ECML PKDD 2025 **Nectar Track**

Institut de Recherche en Informatique de Toulouse
CNRS - INP - UT3 - UT1 - UT2J

UNIVERSITÉ TOULOUSE CAPITOLE

XStacking: Explanation-Guided Stacked Ensemble Learning

Moncef Garouani [1], Ayah Barhrhouj [2], Olivier Teste [3]

[1] IRIT, UMR5505 CNRS, Université Toulouse Capitole, Toulouse, France
[2] LIS, UMR 7020 CNRS, Aix-Marseille University Marseille, France
[3] IRIT, UMR 5505 CNRS, Université de Toulouse, UT2J, Toulouse, France

moncef.garouani@irit.fr

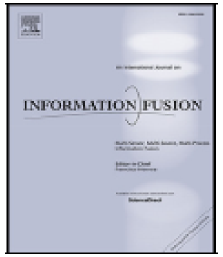
Slides available at mgarouani.fr/talks -> ECMLPKDD 2025 Presentation

ECML PKDD 2025 **Nectar Track**

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus



XStacking : An effective and inherently explainable framework for stacked ensemble learning

Moncef Garouani [a,*], Ayah Barhrhouj [b], Olivier Teste [c]

[a] IRIT, UMR 5505 CNRS, Université Toulouse Capitole, Toulouse, 31000, France
[b] Aix-Marseille University, CNRS, LIS (UMR 7020), Marseille, 13007, France
[c] IRIT, UMR 5505 CNRS, Université Toulouse, UT2J, Toulouse, 31000, France

https://doi.org/10.1016/j.inffus.2025.103358

ARTICLE INFO

Keywords:
Machine learning
Ensemble learning
Stacking
Explainable artificial intelligence
Shapley additive explanations

