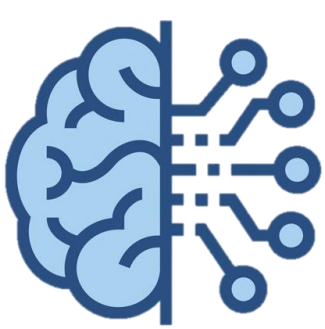
UCLM
UNIVERSIDAD DE CASTILLA-LA MANCHA

Escuela Superior de Informática

Task 3: Predictive Models

Machine Learning Techniques



Pablo Alcázar Morales

Diego Pedregal Hidalgo

Alberto Velasco Mata

First Iteration

Feature Selection

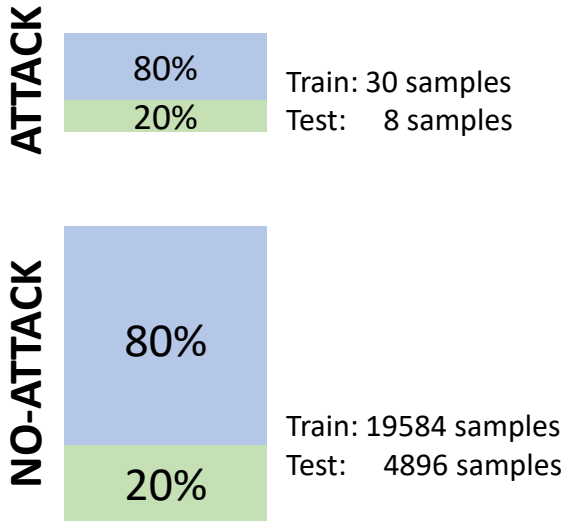
No clue on how physical sensors are affected on attacks

Initially, **all "meaningful" features selected**

UserID
UUID
Version
TimeStemp

Gyroscope (X mean, Z mean, z-x cov, z-y cov)
Magnetic Field (X mean, Z mean, z-x cov, z-y cov)
Pressure (mean)
Linear Acceleration (X mean, Z mean, z-x cov, z-y cov)

