Analysing Facebook User Risk using Machine Learning Algorithm

by

Arnab Barua
15301012
Fahim Adnan
15101023
Ananna Ghosh
19141020

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
April 2020

© 2020. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Arnab Barua
15301012

Fahim Adnan
15101023

Ananna

Ananna Ghosh
19141020

Approval

The thesis/project titled "Analysing Facebook User Risk using Machine Learning Algorithm" submitted by

1. Arnab Barua (15301012)
2. Fahim Adnan (15101023)
3. Ananna Ghosh (19141020)

Of Spring, 2020 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on April 7, 2020.

Examining Committee:

Supervisor:
(Member)

Hossain Arif

Hossain Arif
Assistant Professor
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Md. Saiful Islam

Md. Saiful Islam
Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:

(Member)

Rabiul

Dr. Md. Golam Rabiul Alam,PhD
Associate Professor
Department of Computer Science and Engineering
Brac University

Prof. Mahbub Majumdar
Chairperson
Dept. of Computer Science & Engineering
Brac University

Deputy Head of Department:
(Chair)

Sadia Hamid Kazi
Assistant Professor & Deputy Head, School of Sciences
Department of Computer Science and Engineering
Brac University

Ethics Statement

This research is a requirement for our Honours degree. It has been fully guided and supervised by BRAC University with responsibility for quality and integrity of the research. Hence the ethical consideration has been given full attention.
- The research paper is impartial and free of any plagiarism in primary and secondary level research.
- The informants have been asked for informed consent.
- The confidentiality and anonymity of the research respondents have been and will be respected.
- The participants have knowingly and voluntarily participated in the study.
- Total security (regarding the research work) to the informants or the participants or any relevant entity have been and will be ensured.

Abstract

Now-a-days people exchange their personal information and interact with companions and close relatives in a way which is revolutionized. In any case, the majority of them don't have the foggiest idea how to utilize, where to click, where not to, where to remark, and where not to. A considerable lot of them are posting in Facebook anything they desire and wish. This posting, fellowship and so on once in a while brings shocking occasions like identity theft, phishing, Cyber-wrongdoing and so on. So, Social media security has captured a great concern among the public and authority. At present, many features have been added to reduce the risk of hacking information. It is widely acknowledged that these features have played an important role in the security system. The essential focus point of our paper is on the safety implications of consumers posting their own Facebook information. We have made a survey containing 44 inquiries dependent on Facebook clients' propensity and different things. We have looked at the ongoing information security rupture on Facebook through certain data mining substances. We have targeted three questions about victim of malware, identity theft, and phishing. From, our dataset we will know how many were victim of the three target parameter. We have implemented machine learning algorithms like ANN, XGBoost, SVM, Random Forest, Decision Tree, Gaussian Naive Bayes, Logistic Regression to identify the percentage of how many Facebook accounts are in risk and safe. Moreover, we will compare the best possible approach and worst approach among the algorithms to find the result. Among the models, we see ANN providing us the best result for the three labels with 89.89%, 94.94% and 86.86%. This research illustrates how different machine learning algorithms predicts the risk of Facebook users and which algorithm is most and least suitable to use in this scenario.

Keywords: Identity theft, Phishing, Malware, Data Mining, Machine Learning, Algorithms, ANN, SVM, XGBoost.

Acknowledgement

Firstly, all praise to the Almighty for whom our thesis have been completed without any major interruption. As students of Computer Science and Engineering, we have got tremendous support and help from the Department of Computer Science and Engineering, BRAC University. We pay our respect and gratitude to our supervisor, Mr. Hossain Arif and co-supervisor Md. Saiful Islam for their intellectual support and advice in our work. Also, we are thankful to them who completed the survey and let us prepare the backbone of this research. Last but not the least, we are grateful to our parents for their kind support, for which we are now on the verge of our graduation.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

ACC	Accuracy
ANN	Artificial Neural Network
DT	Decision Tree
FN	False Negative
FP	False Positive
GNB	Gaussian Naive Bayes
ML	Machine Learning
RF	Random Forest
SVM	Support Vector Machine
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
XGB	XGBoost

Chapter 1

Introduction

1.1 Overview

Facebook is the brainchild of then Harvard undergrad Mark Zuckerberg, the social stage was established in 2004 [1]. The stage is a reconciliation of various destinations that was fundamentally engaged around schools, clients who wish to make a record to utilize their school uniform asset locator or URL to enlist on the stage. In the wake of making accounts in these social media, we get associated with our companions and family members alongside other obscure people. We share a few materials like content and pictures comprising of individual data with them. That is the reason associated people get our day by day exercises list creating a protection and security issue. Facebook is such a lot of mainstream online networking that it has 2 Billion dynamic clients every day, where YouTube and Instagram have individually 1.5 billion and 700 million month to month dynamic clients [2]. Facebook by and by increases its clients making it the greatest online social stage on the planet. Clients who make a profile on Facebook present their information to Facebook in the wake of consenting to its Acceptable Use Policy (AUP). The client data permits Facebook to coordinate the client with other profile dependent on the client's data. Moreover, such client information are additionally permitting Facebook to refresh the client's news source and proposals that associate with the client's profile information. To guarantee that the stage is free and meet its obligation commitments the client information is utilized by Facebook to tailor promotion to these clients. What makes Facebook vital to different substances is its abundance of client information that can be utilized to miniaturized scale target clients per their exercises and data on the site. Per its site, Facebook gathers client data, their exercises on and off the site, area, gadget area, equipment, programming, Connection data, for example, the name of your portable administrator or ISP, program type, cell phone number, IP address and so forth. In 2007, Facebook opened its foundation to outsider applications with the targets of expanding client experience and quantifiable profit. Furthermore, the presentation of outsider permits clients to mess around, play tests, shop, and use dating applications on Facebook. Nonetheless, to ensure client data, and conform to the Federal Trade Commission Act Section 5, called the "Unfair or Deceptive Act" [ICLG, n. d] [3]. Facebook guarantees clients acknowledge its term of utilization strategy before the client can get to its foundation. The opening up of Facebook to outsider applications has expanded the association's arrival on venture. In the expressions of Facebook author Mark Zuckerberg,"Until now, social networks have been closed platforms. Today, we're going to end that," [4].Be that as it may, the opening up of Facebook to outsider application presented another sort of security approach worry for Facebook. It rises our worry in regards to the measure of information access and benefit of these outsider applications has over client information. For the correct examination of these security concerns, we will analyze the association's security arrangements that identify with how outsiders get to or the degree of benefit an application has over client information and how clients are using these facilities. Facebook utilizes the Hyper Preprocessor (PHP) a server-side content which it uses to arrange the administrations it renders on its foundation [5]. Facebook in its initial days stored its information in a focal server, in any case, with the headway in cloud innovations, the firm has moved to distributed storage by putting away its information at various information ranches. The association's monstrous information foundation should consistently try to guarantee client fulfillment as of its foundation. According to Nosko, et al."Before it started building its own server farms, Facebook managed its infrastructure by leasing "wholesale" data center space from third-party landlords" [6].That is actually a major concerning security issue. Reason for this work is to manufacture some models to foresee if a Facebook client's record is in security peril or not, we look into and outline the outcome relying upon survey. The poll depends on close to home, security and protection, gadget and post and sharing data [7]. To foresee if a record is at serious risk or not, it is imperative to take their present settings they are utilizing, which is in this way accumulated by means of poll. Yet, a few inquiries can't be fitted in poll, in this manner we gather that data through observing. On the off chance that a record isn't at serious risk, at that point client is sufficient mindful of his record and that record should have been sincerer and follow enough protection and security mindfulness during the utilization of Facebook.

1.2 Motivation

Facebook is one of the highest used sites in the whole world with almost 2.50B monthly active users and almost 1.59B people use it daily from a mobile device[8]. Facebook is known to almost everyone and it consists of users valuable information including photos, address, email, phone number, family members, job, education etc. According to Wagner, k. [9], Facebook disabled 2.2B fake accounts or accounts that are impersonating someone real. Also, among other popular websites, Facebook is the top site that has been used for phishing globally. 48% of this attacks were web-based, whereas 25% were through mobile devices. Moreover, a malware, a software which is designed to damage computer, server, in this case a malicious software which steals information and takes control of one's device, is also common vulnerability Facebook users may face [10]. One of the common malware is, Facebook video virus, which spreads via messenger, can distribute spam messages to contact, promote fake messages containing malware. In short, Facebook users are most vulnerable to fake impersonation, phishing and malware, so, we decided to look up what traits are responsible for letting Facebook users become victim of such vulnerabilities. As, the data we collected consists of such victims, we felt the necessity to test out the predictability of various algorithms which can predict Facebook user's security. Supervised ML algorithms [11] are able to learn from example. Classification algorithms like SVM, DT, RF, XGB, ANN and regression model like LR are some of the most used supervised algorithms which yields acceptable prediction rate in such cases. Therefore, we felt the necessity to see how we can decrease the risk of Facebook users by analyzing the pattern we find from our dataset as we use these models to determine the analysis.

1.3 Aims and Objectives

The main objectives of this thesis:

- Train the collected data from Facebook users consisting of multiple classes and prepare the dataset for building train and test models.
- Implementation of various supervised algorithms like ANN, RF, DT, XGB etc. to predict the vulnerability of the Facebook user.
- To analyze the results of the various models we implemented and find the model with best accuracy to predict the risk of a Facebook user and show why the model worked best.

1.4 Thesis Overview

The remainder of this thesis work is organized as follows.

Chapter 1 shows the introduction of our thesis work. Our motivations behind this thesis, our aims and objectives and finally a short overview of the methodology we followed to complete this paper.

Chapter 2 presents the literature review where we show the background study we have completed for this thesis.

Chapter 3 presents the overall methodology and system workflow as well as methods of collection, visualization and processing of dataset.

Chapter 4 presents the implementation, result and analysis of various approaches.

Chapter 5 concludes the paper with the discussion of further improvement on this development.

Chapter 2

Literature Review

This Chapter consists of the literature review for our thesis and here we have illustrated the background study that we required to complete the thesis.

