Se former autrement

HONORIS UNITED UNIVERSITIES

2022/2023

Prepared at: **ESPRIT**

Supervised by: **Ms. Dorra Trabelsi**

Submitted by: **Jihene Saidi**

Arij Ben Nasr

Mouhib Dalhoumi

Raed Bahria

Youssef Ben Majed

Ahmed ElAmri

Acknowledgment

At the end of this work, we would like to present our most sincere thanks to Ms. Dorra Trabelsi who agreed to supervise us and who gave us the benefit of her extensive knowledge and valuable advice during our project.

However, it should be noted that this work would not have been possible without the invaluable knowledge and know-how acquired in our honorable school "Private Higher School of Engineering and Technology of Tunis".

It is therefore with great pride that we address our most distinguished thanks to all our teachers.

May they find here, as well as anyone who has contributed to the completion of this project, directly or indirectly, the expression of our sincere gratitude.

Finally, we express our most sincere thanks to the members of the jury.

Contents

Table of figures

Summary 1

In a data science project focused on cybersecurity, business understanding would. Involve gaining a comprehensive understanding of the organization's cybersecurity. Needs, challenges, and goals. This includes understanding the nature and scope of foyer threats faced by the organization, its current security infrastructure, and the specific aims of the project (e.g., improving threat detection, reducing data breaches, etc.). It also involves understanding the organization's culture and governance, regulatory requirements, and budget constraints that may affect the project.

How unsafe it could be to live in a digital Era?

How Secure Is Our Data ?

Figure 1 : Cybersecurity as a value solution

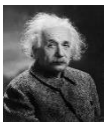
General introduction

The volume of data generated every day is increasing at a surprising rate. Nearly 5 quintillion bytes of data are being created daily. With the rise in data, there has also been a surge in data breaches. Hacking and penetrating a system using various tools have become a significant cause of concern for organizations and individuals Worldwide. Sophisticated data science techniques are now widely used by attackers to break into a system. The question is if data science can be used to take charge of the system, can it be used to prevent it from hacking? The answer is yes; with the use of data science in cyber security, it has become easy to predict vulnerability in a system, which in turn prevents the potential risk of breach by taking appropriate measures.

1. **Business Understanding**

Introduction

Business Understanding: Obviously one of the most critical phases of the data science life cycle.



Einstein is quoted as having said, "If I had an hour to solve a problem, I'd spend 55 minutes thinking about the problem and five minutes thinking about solutions."

That may be a bit extreme, but A well-defined problem often has its own solution, and that solution is usually Obvious and Straightforward. By defining problems properly, you make them easy to solve which means saving time, money, and resources.

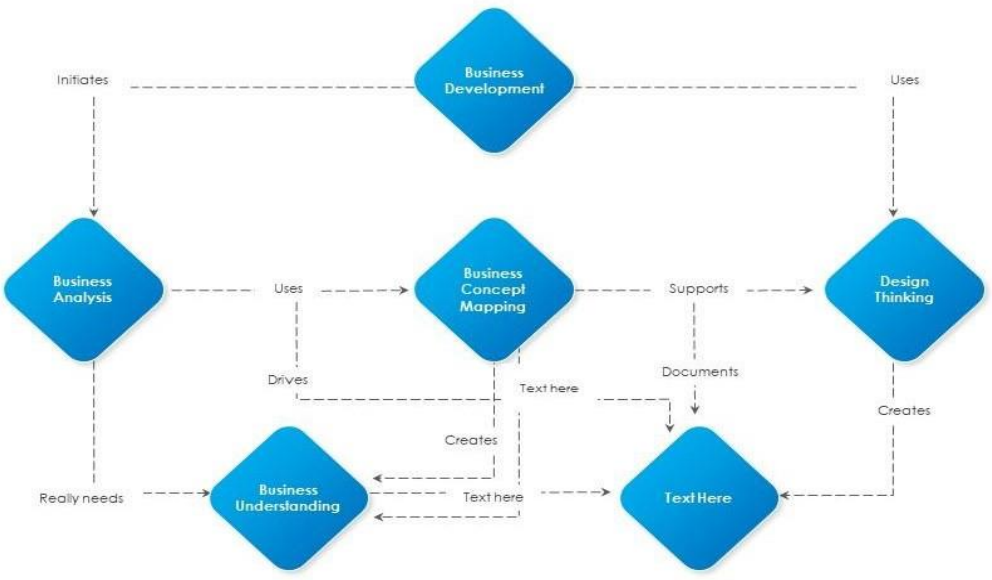


Figure 2 : Business Understanding Concept

1.1 Cybersecurity Understanding

Cybersecurity in a business context refers to the protection of internet connected. systems, including hardware, software, and data from cyber threats such as hacking, phishing, malware, and data breaches. This includes implementing security measures like firewalls, antivirus software, intrusion detection and prevention systems, and encryption, setting up security policies and procedures, ensuring privacy, and keeping the availability and integrity of data, regularly checking, and assessing risk, and having a plan in place for responding to security incidents. The goal is to minimize the impact of security incidents on the business operations, reputation, and financial stability. The field of cybersecurity is constantly evolving to keep up with new technologies and emerging threats.

Figure 3 : Cybersecurity advantages

1.2 Email Phishing Attacus

We've all been the recipient of spam emails before. Spam mail, or junk mail, is a type of email that is sent to a massive number of users at one time, frequently containing cryptic messages, scams, or most dangerously, phishing content. While spam emails are sometimes sent manually by a human, most often, they are sent using a bot. Most popular email platforms, like Gmail and Microsoft Outlook, automatically filter spam emails by screening for recognizable phrases and patterns. A few common spam emails include fake advertisements, chain emails, and

impersonation attempts. While these built-in spam detectors are usually pretty effective, sometimes, a particularly well-disguised spam email may fall through the cracks, landing in your inbox instead of your spam folder. Clicking on a spam email can be dangerous, exposing your computer and personal information about different types of malware. Therefore, it's important to implement additional safety measures to protect your device, especially when it handles sensitive information like user data. A study conducted by PhishMe showed that 91% of cyberattacks started with a phishing email.

