Does the Model Think As We Expect? Exploring ML Model Logic and Trustworthiness Through Decision Rules

Natalia Andrienko, Gennady Andrienko, Bahavathy Kathirgamanathan

Human-Centred AI

INTRODUCTION

- **Key question:** ML models can be accurate, but do they reason as we expect?
- **Why This Matters:**
  - Trust in ML models is not just about accuracy – it's about understanding *why* they make decisions.
  - A model may produce correct predictions while relying on reasoning that differs from human logic.
  - This misalignment can affect model adoption, interpretation, and decision-making in critical applications.
- **Our goal:** support exploring model alignment with human expectations using visual analytics.
  - We consider models represented by systems of decision rules.
- **Focus of this talk:** What insights we gain about trustworthiness when analyzing rule-based ML models.

RUNNING EXAMPLE: CLASSIFICATION MODEL FOR COVID-19 PREDICTION

- Model was developed using a dataset including daily counts and trips for 52 regions during the COVID-19 pandemic period, normalized by population.
- Data aggregated weekly over 64 weeks, excluding initial outbreak phase.
- Discretized data into four levels for disease incidence and population mobility. Low mobility levels indicate restrictions.
- Focus on interdependencies between disease incidence and mobility levels.
  - Increases in disease incidence may lead to reduced mobility through restrictions, which subsequently contribute to a decrease in disease levels.
  - Conversely, relaxed mobility restrictions may result in increased disease incidence.
  - The effects may become noticeable after a delay.

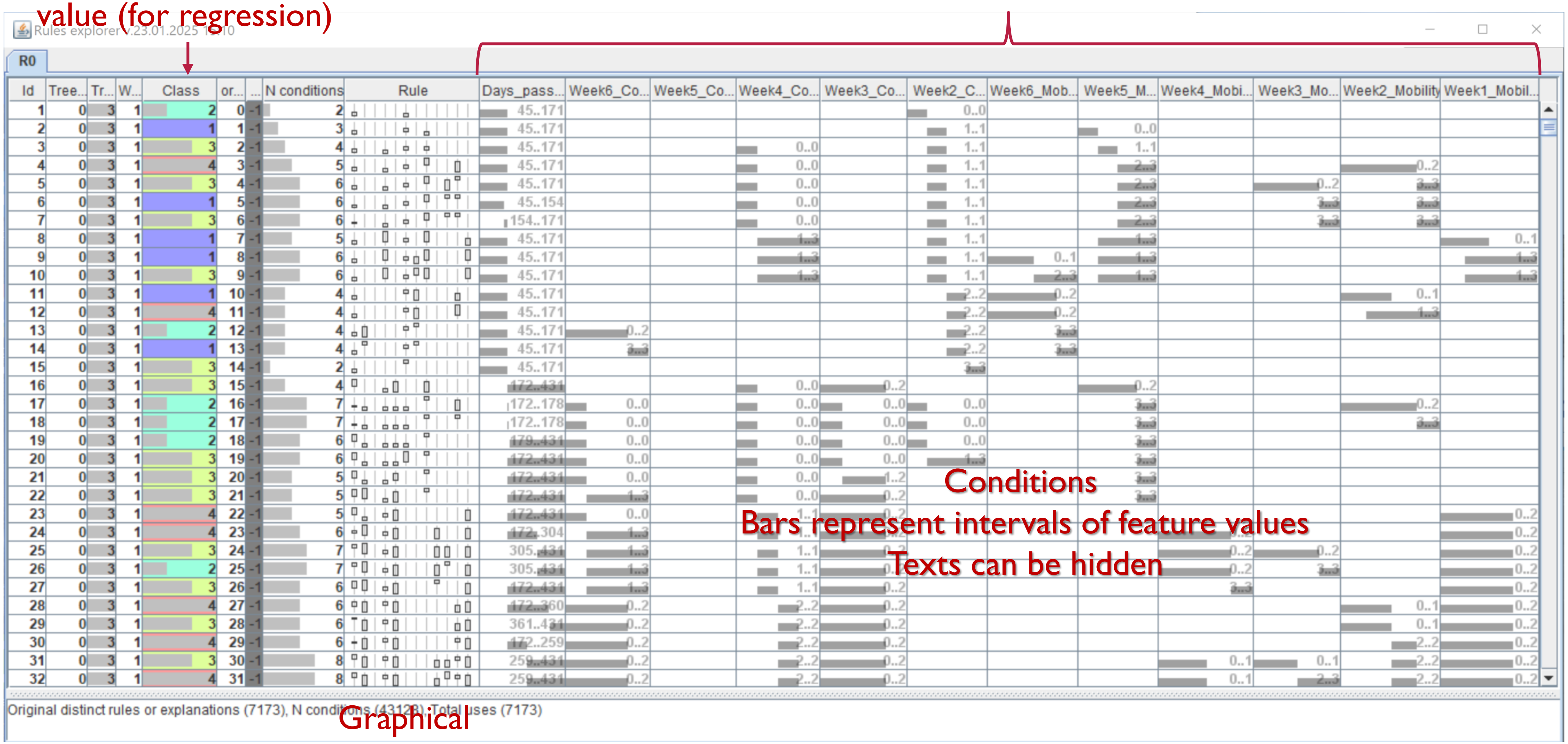
FEATURES USED FOR DERIVING PREDICTIONS

- Temporal features related to COVID-19 levels and mobility trends over six weeks preceding a target event:
  - **COVID-19 Features**: Weekly categorical indicators (Week6_Covid to Week1_Covid) with values c1, c2, c3, and c4, representing increasing severity levels.
  - **Mobility Features**: Weekly categorical indicators (Week6_Mobility to Week1_Mobility) with values m1, m2, m3, and m4, representing mobility levels from low (lockdown) to normal.
- The number of days passed since the start of the pandemic monitoring.
- **Target Class**: The outcome variable categorizing the event into one of the four classes (c1 to c4).
- The categorical values were encoded by numbers from 1 to 4.

REPRESENTATION OF RULES IN A TABLE

Predicted class or value (for regression)

