# Action Rule Mining approach to Analyze Oxford Dataset for Better Recommendation for Policy Making on Covid-19

Titli Sarkar [ [C00222141@louisiana.edu](mailto:C00269380@louisiana.edu) / titli010203[@gmail.com](mailto:seshagiriprabhu@gmail.com) ]

Supervisor: Dr. Satya Katragadda, Dr. Vijay Raghavan

## Problem Statement:

The Oxford Covid-19 Government Response Tracker (OxCGRT) collects systematic information on which governments have taken which measures, and when. This can help decision-makers and citizens understand governmental responses in a consistent way, aiding efforts to fight the pandemic.

Action-rules are mined from association rules which increases the probability that change in values of some flexible attributes in antecedent will enforce change of target variable to desired target value. Such a rule can either be interpreted as a recommendation for an action, or as a counterfactual explanation for the class assigned to the instance. Using action rules will significantly help policy makers to take better decisions.

OxCGRT collects publicly available information on 23 indicators of government response. They also published #of confirmed cases and #confirmed deaths. We need to choose target variables, fixed and flexible variables and action rule mining with those helps us to find out how the target variable is impacted (#of confirmed cases/ #of confirmed deaths) in order to be able to make good policy.

Our analysis part includes studying the impact of different factors on target variable like the following:

i> choice of flexible class,

ii> different thresholds,

iii> best combination of support, confidence and lift etc.

The whitepaper on Association Mining Based Approach to analyze COVID-19 response and Case Growth in the United States by S. Kataragadda et. al. can be found [here](https://drive.google.com/file/d/1T-9kivpEdjYVFWDZQJ4hG1QlfkXczya4/view?usp=sharing).

## Tools and datasets to be used:

**Framework:** A python module for Action Rules [4] - [lucasSycora](https://github.com/lukassykora/actionrules)~~.~~

**Oxford Dataset:** It was provided by Dr. Satya and can be found [here](https://drive.google.com/file/d/1CKx2t3vgxINwPPvY5uweeYLZSCdkjiM7/view?usp=sharing) and the explanation about the dataset can be found [5] [here](https://www.bsg.ox.ac.uk/sites/default/files/2021-05/BSG-WP-2020-034-v3.pdf).

## Data Preprocessing

**Oxford Dataset:** *main\_OxCGRTUS\_timeseries\_all.xlsx* **-** Provided by Dr. Satya [1] , explanation about dataset [2], git repo [3].

### Attributes

**Indices for overall measures (4):**

These four indices aggregate data into a single number – (0 to 100), they reflect the level of the overall government’s response. They measure of how many of the relevant indicators a government has acted upon, and to what degree. Higher is stricter.

* **government\_response\_index** (all indicators)
* **stringency\_index** (all C indicators + H1 which records public information campaigns)
* **containment\_health\_index** (all C and H indicators)
* **economic\_support\_index** (all E indicators)

**Indices for overall measures:**

Each of them focuses on granulated level of government responses in different fields. They are discrete values ranging between 0-5. We have removed all flags.

* **C** (1 to 8) - containment and closure policies
* **H** (1 to 8) - health system policies

**Target:**

* **confirmed\_cases** or
* **confirmed\_deaths**

**Newly Added attributes (3):**

We have added three more attributes from outside.

* **population density** [6] – Measures population per square miles.
* **urban vs rural population** [7] **–** Captures percentage of urbanization in each state
* **hhs regions** [8] – All states are divided in 10 hhs regions, used for analysis in higher level of granularity.

**Data Format:**

* Excel sheets, each sheet represents one attribute
* Data captured from Jan 1, 2020 – April 28, 2021 = 70 weeks

We have removed last two weeks for spurious data. So we have data for **68 weeks**.

* Datacaptured over **51 states**

### Preprocessing

* *Antecedents or attributes* - **Weekly average** across all states, ceiling of average values are considered.
* *Target variables* – **Cases per capita** (for each state) = cases per week / population \*100\*100

where,

cases per week = weekly |cumulative cases today – cumulative cases yesterday|

### Deciding Stable and Flexible Attributes

**Stable**

region, population\_density, urban\_percentage

**Flexible**

We decided to do experiments in two level of granularities, at first we wanted to know how overall government response indices are related to target variables in higher level, then we wanted to know how specific policies on containment and closure policies (C-index) and health system policies (H-index) are related to target variables.