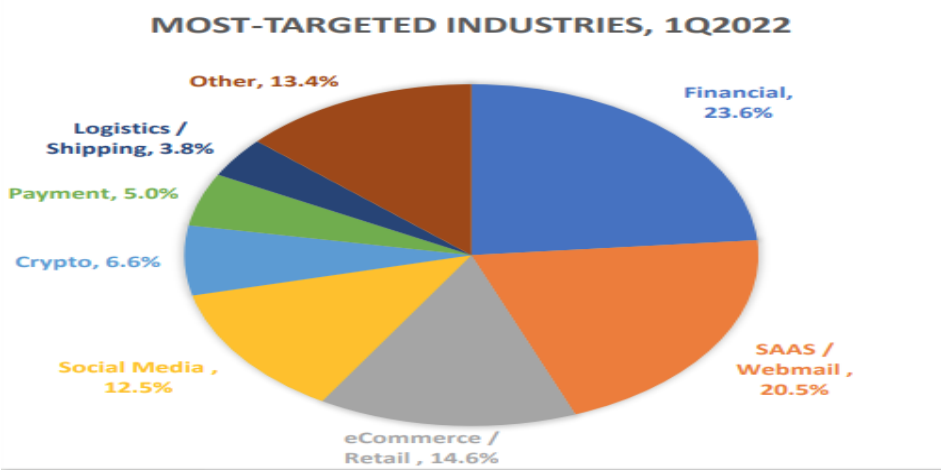


Figure 4 : MOST-TARGETED INDUSTRIES

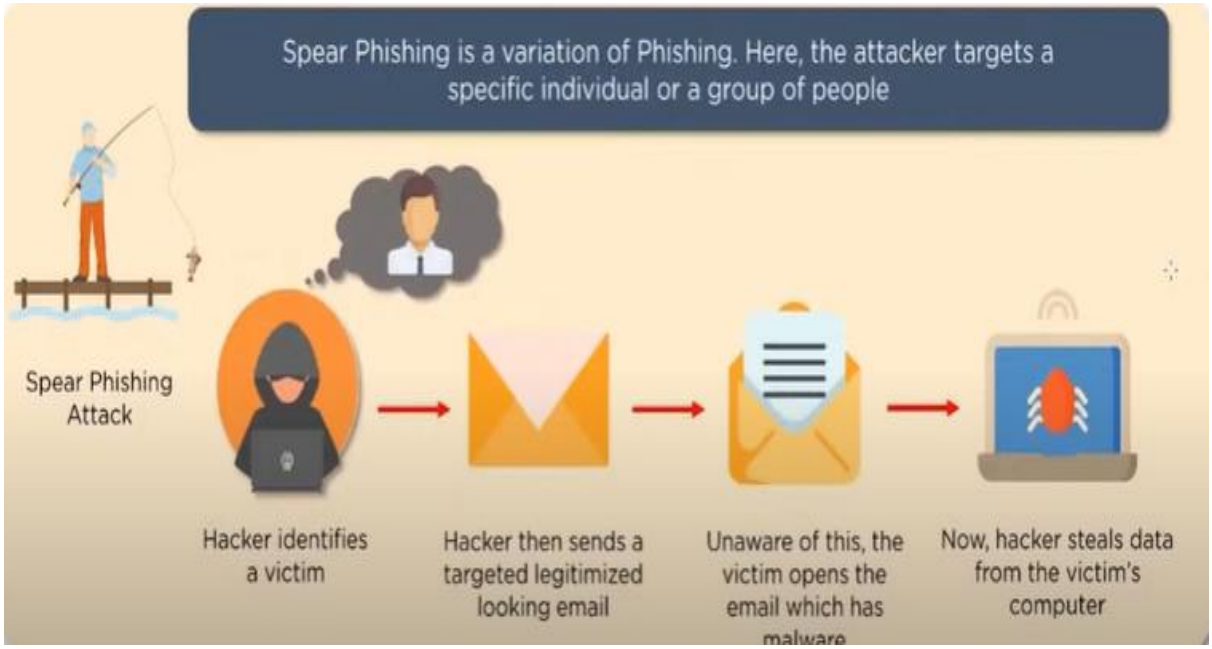


Figure 5 : Phishing Attack steps

Embodiment of events: Let's say that an attacker is targeting your company and you're the security admin responsible for protecting your organization So the attacker is going to start by doing some research on your company on LinkedIn, where they can make a list of employees that they want to target. The attacker then sends personalized emails to these employees trying to get them to click on a link or respond with sensitive information. The victim, one of the employees that's being targeted with this phishing attack is going to go through his inbox in the morning and will find some emails come up and he has to choose whether he thinks the email is a phishing email or a safe email.The email domain in the From field is important to distinguish between a phishing email and a normal email.

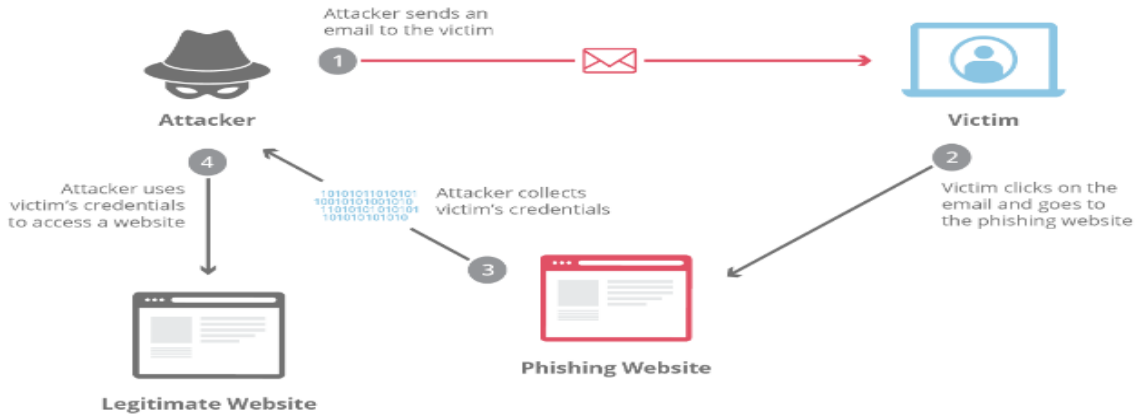


Figure 6 : Phishing event

Consequences of phishing : If the employee responds to the attacker's email or clicks on a link in the email , he may end up giving the attacker access to sensitive data. A report from Intel Security on phishing attacks showed that 97% of users are not able to identify a sophisticated phishing attack . When we are going through our inbox and trying to work, we don't have the time to look at these little details in emails and spot phishing emails.

Impact of phishing : Facebook and Google (€90 million):Between 2013 and 2015, two of the world's biggest tech firms were duped out of $100 million (about €90 million at the time) after falling victim to a fake invoice scam. A Lithuanian man, Evades Rimasauskas, noticed that both organizations use the Taiwanese infrastructure supplier Quanta Computer. He sent a series of bogus multimillion-dollar invoices replicating the supplier over two years, complete with contracts and letters that appeared to have been signed by executives and agents of Facebook and Google. The scam was eventually discovered, and Facebook and Google took legal action. They recovered just under half of the stolen money, while Ramanauskas was

arrested and extradited from Lithuania. In December 2019, he was sentenced to five years in prison. The problem that we are trying to solve is email phishing attacks.

1.3 Project Objectives

To summarize, the primary goal of our project is to ensure the privacy of information, the correctness of data, Hide phishing emails immediately to prevent the user from accidentally opening it and access to authorized users. CIA Triad have served as the industry standard for computer security since the time of the first mainframes.

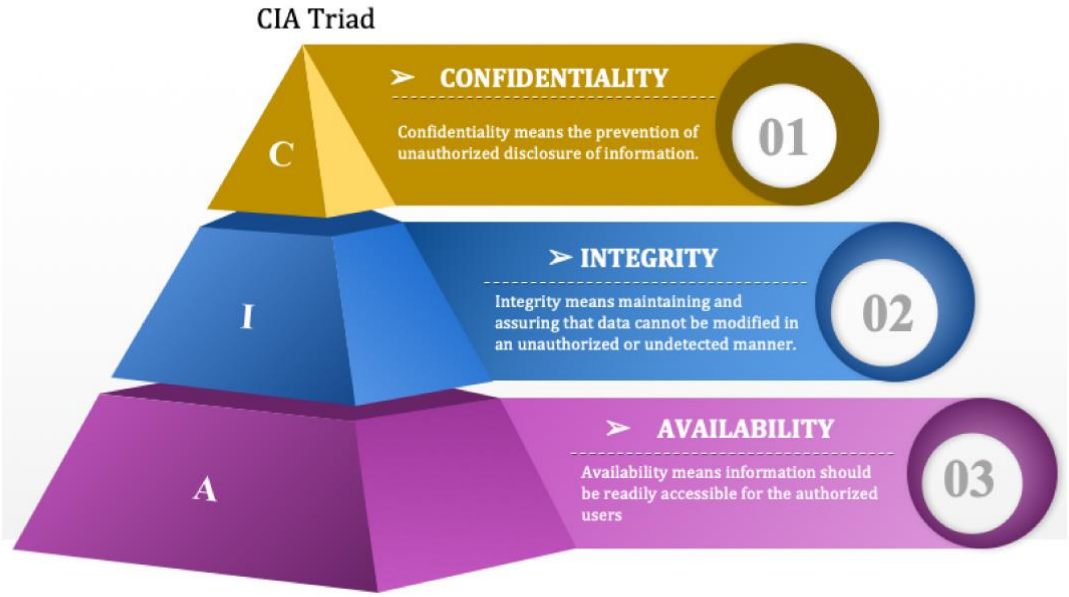


Figure 7 : CIA Triad

1.4 Functional Requirements

Functional requirements are specific tasks and functionality that a system must perform to meet the needs of the stakeholders and Improve security of personal information online.

