# Supplementary Material Referring to the Illustrations

Paper: Explanations Based on Item Response Theory (eXirt): A Model-Specific Method to Explain Tree-Ensemble Model in Trust Perspective

Elements

A

Train

Test

Black box

Model's output

B

Test

Black box

D

Perturbation of model

M1: Baseline test with original model: (No perturbation) — Iteration 1

M2: Test with original model: (Perturbation input 1) — Iteration 2

M3: Test with original model: (Perturbation input 2) — Iteration 3

M4: Test with original model: (Perturbation input 3) — Iteration 4

...

M29: Test with original model: (Perturbation input 4) — Iteration 5

Each model response is evaluated based on different IRT properties, so the theta rank shows the attributes that best explain (global attribute relevance rank) how the model makes its predictions based on **discrimination**, **difficulty**, and **guessing** values.

G

Test dataset instances

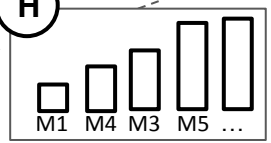
E

| | Instance 1 | Instance 2 | Instance 3 | Instance 4 | Instance ... |
|---|---|---|---|---|---|
| M1 | 1 | 1 | 1 | 1 | … |
| M2 | 1 | 0 | 0 | 0 | … |
| M3 | 1 | 1 | 1 | 0 | … |
| M… | … | … | … | .. | … |
| M29 | 1 | 0 | 1 | 1 | … |

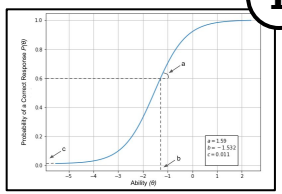
The lowest **Total Score** indicates the best attribute to explain the model.

K

H




I



**Variations:** 7 different ways to perturb the attributes of the model (note, to illustrate the iterations several models are presented, but in practice only 1 model is used).

Note, in this illustrative example, the model has 4 inputs and 7 perturbations were applied to these. So number of respondents is 4*7+1 = 29 (number 1 is the original model without perturbation).

C

Creation of the response matrix for each instance of the test set and each of the iterations of perturbation of model.

**1** correct classifier response.
**0** wrong classifier response.

Performance measure: **accuracy**

F

J

Local explanations created from item parameters estimated by respondents (original model + original model with iterations perturbations).

# Relationship of the terms of IRT and ML

IRT: Questions
ML: Dataset instances

| | Instance 1 | Instance 2 | Instance 3 | Instance … | Instance n |
|---|---|---|---|---|---|
| Model 1 | 1 | 1 | 1 | … | 0 |
| Model 2 | 0 | 1 | 1 | … | 1 |
| Model 3 | 0 | 1 | 0 | .. | 1 |
| Model 4 | 1 | 1 | 0 | .. | 0 |
| Model 5 | 1 | 0 | 0 | … | 1 |
| Model 6 | 1 | 0 | 0 | … | 0 |

IRT: Individuals
ML: Models

IRT: Responses from individuals
ML: Models responses

IRT: Estimation of parameter item

IRT: Item parameters

| | Discrimination | Difficulty | Guessing |
|---|---|---|---|
| Instance 1 | 0.9 | 1.3 | 0.0 |
| Instance 2 | 0.5 | 1.4 | 0.0 |
| Instance 3 | 0.4 | 1.1 | 0.2 |
| Instance … | … | ... | … |
| Instance n | 0.3 | 0.1 | 0.4 |