features



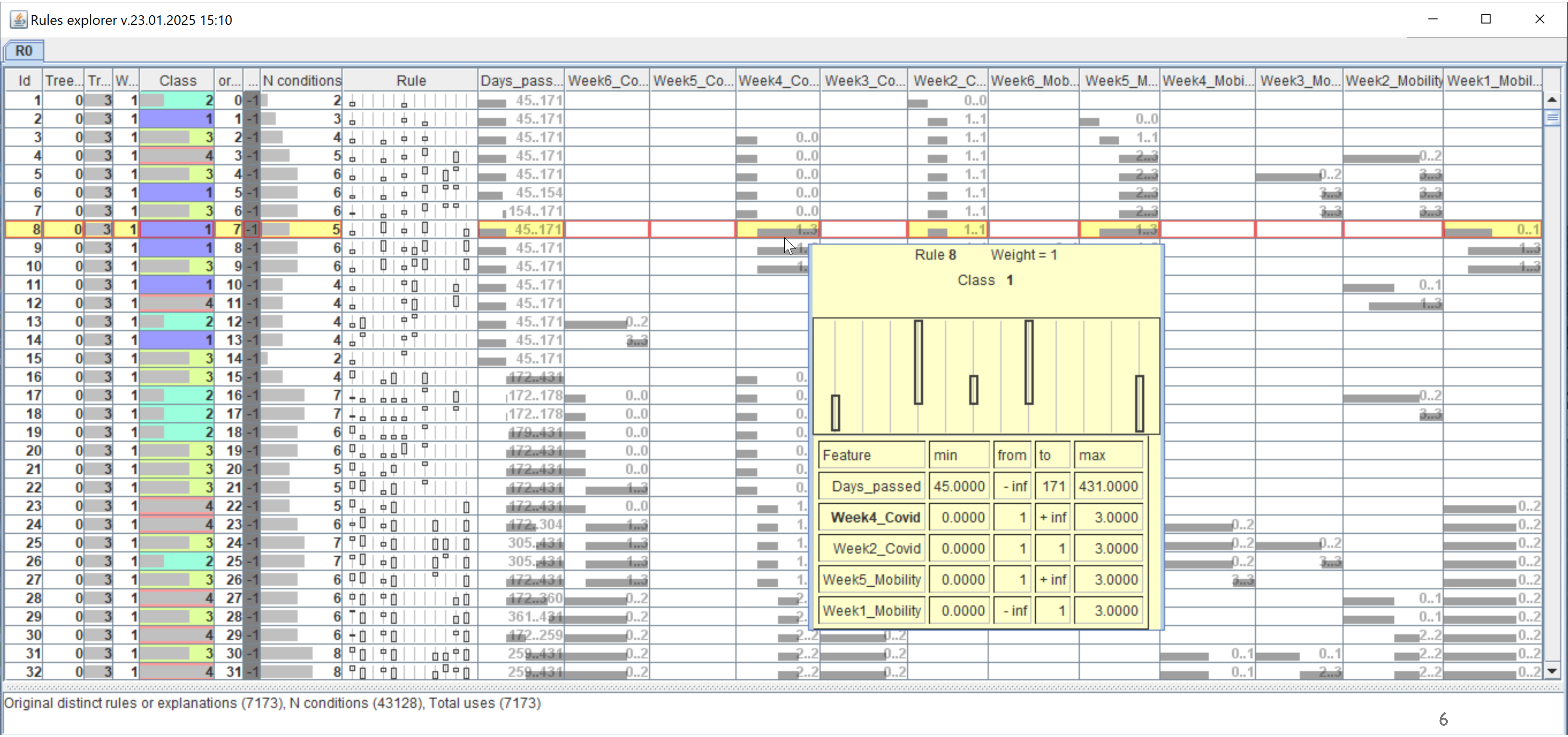
Conditions

Bars represent intervals of feature values

Texts can be hidden

Graphical representations of rules

GRAPHICAL REPRESENTATION OF A RULE



CHALLENGES IN UNDERSTANDING MODEL LOGIC

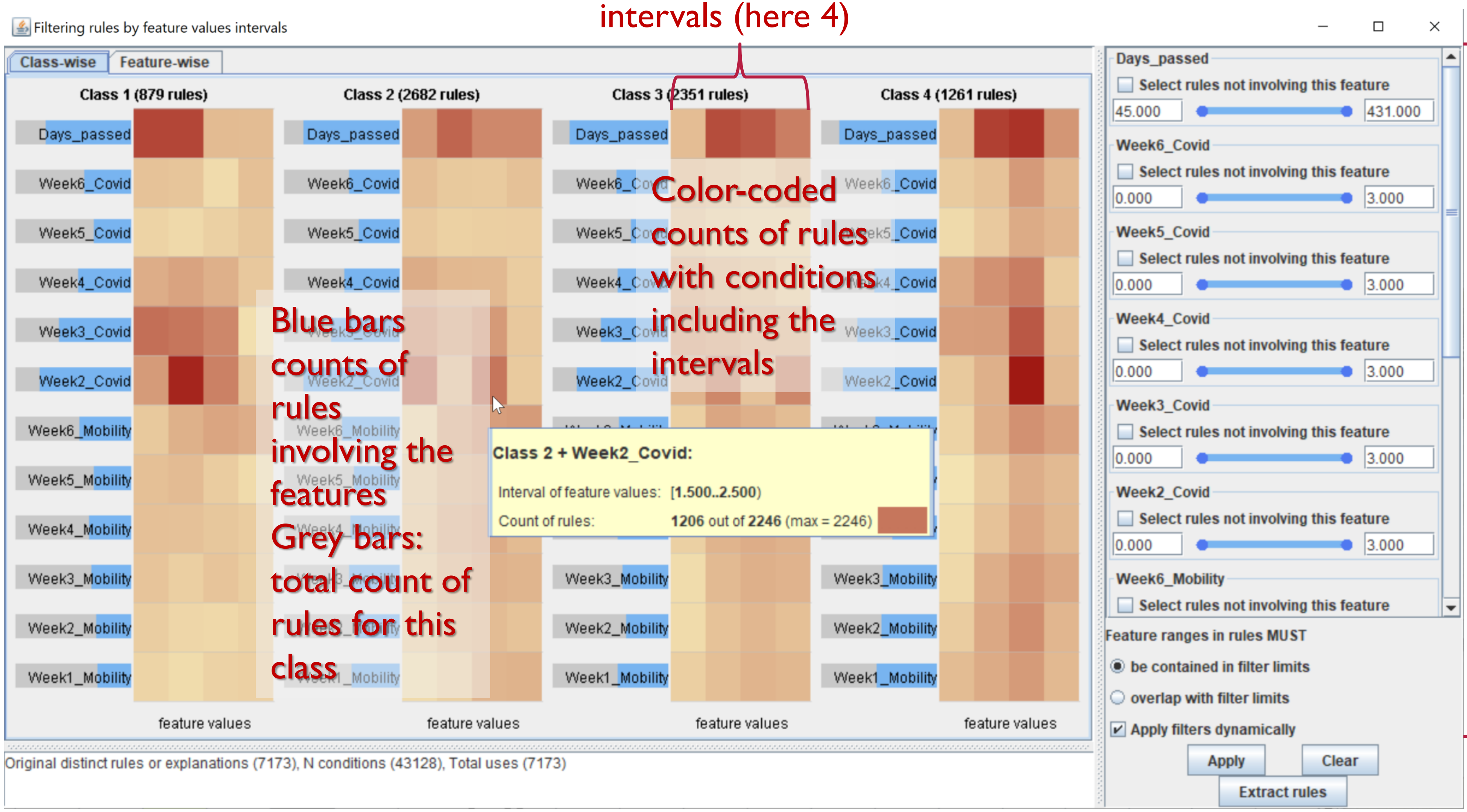
- **We can read and understand the conditions of each rule** >> Representation of a model by rules should allow us to inspect model reasoning, but…
- **The rules are too many** >> detailed examination of individual rules is impractical.
- **The features interact and work jointly** >> investigation of the impacts of individual features on the predictions is insufficient.
- **Our visual analytics solutions:**
    - Provide an overview
    - Enable querying and selection
    - Aggregate and generalize

This research builds on previous work described in paper Adilova, L., Kamp, M., Andrienko, G. and Andrienko, N. "**Re-interpreting rules interpretability**". *International Journal of Data Science and Analytics* (2023). doi:10.1007/s41060-023-00398-5.

DISTRIBUTION OF FEATURE VALUE INTERVALS

Class-wise view

Chosen number of intervals (here 4)

