NTNU

Kunnskap for en bedre verden

Comparative analysis of machine learning methods for analyzing security practice in electronic health records' logs.

By: Prosper K. Yeng, Muhammad A. Fauzi, Bian Yang:

Norwegian University of Science and Technology

Outline

Motivation

Background

Objective/Research Question

Results

Discussion

Hacking scandal shakes Finland - patients pressured for money

Patient information from a Finnish psychotherapy center is going astray after hacking, and several patients have been pressured for money.

Bjørnar Hjellen
@bjornarhjellen
Journalist

Source: **NTB-NRK**

Published Oct 25 at 12:58

securityweek.com/german-hospital-hacked-patient-taken-another-city-dies

Home › Malware

German Hospital Hacked, Patient Taken to Another City Dies

By Associated Press on September 17, 2020

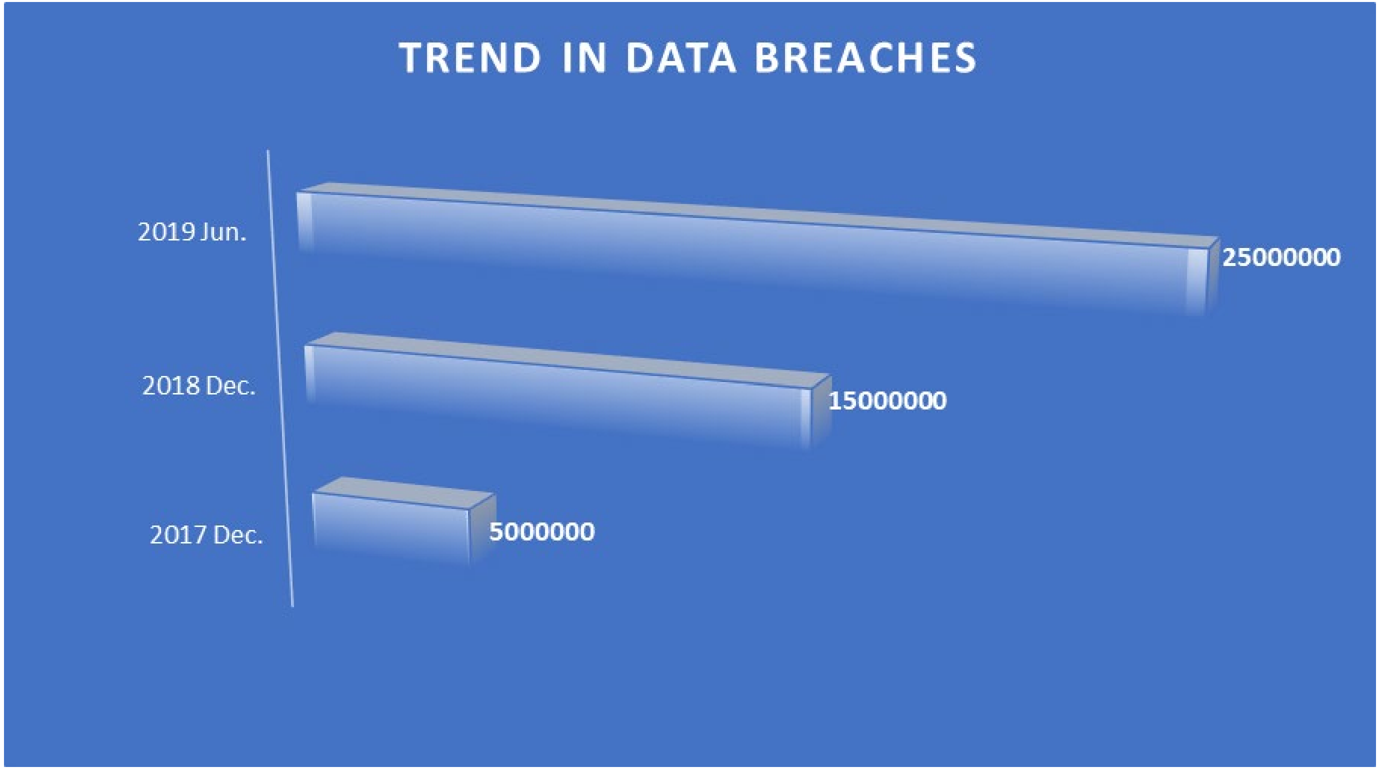
Share Tweet Recommend 389 RSS

German authorities said Thursday that what appears to have been a misdirected hacker attack caused the failure of IT systems at a major hospital in Duesseldorf, and a woman who needed urgent admission died after she had to be taken to another city for treatment.

