Precision Medicine

BLOOD TEST-BASED OPPORTUNISTIC DIAGNOSIS OF DIABETES

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

UC#31 – 2025/07 (v1.1) xtractis.ai

PROBLEM DEFINITION

GOAL

Design an AI-based decision system that accurately establishes a rational and explainable medical diagnosis of diabetes based on the patient's medical characteristics and blood test results.

PROS & BENEFITS

- Identify the parameters actually involved in opportunistic diagnosis of diabetes, i.e. even if the patient is not fasting.
- Enhance medical knowledge by helping healthcare professionals understand the causal relationships between these parameters, their combination, and the disease.
- Help the medical profession identify patients who may be at risk of having diabetes through rapid, systematic and explainable diagnoses, in order to request additional tests reinforcing the diagnosis, if necessary.

REFERENCE DATA

Source:
Mohammed Mustafa

Dataset:
https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

Modeling work carried out by engineering students Yazan El Mahmoud and Amen Allah Nticha (Spring semester 2024) with the UTC-dedicated XTRACTIS® platform

Variable to Predict

The model predicts the patient's condition among 2 modalities:

Normal | **Diabetic**

Potential Predictors

8 variables characterize each patient (medical and demographic data):

gender, age, hypertension, heart disease, smoking history,
BMI, HbA1c level, blood glucose level

Observations

100,000 reference cases of patients.
Each case is associated with a normal condition or a diabetic one.

80,000 cases compose a Learning Dataset for model induction using Training and Validation Datasets. 20,000 cases compose an External Test Dataset to check the top-model's performance on real unknown data and for benchmarking.

Learning Dataset: 80,000 patients | 80.00% 85% for Training, 15% for Validation	
NORMAL	DIABETIC
73,200 | 91.5%	6,800 | 8.5%

External Test Dataset: 20,000 patients | 20.00%	
NORMAL	DIABETIC
18,300 | 91.5%	1,700 | 8.5%

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 rules without chaining.
- Each rule uses from 1 to 5 predictors among the 8 variables that XTRACTIS automatically identified as significant (out of the 8 Potential Predictors).
- Only a few rules are triggered at a time to compute the decision.

☑ **High Predictive Capacity**

It has a good Real Performance (on unknown data). But some other predictors seem missing.

☑ **Ready to Deploy**

It computes real-time predictions up to 70,000 decisions/second, offline or online (API).

XTRACTIS PROCESS

STEPS	Reference Data	**INDUCTION**	XTRACTIS Top-Model	New Cases	**DEDUCTION**	**Normal / Diabetic** Automated Decision (diagnose the patient's condition)
SOFTWARE ROBOTS	XTRACTIS® **REVEAL**			XTRACTIS® **PREDICT**		
OUTCOME	the Decision System + its Structure & Performance Reports			the decision + its Explainability Report unfolding the deductive reasoning		

TOP-MODEL INDUCTION

INDUCTION PARAMETERS

Powered by:
XTRACTIS® REVEAL
v13.2.52316

1. We launch 2,000 inductive reasoning strategies; each strategy is applied to the same single partition of the Learning Dataset (85% Training / 15% Validation) to get a reliable assessment of the descriptive and predictive performances, respectively from Training and Validation Datasets.
2. Each strategy thus generates one unitary model called **Individual Virtual Expert** (IVE).
3. Among the 2,000 induced models, the top-IVE selected is the one that has the best predictive performance, and with the best intelligibility, i.e., with the fewer predictors and rules.

Total number of induced unitary models
2,000 IVEs

Criterion for the induction optimization
F_2-Score

Validation criterion for the top-model selection
F_2-Score

Duration of the process @ Induction Speed FP64
4 days @ 1.13 Tflops

Environmental footprint

57.6 kWh	3.23 kg of CO_2

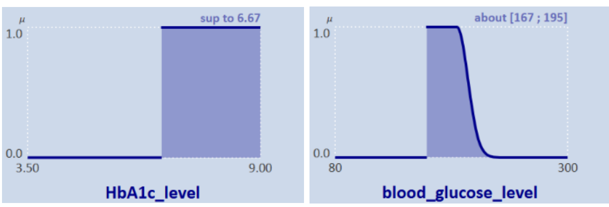
TOP-MODEL STRUCTURE

The top-IVE has an excellent intelligibility (4.88 out of 5): its **12 rules** combine **8 predictors**, with 2.4 predictors per rule on average.