2.1 Overview

While doing our background studies on our research related works, we found out various papers on Facebook user security. In the first phase of our work, we needed to get acquainted with data mining techniques, by which we would be able to collect data from Facebook users[12] and utilize them in this work. Moreover, we have found many research papers on classification and detection of Alzheimer disease where they have used different datasets and different techniques. For example, Gabriela, et al. [2], in their exploration paper, they indicated how understudies utilize web based life website Facebook for day by day scholastic exercises to advance the learning methodology in a superior manner. Additionally, understudies lean toward Facebook in their preparation, examining and other instructive purposes. They found that students utilize Facebook less in scholastic exercises, rather they invested great measure of energy in Facebook for social purposes, for example, to share, tag, chipping in, remain associated with companions, family and family members. Vishwanath, Arun [3], in his proposition work, he centers around ongoing Facebook use with its determinants that prompting online networking phishing assaults. Next to propensities, attitudinal responsibility and mindfulness in security and protection has indicated better in close to home inspirations and cautious to utilization of web based life. Moubarak, et al. [4], researched on specialist understanding relationship dependent on Facebook action of people. As every individuals offer their data [13] in Facebook, understanding companion demand consequently dropped by the specialist. Additionally specialist understanding relationship would be modified if the patient has open access to their primary care physician's Facebook profile. Along these lines, absence of security assurance may make an effect on specialist understanding relationship. Nosko, et al. [6] in their examination work inspected revelation in Facebook profiles that some data are identified with exposure of other delicate individual data. Age and relationship status are two of significant components for example individual data in profiles decline as age expanded and concerning danger of risk for the individuals who are looking for a relationship. Bilge, et al. [14], states that internet based life has billions of enlisted clients, effectively utilizing these stages to share, to keep contact with loved ones, business contacts, instructive purposes and so on. They indicated that genuine clients offer worry to the potential aggressor's fraud assaults so as to get to the huge measure of enrolled clients data. The procedure is by setting up a kinship with focused casualty. For example, sharing freely and keeping contact with obscure individual lead to security and protection worry as utilizing these computerized cross-site profile cloning assault may happen. Govani, et al. [15], in his theory article he clarified, if we assume that the greater part of the Facebook clients knew about the potential results of individual data given openly. Albeit the greater part of the Facebook clients know, feel practical to give data there. Disregarding realizing that they can exact in security settings with respect to perspective on close to home data, they don't step up to the plate and ensure their data. Unique in relation to others we utilized distinctive machine learning methods [16] and way to deal with order the client in regard to their security estimation with the assistance of the information gathered from them and by watching them and also, removed significant elements for being a casualty of different digital assault or digital wrongdoing. On the other hand, in this paper we will analyze the risk of security in Facebook by using Naive Bayes, Support Vector Machine, Decision Tree, Random Forest, ANN, Logistic Regression and XgBoost.

Chapter 3

Workflow and Dataset Description

3.1 Workflow

This following chapter gives a brief idea on the workflow we will follow. From the figure 3.1 we have described our working process from beginning to end. At first we gathered our data doing field survey asking multiple questions about how do people interact with online social media and social networking service Facebook. After fetching data we have decided that we will work on three target features and rest of the features will be treated as input features.[17] Here comes our data pre-processing part. As our data is purely categorical,we have to transform the data into numerical data. Later on, we split our gathered data in train data and test data. We moved to our next part which is preparing our self-designed ANN model and other models based on different machine learning algorithms for training. After that we checked results like accuracy and confusion matrix [18] for different models[19].

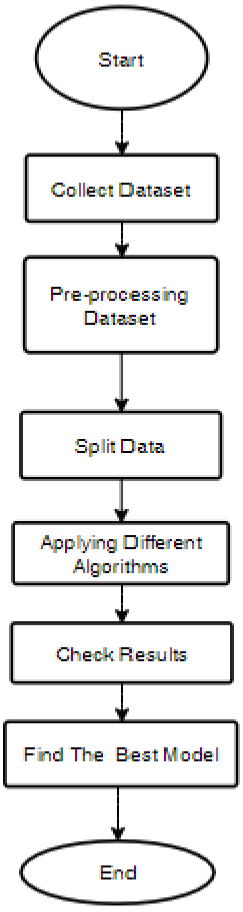


Figure 3.1: Proposed System.

3.2 Dataset Description

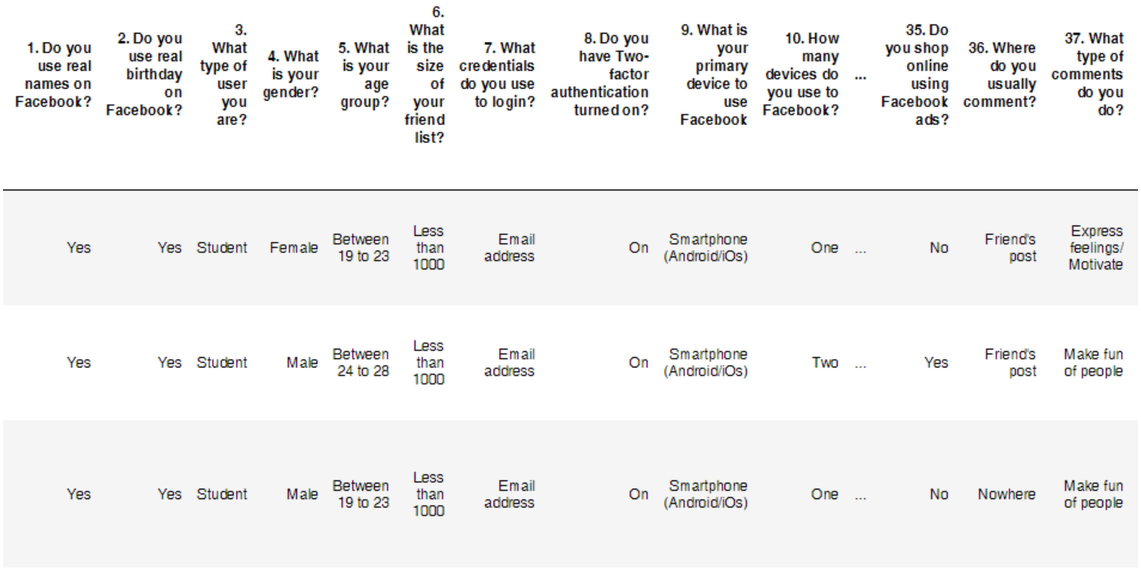
3.2.1 Summary

In our research we gathered information about how our participants interact with a prominent social media Facebook for data analysis and distribution.This dataset contains answers of various topics like use of authentic identity, gender, age, default photo album privacy of 989 subjects. These subjects are aged from 19 to 50. For every subject, information were acquired in single session. The subjects include both men and women. Most of the subjects' age range from 19 to 23. In addition,for diversity, a reliability data set is included that contains middle aged people.

3.2.2 Additional data

Data is stored in a spreadsheet file(Facebook Risk).

Target Features: Our target features corresponds to :
1. Have you ever clicked on unknown links in Facebook and been a victim of malware/virus?
2. Have you seen such people on Facebook who might be impersonating you/uses your personal information/photos etc.?
3. Phishing is a cyber-attack which uses disguised email as a weapon. Have you ever been a victim of Phishing or clicked on link that requests you information/credentials? **Demographics:** Gender (Male/Female), Age.



1. Do you use real names on Facebook?	2. Do you use real birthday on Facebook?	3. What type of user you are?	4. What is your gender?	5. What is your age group?	6. What is the size of your friend list?	7. What credentials do you use to login?	8. Do you have Two-factor authentication turned on?	9. What is your primary device to use Facebook	10. How many devices do you use to Facebook?	...	35. Do you shop online using Facebook ads?	36. Where do you usually comment?	37. What type of comments do you do?
Yes	Yes	Student	Female	Between 19 to 23	Less than 1000	Email address	On	Smartphone (Android/iOs)	One	...	No	Friend's post	Express feelings/ Motivate
Yes	Yes	Student	Male	Between 24 to 28	Less than 1000	Email address	On	Smartphone (Android/iOs)	Two	...	Yes	Friend's post	Make fun of people
Yes	Yes	Student	Male	Between 19 to 23	Less than 1000	Email address	On	Smartphone (Android/iOs)	One	...	No	Nowhere	Make fun of people

Figure 3.2: Dataset Description

3.3 Data Pre-processing

In the data pre-processing part, we have done level encoding and oversampling.

3.3.1 Level Encoding using Scikit-learn(Python)

As we are dealing with categorical variables here, so we have to convert it into numerical variables before we try to plug these variables into machine learning models in Python without pre-processing them first .A categorical variable takes only a limited number of values[20]. For example when we surveyed about the usage of authentic identity and provided two options:"Yes" or "No". In this case, the data is categorical because responses fall into a fixed set of categories. There are mainly two approaches to handle this issue. We can either use One-Hot Encoding or Label Encoding for the purpose. For our research we used Label Encoding[21]. Label encoding can assign different integer to each unique value. This approach assumes an ordering of the categories. For tree-based models like decision trees and random forests, we can expect label encoding to work well with categorical variables.

3.3.2 Over-sampling

To begin, the very first problem we faced is an imbalanced dataset. If so, we can assume that real data are almost balanced however, there were some proportions of bias (mostly due to the collecting method) in the collected data. So, in this case, the best approach is to collect data which are more representative [22]. Three methods are often presented as great ways to balance the dataset before fitting a classifier on it [23]. In short, these mentioned methods act on the given dataset as follows:

1. Under-sampling: Under-sampling consists of sampling from the majority class to keep only a part of these points.

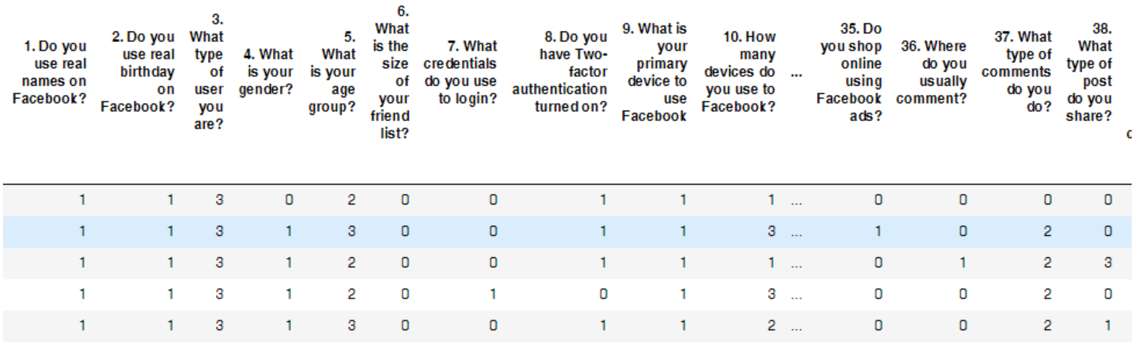
2. Oversampling: Oversampling consists of replicating some points from the minority class to increase its cardinality.

3. Generating synthetic data: Generating synthetic data comprises in creating new artificial points from the minority class (SMOTE method) to increase its cardinality.

So we employed oversampling that is often used to tackle imbalanced classes and that deal with the dataset itself.We avoided under-sampling because of the risks related to under-sampling.

3.3.3 Output

This is the picture after data pre-processing.



1. Do you use real names on Facebook?	2. Do you use real birthday on Facebook?	3. What type of user you are?	4. What is your gender?	5. What is your age group?	6. What is the size of your friend list?	7. What credentials do you use to login?	8. Do you have Two-factor authentication turned on?	9. What is your primary device to use Facebook	10. How many devices do you use to Facebook?	...	35. Do you shop online using Facebook ads?	36. Where do you usually comment?	37. What type of comments do you do?	38. What type of post do you share?
1	1	3	0	2	0	0	1	1	1	...	0	0	0	0
1	1	3	1	3	0	0	1	1	3	...	1	0	2	0
1	1	3	1	2	0	0	1	1	1	...	0	1	2	3
1	1	3	1	2	0	1	0	1	3	...	0	0	2	0
1	1	3	1	3	0	0	1	1	2	...	0	0	2	1

Figure 3.3: Dataset Pre-processing

3.4 Data Visualization

Data visualizations are a great way to understand data distribution.