Email Analysis: The system should have the capability to analyze and extract information from emails, such as URLs, IP addresses, and email headers to Recognize the sources of phishing emails.

Reporting: The system should supply detailed reports on the phishing attacks,

including their types, sources, and targets.

Data storage: The application should store the results of the phishing detection and sentiment classification in a database for future analysis and reporting.

Phishing Detection: The system should be able to detect and categorize phishing emails based on their characteristics, such as the sender, content, and attached files.

Email Classification: The system should have the ability to classify emails into different categories, such as spam, legitimate, phishing, and other types of email attacks.

1.5 Non functional Requirements

Performance: The system must be able to perform its functions on time and with sufficient speed, even when processing large amounts of data

Scalability: The system must be able to scale up or down as the organization's needs change and be able to handle increased workloads without significant degradation of performance.

Reliability: The system must be reliable and robust, with minimal downtime and failures.

Security: The system must have strong security measures in place to protect against cyber threats and data breaches and to Improve security of personal information online

Usability: The system must be easy to use and understand for all relevant stakeholders, including security personnel and end-users.

Maintainability: The system must be easily maintainable, with clear documentation and well-defined processes for updates and maintenance. Interoperability: The system must be able to integrate with other systems and tools used by the organization.

Conclusion

Your understanding of what customers want , combined with your employees' know-how , can be regarded as your knowledge base .Using this knowledge in the right way can help you run your business more efficiently , decrease business risks and exploit opportunities to the full . This is known as the knowledge advantage .

2. Analytic Approach (Data Understanding)

"if we have data we are in business, if we don't have data we are out of business."

2.1 Life Cycle Of Data Science Project

How is a typical data science project expected to be executed across different enterprises both BIG AND SMALL? We will work with the IBM master methodology, it is an agile methodology, there is fast back and forth allowing us to execute the problems quickly from the beginning of the project is a cyclic and iterative methodology with Feedback client.

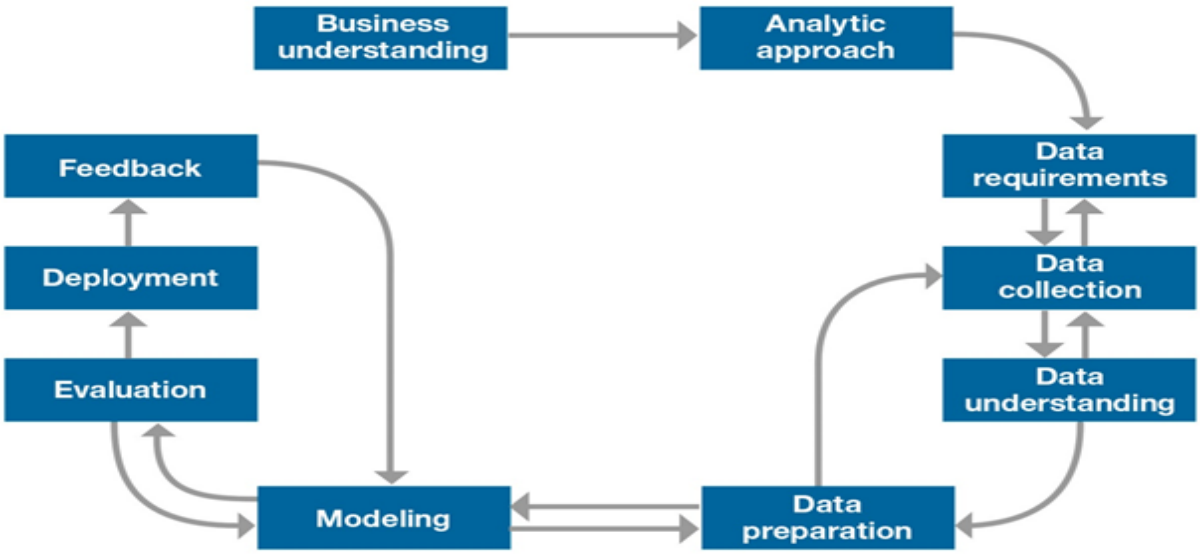


Figure 8 : IBM Master Plan architecture

2.2 Data Science Meaning

Data science is an interdisciplinary(developed) field that is a scientific method, Processes, algorithms, systems to extract knowledge and Insights from structured and unstructured data.

- Structured Data means relational database: Oracle, MySQL, SQLServer
- Unstructured Data means Images, videos, Sound, Text, and they could be stored in non-relational databases such as hadoop, cassandra, mongoDB.

Data science is interrelated to Data Mining and Big Data.

2.3 Data Science In Cybersecurity

In our situation, cybersecurity means data security, the need to prevent threats and spams with respect to the data Data science can be used to predict and prevent email phishing in several ways:

- **Email Content Analysis:** Use Natural Language Processing (NLP) to analyze the content of emails and find ones that have phishing attempts by detecting specific keywords and phrases used in phishing emails.
- **URL analysis:** Analyze URLs in emails to find phishing attempts by checking if they lead to known phishing sites or suspicious domains.
- **Email Authentication:** Verify the authenticity of emails by checking the sender's email address and domain and verifying digital signatures.
- **Machine Learning:** Train machine learning algorithms on past phishing attempts to detect future ones by identifying patterns and anomalies in email behavior.Implementing these techniques can help to reduce the risk of phishing attacks, and keep sensitive information secure.
- **NLP (Natural Language Processing)** is a subfield of data science that deals with the interaction between computers and human languages. NLP techniques are used in various applications, including:
- **Sentiment Analysis:** Analyze the sentiment of text, such as customer reviews or social media posts, to decide the overall tone or opinion of the content.
- **Text Classification:** Classify text into different categories, such as spam/not spam,
- positive/negative sentiment, or relevant/irrelevant topics.

2.4 Data Science Objectives

Data science aims are specific goals related to the use of data and data analysis in a project. In the context of a cybersecurity project, some examples of data science aims might include:

●**Improve threat detection:** The goal of using data analysis to find patterns and anomalies that may show potential cyber threats.

●**Automated threat detection:** The goal of automating the detection of phishing emails.

●**Data visualization:** The goal of visualizing data to help security personnel understand complex security information and make informed decisions.

2.5 Machine Learning Algorithms

Classification Model: When you build a model to predict a certain class or category , you need a way to measure how accurate the predictions are .

- **Precision:** If the model avoids a lot of mistakes in predicting spams and hams, then the model has a high precision.
- **Recall:** If the model avoids a lot of mistakes in predicting spams as hams, then the model has a high recall.
- **F1:** Do you want the model to aim high in both precision and recall, where the model avoids as many mistakes as possible , doing a good job at correctly predicting both spams and hams but if the model has the ability to predict one class and sucks at predicting the other wouldn't it be misleading to look at precision or recall in isolation ?This is where F1 comes in It takes into account both precision and recall . A balance of the two is what F1 scores on. If the model does a good job at accurately predicting both spams and hams, then it will have a high F1 score. In our case we will focus on recall spam or no spam.

Logistic Regression: It is a classification algorithm in machine learning that uses one or more independent variables to decide an outcome. The outcome is measured with a dichotomous variable meaning it will have only two possible outcomes and, in our case, logistic Regression algorithm will detect if the email received is a spam or a ham.

Conclusion

Data Science has left a massive impact on cybersecurity in a short span of time. The amount of data received by each organization is increasing day by day. With the increase in the volume of data, the predictive capabilities of a data science model will also increase. It has become extensively important that the data science team and security team should work collectively at each stage of the process flow. No matter how small or big a company, data is essential to all of them, thus, protecting this data at any cost is key to every organization. The involvement of data science in cyber security has helped to reach a new level of security standards.

3. Data preparation

3.1 Data collection

1-We started with a large archive of email messages that had been installed on the

Spam Assassin database

2-The archive had numerous text files, each of which had the full text of a

single email message, including both the subject and the body of the message.

