



Homeland Security

TEMPORAL IDENTIFICATION OF CRIMINAL PROFILES AND ACTION PHASES FROM COMMUNICATIONS METADATA DURING SURVEILLANCE INVESTIGATIONS

Benchmark vs. Logistic Regression, Random Forest, Boosted Tree & Neural Network

UC#11 – 2025/06 (v6.0)

xtractis.ai

PROBLEM DEFINITION

GOAL Design an AI-based decision system that accurately identifies risky behavior linked to criminal activities by analyzing communication metadata from surveillance investigations, without accessing the content of telephone calls and rationally predicts dangerous Homeland Security situations.

- PROS & BENEFITS**
- ▶ Identify specific metadata characterizing different criminal activities and enhance expert knowledge by helping intelligence specialists understand the causal relationships between the communication profiles and the roles inside criminal organizations.
 - ▶ Help intelligence services detect attacks as early as possible and understand the underlying strategy of the criminals in order to consider measures to thwart future attacks.
 - ▶ Avoid many false alarms thanks to transparent diagnosis.

REFERENCE DATA

Source: Confidential data produced by ATOS-BDS-MCS (EVIDEN)

Variable to Predict The model predicts the type of sender profile [**Banal, Support, Executant, Chief**] and the associated temporal phase phase [**P1 Initialization, P2 Gathering, P3 Planning, P4 Execution**] for a total of 10 feasible combinations (10 possible classes):
BNL | SUP_P2 | SUP_P3 | SUP_P3 | EXEC_P2 | EXEC_P3 | EXEC_P4 | CHIEF_P2 | CHIEF_P3 | CHIEF_P4

Potential Predictors Each communication is described by 29 to 37 metadata. These metadata are combined and aggregated over time to obtain 321 potential predictors [NUM_SMS_2Days: Number of SMS-type communications over the last 2 days, COMVOLUME: Duration of the call in progress..].

Observations 2,492,273 communications within 7 scenarios. Data are divided into a Learning Dataset for model induction using Training, Validation, and Test Datasets, and an External Test Dataset (involving 6 scenarios) to check the top-model's performance on real data and for benchmarking.

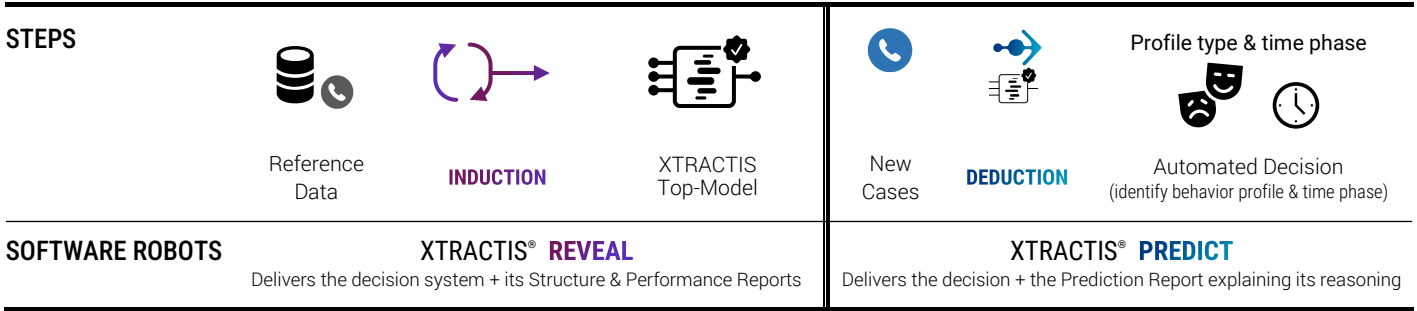
Learning Dataset: 809,554 cases 32.5% (no duplicates)										External Test Dataset: 1,682,719 cases 67.5% (no duplicates)									
Training (434,150 53.63%), Validation (160,399 19.81%), Test (215,005 26.56%)																			
BNL	SUP_P2	SUP_P3	SUP_P4	EXEC_P2	EXEC_P3	EXEC_P4	CH_P2	CH_P3	CH_P4	BNL	SUP_P2	SUP_P3	SUP_P4	EXEC_P2	EXEC_P3	EXEC_P4	CH_P2	CH_P3	CH_P4
57.84%	11.83%	0.95%	0.16%	23.15%	2.19%	0.37%	3.17%	0.30%	0.04%	47.10%	16.23%	0.98%	0.12%	28.72%	1.87%	0.37%	4.26%	0.31%	0.04%

