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

Natalia Andrienko, Gennady Andrienko, Bahavathy Kathirgamanathan

Human-Centred Al

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 (cl to c4).
- The categorical values were encoded by numbers from 1 to 4.

REPRESENTATION OF RULES IN A TABLE

Predicted class or

features

🕌 Rule	alue (for regres	ssion)										Å						\times
R0		Ļ																	٦
Id T	ree Tr W	Class or	. N conditions	Rule		Days pass	Week6 Co	Week5 Co	Week4 C	o We	eek3 Co	Week2 C.	Week6 Mob.	Week5 M.	. Week4 Mobi.	Week3 Mo	Week2 Mobility	Week1 Mot	oil
1	0 3	1 2 0 -	1 2			45171					_	0	0						
2	0 3	1 1 1-	1 3		<u> </u>	45171						1	1	00)				
3	0 3	1 3 2 -	1 4		1111	45171			0	0		1	1	1.1	1				
4	0 3	1 4 3 -	1 5			45171			0	0		1	1	23	ì		02		
5	0 3	1 3 4 -	1 6			45171			0	0		1	1	23	ì	02	33		
6	0 3	1 1 5-1	1 6			45154			0	0		1	1	23	i	33	i 33		
7	0 3	1 3 6 -	1 6 -			154171			0	0		1	1	23	i	33	33		
8	0 3	1 1 7-	1 5	0 4 0		45171			1			1	1	13	ž			(J1
9	0 3	1 1 8 -	1 6	0 + 6	🛛	45171			1	3		1	1 0	1 13	i				
10	0 3	1 3 9 -	1 6	0 0	🛛	45171			1	3		1	1 2	1	i				
11	0 3	1 10 -	1 4	† 🗅		45171						2	2 0	2			01		
12	0 3	1 4 11 -	1 4		🛛	45171						2	20	2			13		
13	0 3	1 2 12 -	1 4 4] 🕈 🗂		45171	02					2	2 3	2					
14	0 3	1 1 13 -	1 4 4	1 † 7		45171	33					2	2 3	2					
15	0 3	1 3 14 -	1 2			45171						3	2						
16	0 3	1 3 15 -	1 4 🛛	0 0		172431			0	0	02			02	2				
17	0 3	1 2 16 -	1 7 +			172178	00		0	0	00	0	0	3	à		02		
18	0 3	1 2 17 -	1 7 +			172178	00		0	0	00	0	0	3	ñ		33		
19	0 3	1 2 18 -	1 6 🖓			179431	00		0	0	00	0	0	300	ñ				
20	0 3	1 3 19 -	1 6 🖓	, 0 7		172431	00		0	0	00	1	2	3	ñ				
21	0 3	1 3 20 -	1 5 🖓	, _0 7		172431	00		0	0	12	<u> </u>	nditio	nc 3mi	ň				
22	0 3	1 3 21 -	1 5 🖓] _0 7		172431	13		0	0	02			15 Juli	ñ				
23	0 3	1 4 22 -	1 5 🖓	, + <u>(</u>)		172431	00		Barc	"ho	Drocc	nt in	torvale	of for	turo)2
24	0 3	1 4 23 -	1 6 🕂 🗍] 4 []	0 0	172.304	13		Dais	"I E	prese		Lei vais	UT lea	iture va	nues)2
25	0 3	1 3 24 -	1 7 9] 4 []		305.431	13		1	1	12				0	02	2)2
26	0 3	1 2 25 -	1 7 9] 4 []	ċŸ∣ċ	305.431	13		1	1	0.2	exts o	can de	niaaer	0	2 300	M)2
27	0 3	1 3 26 -	1 6 🛛] 4 []	۵ ۵	172431	13		1	1	02				2m	2)2
28	0 3	1 4 27 -	1 6 中位] 🕈 🖸		1723 60	02		2	2	02						01)2
29	0 3	1 3 28 -	1 6 1	1 🕈 🗋		361434	02		2	2	02						01)2
30	0 3	1 4 29 -	1 6 + (9011	I P D	472259	02		2	2	02						22		12
31	0 3	1 3 30 -	1 8 9 0	9000	6640	259.431	02		2	2	02				0	01	22)2
32	0 3	1 4 31 -	1 8 9 0	P []	690	259.431	02		2	2	02				0	1 23	22		J2 🔻
Original	distinct rule	s or explanations (7	173), N conditions	raphi	cal	ses (7173)								*****					