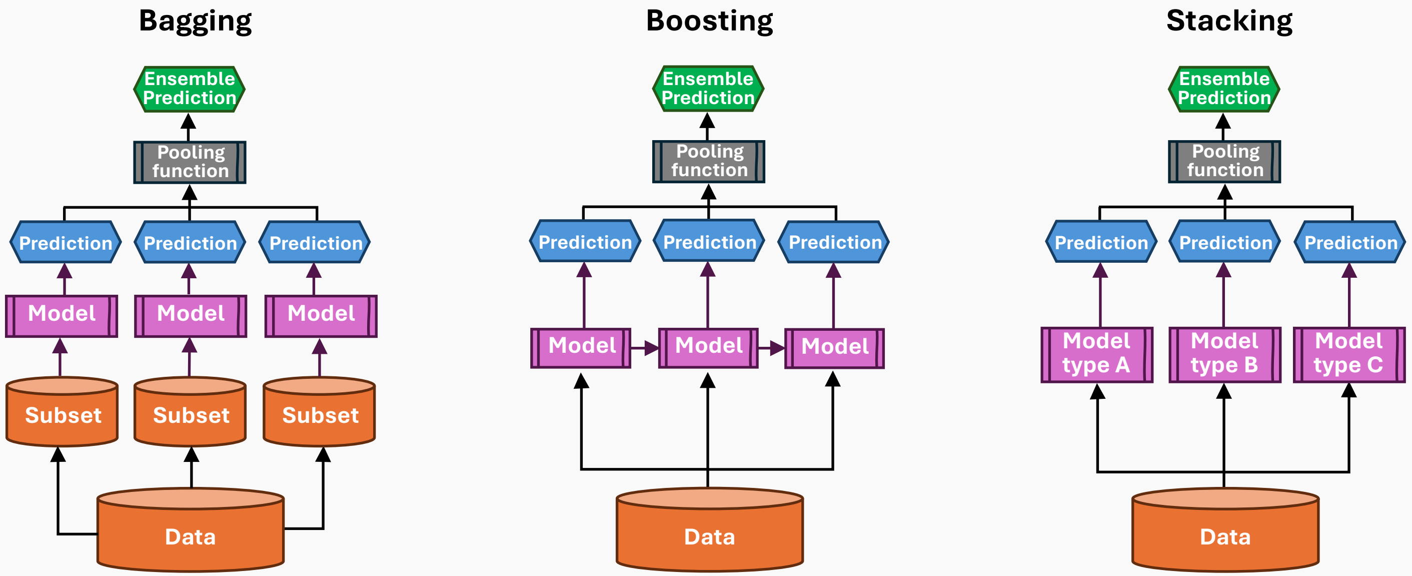
ABSTRACT

Ensemble Machine Learning (EML) techniques, especially stacking, have proven effective in boosting predictive performance by combining several base models. However, traditional stacked ensembles often face challenges in predictive effectiveness of the learning space and model interpretability, which limit their practical application. In this paper, we introduce *XStacking*, an effective and inherently explainable framework that addresses these limitations by integrating dynamic feature transformation with model-agnostic Shapley Additive Explanations. XStacking is designed to enhance both effectiveness and transparency, ensuring high predictive accuracy and providing clear insights into model decisions. We evaluated the framework on 29 benchmark datasets for classification and regression tasks, showing its competitive performance compared to state-of-the-art stacked ensembles. Furthermore, XStacking interpretability features offer actionable insights into feature contributions and decision pathways, making it a practical and scalable solution for applications where both high performance and model transparency are critical.

PLAN

Ensemble learning

- **Idea**: Combine predictions from multiple models → better accuracy & robustness
- **Why it works**:
  - Reduces variance & bias
  - Captures diverse data patterns
- **Common approaches**:
  - Bagging – parallel, variance reduction (e.g., Random Forest)
  - Boosting – sequential, bias reduction (e.g., XGBoost, AdaBoost)
  - Stacking – meta-learning on predictions of base models



Stacking – Inner Working & Advantages

❖ **How it works:**

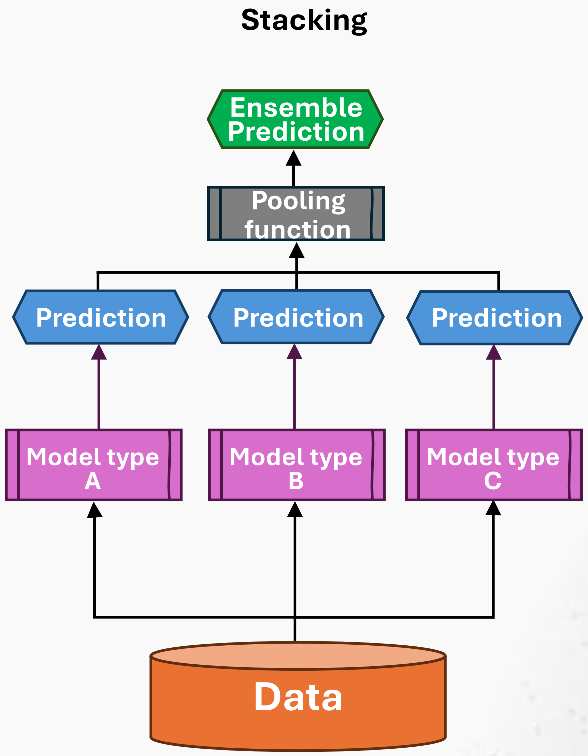
- Train multiple diverse base models on the same dataset
- Each base model produces predictions for all samples
- Create a new dataset from these predictions
- Train a meta-learner on this new dataset to make final predictions

❖ **Key idea:**

- Meta-learner combines complementary insights
- Learns to weight base models' outputs for better accuracy

❖ **Advantages:**

- Combines strengths of heterogeneous models (not limited to same model family)
- Captures complementary patterns in **data** & strengths of diverse **models**
- More flexible and more powerful in complex problems beyond bagging or boosting
- Proven effective across domains: healthcare, finance, NLP, etc.
- More flexible than Bagging (reduces variance) and Boosting (reduces bias).

