DATASETS FOR NETWORK SECURITY

Setinský Jiří, xsetin00@stud.fit.vutbr.cz

Excel@FIT 2023

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

MOTIVATION

- Effectively reduce datasets
- Quicker learning process
- Condense fundamental data for learning
- Periodically refresh datasets with new data
- Sustain minimal dataset

FRAMEWORK DESIGN

Figure 1: Report generation.

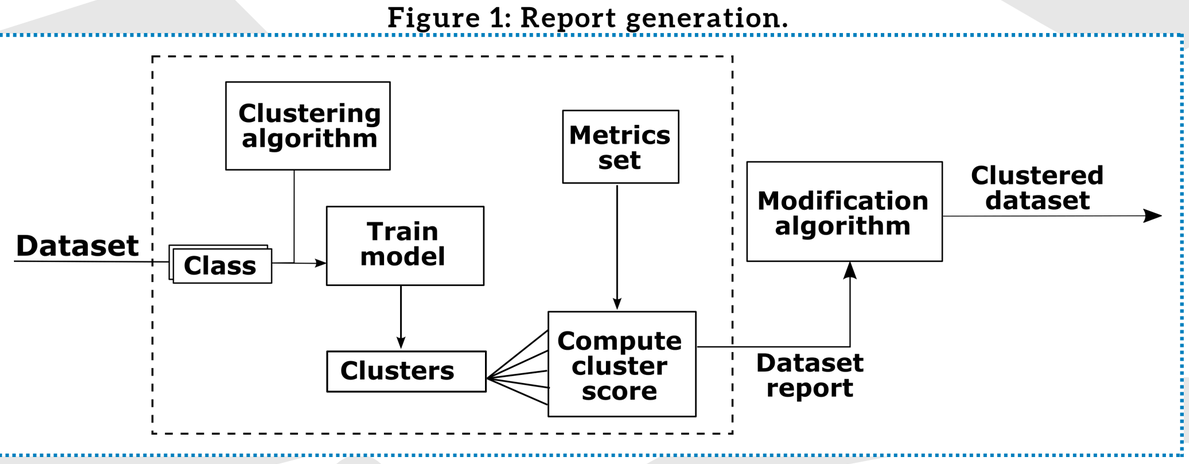
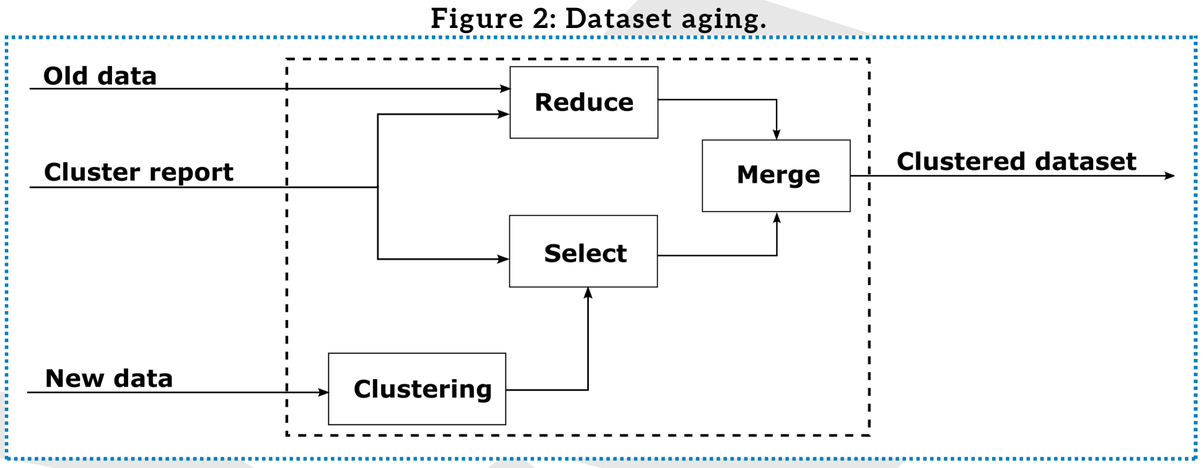


Figure 2: Dataset aging.



