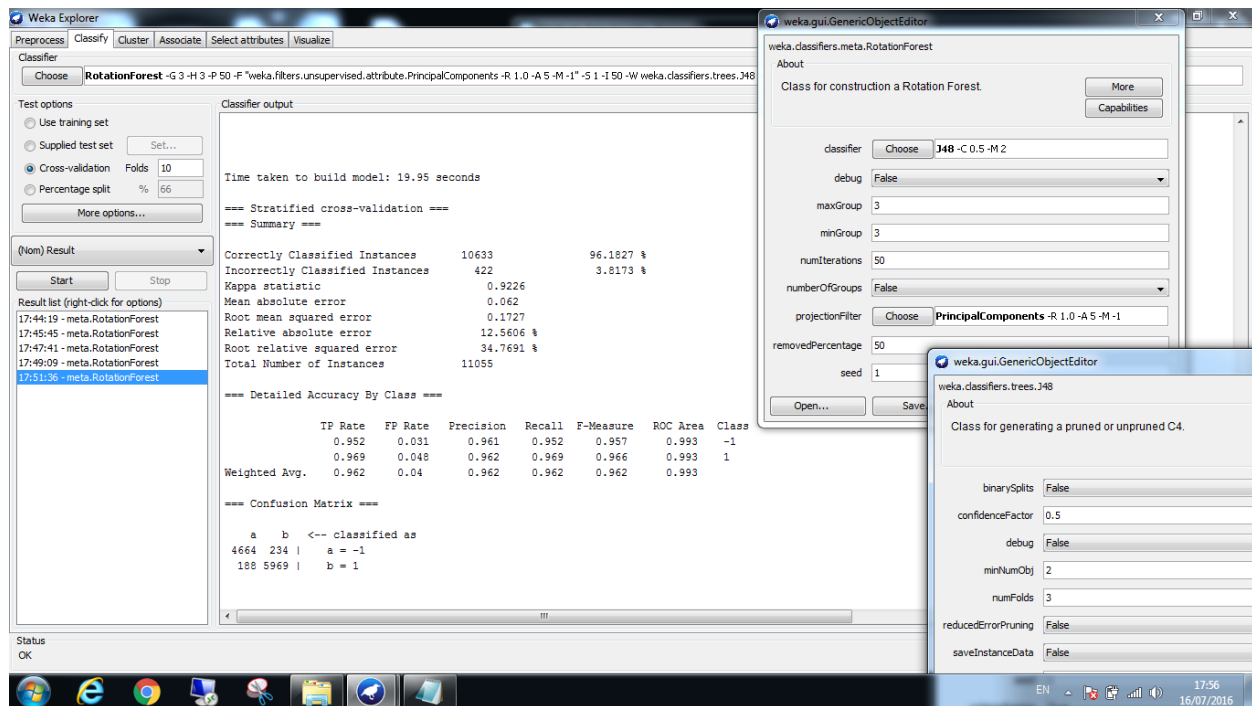


Figure 24: The best performing algorithm using 15 attributes “RotationForest” (with the used parameters’ values).

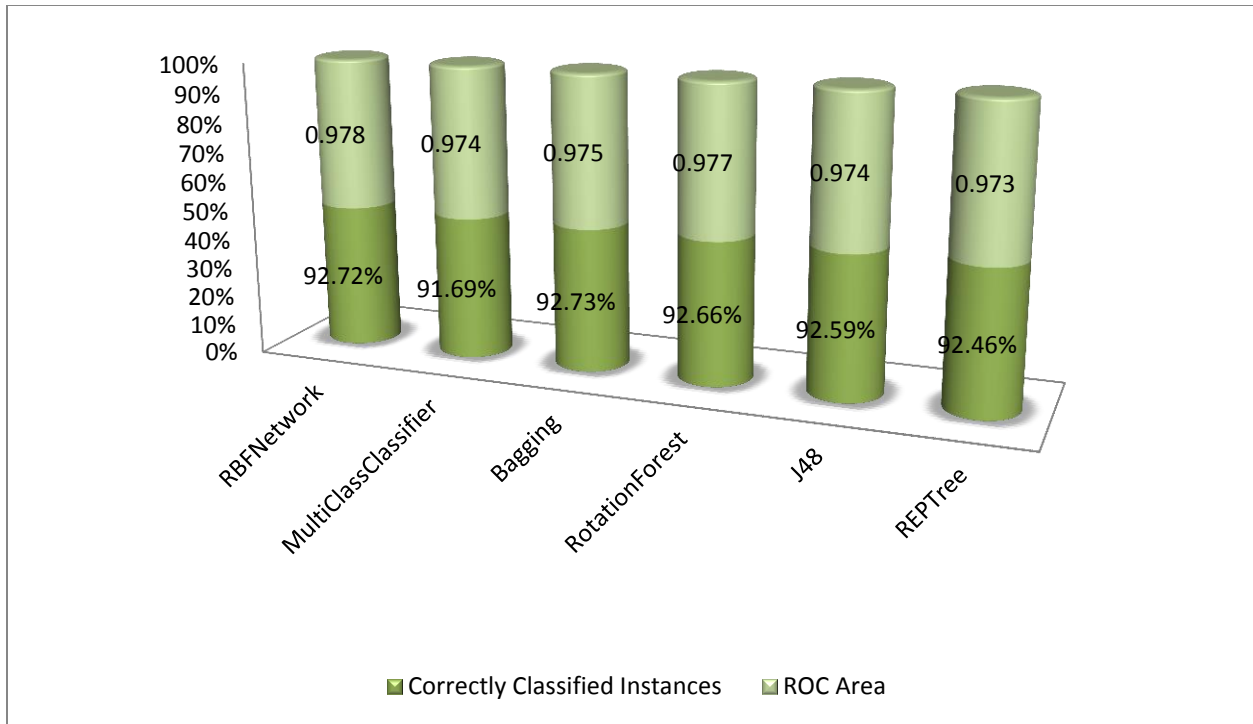


Figure 25: Accuracy % of the best 6 performing algorithms using 5 attributes.