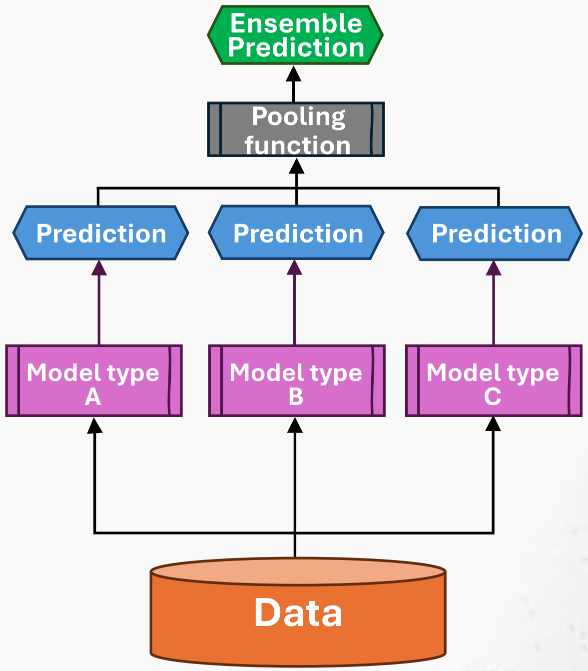


Stacking — Limitations

❖ **Learning space issues:**
- Meta-learner input = base models' predictions
- May be **insufficiently informative** if :
  - predictions are limited in diversity
  - Base models are highly correlated or redundant

❖ **Transparency challenges:**
- Meta-learner and base models often "**black boxes**"
- Hard to trace how predictions are derived
- Complex interactions between models make feature contribution unclear
- Low interpretability → reduced trust and accountability

Stacking



XStacking: Explanation-Guided Stacked Ensemble Learning

- **Goal:** Improve <u>predictive effectiveness of the learning space</u> and stacking model <u>interpretability</u>

- **Inspiration:**
  - Like in **human decision-making**:

    *We don't just consider people's recommendations — we consider **why** they made them*
  - Human decision-making uses **rationales & explanations**

- **Key Idea:**
  - Build a stacking ensemble guided by explanations, not just final predictions
  - Mimic human reasoning: combines both outputs and rationales of base models
  - Use **feature importance** (e.g., **Shapley values**) from each base classifier
  - Concatenate explanations from all base models → **enriched learning space**
  - Train meta-learner using explanations form base learners

XStacking: Explanation-Guided Stacked Ensemble Learning

❖ **Goal:** Improve predictive effectiveness of the learning space and stacking model interpretability

❖ **Inspiration:**

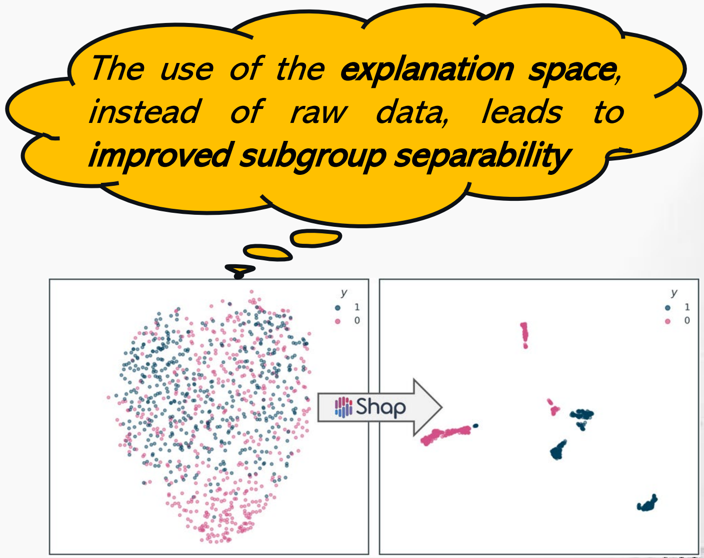
- Like in **human decision-making**:

  *We don't just consider people's recommendations — we consider **why** they made them*

- Human decision-making uses **rationales** + **explanations**

❖ **Key Idea:**

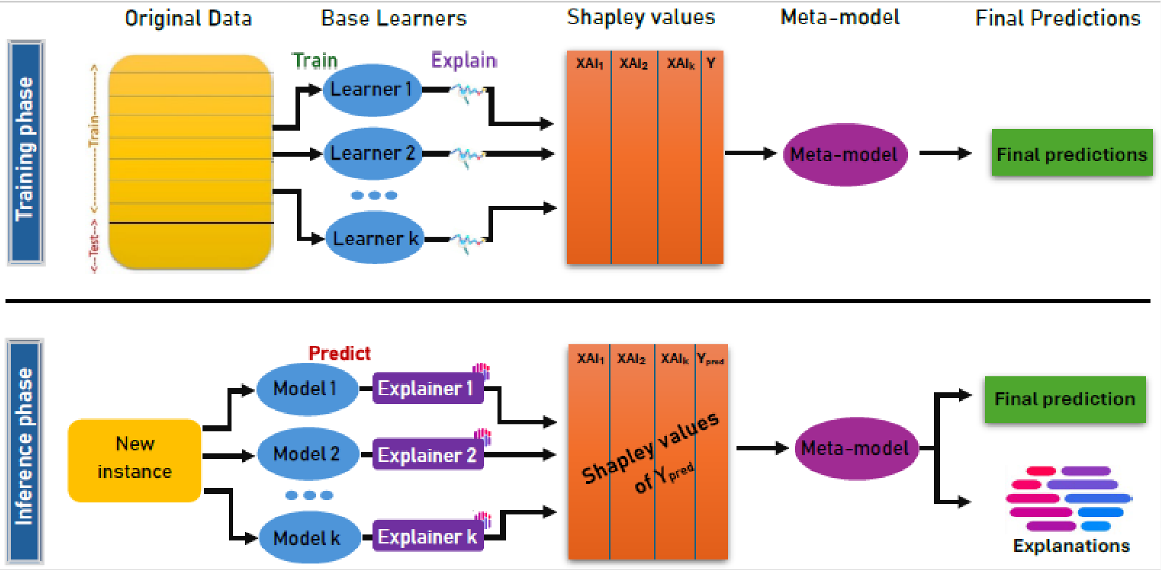
- Build a stacking ensemble guided by explanations, not just final predictions
- Mimic human reasoning: combines both outputs and rationales of base models
- Use **feature importance** (e.g., **Shapley values**) from each base classifier
- Concatenate explanations from all base models → **enriched learning space**
- Train meta-learner using explanations form base learners



https://doi.org/10.1007/978-3-030-93733-1_29

XStacking: Explanation-Guided Stacked Ensemble Learning

❖ **Pipeline:**
- Train multiple base classifiers on original data
- Generate predictions' explanations for each sample
- Concatenate explanations → **build new enriched dataset**
- Train stacking meta-learner on enriched input



❖ **Benefits:**
- Captures **complementary knowledge** from diverse models
- Provides **richer**, **more informative** input for meta-learner
- Bridges gap between **accuracy** and **interpretability**

Experimental study

XStacking: Explanation-Guided Stacked Ensemble Learning

❑ Experimental study

Datasets

- **29** datasets from PMLB benchmark
- **17** classification, **12** regression tasks
- Varied domains: categorical, ordinal, and continuous features
- Train/test split: 70% / 30%