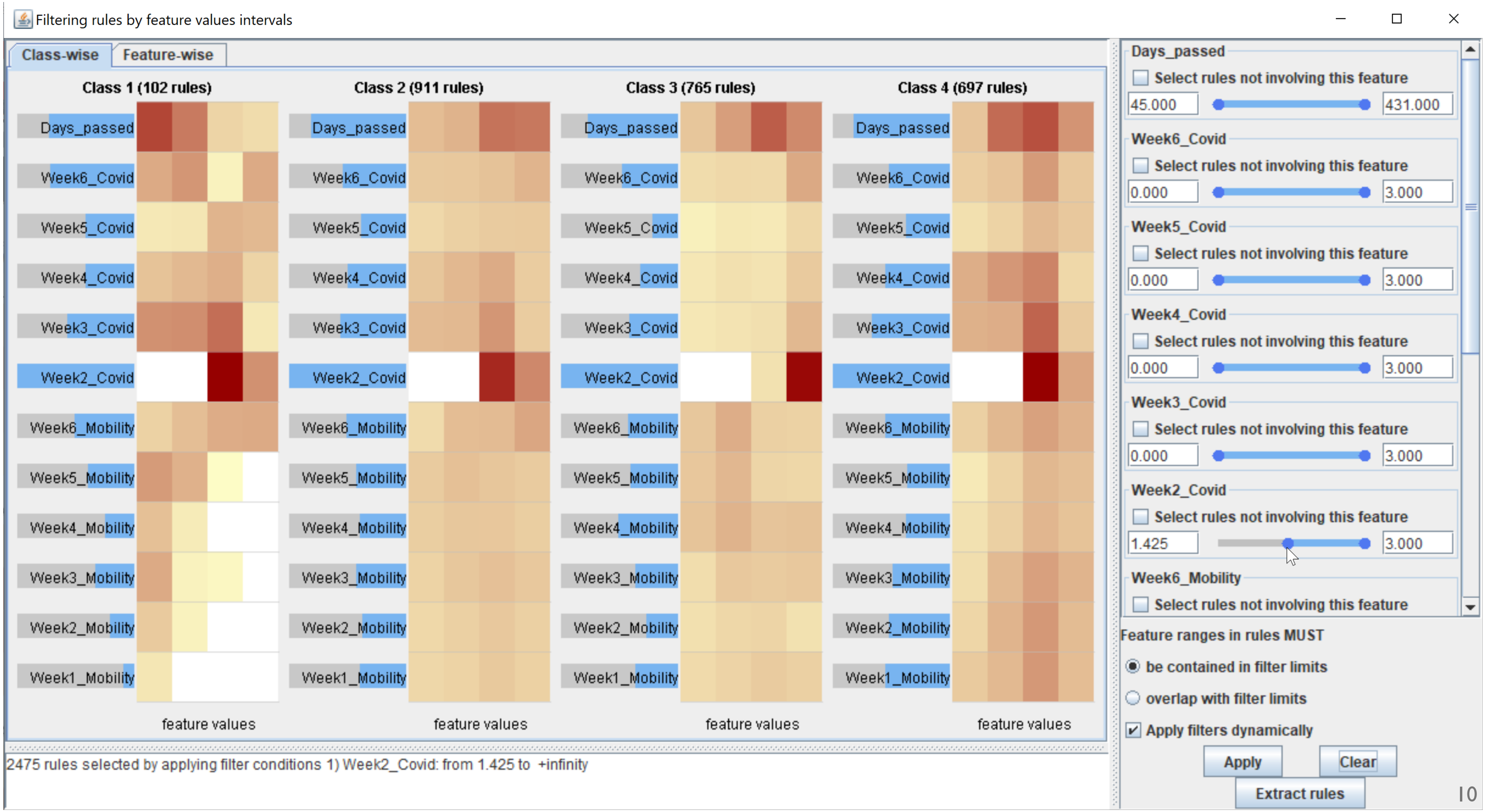


Controls for interactive filtering

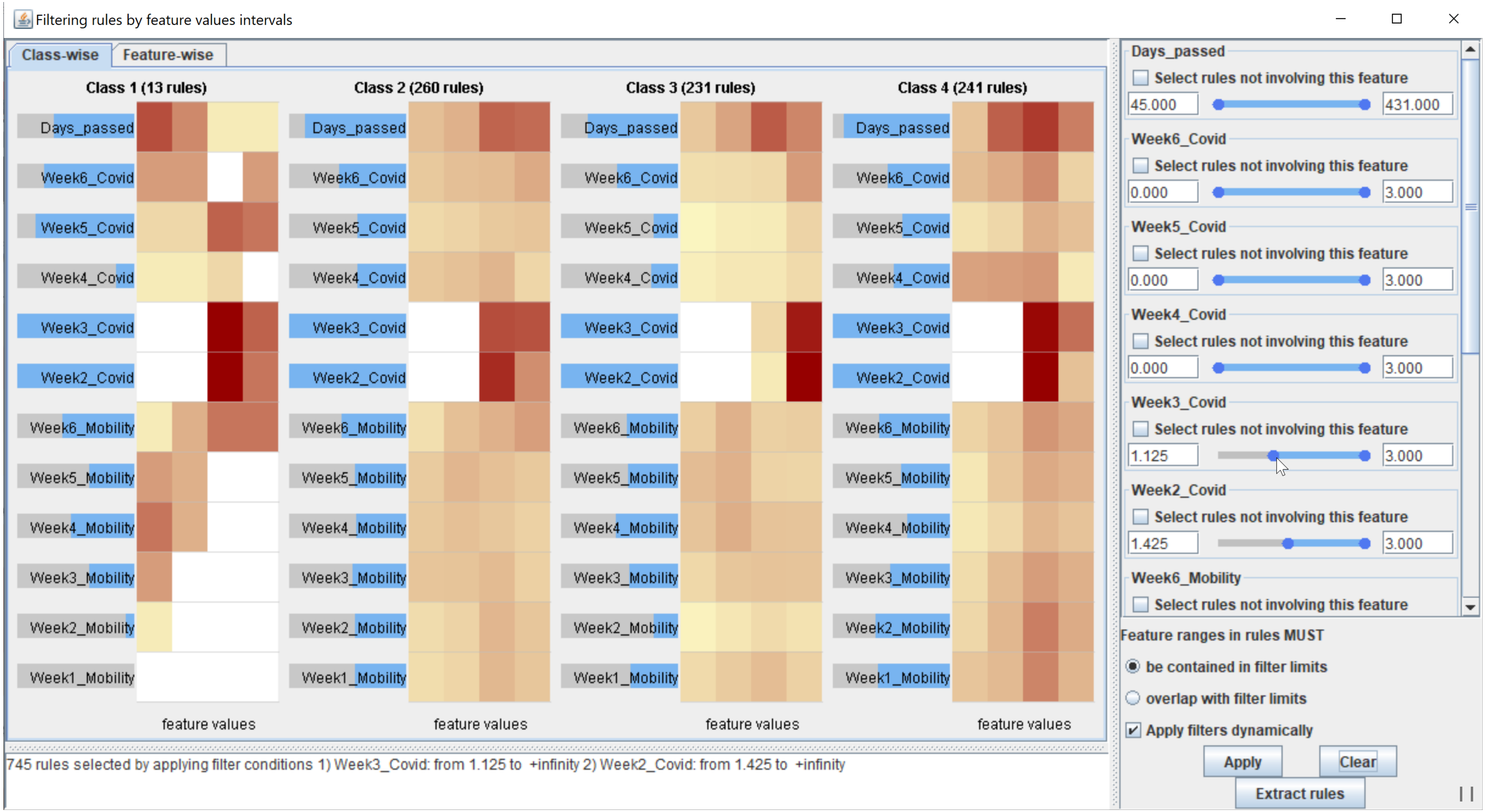
DISTRIBUTION OF FEATURE VALUE INTERVALS

Feature-wise view

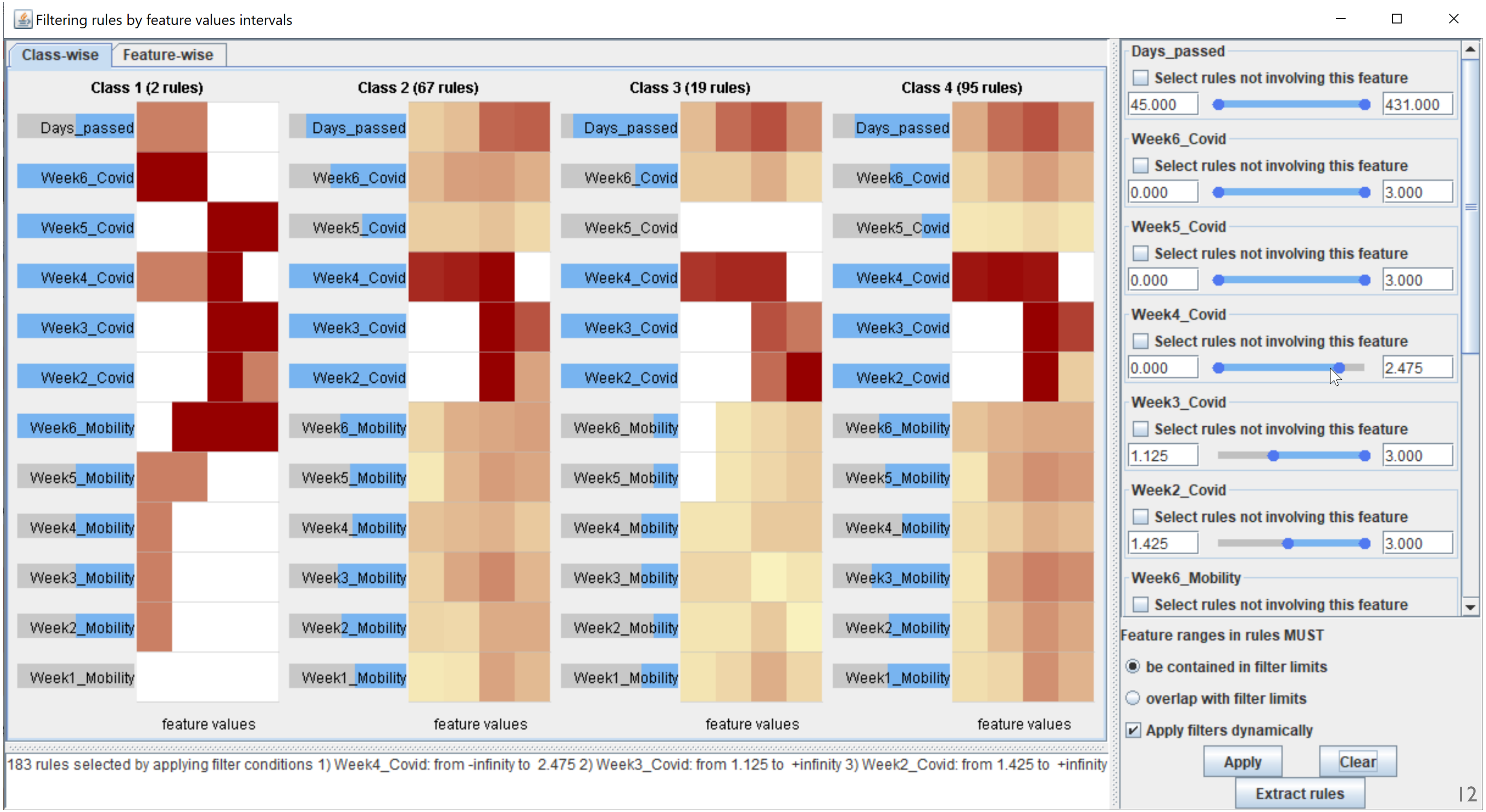
INTERACTIVE FILTERING



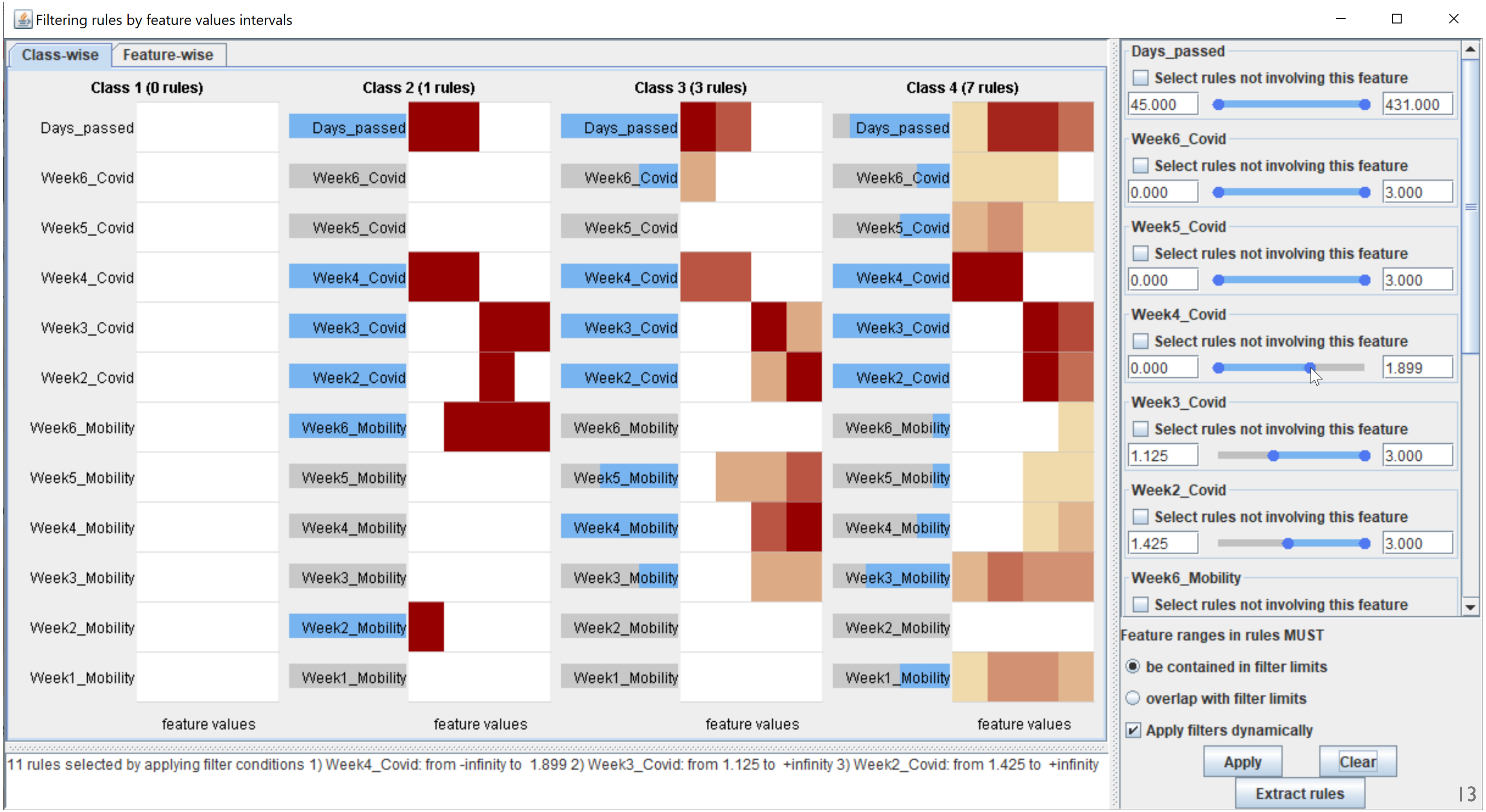
INTERACTIVE FILTERING



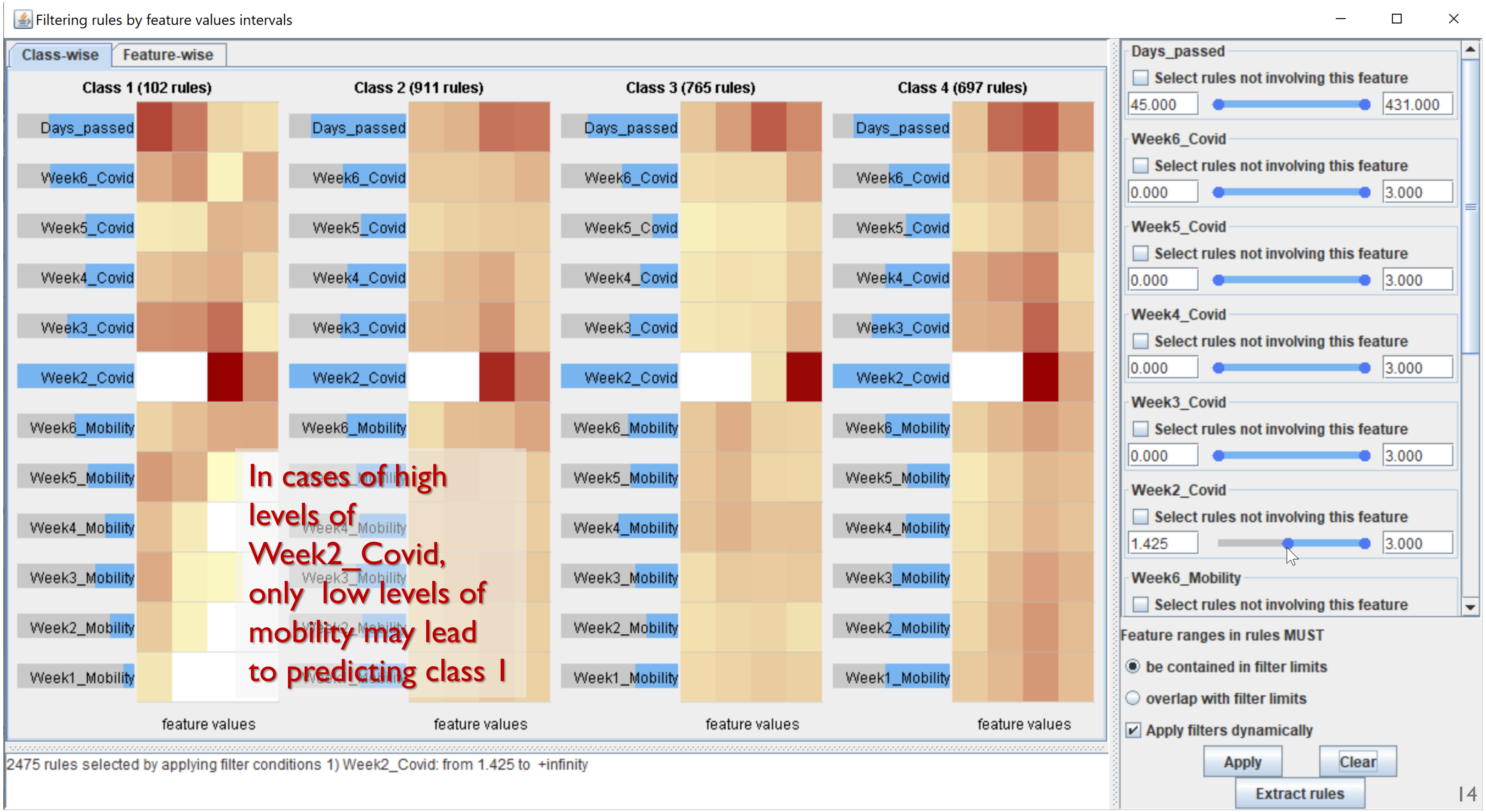
INTERACTIVE FILTERING



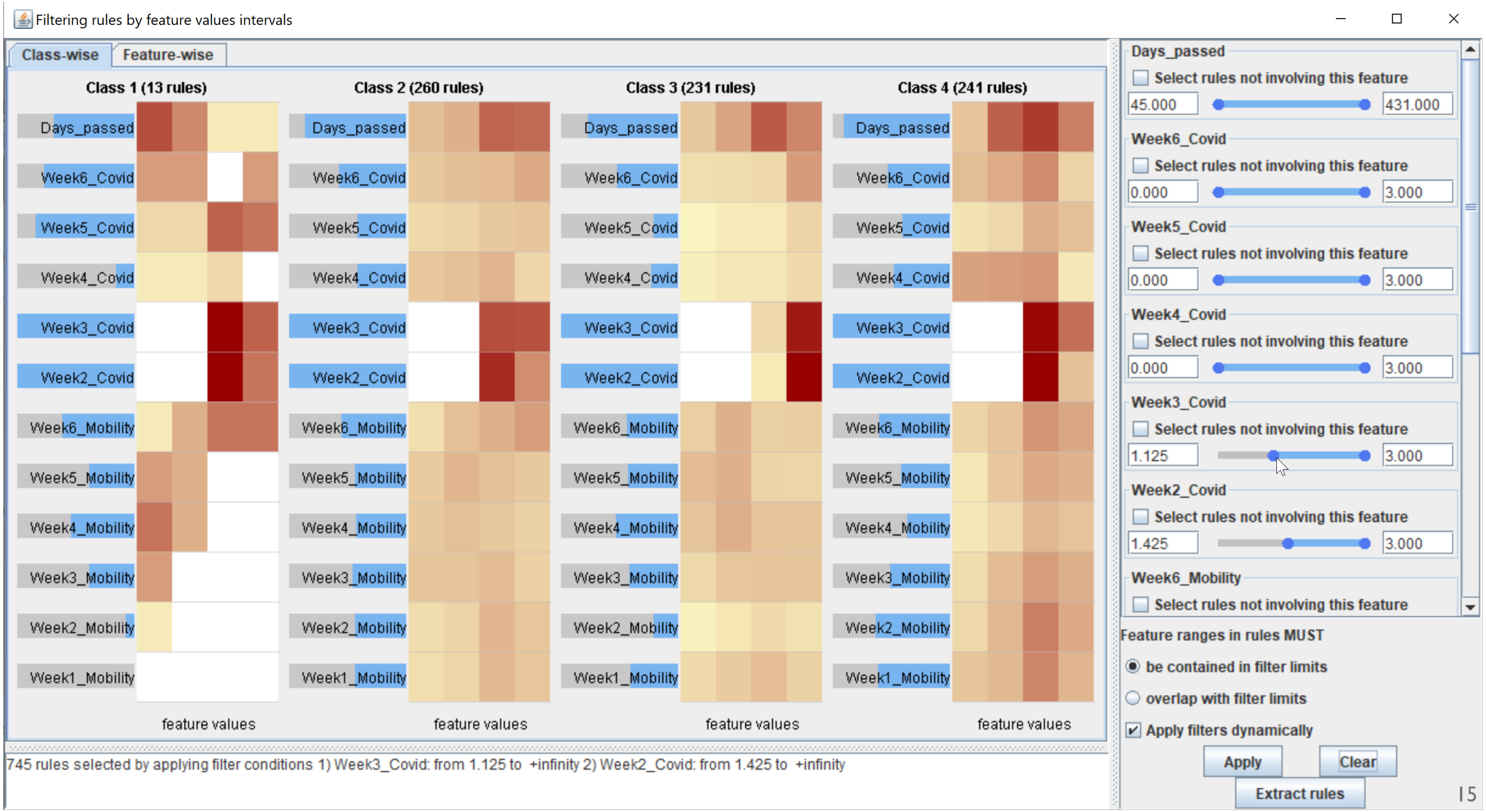
INTERACTIVE FILTERING



RELATIONSHIPS BETWEEN FEATURES



COMBINED EFFECTS OF 2 OR MORE FEATURES

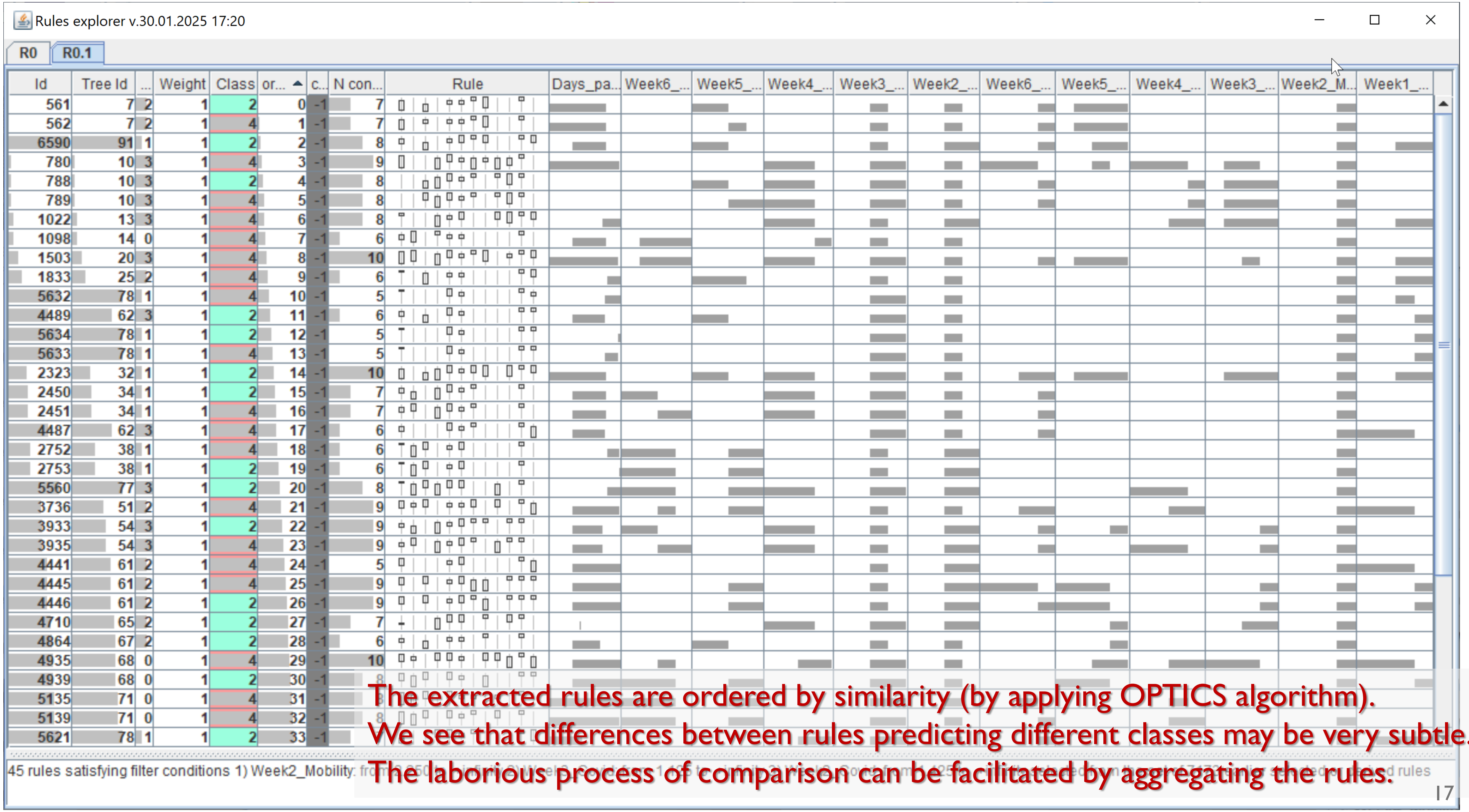


SURPRISING FINDINGS

21 rules predict decrease of pandemic level from classes 3 or 4 to class 2 despite high level of mobility

The distributions of the feature values in the two groups of rules appear quite similar

EXTRACTED RULES SATISFYING THE FILTER

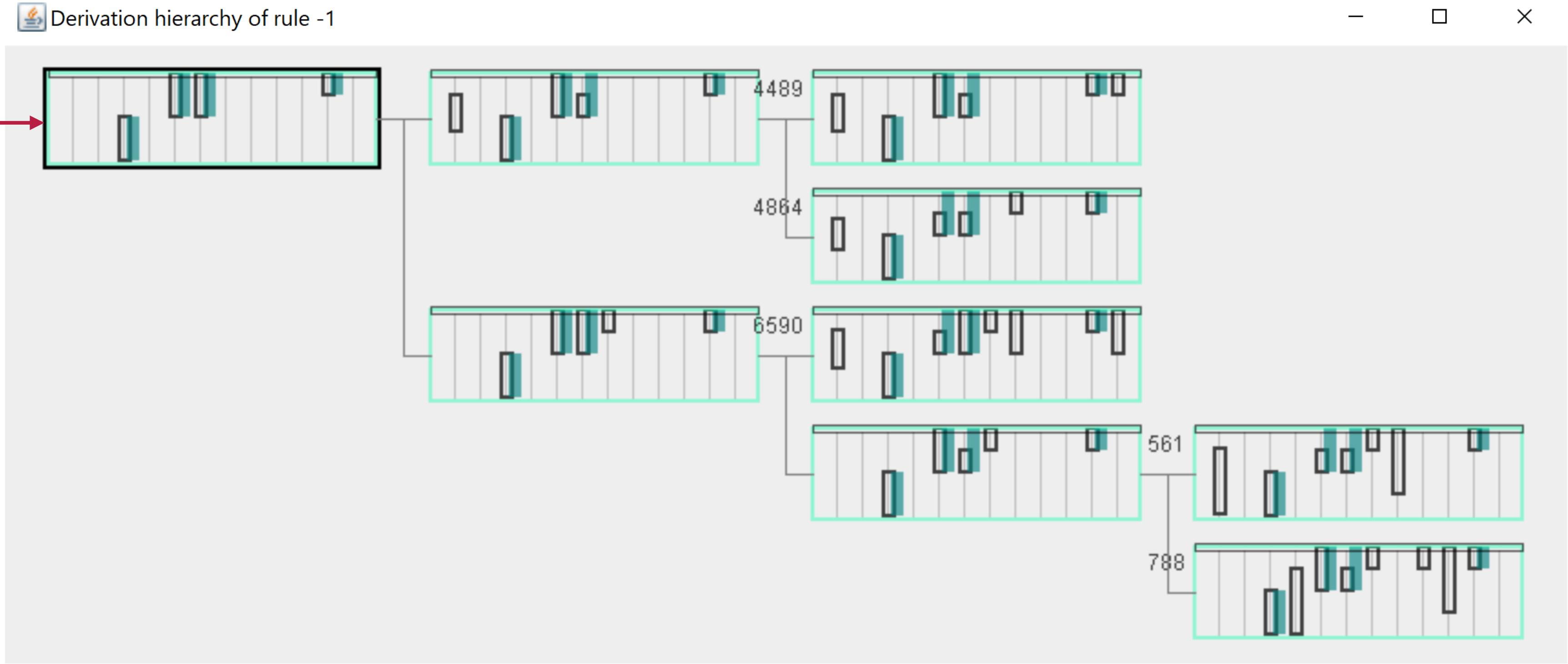


Id	Tree Id	...	Weight	Class	or...	c...	N con...
561	7	2	1	2	0	-1	7
562	7	2	1	4	1	-1	7
6590	91	1	1	2	2	-1	8
780	10	3	1	4	3	-1	9
788	10	3	1	2	4	-1	8
789	10	3	1	4	5	-1	8
1022	13	3	1	4	6	-1	8
1098	14	0	1	4	7	-1	6
1503	20	3	1	4	8	-1	10
1833	25	2	1	4	9	-1	6
5632	78	1	1	4	10	-1	5
4489	62	3	1	2	11	-1	6
5634	78	1	1	2	12	-1	5
5633	78	1	1	4	13	-1	5
2323	32	1	1	2	14	-1	10
2450	34	1	1	2	15	-1	7
2451	34	1	1	4	16	-1	7
4487	62	3	1	4	17	-1	6
2752	38	1	1	4	18	-1	6