representations of rules

GRAPHICAL REPRESENTATION OF A RULE

🛃 Rules explorer v.23.01.2025 15:10

RO

ld Ti	ree Tr W	Class or N conditions	Rule	Days_pass	Week6_Co	Week5_Co	Week4_Co	. Week3_Co	Week2_C	Week6_Mo	b Week5_M	Week4_Mobi	. Week3_Mo	Week2_Mobility	Week1_Mobil.	
1	0 3 1	2 0 -1 2		45171					00							
2	0 3 1	1 1 -1 3	┢ ┍ ┢	45171					11		00					
3	0 3 1	3 2 -1 4	┢ ┢ ┢ ቀ	45171			00)	11		11					
4	0 3 1	4 3 -1 5	ь <u>ь</u> ф Ф <u>ф</u>	45171			00)	11		23			02		
5	0 3 1	3 4 -1 6	6 6 6 7 6 7	45171			00)	11		23		02	33		
6	0 3 1	1 5 -1 6		45154			00)	11		23		33	33		
7	0 3 1	3 6 -1 6	+	∎154171			00)	11		23		3.3	33		
8	0 3 1	1 7 -1 5	<u> </u>	45171			13		11		13				01	1
9	0 3 1	1 8 -1 6	<u>+ Q + 6 Q Q</u>	45171				-	Rule 8	Weight - 1					1.1	â.
10	0 3 1	3 9 -1 6	<u> </u>	45171			1.		IXUIC 0	Weight - 1					1.7	â.
11	0 3 1	1 10 -1 4	<u> </u>	45171					Clas	s 1				01		1
12	0 3 1	4 11 -1 4	<u> </u>	45171										13		1
13	0 3 1	2 12 -1 4		45171	02		ſ					1				1
14	0 3 1	1 13 -1 4		45171	33											
15	0 3 1	3 14 -1 2		45171												1
16	0 3 1	3 15 -1 4	P	172431			0.		11 1		- h					1
17	0 3 1	2 16 -1 7		172178	00		0.							02		1
18	0 3 1	2 17 -1 7	+	172178	00		0.		ТІТ	T				33		1
19	0 3 1	2 18 -1 6		179431	00		0.	<u> </u>			<u> </u>					1
20	0 3 1	3 19 -1 6	₽₽₽₽₽	172431	00		0.	Feature	min	from to	max					1
21	0 3 1	3 20 -1 5	Ρ	172431	00		0.	- outaro								1
22	0 3 1	3 21 -1 5	PQ	172431	13		0.	Days_passe	d 45.0000	- inf 171	431.0000					1
23	0 3 1	4 22 -1 5	₽₩ ₩0 0	172431	00		1.								07	2
24	0 3 1	4 23 -1 6	+0 +0 0 0	172.304	13		1.	Week4_Cov	id 0.0000	1 + IN	3.0000	02	-		07	2
25	0 3 1	3 24 -1 7	90140110010	305.431	13		1.	Week2 Cov	0 0000 bi	1 4	3 0000	02	02		07	2
26	0 3 1	2 25 -1 7	٩٩ +٥ ۵	305.431	13		1.	110012_001	0.0000		0.0000	02	33		07	2
27	0 3 1	3 26 -1 6	PQ +0 "	172431	13		1.	Week5_Mobil	ity 0.0000	1 + in	f 3.0000	33	Ŕ		07	2
28	0 3 1	4 27 -1 6	Ф <u>0</u> Ф <u>0</u> <u></u> d	172360	02		2.							01	07	2
29	0 3 1	3 28 -1 6	TO +O	361434	02		2.	Week1_Mobil	ity 0.0000	- inf	3.0000			01	07	2
30	0 3 1	4 29 -1 6	+0 +0 +	172259	02		2.2	02						22	07	2
31	0 3 1	3 30 -1 8	90190116690	259.431	02		22	02				01	01	22	07	2
32	0 3 1	4 31 -1 8	P0 +0 + P+0	259.431	02		22	02				01	23	22	07	2 -

Original distinct rules or explanations (7173), N conditions (43128), Total uses (7173)