Its Structure Report reveals all the internal logic of the decision system and ensures that the model is understandable. It is a transparent model that can be audited by the expert and certified by the regulator before deploying to end-users, thus AI Act-compliant!

PREDICTORS

- 8 features (out of 8)
- 4 continuous + 4 nominal variables
- Ranked by impact significance (3 strong, 2 medium & 3 weak signals): #1 blood_glucose_level … / #8 gender
- Labeled by fuzzy and binary classes
  Examples: **binary interval** "sup to 6.67";
  fuzzy interval "about [167 ; 195]"



RULES

- 12 connective fuzzy rules without chaining (aggregated into 2 disjunctive fuzzy rules)
- 1 to 5 predictors per rule (on average, 2.4 predictors per rule)
- Example: **fuzzy rule R4** uses 2 predictors to conclude "Normal". 11 other rules complete this model.

IF	age	IS	inf to about 42.9
AND	blood_glucose_level	IS	[167 ; about 195]
THEN	Diagnosis	IS	Normal

Literally, the patient's condition is normal ***if*** *the patient is about under 43 years old* ***and*** *the level of blood glucose level is greater than 167 and less than about 195 mg/dL.*

TOP-MODEL PERFORMANCE

The top-IVE performances, measured in Training/Validation, then in External Test on ETD, guarantee the model's predictive and real performances.

Performance Type Dataset	**DESCRIPTIVE** 85% Training	**PREDICTIVE** 15% Validation	**REAL** External Test
F_2-Score	**80.16%**	**80.06%**	**80.41%**
Classification Error	5.72%	5.70%	5.76%

EXPLAINED PREDICTIONS FOR 2 UNKNOWN CASES

Powered by: XTRACTIS® PREDICT v14.0.55721

CASE

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

DEDUCTIVE INFERENCE OF RULES

REAL-TIME DECISION

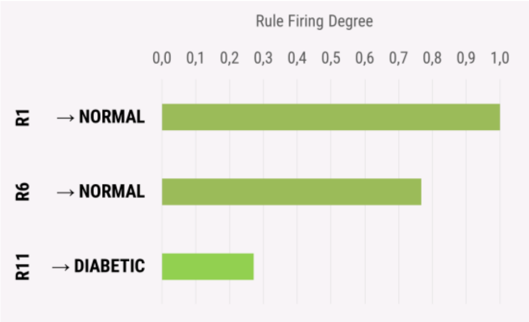
PATIENT #51,143	
age	74.0
bmi	21.2
HbA1c_level	5.70
Blood_glucose_level	158
gender	Female
hypertension	No
heart_disease	No
smoking_history	Former

Actual Value	NORMAL

For this patient, 3 rules are triggered:

R1 at 1.000 and **R6** at 0.767 to conclude NORMAL,
and
R11 at 0.271 to conclude DIABETIC

The 9 other rules are not activated.



NUMBER OF TRIGGERED RULES

3 / 12

FUZZY PREDICTION

{ NORMAL | 1.000,
DIABETIC | 0.271 }

FINAL PREDICTION

{ NORMAL }

The system delivers the correct decision compared to the actual case

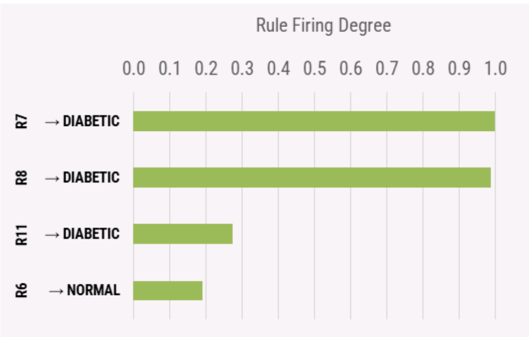
PATIENT #86,892	
age	71.0
bmi	29.0
HbA1c_level	8.20
Blood_glucose_level	220
gender	Male
hypertension	No
heart_disease	No
smoking_history	Current

Actual Value	DIABETIC

For this patient, 4 rules are triggered:

R7 at 1.000, **R8** at 0.988, and **R11** at 0.273
to conclude DIABETIC, and
R6 at 0.191 to conclude NORMAL.