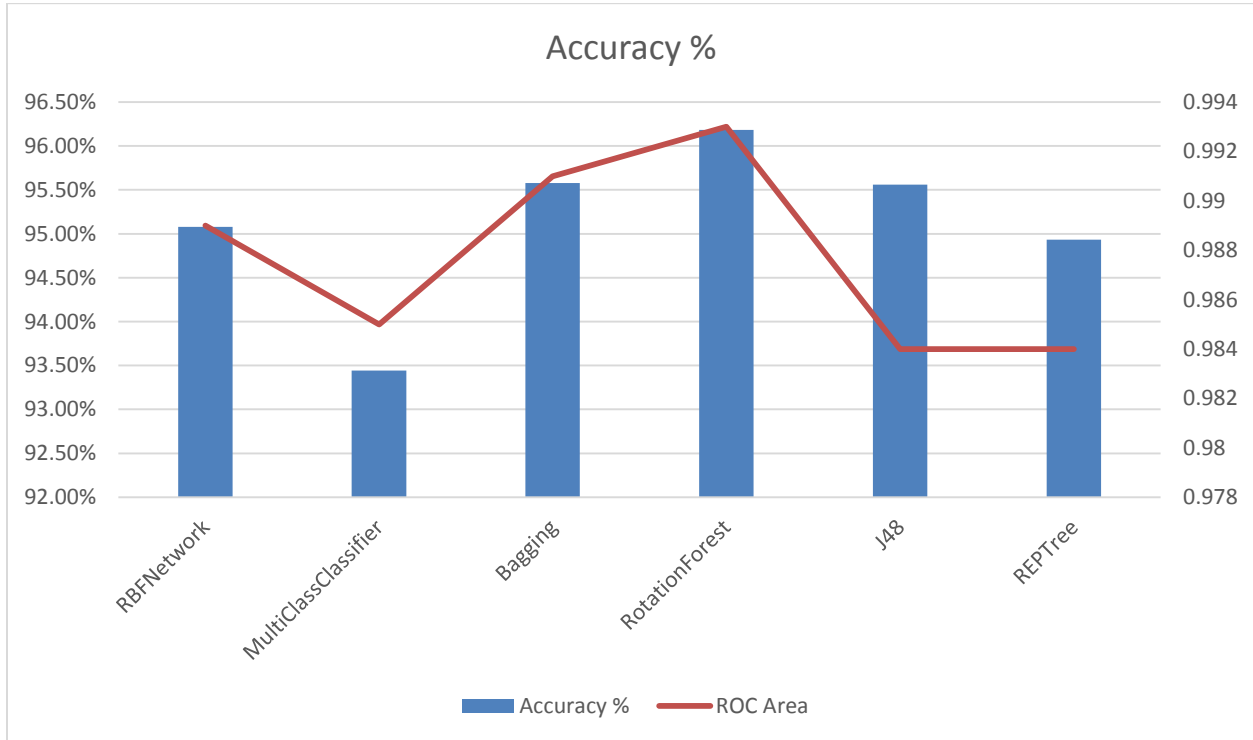


Figure 26: Accuracy % of the best 6 performing algorithms using 15 attributes.