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) Filtering rules by feature values intervals \times Days passed Class-wise Feature-wise Select rules not involving this feature Class 3 (2351 rules) Class 1 (879 rules) Class 2 (2682 rules) Class 4 (1261 rules) 45.000 431.000 Days_passed Days_passed Days_passed Days_passed Week6 Covid Select rules not involving this feature Week6 Covid Week6 Covid Week6_Co Color-coded Week6_Covid 0.000 3.000 Week5_Cocounts of rulesek5_Covid Week5 Covid Week5 Covid Week5 Covid Select rules not involving this feature Week4 cowith conditionsk4 Covid Week4 Covid Week4_Covid 3.000 0.000 **Blue bars** Weeks councluding the Weeks Covid Week4 Covid Week3 Covid Controls for Select rules not involving this feature intervals counts of 3.000 0.000 Week2 Co Week2_Covid Week2 Covid interactive rules Week3_Covid 101 10 10 11 filtering Week6_Mobility Select rules not involving this feature involving the Class 2 + Week2 Covid: 0.000 3.000 Week5 Mobility features Interval of feature values: [1.500..2.500) Week2 Covid 1206 out of 2246 (max = 2246) Select rules not involving this feature Count of rules: Week4 Mobility Grey bars: 3.000 0.000 total count of Week3 Mobility Week3_Mobility Week3 Mobility Week6 Mobility Select rules not involving this feature rules for this Week2_Mobility Week2 Mobility Week2_Mobility Feature ranges in rules MUST class Mobility be contained in filter limits Week1_Mobility Week1 Mobility Week1 Mobility overlap with filter limits feature values feature values feature values feature values Apply filters dynamically Apply Clear Original distinct rules or explanations (7173), N conditions (43128), Total uses (7173)

8

Extract rules

DISTRIBUTION OF FEATURE VALUE INTERVALS Feature-wise view



9

left Filtering rules by feature values intervals



left Filtering rules by feature values intervals



left Filtering rules by feature values intervals



left Filtering rules by feature values intervals

- 🗆 X



RELATIONSHIPS BETWEEN FEATURES

Filtering rules by feature values intervals

– 🗆 ×



COMBINED EFFECTS OF 2 OR MORE FEATURES

left Filtering rules by feature values intervals



SURPRISING FINDINGS

Siltering rules by feature values intervals



- 🗆 🗙

Extract rules

EXTRACTED RULES SATISFYING THE FILTER

🕌 Rules explorer v.30.01.2025 17:20

- 🗆 X

0 8	01																							
		1										1	1		1									
ld	Tree Id	. Weight	Class	or 🔺	• C	N con		Ru	le		Days_pa	Week6	. Week5	Week4	Week3	Week2	Week6	Week5	Week4	Week3	Week2_M.	. Week	c1	
561	7	2 1	1 2	(0 -1	7	0 6	🕈 🕈	┦ 🏻 🖛	┦│													4	•
562	7	2 1	1 4	1	1 -1	7	0 9	99	┦ 🏻 🖛	┦│														
6590	91	1 1	1 2		2 -1	8	P d	🕈 🖣	┦ ┦ │ │	Π Π												<u> </u>		
780	10	3 1	1 4		3 -1	9		0 9 9	0 9 0 9	1														
788	10	3 1	1 2	4	4 -1	8			7 7 Q	7														
789	10	3 1	1 4		5 -1	8			7 7 Q	7														
1022	13	3 1	1 4	(6 -1	8		0 9 9	🛛 🖓	7 0												<u> </u>		
1098	14	0 1	1 4	1	7 -1	6	φQΙ	7 9 9																
1503	20	3 1	1 4		8 -1	10	ΟQΙ	0 9 9	¶ Q 🛉	7 0												<u> </u>		
1833	25	2 1	1 4	9	9 -1	6	<u> </u>	99		7 9												<u> </u>		
5632	78	1 1	1 4	10	0 -1	5	TII			7 9												<u> </u>		
4489	62	3 1	1 2	11	1 -1	6	<u> </u>																	
5634	78	1 1	1 2	12	2 -1	5	TII																_	_
5633	78	1 1	1 4	13	3 -1	5	TII				-													
2323	32	1 1	1 2	14	4 -1	10	0 d		9010	99												<u> </u>		
2450	34	1 1	1 2	1	5 -1	7	96	0 9 9		"												4		
2451	34	1 1	1 4	10	6 -1	7	<u> р</u> Р	0 9 9		1														
4487	62	3 1	1 4	17	7 -1	6	<u> </u>			<u> </u>												<u> </u>		
2752	38	1 1	1 4	18	8 -1	6	<u> </u>	<u> + P</u>		<u> </u>														
2753	38	1 1	1 2	19	9 -1	6	<u> </u>	<u> + P</u>		<u> </u>												4		
5560	77 :	3 1	1 2	20	0 -1	8	<u> </u>	000		<u> </u>												4		
3736	51	2 1	1 4	2'	1 -1	9	P P P	99	<u> </u>	<u> </u>														
3933	54	3 1	1 2	27	2 -1	9	9 6	<u>0 + P</u>		1														
3935	54	3 1	1 4	23	3 -1	9	₽ ₽	<u>0 + P</u>	1 0 1															
4441	61	2 1	1 4	24	4 -1	5		+ -		<u><u> </u></u>												-		
4445	61	2 1	1 4	25	5 -1	9	P P	 																
4446	61	2 1	1 2	26	6 -1	9	$P \mid P$		<u>" "</u>															
4710	65	2 1	1 2	27	7 -1	7	+	0 9 9		1	1											4	_	
4864	67	2 1	1 2	28	8 -1	6	₽ 	99		1													_	
4935	68	D 1	1 4	29	9 -1	10	P 🕈	999	9 9 6	ĨÓ														
4939	68	0 1	1 2	30	0 -1	8	ΡċΨ	99		77														
5135	71	0 1	1 4	31	1 -1	B	ne 🛛	ext	rac	гес	1 rules	are o	ordere	VO DS	simila	rity (r	IN ADD	iving (S algo	orithm	11.		