3-we wrote a Python script to parse these text files and extract the subject and body of

each message, then we organized the extracted information into two separate datasets one containing spam messages and the other containing legitimate (ham) messagesin CSV format.

4-We used “pandas” to read in the two datasets, and then concatenated them into a single

DataFrame , then we saved the concatenated DataFrame to a new CSV file.
In the end we got because of this process the following data set with 33 columns distributed as follows:

● two features: subject and body and 29 unnamed columns

● Label: spam or ham

This dataset has 16351 records.

3.1.1 Sentiment analysis dataset

This project involved analyzing Email content. By understanding the words involved in the mail body, we are going to predict whether an email has cyberbullying content or not, and if it is a cyberbullying email then predicting the nature of the cyberbullying into 6 Categories:

● Age

● Ethnicity

● Gender

● Religion

● Other Cyberbullying

The data has been balanced to have ~8000 of each class.

Trigger Warning These tweets either describe a bullying event or are the offense themselves, therefore explore it to the point where you feel comfortable.

The dataset we are going to use is: Cyberbullying Classification data from Kaggle: https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset?select=spam.csv

Phishing Site URLs | Kaggle

● Data has 549,346 entries.

● There are four columns.

● Label column is prediction col which has 4 categories

1. Benign: This refers to a URL that is not malicious and does not pose a threat to the user's security. These are typically legitimate websites that are safe to access.
2. Defacement: This refers to a URL that has been changed or vandalized by an attacker. The attacker may have altered the website's appearance or content, but the site may not necessarily have malware.
3. phishing: This refers to a URL that is designed to trick users into divulging sensitive information, such as usernames, passwords, or credit card numbers. These URLs often mimic legitimate websites, such as banking or email sites, and are used to steal personal information.
4. Malware: This refers to a URL that has malicious code designed to harm the user's computer or steal sensitive information. Malware URLs may be used to install viruses, spyware, or ransomware on the user's device.

There is no missing value in the dataset. https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset?select=spam.csv

Spam Collection is a set of tagged messages that have been collected for SMS Spam research. It has one set of SMS messages in English of 5,574 messages, tagged according to being ham (legitimate) or spam.

3.1.2 web scraping

Visualize internal links, it will show all redirect links. Scrape any website.

● First, set up the Chrome web driver so we can scrape dynamic web pages. Chrome web driver

● WebDriver tool used for automated testing of web apps across many browsers. It supplies capabilities for navigating web pages, user input and more.

Data augmentation used in our case is the process of augmenting the dataset we have with added data. This added data here is the internal links that redirect the user to the server. We used data augmentation in machine learning algorithms to improve their performance.

3.2 Data storage

Data import module

The aim of this module is to import data to enrich the network. The data in question is the four entities that compose an email basically, the source, the subject, the URL and the content. The module had to

be able to process CSV files.

The import module initially consists of four steps:

- Sourcing: We select the source file and the type of data.
- Mapping: We create the link with the database, which means that all the information in a column of the CSV file will be stored in a specific location in JSON format. The mapping is made to allow milling data in any form. The mapping makes the translation between a lambda format and the database architecture and ontology of our project.
- Preview: The data that will be imported into the email database is displayed in JSON format. This allows us to check that we have assigned the data to the right place.
- Conversion: Finally, if we are satisfied, we save the data in the database (This part has been separated from the module) Afterward, we improved this module, with the different files we were given to import. Each of these files had a particularity that had not been considered before.

3.2.2 Data schemas in NoSQL databases

Existing NoSQL solutions can be grouped into four main families:

- Key-value: The data is simply stored as a pair of keys and values. The best-known solutions are Redis, Riak, and Voldemort, created by LinkedIn. These engines offer simplified functionalities with less functional richness in terms of queries and excellent performance thanks to their simplified access model.
- •Column-oriented engines: This model implements a structure close to the table. Data is represented in rows and separated by columns, which are variable. Indeed, the number of columns can vary from one record to another. This model is more used for important volumes. There are Hbase and Casandra as solutions.
- •Document-oriented engines: This family is an extension of the key/value family by
- associating a key with a hierarchical document such as XML or JSON (JavaScript Object Notation). For this model, the most popular implementations are CouchDB (by Apache) and MongoDB.
- •Oriented Graph: This model of data representation is based on the theory of graphs. It is based on the notion of nodes and the relations and properties that are attached to them. This model eases the representation of the real world, which makes it suitable for processing data from social networks.

3.2.3 MongoDB database

MongoDB is a document-oriented database management system. Developed in C++ and distributed under the AGPL license (free license) since 2007 by 10gen (a company working on cloud computing tools), it is very admirably adapted to web applications. MongoDB has been adopted by several big names in IT, such as Foursquare, SAP, or even GitHub. With its ease of use and remarkable performance, MongoDB is one of the most widely used document-oriented databases.

Documents under MongoDB:MongoDB is a document database in which documents are grouped in the form of collections. The collections are the equivalent of SQL tables. It is possible to have each document in JSON format (MongoDB uses a more compact binary variant of JSON called BSON for its internal storage). Each document has a unique key to find it in the collection.

3.3 Realisation

3.3.1 spam detection

3.3.1.1 Natural Language Processing NLP:

Natural language processing refers to the technical branch of computer science, or artificial intelligence, that is concerned with providing computer systems with the ability to comprehend digital text and spoken remarks - in much the same way that human beings can accomplish.

this subfield of artificial intelligence (AI) combines computational linguistics, or the rule-based modeling of human languages, statistical modeling, machine-based learning, and deep learning benchmarks. Jointly, these advanced technologies enable computer systems to process human languages via the form of voice or text data. The desired outcome or purpose is to 'understand' the full significance of the respondent's messaging, alongside the speaker or writer's objective and belief.

NLP Application to Detect SPAM

Natural language processing-based solutions comprise language translation,, emotive or sentiment analysis and automatic text categorization .

Tokenization:

a key concept that allows models to better understand large blocks of text by breaking them up into smaller, more digestible pieces.

for a complex email text with many action words and phrases. We might think that it's better to split into sentences first and then process it sentence-by-sentence or further break the statements into words.

For this task we could use sent_tokenize from nltk:

```
def transform_text(text):
    text = text.lower()
    text = nltk.word_tokenize(text)
```

Figure 9 : Tokenization

Stop words and Punctuation:

Stop Words: A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.

We would not want these words to take up space in our database or taking up valuable processing time. For this, we can remove them easily, by storing a list of words that you consider stopping words. NLTK(Natural Language Toolkit) in python has a list of stop words stored in 16 different languages

```
import string
string.punctuation
```

```
'!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'
```

Figure 10 : String punctuation

```
for i in text:
    if i not in stopwords.words('english') and i not in string.punctuation:
        y.append(i)

text = y[:]
y.clear()
```

Figure 11 : Stop words elimination

Stemming

The process of removing affixes from a word so that we are left with the stem of that word is called stemming. For example, consider the words ‘love’, ‘loving’, and ‘loved’, all convert into the root word ‘love’ after stemming is implemented on them.

```
from nltk.stem.porter import PorterStemmer
ps = PorterStemmer()
ps.stem('loving')
```

```
'love'
```

Figure 12 : Stemming result

INPUT/OUTPUT:

The result after removing stop words and punctuation, tokenization and stemming will be as follows:

```
transform_text("I loved the YT lectures on Machine Learning . How about you? ")
```

```
'love yt lectur machin learn'
```

Figure 13 : NLP Application final result

3.1.1.2 Plotting Wordcloud

It's important to remember that while word clouds are useful for visualizing common words in a text or data set words with higher frequency of appearance will seem larger than other words.