MODEL TYPE Regression **Multinomial Classification** Binomial Classification Scoring

XTRACTIS-INDUCED DECISION SYSTEM

- ☑ **Intelligible Model, Explainable Decisions**
 - ▶ The top-model is a decision system composed of 12 gradual rules without chaining.
 - ▶ Each rule uses from 3 to 12 predictors among the 24 variables that XTRACTIS automatically identified as significant (out of the 321 Potential Predictors).
 - ▶ Only a few rules are triggered at a time to compute the decision.
- ☑ **High Predictive Capacity** It has a good Real Performance for all 6 External Test Dataset scenarios (on unknown data).
- ☑ **Ready to Deploy** It computes real-time predictions up to 70,000 decisions/second, offline or online (API).

XTRACTIS PROCESS



TOP-MODEL INDUCTION

INDUCTION PARAMETERS

Powered by:



- We launch 464 inductive reasoning strategies; each strategy is applied to the same single partition of the learning dataset (53.6% Training / 19.8% Validation / 26.6% Test) to get a reliable assessment of the descriptive and predictive performances, respectively from Training and Validation Datasets.
- Each strategy thus generates one unitary model called **Individual Virtual Expert (IVE)**.
- Among the 464 induced models, the top-IVE selected is the one that has the best predictive performance, close to its descriptive performance, and with the best intelligibility, i.e., with the fewer predictors and rules.

Total number of induced unitary models

464 IVEs

Criterion for the induction optimization

Average F₂-Score

Validation criterion for the top-model selection

Average F₂-Score

Duration of the process @ Induction Speed FP64

35 days @ 24.01 Tflops

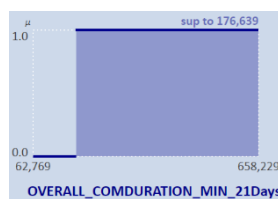
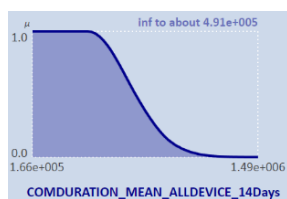
TOP-MODEL STRUCTURE

The top-IVE has a very good intelligibility as it has **12 rules** combining **24 predictors**, with 6.3 predictors per rule on average.

Its Structure Report reveals all the internal logic of the decision system and ensures that the model is understandable by the human expert. It is a transparent model that can be audited and certified before deployment to end-users.

PREDICTORS

- 24 continuous metadata (out of 321)
- Ranked by impact significance (7 strong, 4 medium & 13 weak signals):
 #1 `NUM_UNIQUE_USED_DEVICE_SMS_14Days .../.../`
 #11 `VARPRC_OVERALL_NUM_VOICE_14_21Days ...`
- Labeled by fuzzy and binary classes
 Examples: **binary interval** "sup to 176,639";
fuzzy interval "inf to about 4.91e+005"



RULES

- 12 connective fuzzy rules without chaining (aggregated into 10 disjunctive fuzzy rules)
- 3 to 12 predictors per rule (on average, 6.3 predictors per rule)
- Example: **fuzzy rule R6** uses 4 predictors and concludes "EXEC_P2" (Profile Executant, Phase Grouping). 11 other fuzzy rules complete this model.

```

IF COMDURATION_MEAN_ALLDEVICE_14D IS inf to ~4.91e+005
AND NUM_UNIQUE_USED_DEVICE_1D IS sup to ~1.99
AND OVERALL_COMDURATION_MIN_21D IS sup to 176,639
AND VARPRC_OVERALL_NUM_SMS_3_7D IS sup to ~-57.0
THEN Sender_Profile_Phase IS EXEC_P2
    
```

Literally, intercepted communication is that of an EXECUTANT in Gathering Phase if his mean duration of communication, during the last 14 days, is inferior to 8'11", and his number of different devices used, during the last 24h, is 2 or more, and the group's minimum duration of communication, during the last 21 days, is superior to 2'57", and the relative change in group's number of SMS, between the last 3 and the last 7 days, is superior to -57%.