5139 71 0 1 4 32 1 5621 78 1 1 2 33 1 We see that differences between rules predicting different classes may be very subtle.

45 rules satisfying filter conditions 1) Week2_Mobility france aborious process of comparison can be facilitated by aggregating the rules rules

ITERATIVE AGGREGATION AND GENERALISATION OF RULES



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
18	3325 2	1	4 1	0 -1	6	T D P P	🕆 🗗	1	(0 100.00	0)	1											
67	9094 0	1	4 1	1 -1	7	0 79+0	🕆 🗗	1	(0 100.00	0)	1											
	-1 -1 -1	3	4 5	2 -1	5	P P P	🗂 🖞	5	(0100.00	5	5	3											
	-1 -1 -1	3	2 3	3 -1	5	P P P b	🕇	3		75.000	4	Ļ	3											
	-1 -1 -1	4	4 4	4 -1	4	7 0 0	🕇	4	(0 100.00	6	5	3						D:		hat		2	21
	-1 -1 1	2	2 3	5 -1	4	<u> </u>		3	(0 100.00	3	3	2						Jiner	ence	s bet	ween	Z SIN	niiar
	-1 -1 -1	2	2 2	6 -1	6	₽ġ ġ₽₽		3	(0 100.00	2	2	2						aulos	prodi	cting	dicti	act d	20226
	-1 -1 -1	2	4 3	The	co rui	los hav		CODT	one	0100.00	3	3	2						uies	predi	cung	uistii		asses
	-1 -1 -1	2	2 2	1 816	serų			серц		0100.00	2	2	2						rannc	t be	grour	nded	in do	main
	-1 -1 -1	5	2 5	9 -1	4			7	(0 100.00	8	3	4								o. o u.			
	-1 -1 -1	2	4 3	10 -1	5			3	(0100.00	3	3	2						know	ledge				
	-1 -1 3	2	4 2	11 -1	7	0 0000	¢ 🖱	2	(0 100.00	2	2	2					- 10 C		<u> </u>		_	_	
	-1 -1 -1	2	4 2	12 -1	6			2		0100.00	2	2	2											1
	-1 -1 -1	3	4 3	13 -1	4			6		275.000	4	L	3											1
	-1 -1 -1	2	2 2	14 -1	8	0 0 9 9 9 9	🛛 🖱	2	(0 100.00	2	2	2											1
	-1 -1 -1	2	2 2	15 -1	5		<u> ¶ </u>	3	(0100.00	2	2	2											
	-1 -1 0	2	2 2	16 -1	7	06 940		2	(0100.00	2	2	2											