attack

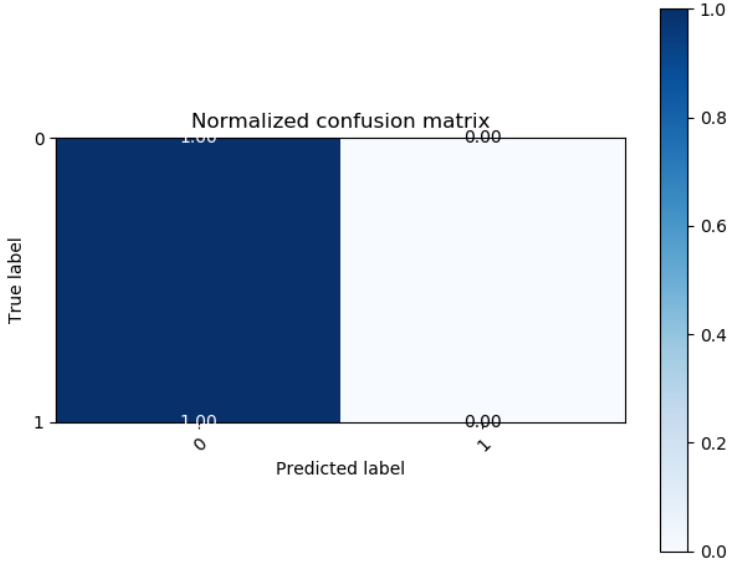


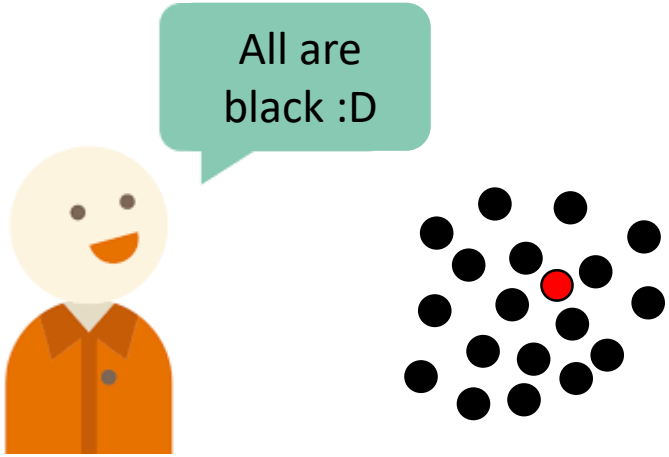
First Iteration

Naïve Bayes

Tried different classifier models

All give really bad results due to **unbalanced data**



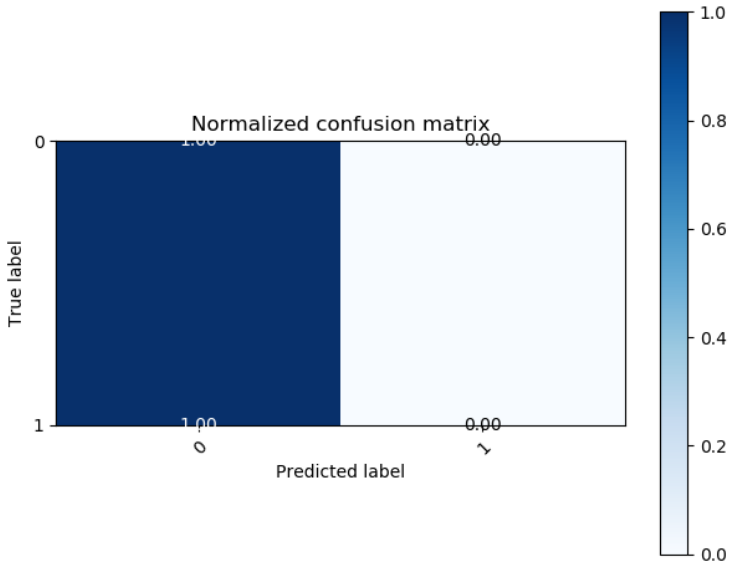


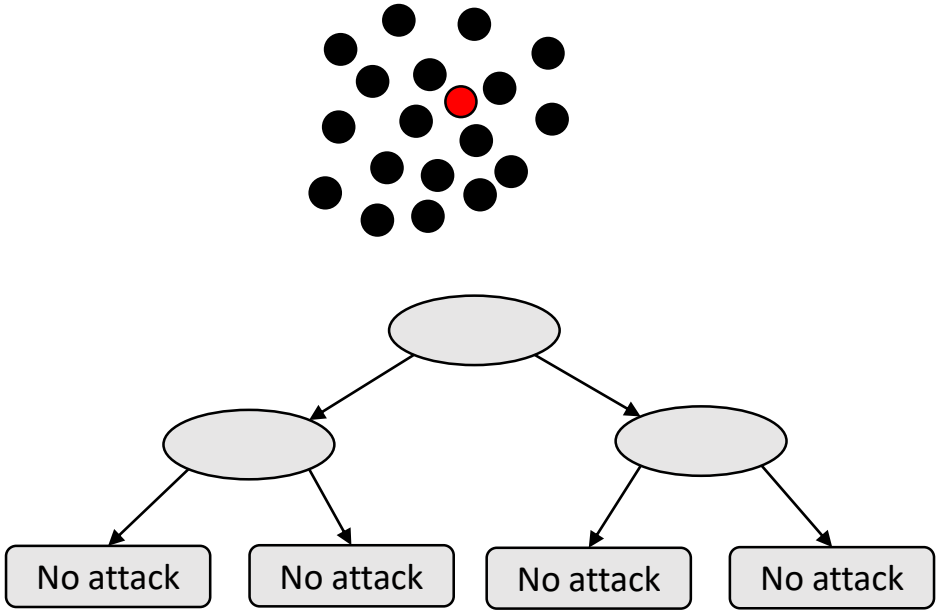
Accuracy: 99.84%

First Iteration

Decision Trees

Unbalanced data leads to practically the **same result**



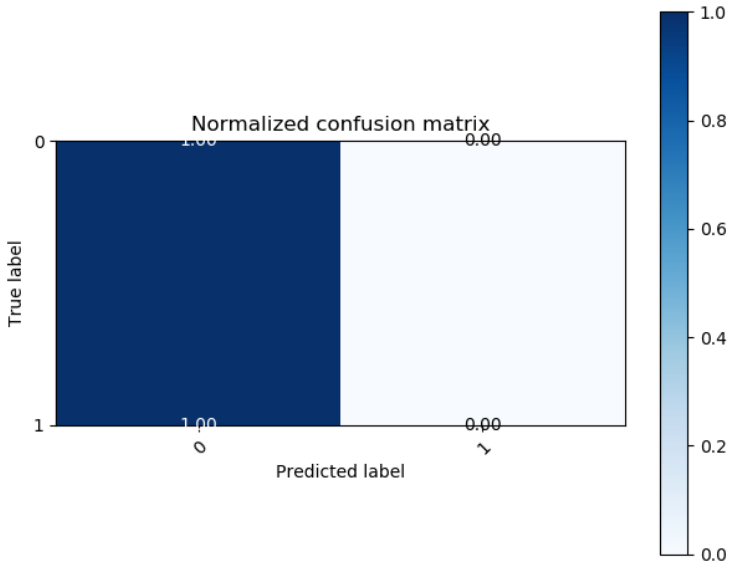


Accuracy: 99.84%