TOP-MODEL PERFORMANCE

The top-IVE performances, measured in Training / Validation / Test, then in External Test on reference data for each of the 6 scenarios, guarantee the model's predictive and real performances.

Performance Type
Dataset
Average F₂-Score
Classification Error

DESCRIPTIVE
53.6% Training
92.37%
1.41%

PREDICTIVE
19.8% Validation
90.40%
1.56%

REAL
26.6% Test
89.82%
1.06%

REAL
External Test
87.23%
0.64%

EXPLAINED PREDICTIONS FOR 2 UNKNOWN CASES

CASE

(from the External Test Dataset, i.e., not included in the Learning Dataset)

ihfgwmqida_2014-05-23
16:17:47.166



actual value = CHIEF_P4

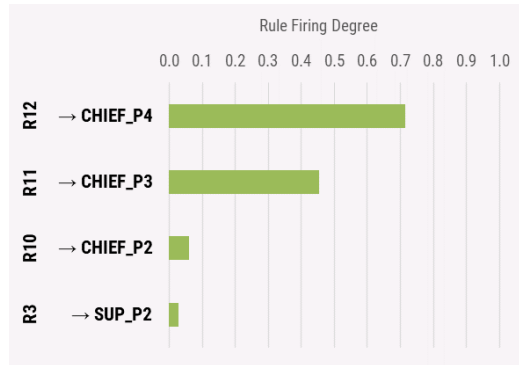
COMDURATION_MEAN_7D	1.83e+005
NUM_UNIQUE_TMSI_RECEIVER_SMS_3D	Missing Value
NUM_UNIQUE_USED_DEVICE_SMS_14D	6.00
NUM_VOICE_ALLDEVICE_1D	4.0
...	...
OVERALL_COMDURATION_MIN_21D	80,186
VARPRC_NUM_UNIQUE_TMSI_RECEIVER_1_2D	0.0
VARPRC_NUM_VOICE_ALLDEVICE_7_14D	-61.2
VARPRC_OVERALL_NUM_SMS_1_2D	-49.7
VARPRC_OVERALL_NUM_VOICE_7_14D	-50.7

DEDUCTIVE INFERENCE OF RULES

For this communication, 4 rules are triggered:

R12 at 0.715, R11 at 0.453, R10 at 0.061 and R3 at 0.027

The 8 other rules are not activated.



AUTOMATED DECISION

NUMBER OF TRIGGERED RULES

4 / 12

FUZZY PREDICTION

{ CHIEF_P4 | 0.715,
CHIEF_P3 | 0.453,
CHIEF_P2 | 0.061,
SUP_P2 | 0.027 }

FINAL PREDICTION

{ CHIEF_P4 }

The system delivers the correct diagnosis compared to that given by the intelligence expert:

Profile CHIEF,
Phase EXECUTION



wavziguktqy_2013-09-24
21:54:39.903



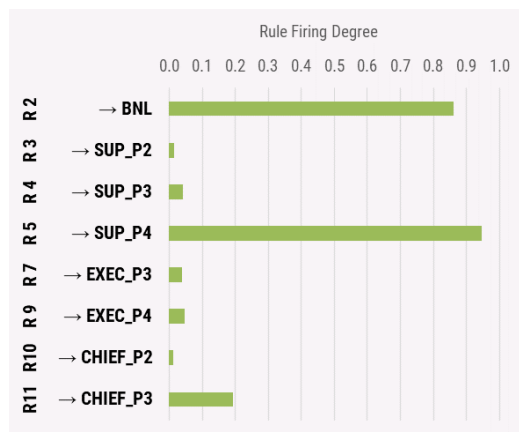
actual value = SUP_P4

COMDURATION_MEAN_7D	3.74E+05
NUM_UNIQUE_TMSI_RECEIVER_SMS_3D	Missing Value
NUM_UNIQUE_USED_DEVICE_SMS_14D	2.00
NUM_VOICE_ALLDEVICE_1D	2.0
...	...
OVERALL_COMDURATION_MIN_21D	64,285
VARPRC_NUM_UNIQUE_TMSI_RECEIVER_1_2D	0.0
VARPRC_NUM_VOICE_ALLDEVICE_7_14D	-33.3
VARPRC_OVERALL_NUM_SMS_1_2D	-46.6
VARPRC_OVERALL_NUM_VOICE_7_14D	-50.6

For this communication, 8 rules are triggered:

R5 at 0.946, R2 at 0.860, R11 at 0.194...

The 4 other rules are not activated.



NUMBER OF TRIGGERED RULES

8 / 12

FUZZY PREDICTION

{ SUP_P4 | 0.946
BNL | 0.860,
CHIEF_P3 | 0.194,
EXEC_P4 | 0.048,
SUP_P3 | 0.041
EXEC_P3 | 0.039
SUP_P2 | 0.014
CHIEF_P2 | 0.013 }

FINAL PREDICTION


{ SUP_P4 }

The system delivers the correct diagnosis compared to that given by the intelligence expert, although it considered that it could also be a Banal behavior with a closer possibility:

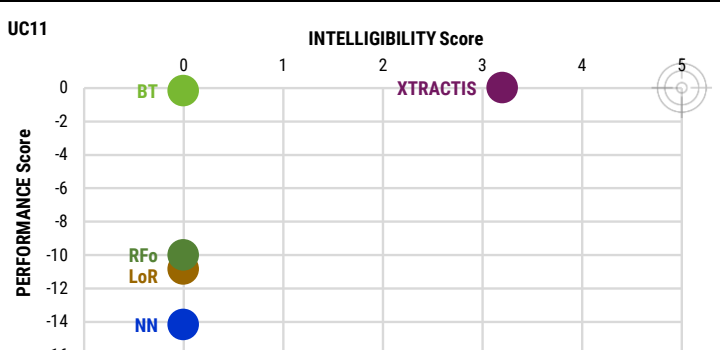
Profile SUPPORT,
Phase EXECUTION



TOP-MODELS BENCHMARK

	XTRACTIS 	LOGISTIC REGRESSION	RANDOM FOREST	BOOSTED TREE	NEURAL NETWORK	
MODELING PARAMETERS	MODELS RELEASE	2023/01	2023/01	2023/01	2023/01	
	ALGORITHM VERSION	XTRACTIS REVEAL 12.2.44349	Python 3.9 Scikit-Learn 1.3.0	Python 3.9 LightGBM 3.3.2	Python 3.9 LightGBM 3.3.2	Python 3.9 TensorFlow 2.10.0 Keras 2.10.0
	CROSS-VALIDATION TECHNIQUE	All explored strategies for all algorithms use the same single-split of the Learning Dataset: 60% Training 20% Validation 20% Test				
	NUMBER OF EXPLORED STRATEGIES⁽¹⁾	464 induction strategies	1,000 data analysis strategies	1,000 ML strategies	1,000 ML strategies	1,000 ML strategies
	TOP-MODEL SELECTION⁽²⁾	Top-IVE among 464 IVEs	Top-IVE among 1,000 IVEs	Top-IVE among 1,000 IVEs	Top-IVE among 1,000 IVEs	Top-IVE among 1,000 IVEs