The Duesseldorf University Clinic's systems have been disrupted since last Thursday. The hospital said investigators have found that the source of the problem was a hacker attack on a weak spot in "widely used commercial add-on software," which it didn't identify.

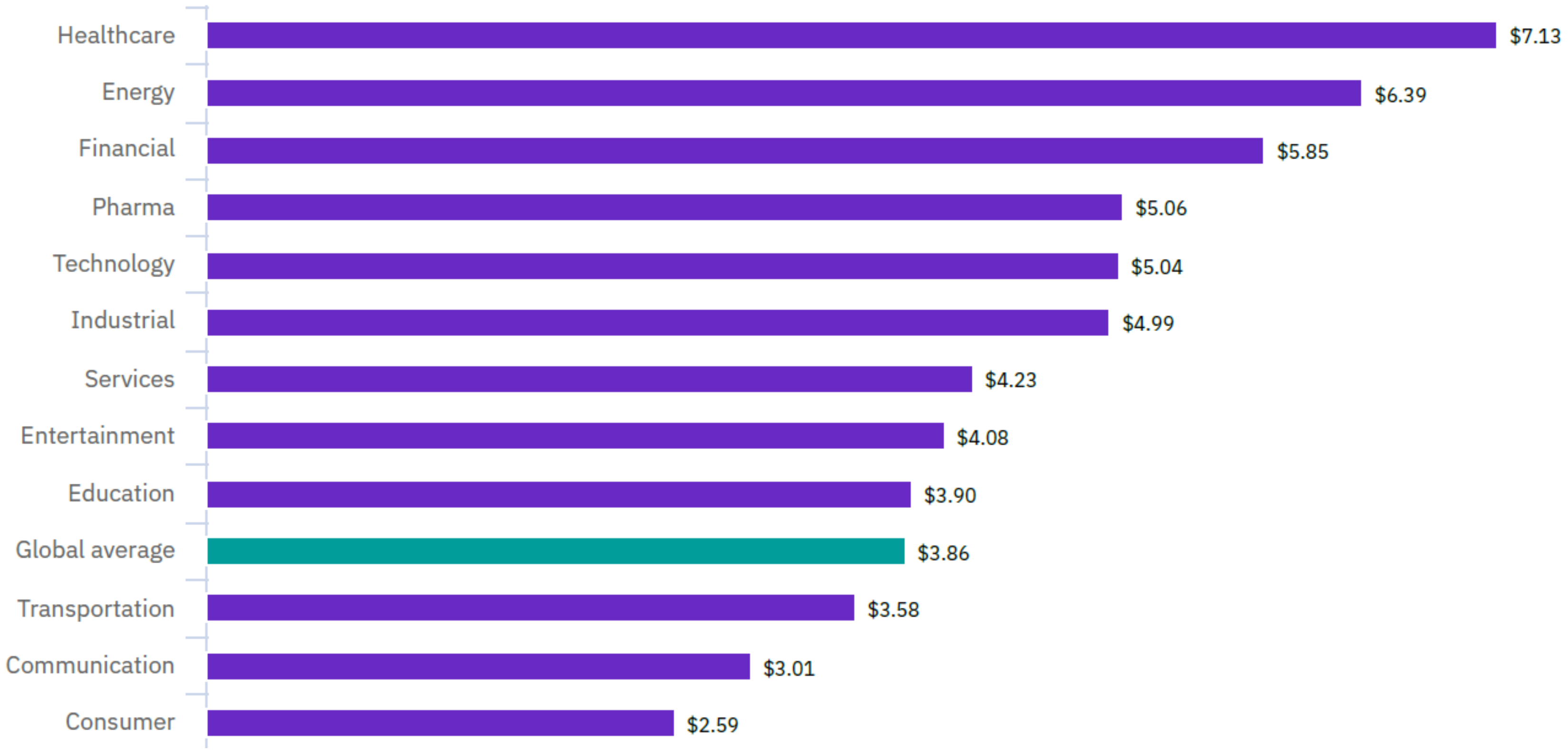
As a consequence, systems gradually crashed and the hospital wasn't able to access data; emergency patients were taken elsewhere and operations postponed.