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:

ld	. Tr	Weight	Class	N uses	or	<u>م</u>	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	3	2	3		0 -1	1 5	P PP <u> </u> ¶	3	1	75.00	4	3												
26	2	1	2	1		1 -1	1 9	$P P + P^{P} P^{P}$	1	0	100.0	0	0												
36	3 0	1	2	1		2 -1	1 8	₽ <u></u>	1	0	100.0	0	0												
48	2 1	1	2	1		3 -1	1 10	T90+9+100071	1	0	100.0	0	0												
56	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
183325 2	1	4 1	0 -1	6	[]] 1	0	100.00	0	1												
6790 94 0	1	4 1	1 -1	7 🗘	4 4 4 4] 1	0	100.00	0	1												
-1 -1 -1	3	4 5	2 -1	5 P] 5	0	100.00	5	3												
-1 -1 -1	3	2 3	3 -1	5 P	🛛 🖓 🖓 🖞 🖑	3		75.000	4	3												
-1 -1 -1	4	4 4	4 -1	4		4	0	100.00	6	3								onco	hot		2 cin	ailan
-1 -1 1	2	2 3	5 -1	4 T		3	0	100.00	3	2							Diller	ence	s bet	ween	z sin	niiar
-1 -1 -1	2	22	6 -1	6 🛛		3	0	100.00	2	2							nulas	predi	cting	distir	het d	20226
-1 -1 -1	2	4 3	The	ရေ က်ခြ	s have e	vcenti	ons ⁰	100.00	3	2							uics	pica	cung	uistii		asses
-1 -1 -1	2	2 2	84		S mare e	L CP4		100.00	2	2							canno	t be	grour	nded	in do	main
-1 -1 -1	5	2 5	9 -1	4		7	0	100.00	8	4									0			
-1 -1 -1	2	4 3	10 -1	5		3	0	100.00	3	2							know	ledge				
-1 -1 3	2	4 2	11 -1	7 🛛	<u> </u>	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	<u> + P T T</u>	3	0	100.00	2	2												
-1 -1 0	2	2 2	16 -1	70	<u> </u>	2	0	100.00	2	2												
J																						

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

RO	R0.1	R0.1.	1 R().1.1.1	R0.2	R0.	.3																		
Id	Tree)	Class	🔺	N con		Rule	N	I right N	wro Coher	. N unit	Depth	Days	Week	Week5	Week	Week3	Week	Week6	Week5.	Week4	Week3	Week2	Week1.	
1	0	3 1	2	0 -1	2	4	_		393	86 82.04	. (D	1			~									•
4317	60	1 1	2	1 -1	2	4	_		168	6572.103	. ()	1			12									
5382	75	0 1	2	2 -1	3	\downarrow	_	01	9	8 52.941	. ()	1					1	Rule	1 V	Veight = 1			1	7
6604	92	0 1	2	3 -1	4	+11		0	7	187.500	. ()	1						rvure	•	reigin – i				H
5383	75	0 1	2	4 -1	3	Π		01	37	588.095	. ()	1							CI	ass 2				
5869	82	1 1	3	5 -1	4	+	_	0d	4	180.000.	. (0	1							N us	ses: 1				
2156	30	2 1	2	6 -1	3	+11		🛛	9	950.000	. ()	1												
4838	67	2 1	2	7 -1	4	+	_	Q (j	7	653.846	. ()	1						N dis	stinct obje	ects: 1				
1851	26	2 1	2	8 -1	4	+	_0		3	442.857	. ()	1												
4385	61	2 1	2	9 -1	3	6			61	1184.722	. ()	1			0.1			0.00 N			202 N.			
5513	77	3 1	2	10 -1	3			"	31	1272.093	. ()	1			Cone	erence (ru	e-based	0.82 N	correctly	covered	393 N N	rongly co	vered 86	1
6546	91	1 1	2	11 -1	3	6		"	64	1482.051.	. ()	1												
1468	20	3 1	2	12 -1	4	6 6	1411	"	28	584.848	. ()	1					T T							
2428	34	1 1	2	13 -1	4			1	1	1 50.000	. ()	1												
1256	17	1 1	2	14 -1	4	+		7	3	350.000	. ()	1												
1469	20	3 1	2	15 -1	4	+ 0	<u> </u>	7	3	350.000	. (0	1												
6470	90	3 1	2	1 6 -1	6	4	0.0	۳ Q	2	166.666	. (0	1					n l							
1532	21	1 1	2	17 -1	4	+		7	2	340.000	. ()	1												
5453	76	1 1	2	1 8 -1	4			111	12	192.307	. ()	1					ų		<u> </u>					
1797	25	2 1	2	1 9 -1	4	+		┦	2	2 50.000	. ()	1												
4892	68	0 1	2	20 -1	4	411			7	7 50.000	. ()	1			Feat	ure	min	from	to ma	ax				
3912	54	3 1	2	21 -1	5	411	0_0	7	0	1 0.00000) (0	1												
6686	93	0 1	2	22 -1	5	411			29	878.378	. ()	1			Days	s_passed	45.000	0 - inf	171 43	1.0000				
3686	51	2 1	2	23 -1	5	4		10P	8	188.888	. ()	1			10/00	ko Onvid	0.000		0	2 0000				
5953	83	3 1	2	24 -1	6	+ 6	140		2	166.666	. ()	1			wee	ekz_Covid	0.000	0 - INI	0	3.0000				
1852	26	2 1	3	25 -1	4	<u>+ </u>	_0		3	350.000	. ()	1			_									T
1583	22	3 1	1	26 -1	2				23	874.193	. ()	1												
4205	58	0 1	2	27 -1	4		0	0	9	660.000	. ()	1												
6605	92	0 1	3	28 -1	3	+		🖷	9	7 56.250	. ()	1												
5048	70	0 1	3	29 -1	5	+		Π	1	233.333.	. ()	1												
5099	71	0 1	2	30 -1	5	4		0 0	18	675.000	. ()	1												-