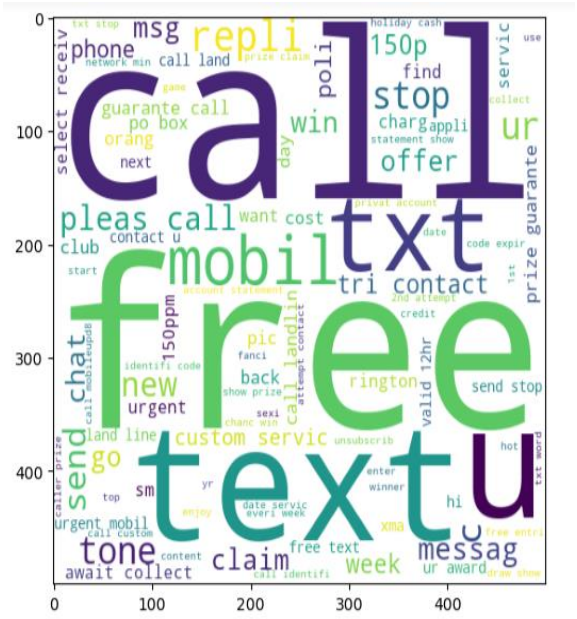


Figure 14 : SPAM wordcloud

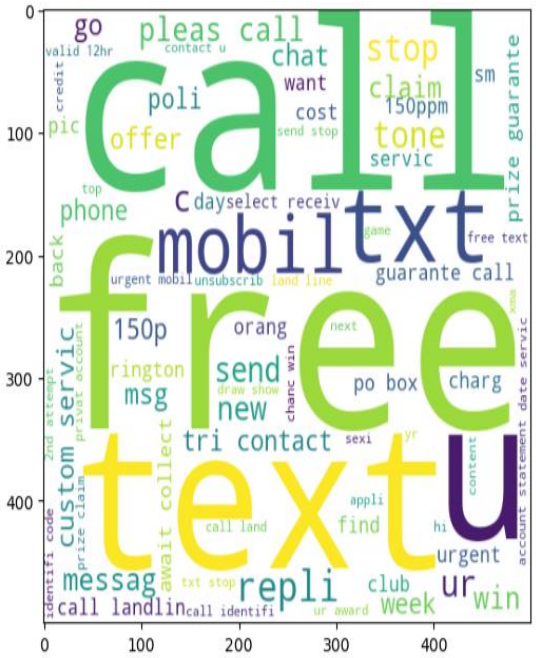


Figure 15 : HAM wordcloud

3.3.2 Model Building

```
from sklearn.feature_extraction.text import CountVectorizer,TfidfVectorizer
cv = CountVectorizer()
tfidf = TfidfVectorizer(max_features=3000)
```

```
X = tfidf.fit_transform(df['transformed_text']).toarray()
```

```
array([[0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       ...,
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.]])
```

Figure 16: Data vectorization

Train Test Split : 80% data for train , 20 % data for test.

3.3.1.1 Naive Bayes Algorithm

Naive Bayesian classifiers mostly used in text classification (due to better result in multi class problems and independence rule) have higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam email) and sentiment analysis (in social media analysis, to identify positive and negative customer sentiments)

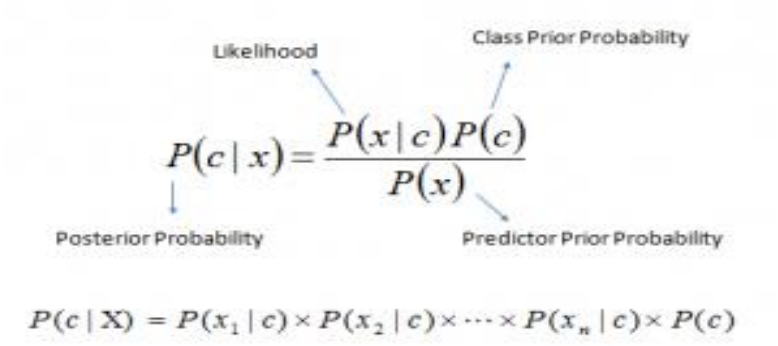


Figure 17 : Naive Bayes Formula

- **Gaussian Naive Bayes:** gaussiannb is used in classification tasks and it assumes that feature values follow a gaussian distribution.
- **Multinomial Naive Bayes:** It is used for discrete counts. For example, let's say, we have a text classification problem. Here we can consider Bernoulli trials which is one step further and instead of "word occurring in the document", we have "count how often word occurs in the document", you can think of it as "number of times outcome number it is observed over the n trials".
- **Bernoulli Naive Bayes:** The binomial model is useful if your feature vectors are boolean (i.e. zeros and ones). One application would be text classification with 'bag of words' model where the 1s & 0s are "word occurs in the document" and "word does not occur in the document" respectively.

```
0.8694390715667312
[[788 108]
 [ 27 111]]
0.5068493150684932
```

Figure 18 : GaussianNB metrics

- GuassianNB has an accuracy 0.86 and precision 0.50

```
0.9709864603481625
[[896   0]
 [ 30 108]]
1.0
```

Figure 19 : MultinomialNB metrics

- MultinomialNB has an accuracy 0.97 and precision 1.0

```
0.9835589941972921
[[895   1]
 [ 16 122]]
0.9918699186991869
```

Figure 20 : BernoulliNB metrics

- BernoulliNB has an accuracy 0.98 and precision 0.99

tfidf --> MNB ,we will work with multinomialNB because we have more precision and accuracy

3.3.1.2 Other classification models

```
svc = SVC(kernel='sigmoid', gamma=1.0)
knc = KNeighborsClassifier()
mnb = MultinomialNB()
dtc = DecisionTreeClassifier(max_depth=5)
lrc = LogisticRegression(solver='liblinear', penalty='l1')
rfc = RandomForestClassifier(n_estimators=50, random_state=2)
abc = AdaBoostClassifier(n_estimators=50, random_state=2)
bc = BaggingClassifier(n_estimators=50, random_state=2)
etc = ExtraTreesClassifier(n_estimators=50, random_state=2)
gbdt = GradientBoostingClassifier(n_estimators=50,random_state=2)
xgb = XGBClassifier(n_estimators=50,random_state=2)
```

Figure 21 : Classification models

	Algorithm	Accuracy	Precision
1	KN	0.905222	1.000000
2	NB	0.970986	1.000000
5	RF	0.974855	0.982759
0	SVC	0.975822	0.974790
8	ETC	0.974855	0.974576
4	LR	0.958414	0.970297
10	xgb	0.971954	0.943089
6	AdaBoost	0.960348	0.929204
9	GBDT	0.947776	0.920000
7	BgC	0.957447	0.867188
3	DT	0.932302	0.833333

Figure 22 : Classification evaluation metrics

- MultinomialNB , SVC , ETC are the best models for classification.
- Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output.

3.3.2 URL Study

3.3.2.1 Plotting Wordcloud

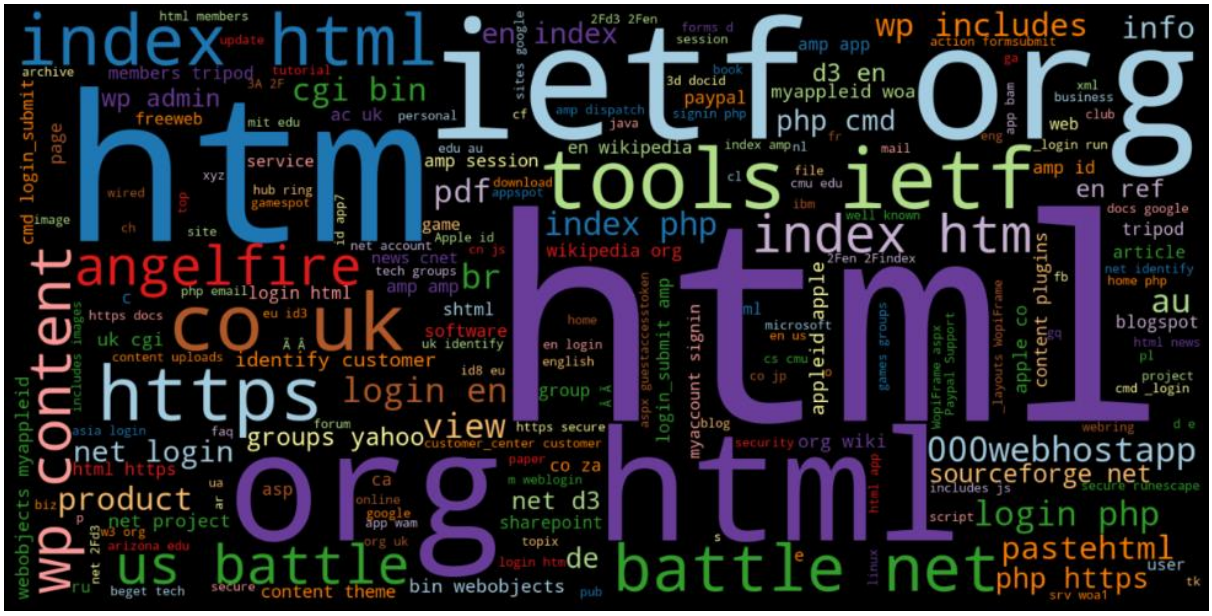


Figure 23 : Phish_url wordcloud

the reason for html https apperance that phishing urls try to appear like original websites.