3.4.1 Visualization of first target

We explored data distribution of 'Have you ever clicked on unknown links in Facebook and been a victim of malware/virus?'

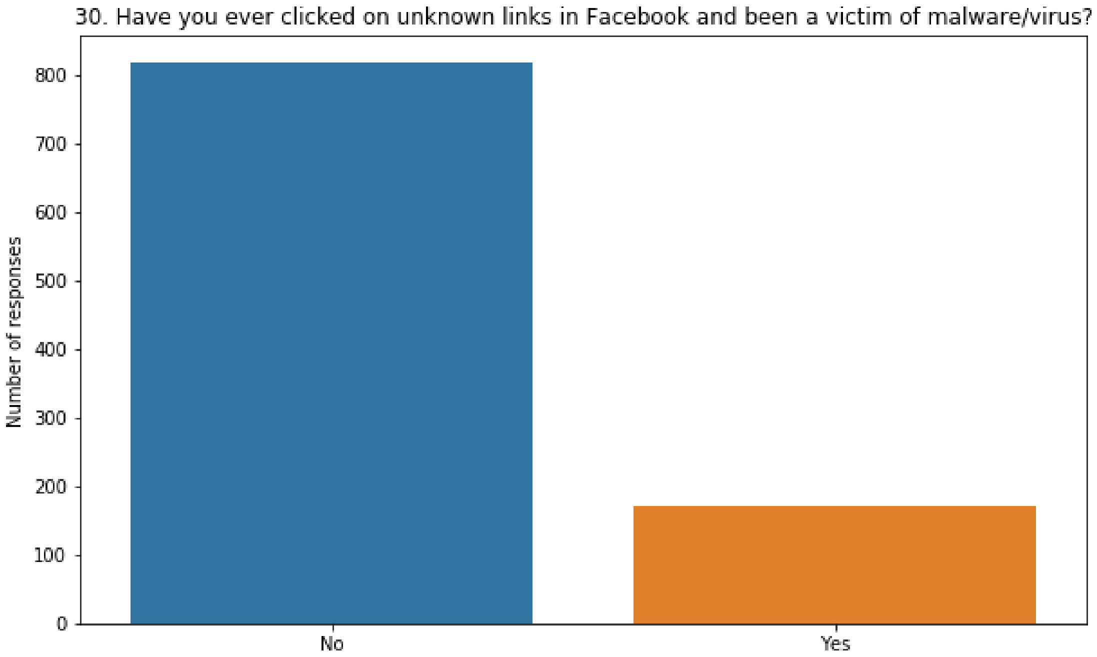


Figure 3.4: Dataset Visualization

3.4.2 Visualization of second target

We explored data distribution of ’ Have you seen such people on Facebook who might be impersonating you/uses your personal information/photos etc.?’

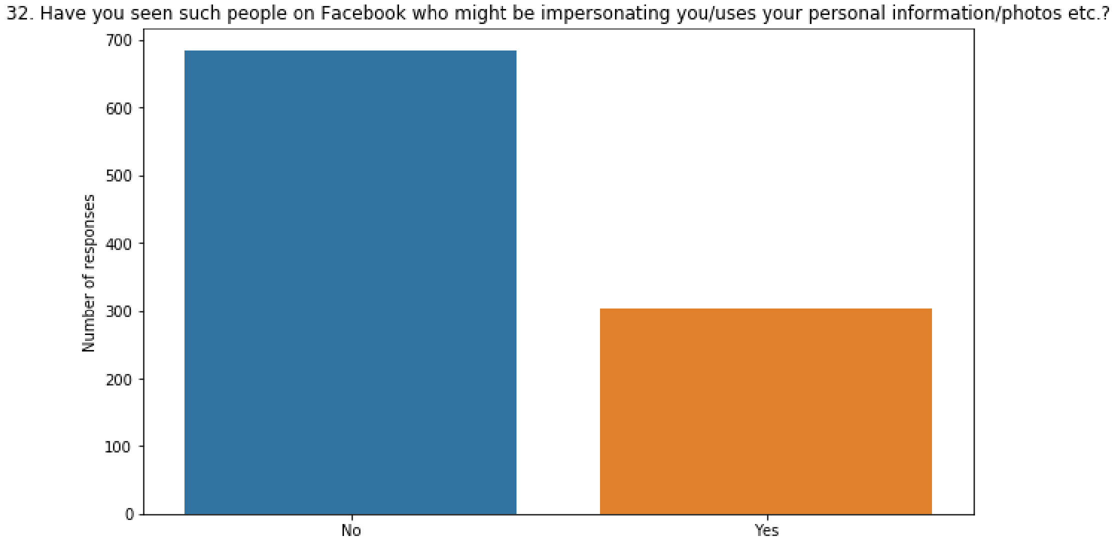


Figure 3.5: Dataset Visualization

3.4.3 Visualization of third target

We explored data distribution of ' Have you seen such people on Facebook who might be impersonating you/uses your personal information/photos etc.?'

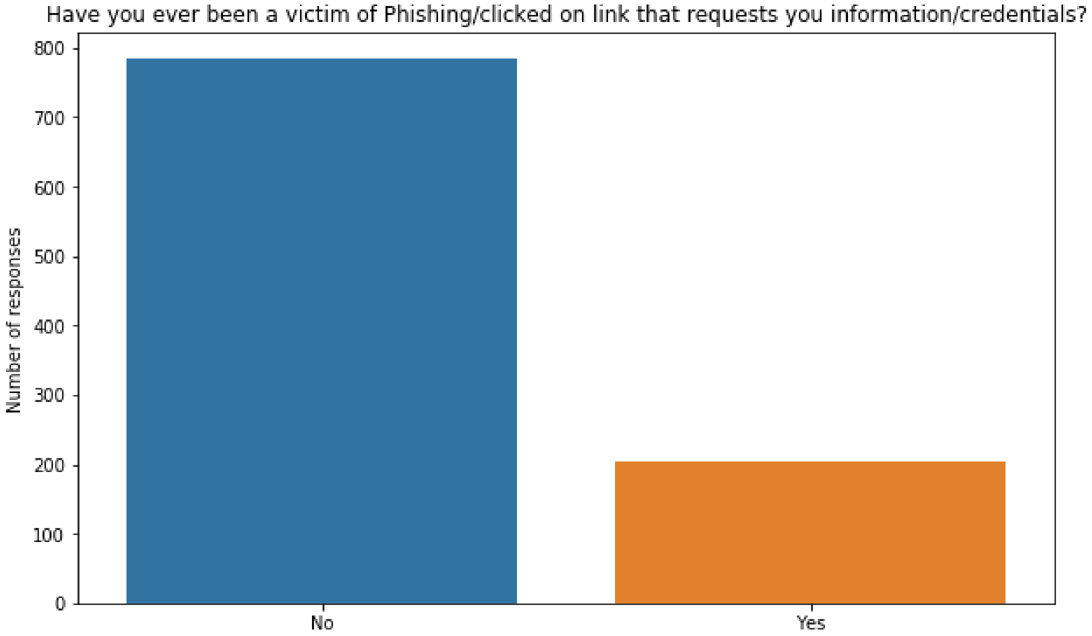


Figure 3.6: Dataset Visualization

Chapter 4

Implementation and Result

This chapter provides a complete overview of the models that we have implemented and the outcome of our thesis.

4.1 Applying proposed models

By pre-processing the data, we can apply our self-designed ANN model and other machine learning models to calculate the accuracy, confusion matrix , model accuracy and model loss graph.

4.1.1 Accuracy equation

Accuracy equation:

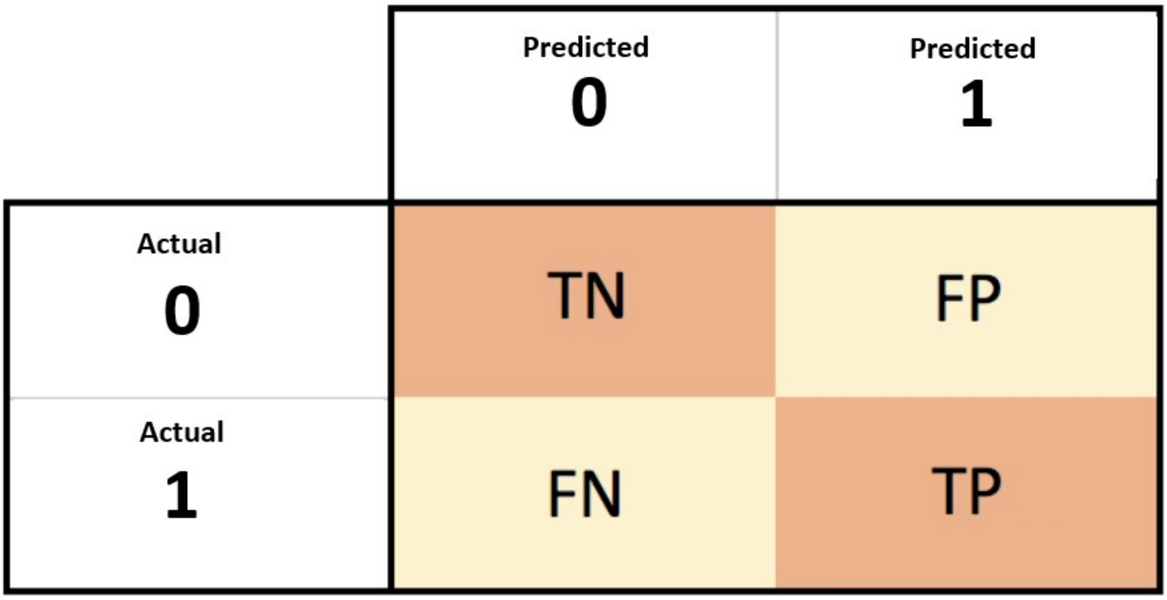
$$\begin{aligned} ACC &= \frac{TP+TN}{P+N} \\ &= \frac{TP+TN}{TP+TN+FP+FN} \end{aligned} \tag{4.1}$$

Where-

- ACC = Accuracy
- TP = True Positive
- TN = True Negative
- P = Condition Positive
- N = Condition Negative
- FP = False Positive
- FN = False Negative

4.1.2 Confusion Matrix

Confusion matrix is a tool to compute the performance of ML classification problems, especially when there are multiple target parameters. Confusion matrix show results of M-class classification problems, as confusion matrix can determine the similarities and variance between classifier outputs and the true ones outputs. In our research, we demonstrate confusion matrix of each algorithm's prediction to determine the effectiveness and performance of the algorithm, which then can be used to select the best algorithm. Moreover, confusion matrix shows the error that are made by the classifier and what types of error are those. Where-



	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Figure 4.1: Confusion matrix

- TP = True Positive(Correctly detected)
- TN = True Negative(Correctly rejected)
- FP = False Positive(Incorrectly detected.)
- FN = False Negative(Incorrectly rejected)

4.2 Self designed ANN architecture

4.2.1 Self designed ANN architecture Concept

Neurons

Like other neural networks, ANNs are made out of "artifical" neurons. With activation state this neurons can get the input and produces output. Primary datas for ANNs are most likely to be image, documents etc[24]. A conclusive outputs accomplish the task, for instance, seeing an article in a picture. The significant property of the activation work is that it gives a smooth, differential progress as input esteems change, for instance for a slight change in input, it creates a little change in output.