The 8 other rules are not activated.



NUMBER OF TRIGGERED RULES

4 / 12

FUZZY PREDICTION

{ DIABETIC | 1.000,
NORMAL | 0.191 }

FINAL PREDICTION

{ DIABETIC }

The system delivers the correct decision compared to the actual case

TOP-MODELS BENCHMARK

MODELING PARAMETERS

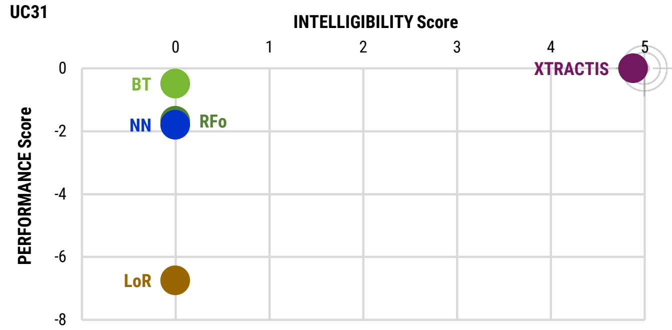
	XTRACTIS	LOGISTIC REGRESSION	RANDOM FOREST	BOOSTED TREE	NEURAL NETWORK
MODELS RELEASE	2024/09	2024/09	2024/09	2024/09	2024/09
ALGORITHM VERSION	XTRACTIS **REVEAL** 13.2.52316	XTRACTIS **BENCHMARK** module embedding Python 3.9.10 | Scikit-Learn 1.3.0 | LightGBM 3.3.2 | TensorFlow 2.10.0 | Keras 2.10.0			
CROSS-VALIDATION TECHNIQUE	All explored strategies for all algorithms use the same single-split of the Learning Dataset: 85% Training | 15% Validation				
NUMBER OF EXPLORED STRATEGIES(1)	2,000 induction strategies	2,000 ML strategies			
TOP-MODEL SELECTION(2)	Top-IVE among 2,000 IVEs	Top-IVE among 2,000 IVEs			

TOP-MODEL STRUCTURE

	XTRACTIS	LOGISTIC REGRESSION	RANDOM FOREST	BOOSTED TREE	NEURAL NETWORK
NUMBER OF PREDICTORS (out of 8 Potential Predictors)	**8**	**15** 2 nominal variables are decomposed into 9 binary variables	**8**	**8**	**15** 2 nominal variables are decomposed into 9 binary variables
AVERAGE NUMBER OF PREDICTORS PER RULE OR EQUATION	**2.4** per rule	**15.0** per equation	**5.5** per rule	**3.8** per rule	**15.0** per equation
STRUCTURE OF THE DECISION SYSTEM	**12** rules without chaining (aggregated into 2 disjunctive fuzzy rules) Only a few rules are triggered at a time to compute a decision	**1** linear equation	**8** trees without chaining **1,416** binary rules	**1** chain of **69** trees **2,500** binary rules Tree #N corrects the error of the N-1 previous trees	**5** hidden layers | **156** hidden nodes **157** equations 156 unintelligible synthetic variables

TOP-MODEL SCORES

	Random(3)	XTRACTIS	LoR	RFo	BT	NN
INTELLIGIBILITY Score(4)		4.88	0.00	0.00	0.00	0.00
IVE Real Perf. (F_2-Score) in External Test	*10.65*	80.41	73.66	78.73	79.92	78.61
Gap to Leader in External Test	-69.76	0.00	-6.75	-1.68	-0.49	-1.80
PERFORMANCE Score(4)		**0.00**	**-6.75**	**-1.68**	**-0.49**	**-1.80**



(1) For all algos: on exactly the same Learning Dataset. All models are optimized according to their Validation F_2-Score.
(2) All top-models are selected according to their Validation F_2-Score while checking that it remains close to their Training F_2-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 Scores

AI Technique #i	T_i	$i \in [1 ; n]$ n = number of AI Techniques benchmarked in terms of data-driven modeling = 5
Benchmark #k	B_k	$k \in [1 ; p]$ p = number of Benchmarks for the Use Case $\in \{1, 2, 3\}$

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 <u>and</u> 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; F_1-Score for a Binomial Classification; Average F_1-Score or Average F_2-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: $PS(T_i, B_k) = PC(T_i, B_k) - Best\_PC(B_k)$.