2753	38	1	1	2	19	-1	6
5560	77	3	1	2	20	-1	8
3736	51	2	1	4	21	-1	9
3933	54	3	1	2	22	-1	9
3935	54	3	1	4	23	-1	9
4441	61	2	1	4	24	-1	5
4445	61	2	1	4	25	-1	9
4446	61	2	1	2	26	-1	9
4710	65	2	1	2	27	-1	7
4864	67	2	1	2	28	-1	6
4935	68	0	1	4	29	-1	10
4939	68	0	1	2	30	-1	8
5135	71	0	1	4	31	-1	8
5139	71	0	1	4	32	-1	8
5621	78	1	1	2	33	-1	[illegible]

The extracted rules are ordered by similarity (by applying OPTICS algorithm).
We see that differences between rules predicting different classes may be very subtle.
The laborious process of comparison can be facilitated by aggregating the rules.

ITERATIVE AGGREGATION AND GENERALISATION OF RULES

This final rule was derived from 5 original rules.
The glyph representing the rule is interactively selected.
The conditions from this rule are represented by blue vertical bars in all glyphs for comparison.



Rule R_1 *covers* rule R_2 if R_1 is applicable to all instances where R_2 is applicable.
R_1 is *more general* than R_2.
The operation of uniting 2 or more rules creates a more general rule covering each of the original rules.

Uniting rules predicting the same class:

- Identical conditions remain in the resulting rules.
- Differing intervals are joined by creating a covering interval.
- Conditions with features missing in one of the rules are omitted.