IRT: Θ estimation

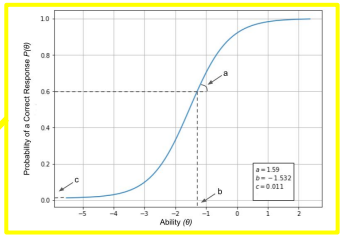
IRT: Individuals
ML: Models

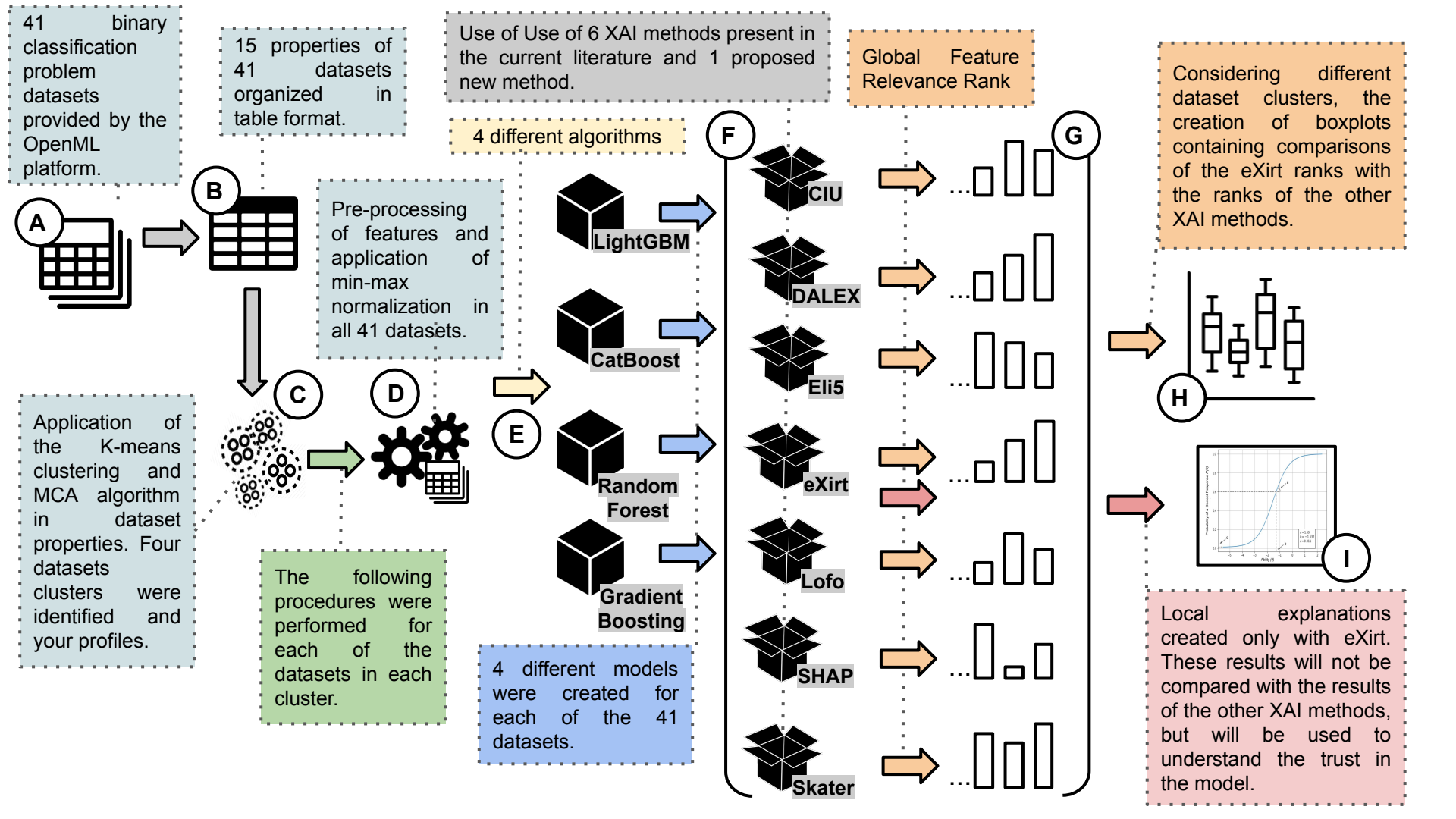
IRT: Θ

| | Score |
|---|---|
| Model 2 | 21.5 |
| Model 1 | 14.3 |
| Model 6 | 13.4 |
| Model 4 | 9.2 |
| Model 5 | 4.4 |
| Model 3 | -1.1 |