Performance Score of T_i

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

<u>Remark:</u>

- 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):

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

  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):

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

  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):

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

  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):

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

  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

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

<u>Remarks:</u>

- 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 Ti 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 Specificity	Sensitivity	Specificity	PPV	NPV	**F_2-Score**	Refusal
RANDOM MODEL *Number of Random Permutations (P-value) = 100,000 (0.001)*								
Performance against chance in External Test	*15.19%*	*10.65%*					***10.65%***	
XTRACTIS TOP-MODEL								
Descriptive Performance (Training)	5.72%	86.70%	86.70%	94.98%	61.58%	98.72%	80.16%	1.67%
Predictive Performance (Validation)	5.70%	86.50%	86.50%	95.03%	61.70%	98.70%	80.06%	1.68%
Real Performance (External Test)	5.76%	87.21%	87.21%	94.89%	61.30%	98.76%	**80.41%**	1.55%
LOGISTIC REGRESSION TOP-MODEL								
Descriptive Performance (Training)	6.91%	79.81%	79.81%	94.33%	56.65%	98.05%	73.78%	
Predictive Performance (Validation)	6.77%	79.41%	79.41%	94.52%	57.37%	98.02%	73.74%	
Real Performance (External Test)	6.94%	79.71%	79.71%	94.30%	56.51%	98.04%	**73.66%**	
RANDOM FOREST TOP-MODEL								
Descriptive Performance (Training)	5.95%	88.84%	88.84%	94.54%	60.17%	98.92%	81.11%	
Predictive Performance (Validation)	6.08%	87.16%	87.16%	94.55%	59.78%	98.75%	79.85%	
Real Performance (External Test)	6.46%	86.41%	86.41%	94.21%	58.09%	98.68%	**78.73%**	
BOOSTED TREE TOP-MODEL								
Descriptive Performance (Training)	5.36%	87.51%	87.51%	95.30%	63.38%	98.80%	81.32%	
Predictive Performance (Validation)	5.35%	86.37%	86.37%	95.42%	63.66%	98.69%	80.62%	
Real Performance (External Test)	5.85%	86.65%	86.65%	94.85%	60.99%	98.71%	**79.92%**	
NEURAL NETWORK TOP-MODEL								
Descriptive Performance (Training)	5.84%	84.91%	84.91%	95.02%	61.32%	98.55%	78.85%	
Predictive Performance (Validation)	5.73%	85.49%	85.49%	95.09%	61.80%	98.60%	79.40%	
Real Performance (External Test)	5.98%	84.94%	84.94%	94.86%	60.57%	98.55%	**78.61%**	

Corrective Induction, a virtuous R&D cycle with XTRACTIS, made possible by the intelligibility of XTRACTIS decision systems!

This "Refusal" metric means that for 1.67% of known patients (Training Dataset), XTRACTIS refuses to deliver a decision/prediction. This refusal level is similar for unknown patients (Validation and External Test Datasets).
In such cases, the modeler should ask the medical experts to check their diagnosis in the Reference Dataset, comparing their reasoning with that unfolded in XTRACTIS PREDICT's Prediction Report, and if necessary, correct this diagnosis in the Learning Dataset.
If corrections are made, the modeler should re-run XTRACTIS REVEAL with the updated Learning Dataset to induce a new, more robust and more intelligible top-model. If doctors were to confirm their diagnosis, there would be conflicting cases, i.e., patients similar for all the 8 predictors but receiving different diagnoses.
This would mean that some predictors would be missing to be able to differentiate between the cases.