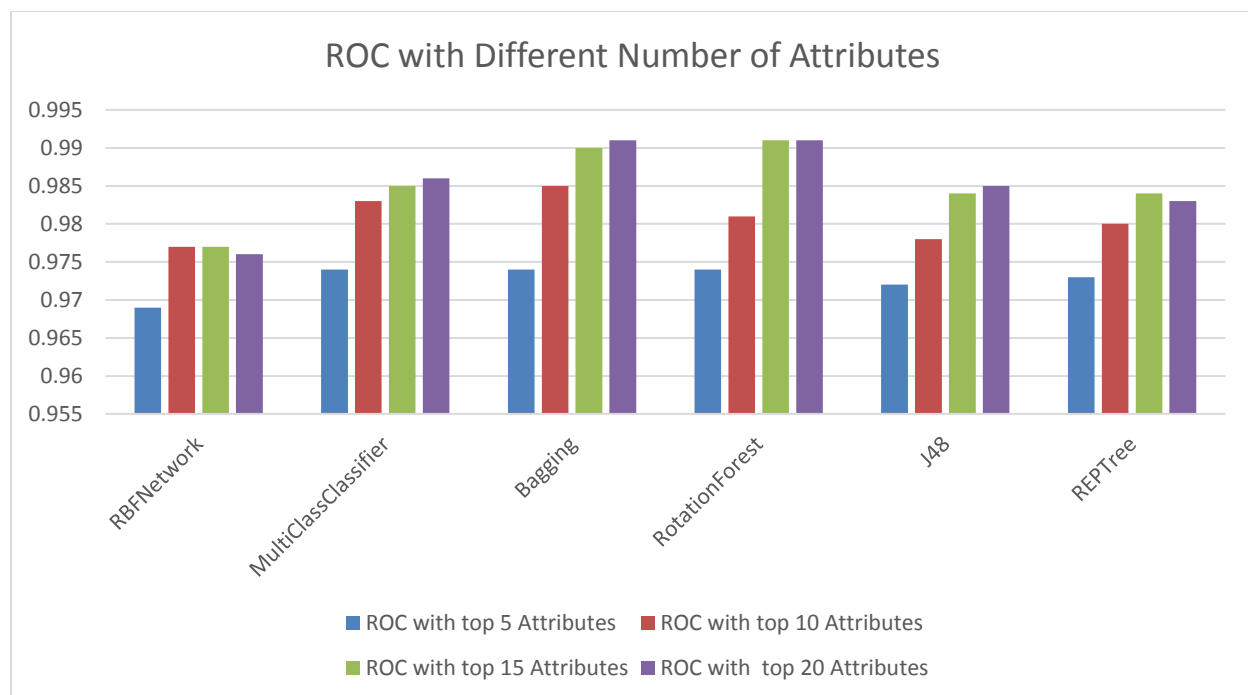


Figure 27: ROC value using different number of attributes.

9. Discussion of the Testing Results and Performance of the Algorithms

9.1 The Output and the Results

As can be noticed from the previous sections, getting the best performance and results requires different ways of testing. The algorithms should be tested with different combinations of attributes as well as trying different parameter values. The results showed that the dataset pre-processing stage is a very important stage that should be conducted before running the algorithms against the dataset. Even though some algorithms performed quite well before this stage, many other algorithms did not. In addition, all of the algorithms did not perform in to their best manner as the later sections have shown.

Selecting the most appropriate attributes for the algorithms is not an easy task. Especially, when the number of attributes in the dataset is quite high; as in this dataset. The three attribute evaluators I used have selected different combinations of attributes. While the first evaluator selected only 9 attributes, the second evaluator selected 23 attributes. There is a big difference between the numbers of selected attributes by these two evaluators. However, the good thing was that none of the 7 excluded attributes by the second evaluator was selected by the first evaluator. This gave me an indication that both evaluators at least agree on the excluded attributes. The third evaluator did not select specific attributes, but has ranked all the attributes and assigned a ranking score for each of them. This gave me a chance to select different combinations of attributes to use for the testing. In fact, this step was very important and helped me in exploring the attributes and finding the best results.

I have done quite good amount of testing with different combinations of attributes. Running the algorithms using the attributes that belong to different categories of websites' features did not show very good results. However, the results were interesting and indicated that there are some attributes of more importance to the performance of the algorithms than other attributes. In addition, the testing results showed that most of the algorithms produced better results as the number of used

attributes was increased. However, this was applicable only to some extent. When the number of selected attributes exceeded 15 attributes, the performance of the algorithms did not improve much. The best performance for most of the algorithms was obtained when I used the top 15 attributes which were ranked by the OneR attribute evaluator. However, there was not a big difference in the performance of the algorithms when moving from 5 to 15 attributes. Therefore, using only 5 attributes instead of all the 30 attributes is indeed sufficient to distinguish between legitimate and phishing websites. These 5 attributes can be considered as the most discriminative features among all of the other features.

While using the top 5 and 15 attributes have showed good results, the algorithms still did not perform in their best manner. Exploring the different parameters' values for each of the algorithms had even improved the results of the algorithms. As can be noticed from tables 18 and 19, changing the parameters' values had contributed as well to reaching the best performance of the algorithms. Using only 5 attributes, the RBFNetwork and the RotationForest algorithms showed very good performance. These two algorithms – as well as other algorithms – can classify the websites with an accuracy of over 92%.

The performance of the RBFNetwork and the RotationForest algorithms is very encouraging. Getting a classification model with an accuracy of 92% is promising. In addition, using only 5 features or attributes of a website in order to make this classification is reasonable. Using only 5 attributes instead of 30 attributes does not take long time to perform the classification process. This model can be used to protect many Internet users while they are surfing the Internet. This model can be embedded in web browsers. It can also be running as a small software agent in the users' computers. Whenever a user visits a website, the 5 features need to be extracted and then the model performs the classification task. The user is then notified whether the intended website to be visited is a legitimate or a phishing website. This model can be designed to work automatically while users are surfing the Internet. This approach can help users to browse the Internet safely and not to fall victims to phishing attacks.

9.2 Performance and Functionality of the Algorithms

While some algorithms did not perform very well, most of the other algorithms have shown good classification results. The 6 best performing algorithms are the RBFNetwork, MultiClassClassifier, Bagging, RotationForest, J48 and REPTree. Most of these 6 algorithms have produced models with over 92% classification accuracy. While the performance of these algorithms is quite high, the type of the algorithm and the classification method used by each of them is not the same as that of the others. This subsection is designated to explore some of the algorithms, and compare their internal functionalities.

The RBFNetwork Algorithm

The RBFNetwork algorithm is one of the neural network machine learning algorithms. The Radial Basis Function (RBF) neural network algorithm is a three-layer (input layer, hidden layer, and output layer) network which has one single hidden layer. It uses a mix of linear and non-linear learning algorithms. Unlike the classical neural network algorithms, the RBFNetwork algorithm uses more neurons in its hidden layer. The strength of this algorithm is based on its capability to self-adapt the allocation of the neurons in the hidden layer based on the classification problem. This allocation process of course depends on the size, type and distribution of the samples in the training dataset. In general, the components and structure of the RBFNetwork algorithm enables it to learn fast and to produce good classification results (Jia et al., 2014).