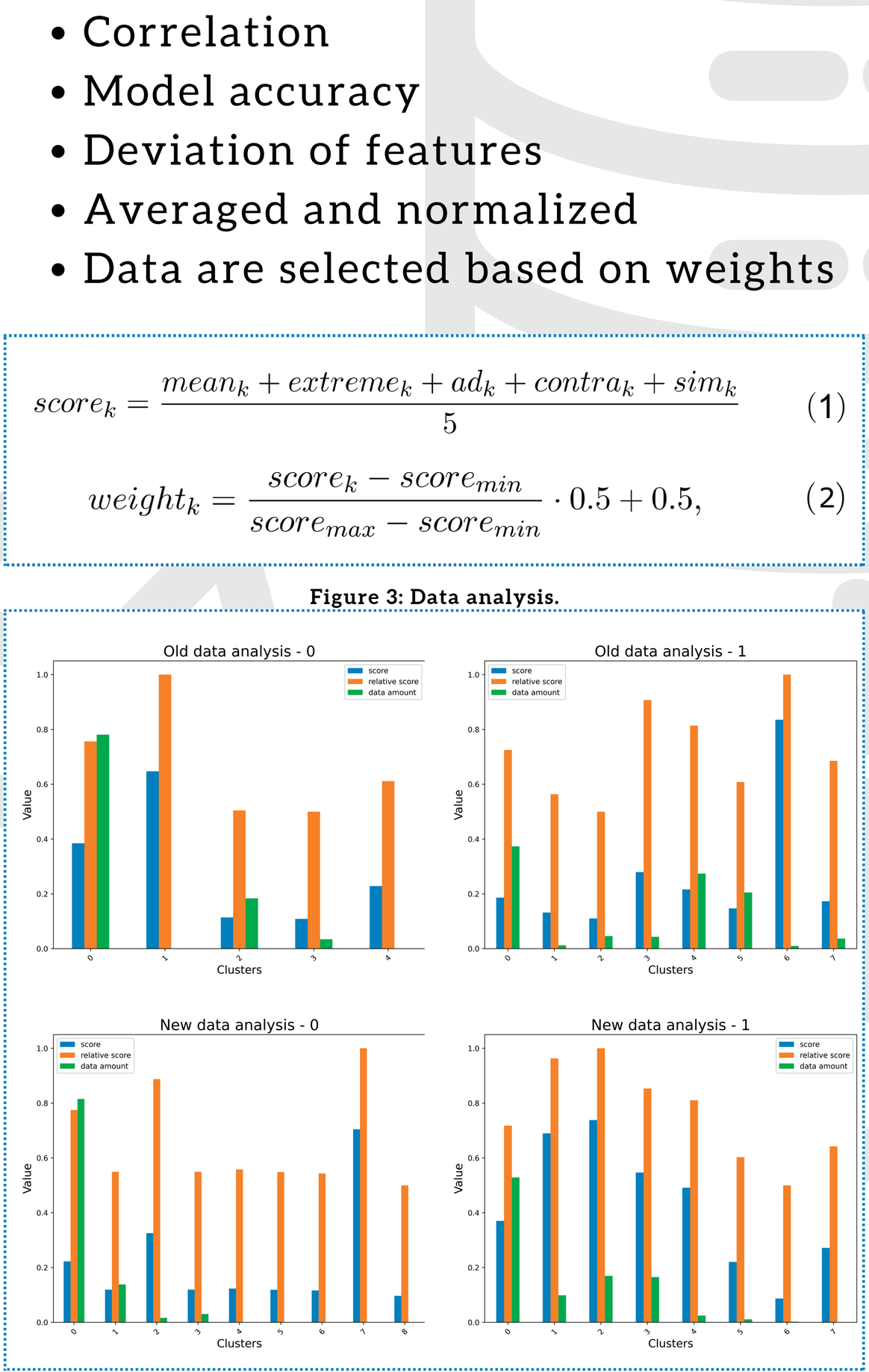
REPORT COMPUTATION

- Correlation
- Model accuracy
- Deviation of features
- Averaged and normalized
- Data are selected based on weights

$$score_k = \frac{mean_k + extreme_k + ad_k + contra_k + sim_k}{5} \quad (1)$$

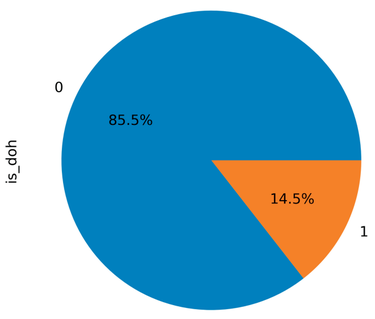
$$weight_k = \frac{score_k - score_{min}}{score_{max} - score_{min}} \cdot 0.5 + 0.5, \quad (2)$$

Figure 3: Data analysis.