A RESULT OF AGGREGATING 45 SELECTED RULES

Coherence threshold = 0.75 (fraction of allowed exceptions)

Rules explorer v.30.01.2025 17:20

R0 | R0.1 | R0.1.1

Id	...	...	Weight	Class	...	... ▲	c...	N con...	Rule	N right...	N wro...	Coher...	N unit...	Depth ...	Days_...	Week...	Week5...	Week...	Week3...	Week...	Week6...	Week5...	Week4...	Week3...	Week2...	Week1...
1833	25	2	1	4	1	0	-1	6		1	0	100.00...	0	1												
6790	94	0	1	4	1	1	-1	7		1	0	100.00...	0	1												
-1	-1	-1	3	4	5	2	-1	5		5	0	100.00...	5	3												
-1	-1	-1	3	2	3	3	-1	5		3	1	75.000...	4	3												
-1	-1	-1	4	4	4	4	-1	4		4	0	100.00...	6	3												
-1	-1	1	2	2	3	5	-1	4		3	0	100.00...	3	2												
-1	-1	-1	2	2	2	6	-1	6		3	0	100.00...	2	2												
-1	-1	-1	2	4	3	7	-1	5		3	0	100.00...	3	2												
-1	-1	-1	2	2	2	8	-1	4		2	0	100.00...	2	2												
-1	-1	-1	5	2	5	9	-1	4		7	0	100.00...	8	4												
-1	-1	-1	2	4	3	10	-1	5		3	0	100.00...	3	2												
-1	-1	3	2	4	2	11	-1	7		2	0	100.00...	2	2												
-1	-1	-1	2	4	2	12	-1	6		2	0	100.00...	2	2												
-1	-1	-1	3	4	3	13	-1	4		6	2	75.000...	4	3												
-1	-1	-1	2	2	2	14	-1	8		2	0	100.00...	2	2												
-1	-1	-1	2	2	2	15	-1	5		3	0	100.00...	2	2												
-1	-1	0	2	2	2	16	-1	7		2	0	100.00...	2	2												

These rules have exceptions

Differences between 2 similar rules predicting distinct classes cannot be grounded in domain knowledge.

17 aggregated and generalized rules obtained from the set of 45 earlier selected or derived rules using the following parameter settings: min coherence = 0.750; stepwise aggregation starting from 1.000 with step 0.050

Rules covered by one of the "rough" rules:

Id	...	Tr...	Weight	Class	N uses	or... ▲	...	N con...	Rule	N ri...	N wr...	Cohe...	N unit...	Dept...	Days_...	Week...	Week5...	Week...	Week3...	Week...	Week6...	Week...	Week4_...	Week...	Week2_...	Week1_...
1	-1	-1	3	2	3	0	-1	5		3	1	75.00...	4	3												
2	61	2	1	2	1	1	-1	9		1	0	100.0...	0	0												
3	68	0	1	2	1	2	-1	8		1	0	100.0...	0	0												
4	82	1	1	2	1	3	-1	10		1	0	100.0...	0	0												
5	61	2	1	4	1	4	-1	9		1	0	100.0...	0	0												

Exception (a rule predicting a different class)

A RESULT OF AGGREGATING 45 SELECTED RULES

Coherence threshold = 0.75 (fraction of allowed exceptions)

Rules explorer v.30.01.2025 17:20

R0 | R0.1 | R0.1.1

Id	...	...	Weight	Class	...	...	c...	N con...	Rule	N right...	N wro...	Coher...	N unit...	Depth ...	Days_...	Week...	Week5...	Week...	Week3...	Week...	Week6...	Week5...	Week4...	Week3...	Week2...	Week1...
1833	25	2	1	4	1	0	-1	6		1	0	100.00...	0	1												
6790	94	0	1	4	1	1	-1	7		1	0	100.00...	0	1												
-1	-1	-1	3	4	5	2	-1	5		5	0	100.00...	5	3												
-1	-1	-1	3	2	3	3	-1	5		3	1	75.000...	4	3												
-1	-1	-1	4	4	4	4	-1	4		4	0	100.00...	6	3												
-1	-1	1	2	2	3	5	-1	4		3	0	100.00...	3	2												
-1	-1	-1	2	2	2	6	-1	6		3	0	100.00...	2	2												
-1	-1	-1	2	4	3	7	-1	5		3	0	100.00...	3	2												
-1	-1	-1	2	2	2	8	-1	4		2	0	100.00...	2	2												
-1	-1	-1	5	2	5	9	-1	4		7	0	100.00...	8	4												
-1	-1	-1	2	4	3	10	-1	5		3	0	100.00...	3	2												
-1	-1	3	2	4	2	11	-1	7		2	0	100.00...	2	2												
-1	-1	-1	2	4	2	12	-1	6		2	0	100.00...	2	2												
-1	-1	-1	3	4	3	13	-1	4		6	2	75.000...	4	3												
-1	-1	-1	2	2	2	14	-1	8		2	0	100.00...	2	2												
-1	-1	-1	2	2	2	15	-1	5		3	0	100.00...	2	2												
-1	-1	0	2	2	2	16	-1	7		2	0	100.00...	2	2												

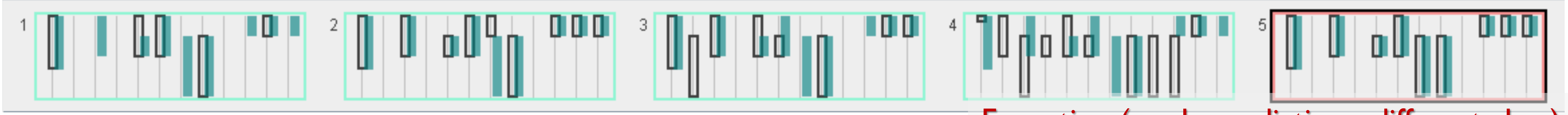
These rules have exceptions

Differences between 2 similar rules predicting distinct classes cannot be grounded in domain knowledge.

17 aggregated and generalized rules obtained from the set of 45 earlier selected or derived rules using the following parameter settings: min coherence = 0.750; stepwise aggregation starting from 1.000 with step 0.050

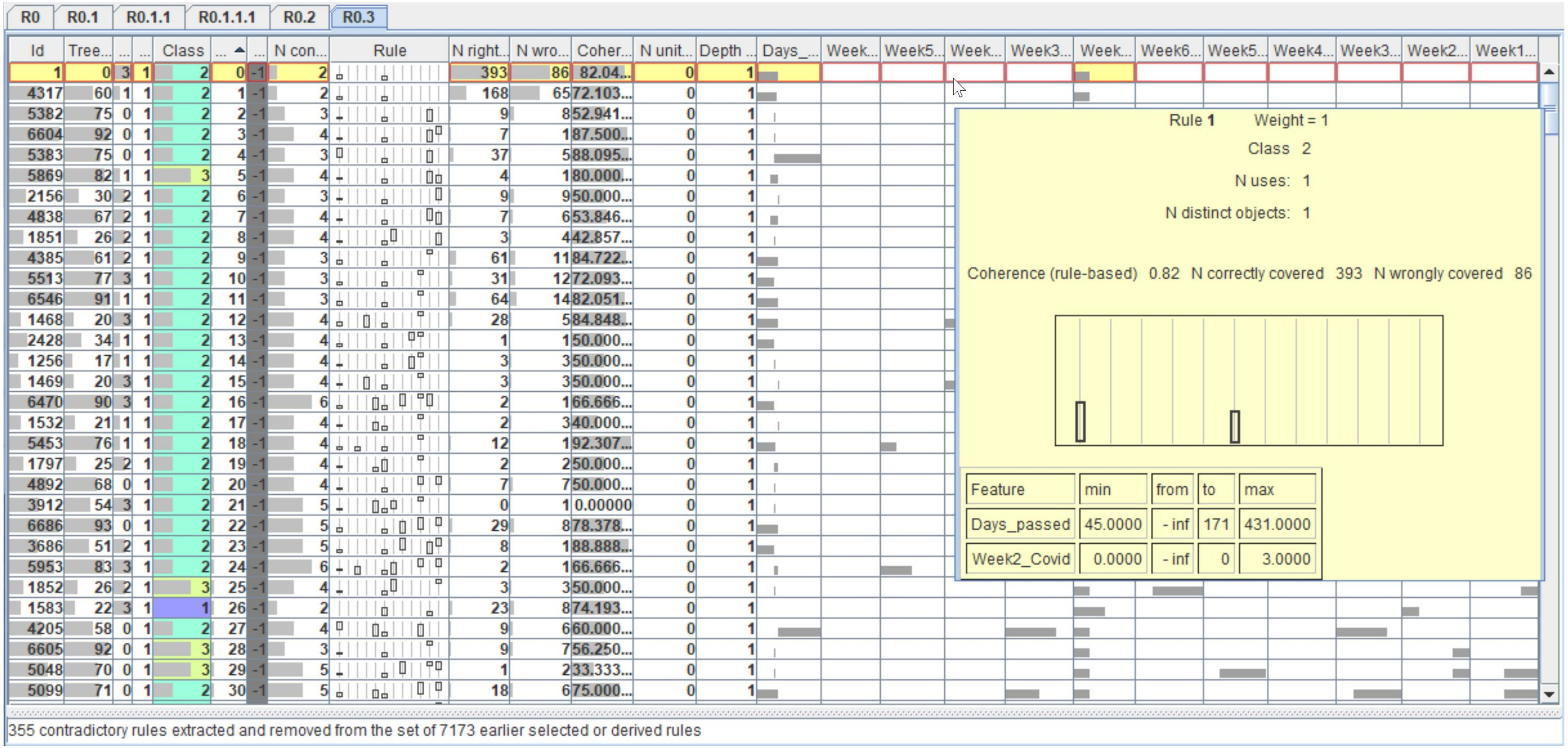
Rules covered by one of the "rough" rules:

All coverages of rule # 1



Exception (a rule predicting a different class)