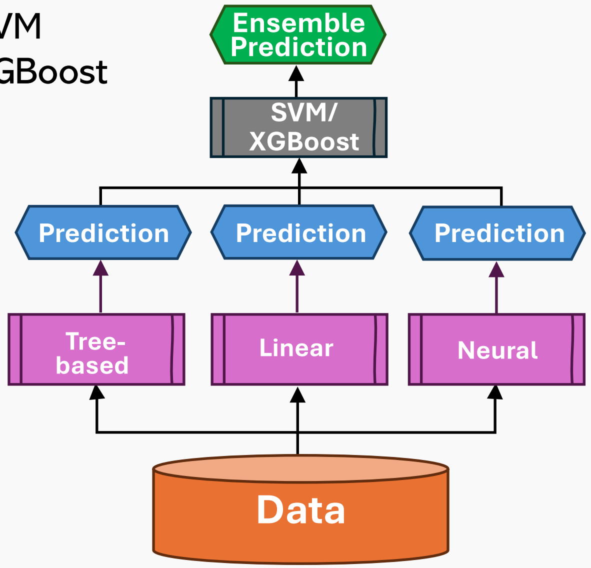
Predictive Models

❑ **Base learners:**

- Tree-based: Decision Trees
- Linear: Logistic / Linear Regression
- Neural: Multilayer Perceptrons

❑ **Meta-Learners:**

- SVM
- XGBoost



Baseline

- Traditional Stacking (Scikit-learn)
- Same datasets, same setup
- Assess improvement from explanation-based learning space

XStacking: Explanation-Guided Stacked Ensemble Learning

❑ **Experimental study:** *research questions*

- **RQ1 (Effectiveness)** — How effective is XStacking for classification and regression tasks?
- **RQ2 (Efficiency)** — How computationally efficient is XStacking?
- **RQ3 (Explainability)** — How explainable are the results produced by XStacking?

XStacking: Explanation-Guided Stacked Ensemble Learning

❑ Experimental study: *Effectiveness of XStacking (RQ1)*

Dataset	SVM Meta Learner		XGB Meta Learner	
	Stacking	XStacking	Stacking	XStacking
iris	1	1	1	1
digits	0.9666	0.9729	0.9648	0.9685
wine	0.9629	0.9629	0.9444	0.9444
breast_cancer	0.9649	0.9824	0.9707	0.959
Personal Loan	0.984	0.984	0.984	0.9873
diabetes	0.7705	0.7489	0.7705	0.7229
adult	0.8366	0.8454	0.8366	0.8624
chess	0.9927	0.9927	0.9927	0.9927
mushroom	1	1	1	1
vehicle	0.7204	0.7795	0.7362	0.7519
cmc	0.5633	0.5656	0.561	0.5701
splice	0.8766	0.8923	0.8766	0.9007
car	0.9653	0.9653	0.9653	0.9556
churn	0.902	0.9086	0.9026	0.9466
shuttle	0.9992	0.9992	0.9993	0.9993
hypothyroid	0.9768	0.9799	0.9768	0.9778
Breast-cancer	0.7325	0.7558	0.7325	0.7441
	09/17 best		09/17 best	
	16/17 equal or better		14/17 equal or better	

Table 2. *Comparison of XStacking and baseline performance in classification tasks based on Accuracy.*

Dataset	SVM Meta Learner		XGB Meta Learner	
	Stacking	XStacking	Stacking	XStacking
wind	10.1596	9.7476	10.1669	9.1771
cpu_small	22.4364	11.354	16.6171	7.6828
ESL	0.4230	0.2928	0.3444	0.3060
ERA	2.802	2.7079	2.721	2.6591
LEV	0.4954	0.4880	0.4922	0.5003
pol	15.0157	13.923	16.1196	15.5640
puma8NH	11.1515	10.9882	10.61	11.0955
satellite_image	0.8952	0.5294	0.7886	0.5054
pm10	0.69	0.6962	0.6901	0.5892
pollen	2.2528	2.1962	2.1997	2.4380
Abalone	4.5547	4.4766	4.3444	4.6212
Wine_Quality	0.4705	0.4384	0.3841	0.3562
	11/12 best		08/12 best	
	11/12 equal or better		08/12 equal or better	