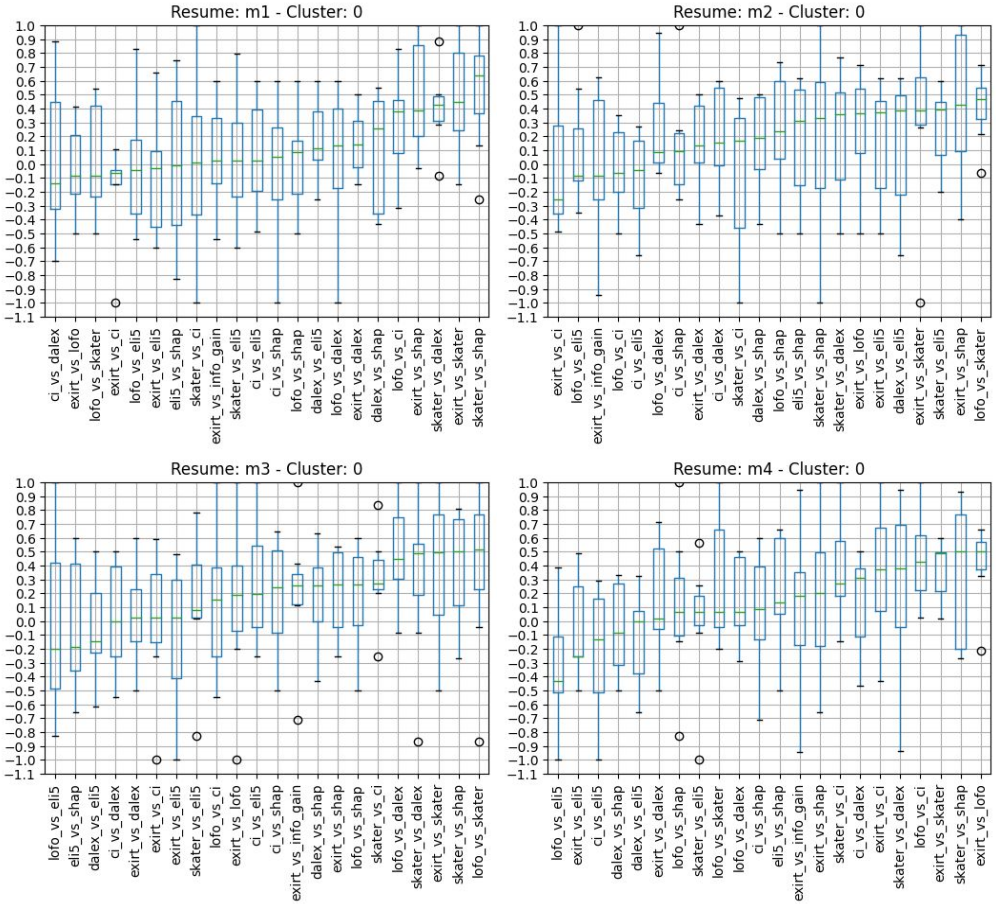
IRT: ICC plot

IRT: ICC result

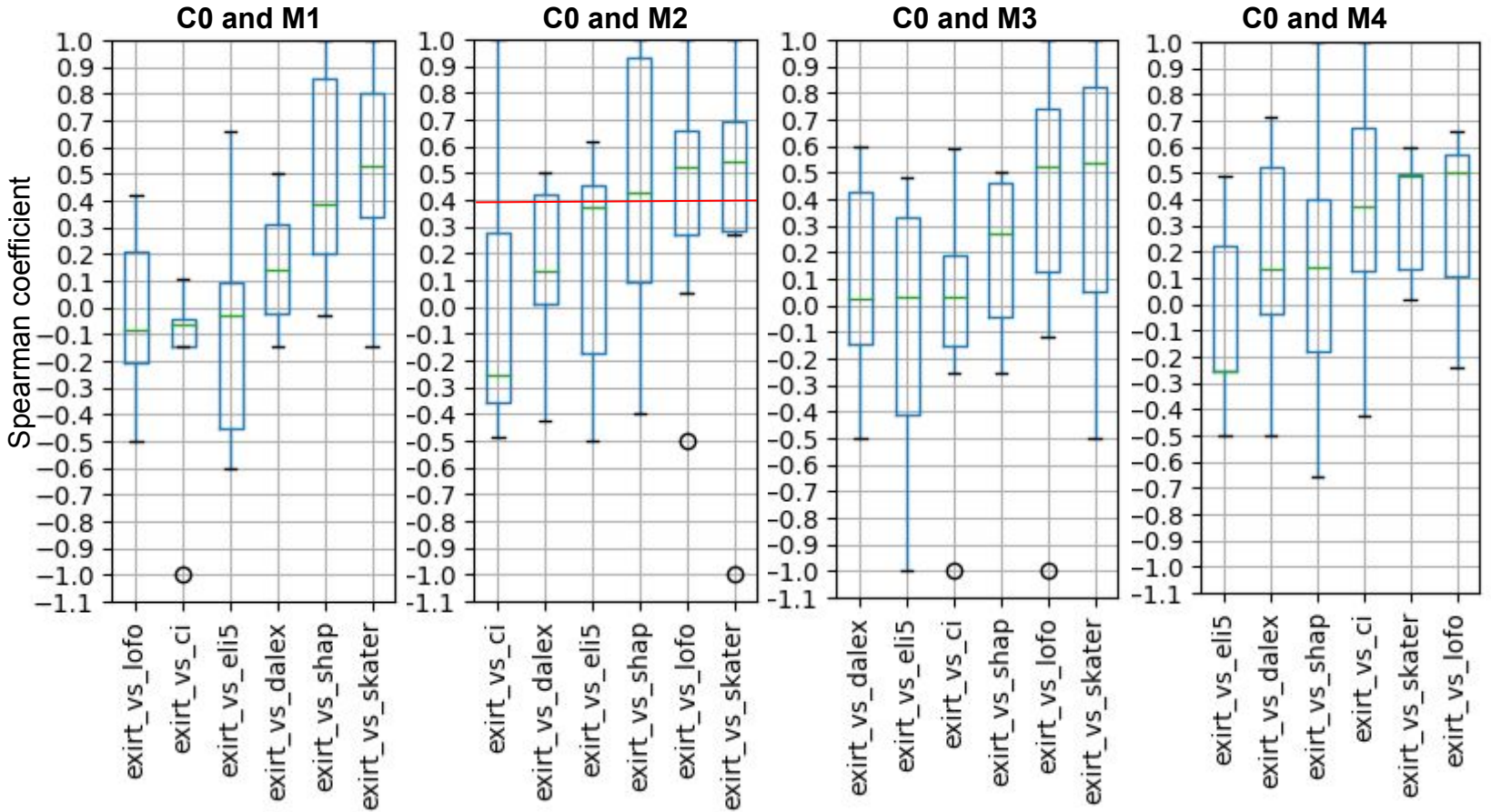



41 binary classification problem datasets provided by the OpenML platform.
15 properties of 41 datasets organized in table format.
Use of Use of 6 XAI methods present in the current literature and 1 proposed new method.
Global Feature Relevance Rank
Considering different dataset clusters, the creation of boxplots containing comparisons of the eXirt ranks with the ranks of the other XAI methods.
A
B
Pre-processing of features and application of min-max normalization in all 41 datasets.
4 different algorithms
F
G
LightGBM
CatBoost
Random Forest
Gradient Boosting
CIU
DALEX
Eli5
eXirt
Lofo
SHAP
Skater
H
C
D
E
Application of the K-means clustering and MCA algorithm in dataset properties. Four datasets clusters were identified and your profiles.
The following procedures were performed for each of the datasets in each cluster.
4 different models were created for each of the 41 datasets.
I
Local explanations created only with eXirt. These results will not be compared with the results of the other XAI methods, but will be used to understand the trust in the model.

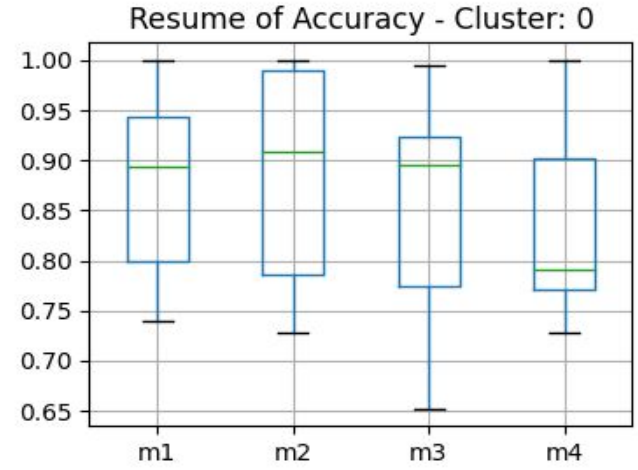
# All pair comparisons of feature relevance ranks for models (M1 to M4) based on cluster 0 datasets

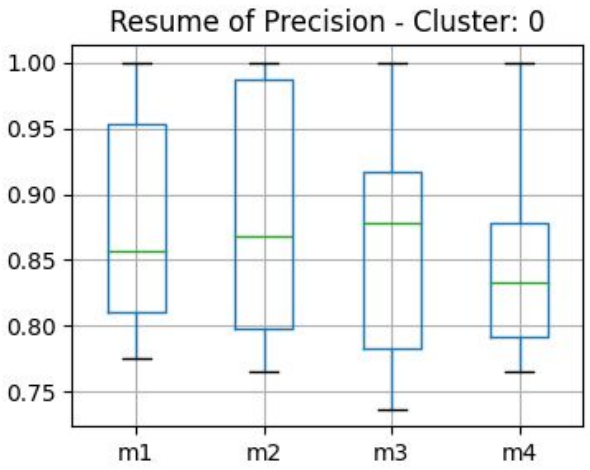


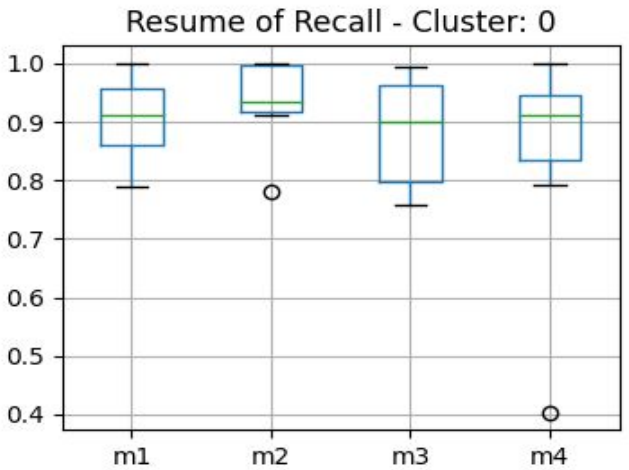
# Only *eXirt* comparisons of feature relevance ranks for models (M1 to M4) based on cluster 0 datasets