DETECTING AND REMOVING CONTRADICTORY RULES

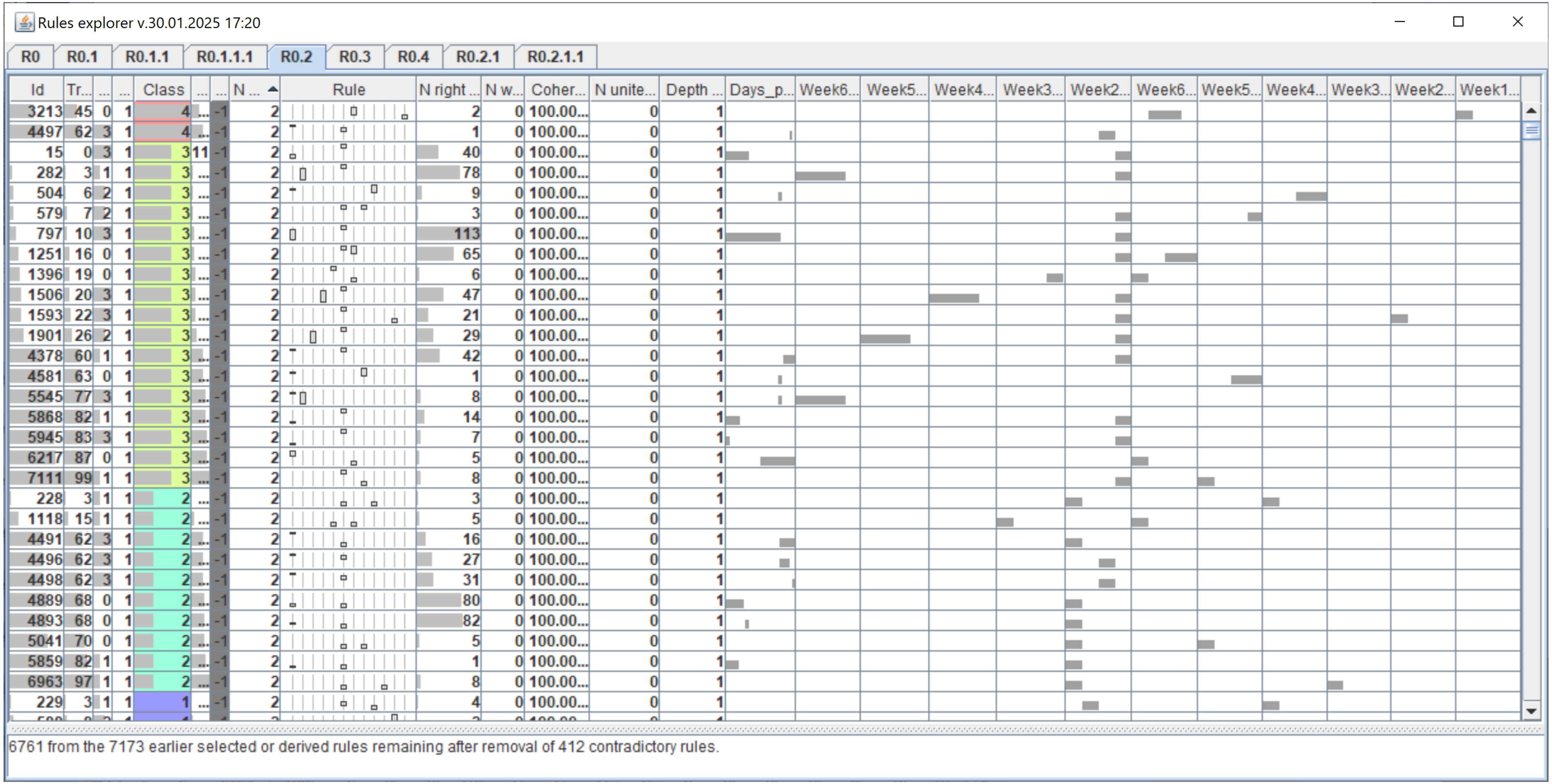


Id	Tree...	...	...	Class	...	...	N con...	Rule	N right...	N wro...	Coher...	N unit...	Depth ...
1	0	3	1	2	0	-1	2		393	86	82.04...	0	1
4317	60	1	1	2	1	-1	2		168	65	72.103...	0	1
5382	75	0	1	2	2	-1	3		9	8	52.941...	0	1
6604	92	0	1	2	3	-1	4		7	1	87.500...	0	1
5383	75	0	1	2	4	-1	3		37	5	88.095...	0	1
5869	82	1	1	3	5	-1	4		4	1	80.000...	0	1
2156	30	2	1	2	6	-1	3		9	9	50.000...	0	1
4838	67	2	1	2	7	-1	4		7	6	53.846...	0	1
1851	26	2	1	2	8	-1	4		3	4	42.857...	0	1
4385	61	2	1	2	9	-1	3		61	11	84.722...	0	1
5513	77	3	1	2	10	-1	3		31	12	72.093...	0	1
6546	91	1	1	2	11	-1	3		64	14	82.051...	0	1
1468	20	3	1	2	12	-1	4		28	5	84.848...	0	1
2428	34	1	1	2	13	-1	4		1	1	50.000...	0	1
1256	17	1	1	2	14	-1	4		3	3	50.000...	0	1
1469	20	3	1	2	15	-1	4		3	3	50.000...	0	1
6470	90	3	1	2	16	-1	6		2	1	66.666...	0	1
1532	21	1	1	2	17	-1	4		2	3	40.000...	0	1
5453	76	1	1	2	18	-1	4		12	1	92.307...	0	1
1797	25	2	1	2	19	-1	4		2	2	50.000...	0	1
4892	68	0	1	2	20	-1	4		7	7	50.000...	0	1
3912	54	3	1	2	21	-1	5		0	1	0.00000	0	1
6686	93	0	1	2	22	-1	5		29	8	78.378...	0	1
3686	51	2	1	2	23	-1	5		8	1	88.888...	0	1
5953	83	3	1	2	24	-1	6		2	1	66.666...	0	1
1852	26	2	1	3	25	-1	4		3	3	50.000...	0	1
1583	22	3	1	1	26	-1	2		23	8	74.193...	0	1
4205	58	0	1	2	27	-1	4		9	6	60.000...	0	1
6605	92	0	1	3	28	-1	3		9	7	56.250...	0	1
5048	70	0	1	3	29	-1	5		1	2	33.333...	0	1
5099	71	0	1	2	30	-1	5		18	6	75.000...	0	1

Feature	min	from	to	max
Days_passed	45.0000	- inf	171	431.0000
Week2_Covid	0.0000	- inf	0	3.0000

Our model contains 355 contradictory rules. Some of them include only a few conditions. According to the domain and/or commonsense knowledge, these conditions cannot be sufficient for making valid predictions.

MANY NON-CONTRADICTORY RULES ALSO INCLUDE TOO FEW CONDITIONS



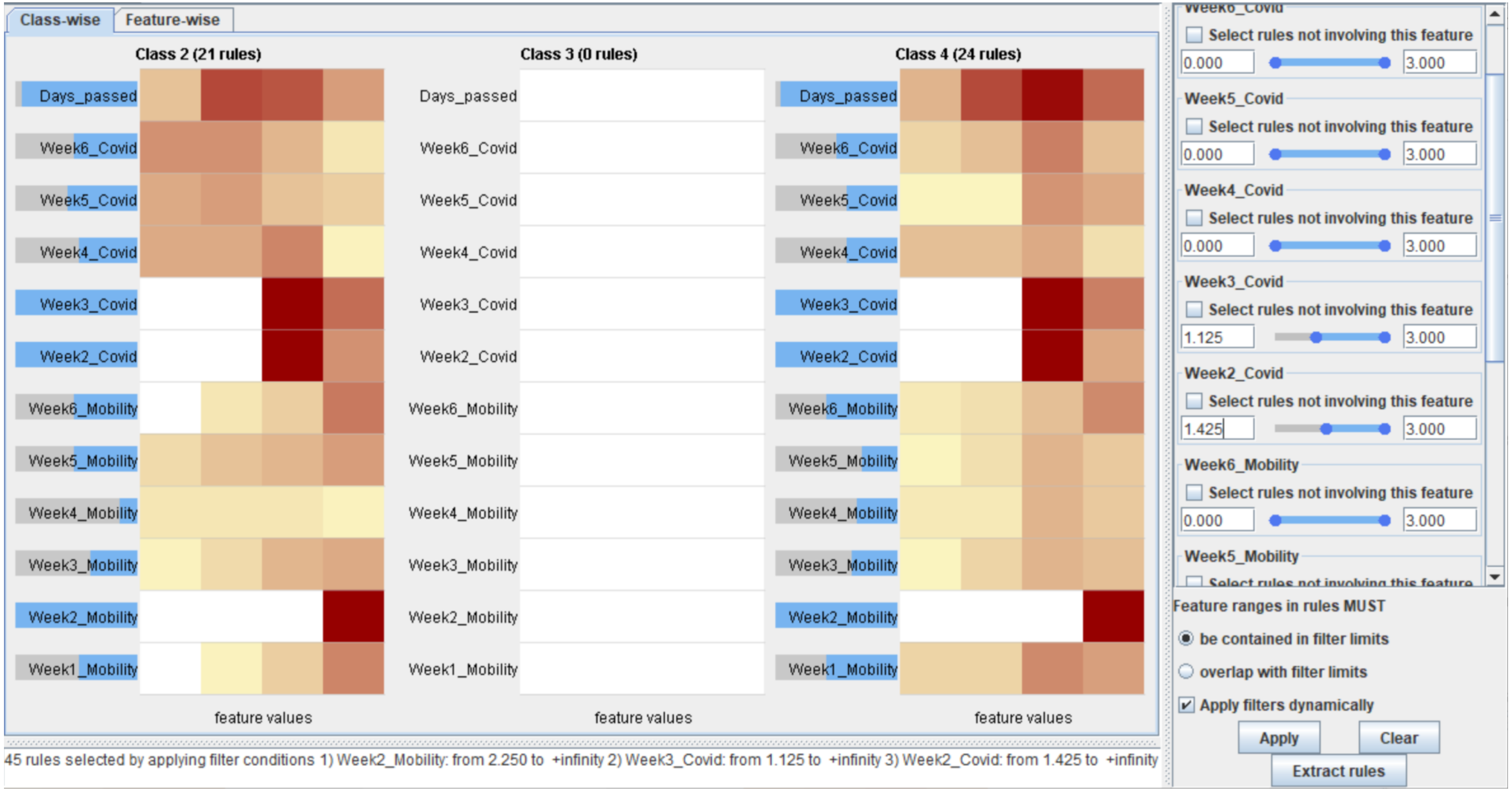
Rules explorer v.30.01.2025 17:20

R0 | R0.1 | R0.1.1 | R0.1.1.1 | R0.2 | R0.3 | R0.4 | R0.2.1 | R0.2.1.1

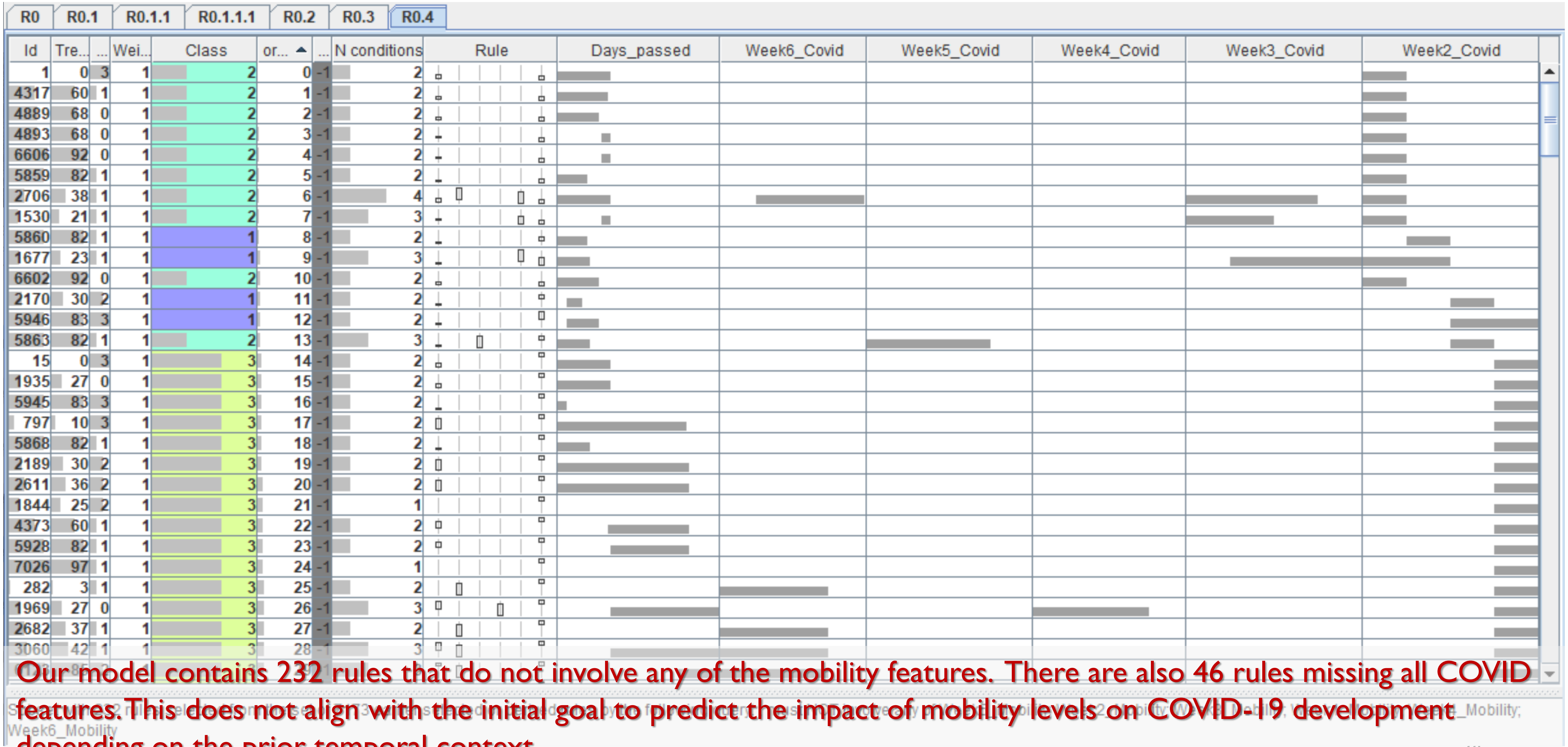
Id	Tr...	...	...	Class	...	...	N ...	Rule	N right ...	N w...	Coher...	N unite...	Depth ...	Days_p...	Week6...	Week5...	Week4...	Week3...	Week2...	Week6...	Week5...	Week4...	Week3...	Week2...	Week1...
3213	45	0	1	4	...	-1	2		2	0	100.00...	0	1												
4497	62	3	1	4	...	-1	2		1	0	100.00...	0	1												
15	0	3	1	3	11	-1	2		40	0	100.00...	0	1												
282	3	1	1	3	...	-1	2		78	0	100.00...	0	1												
504	6	2	1	3	...	-1	2		9	0	100.00...	0	1												
579	7	2	1	3	...	-1	2		3	0	100.00...	0	1												
797	10	3	1	3	...	-1	2		113	0	100.00...	0	1												
1251	16	0	1	3	...	-1	2		65	0	100.00...	0	1												
1396	19	0	1	3	...	-1	2		6	0	100.00...	0	1												
1506	20	3	1	3	...	-1	2		47	0	100.00...	0	1												
1593	22	3	1	3	...	-1	2		21	0	100.00...	0	1												
1901	26	2	1	3	...	-1	2		29	0	100.00...	0	1												
4378	60	1	1	3	...	-1	2		42	0	100.00...	0	1												
4581	63	0	1	3	...	-1	2		1	0	100.00...	0	1												
5545	77	3	1	3	...	-1	2		8	0	100.00...	0	1												
5868	82	1	1	3	...	-1	2		14	0	100.00...	0	1												
5945	83	3	1	3	...	-1	2		7	0	100.00...	0	1												
6217	87	0	1	3	...	-1	2		5	0	100.00...	0	1												
7111	99	1	1	3	...	-1	2		8	0	100.00...	0	1												
228	3	1	1	2	...	-1	2		3	0	100.00...	0	1												
1118	15	1	1	2	...	-1	2		5	0	100.00...	0	1												
4491	62	3	1	2	...	-1	2		16	0	100.00...	0	1												
4496	62	3	1	2	...	-1	2		27	0	100.00...	0	1												
4498	62	3	1	2	...	-1	2		31	0	100.00...	0	1												
4889	68	0	1	2	...	-1	2		80	0	100.00...	0	1												
4893	68	0	1	2	...	-1	2		82	0	100.00...	0	1												
5041	70	0	1	2	...	-1	2		5	0	100.00...	0	1												
5859	82	1	1	2	...	-1	2		1	0	100.00...	0	1												
6963	97	1	1	2	...	-1	2		8	0	100.00...	0	1												
229	3	1	1	1	...	-1	2		4	0	100.00...	0	1												