The Bagging Algorithm

The name of this algorithm illustrates the idea behind its functionality. The technique used here is the bagging (grouping) of multiple algorithms to solve a prediction or a classification problem. The bagging technique can be applied to different groups of algorithms. For example, bagging can be applied on decision trees as well as for Naïve Bayes algorithms. The final result taken is the average result of all of the used algorithms in the bag (Breiman, 1996, Tu et al., 2009). It was found that the

performance of the bagged algorithms is usually better than the performance of the individual algorithms. The idea here is the same as consulting a group of experts rather than just consulting only one expert (Tu et al., 2009). In most of the times, getting the views of multiple experts produces more and better insights, and leads to better decisions. For example, the results of one of the practical studies to compare multiple supervised learning algorithms showed that using bagged decision trees outperformed the performance of each single tree used individually (Caruana and Niculescu-Mizil, 2006). In another study to identify the illness of heart disease, the classification results of the used bagging algorithms were better than the results of the single decision trees (Tu et al., 2009). In the case of the phishing websites dataset used in this dissertation, the bagging algorithm performed very well with most of the combination sets of the attributes. This is due to the internal functionality of the bagging technique.

The RotationForest Algorithm

As the name indicates, the RotationForest algorithm continuously keeps spinning over a forest or a group of individual algorithms using different combinations of feature sets. This continuous rotation encourages the simultaneous improvement of the accuracy of the individual algorithms as well as the diversity within the forest. Diversity is encouraged through the extraction of different features for each of the individual base algorithms or classifiers. The base classifiers are usually – but not necessarily – decision trees algorithms. Overall, the rotation forest technique will produce individual classifiers that are very accurate and yield to low error rates (Rodriguez et al., 2006). In one of the experimental studies for the classification of hyperspectral remote sensing images, the results showed that the RotationForest algorithm produced more accurate results than bagging, AdaBoost, and Random Forest algorithms (Xia et al., 2014).

One very important point about the RBFNetwork and the RotationForest algorithms is that they seem to perform well in most of the cases due to their internal functionality. This means that they both gave good classification results with most combinations of attributes. This is a very good indication that these two algorithms can adapt themselves and work well with different combinations of features. In fact, these

are the kind of the algorithms that we are looking for here due to the continuous changes of the features of phishing websites. As mentioned earlier, phishers tend to always use new phishing tricks and techniques. The RBFNetwork and the RotationForest algorithms had proved to work very well with most of the features' combinations. Therefore, we should also expect them to perform well even with newly introduced phishing features.

The J48 and REPTree Algorithms

The J48 and REPTree algorithms are two of the common decision tree algorithms. Decision tree algorithms are some of the earlier machine learning techniques. These kinds of algorithms construct trees (or rule-based trees) in order to solve different medical, lingual, financial, scientific and many other classification problems. Usually, the decision tree is constructed by a top-down or general-to-specific approach. The constructed decision tree based on the training stage will then be applied on the dataset instances to perform the classification task. The process starts by using the root node to classify the dataset instances. If the root node is sufficient to classify all the instances, then the process is completed. Otherwise, more nodes and leaves are added to the tree recursively until all the instances belong to one of the constructed classes. J48 is in fact the C4.5 decision tree algorithm; which was an extension of the very popular ID3 modelling system. REPTree (Reduced Error Pruning Tree) algorithm is almost similar to the J48 algorithm in which it uses the C4.5 algorithm internally. It is a fast decision tree algorithm. It builds its decision tree by information gain or by variance reduction. Both the J48 and the REPTree algorithms showed almost the same performance (Apté and Weiss, 1997, Mohamed et al., 2012, Patil and Sherekar, 2013).

Conclusion

In this dissertation report, the phishing crime was discussed and introduced to users. The sophisticated nature and the continuous change in shape and design of phishing attacks made them very dangerous to Internet users and organisations. Different countermeasures had been implemented to fight phishing. Among these countermeasures were the black lists and the anti-phishing plugins. These countermeasure tools and techniques proved to be ineffective in protecting Internet users and organisations. For example, the black lists approach acted only after the phishing website had been discovered and added to the black list. This means that black lists were not able to protect users from newly created phishing websites. Anti-phishing plugins on the other hand did work automatically on the fly, but they had some detection limitations. The function of those plugins was based on the implementation instructions when those plugins were designed and created. Those plugins could not detect phishing websites that utilised new tricks and techniques not covered by the plugin instructions. It was necessary that those plugins get updated regularly with new instructions.

The limitations of the different previous phishing countermeasure approaches were mainly due to the lack of the up-to-date knowledge about the new phishing tricks and techniques utilised by phishers. Another draw-back reason was that some of those approaches required the intervention of users. It could not be always guaranteed that users do act, and whether they acted correctly or not if they do so. Due to these limitations and ineffectiveness of such approaches in protecting users, there was a real need for an automatic tool that checks and evaluates a website on the fly, and decides whether this website is a legitimate or a phishing one. Using business analytics tools and techniques proved to be very effective in distinguishing between legitimate and phishing websites based on the websites' features.

In this dissertation report, I applied business analytics techniques on a phishing websites dataset using Weka in order to explore different classification algorithms, and to develop a model that can protect Internet users from phishing. The algorithms I

tested belonged to different main categories such as bayes, functions, meta, rules and trees. In general, most of the algorithms performed quite well – with some algorithms showing better results than the others. The final results showed that there was no need to use all the 30 attributes or website features in order to decide whether the website was a legitimate or a phishing one. Algorithms such as the RBFNetwork and the RotationForest as well as some tree algorithms like the J48 and the REPTree showed very good classification results. For example, the RBFNetwork and the RotationForest algorithms could correctly classify the websites with an accuracy of over 92% using only 5 attributes. The models produced by these algorithms can be utilised to automatically protect Internet users from phishing attacks while they are surfing the Internet.

