**Cybercrime Awareness among Saudi Nationals: Dataset**

Machine Learning Capstone Project

By:

Hala Al-Souly

Mohammed Al-Shihab

Mustafa Khurida

Amani Almutairi

Asma Alghamdi

**Abstract**

The idea of the project is to collect data by making a survey to measure the level of cybersecurity awareness of cyber- crime in Saudi Arabia. The survey recruited 1230 Saudi nationals with different backgrounds and relies on knowledge and attitude aspects. The collected data were gathered between August and October of 2019. The survey contains several questions, which are listed below. Personal and skills information, Cybersecurity Activities which aims to assess current Information technology knowledge, Cybercrime Consciousness to measure what subjects believe and their opinion, Case Reports, which aimed to evaluate subjects and reactions when they faced a cybercrime incident. The purpose of the project is to study, analyze, building models and explore data to obtain the information needed to educate people about cybercrime and to inform the relevant authorities.

**Background of the study**

Many people think that security is not outside the scope of its traditional concept. But the concepts have recently expanded to include many types of security, foremost of which is cyber security, which has become an urgent necessity after the emergence of the Fourth Industrial Revolution or what is known as the data revolution, because the Internet space has become teeming with electronic transactions and transactions that need to be encrypted and secure those transactions.

The concept of cybersecurity is more comprehensive and broader than information security, especially if we considered data protection and various informational applications such as e-government, e-commerce, e-health, e-education and information society. At present, the vision of the Kingdom of Saudi Arabia aims to develop and teach a lot in the field of information technology, especially cybersecurity. Several platforms and academies have been established by the government to educate people about the importance of the field of cyber security, including the National Cyber ​​Security Authority, the Saudi Digital Academy, the Tuwaiq platform and many more platforms. Also, there are many cyber security courses given to all age groups to raise awareness in protecting their information. This survey was conducted to study several criteria to see people's awareness of cybercrime, including the level of study, age group, administrative regions, purpose of using the Internet, level of Internet skills, system used, practices used in the Internet, type of communication, security knowledge, type of device , the software used to protect against viruses, have you been a victim before in cybercrime or not, and many questions that aim to see the level of awareness and knowledge and also see the practices that made you exposed to a victim of cybercrime

**Business Problem**

It turns out that there are some who have no awareness of cybercrime and do not know how to protect their information. There are also bad practices in the use of computers and the Internet, including: those who browse in unsafe sites without their knowledge, and there are some who do not use virus protection programs in their devices. There are also those who use easy passwords and make them match personal information, which means it is easy to fall victim to an information crime. There are some people who do not read the terms and conditions for registering on some sensitive sites and they should be alerted about this.

Therefore, there are several points in the data that indicate a lack of awareness of cybercrime, which needs to be improved and developed.

**Motivation for the Project**

Talking about how to raise societal awareness of obtaining safety and security in the world of the Internet and protecting the digital identity, we should draw attention to the Internet access guidelines according to age groups and education level. Finding bad practices in the use of the Internet through the data collected to raise awareness. The data are important since they evaluate the level of awareness of Saudis based on aspects not covered in previous studies, such as using participants of different backgrounds, regions and expertise in utilizing technology. The data can be useful for Saudi authorities such as the National Cybersecurity authority as well as researchers who are interested in the cybersecurity field. The data can be utilized for educational purposes in terms of short courses and training.

**Approaches and Methodologies**

The project uses machine learning technology to see the criteria that cause the victim of a cybercrime through the practices used by the user on the Internet and computer use. The technology used for the project is supervised learning and unsupervised concept to predict the outcomes of being a victim of cybercrime or not.

**Exploratory Data Analysis**

Number of Victims of cybercrime, this shows numbers of victim approximately above 200 and less than 400 and the most are not victim

Chart, bar chart

Description automatically generated

Victim by Age, Gender, Education and using public Wi-Fi, this shows the most age

between 18-20 are the victims.

For the gender, the males victims are more than females.

For the education level, the undergraduate (Diploma, BSC) are the most victims

For using Public Wi-Fi, sometimes are the most choice.

Chart, bar chart

Description automatically generated

People who install software updates regularly in the 18-29 age group, the most which mean the other age group need to be aware and take courses training

Chart, bar chart, histogram

Description automatically generated

People who use personal information in their password, always 21% and often 30% which is a lot use bad practice in making password