Table 1. *Comparison of XStacking and baseline performance in regression tasks based on MSE.*

❖ **Wilcoxon signed-rank test: performance between traditional stacking and XStacking is statistically significant ($p < 0.01$) in terms of accuracy across all datasets.**

XStacking: Explanation-Guided Stacked Ensemble Learning

❑ Experimental study: *Computational Efficiency of XStacking (RQ2)*

Dataset properties	Meta-learner	Stacking	XStacking
D<15 | m<1700	SVM	4,14	1012,83
	XGBoost	7,81	947,04
D>15 | m>1700	SVM	608,87	3740
	XGBoost	611,02	3250

Table 3. *Comparison of the average runtime, in seconds, of the XStacking method against the state-of-the-art stacking ensemble learning.*

- **SHAP-based explanations** → added computational overhead.
- **Trade-off**: Slightly longer runtime ↔ Ensured interpretability.

XStacking: Explanation-Guided Stacked Ensemble Learning

❑ Experimental study: *Explainability and reliability of XStacking (RQ3)*

❖ Case 1: All Base Models Have High Accuracy (IRIS Dataset)

❖ **Context**

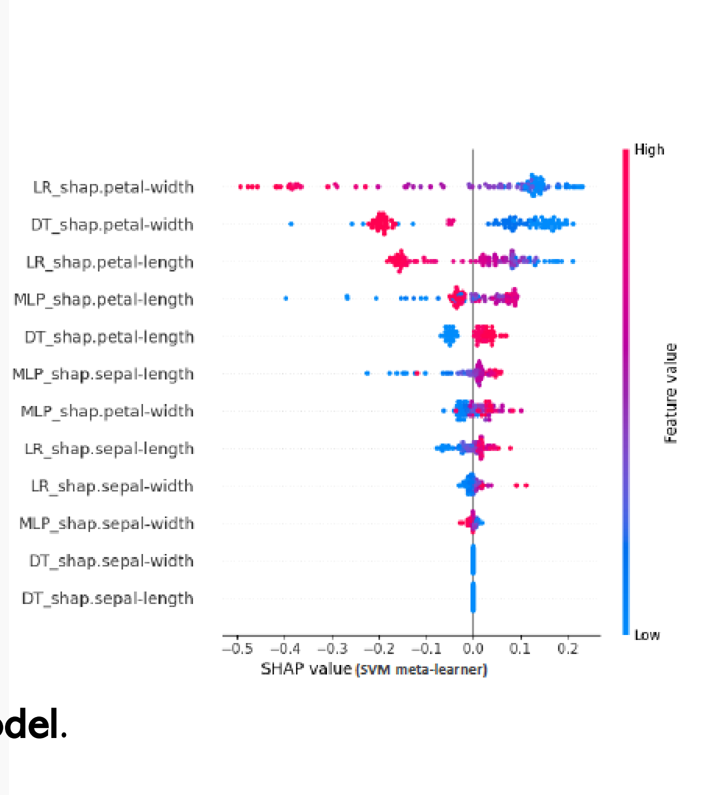
- Base learners: **MLP (0.9523)**, **LR (0.9619)**, **DT (0.9714)**.
- Accuracy is consistently high across all models.