References

- ABURROUS, M., HOSSAIN, M. A., DAHAL, K. & THABTAH, F. 2010. Experimental Case Studies for Investigating E-Banking Phishing Techniques and Attack Strategies. *Cognitive Computation*, 2, 242-253.
- ACITO, F. & KHATRI, V. 2014. Business analytics: Why now and what next? *Business Horizons*, 57, 565-570.
- ADOMAVICIUS, G. & TUZHILIN, A. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17, 734-749.
- AGGARWAL, N., KUMAR, A., KHATTER, H. & AGGARWAL, V. 2012. Analysis the effect of data mining techniques on database. *Advances in Engineering Software*, 47, 164-169.
- ALKHOZAE, M. G. & BATARFI, O. A. 2011. Phishing websites detection based on phishing characteristics in the webpage source code. *International Journal of Information and Communication Technology Research*, 1.
- ALSHARNOUBY, M., ALACA, F. & CHIASSON, S. 2015. Why phishing still works: User strategies for combating phishing attacks. *International Journal of Human-Computer Studies*, 82, 69-82.
- ANDERSON, T. 2008. Towards a theory of online learning. *Theory and practice of online learning*, 2, 15-44.
- ANTI-PHISHING WORKING GROUP, A. 2014a. Phishing Activity Trends Reports - 2nd Quarter 2014.
- ANTI-PHISHING WORKING GROUP, A. 2014b. Phishing Activity Trends Reports - 4th Quarter 2013.
- APTÉ, C. & WEISS, S. 1997. Data MiningData mining with decision trees and decision rules. *Future Generation Computer Systems*, 13, 197-210.
- BARRETT, M. A., HUMBLET, O., HIATT, R. A. & ADLER, N. E. 2013. Big data and disease prevention: From quantified self to quantified communities. *Big data*, 1, 168-175.
- BELLMAN, S., LOHSE, G. L. & JOHNSON, E. J. 1999. Predictors of online buying behavior. *Communications of the ACM*, 42, 32-38.
- BREIMAN, L. 1996. Bagging predictors. *Machine learning*, 24, 123-140.
- BRYANT, R., KATZ, R. H. & LAZOWSKA, E. D. 2008. Big-data computing: creating revolutionary breakthroughs in commerce, science and society. December.
- BRYNJOLFSSON, E. & HITT, L. M. 2000. Beyond computation: Information technology, organizational transformation and business performance. *The Journal of Economic Perspectives*, 14, 23-48.
- CARUANA, R. & NICULESCU-MIZIL, A. 2006. An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd international conference on Machine learning*. Pittsburgh, Pennsylvania, USA: ACM.
- CHANG, C.-C. & LIN, C.-J. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2, 27.
- CHAUDHURI, S., DAYAL, U. & NARASAYYA, V. 2011. An overview of business intelligence technology. *Commun. ACM*, 54, 88-98.
- CHEN, H., CHIANG, R. H. & STOREY, V. C. 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS quarterly*, 36, 1165-1188.
- DAHAMIJA, R., TYGAR, J. D. & HEARST, M. 2006. Why Phishing Works.
- DAVENPORT, T. H. 2006. Competing on analytics. *harvard business review*, 84, 98.
- FERRANTI, J. M., LANGMAN, M. K., TANAKA, D., MCCALL, J. & AHMAD, A. 2010. Bridging the gap: leveraging business intelligence tools in support of patient safety and financial effectiveness. *Journal of the American Medical Informatics Association*, 17, 136-143.