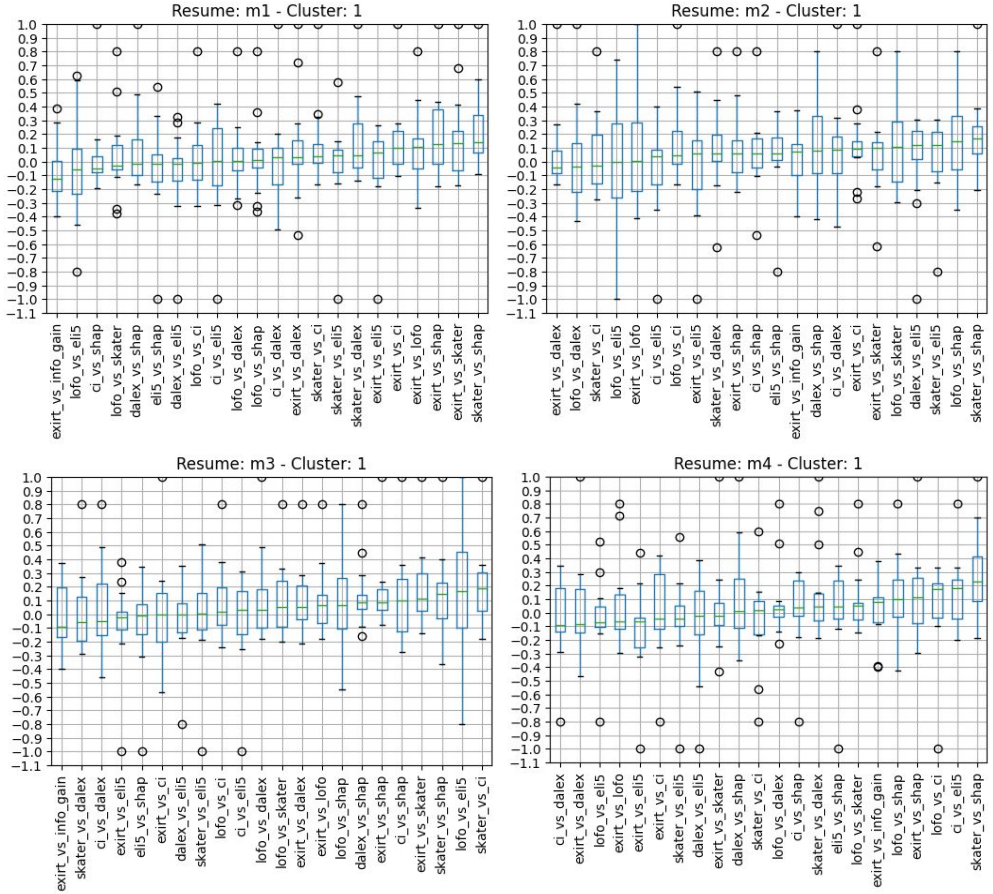
# Data regarding the performance of models M1 to M4 based on cluster 0 datasets