6761 from the 7173 earlier selected or derived rules remaining after removal of 412 contradictory rules.

The rules predicting decrease of pandemic level despite high mobility remain after removing the contradictions



RULES MISSING ALL MOBILITY FEATURES



Our model contains 232 rules that do not involve any of the mobility features. There are also 46 rules missing all COVID features. This does not align with the initial goal to predict the impact of mobility levels on COVID-19 development depending on the prior temporal context.

APPLICATION OF RULES TO DATA

Full set of 7173 original rules

17 misclassified instances (3.63%)

A (original data) | version A.1 | version A.2 | version A.3

Rec...	Ori...	Pre...	...	Wei...	Wei...	Wei...	Wei...	Day...	We...	We...	We...	We...	We...	We...	We...	We...	We...	We...	We...
06_...	2	3	no	5	9	82	6	399	2	1	1	1	1	2	2	2	3	3	2
06_...	3	1	no	48	12	35	5	392	2	2	1	1	1	1	2	2	2	3	3
08_...	3	2	no	0	58	43	0	357	3	3	3	3	3	1	1	1	1	1	1
10_...	3	2	no	0	59	6	36	434	1	2	2	2	2	2	3	2	3	3	2
13_...	2	3	no	0	10	90	0	364	3	3	3	3	3	1	2	2	2	2	2
16_...	2	3	no	0	13	87	0	364	3	3	3	3	3	1	1	1	1	2	2
17_...	2	1	no	54	46	0	0	63	1	2	2	2	2	3	0	0	0	0	0
25_...	4	2	no	4	71	7	29	392	3	2	2	2	2	2	2	2	2	2	2
27_...	1	2	no	0	100	0	0	49	0	0	1	2	2	3	3	3	0	0	0
28_...	2	1	no	59	41	0	0	63	3	3	3	2	2	3	0	0	0	0	0
30_...	3	2	no	0	58	43	0	357	3	3	3	3	3	1	1	1	1	1	1
39_...	3	2	no	0	65	35	0	175	0	0	0	0	0	3	3	3	3	3	3
42_...	2	1	no	53	47	0	0	91	2	2	1	1	0	0	0	0	0	0	0
44_...	3	2	no	0	59	41	0	147	0	0	0	0	0	2	2	2	3	3	3
46_...	3	2	no	0	58	43	0	357	3	3	3	3	3	1	1	1	1	1	1
49_...	4	2	no	0	56	8	36	210	1	1	1	2	2	3	3	3	3	3	3
50_...	3	2	no	0	79	22	0	147	0	0	0	0	0	2	2	2	2	2	2
01_...	1	1	yes	59	41	0	0	63	3	3	3	2	2	3	0	0	0	0	0
01_...	2	2	yes	0	98	2	0	56	2	3	3	3	2	3	3	0	0	0	0
01_...	2	2	yes	0	79	22	0	147	0	0	0	0	0	2	2	2	2	2	2
01_...	3	3	yes	0	0	100	0	42	0	1	2	3	3	3	3	3	3	0	0
01_...	3	3	yes	3	10	82	16	161	0	0	0	0	1	2	2	2	2	2	2
01_...	3	3	yes	2	7	83	9	196	1	2	2	3	3	2	2	1	0	0	1
01_...	3	3	yes	0	0	100	0	280	3	3	3	3	3	2	2	2	2	2	2
01_...	3	3	yes	0	0	100	0	357	2	2	3	3	3	1	1	2	2	2	2
01_...	3	3	yes	0	0	101	0	434	2	2	3	3	3	2	2	2	2	2	2
01_...	4	4	yes	1	4	14	81	175	0	0	1	1	2	2	2	2	2	2	1
01_...	4	4	yes	0	0	0	100	238	2	2	2	2	2	2	2	2	2	2	2
01_...	4	4	yes	0	0	0	100	329	2	2	2	2	2	2	1	2	1	1	1
01_...	4	4	yes	0	16	6	79	406	2	2	2	2	2	2	2	2	2	2	2
02_...	1	1	yes	100	1	0	0	70	2	2	2	2	1	0	0	0	0	0	0
02_...	1	1	yes	63	15	22	1	392	2	2	1	1	1	2	2	2	2	2	3
02_...	2	2	yes	2	97	0	1	56	1	2	2	2	2	3	3	0	0	0	0
02_...	2	2	yes	0	100	1	0	161	0	0	0	0	0	2	2	2	2	2	2
02_...	2	2	yes	3	92	5	0	364	3	3	3	3	2	1	1	1	2	2	2
02_...	2	2	yes	0	71	29	0	406	1	1	1	1	0	2	2	2	3	3	2

Data records with predictions obtained by applying 7173 rules described as Original distinct rules or explanations (7173), N conditions (43128), Total uses (7173)
Previous version of the data: Set with 468 original data records loaded from file C:\CommonGISprojects\Lamarr\model_rules\rules-Covid\final_data_num_classes.csv

Counts | Percentages

	1	2	3	4	Σ
1	53	1	0	0	54
2	3	121	3	0	127
3	1	7	174	0	182
4	0	2	0	103	105
Σ	57	131	177	103	468

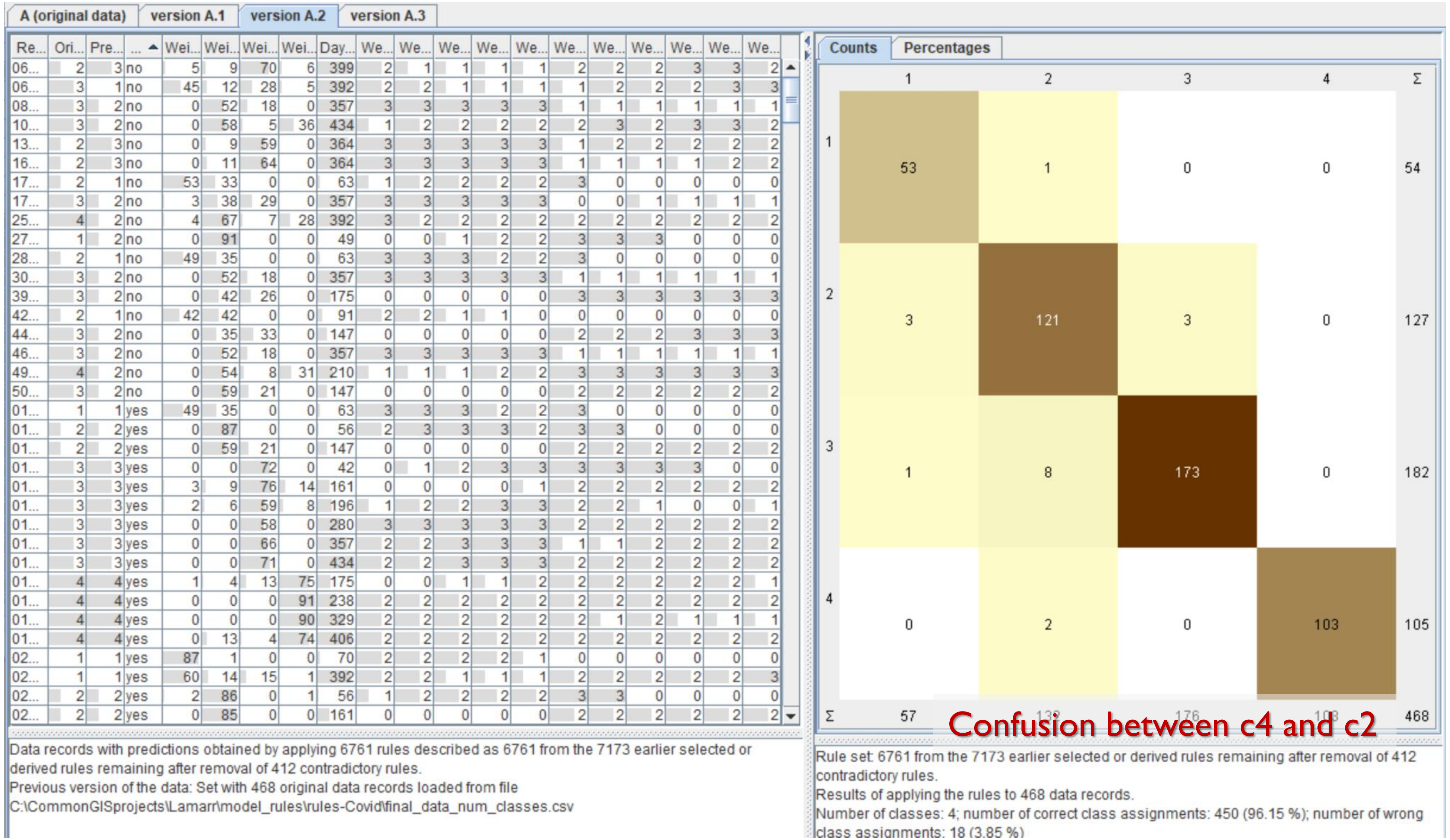
Confusion between c4 and c2

Rule set: Original distinct rules or explanations (7173), N conditions (43128), Total uses (7173)
Results of applying the rules to 468 data records.
Number of classes: 4; number of correct class assignments: 451 (96.37 %); number of wrong class assignments: 17 (3.63 %)