355 contradictory rules extracted and removed from the set of 7173 earlier selected or derived rules

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.30.01.202	5 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 D	epth Dav	s p	Week6	Week5	Week4	Week3	Week2	Week6	Week5	Week4	Week3	Week2	Week1.	
3213 45 0 1 4	-1 2	0		2 (100.00	0	1													
4497 62 3 1 4	-1 2 1	<u> </u>		1 (100.00	0	1													
15 0 3 1 311	-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 (100.00	0	1													1
579 7 2 1 3	-1 2			3 (100.00	0	1													
797 10 3 1 3	-1 2 0			113 (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 (100.00	0	1													
1506 20 3 1 3	-1 2	₫ ┦		47 (100.00	0	1													
1593 22 3 1 3	-1 2			21 (100.00	0	1													
1901 26 2 1 3	-1 2 []			29 (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	1												
5545 77 3 1 3	-1 2 †0			8 (100.00	0	1	1												
5868 82 1 1 3	-1 2 \downarrow 🗌			14 (100.00	0	1													
5945 83 3 1 3	-1 2			7 (100.00	0	1													
6217 87 0 1 <u>3</u>	-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 (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 1	 		27 (100.00	0	1													-
4498 62 3 1 2	-1 2 🚺	 		31 (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												Ļ	_
5041 70 0 1 2	-1 2			5 (100.00	0	1												<u> </u>	
5859 82 1 1 2	-1 2 \downarrow 🗌			1 (100.00	0	1												L	_
<u>6963 97 1 1 2</u>	-1 2			8 (100.00	0	1													
229 3 1 1 1	-1 2			4 (100.00	0	1												L	
					400.00															

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

Class-wise Feature-wise			Weeko_Covid
Class 2 (21 rules)	Class 3 (0 rules)	Class 4 (24 rules)	Select rules not involving this feature
Days_passed	Days_passed	Days_passed	Week5_Covid
Week6_Covid	Week6_Covid	Week6_Covid	Select rules not involving this feature 0.000 3.000
Week5_Covid	Week5_Covid	Week5_Covid	Week4_Covid
Week4_Covid	Week4_Covid	Week4_Covid	0.000 • 3.000
Week3_Covid	Week3_Covid	Week3_Covid	Week3_Covid Select rules not involving this feature
Week2_Covid	Week2_Covid	Week2_Covid	1.125 3.000 Week2_Covid
Week8_Mobility	Week6_Mobility	Week6_Mobility	Select rules not involving this feature
Week5_Mobility	Week5_Mobility	Week5_Mobility	Week6_Mobility
Week4_Mobility	Week4_Mobility	Week4_Mobility	Select rules not involving this feature 0.000 3.000
Week3_ <mark>Mobility</mark>	Week3_Mobility	Week3_Mobility	Week5_Mobility
Week2_Mobility	Week2_Mobility	Week2_Mobility	Feature ranges in rules MUST be contained in filter limits
Week1_Mobility	Week1_Mobility	Week1_Mobility	overlap with filter limits
feature values	feature values	feature values	Apply filters dynamically
45 rules selected by applying filter conditions	1) Week2_Mobility: from 2.250 to +infinity 2) Week3_Covid:	from 1.125 to +infinity 3) Week2_Covid: from 1.425 to +	infinity Extract rules