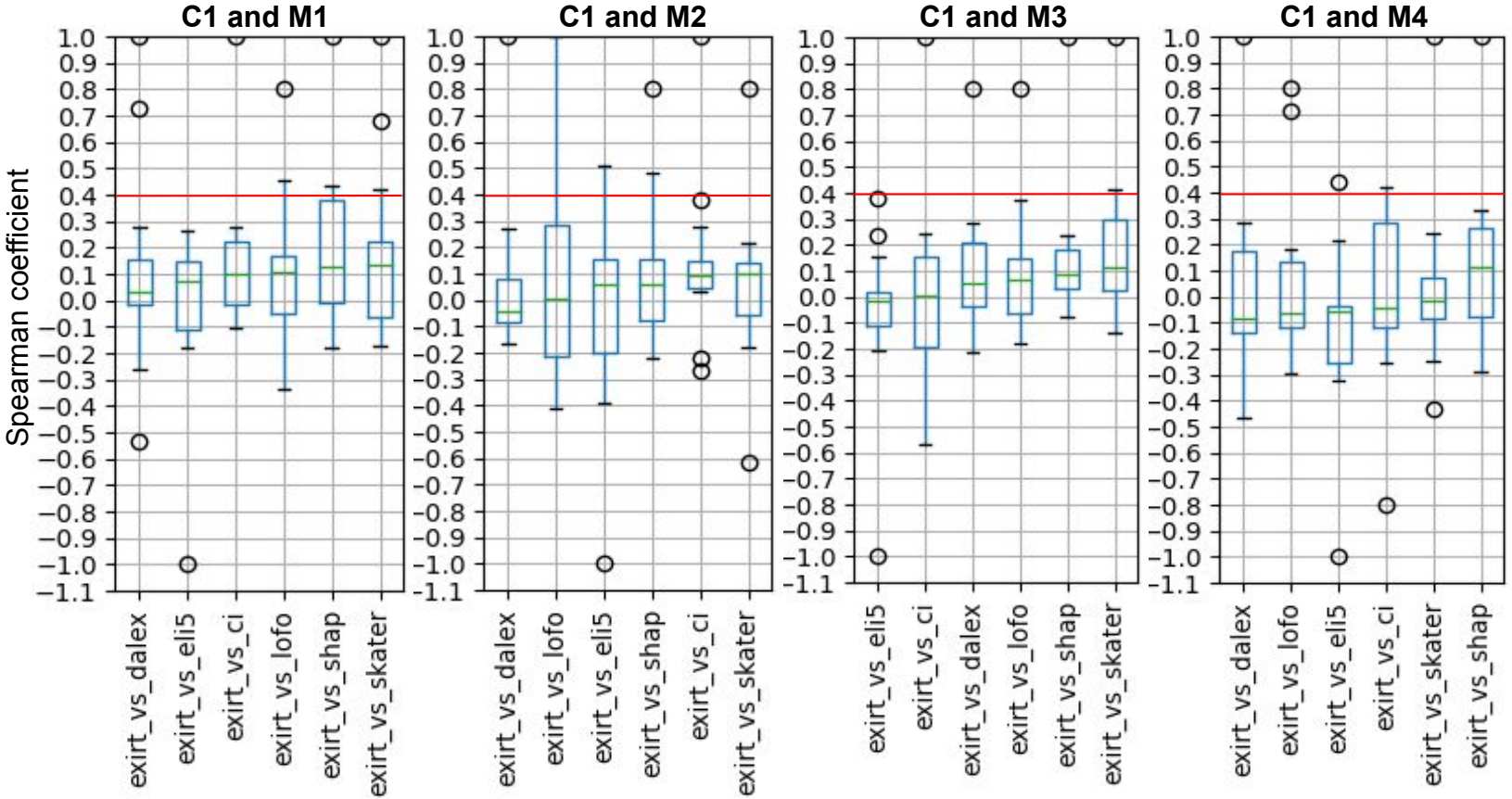



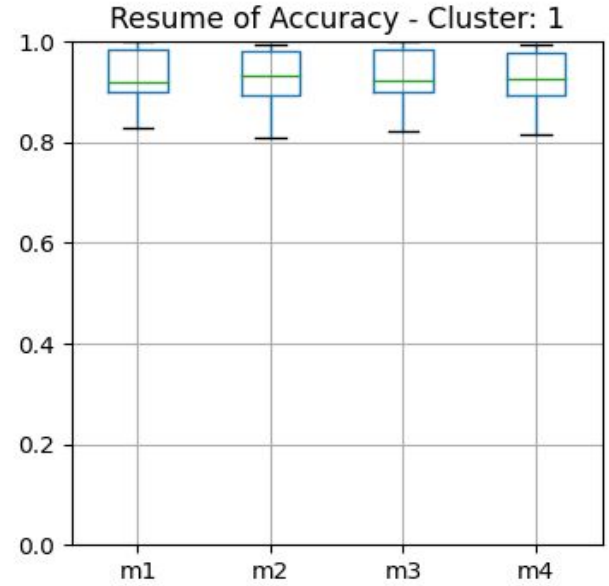
# All pair comparisons of feature relevance ranks for models (M1 to M4) based on cluster 1 datasets

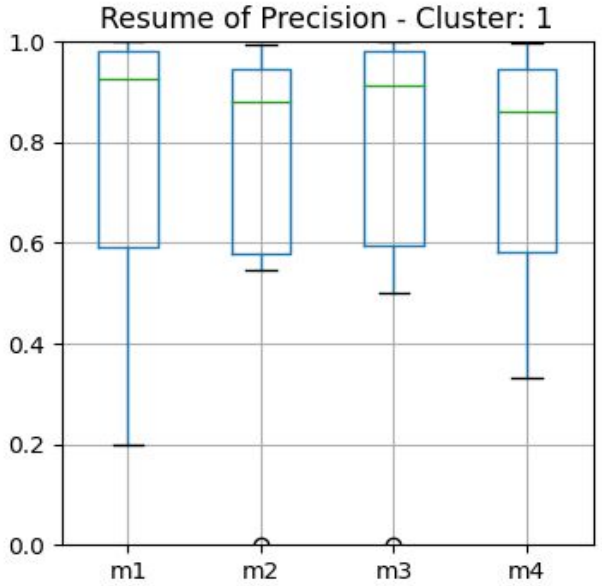


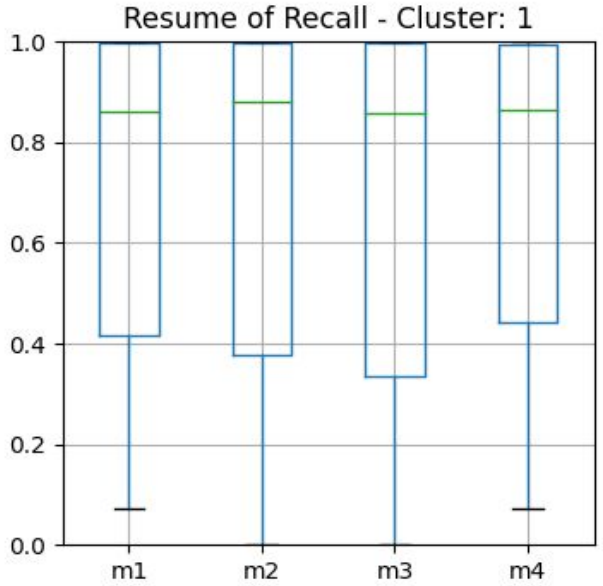
# Only *eXirt* comparisons of feature relevance ranks for models (M1 to M4) based on cluster 1 datasets



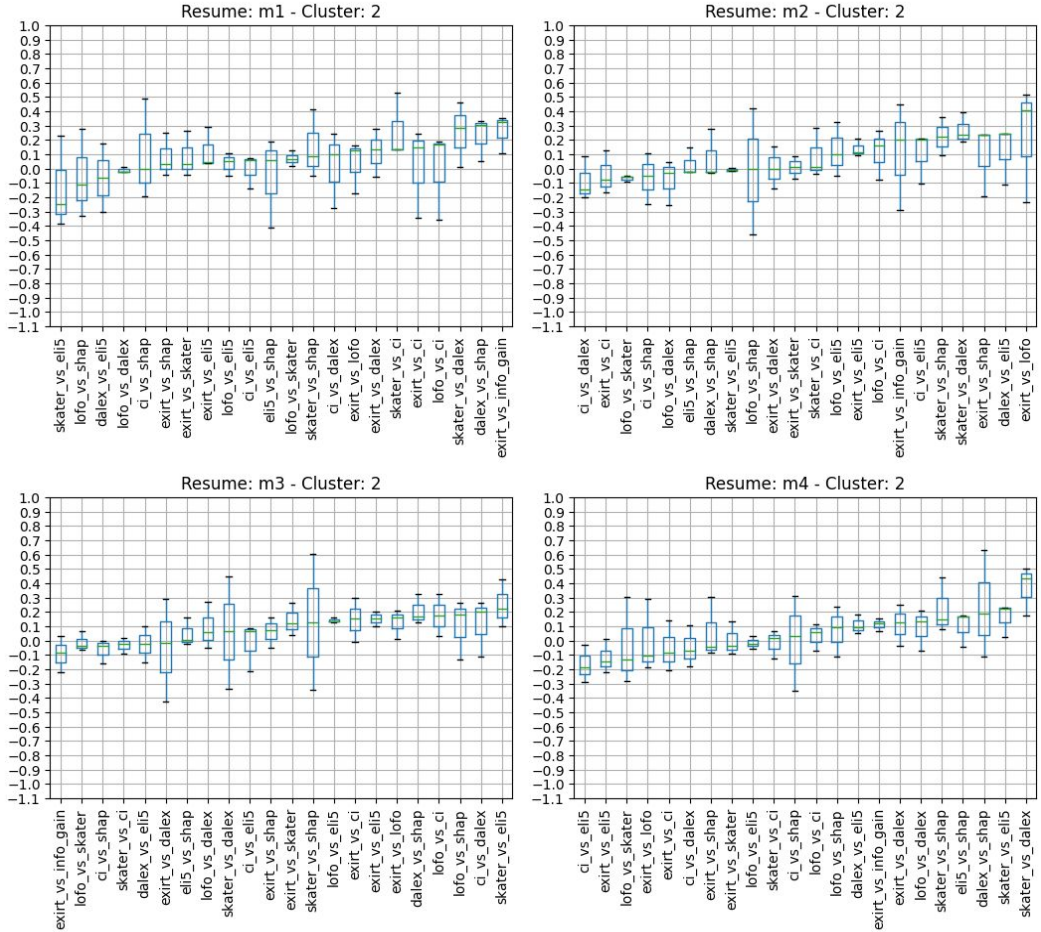
# Data regarding the performance of models M1 to M4 based on cluster 1 datasets