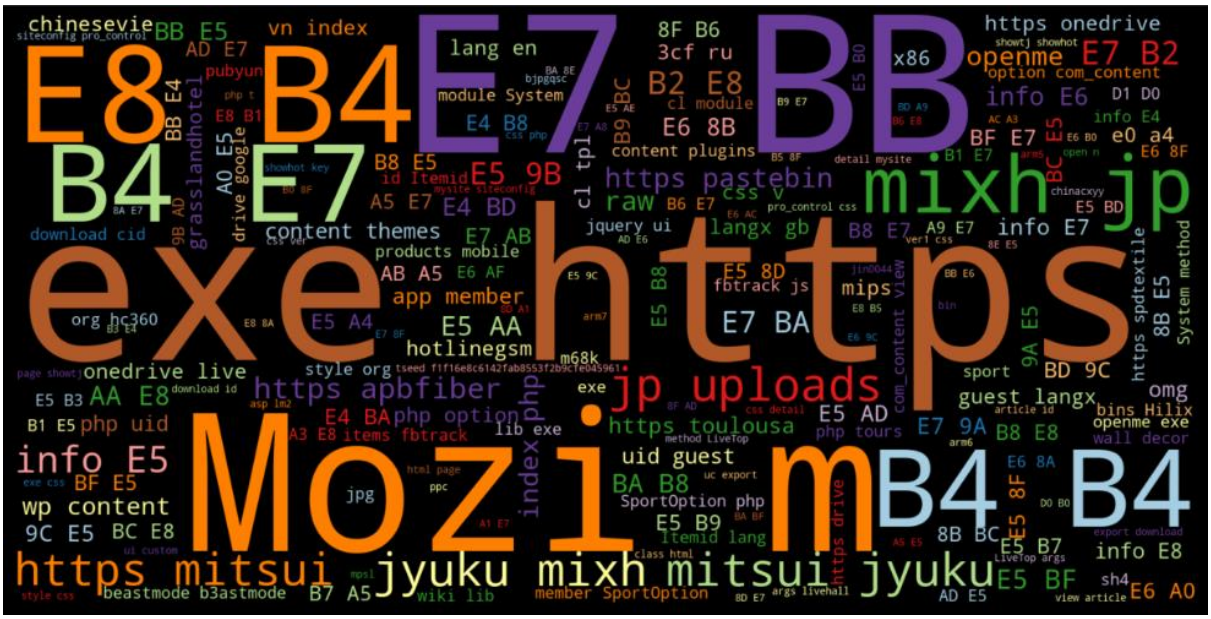


Figure 24 : Malware_url wordcloud

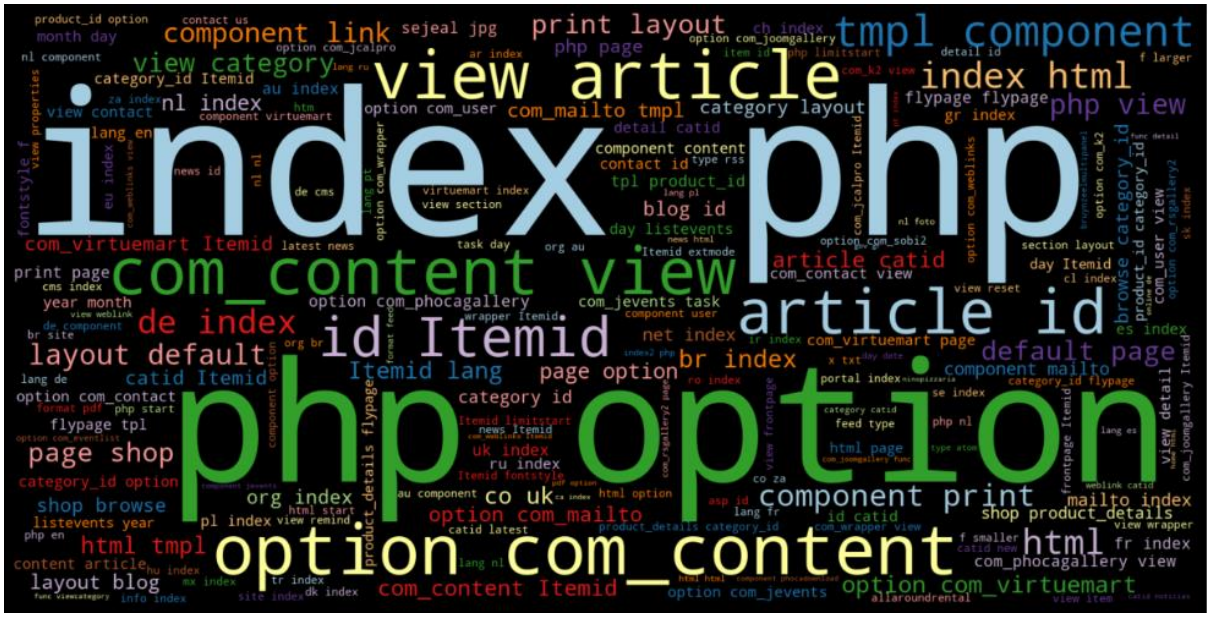


Figure 25 : Deface_url wordcloud

- defacement try to modify original websites through development techniques php option index

Figure 26 : Safe_url wordcloud

3.3.3 Feature Engineering

- create different lexical features from URLs to use them in machine learning.

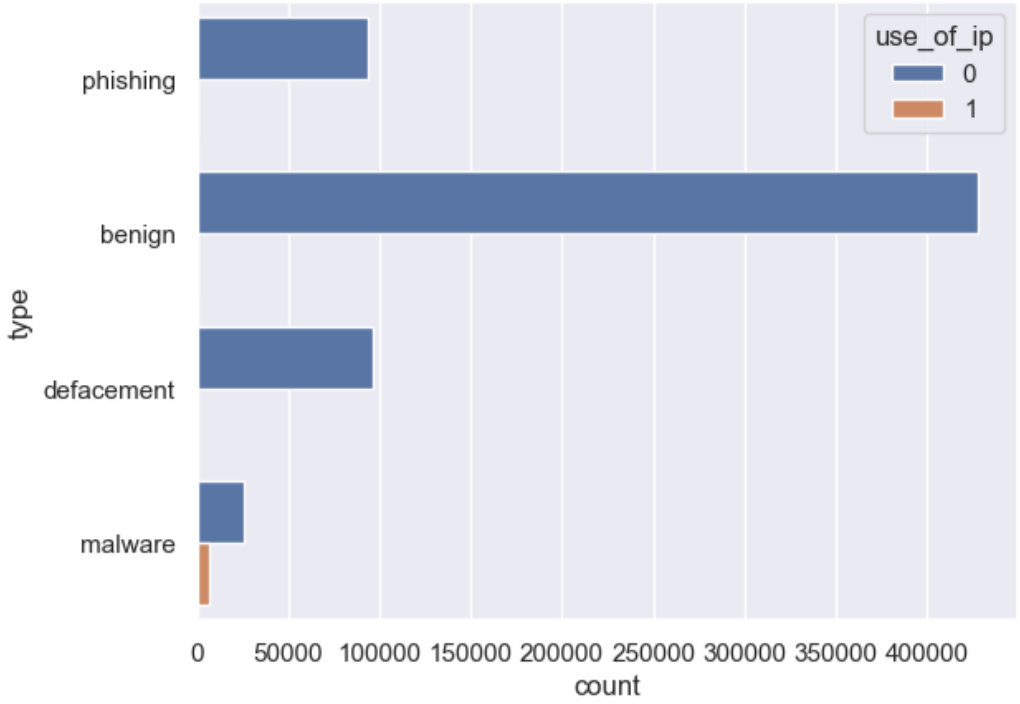


Figure 27 : Distribution of use_of_ip

Interpretation:

- only malware contains ip addresses which is a very good distinction.
- A malicious IP has been positively associated with malicious activity

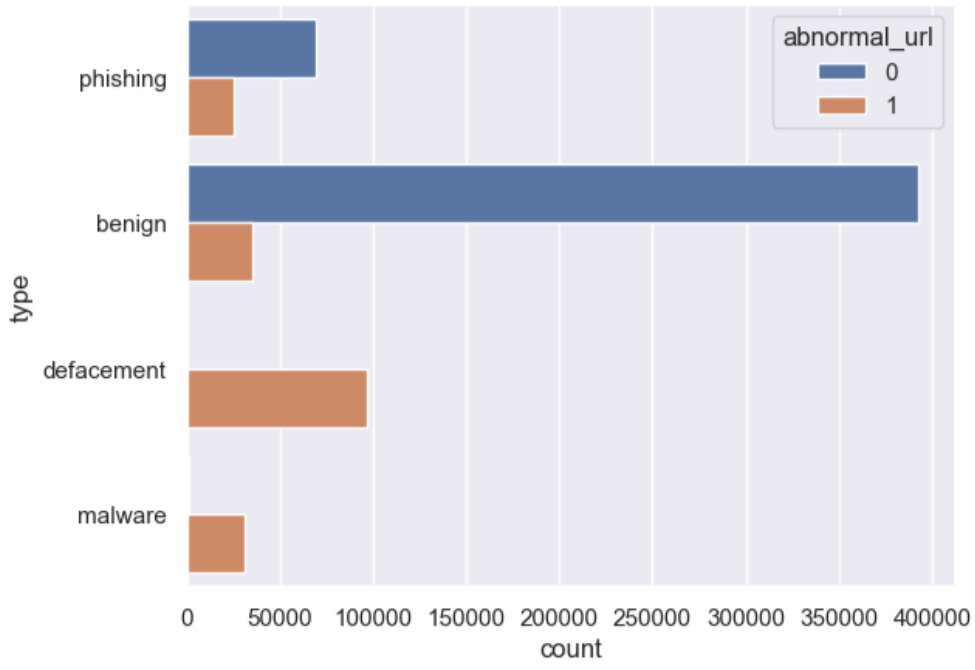


Figure 28 : Distribution of abnormal url

Interpretation:

- Defacement has the highest abnormal urls
- For the others similar significant abnormal urls

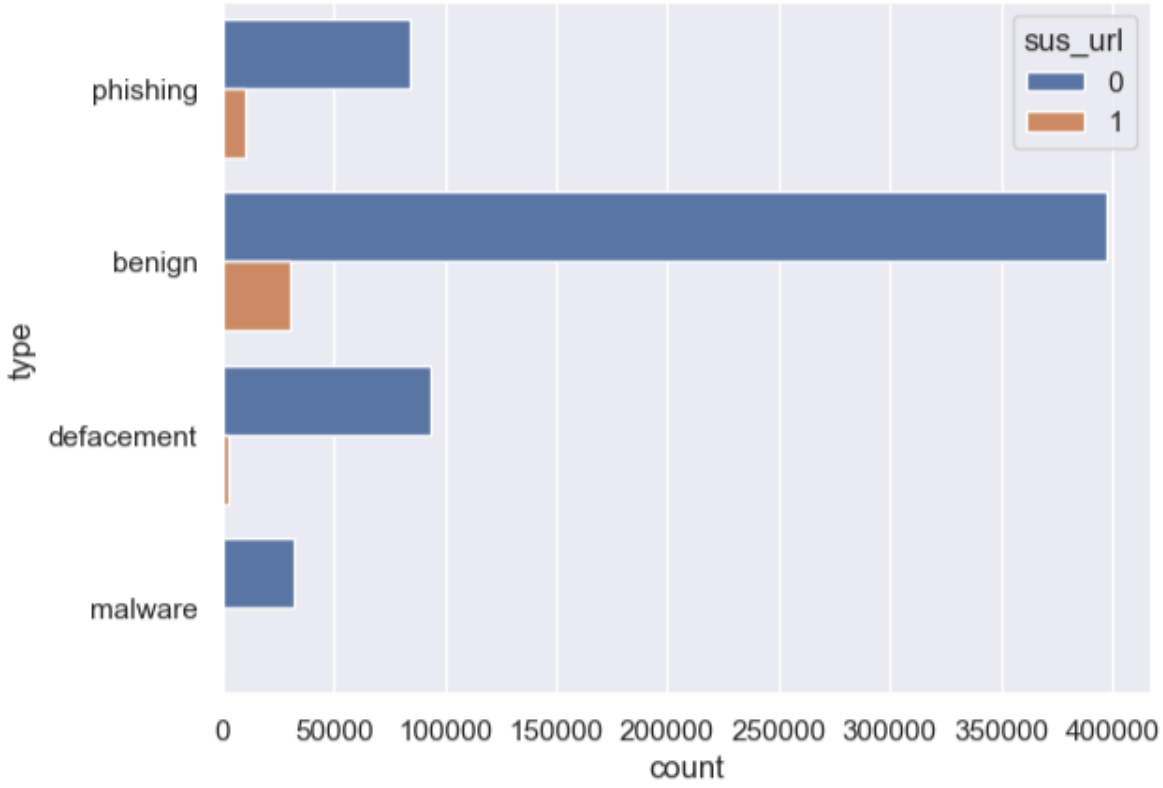


Figure 29 : Distribution of Suspicious URL

Interpretation

- This indicator may appear when a link contains a combination of characters that is considered unusual
- phishing urls used also keywords for transactions but the highest part for the safe urls because most of our transaction payment via email.

Train Test Split: 80% data for train, 20 % data for test.

3.3.3.1 Model Building

3.3.3.1.1 Random Forest Algorithm

The Random Forest is a powerful tool for classification problems, but as with many machine learning algorithms, it can take a little effort to understand exactly what is being predicted and what it means in context. Luckily, Scikit-Learn makes it pretty easy to run a Random Forest and interpret the results.