Weights and Connections

Like other neural networks ANN contains the output of one neuron was a result of another neuron's input. The neurons are unique in the data and connection they contain. All the connections are assigned a particular weight with own significance.

Activation Function

In our model we have used two different activation functions.

1. Sigmoid or Logistic Function
2. ReLu Function

Sigmoid Function: The mathematical term for sigmoid function is

$$f(x) = \frac{1}{1 + exp(-x)} \tag{4.2}$$

Its range is somewhere in the range of 0 and 1 and it is a s-shaped curve. Applying this activation work is nearly easy yet it has vanishing gradient issue and its yield is not zeros focused as a result the gradient updates go excessively far various way which makes streamlining more earnestly.

Relu Function: The mathematical function of ReLu is-

$$ReLu = max(0, x) \tag{4.3}$$

Utilizing this capacity is extremely straightforward and effective and it maintains a strategic distance from the vanishing gradient issue. In light of this explanation we have utilized this activation work for our models. Yet, it has an issue, it very well may be just applied in the hidden layers yet contrasting and the other two activation functions it appears to be better.
In the model of ours, we have utilized relu work with each hidden layer. We have also utilized the sigmoid capacity with final layer.

4.2.2 Result of our self-designed ANN architecture

After our model was ready, we calculated the following things from our data:

Accuracy

We have listed our accuracy for our targets features in the table below.

Table 4.1: Accuracy for ANN

	First Label	**Second Label**	**Third Label**
Accuracy	89.89	94.94	86.86

Confusion Matrix for First Target Feature:

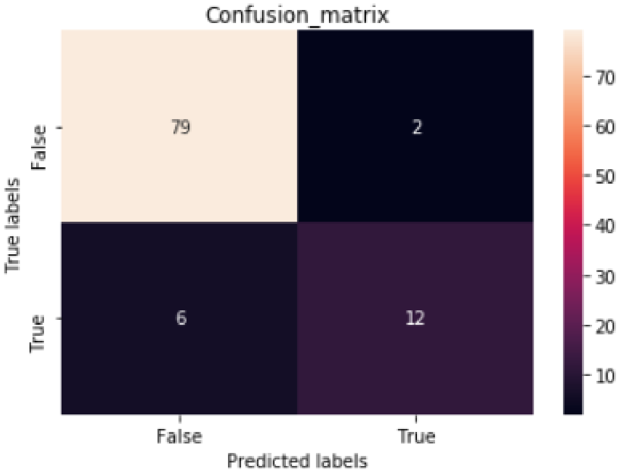


Figure 4.2: Confusion matrix interpretation of ANN architecture.

Confusion Matrix for Second Target Feature:

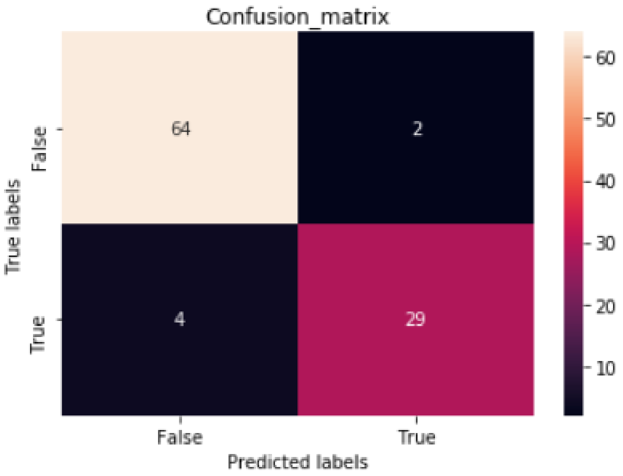


Figure 4.3: Confusion matrix interpretation of ANN architecture.

Confusion Matrix for Third Target Feature:

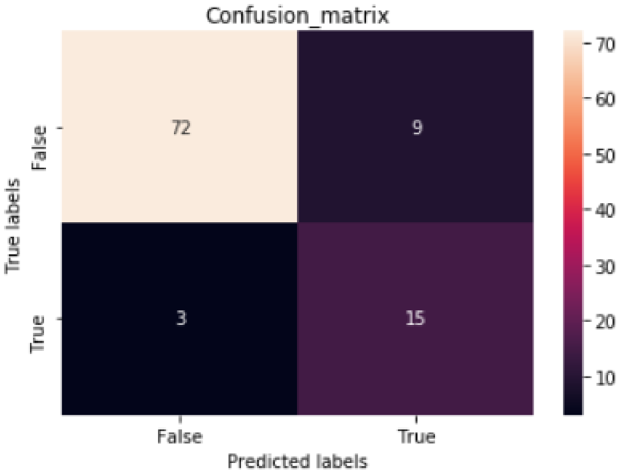


Figure 4.4: Confusion matrix interpretation of ANN architecture.

If we interpret this confusion matrix,overall performance of the model is very good we can say from this performance analysis.

Model accuracy and model loss:

Model accuracy and model loss for First target:

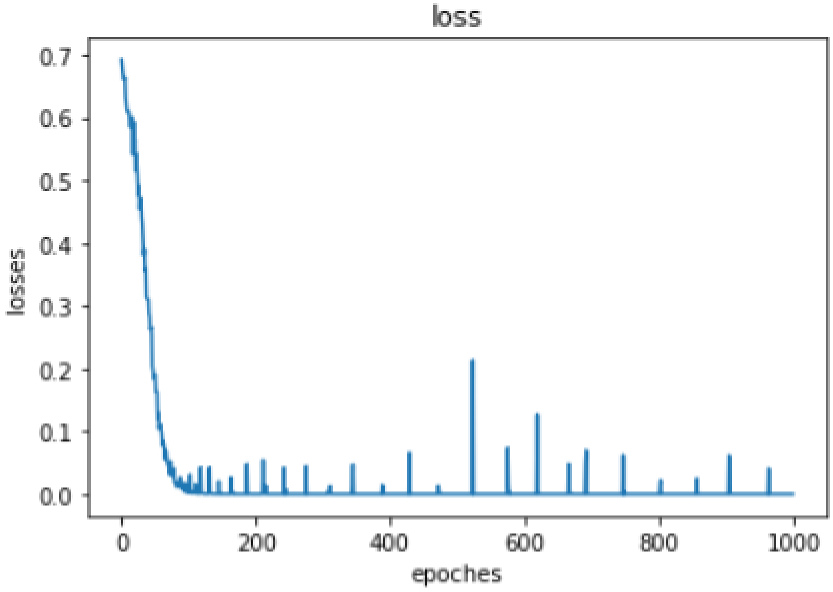


Figure 4.5: Model loss for first target

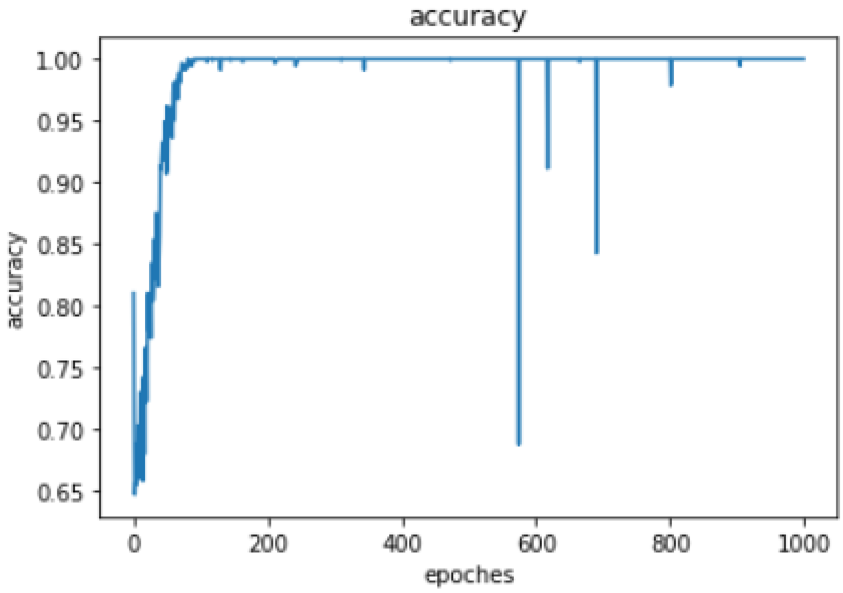


Figure 4.6: Model accuracy for first target

Model accuracy and model loss for Second target:

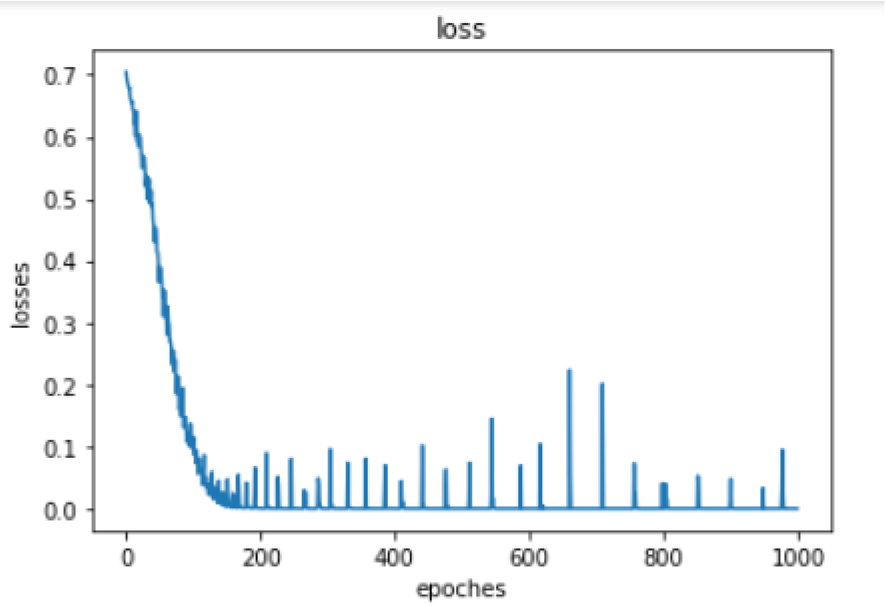


Figure 4.7: Model loss for second target

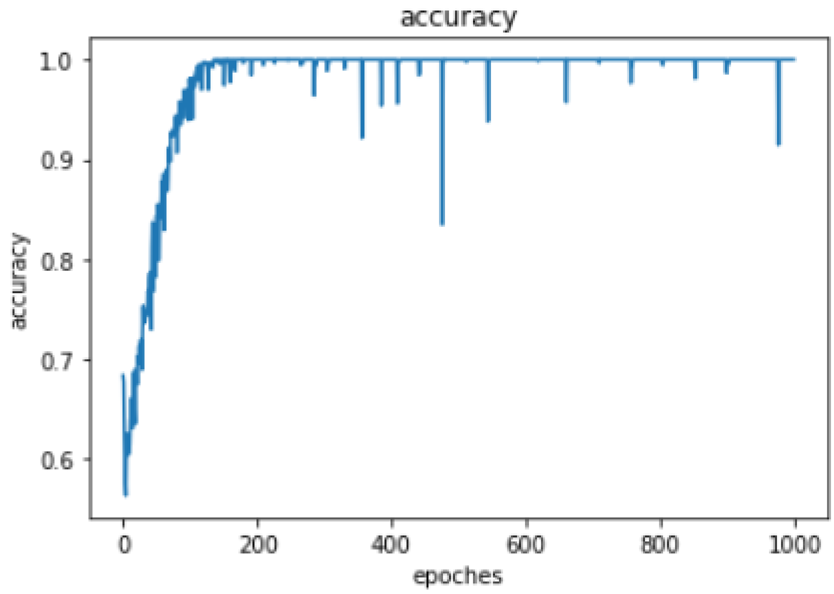


Figure 4.8: Model accuracy for second target

Model accuracy and model loss for Third target:

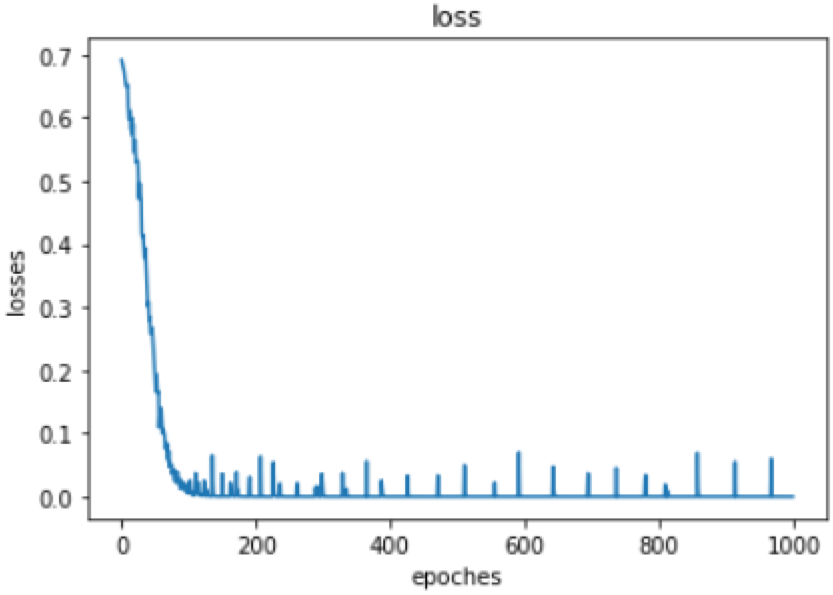


Figure 4.9: Model loss for third target

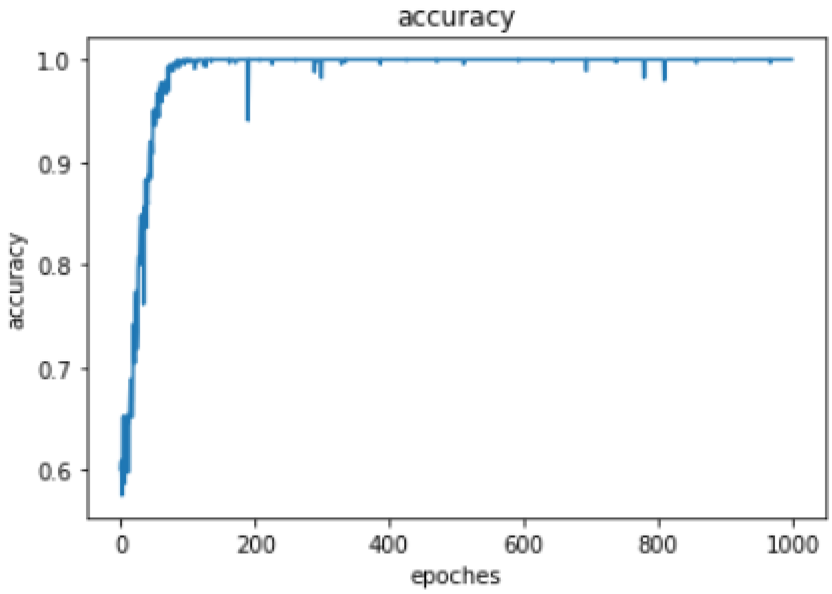


Figure 4.10: Model accuracy for third target

4.3 Decision-Tree Classifier

4.3.1 Decision-Tree Classifier Concept

Decision Tree is a significant Supervised ML algorithm for classifying. It can split data continuously in regards to specific specification. It is used universally for ML, statistics, data mining etc.[22]. It utilizes a structure similar to tree, where the leaves and branches represents class label and output respectively[25]. Decision trees are able to yield better results with minimum computation power. Moreover, DTs can work easily with continuous and categorical data as well [26].

4.3.2 Result of our Decision-Tree Classifier

After our model was ready, we calculated the following things from our data:

Accuracy

We have listed our accuracy for our targets features in the table below.

Table 4.2: Accuracy for Decision-Tree Classifier

	First Label	**Second Label**	**Third Label**
Accuracy	78.78	57.57	67.67

Confusion Matrix Interpretation:

Confusion Matrix for First Target Feature:

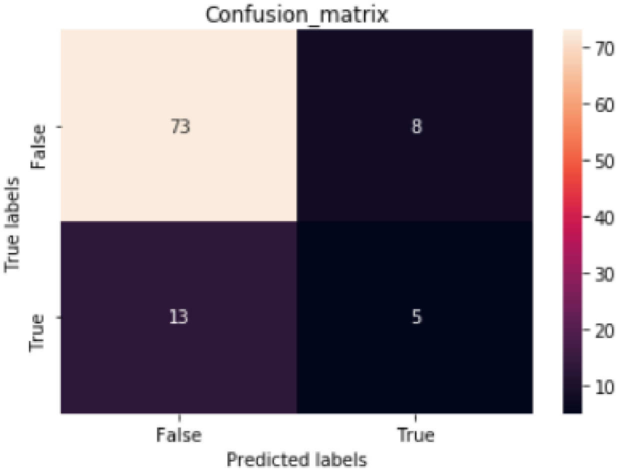


Figure 4.11: Confusion matrix interpretation of Decision-Tree Classifier.

Confusion Matrix for Second Target Feature:

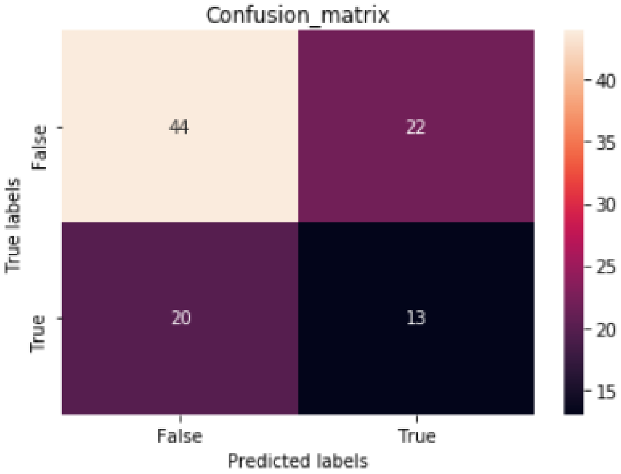


Figure 4.12: Confusion matrix interpretation of Decision-Tree Classifier

Confusion Matrix for Third Target Feature:

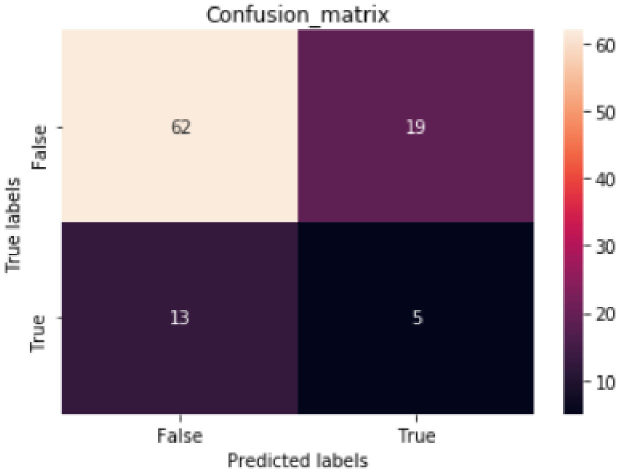


Figure 4.13: Confusion matrix interpretation of Decision-Tree Classifier

4.4 Gaussian Naive Bayes

4.4.1 Gaussian Naive Bayes Concept

Naive Bayes can be loosened up to real-valued attributes, most normally by anticipating a Gaussian distribution.
This expansion or procedure of Naive Bayes is known as Gaussian Naive Bayes[27]. Various functions can be used to evaluate the distribution of the information, anyway the Gaussian (or Normal distribution) is the most direct to work with considering the way that one simply need to assess the mean and the standard deviation from training information [28].

4.4.2 Result of our Gaussian Naive Bayes

After our model was ready, we calculated the following things from our data:

Accuracy

We have listed our accuracy for our targets features in the table below.

Table 4.3: Accuracy for Gaussian Naive Bayes

	First Label	**Second Label**	**Third Label**
Accuracy	75.75	64.64	75.75

Confusion Matrix Interpretation:

Confusion Matrix for First Target Feature:

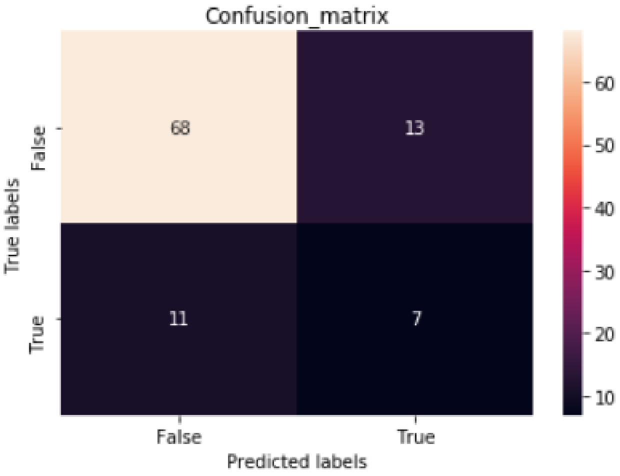


Figure 4.14: Confusion matrix interpretation of Gaussian Naive Bayes

Confusion Matrix for Second Target Feature:

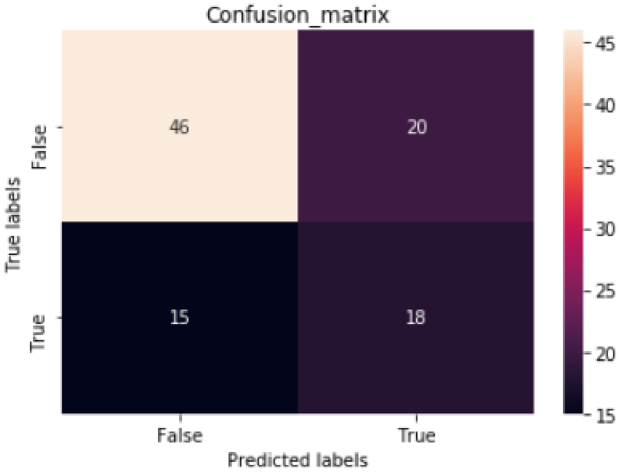


Figure 4.15: Confusion matrix interpretation of Gaussian Naive Bayes

Confusion Matrix for Third Target Feature:

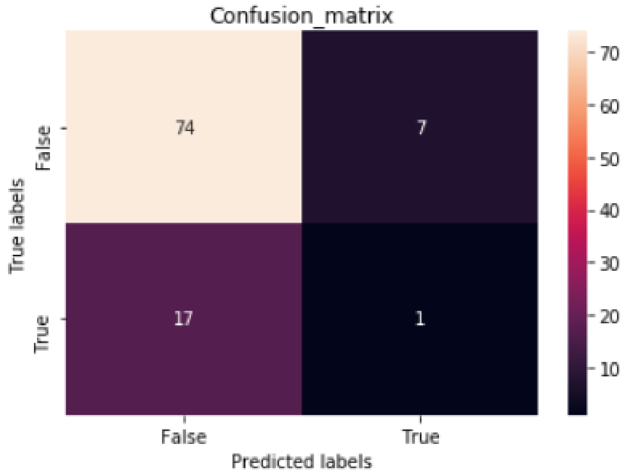


Figure 4.16: Confusion matrix interpretation of Gaussian Naive Bayes

4.5 Support Vector Machine

4.5.1 Support Vector Machine Concept

SVMs are known to be as supervised learning models, when we talk about machine learning. This includes related learning algorithms for classification and regression analysis that are capable to dissect data utilized[29].. An SVM training algorithm builds a model by making it a non-probabilistic binary linear classifier, that allocates new examples to one category or the another, when given a set of training examples those of which are individually marked as belonging to one or the other of two categories. However, there are other existent ways such as the Platt scaling that use SVM in a probabilistic classification setting.[30]. In addition to it, an SVM model can be a depiction of the examples as points in space, those of which are mapped such that the examples of the separate categories are divided by a clear visible gap that is as wide as possible. In that similar space new examples are then mapped and then categorized on the basis of the side of the gap they fall based on the prediction. Apart from performing line classification, by using the kernel trick, SVMs can also efficiently perform a non-linear classification. They do this so by mapping their inputs into high-dimensional feature spaces. An unsupervised learning approach is to be followed when the data's are unlabeled, as supervised learning is not possible in this case. An unsupervised learning approach tries grouping the natural accumulation of data. After that, these data's are then mapped in accordance to these formed groups. One of the most widely accepted clustering algorithms in industrial applications to categorize unlabeled data is the support-vector clustering algorithm which is created by Hava Siegelmann and Vladimir Vapnik; it applies the statistics of support vectors.

4.5.2 Result of our Support Vector Machine

After our model was ready, we calculated the following things from our data:

Accuracy

We have listed our accuracy for our targets features in the table below.

Table 4.4: Accuracy for Support Vector Machine

	First Label	**Second Label**	**Third Label**
Accuracy	81.81	66.66	81.81

Confusion Matrix Interpretation:

Confusion Matrix for First Target Feature:

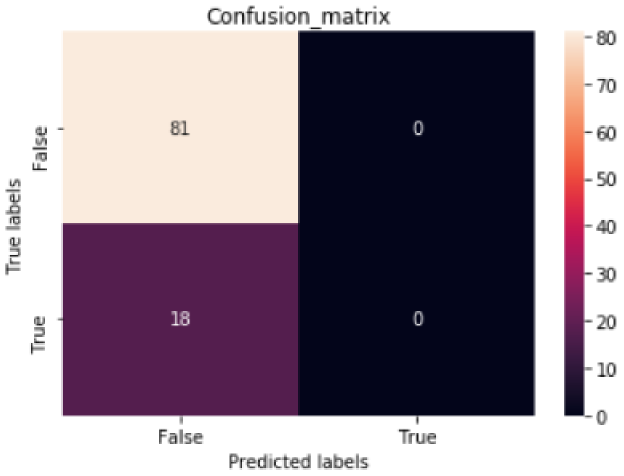


Figure 4.17: Confusion matrix interpretation of Support Vector Machine

Confusion Matrix for Second Target Feature:

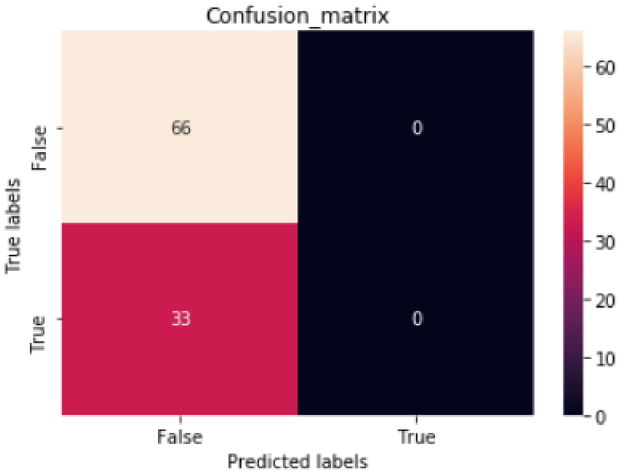


Figure 4.18: Confusion matrix interpretation of Support Vector Machine

Confusion Matrix for Third Target Feature:

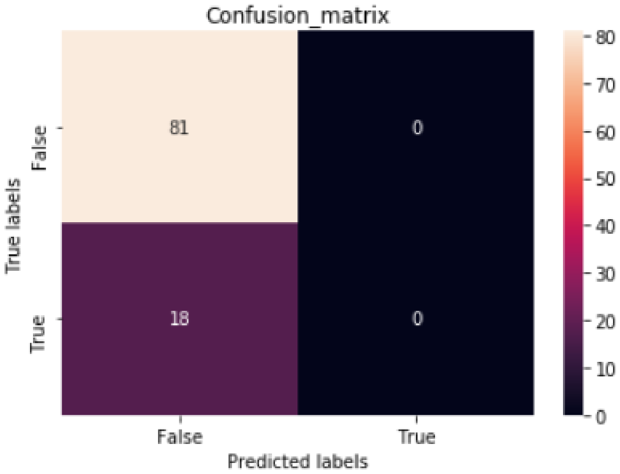


Figure 4.19: Confusion matrix interpretation of Support Vector Machine

4.6 Random forest

4.6.1 Random forest Concept

The name Random forest speaks for itself. It includes a significant number of distinct decision trees, and each of these individual decision trees function as an ensemble [31]. The class bearing the most number of given votes becomes our forecasted model. This happens so as every individual tree that are in the random forest gives out prediction of the class [32]. The underlying notion that we can deduct from random forest is- the wisdom of crowds. It may sound very lucid but this concept has utmost power t. One might wonder, why the random forest works in such a good way?—we can explain this if we shed some light in terms of data sciences: The answer to our question is- the low correlation amongst the models. This can be well explained if we take an example of investments, it consists of bonds and stocks, if we look at each of them individually they have a low correlation but when they are together as portfolio we realize that-the sum is greater than the parts[33]. That is, the individual models are more prone to inaccuracy in comparison to the uncorrelated models that produced as a whole. Now, getting back to our concept, we can say that, the greater number of relatively uncorrelated models (trees) will always be more significant that each of the individual model. The reason for occurrence of such phenomena is that the trees shield themselves from their own errors until and unless each of them do not error in the same way they are good to go. Some trees may in fact be wrong, but however the others may as well be right, so as a result, in the end the trees move in the correct direction when as a group.

4.6.2 Result of our Random forest

After our model was ready, we calculated the following things from our data:

Accuracy

We have listed our accuracy for our targets features in the table below.

Table 4.5: Accuracy for Random forest

	First Label	**Second Label**	**Third Label**
Accuracy	80.80	63.63	81.81

Confusion Matrix Interpretation:

Confusion Matrix for First Target Feature:

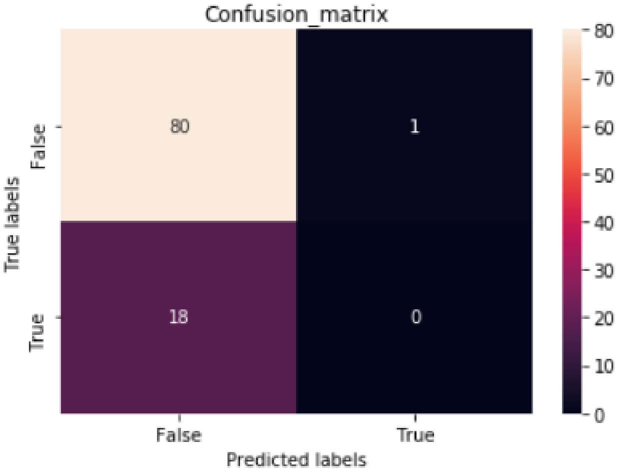


Figure 4.20: Confusion matrix interpretation of Random forest

Confusion Matrix for Second Target Feature:

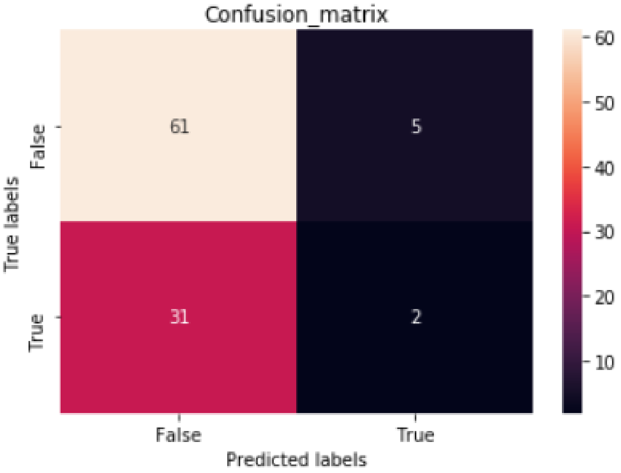


Figure 4.21: Confusion matrix interpretation of Random forest

Confusion Matrix for Third Target Feature:

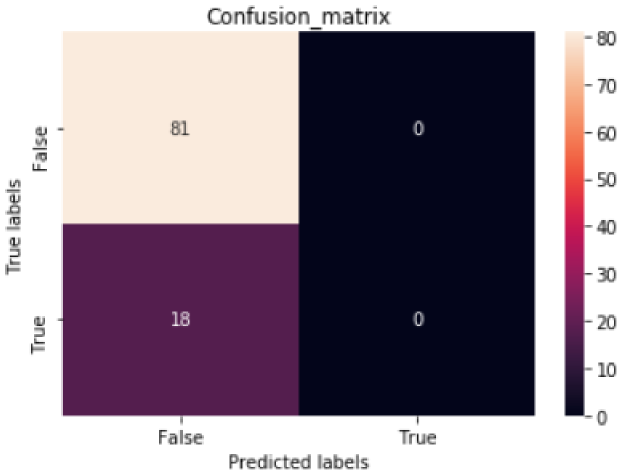


Figure 4.22: Confusion matrix interpretation of Random forest

4.7 Logistic Regression

4.7.1 Logistic Regression Concept

Logistic regression is very different than linear regression. Linear regression gives us an output of continuous number values whereas, logistic regression, works by utilizing the logistic sigmoid function. It works so by returning to a probability value that changes its output such that it can be mapped to two or more discrete classes[34]. So, we can tell that, logistic regression is an algorithm for a discrete set of classes to allocate observations [35]. Logistic regression uses a logistic function. By using that function it estimates probabilities by measuring the relationship amongst the categorical dependent variable and one or more multiple number of independent variables. This is what we know as the cumulative distribution function of logistic distribution. Which is why, we can say that it assumes similar set of problems as probit [36] regression using corresponding techniques. The linear model uses a cumulative normal distribution curve. In similar ways, when it comes to the latent variable interpretations between these two methods, we can say that, the linear assumes a standard logistic distribution of errors whereas the logistic assumes a standard normal distribution of errors. Linear model is in itself a much generalized model. However, the logistic regression can be considered to be a unique case of the linear model, which explains why it is so similar to linear model. But even after having certain similarities, the basis of assumption of logistic regression is quite distinct, when it to accounting for the relationship that exists between the depended variables and independent variables.