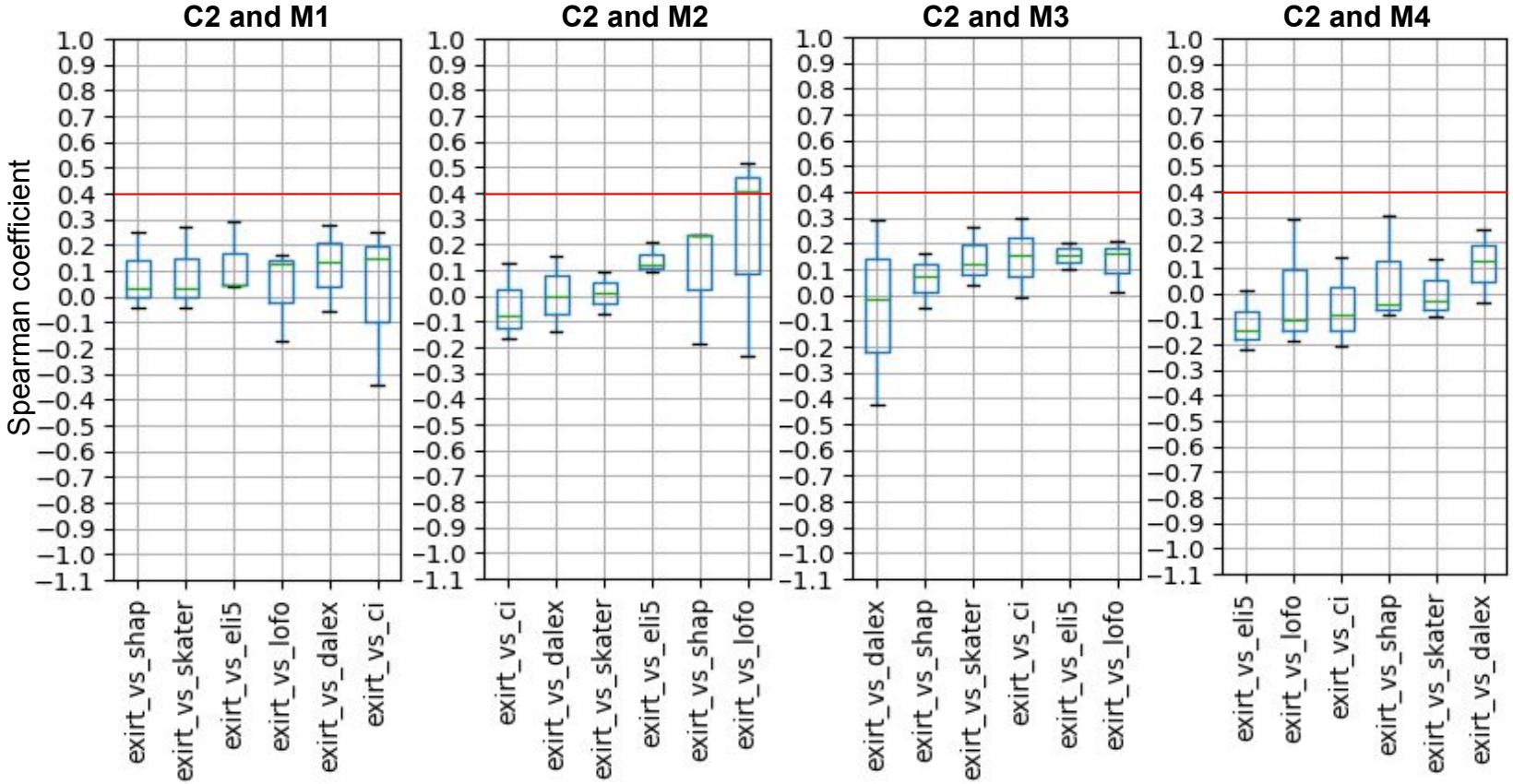



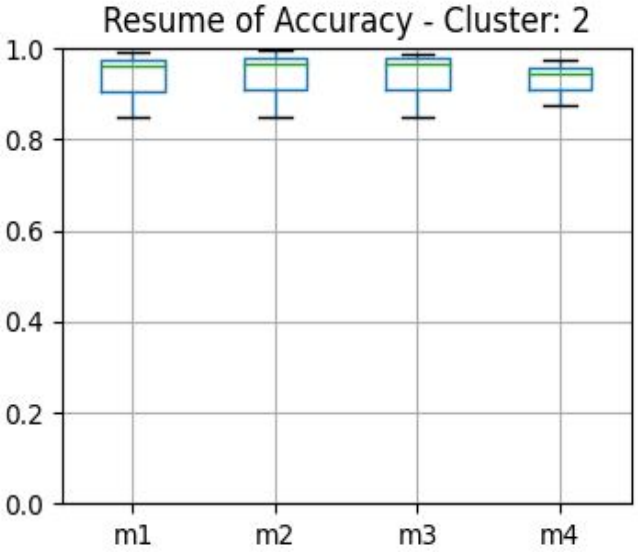
# All pair comparisons of feature relevance ranks for models (M1 to M4) based on cluster 2 datasets