* **Level 1:** stringency\_index, government\_response\_index, containment\_health\_index, economic\_support\_index
* **Level 2:** Remove Level 1 attribues and use (separately):
  + C-indices (containment and closure policies)
  + H-indices (health system policies)

### Discretization

We need to discretize the attributes and target variables to make them suitable for numerical evaluation and implementation in a uniform representation across the dataset. We observed the distribution of each attribute and found out that not all of them have same distribution. We have used different range and bin boundaries for each of them depending on the data distribution.

* C and H attributes are already discretized values. So we took only weekly average and took the ceiling.
* Details of the other variables are listed in *Table 1*.

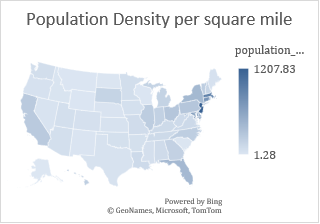
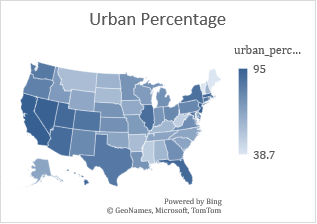
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Attribute** | **Type** | **Discretized Values** | **Percentage** | **Actual Values** |
| confirmed\_cases\_discretized | Target | high | 100 | 135.81 to 1254.23 |
| confirmed\_cases\_discretized | Target | medium | 66 | 38.30 to 135.81 |
| confirmed\_cases\_discretized | Target | low | 33 | 0.00 to 38.30 |
|  |  |  |  |  |
| confirmed\_deaths\_discretized | Target | high | 100 | 2.25 to 43.63 |
| confirmed\_deaths\_discretized | Target | medium | 66 | 0.67 to 2.25 |
| confirmed\_deaths\_discretized | Target | low | 33 | 0.00 to 0.67 |
|  |  |  |  |  |
| population\_density\_discretized | Stable | high | 100 | 174.47 to 1207.83 |
| population\_density\_discretized | Stable | medium | 66 | 58.03 to 174.47 |
| population\_density\_discretized | Stable | low | 33 | 1.28 to 58.03 |
|  |  |  |  |  |
| urban\_percentage\_discretized | Stable | high | 100 | 73.75 to 95.00 |
| urban\_percentage\_discretized | Stable | medium | 50 | 64.80 to 73.75 |
| urban\_percentage\_discretized | Stable | low | 25 | 38.70 to 64.80 |
|  |  |  |  |  |
| region | Stable | Region1\_Boston | | 'Connecticut', 'Maine', 'Massachusetts', 'New Hampshire', 'Rhode Island', 'Vermont' |
| region | Stable | Region2\_NewYork | | 'New Jersey', 'New York' |
| region | Stable | Region3\_Philadelphia | | 'Delaware', 'Maryland', 'Pennsylvania', 'Virginia', 'West Virginia' |
| region | Stable | Region4\_Atlanta | | 'Alabama', 'Florida', 'Georgia', 'Kentucky', 'Mississippi', 'North Carolina', 'South Carolina', 'Tennessee' |
| region | Stable | Region5\_Chicago | | 'Illinois', 'Indiana', 'Michigan', 'Minnesota', 'Ohio', 'Wisconsin' |
| region | Stable | Region6\_Dallas | | 'Arkansas', 'Louisiana', 'New Mexico', 'Oklahoma', 'Texas' |
| region | Stable | Region7\_KansasCity | | 'Iowa', 'Kansas', 'Missouri', 'Nebraska' |
| region | Stable | Region8\_Denver | | 'Colorado', 'Montana', 'North Dakota', 'South Dakota', 'Utah', 'Wyoming' |
| region | Stable | Region9\_SanFrancisco | | 'Arizona', 'California', 'Hawaii', 'Nevada' |
| region | Stable | Region10\_Seattle | | 'Alaska', 'Idaho', 'Oregon', 'Washington' |
|  |  |  |  |  |
| stringency\_index\_discretized | Flexible | 5 | 100 | 68.52 to 93.52 |
| stringency\_index\_discretized | Flexible | 4 | 80 | 59.91 to 68.52 |
| stringency\_index\_discretized | Flexible | 3 | 60 | 53.97 to 59.91 |
| stringency\_index\_discretized | Flexible | 2 | 40 | 43.42 to 53.97 |
| stringency\_index\_discretized | Flexible | 1 | 20 | 0.00 to 43.42 |
|  |  |  |  |  |
| containment\_health\_index\_discretized | Flexible | 5 | 100 | 65.48 to 79.66 |
| containment\_health\_index\_discretized | Flexible | 4 | 80 | 60.24 to 65.48 |
| containment\_health\_index\_discretized | Flexible | 3 | 60 | 55.72 to 60.24 |
| containment\_health\_index\_discretized | Flexible | 2 | 40 | 46.43 to 55.72 |
| containment\_health\_index\_discretized | Flexible | 1 | 20 | 0.00 to 46.43 |
|  |  |  |  |  |
| economic\_support\_index\_discretized | Flexible | 5 | 100 | 75.0 to 100.0 |
| economic\_support\_index\_discretized | Flexible | 4 | 80 | 62.5 to 75.0 |
| economic\_support\_index\_discretized | Flexible | 3 | 60 | 37.5 to 62.5 |
| economic\_support\_index\_discretized | Flexible | 2 | 40 | 25.0 to 37.5 |
| economic\_support\_index\_discretized | Flexible | 1 | 20 | 0.00 to 25.0 |
|  |  |  |  |  |
| government\_response\_index\_discretized | Flexible | 5 | 100 | 65.44 to 80.21 |
| government\_response\_index\_discretized | Flexible | 4 | 80 | 59.96 to 65.44 |
| government\_response\_index\_discretized | Flexible | 3 | 60 | 54.81 to 59.96 |
| government\_response\_index\_discretized | Flexible | 2 | 40 | 45.31 to 54.81 |
| government\_response\_index\_discretized | Flexible | 1 | 20 | 0.00 to 45.31 |