TOP-MODEL STRUCTURE	NUMBER OF PREDICTORS (out of 321 Potential Predictors)	24	321	299	313	321
	AVERAGE NUMBER OF PREDICTORS PER RULE OR EQUATION	6.3 per rule	321.0 per equation	8.3 per rule	6.1 per rule	117.6 per equation
	STRUCTURE OF THE DECISION SYSTEM	12 fuzzy rules without chaining (aggregated into 10 disjunctive fuzzy rules) Only a few rules are triggered at a time to compute a decision	10 linear equations	500 trees without chaining 20,216 binary rules	10 chains of 309 trees each 49,797 binary rules Tree #N corrects the error of the N-1 previous trees	2 hidden layers 22 hidden nodes 32 equations 22 unintelligible synthetic variables

TOP-MODEL SCORES		Random ⁽³⁾	XTRACTIS	LoR	RFo	BT	NN	UC11 	
	INTELLIGIBILITY Score⁽⁴⁾			3.20	0.00	0.00	0.00		0.00
	IVE Real Perf. (Average F ₂ -Score) in Test			89.82	78.89	77.83	89.54		84.20
Gap to Leader in Test			0.00	-10.93	-11.99	-0.28	-5.62		
IVE Real Perf. (Average F ₂ -Score) in External Test	7.79%		87.23	76.46	79.19	87.14	64.47		
Gap to Leader in External Test			0.00	-10.77	-8.04	-0.09	-22.76		
IVE Average Real Performance			88.53	77.68	78.51	88.34	74.34		
PERFORMANCE Score⁽⁴⁾			0.00	-10.85	-10.02	-0.18	-14.19		

(1) For all algos: on the same Learning Dataset. All Models are optimized according to their Validation Average F₂-Score.

(2) All top-models are selected according to their Validation Average F₂-Score while checking that it remains close to their Training Average F₂-Score.

(3) Baseline performances that models must exceed to perform better than chance (P-value = 0.001; 100,000 models generated by random permutation of the output values). The value of each performance criterion is generally achieved by a different random model.

(4) See Appendices for explanations and detailed results. Performance Scores are calculated on all available unknown data.

More Use Cases:
xtractis.ai/use-cases/

APPENDIX 1 – Calculation of the Intelligibility × Performance

AI Technique #i	T _i	i ∈ [1 ; n] n = number of AI Techniques benchmarked in terms of data-driven modeling = 5
Benchmark #k	B _k	k ∈ [1 ; p] p = number of Benchmarks for the Use Case ∈ {1, 2, 3}

Remarks:

- In case of a small number of reference data, a CVE model (College of Virtual Experts) is generated by each explored strategy of T_i, generally via an N×K-fold cross validation. In this case, a Benchmark is led with the top-CVE on the External Test Dataset (ETD, composed of unknown reference cases). Then, a top-IVE model (Individual Virtual Expert) is generated from the top-CVE, through the XTRACTIS® reverse-engineering process, or for the other T_i, by applying the top-strategy, which has generated the top-CVE, on the Training and Validation Datasets. And a second Benchmark is led with this top-IVE on the same ETD.
- In case of a huge number of reference data, an IVE is generated by each explored strategy of T_i, via a 1-split validation. In this case, Benchmarks are led with the top-IVE on the Test Dataset (TD, composed of unknown reference cases) and on the available ETDs.
- Each Benchmark uses the latest versions of the following algorithms available at the date of the benchmark. XTRACTIS®: REVEAL; Logistic Regression: Python, Scikit-Learn; Random Forest & Boosted Tree: Python, LightGBM; Neural Network: Python, TensorFlow, Keras.
- Each B_k uses exactly the same TD and ETD for each T_i model.
- No Regression models can be obtained by Logistic Regression. So, this Data Analysis technique is benchmarked only for Classification or Scoring problems.
- The Holy Grail for critical AI-based decision systems is to obtain a model with the highest Performance and the highest Intelligibility scores (top-right corner of the graph).