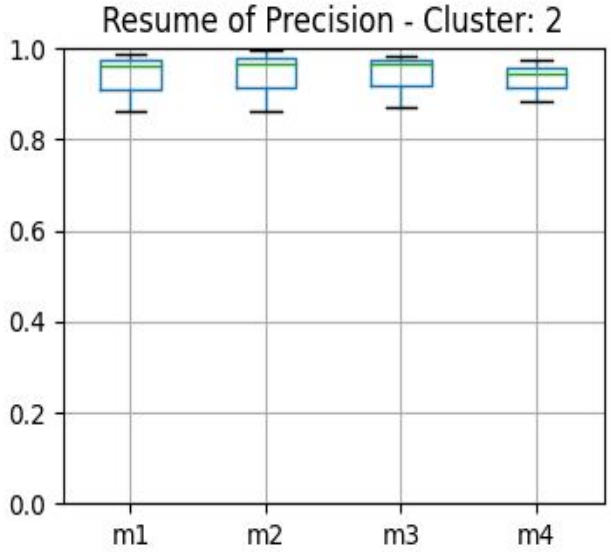


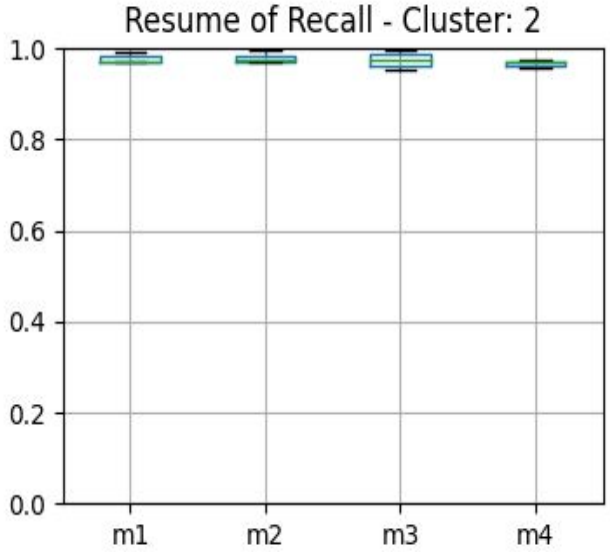
# Only *eXirt* comparisons of feature relevance ranks for models (M1 to M4) based on cluster 2 datasets



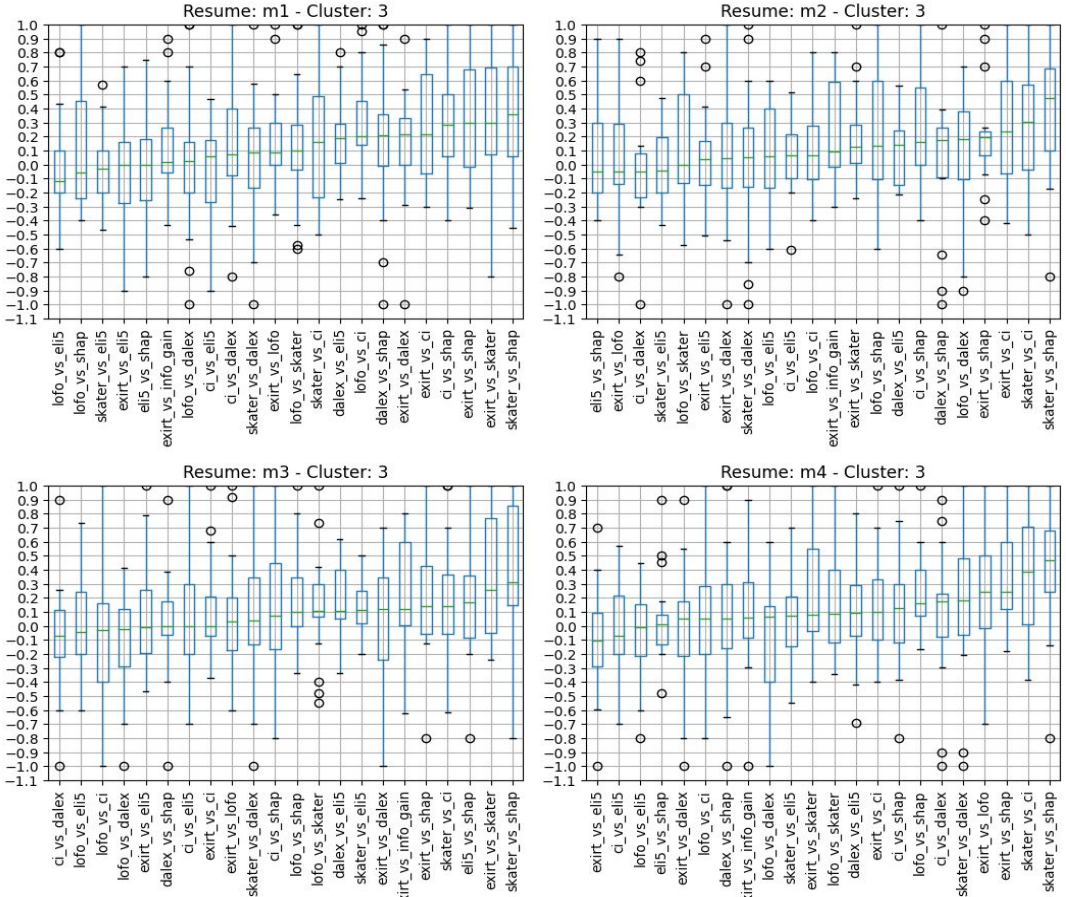
# Data regarding the performance of models M1 to M4 based on cluster 1 datasets