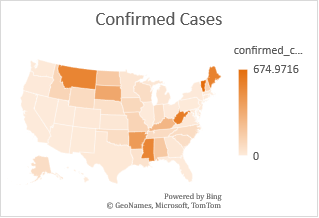
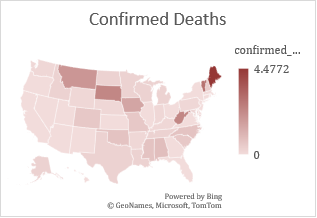
*Table 1: Details of attributes*

## Results

First, we wanted to analyze data before implementing action rules:

### General Overview across United States

### Global Comparison

Next, we wanted to observe how different flexible attributes are related to target variables. Data for 68 weeks across all states (average) are plotted. First, four overall government response indices are plotted with target variables, then, C and H indices are plotted.

We can see that there is no strong correlation is observed between target variable and any of the flexible attributes.

### Action Rules

#### Step 1 - Classification rules from frequent itemset

First, we wanted to observe which combination of support and confidence generates a feasible number of association rules to focus on which support and confidence values to be used for action rules generation.

#of total transactions = 3400

We have used the Apriori module from lucasSycora to generate association rules for each combination of support and confidence in the following range and saved the results in a file */data/specific index/parameter\_combos.txt*:

for sup in range(100 to 1 in steps of 5):

for cnf in range (100 to 1 in steps of 10):

actionRDiscovery.decisions.prepare\_data\_fim(attributes, target)

actionRDiscovery.decisions.fit\_fim\_apriori(conf=cnf, support=sup)

transactions = actionRDiscovery.decisions.transactions

assoRules = actionRDiscovery.decisions.rules

fout.writelines([str(sup), "\t", str(cnf), "\t", str(len(transactions)), "\t", str(len(assoRules)), "\n"])

We observed that for each type of attributes, the best combination is support ranging from (10-15) and confidence ranging from (60-70). Any action rules generated with lower support and confidence values than these does not hold up for a strong actin rule.