- FRANK, E., HALL, M., HOLMES, G., KIRKBY, R., PFAHRINGER, B., WITTEN, I. H. & TRIGG, L. 2005. *Weka. Data Mining and Knowledge Discovery Handbook*. Springer.
- GASTELLIER-PREVOST, S., GRANADILLO, G. G. & LAURENT, M. Decisive heuristics to differentiate legitimate from phishing sites. *Network and Information Systems Security (SAR-SSI)*, 2011 Conference on, 2011. IEEE, 1-9.
- GRAEFF, T. R. & HARMON, S. 2002. Collecting and using personal data: consumers' awareness and concerns. *Journal of Consumer Marketing*, 19, 302-318.
- GRUMAN, G. 2006. The Four Stages of Enterprise Architecture. Available: <http://www.cio.com/article/2443291/service-oriented-architecture/the--four-stages-of--enterprise--architecture.html>.
- HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P. & WITTEN, I. H. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11, 10-18.
- HE, M., HORNG, S.-J., FAN, P., KHAN, M. K., RUN, R.-S., LAI, J.-L., CHEN, R.-J. & SUTANTO, A. 2011. An efficient phishing webpage detector. *Expert Systems with Applications*, 38, 12018-12027.
- HERZBERG, A. & GBARA, A. 2004. Trustbar: Protecting (even naive) web users from spoofing and phishing attacks. *Cryptology ePrint Archive*, Report 2004/155. <http://eprint.iacr.org/2004/155>.
- JEE, K. & KIM, G.-H. 2013. Potentiality of big data in the medical sector: focus on how to reshape the healthcare system. *Healthcare informatics research*, 19, 79-85.
- JIA, W., ZHAO, D., SHEN, T., SU, C., HU, C. & ZHAO, Y. 2014. A New Optimized GA-RBF Neural Network Algorithm. *Computational Intelligence and Neuroscience*, 2014, 6.
- KAUSAR, F., AL-OTAIBI, B., AL-QADI, A. & AL-DOSSARI, N. 2014. Hybrid Client Side Phishing Websites Detection Approach. *International Journal of Advanced Computer Science and Applications*.
- KIRDA, E. & KRUEGEL, C. Protecting users against phishing attacks with antiphish. *Computer Software and Applications Conference*, 2005. COMPSAC 2005. 29th Annual International, 2005. IEEE, 517-524.
- MITHAS, S., LEE, M. R., EARLEY, S., MURUGESAN, S. & DJAVANSHIR, R. 2013. Leveraging Big Data and Business Analytics [Guest editors' introduction]. *IT Professional*, 15, 18-20.
- MOHAMED, W. N. H. W., SALLEH, M. N. M. & OMAR, A. H. A comparative study of reduced error pruning method in decision tree algorithms. *Control System, Computing and Engineering (ICCSCE)*, 2012 IEEE International Conference on, 2012. IEEE, 392-397.
- MOHAMMAD, R. M., THABTAH, F. & MCCLUSKEY, L. An assessment of features related to phishing websites using an automated technique. *Internet Technology And Secured Transactions*, 2012 International Conference for, 2012. IEEE, 492-497.
- MOHAMMAD, R. M., THABTAH, F. & MCCLUSKEY, L. 2013. Predicting phishing websites based on self-structuring neural network. *Neural Computing and Applications*, 25, 443-458.
- MOHAMMAD, R. M., THABTAH, F. & MCCLUSKEY, L. 2014a. Intelligent rule-based phishing websites classification. *IET Information Security*, 8, 153-160.
- MOHAMMAD, R. M., THABTAH, F. & MCCLUSKEY, L. 2014b. Predicting phishing websites based on self-structuring neural network. *Neural Computing and Applications*, 25, 443-458.
- MOHAMMAD, R. M., THABTAH, F. & MCCLUSKEY, L. 2015. Tutorial and critical analysis of phishing websites methods. *Computer Science Review*, 17, 1-24.
- MORENO-TORRES, J. G., SÁEZ, J. A. & HERRERA, F. 2012. Study on the impact of partition-induced dataset shift on-fold cross-validation. *IEEE Transactions on Neural Networks and Learning Systems*, 23, 1304-1312.
- PATIL, T. R. & SHEREKAR, S. 2013. Performance analysis of Naive Bayes and J48 classification algorithm for data classification. *International Journal of Computer Science and Applications*, 6, 256-261.
- PURKAIT, S. 2012. Phishing counter measures and their effectiveness – literature review. *Information Management & Computer Security*, 20, 382-420.

- RAGHUPATHI, W. 2010. Data Mining in Health Care. In: KUDYBA, S. (ed.) *Healthcare Informatics: Improving Efficiency and Productivity*. Francis: Francis.
- RODRIGUEZ, J. J., KUNCHEVA, L. I. & ALONSO, C. J. 2006. Rotation forest: A new classifier ensemble method. *IEEE transactions on pattern analysis and machine intelligence*, 28, 1619-1630.
- SAHAY, B. & RANJAN, J. 2008. Real time business intelligence in supply chain analytics. *Information Management & Computer Security*, 16, 28-48.
- SHAHRIAR, H. & ZULKERNINE, M. 2012. Trustworthiness testing of phishing websites: A behavior model-based approach. *Future Generation Computer Systems*, 28, 1258-1271.
- SHEKOKAR, N. M., SHAH, C., MAHAJAN, M. & RACHH, S. 2015. An Ideal Approach for Detection and Prevention of Phishing Attacks. *Procedia Computer Science*, 49, 82-91.
- TRKMAN, P., MCCORMACK, K., DE OLIVEIRA, M. P. V. & LADEIRA, M. B. 2010. The impact of business analytics on supply chain performance. *Decision Support Systems*, 49, 318-327.
- TU, M. C., SHIN, D. & SHIN, D. A comparative study of medical data classification methods based on decision tree and bagging algorithms. Dependable, Autonomic and Secure Computing, 2009. DASC'09. Eighth IEEE International Conference on, 2009. IEEE, 183-187.
- VOSEN, G. 2013. Big data as the new enabler in business and other intelligence. *Vietnam Journal of Computer Science*, 1, 3-14.
- WAIKATO, M. L. G. A. T. U. O. *Weka Download* [Online]. Available: <http://www.cs.waikato.ac.nz/ml/weka/downloading.html> [Accessed].
- WARD, M. J., MARSOLO, K. A. & FROEHLE, C. M. 2014. Applications of business analytics in healthcare. *Business Horizons*, 57, 571-582.
- WHITTAKER, C., RYNER, B. & NAZIF, M. 2010. Large-scale automatic classification of phishing pages.
- XIA, J., DU, P., HE, X. & CHANUSSOT, J. 2014. Hyperspectral remote sensing image classification based on rotation forest. *IEEE Geoscience and Remote Sensing Letters*, 11, 239-243.
- ZHANG, D., YAN, Z., JIANG, H. & KIM, T. 2014. A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites. *Information & Management*, 51, 845-853.