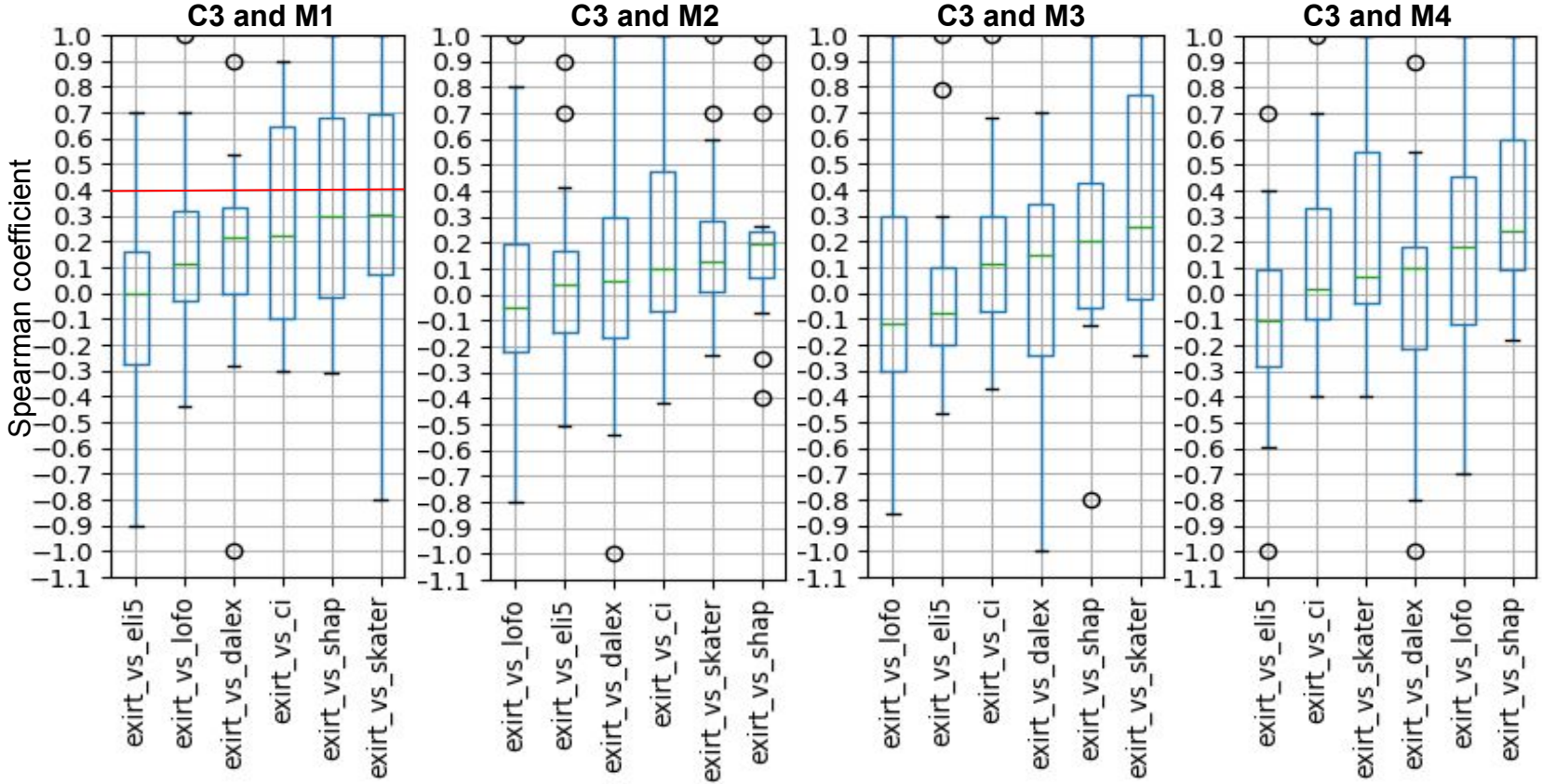



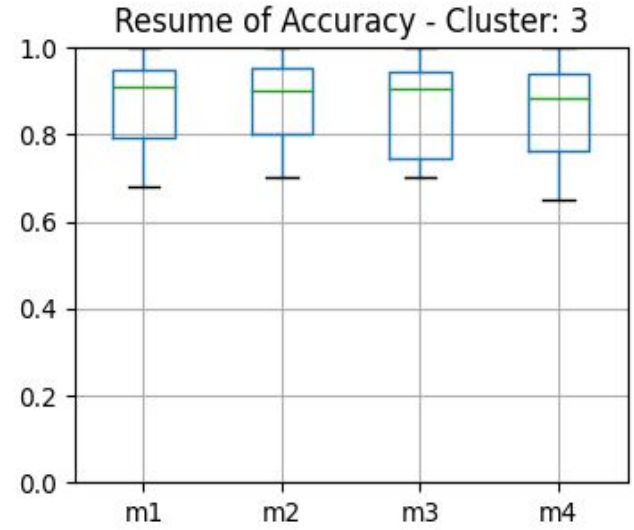
# All pair comparisons of feature relevance ranks for models (M1 to M4) based on cluster 3 datasets

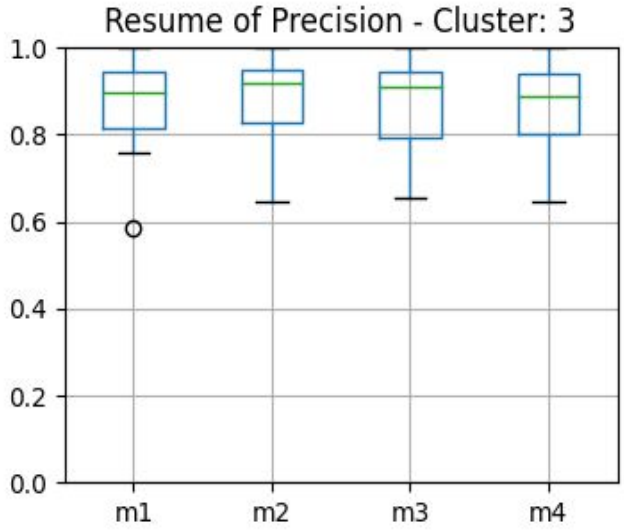


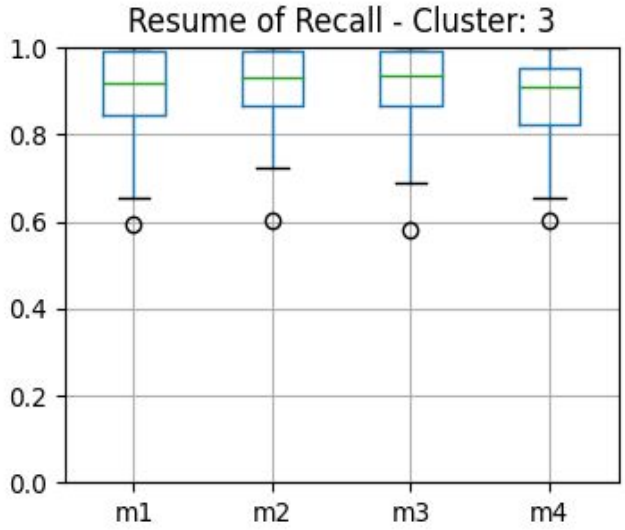
# Only *eXirt* comparisons of feature relevance ranks for models (M1 to M4) based on cluster 3 datasets



# Data regarding the performance of models M1 to M4 based on cluster 3 datasets



# Item Characteristic Curve of the credit-g dataset (test) for models M1 to M4