RULES MISSING ALL MOBILITY FEATURES

R0	R0.1	R0.1.1	R0.1.1.1	R0.2	R0.3 F	R0.4						
Id	Tre	Wei	Class	or 🔺	N conditio	ns Rule	Days_passed	Week6_Covid	Week5_Covid	Week4_Covid	Week3_Covid	Week2_Covid
1	0	3 1	2	0 -1	1	2						
4317	60	1 1	2	1 -1	1	2						
4889	68	0 1	2	2 -1	1	2						
4893	68	0 1	2	3 -1	1	2 - .						
6606	92	0 1	2	4 -1	1	2 +						
5859	82	1 1	2	5 -1	1	2 \downarrow 🛛 🖉 🔒						
2706	38	1 1	2	6 -1	1	4 4 0 0 4						
1530	21	1 1	2	7 -1	1	3 +						
5860	82	1 1	1	8 -1	1	2 🖵 🕂						
1677	23	1 1	1	9 -1	1	3 🛓 🛛 🖓 🗄						
6602	92	0 1	2	10 -1	1	2						
2170	30	2 1	1	11 -1	1	2	-					
5946	83	3 1	1	12 -1	1	2 +						
5863	82	1 1	2	13 -1	1	3 0 +						
15	0	3 1	3	14 -1	1	2						
1935	27	0 1	3	15 -1	1	2						
5945	83	3 1	3	16 -1	1	2						
797	10	3 1	3	17 -1	1	2 0 "						
5868	82	1 1	3	18 -1		2						
2189	30	2 1	3	19 -1		2 0						
2611	36	2 1	3	20 -1		2						
1844	25	2 1	3	21 -		1 T						
43/3	60	1 1	3	22 -								
5928	82	1 1	3	23 -		2 9 T						
/026	97	1 1	3	24 -								
282	3	1 1	3	25 -	_	2						
1909	21		3	20 -	_	3 4						
2082	37		3	21 -	_	2						
3000	42			- 20				f the meshilt				
	ur∘m	iodel C	ontain	5 252	rules t	that do not	involve any o	T the modility	reatures. II	nere are also	to rules miss	ing all COVID
sfea	ture	s: This	does	not al	ign wit	h•the initial	goal to predi	ct the impact	of mobility	levels on CO	VID-19 devel	opment_Mobility;
Week	6_Mobi	lity			0		0	et alle illipad				
de	Deno	ding or	າ the c	rior t	empor	al context.						上竹

APPLICATION OF RULES TO DATA Full set of 7173 original rules

17 misclassified instances (3.63%)



wrong class assignments: 17 (3.63 %)

C:\CommonGISprojects\Lamarr\model rules\rules-Covid\final data num classes.csv

APPLICATION OF RULES TO DATA Subset of 6761 non-contradictory original rules [8 misclassified instances (3.85%)]

