AIROV 2026 · MACHINE LEARNING · META-LEARNING

Explainable Selection of Machine Learning Algorithms in Social Sciences

Dijana Oreški, Luka Katava, Alen Kišić · AIRoV – Austrian Symposium on AI, Robotics, and Vision

MOTIVATION

The algorithm selection problem

Practitioners in education and business often lack ML expertise. Trial-and-error algorithm selection is inefficient and non-transparent. Meta-learning can help, but existing systems remain black boxes.

METHODOLOGY

Pipeline

45 datasets (education + business)

↓

179 meta-features extracted (pymfe)

↓

Mutual info + Kendall's τ selection

↓

Top 15 meta-features retained

↓

Scoring meta-model (linear, interpretable)

↓

Algorithm recommendation + explanation

CANDIDATE ALGORITHMS

Decision Tree · KNN · MLP · Ridge

45	179	15
DATASETS	META-FEATURES	SELECTED

RESULTS · CONFUSION MATRIX

Model performance

Accuracy 0.556 (LOO). DT and Ridge are predicted most reliably; KNN shows the most confusion with other classes.

	DT	KNN	MLP	Ridge
DT	5	3	0	0
KNN	4	5	3	1
MLP	1	1	7	1
Ridge	2	2	2	8

RESULTS · META-FEATURE IMPORTANCE

Key features per algorithm

	DT	KNN	MLP	Ridge
ch	-0.13	0.18	-0.21	0.12
eigenvalues.mean	0.12	0.02	-0.12	-0.01
var.mean	0.12	0.02	-0.12	-0.01
joint_est.sd	-0.06	-0.20	0.07	0.19
attr_ent.sd	-0.01	-0.13	-0.11	0.24
n4.mean	-0.10	0.04	-0.14	0.17
iq_range.mean	0.09	0.05	-0.12	-0.03
two_itemset.sd	0.03	0.14	-0.13	-0.04
cor.sd	0.03	0.35	-0.11	-0.26
leaves_corrob.sd	0.09	0.05	0.21	-0.32
ns_ratio	-0.04	-0.20	0.29	-0.03
best_node.sd	0.14	0.18	-0.01	-0.28
min.sd	0.11	0.07	-0.25	0.07
var.sd	0.05	0.02	-0.07	0.01
cor.mean	-0.04	0.27	-0.01	-0.22

strong negative · neutral · strong positive

KEY FINDINGS

What drives recommendations?

- cor.sd is the single most influential meta-feature — confirmed by both sensitivity analysis and leave-one-out (LOO) validation.
- High variability in attribute correlations (cor.sd=0.35) strongly favours KNN — consistent with distance-based methods benefiting from structured inter-variable relationships.
- Model-based features (leaves_corrob.sd, best_node.sd) negatively correlate with Ridge (–0.32, –0.28), pointing to structural dataset properties influencing linear model suitability.
- ns.ratio and leaves_corrob.sd associate with MLP (0.29, 0.21), suggesting datasets with complex boundaries favour higher-capacity models.
- Most meta-features act collectively — recommendations emerge from joint patterns, not a single dominant signal.

MODEL DESIGN

Why interpretable?

The scoring meta-model uses explicit Kendall's τ-derived weights. Each meta-feature contributes linearly to the final score — making every recommendation fully traceable back to dataset characteristics. No black box.

TAKEAWAY

Explainable meta-learning can recommend ML algorithms transparently — building trust with non-expert practitioners in education and business.

LIMITATIONS + FUTURE WORK

Small meta-dataset (45 datasets), limited algorithm set. Future work: expand the repository, broaden algorithm coverage, and explore SHAP and other XAI techniques for deeper explanation.