First Iteration

Random Forests

Once again…



It's time to

BALANCE

Accuracy: 99.84%

Second Iteration
Improvement

Balance Dataset

Clustering 0s

K-Means: $\sqrt{20000} \approx 141$ → group 0s in 141 clusters

- New dataset: 141 no-attacks (cluster centers) and (initial) 30 attacks

Better results with this one

Random selection of 0s

Downsample no-attacks by randomly selecting samples

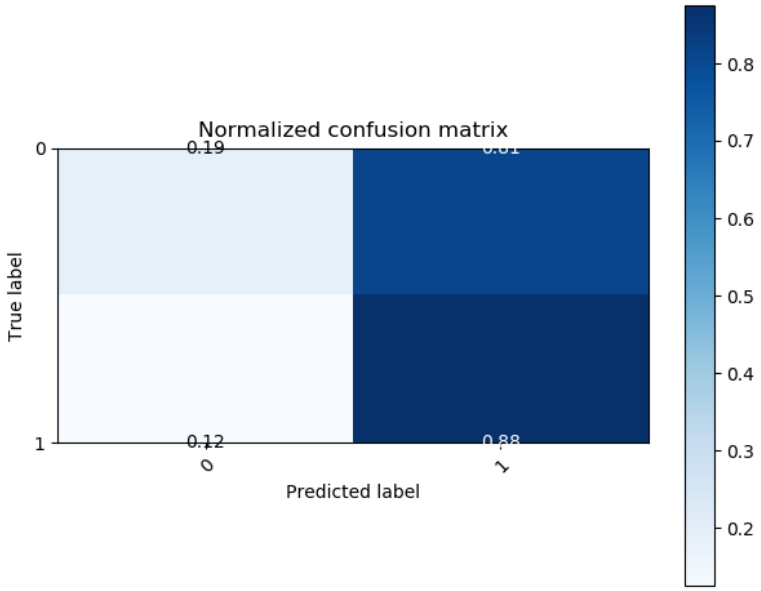
- New dataset: ~200 no-attacks and 30 attacks