The hospital said that that "there was no concrete ransom demand." It added that there are no indications that data is irretrievably lost and that its IT systems are being gradually restarted.


TREND IN DATA BREACHES
2019 Jun.
25000000
2018 Dec.
15000000
2017 Dec.
5000000

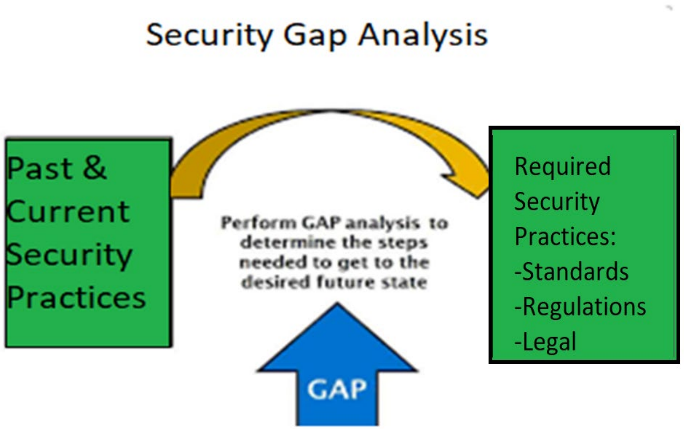
Average total cost of a data breach by industry

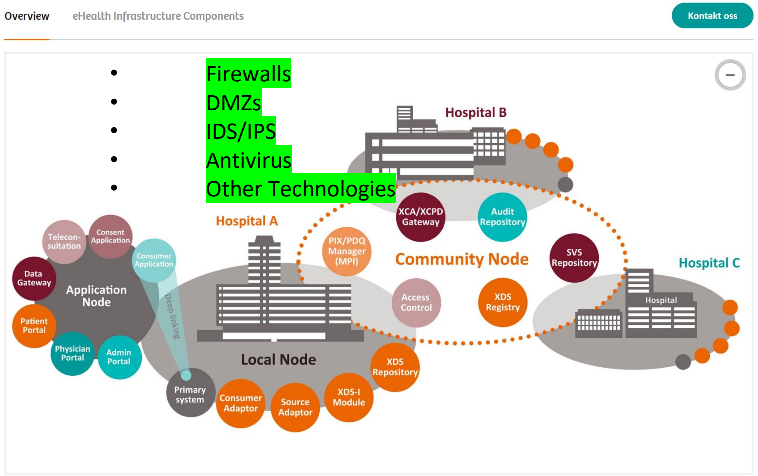
Measured in US$ millions



NTNU

What is the goal and Why?