A (0	riginal da	ata)	version A.1	V	ersio	n A.2	vers	sion A.3																
Re	Ori Pr	re	▲ Wei We	i We	ei W	/eiD	ay W	e We.	. We	We.	. We.	. We	We.	We	. We	We \	Ve	1	Counts	Percentag	es			
06	2	3 no	5	9	70	6	399	2	1	1	1 1	2	1	2 2	2	3 3	2 🔺			4	2	2		-
06	3	1 no	45 1	2	28	5	392	2	2	1	1 1	1		2 2	2	2 3	3	100		1	2	3	4	2
08	3	2 no	0 5	2	18	0	357	3	3	3 3	3 3	3 1	1	1 1	1	1 1	1	100						
10	3	2 no	0 5	8	5	36	434	1 :	2	2 3	2 2	2 2		3 2	2	3 3	2	100						
13	2	3 no	0	9	59	0	364	3	3	3 3	3 3	3 1	1	2 2	2	2 2	2	1000	1					
16	2	3 no	0 1	1	64	0	364	3	3	3 3	3 3	3 1		1 1	1	1 2	2	2000		53	1	0	0	54
17	2	1 no	53 3	33	0	0	63	1 :	2	2	2 2	2 3	(0 ()	0 0	0	100						
17	3	2 no	3 3	8	29	0	357	3	3	3	3 3	3 0	(0	1	1 1	1	1000						
25	4	2 no	4 6	7	7	28	392	3	2	2	2 2	2 2		2 2	2	2 2	2	100						
27	1	2 no	0 9	1	0	0	49	0 (0	1 :	2 2	2 3		3 3	3	0 0	0	100						
28	2	1 no	49 3	5	0	0	63	3	3	3	2 2	2 3	(0 0)	0 0	0	1000						
30	3	2 no	0 5	2	18	0	357	3	3	3 3	3 3	3 1		1 1	1	1 1	1	2000						
39	3	2 no	0 4	12	26	0	175	0 (0	0 (0 (3		3 3	3	3 3	3	100	2					
42	2	1 no	42 4	2	0	0	91	2	2	1	1 (0 0	(0 ()	0 0	0	1000		3	121	3	0	127
44	3	2 no	0 3	35	33	0	147	0 1	0	0	0 (2		2 2	2	3 3	3	2000						
46	3	2 no	0 5	2	18	0	357	3	3	3	3 3	3 1		1 1	1	1 1	1	100						
49	4	2 no	0 5	64	8	31	210	1	1	1	2 2	2 3		3 3	3	3 3	3	1000						
50	3	2 no	0 5	9	21	0	147	0	0	0	0 (2		2 2	2	2 2	2	100						
01	1	1 yes	49 3	15	0	0	63	3	3	3	2 2	2 3	(0 ()	0 0	0	100						
01	2	2 yes	8 0 8	37	0	0	56	2	3	3	3 2	2 3		3 ()	0 0	0	2000						
01	2	2 yes	0 5	9	21	0	147	0	0	0	0 (2		2 2	2	2 2	2	100	3					
01	3	3 yes	0	0	72	0	42	0	1	2	3 3	3 3		3 3	3	3 0	0	100		1	8	173	0	182
01	3	3 yes	3	9	76	14	161	0	0	0	0 1	2	1	2 2	2	2 2	2	1000						
01	3	3 yes	2	6	59	8	196	1 :	2	2	3 3	3 2		2 1	1	0 0	1	2000						
01	3	3 yes	0	0	58	0	280	3	3	3	3 3	3 2	1	2 2	2	2 2	2	100						
01	3	3 yes	0	0	66	0	357	2	2	3	3 3	3 1		1 2	2	2 2	2	1000						_
01	3	3 yes	0	0	71	0	434	2	2	3 3	3 3	3 2		2 2	2	2 2	2	100						
01	4	4 yes	1	4	13	75	175	0	0	1	1 2	2 2	1	2 2	2	2 2	1	100						
01	4	4 yes	0	0	0	91 3	238	2	2	2	2 2	2 2		2 2	2	2 2	2	1000	4					
01	4	4 yes	0	0	0	90	329	2	2	2	2 2	2 2		1 2	2	1 1	1	100		0	2	0	103	105
01	4	4 yes	0 1	3	4	74	406	2	2	2	2 2	2 2	1	2 2	2	2 2	2	100						
02	1	1 yes	87	1	0	0	70	2	2	2	2 1	0	(0 ()	0 0	0	1000						
02	1	1 yes	60 1	4	15	1	392	2 1	2	1	1 1	2		2 2	2	2 2	3	1000						
02	2	2 yes	2 8	6	0	1	56	1 :	2	2	2 2	2 3		3 ()	0 0	0	100						-
02	2	2 yes	8 0 8	15	0	0	161	0	0	0	0 0	2	1	2 2	2	2 2	2 🗸		Σ	57	onfusion	notwicen c	4 and c?	468
1000000									an a				ana	an a							onusion	Jermeen C		
Data	ecords v	with pre	dictions obta	ained	by a	pplying	g 6761	rules de	escrib	oed as	6761 f	rom th	e 717	3 earl	ier se	elected or		1111	Rule set: 6	761 from the	7173 earlier selected	or derived rules remai	ining after removal of	f 412
derive	d rules r	emain	ing after rem	oval	of 412	2 contr	adictor	y rules.										1000	contradicto	ry rules.				
Previo	us versi	on of th	ie data: Set v	with 4	68 or	iginal	data re	cords lo	adeo	trom f	lle							2222	Results of	applying the n	ules to 468 data recor	ds.		

C:\CommonGISprojects\Lamarr\model_rules\rules-Covid\final_data_num_classes.csv

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



Number of classes: 4; number of correct class assignments: 450 (96.15 %); number of wrong class assignments: 18 (3.85 %)

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.

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: 25_c4_7_15/03/2021_02/05/2021

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?