PERFORMANCE Score

For each B_k, we calculate the values of the Performance Criterion (PC) on the same ETD for all the T_i top-CVEs; and on the same TD and ETDs for all the T_i top-IVEs. The PC is: RMSE in percentage for a Regression; F₁-Score for a Binomial Classification; Average F₁-Score or Average F₂-Score for a Multinomial Classification; Gini index for a Scoring. Then, we compare the value of the PC of each T_i top-CVE (resp. top-IVE) to the best value of this PC reached by the best T_i top-CVE (resp. top-IVE) on ETD (resp. on TD and ETDs).

For Regression, we calculate for each T_i top-model (CVE and IVE): PS(T_i, B_k) = Best_PC(B_k) - PC(T_i, B_k).

For Classification and Scoring, we calculate for each T_i top-model: PS(T_i, B_k) = PC(T_i, B_k) - Best_PC(B_k).

$$\text{Performance Score of } T_i$$

$$\text{PS}(T_i) = \text{Mean}(\text{PS}(T_i, B_k))_{k \in [1 ; p]}$$

Remark:

- Each PS varies theoretically from -100 (Lowest Score) to 0 (Highest Score), but practically between -50 and 0.

INTELLIGIBILITY Score

We consider the T_i top-IVE. Its Intelligibility Score IS(T_i) is valued from 0.00 to 5.00 regarding the structure of the model: number of predictors, classes, rules, equations, trees, synthetic variables, modalities to predict for classifications (or numeric variables to predict for regressions or scoring). The more compact the model, the higher its IS.

The IS of each T_i is obtained by accumulating the following five penalty values to the ideal IS value of 5.00 (each penalty has a null or a negative value):

- Penalty 1 (logarithmic penalty regarding the number of predictors):

$$\text{Pen1}(T_i) = \min(0, 1 - \log_{10} \text{number of predictors})$$

Examples: Pen1 = 0.00 for up to 10 predictors
Pen1 = -3.00 for 10.000 predictors

- Penalty 2 (linear penalty regarding the average number of rules or equations per variable or modality to predict):

$$\text{Pen2}(T_i) = \min\left(0, \frac{1 - \text{average number of rules or equations per variable or modality to predict}}{40}\right)$$

Examples: Pen2 = 0.00 for 1 rule or equation per variable or modality to predict on average
Pen2 = -3.00 for 121 rules or equations per variable or modality to predict on average

- Penalty 3 (linear penalty regarding the average number of predictors per rule or equation):

$$\text{Pen3}(T_i) = \min\left(0, \frac{9 - 3 \times \text{average number of predictors per rule or equation}}{7}\right)$$

Examples: Pen3 = 0.00 for up to 3.0 predictors per rule or equation on average
Pen3 = -3.00 for 10.0 predictors per rule or equation on average

- Penalty 4 (linear penalty regarding the number of trees per chain, here for BT only):

$$\text{Pen4}(T_i) = \min(0, 1 - \text{number of trees per chain})$$

Examples: Pen4 = 0.00 for 1 tree per chain
Pen4 = -3.00 for 4 trees per chain

- Penalty 5 (maximum penalty due to unintelligibility of synthetic variables, here for NN only):

$$\text{Pen5}(T_i) = -5$$

Intelligibility Score of T_i

$$\text{IS}(T_i) = \max(0.00, 5.00 + (\text{Pen1} + \text{Pen2} + \text{Pen3} + \text{Pen4} + \text{Pen5}))$$

Remarks:

- For the difference between the Intelligibility and the Explainability of a model, please see the XTRACTIS® Brochure, page 7.
- The real complexity of the process/phenomenon under study is intrinsic, i.e., it could not be reduced or simplified, but only discovered; thus, the top-model will be complex if the process/phenomenon turns out to be complex [Zalila 2017]. Consequently, for some complex process/phenomenon, IS can be equal to 3.00 or less, even if T_i natively produces intelligible models (XTRACTIS, Random Forest).
- For similar structures, the Boosted Tree model is always less intelligible than the Random Forest one, as it is composed of chains of trees, instead of a college of trees (see Penalty 4).
- Neural Network model has always the lowest IS of 0.00, because it uses synthetic unintelligible variables (hidden nodes) in addition to all the potential predictors (see Penalty 5).