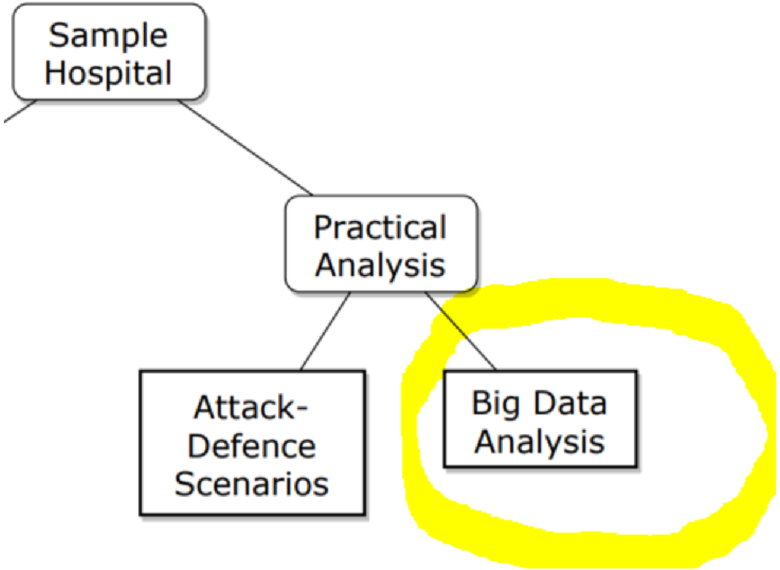




More attention on tech. measures than the human firewall!

What is the objective of this paper?

- To compare machine learning classification methods towards anaysing healthcare security practice.



What is healthcare security practice?

Healthcare:

- Behaviors required to put up by healthcare staff in order to comply with CIA requirements of information systems

Big data context:

- Traces of users' electronic accesses (logs) which can be reconstructed to form individuals unique access profiles

What is the Scope Research Question and contribution?

Amid various machine learning methods,
which of the methods is suitable for analysing
healthcare security practice in EHR logs?

Method: Data simulation (Normal data)

NTNU

Security requirement in EHR

Case study: Norway

NTNU

TABLE I: List of Departments

ID	Name
0	IT
1	Finance
2	Administration
3	Laboratory
4	Pharmacy
5	Out Patients Ear-Nose-Throat
6	Out Patients Eyes
7	Out Patients Tooth
8	Out Patients Child
9	Out Patients Orthopedic
10	Out Patients Neurological
11	Out Patients Gynecological
12	Out Patients Diabetes
13	Out Patients Rheumatology
14	Out Patients Cancer
15	Emergency
16	In Patients Ward1
17	In Patients Ward2
18	In Patients Ward3

NTNU

TABLE II: List of Roles

ID	Name	Code
0	Head of IT	HIT
1	Technical Support	TS
2	Head of Finance	HF
3	Finance Staff	FS
4	Head of Administration	HA
5	Staff of Administration	SA
6	Head of Lab	HL
7	Lab Assistant	LA
8	Head of Pharmachy	HP
9	Pharmacy Assistant	PA
10	Doctor	DO
11	Nurse	NU

NTNU

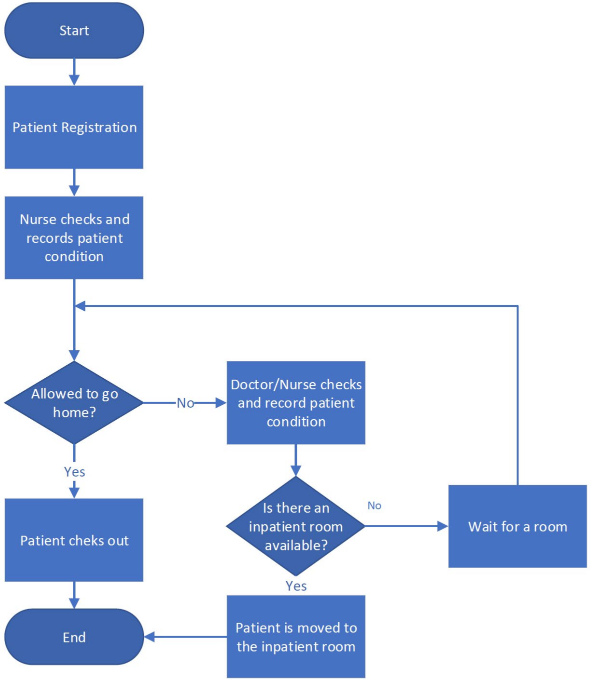
TABLE IV: Three 8-hours shift

Department	Roles (number of employe
Emergency	DO(2), NU(7)
In Patients Ward1	NU(2)
In Patients Ward2	NU(2)
In Patients Ward3	NU(2)