RESULTS

Old dataset balance: 1 581 271



New dataset balance: 2 055 955

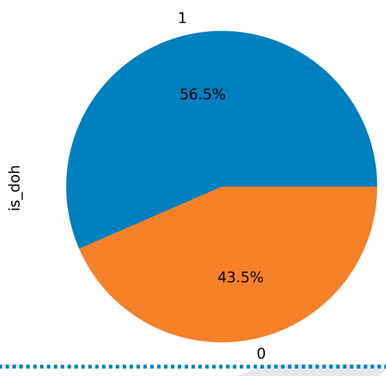


Table 3: Old model

	precision	recall	f1-score	support
negative	0.7043	0.9685	0.8155	1 497 558
positive	0.8711	0.3432	0.4924	927 262
accuracy	0.7294	0.7294	**0.7294**	2 424 820
macro avg	0.7877	0.6559	**0.6540**	2 424 820
weighted avg	0.7681	0.7294	**0.6920**	2 424 820

Table 4: New model

	precision	recall	f1-score	support
negative	0.9341	0.8189	0.8727	1 497 558
positive	0.7561	0.9068	0.8246	927 262
accuracy	0.8525	0.8525	**0.8525**	2 424 820
macro avg	0.8451	0.8628	**0.8487**	2 424 820
weighted avg	0.8661	0.8525	**0.8543**	2 424 820

Clustered dataset balance: 1 581 269

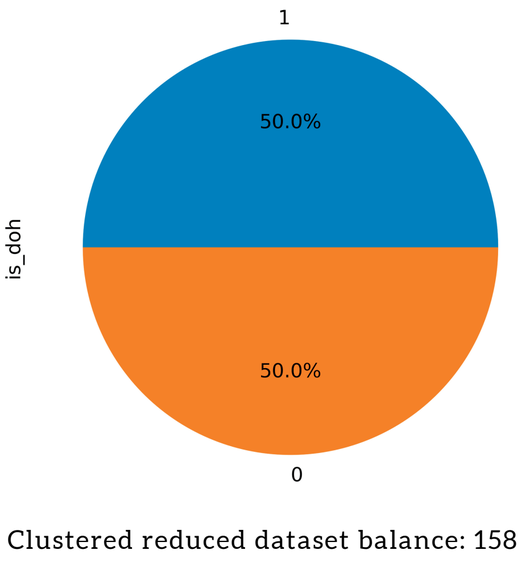


Table 5: Random sampled model

	precision	recall	f1-score	support
negative	0.9671	0.9181	0.9420	1 497 558
positive	0.8778	0.9495	0.9122	927 262
accuracy	0.9301	0.9301	**0.9301**	2 424 820
macro avg	0.9224	0.9338	**0.9271**	2 424 820
weighted avg	0.9329	0.9301	**0.9306**	2 424 820

Table 6: Clustered model

	precision	recall	f1-score	support
negative	0.9673	0.9223	0.9442	1 497 558
positive	0.8832	0.9496	0.9152	927 262
accuracy	0.9327	0.9327	**0.9327**	2 424 820
macro avg	0.9252	0.9359	**0.9297**	2 424 820
weighted avg	0.9351	0.9327	**0.9331**	2 424 820

Clustered reduced dataset balance: 158 125

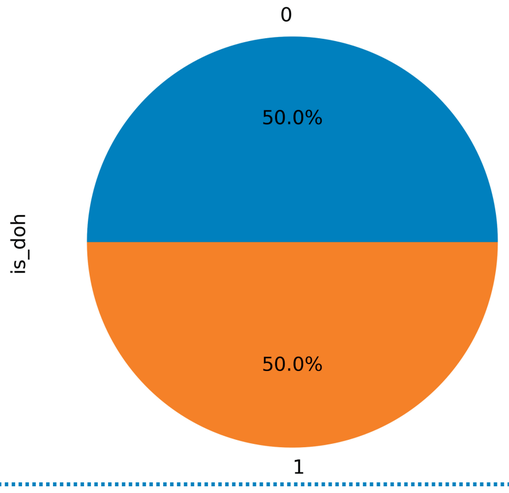


Table 1: Random sampled model reduced

	precision	recall	f1-score	support
negative	0.9634	0.9214	0.9419	1 497 558
positive	0.8814	0.9435	0.9114	927 262
accuracy	0.9298	0.9298	**0.9298**	2 424 820
macro avg	0.9224	0.9324	**0.9266**	2 424 820
weighted avg	0.9320	0.9298	**0.9302**	2 424 820

Table 2: Clustered model reduced

	precision	recall	f1-score	support
negative	0.9704	0.9307	0.9501	1 497 558
positive	0.8950	0.9541	0.9236	927 262
accuracy	0.9397	0.9397	**0.9397**	2 424 820
macro avg	0.9327	0.9424	**0.9369**	2 424 820
weighted avg	0.9416	0.9397	**0.9400**	2 424 820