Appendix A: List of Figures

Figure No.	Description	Page
Figure 1	The dataset attributes and their values in the ARRF file	35
Figure 2	The actual attributes' values of the dataset in the ARRF file	36
Figure 3	The dataset file once loaded into Weka	37
Figure 4	Error displayed by Weka to indicate a missing "@" symbol	37
Figure 5	Details displayed for a specific attribute in the list	39
Figure 6	Details displayed for a specific attribute in the list	39
Figure 7	Output of the BayesNet algorithm (without dataset pre-processing)	44
Figure 8	Output of the HNB algorithm (without dataset pre-processing)	45
Figure 9	Output of the Logistic algorithm (without dataset pre-processing)	45
Figure 10	Output of the RBFNetwork algorithm (without dataset pre-processing)	46
Figure 11	Output of the Bagging algorithm (without dataset pre-processing)	46
Figure 12	Output of the MultiClassClassifier algorithm (without dataset pre-processing)	47
Figure 13	Output of the RotationForest algorithm (without dataset pre-processing)	47
Figure 14	Output of the ClassificationViaClustering algorithm (without dataset pre-processing)	48
Figure 15	Output of the Winnow algorithm (without dataset pre-processing)	48
Figure 16	Output of the Ridor algorithm (without dataset pre-processing)	49
Figure 17	Output of the BFTree algorithm (without dataset pre-processing)	49
Figure 18	Output of the J48 algorithm (without dataset pre-processing)	50
Figure 19	Output of the REPTree algorithm (without dataset pre-processing)	50
Figure 20	Tree graph of the J48 algorithm (without dataset pre-processing)	51
Figure 21	Tree graph of the REPTree algorithm (without dataset pre-processing)	51

Figure 22	The top 5 and 15 ranked attributes, their number and their ranking score	70
Figure 23	The best performing algorithm using 5 attributes “RBFNetwork” (with the used parameters’ values)	74
Figure 24	The best performing algorithm using 15 attributes “RotationForest” (with the used parameters’ values)	74
Figure 25	Accuracy % of the best 6 performing algorithms using 5 attributes	75
Figure 26	Accuracy % of the best 6 performing algorithms using 15 attributes	75
Figure 27	ROC value using different number of attributes	76

Appendix B: List of Tables

Table No.	Description	Page
Table 1	Details of the dataset attributes	19
Table 2	The 13 different algorithms chosen to run on the dataset	41
Table 3	Summary of the results of the algorithms (run without dataset pre-processing)	42
Table 4	The output results of the Attributes' Evaluators	53
Table 5	Summary of the results of the algorithms (using attributes selected by CFS Subset Evaluator)	55
Table 6	Summary of the results of the algorithms (using attributes selected by Consistency Subset Evaluator)	57
Table 7	Summary of the results of the algorithms (using the top 5 attributes ranked by the OneR Attribute Evaluator)	59
Table 8	Summary of the results of the algorithms (using the top 10 attributes ranked by the OneR Attribute Evaluator)	60
Table 9	Summary of the results of the algorithms (using the top 15 attributes ranked by the OneR Attribute Evaluator)	61
Table 10	Summary of the results of the algorithms (using the top 20 attributes ranked by the OneR Attribute Evaluator)	62
Table 11	Summary of the results of the algorithms (using attributes number 1 to 12 that belong to the Address Bar Category)	64
Table 12	Summary of the results of the algorithms (using attributes number 13 to 18 that belong to the Abnormality Category)	65
Table 13	Summary of the results of the algorithms (using attributes number 19 to 23 that belong to the HTML and JavaScript Category)	66
Table 14	Summary of the results of the algorithms (using attributes number 24 to 30 that belong to the Domain Category)	67
Table 15	The similar results of all algorithms (when using only top one or two	68

	attributes)	
Table 16	The top 5 and 15 ranked attributes used for the final testing	69
Table 17	The different parameters tested for each algorithm	71
Table 18	Summary of the best results of the 6 algorithms along with the parameters' values (using top 5 attributes)	71
Table 19	Summary of the best results of the 6 algorithms along with the parameters' values (using top 15 attributes)	72

Appendix C: Applications of Business Analytics/Business Intelligence

Today, it is very hard to find a successful enterprise that is not employing business intelligence techniques in running its business and in making strategic decisions. Business analytics technologies are widely used within different sectors. They are adopted in many financial institutions, supply-chain industries, transportation management, telecommunication companies, health care services and educational institutions (Chaudhuri et al., 2011). This section is designated for illustrating the power of business analytics in different business areas.

C.1 Applications of BA/BI in Commercial Enterprises

C.1.1 Business Analytics: the Main Tool to Success

Very well-known competitive enterprises such as Amazon, Capital One, Harrah's, the Boston Red Sox and other enterprises have shaped their successful business based on their strategic and wise collection, analysis and utilisation of data. These enterprises are competing heavily on analytics because it is main driver of their success. In one of the surveys conducted by Bloomberg BusinessWeek, it was found that 97% of companies having profits over \$100 million were utilising some kinds of business analytics techniques (Chen et al., 2012). Unlike other organisation that have employed analytics on the departmental levels, these enterprises have utilised analytics on the enterprise level. This means that these enterprises have implemented analytics on every aspect related to the enterprise; this includes products, services, customers, employees, management, assets, buildings and so on. Most importantly, this enterprise level approach is supported and driven by the top management (Davenport, 2006). These enterprises have reached a level of knowledge that enables them to accurately predict what products their customers are interested in, how much money their customers are willing to spend on such products and what motivates their customers to buy such products. In addition, analytics has enabled these enterprises to correctly and successfully select the types of promotions their customers – individuals and groups –

are more likely to accept and take. Not only that, but analytics has also enabled them to predict the best time to send such promotions (Davenport, 2006).