TABLE III: Regular Shift

ID	Department	Roles (number of employee
0	IT	HIT(1), TS(2)
1	Finance	HF(1), FS(4)
2	Administration	HA(1), SA(2)
3	Laboratory	HL(1), LA(5)
4	Pharmacy	HP(1), PA(2)
5	Out Patients Ear-Nose-Throat	DO(1), NU(2)
6	Out Patients Eyes	DO(1), NU(2)
7	Out Patients Tooth	DO(1), NU(2)
8	Out Patients Child	DO(1), NU(2)
9	Out Patients Orthopedic	DO(1), NU(2)
10	Out Patients Neurological	DO(1), NU(2)
11	Out Patients Gynecological	DO(1), NU(2)
12	Out Patients Diabetes	DO(1), NU(2)
13	Out Patients Rheumatology	DO(1), NU(2)
14	Out Patients Cancer	DO(1), NU(2)
16	In Patients Ward1	DO(1)
17	In Patients Ward2	DO(1)
18	In Patients Ward3	DO(1)

Inpatient flow



EHR rules by the healthcae code of conduct in Norway:

- Accessing patients records is only allowed for therapeutic purposes
- Access is given to only those with an official need to use,
- Self-authorization or ”break the glass” scenarios is allowed but the necessary measures should be provided,
- All of the activities related to access of the personal health data (register, update, edit, delete etc)must be logged

Attributes and features

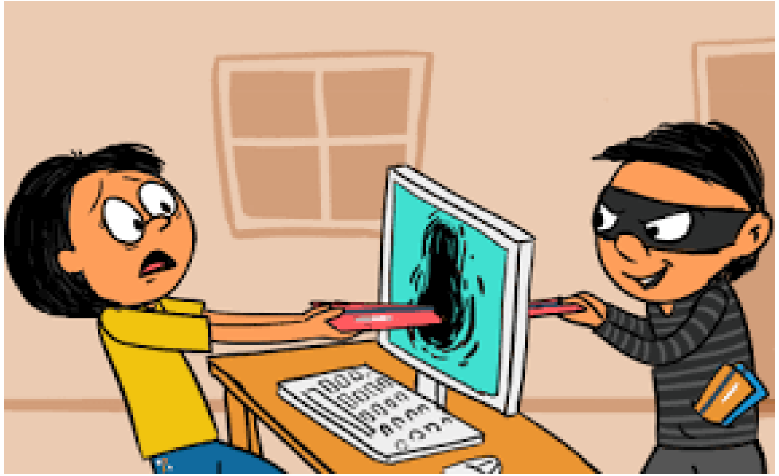
TABLE V: Record Fields

Number	Field Name	Description
1	startAccessTime	The time employee start to acces the patient record. format = 'dd/mm/yyy HH:mm tt'
2	endAccessTime	The time employee end the patient record access. format = 'dd/mm/yyy HH:mm tt'
3	employeeID	The ID of the employee who access the patient record
4	roleID	The role of the employee who access the patient record
5	patientID	The ID of the patient whose record is being accessed by employee
6	activityID	The ID of the activity (1: Create, 2: Read, 3:Update, 4: Delete)
7	employeeDepartmen-tID	The department of the employee who access the patient record
8	employeeOrganiza-tionID	The organization of the employee who access the patient record
9	osID	The OS of the computer used by the employee to access patient record
10	deviceID	The ID of the computer used by the employee to access patient record
11	browserID	The browser used by the employee to access patient record
12	ipAddress	The IP Address of the computer used by the employee to access patient record

TABLE VI: Dataset feature names and descriptions

Feature Name	Description
number of create	Number of 'create' transactions conducted in a single day
number of read	Number of 'read' transactions conducted in a single day
number of update	Number of 'update' transactions conducted in a single day
number of delete	Number of 'delete' transactions conducted in a single day
number of patient record	Number of access to the patient records in a single day
number of unique patient	Number of unique patients whose records has been accessed in a single day
number of modules	Number of kind of modules in the information system accessed in a single day
number of report module	Number of transactions conducted in the report module in a single day
number of finance module	Number of transactions conducted in the finance module in a single day
number of patient module	Number of transactions conducted in the patient management module in a single day
number of lab module	Number of transactions conducted in the laboratory module in a single day
number of pharmacy module	Number of transactions conducted in the pharmacy module in a

Abnormal data simulation



- Access by identity theft.
  - The attacker will access more data than legitimate users
  - Attackers sometimes not follow the flows.