Chart, pie chart

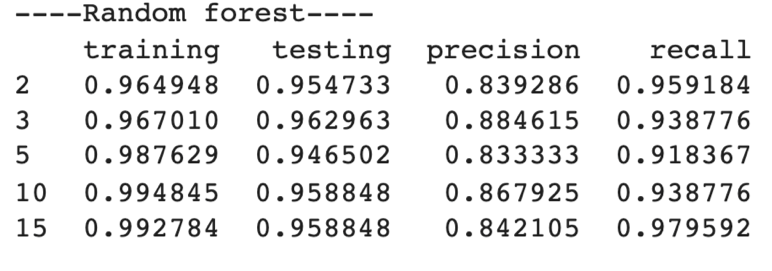
Description automatically generated

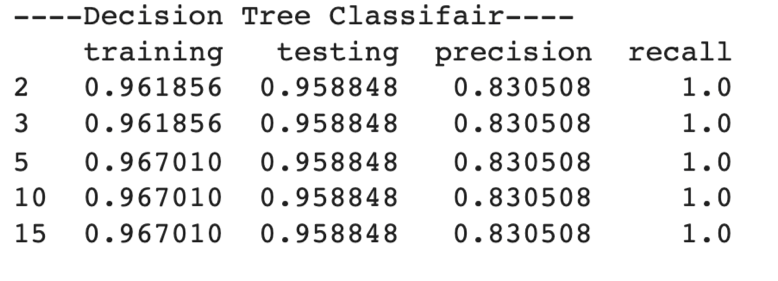
**Models:**

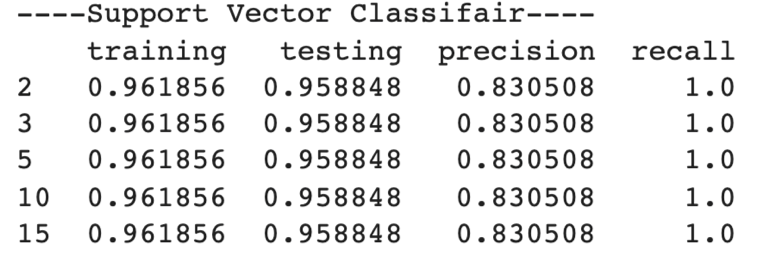
Supervised models:

In supervised we tried to run the code with multiple features with 10 features, 5 features, 20 features by using a for loop, and we change the number of features we used to see who give us the best result. In supervised we using 4 different models: Logistic Regression, Decision Tree Classifair, Random Forest and Support Vector Classifair.

Here some result of them:







As you can see Random forest was the best models in supervised model.

Unsupervised models:

In unsupervised we apply two models: PCA, Kmean. In PCA we take 9 components for it, here some result for the measure without and with PCA:

Measure how long it takes to train WITHOUT PCA

13.5 ms ± 44.2 µs per loop (mean ± std. dev. of 7 runs, 100 loops each)

Measure how long it takes to train WITH PCA

3.92 ms ± 17.5 µs per loop (mean ± std. dev. of 7 runs, 100 loops each)

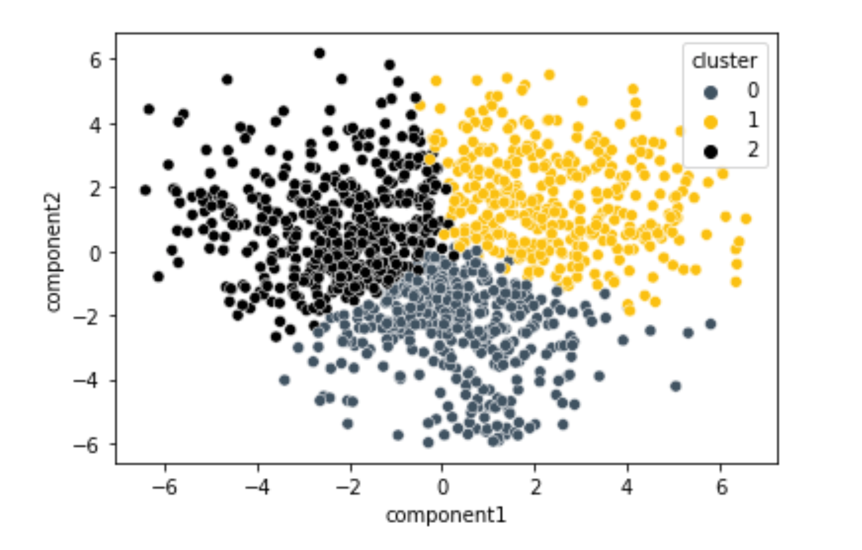
And this is the accuracy of PCA, still Random Forest give us the best approach

PCA Training accuracy: 0.9350515463917526

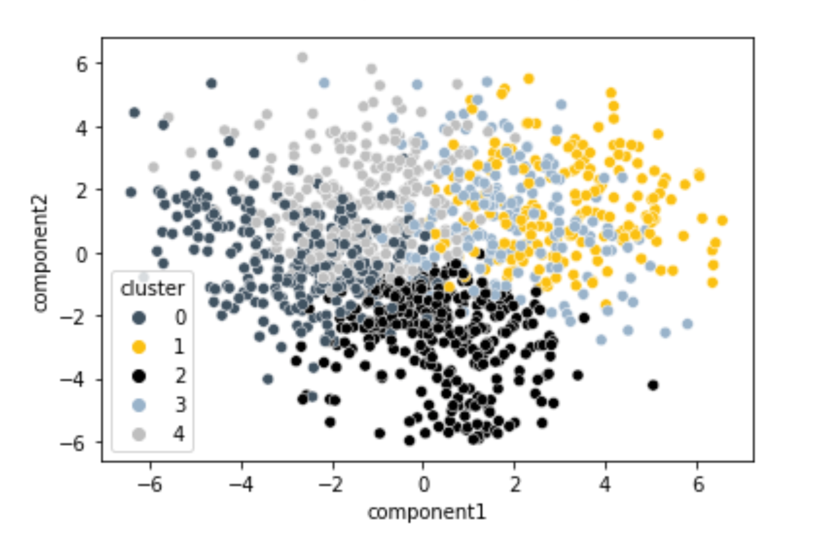
PCA Testing accuracy: 0.9259259259259259

Kmean when we apply the Elbow Graph, we found the cluster stop on 3 and 5 so we decide to draw each one.

Three cluster:



Five cluster:



As you can see 3 cluster give US the best result.

Deep learning:

In deep learning we used ANN (Artificial Neural Network). We add 2 hidden layers with 5 unit for each one and for epochs = 200 so the result was the best, of Corse we can add more layers but we are happy with the result. As you can see the ANN was close to Random Forest.