4.7.2 Result of our Logistic Regression

After our model was ready, we calculated the following things from our data:

Accuracy

We have listed our accuracy for our targets features in the table below.

Table 4.6: Accuracy for Logistic Regression

	First Label	Second Label	Third Label
Accuracy	80.80	65.65	81.81

Confusion Matrix Interpretation:

Confusion Matrix for First Target Feature:

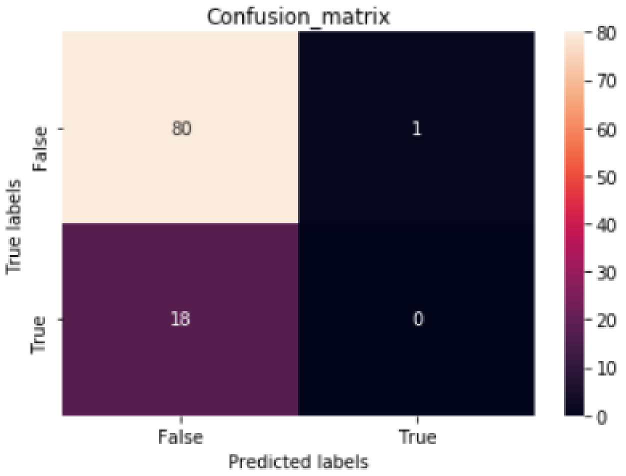


Figure 4.23: Confusion matrix interpretation of Logistic Regression

Confusion Matrix for Second Target Feature:

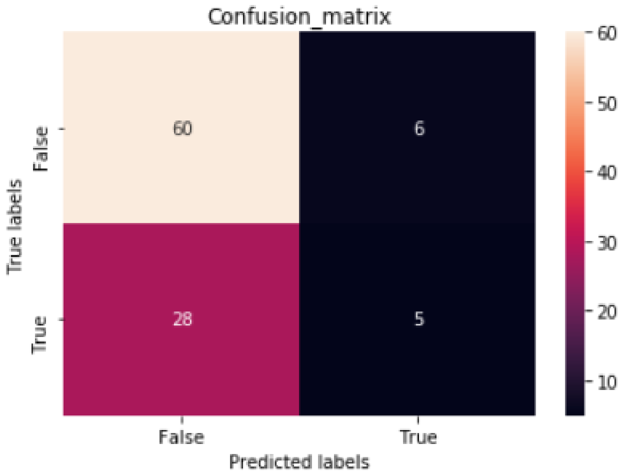


Figure 4.24: Confusion matrix interpretation of Logistic Regression

Confusion Matrix for Third Target Feature:

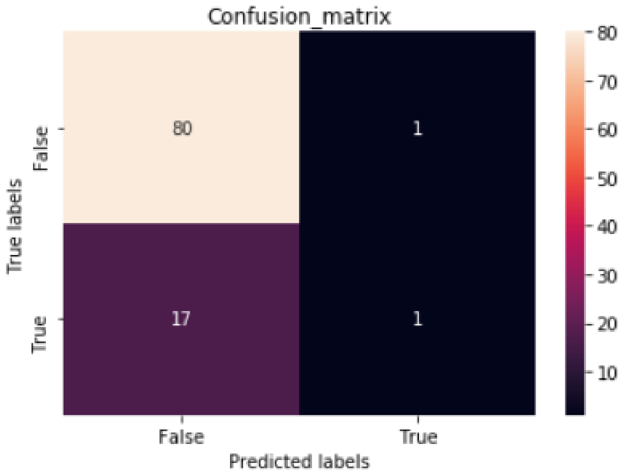


Figure 4.25: Confusion matrix interpretation of Logistic Regression

4.8 Xgboost

4.8.1 Xgboost Concept

Gradient boosting is a machine learning procedure for regression and classification problems, which delivers a prediction model as a troupe of feeble prediction models, commonly decision trees. It constructs the model in a stage-wise design like other boosting methods do, and it sums them up by permitting optimization of an arbitrary differential loss function. [37] XGBoost is mainly a ensemble Machine Learning algorithm based on decision-tree, which uses a gradient boosting framework. When it come to prediction problems which includes unstructured data like images, text, etc. artificial neural networks are more likely to overpower most of the other algorithms or frameworks [37]. Be that as it may, with regards to data with structure and small to medium data, decision tree based algorithms are viewed as top tier right now [38]. When it comes to boosting, the trees are built progressively with the ultimate objective that each subsequent tree intends to diminish the errors of the past tree. Each tree gains from its precursors and updates the lingering blunders. In this manner, the tree that gets next in the progression will pick up from a refreshed adaptation of the residuals [39]. The base learners in boosting are frail learners in which the inclination is high, and the prescient power is just somewhat better than arbitrary theorizing. All of these powerless students contributes some essential data for prediction, enabling the boosting framework to convey a solid learner by effectively combining these frail learners. The last solid learner chops down both the bias and the difference.

4.8.2 Result of our Xgboost

After our model was ready, we calculated the following things from our data:

Accuracy

We have listed our accuracy for our targets features in the table below.

Table 4.7: Accuracy for Xgboost

	First Label	**Second Label**	**Third Label**
Accuracy	81.81	64.64	76.76

Confusion Matrix Interpretation:

Confusion Matrix for First Target Feature:

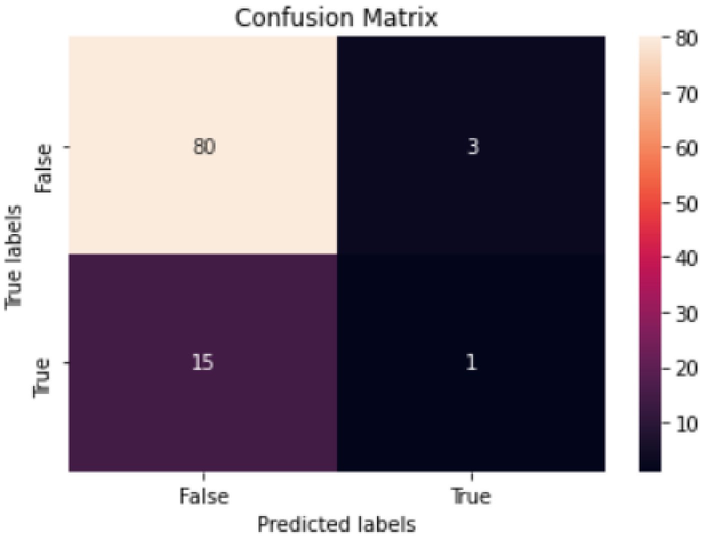


Figure 4.26: Confusion matrix interpretation of Xgboost

Confusion Matrix for Second Target Feature:

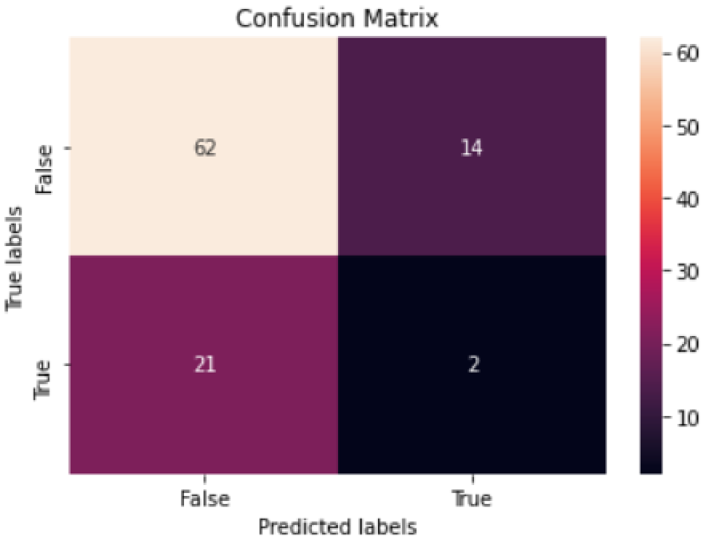


Figure 4.27: Confusion matrix interpretation of Xgboost

Confusion Matrix for Third Target Feature:

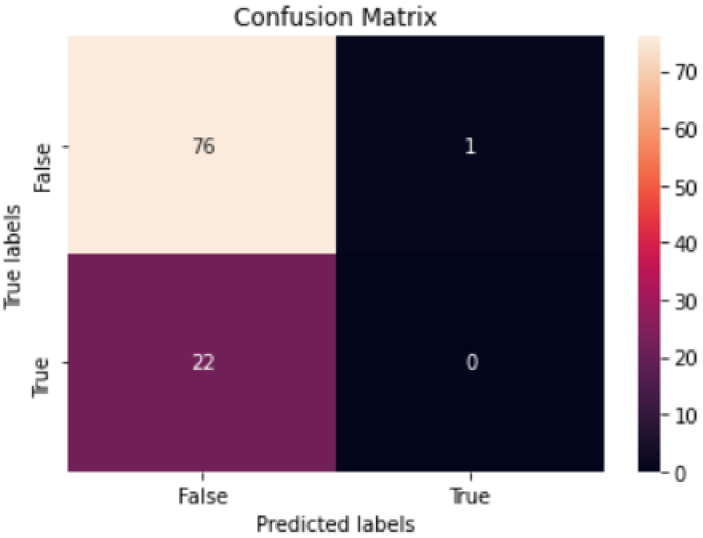


Figure 4.28: Confusion matrix interpretation of Xgboost

4.9 Comparison between Machine Learning Algorithms

From the comparison table,we can see that our self-designed ANN model have over all performed well on our data. On the other hand Decision-Tree Classifier and Gaussian Naive Bayes have performed moderately well on our model. With slightly different result in same accuracy other algorithms have shown relatively poor performance on our data.

Table 4.8: Comparison Table.