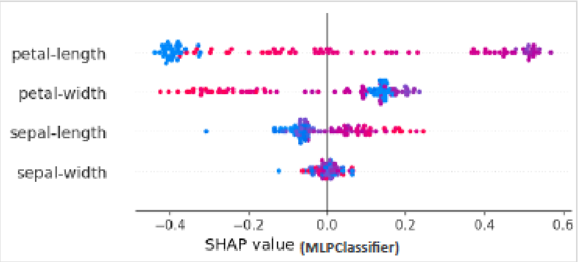
❖ **SHAP Insights**

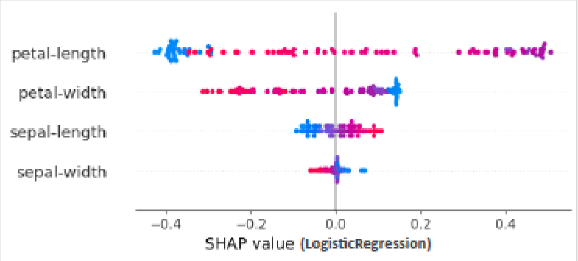
- **Petal length** and **petal width** dominate across all models.
- Meta-learner **aligns with base model rankings**.
- Equal contributions → balanced feature importance in meta-model.

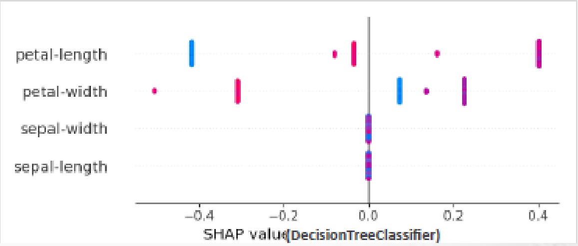
❖ **Implication**

- When base models perform similarly well,
→ Meta-learner combines insights **without bias toward one model**.
- Features importance remains **stable** across learners.









XStacking: Explanation-Guided Stacked Ensemble Learning

❑ Experimental study: *Explainability and reliability of XStacking (RQ3)*

❖ Case 2: Base Models Have Very Different Accuracies (CAR Dataset)

❖ **Context**

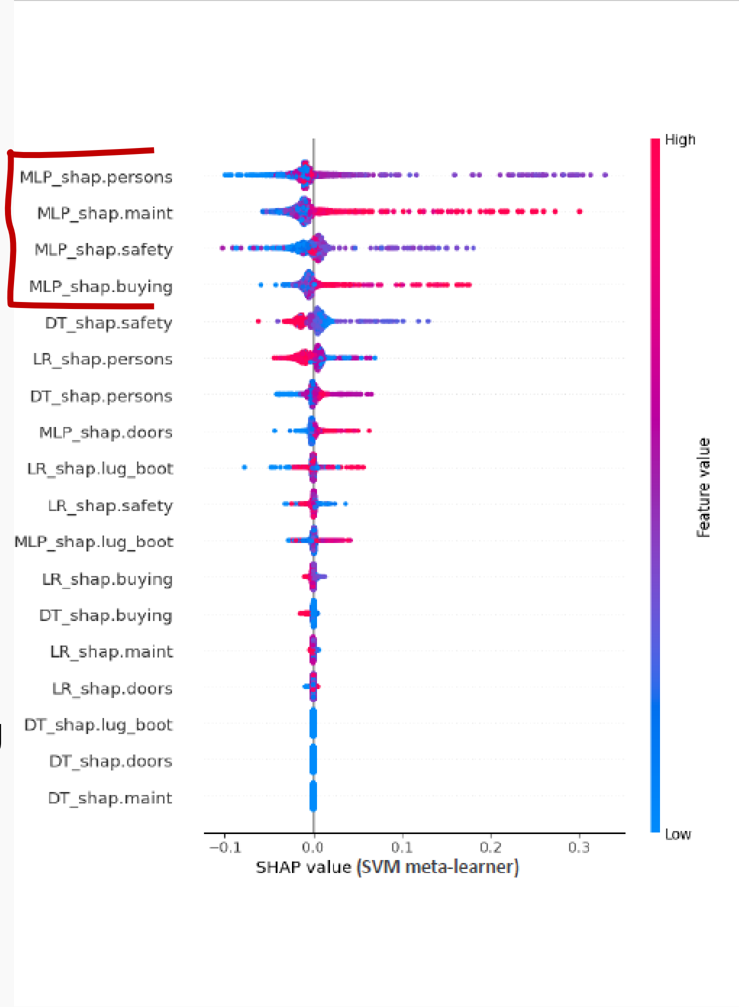
- Base learners: **DT (0.69), LR (0.79), MLP (0.99)**.
- Performance varies greatly.