Data Processing cont...

From this data simulation,

- 283,678 logs were created
- Legitimate access were 274,983
- Fraudulent access were 8,695

Data Processing cont...

we process the logs data into 24-hour blocks so that an instance represents the

cumulative activity of a user in a single day.

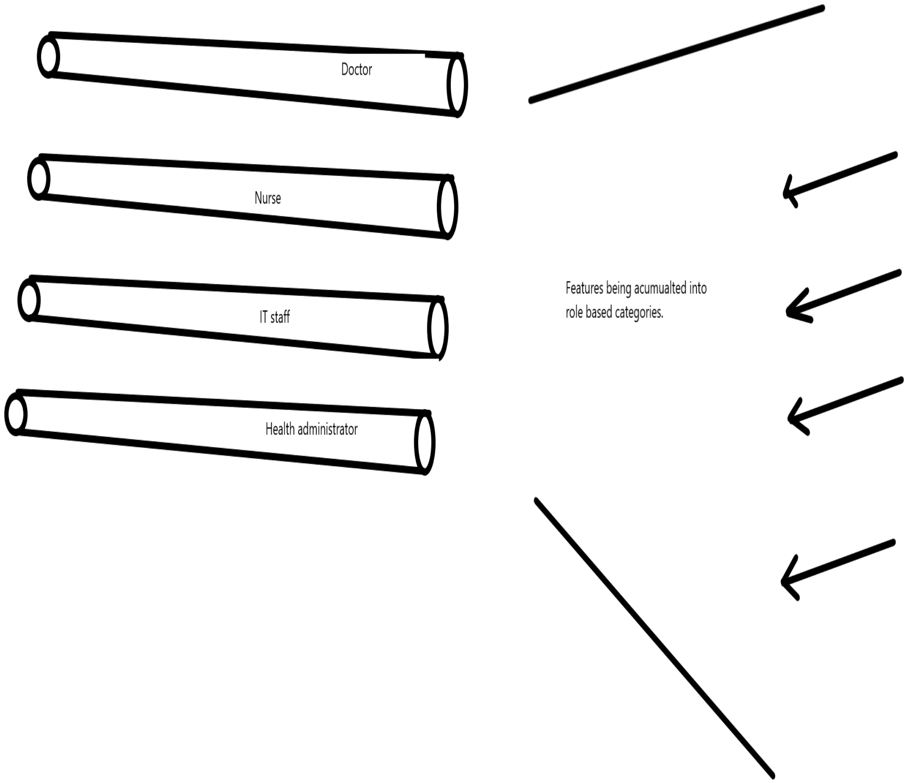
- 24,286 of them are considered normal
- 362 of them are considered an anormal

Role classification model

- classify the cumulative user activity in a single day into one of the 12 categories
- The model was used to classify the cumulative activity of a user in a single

Day

- The model was then trained and validated with cross-fold validation



Anomaly Detection

Hard classification

- we classify each instance into one category (role)
- If an instance is classified into their actual role, then the instance is considered normal.
- If the instance was misclassified, then that instance was considered abnormal practice

Soft classification

- It gives tolerance for the user to act like users from other roles because some roles have quite similar activities.
- The classifier compute the probability of the user’s instance belong to their role class.
- If the probability is above a particular threshold, then it is considered normal.
- Otherwise, it will be considered an anomaly

Algorithms compared

- Multinomial Naive Bayes(multnb),
- Bernoulli Naive Bayes (bernnb),
- Support Vector Machine (svm),
- Neural Network (nn),
- K-Nearest Neighbours(knn),
- Logistic Regression (lr),
- Random Forest (rf),
- Decision Tree (dt).