Model Name	First Label Accuracy	Second Label Accuracy	Third Label Accuracy
ANN	89.89%	94.94%	86.86%
Decision-Tree	75.75%	64.64%	75.75%
Gaussian Naive Bayes	81.81%	66.66%	81.81%
Support Vector Machine	80.80%	63.63%	81.81%
Random forest	80.80%	65.65%	81.81%
Logistic Regression	80.80%	65.65%	81.81%
Xgboost	81.81%	64.64%	76.76%

4.9.1 Reasons behind superior performance of ANN

Among many advantages, ANNs show the following key traits, for which ANN is our best model for building the result of the research.

1. Interestingly, ANNs can utilise data to learn and model data which are not linear and has complex relationships[40], making it far more cruicial for our research as in real-life most of the data we visit has inputs and outputs which are complex as well as non-linear.

2. ANNs have the ability to generalize — As ANN starts to identify the data, primary inputs and the relationship among them, Ann is able to show similarities which are unseen and which occurs in unseen or unlisted data. Therefore, the model is more likely to generalize and make predictions on unseen data.

3. Among various other techniques of prediction, ANN shows barely any restrictions to the input values (as they are distributed). Moreover, there are numerous studies and findings showing that ANNs are a better model heteroskedasticity, meaning the data which have high volatility and non-constant variance, as it learns all the hidden relationships, similarities of the data without showing any permanent similarity or relation.

4.9.2 Why are we claiming Decision-Tree and Gaussian Naive Bayes are better than other models except ANN?

If we look at accuracy of other models,it may seem like we made a mistake claiming Decision-Tree and Gaussian Naive Bayes are better than other models except ANN. One of the most unavoidable and crucial metric is accuracy, however, it can be sometimes misleading, so it should implemented with care alongside with confusion matrix [41]. SVM,Random forest and rest of models showed **naive behavior**.As our dataset is imbalanced ,those models actually predicted all test samples as negative. Those model failed to predict positive samples correctly at all.But as our dataset mostly consists of negative samples,accuracy is high.In reality those models have not learned anything.

On the contrary,Decision-Tree and Gaussian Naive Bayes learned to some extent to distinguish between positive and negative samples.

4.9.3 Possible solutions to better the performance of our models

Firstly, when working with imbalanced dataset the first impression is that the data do not signifies reality: in this cases, in our gathered dataset, we can safely assume that the real data is almost balanced, however, slight bit of the data has proportions bias, which may be a cause of how we collected the data. In such case, what is more advised to do is gather more data which represents reality.

Getting more features and generating more synthetic data is our best option to tackle imbalanced data.

Chapter 5

Conclusion

The main purpose of our research was to see how various supervised algorithms predict our three targets, Facebook user's chance of becoming a victim of identity impersonating, phishing and victim of malware based on the dataset as well as, determining which algorithm performs the best predicating the three targets. In our paper, we showed this prediction using seven algorithms. We were able to achieve our objectives through comparing the algorithms for finding the best possible approach based on our survey data. We got ANN as the best possible approach and Decision Tree as the worst approach. In this research work, we used supervised machine learning algorithms and came to know about how different algorithm yields different outcome and we demonstrated how the best resulting algorithm yielded such results. We hope this paper can provide readers with a better understanding of Facebook risk analysis and create more options to secure their personal information.

Future Plan

In this research we have analyzed Facebook user risk by using different algorithms. Our future plan is to work on other social media platforms like Twitter, Instagram etc. using these algorithms. Also we are interested to gather more data than we have collected for this project. We also plan to work on these models with different parameters to determine if the accuracy and performance of these algorithms can be increased and predict more accurately. In future, we would also like to implement our prediction algorithms to predict Facebook or any other social media user's security, by letting them complete a updated questionnaire and they can see where to be more careful so that their social media is not vulnerable anymore.

Bibliography

[1] S. Aboulhosn, *18 facebook statistics every marketer should know in 2020*, Mar. 2020. [Online]. Available: https://sproutsocial.com/insights/facebook-stats-for-marketers/.

[2] G. Grosseck, R. Bran, and L. Tiru, "Dear teacher, what should i write on my wall? a case study on academic uses of facebook", *Procedia-Social and Behavioral Sciences*, vol. 15, pp. 1425–1430, 2011.

[3] A. Vishwanath, "Habitual facebook use and its impact on getting deceived on social media", *Journal of Computer-Mediated Communication*, vol. 20, no. 1, pp. 83–98, 2015.

[4] G. Moubarak, A. Guiot, Y. Benhamou, A. Benhamou, and S. Hariri, "Facebook activity of residents and fellows and its impact on the doctor–patient relationship", *Journal of medical ethics*, vol. 37, no. 2, pp. 101–104, 2011.

[5] D. M. Freeman, "Using naive bayes to detect spammy names in social networks", *Proceedings of the 2013 ACM workshop on Artificial intelligence and security - AISec 13*, 2013. DOI: 10.1145/2517312.2517314.

[6] A. Nosko, E. Wood, and S. Molema, "All about me: Disclosure in online social networking profiles: The case of facebook", *Computers in human behavior*, vol. 26, no. 3, pp. 406–418, 2010.

[7] A. Subroto and A. Apriyana, "Cyber risk prediction through social media big data analytics and statistical machine learning", *Journal of Big Data*, vol. 6, no. 1, Jul. 2019. DOI: 10.1186/s40537-019-0216-1.

[8] S. Aboulhosn, *18 facebook statistics every marketer should know in 2020*, Mar. 2020. [Online]. Available: https://sproutsocial.com/insights/facebook-stats-for-marketers/.

[9] [Online]. Available: https://www.bloomberg.com/news/articles/2019-05-23/facebook-removed-2-2-billion-fake-accounts-in-first-quarter.

[10] S. Kumari and S. Singh, "A critical analysis of privacy and security on social media", *2015 Fifth International Conference on Communication Systems and Network Technologies*, 2015. DOI: 10.1109/csnt.2015.21.

[11] M. Injadat, F. Salo, and A. B. Nassif, "Data mining techniques in social media: A survey", *Neurocomputing*, vol. 214, pp. 654–670, 2016.

[12] Y. Zhao, "Data mining techniques.", 2015.

[13] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques.* Elsevier, 2011.

[14] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda, “All your contacts are belong to us: Automated identity theft attacks on social networks”, Jan. 2009, pp. 551–560. DOI: 10.1145/1526709.1526784.

[15] T. Govani and H. Pashley, “Student awareness of the privacy implications when using facebook”, *Unpublished paper presented at the “Privacy poster fair” at the Carnegie Mellon university school of library and information science*, vol. 9, pp. 1–17, 2005.

[16] K. Sethi, A. Gupta, G. Gupta, and V. Jaiswal, “Comparative analysis of machine learning algorithms on different datasets”, Apr. 2019.

[17] S. A. Alasadi and W. S. Bhaya, “Review of data preprocessing techniques in data mining”, *Journal of Engineering and Applied Sciences*, vol. 12, no. 16, pp. 4102–4107, 2017.

[18] S. Visa, B. Ramsay, A. Ralescu, and E. Knaap, “Confusion matrix-based feature selection.”, vol. 710, Jan. 2011, pp. 120–127.

[19] M. Sadiku, A. Shadare, S. Musa, C. Akujuobi, and R. Perry, “Data visualization”, *International Journal of Engineering Research and Advanced Technology (IJERAT)*, vol. 12, pp. 2454–6135, Dec. 2016.

[20] B. M. Shahshahani and D. A. Landgrebe, “The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon”, *IEEE Transactions on Geoscience and remote sensing*, vol. 32, no. 5, pp. 1087–1095, 1994.

[21] A. Blum and T. Mitchell, “Combining labeled and unlabeled data with co-training”, in *Proceedings of the eleventh annual conference on Computational learning theory*, 1998, pp. 92–100.

[22] E. Keogh, L. Keogh, and J. C. Handley, “Compression-based data mining”, *Encyclopedia of Data Warehousing and Mining, Second Edition*, pp. 278–285, DOI: 10.4018/978-1-60566-010-3.ch045.

[23] J. Hernandez, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, “An empirical study of oversampling and undersampling for instance selection methods on imbalance datasets”, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications Lecture Notes in Computer Science*, pp. 262–269, 2013. DOI: 10.1007/978-3-642-41822-8_33.

[24] K. Bala, “Classification using ann: A review”, *International Journal of Computational Intelligence Research*, vol. 13, no. 7, pp. 1811–1820, 2017.

[25] “A survey on decision tree algorithms of classification in data mining”, *International Journal of Science and Research (IJSR)*, vol. 5, no. 4, pp. 2094–2097, May 2016. DOI: 10.21275/v5i4.nov162954.

[26] J. R. Quinlan, “Induction of decision trees”, *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[27] D. M. Freeman, “Using naive bayes to detect spammy names in social networks”, *Proceedings of the 2013 ACM workshop on Artificial intelligence and security - AISec 13*, 2013. DOI: 10.1145/2517312.2517314.

[28] C. Ratanamahatana and D. Gunopulos, "Feature selection for the naive bayesian classifier using decision trees.", *Applied Artificial Intelligence*, vol. 17, pp. 475–487, May 2003. DOI: 10.1080/713827175.

[29] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines", *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[30] T. Evgeniou and M. Pontil, "Support vector machines: Theory and applications", in *Advanced Course on Artificial Intelligence*, Springer, 1999, pp. 249–257.

[31] M. Al-Janabi and P. Andras, "A systematic analysis of random forest based social media spam classification", in *International Conference on Network and System Security*, Springer, 2017, pp. 427–438.

[32] N. Matloff, "Multiclass classification problems", *Statistical Regression and Classification*, pp. 179–213, 2017. DOI: 10.1201/9781315119588-5.

[33] L. Breiman, "Random forests", *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[34] G. Chauhan, *All about logistic regression*, Oct. 2018. [Online]. Available: https://towardsdatascience.com/logistic-regression-b0af09cdb8ad.

[35] J. R. Brzezinski and G. J. Knafl, "Logistic regression modeling for context-based classification", in *Proceedings. Tenth International Workshop on Database and Expert Systems Applications. DEXA 99*, IEEE, 1999, pp. 755–759.

[36] *(tutorial) understanding logistic regression in python.* [Online]. Available: https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python.

[37] S. Ramraj, N. Uzir, R. Sunil, and S. Banerjee, "Experimenting xgboost algorithm for prediction and classification of different datasets", *International Journal of Control Theory and Applications*, vol. 9, pp. 651–662, 2016.

[38] J. Swain, "A hybrid approach to thematic analysis in qualitative research: Using a practical example", 2018. DOI: 10.4135/9781526435477.

[39] *Article: Fitting data with xgboost.* [Online]. Available: https://www.actuaries.org.uk/news-and-insights/news/article-fitting-data-xgboost.

[40] J. Mahanta, *Introduction to neural networks, advantages and applications*, Jul. 2017. [Online]. Available: https://towardsdatascience.com/introduction-to-neural-networks-advantages-and-applications-96851bd1a207.

[41] M. Mohammadrezaei, M. E. Shiri, and A. M. Rahmani, "Identifying fake accounts on social networks based on graph analysis and classification algorithms", *Security and Communication Networks*, vol. 2018, pp. 1–8, May 2018. DOI: 10.1155/2018/5923156.