Sorbonne University
Master 2 MS2A – Mathematics, Statistics and Learning

Data Challenge 2025 – Gas Detection

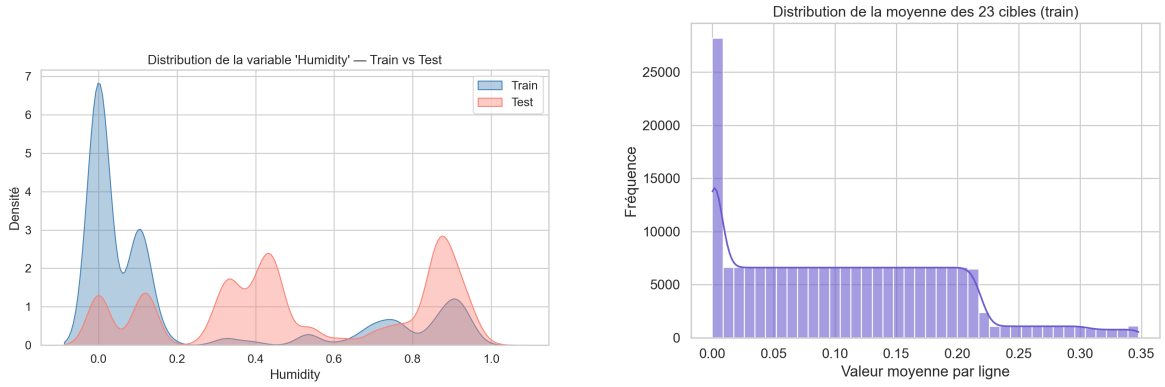
Summary Report – Tom DE OLIVEIRA

November 8, 2025

The **Data Challenge 2025 (ENS / Bertin Technologies)** focuses on simultaneously predicting alarm levels for different categories of toxic gases.

1. Data Preprocessing

Exploratory analysis: The exploratory analysis first confirmed the overall quality of the dataset. All variables are numerical, with no missing values or major anomalies, and sensor distributions are globally consistent. However, several critical points emerged during this phase. First, Figure 1 (left) highlights a **significant distribution shift in the *Humidity*** variable between the training and test sets. This shift, already mentioned in the challenge description, reveals a structural *data shift*: test samples correspond to physical conditions rarely observed in the training data. Further analysis across all sensors showed that this phenomenon extended, to a lesser extent, to other variables as well, confirming a **global train/test shift**. On the target side, variable ***c15*** was found to be constant (filled with zeros) across the entire training set, making it uninformative for learning. Finally, Figure 1 (right) presents the **distribution of the mean of the 23 targets per sample**. It is highly asymmetric, with a majority of low-concentration samples and a few high values. This imbalance suggests an uneven density between low and high-value regions, an aspect that influenced the validation and error-weighting strategy discussed later.



(a) Train/test shift on *Humidity*.

(b) Mean of the 23 targets (train).

Figure 1: Initial exploration: (left) humidity shift, (right) stratification motivation.

Data cleaning and preprocessing: Following the exploratory analysis, several adjustments were implemented to mitigate the effects of the *data shift* and improve generalization stability. First, the *Humidity* variable was **removed** from most model variants. Its distribution, highly different between train and test, introduced strong instability in cross-validation and degraded leaderboard performance. Alternative approaches were tested to keep it (rank transformation, quantile normalization, shared outlier detection), but none produced stable behavior, confirming that its removal offered the best bias/variance tradeoff. Second, a **global quantile clipping** (1^{st}–99^{th} percentiles) was applied to all numerical variables to reduce the influence of extreme values and alleviate distributional differences.

This also improved the robustness of ensemble estimators. Finally, additional descriptors were introduced using ***row-wise* features**: means, medians, standard deviations, interquartile ranges (IQR), mean absolute deviations (MAD), and L1/L2 norms computed per sample. These aggregates capture invariant statistical properties of the sensor measurements and proved particularly effective in non-stationary contexts, significantly improving robustness to the train/test shift. These transformations formed the foundation of the pipeline used for model selection and subsequent experiments.

2. Model Selection Phase

I compared several model families: *Random Forest*, *Extra Trees*, *XGBoost*, and *LightGBM*. Boosting methods showed strong local performance but high variability on the public leaderboard (sensitive to distribution shifts). *Random Forest* served as a robust baseline, while the **Extra Trees model** stood out for its speed on a large dataset ($\sim$330k rows) and robustness to noisy distributions.

Several **Extra Trees variants** were trained: with/without *row-wise* features, limited depth, bootstrap, and different random seeds. Evaluation was performed using the **Weighted RMSE** metric (weight 1.2 for $y \geq 0.5$), as required by the challenge. Because this metric depends on the true test distribution, local validation was imperfect; model changes were empirically validated through the two daily submissions allowed.

3. Final Model

The final pipeline is an **ensemble of six ExtraTreesRegressors** (multi-output), trained on stratified folds. Predictions were combined using an **optimized blending** (random sampling of weights on the simplex), followed by a **linear calibration** per target and a **shrinkage** toward the training mean ($\alpha \approx 0.95$). A final *clipping* ensured outputs remained in $[0, 1]$.

Scores: Public: **0.1400** / Private: **0.1520**

Table 1 summarizes the performance evolution across experiments.

Date	Method	Comment (params)	Public Score
28/09	RF (first submission)	simple baseline	0.1565
21/10	RF (K-Fold)	depth/leaves tuning	0.1540
31/10	RF + RW	added *row-wise* features (improved robustness)	0.1488
01/11	ExtraTrees	n_estimators=300, max_features=0.8	0.1417
07/11	ET_BLEND	6 variants (+RW, seeds, bootstrap)	**0.1400**

Table 1: Selected experiments and public leaderboard progression.

4. Challenges and Achievements

Challenges. The train–test shift (notably *Humidity*) made local validation unreliable; domain adaptation attempts (e.g., *CORAL*, adversarial validation) did not yield stable gains in this context. The **Weighted RMSE** metric, penalizing high-concentration errors more strongly, further widened the gap between OOF and leaderboard scores.

Achievements. I designed a **reproducible and modular pipeline**: quantile clipping, *Humidity* removal, ***row-wise*** features (strong robustness lever), diverse **Extra Trees** variants, optimized **blending**, **calibration**, and **shrinkage**. Iterative submissions guided exploration despite a metric that was difficult to *approximate* outside the competition platform.

The submitted notebook fully reproduces the pipeline and the final submission.