C.1.2 Shared Characteristics of the Competing Enterprise

Today, business analytics is not attracting only retail companies, but also finance companies, travel agencies, sport clubs, entertainment organisations and many other companies from different sectors (Davenport, 2006). Among all of these companies, only few enterprises succeed to fully utilise analytics and reach the business modularity stage. In this stage, all enterprise business processes and their supporting technologies become modules that can be reused for efficiency and recombined for agility (Gruman, 2006). In their study on analysing the top 32 organisations utilising business analytics, the authors in (Davenport, 2006) found that only 11 companies are fully employing analytics in their business at the enterprise level. They also found that these few 11 companies share some common characteristics that are not found on other companies. First, most of those 11 companies tend to extensively use modelling and optimisation. While normal companies use only in-house data to produce basic statistics – average sales and profit for example, those leading enterprises utilise internal (from their systems) and external (from blogs and social media sites) data in order to get the full picture about their customers. They also use many predictive modelling techniques to extract accurate knowledge about their customers which brings them the highest profit possible (Davenport, 2006). For example, to maximise the number of potential customers signing for credit cards, Capital One conducts around 30,000 experiments each year. Second, the focus of those leading companies is directed towards the enterprise level on every aspect of the business. While the marketing side usually requires more social and communication skills, these companies utilise data-driven marketing approaches as well. Third, the authors also found that the CEOs (Chief Executive Officers) of most of these leading enterprises were very talented executives with the desire to change. Examples of the very successful CEOs include Jeff Bezos of Amazon, Loveman of Harrah's and Rich Fairbank of Capital One. It was not just the leadership skills, but those CEOs were very enthusiastic to change and believed in the quantitative approaches and numbers (Davenport, 2006).

C.2 Applications of BA/BI in the Healthcare Sector

The utilisation of information technology in the health care systems has grown rapidly worldwide. Nowadays, many countries use electronic records to store and retrieve patients' medical information from the various health care systems. In the United States, the use of electronic health records has almost doubled within only 4 years; from 2008 to 2012, and almost 44% of U.S hospitals were already using at least the basic electronic health services (Ward et al., 2014). The extensive deployment of electronic systems in the health care sector has led to the generation of huge volumes of data. It was reported that the U.S healthcare system alone produced about 150 Exabyte of data in 2011, and was expected to reach the Zettabyte level within few years (Raghupathi, 2010). Storing and retrieving data was not an issue, but utilising this data to improve the health care services was indeed a big challenge. However, the emergence of business analytics created many opportunities for the development, improvement and customer satisfaction in the health care sector. Business analytics technologies helped in transforming this huge data into valuable knowledge which helped in making faster and better medical decisions. Besides the advantages on the medical side, business analytics also contributes to great reduction in healthcare operational cost. It was estimated that applying analytics tools in the U.S healthcare sector has led to reduction cost of about \$300 billion each year (Raghupathi, 2010).

The utilisation of analytics in the healthcare sector has improved the healthcare services in different ways. It has contributed to the enhancements of faster disease discovery, treatment efficiency and healthcare service delivery (Ward et al., 2014). In fact, analytics has shifted the way most healthcare providers are functioning. Instead of fighting diseases, the objectives have changed now to prevent them from occurring based on accurate evidence and data-driven diagnosis and treatment (Chen et al., 2012). Evidence-based diagnosis and treatment involves the utilisation of clinical information as well as historical health details in order to decide on the best medical treatment and procedures for individual patients (Jee and Kim, 2013). For example, applying analytical techniques to patients' genetic data led to the early discovery and treatment of potential diseases. On the other hand, using visual control charts –

supported by analytics – helped to improve different services such as room utilisation and patient waiting time (Ferranti et al., 2010). Predicting and planning for patient flow in clinics and emergency units have been one of the main fields targeted by analytics. Applying analytics helped in creating smooth patient flow and in prioritising patients based on their illness severity (Ward et al., 2014). In addition to supporting and offering better and higher quality medical treatments, analytics has also played a great role in healthcare cost reduction. Not only cost reduction for patients, but also for governments spending on the healthcare sector (Ferranti et al., 2010, Jee and Kim, 2013).

C.3 Applications of BA/BI in the Supply Chain Sector

Supply chain has become one of the main components in the production and distribution of products and services. The sophisticated multiple procedures and processes involved in supply chain contribute heavily to the great complexity of managing this type of business. Different organisations, people, services and products are involved in supply chain – hence the amount of data and information to be processed is really huge and from different sources (Trkman et al., 2010). Supply chain management has always been a burden since the goal is performance improvement and optimisation. Proper supply chain management is one of the main factors for minimising the operational cost and maximising the financial outcomes through optimisation of inventory and sales. This complex management task includes various processes such as planning, identifying targets and measures, communication, monitoring and reporting. Until today, different supply chain management systems have been implemented and used. However, many of them were not effective in achieving competitive advantages, despite the huge investments on them (Sahay and Ranjan, 2008). Business analytics helped in reducing this management burden. Business analytics tools and techniques proved to be very effective in analysing huge supply chain data and generate effective and efficient solutions to different problems related to the supply chain process. Based on the insights generated by business analytics, executives and managers can make faster decisions and take proper actions (Sahay and Ranjan, 2008).

Appendix D: Weka Datasets Websites and Repositories

Below are some of the websites and repositories for many datasets that work in Weka:

- UCI Archive
<https://archive.ics.uci.edu/ml/datasets.html>
- The University of Waikato Weka Dataset Repository
<http://www.cs.waikato.ac.nz/ml/weka/datasets.html>
- UCI Knowledge Discovery in Databases Archive
<http://kdd.ics.uci.edu/>
- Dr. Gary M. Weiss Page (Associate Professor & Director of the [WISDM Lab](#))
<http://storm.cis.fordham.edu/~gweiss/data-mining/datasets.html>
- Seasar Repositories
<http://repository.seasar.org/Datasets/UCI/arff/>