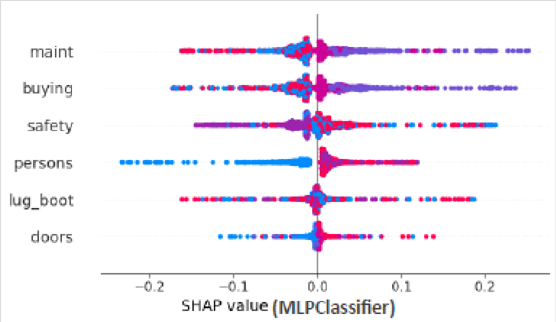
❖ **SHAP Insights**

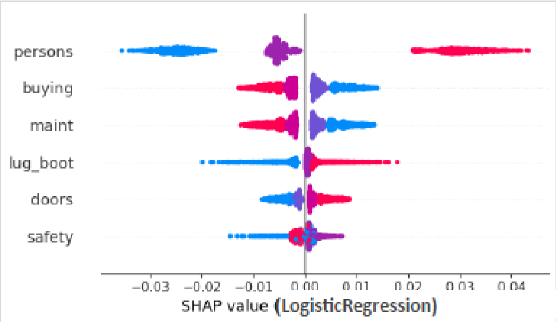
- **MLP dominates** meta-model feature importance due to very high accuracy.
- LR provides **supplementary linear insights**.
- DT's contribution is **minimal** (low accuracy).

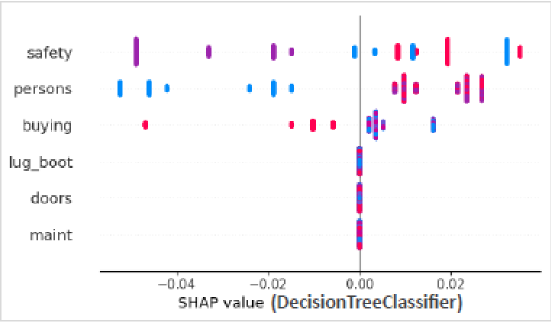
❖ **Implication**

- When base models have unequal accuracies,
→ Meta-learner assigns **greater weight to the best performing models**.
- Interpretability reflects **performance-driven weighting**.









Conclusion & Future works

❖ **What we proposed**

- **XStacking**: Enhances stacked ensemble learning with **explanation guided meta-earning.**
- Integrates **SHAP-based explanations** directly into stacking.

❖ **Key outcomes**

- **Higher accuracy** than traditional stacking.
- **Built-in interpretability** at both base and meta-model levels.
- Scales to diverse datasets & model families.

❖ **Why it matters**

- Enables **transparent & trustworthy** ensemble learning.
- Facilitates **debugging** and **feature-level insights**.
- Suitable for **research** & **real-world applications**.

❖ **Future directions**

- Speed up explanations' computation for **large-scale ensembles**.
- Extend to **deep learning** & **gradient-based attributions**.

THANK YOU!

ECML PKDD 2025 **Nectar Track**

Institut de Recherche en Informatique de Toulouse
CNRS - INP - UT3 - UT1 - UT2J

UNIVERSITÉ TOULOUSE CAPITOLE

XStacking: Explanation-Guided Stacked Ensemble Learning

Moncef Garouani [1], Ayah Barhrhouj [2], Olivier Teste [3]

[1] IRIT, UMR5505 CNRS, Université Toulouse Capitole, Toulouse, France
[2] LIS, UMR 7020 CNRS, Aix-Marseille University Marseille, France
[3] IRIT, UMR 5505 CNRS, Université de Toulouse, UT2J, Toulouse, France

moncef.garouani@irit.fr

Slides available at mgarouani.fr/talks -> ECMLPKDD 2025 Presentation