APPLICATION OF RULES TO DATA

Subset of 6761non-contradictory original rules

18 misclassified instances (3.85%)



Re...	Ori...	Pre...	...	Wei...	Wei...	Wei...	Wei...	Day...	We...	We...	We...	We...	We...	We...	We...	We...	We...	We...	We...
06...	2	3	no	5	9	70	6	399	2	1	1	1	1	2	2	2	3	3	2
06...	3	1	no	45	12	28	5	392	2	2	1	1	1	1	2	2	2	3	3
08...	3	2	no	0	52	18	0	357	3	3	3	3	3	1	1	1	1	1	1
10...	3	2	no	0	58	5	36	434	1	2	2	2	2	2	3	2	3	3	2
13...	2	3	no	0	9	59	0	364	3	3	3	3	3	1	2	2	2	2	2
16...	2	3	no	0	11	64	0	364	3	3	3	3	3	1	1	1	1	2	2
17...	2	1	no	53	33	0	0	63	1	2	2	2	2	3	0	0	0	0	0
17...	3	2	no	3	38	29	0	357	3	3	3	3	3	0	0	1	1	1	1
25...	4	2	no	4	67	7	28	392	3	2	2	2	2	2	2	2	2	2	2
27...	1	2	no	0	91	0	0	49	0	0	1	2	2	3	3	3	0	0	0
28...	2	1	no	49	35	0	0	63	3	3	3	2	2	3	0	0	0	0	0
30...	3	2	no	0	52	18	0	357	3	3	3	3	3	1	1	1	1	1	1
39...	3	2	no	0	42	26	0	175	0	0	0	0	0	3	3	3	3	3	3
42...	2	1	no	42	42	0	0	91	2	2	1	1	0	0	0	0	0	0	0
44...	3	2	no	0	35	33	0	147	0	0	0	0	0	2	2	2	3	3	3
46...	3	2	no	0	52	18	0	357	3	3	3	3	3	1	1	1	1	1	1
49...	4	2	no	0	54	8	31	210	1	1	1	2	2	3	3	3	3	3	3
50...	3	2	no	0	59	21	0	147	0	0	0	0	0	2	2	2	2	2	2
01...	1	1	yes	49	35	0	0	63	3	3	3	2	2	3	0	0	0	0	0
01...	2	2	yes	0	87	0	0	56	2	3	3	3	2	3	3	0	0	0	0
01...	2	2	yes	0	59	21	0	147	0	0	0	0	0	2	2	2	2	2	2
01...	3	3	yes	0	0	72	0	42	0	1	2	3	3	3	3	3	3	0	0
01...	3	3	yes	3	9	76	14	161	0	0	0	0	1	2	2	2	2	2	2
01...	3	3	yes	2	6	59	8	196	1	2	2	3	3	2	2	1	0	0	1
01...	3	3	yes	0	0	58	0	280	3	3	3	3	3	2	2	2	2	2	2
01...	3	3	yes	0	0	66	0	357	2	2	3	3	3	1	1	2	2	2	2
01...	3	3	yes	0	0	71	0	434	2	2	3	3	3	2	2	2	2	2	2
01...	4	4	yes	1	4	13	75	175	0	0	1	1	2	2	2	2	2	2	1
01...	4	4	yes	0	0	0	91	238	2	2	2	2	2	2	2	2	2	2	2
01...	4	4	yes	0	0	0	90	329	2	2	2	2	2	2	1	2	1	1	1
01...	4	4	yes	0	13	4	74	406	2	2	2	2	2	2	2	2	2	2	2
02...	1	1	yes	87	1	0	0	70	2	2	2	2	1	0	0	0	0	0	0
02...	1	1	yes	60	14	15	1	392	2	2	1	1	1	2	2	2	2	2	3
02...	2	2	yes	2	86	0	1	56	1	2	2	2	2	3	3	0	0	0	0
02...	2	2	yes	0	85	0	0	161	0	0	0	0	0	2	2	2	2	2	2

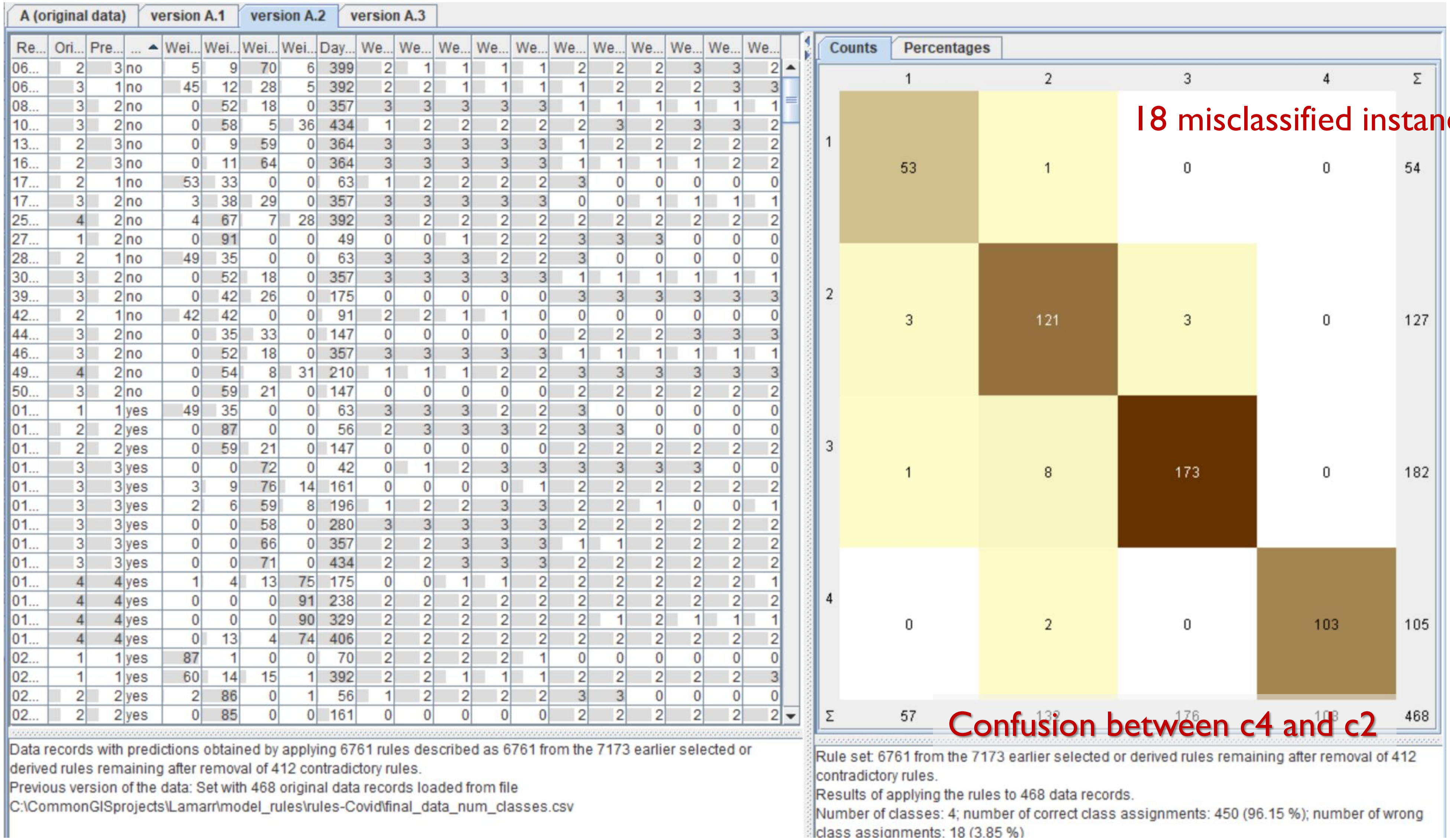
	1	2	3	4	Σ
1	53	1	0	0	54
2	3	121	3	0	127
3	1	8	173	0	182
4	0	2	0	103	105
Σ	57	[illegible]	[illegible]	[illegible]	468

Data records with predictions obtained by applying 6761 rules described as 6761 from the 7173 earlier selected or derived rules remaining after removal of 412 contradictory rules.
Previous version of the data: Set with 468 original data records loaded from file C:\CommonGISprojects\Lamarr\model_rules\rules-Covid\final_data_num_classes.csv

Rule set: 6761 from the 7173 earlier selected or derived rules remaining after removal of 412 contradictory rules.
Results of applying the rules to 468 data records.
Number of classes: 4; number of correct class assignments: 450 (96.15 %); number of wrong class assignments: 18 (3.85 %)

APPLICATION OF RULES TO DATA

6600 non-contradictory rules after removing the rules missing mobility features



18 misclassified instances (3.85%)

Confusion between c4 and c2

WHERE THE CONFUSION HAPPENS

Here the pandemic level remains stable at c2 while the level of mobility is reduced (m2). The prediction of class c2 rather than c4 appears reasonable.

Record ID: 25_c4_7_15/03/2021_02/05/2021
Row: 242

Column	Value
Original Class/Value	4.0
Predicted Class/Value	2
Match?	no
Weight 1	4
Weight 2	62
Weight 3	7
Weight 4	28
Days_passed	392
Week6_Covid	3
Week5_Covid	2
Week4_Covid	2
Week3_Covid	2
Week2_Covid	2
Week6_Mobility	2
Week5_Mobility	2
Week4_Mobility	2
Week3_Mobility	2
Week2_Mobility	2
Week1_Mobility	2

Record ID: 49_c4_10_14/09/2020_22/11/2020
Row: 457

Column	Value
Original Class/Value	4.0
Predicted Class/Value	2
Match?	no
Weight 1	0
Weight 2	54
Weight 3	8
Weight 4	29
Days_passed	210
Week6_Covid	1
Week5_Covid	1
Week4_Covid	1
Week3_Covid	2
Week2_Covid	2
Week6_Mobility	3
Week5_Mobility	3
Week4_Mobility	3
Week3_Mobility	3
Week2_Mobility	3
Week1_Mobility	3

Here the pandemic level slightly increased from c1 to c2 while the mobility level remains relatively high (m3). Here an increase of the pandemic level would be expected.

Evidently, the 21 rules predicting class c2 when mobility in week -2 is high are not responsible for these confusions. Hence, there were training data instances corresponding to these rules.

WHAT WE HAVE LEARNT ABOUT THE MODEL

- Unwanted behaviour: making predictions based on insufficient (too few) conditions
- Unwanted behaviour: making predictions while ignoring mobility features or prior pandemic levels
- Unwanted property: contradictions among the rules – multiple rules predicting different classes can be applied to the same instances
  - Removal of the contradictions, rules ignoring mobility or pandemic features, and remaining rules with less than 3 conditions only slightly (by 0.22%) decreases the model accuracy for an available test dataset.
    This could be acceptable for the sake of improving model logic.
- Unwanted property: some rules are not justifiable by domain logic or common sense.
  - However, they seem to be in accord with the training and test data.
    Hence, there are real cases contradicting the logical expectations.

GENERAL INSIGHTS: DIFFERENCES BETWEEN ML AND HUMAN REASONING

- A model may reach correct conclusions but may not "think" like a human.
- Some rules may lack domain-relevant conditions yet still function well.
- Adjusting or filtering rules based on human logic might not significantly affect accuracy—suggesting redundancy or alternative reasoning paths.

CONCLUSIONS

- Trustworthiness is not just about accuracy—it's about understanding ***why*** the model makes decisions.
- How should we balance human logic vs. data-driven inferences when interpreting and explaining models?
- **Open question**: Should ML models be adjusted to align better with human reasoning, even if accuracy does not improve or may even slightly degrade?
  - If so, how can we incorporate domain knowledge and human logic at the stage of model training?