In this report we will walk through the process of training a straightforward Random Forest model and evaluating its performance using confusion matrices and classification reports.

	precision	recall	f1-score	support
benign	0.97	0.98	0.98	85621
defacement	0.98	0.99	0.99	19292
phishing	0.99	0.95	0.97	6504
malware	0.91	0.86	0.88	18822
accuracy			0.97	130239
macro avg	0.96	0.95	0.95	130239
weighted avg	0.97	0.97	0.97	130239

accuracy: 0.966

Figure 30 : Classification reports RFA

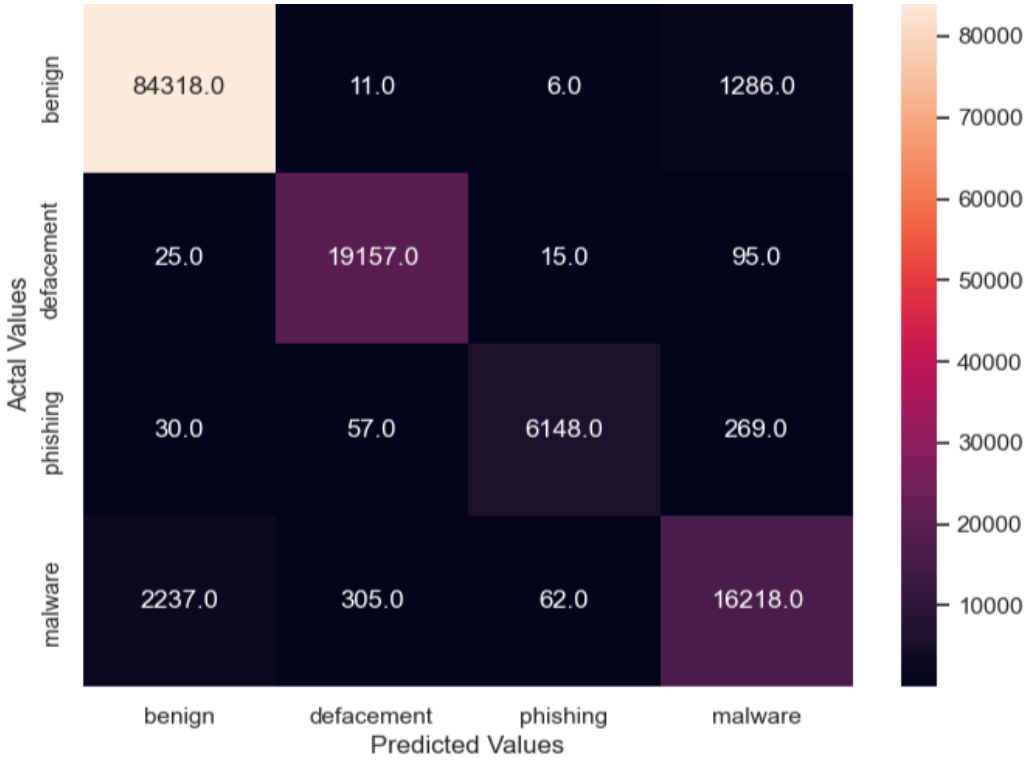


Figure 31 : Confusion matrix RFA

Random Forest Algorithm has an accuracy of 0,966. Then, it is a good model for classification.

3.3.3.1.2 Light GBM Classifier

LightGBM is a distributed and high-performance gradient boosting framework based on decision tree algorithms, used for ranking, classification, and many other machine learning tasks.

	precision	recall	f1-score	support
benign	0.97	0.99	0.98	85621
defacement	0.96	0.99	0.98	19292
phishing	0.97	0.90	0.93	6504
malware	0.90	0.83	0.86	18822
accuracy			0.96	130239
macro avg	0.95	0.93	0.94	130239
weighted avg	0.96	0.96	0.96	130239

accuracy: 0.959

Figure 32 : Classification report Light GBM

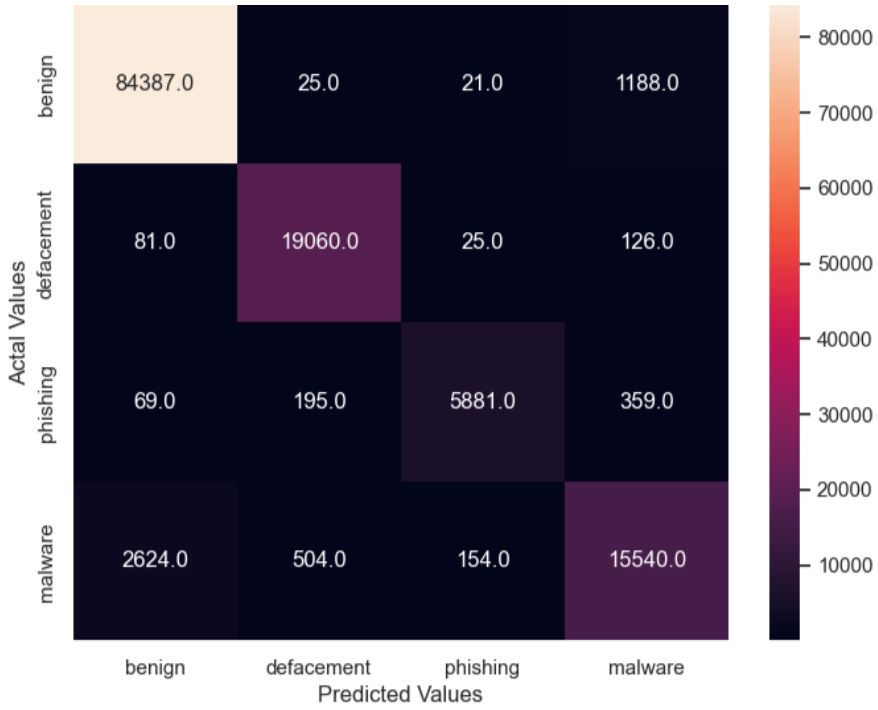


Figure 33 : Confusion matrix Light GBM

Light GBM model has an accuracy of 0,959. Then, it is a good model for classification.

3.3.3.1.3 XGboost Classifier

XGBoost is an optimized open-source software library that implements optimized distributed gradient boosting machine learning algorithms under the Gradient Boosting framework.

	precision	recall	f1-score	support
benign	0.97	0.99	0.98	85621
defacement	0.97	0.99	0.98	19292
phishing	0.97	0.92	0.94	6504
malware	0.91	0.83	0.87	18822
accuracy			0.96	130239
macro avg	0.96	0.93	0.94	130239
weighted avg	0.96	0.96	0.96	130239

accuracy: 0.962

Figure 34 : Classification report XGboost

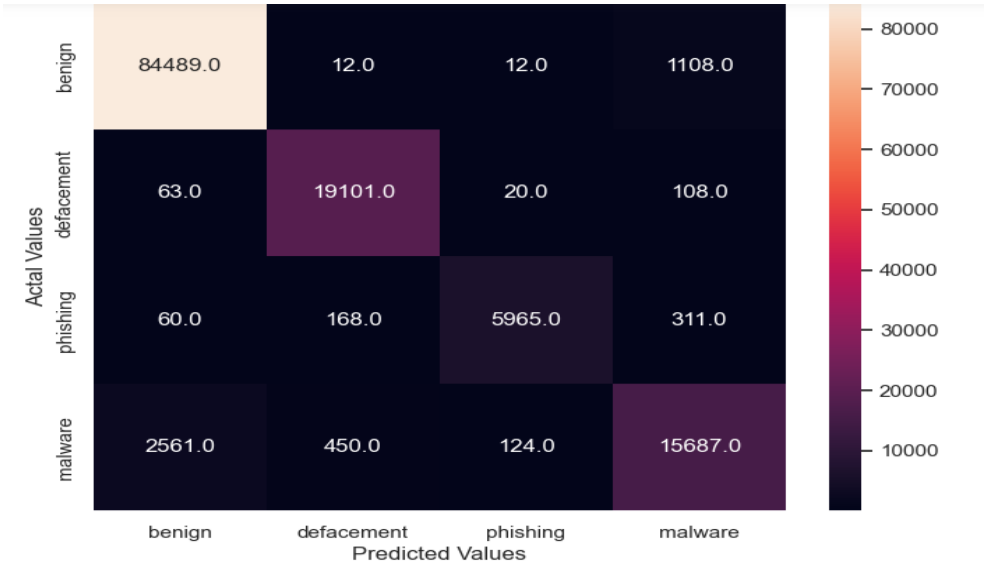


Figure 35 : Confusion matrix XGboost

XGboost model has an accuracy of 0,962. Then, it is a good model for classification.

Conclusion

A comparison of three learning models confirms that Random Forest Classifier is the best model for classification.

In conclusion, this project developed a solution for detecting phishing emails using deep learning, and machine learning algorithms. The solution provides a scalable and adaptable approach to detecting phishing emails and improves the accuracy and efficiency of the detection process. Our solution has significant implications for the field of cybersecurity and could be applied to a wide range of applications beyond email phishing detection.