Performance measures

Actual \ Predicted	Anomaly	Normal
Anomaly	TP	FN
Normal	FP	TN

$$Acc = \frac{TP + TN}{TP + FP + FN + TN}$$

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F_1 = 2\frac{P \cdot R}{P + R}$$

Precision, p: Number of instances that are labeled as anormaly,how many are actually anomaly?

Recall, R: Number of all instances that are actually anomaly, how many of those are correctly predicted

F1 Mesure: is the harmonic mean(average) of the precision and recall

Findings

- Hard classification None normalized

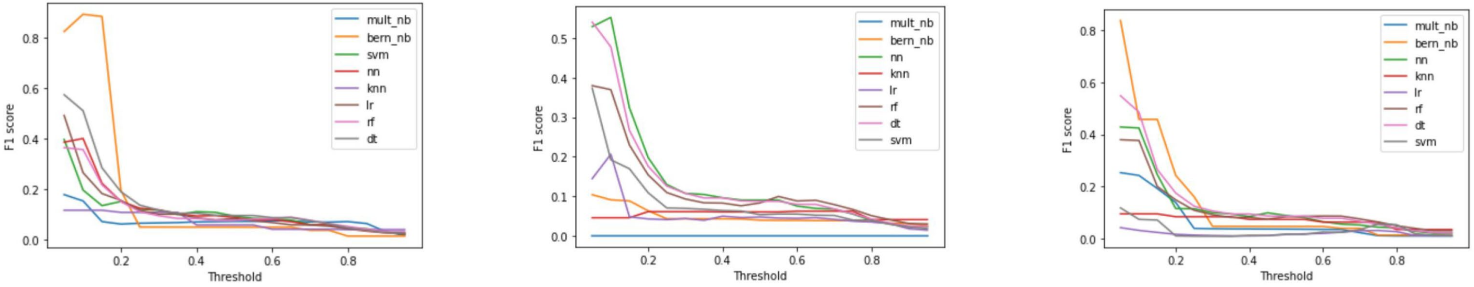
Method	Acc	Prec	Rec	F1
multnb	0.880	0.037	0.698	0.071
bernnb	0.776	0.025	0.868	0.048
nn	0.909	0.045	0.642	0.084
knn	0.873	0.030	0.585	0.057
lr	0.891	0.046	0.792	0.087
rf	0.913	0.041	0.547	0.076
dt	0.913	0.050	0.679	0.093
svm	0.909	0.046	0.660	0.086

Soft classification (F1- score)

Method	None Normalised data	Normalized data (Z-score)	Normalized data (Min-Max)
multnb	0.152	-	0.243
bernnb	0.893	0.091	0.457
nn	0.208	0.548	0.214
knn	0.375	0.046	0.095
lr	0.115	0.206	0.032
rf	0.264	0.377	0.355
dt	0.383	0.482	0.485
svm	0.507	0.184	0.075

NTNU

Soft classification (F1- score)…



(a) Result on None Normalised Data (b) Result on Normalized Data (Z-Score) (c) Result on Normalized Data (Min-Max)

Fig. 8: F1-score of Anomaly Detection using Soft Classification

Performance

- Generally, the Soft Classification approach achieved better performance than the Hard Classification approach
- Bernoulli Naive Bayes on the None Normalised data performed better with an F1-score of 0.893.

All of the methods obtained a high recall and accuracy but low precision and F1-score.

This high recall means that the method from this work can be a good tool to narrow down the data for further manual investigation.

Soft Classification approach performed better than the Hard Classification approach because it provides some tolerances as roles in different activities can be very similar.

Future works

- future works on further processing the anomalies to detect malicious activities.
- Additional, as labeled real data can be difficult to get, it is also important to compare unsupervised methods for the detection of anomalies and maliciousness in the context of big data.

Thank You For Your Attention!

Questions?

PROSPER K. YENG

PhD. Candidate

Norwegian University of Science and Technology

E-mail: Prosper.Yeng@ntnu.no