APPENDIX 2 – Use Case Results (all Performance criteria of all Top-Models)

Performance Criterion	Classification Error	Min. Sensitivity	Average Sensitivity	Min. PPV	Average PPV	Average F₂-Score	Weighted Av. F ₂ -Score	Refusal
RANDOM MODEL								
<i>Number of Random Permutations (P-value) = 100,000 (0.001)</i>								
<i>Performance against chance</i>	81.55%	0.37%	7.79%	0.37%	7.79%	7.79%	18.45%	
XTRACTIS TOP-MODEL								
Descriptive Performance (Training)	1.41%	74.22%	93.01%	56.22%	92.07%	92.37%	98.59%	0 (0.00%)
Predictive Performance (Validation)	1.56%	51.21%	89.63%	79.06%	94.84%	90.40%	98.42%	0 (0.00%)
Real Performance (Test)	1.06%	60.97%	90.75%	49.44%	90.87%	89.82%	98.92%	2 (0.00%)
Real Performance (External Test for 6 scenarios)	0.64%	46.37%	87.48%	54.63%	87.85%	87.23%	99.36%	0 (0.00%)
LOGISTIC REGRESSION TOP-MODEL								
Descriptive Performance (Training)	0.91%	76.09%	93.11%	91.11%	97.84%	93.93%	99.09%	
Predictive Performance (Validation)	4.21%	54.57%	82.36%	29.04%	82.37%	80.44%	95.74%	
Real Performance (Test)	2.45%	44.69%	84.64%	22.34%	81.36%	78.89%	97.52%	
Real Performance (External Test for 6 scenarios)	8.11%	44.91%	80.62%	26.85%	80.06%	76.46%	91.75%	
RANDOM FOREST TOP-MODEL								
Descriptive Performance (Training)	0.31%	93.49%	98.29%	80.48%	95.06%	97.58%	99.70%	
Predictive Performance (Validation)	8.42%	30.47%	80.66%	24.15%	73.28%	76.35%	91.82%	
Real Performance (Test)	7.18%	17.76%	85.56%	11.63%	68.95%	77.83%	93.14%	
Real Performance (External Test for 6 scenarios)	12.20%	55.29%	86.68%	20.55%	68.28%	79.19%	88.29%	
BOOSTED TREE TOP-MODEL								
Descriptive Performance (Training)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
Predictive Performance (Validation)	2.92%	30.28%	84.92%	66.63%	91.97%	84.91%	96.97%	
Real Performance (Test)	2.08%	42.36%	89.65%	59.03%	94.12%	89.54%	97.80%	
Real Performance (External Test for 6 scenarios)	3.93%	49.74%	87.26%	57.03%	92.97%	87.14%	95.97%	
NEURAL NETWORK TOP-MODEL								
Descriptive Performance (Training)	0.64%	78.52%	95.67%	82.20%	95.59%	95.55%	99.35%	
Predictive Performance (Validation)	2.90%	56.64%	87.76%	77.30%	89.45%	87.83%	97.06%	
Real Performance (Test)	1.81%	61.68%	87.57%	33.50%	82.59%	84.20%	98.16%	
Real Performance (External Test for 6 scenarios)	18.09%	23.19%	69.84%	32.24%	70.28%	64.47%	81.43%	

The entirety of this document is protected by copyright. All rights are reserved, particularly the rights of reproduction and distribution. Quotations from any part of the document must necessarily include the following reference:
 Zalila, Z., Idagrai Labs & Xtractis (2019-2025). XTRACTIS® the Reasoning AI for Trusted Decisions. Use Case #11 | Homeland Security: Temporal Identification of Criminal Profiles and Action Phases from Communications Metadata during Surveillance Investigations – Benchmark vs. Logistic Regression, Random Forest, Boosted Tree & Neural Network. IDAGRAI LABS, June 2025, v6.0, Compiègne, France, 6p.