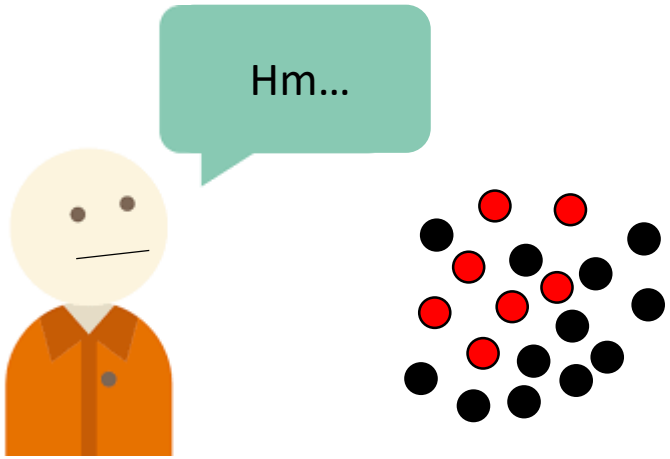
Second Iteration

Naïve Bayes

Tends to predict attacks

However, it seems like balancing the data did its job

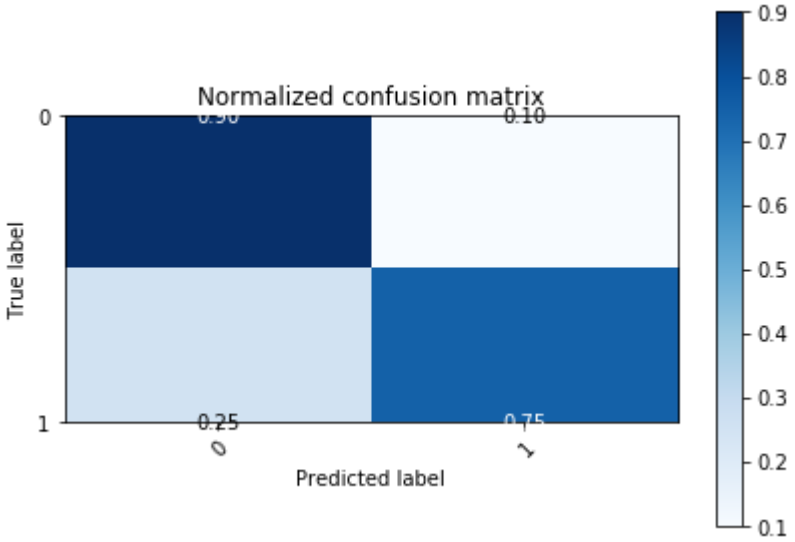




Accuracy: 18.8%

Second Iteration

Decision Trees



75% of attacks in test are detected

In this case it might be better to have **false positives** rather than **false negatives**

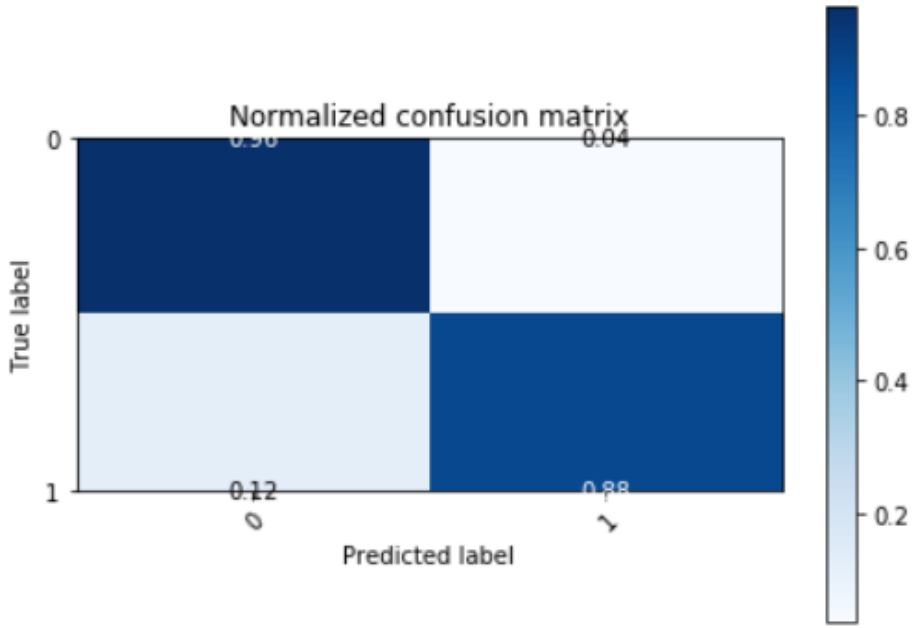
Maybe Random Forests can do better (as they usually do…)

Accuracy: 90.07%

Second Iteration

Random Forests

They do!



		Predicted	
		0	1
Actual	0	4716	180
	1	1	7

Accuracy: 96.0%

Observations

- **Balancing** the dataset was the key
- ***Random Forest*** gives the best results
- However, it might be interesting to **obtain more false positives** if that leads to **detect all attacks**
- We could try to **select those features RF gives more relevance to** (third iteration focusing on feature selection)