#### Step 2 – Generate Action Rules

Action rules can be generated in two steps:

**Phase 1: Classification Rule Mining**



where a1, b1, c1, e1 are values of features (attributes) in the antecedent of the rule, and d1 is a value of the target variable predicted by the consequent of the rule.

Each discovered rule is accompanied by values of support and confidence that express its quality. = #of transactions that match both the antecedent and consequent of the given rule, =  = ratio between #of transactions that satisfy the antecedent as well as the consequent to #of transactions that satisfy the antecedent.

**Phase 2: Generation of Action Rules**



where is a fixed condition (set of stable attributes), which describes the object,  is the proposed to change to a subset of user-designated flexible attributes and represents the implied change to the target attribute.

The quality of action rules can be represented using the following measures. Let r be an action rule, which was generated from two classification rules r1 and r2:







Note that the definitions of support and confidence for action rules are different from these definitions for classification rules. Following is the general code for action rule discovery, we need to use different parameters for different cases:

function actionRuleDiscovery(conf, supp, level, targetname):

actionRDiscovery = ActionRulesDiscovery()

actionRDiscovery.load\_pandas(df)

actionRDiscovery.fit(stable\_attributes = stable,

flexible\_attributes = flexible,

consequent = targetname,

conf=conf,

supp=supp,

desired\_classes = [level],

)

pretty\_ar = actionRDiscovery.get\_pretty\_action\_rules()

print(level, conf, supp, "Rules:", pretty\_ar)

We have generated action rules for both target variables *confirmed\_cases* and *confirmed\_deaths.*

The stable attributes and target name is are fixed in all the cases:

stable attributes = *region, population\_density\_discretized, urban\_percentage\_discretized*

target name *= confirmed\_deaths\_discretized* or *confirmed\_deaths\_discretized*

#### Experiments using Government Response Index

At the first stage we wanted to evaluate how target variable is affected by four main government response index which captures the overall government responses in all broad categories in different sectors. Here the flexible attributes used are as follows:

flexible attributes = *stringency\_index\_discretized, government\_response\_index\_discretized,containment\_health\_index\_discretized, economic\_support\_index\_discretized*

Our goal is to find meaningful action rules which can help in improved recommendation. Therefore, though we have generated and studied action rules with all kinds of targets: *high*, *medium,* and *low*, but we focused on action rules with target *low* which means how the change of flexible attributes changes target *confirmed\_cases/ confirmed\_deaths* from *high* or *medium* to *low*.

If we study relationship between each government response index and a target variable, we can notice that there is not a strong co-relation.

The best possible rules for target = *low* were generated with maximum:

confirmed\_deaths = low where Support = 10, confidence = 50 => #of actionrules = 3

Rule 1: If attribute 'population\_density\_discretized' is 'low', attribute 'economic\_support\_index\_discretized' value '2' is changed to '1', then 'confirmed\_deaths\_discretized' value 'high' is changed to 'low' with support: 0.12735294117647059, confidence: 0.4859139414802065 and uplift: 0.11061344537815124.

Rule 2: If attribute 'population\_density\_discretized' is 'low', attribute 'economic\_support\_index\_discretized' value '2' is changed to '1', attribute 'government\_response\_index\_discretized' value '3' is changed to '1', then 'confirmed\_deaths\_discretized' value 'high' is changed to 'low' with support: 0.10676470588235294, confidence: 0.5276162790697675 and uplift: 0.10676470588235294.

Rule 3: If attribute 'population\_density\_discretized' is 'low', attribute 'economic\_support\_index\_discretized' value '2' is changed to '1', attribute 'stringency\_index\_discretized' value '3' is changed to '1', then 'confirmed\_deaths\_discretized' value 'high' is changed to 'low' with support: 0.10529411764705883, confidence: 0.5512020047570506 and uplift: 0.10485698866702643.

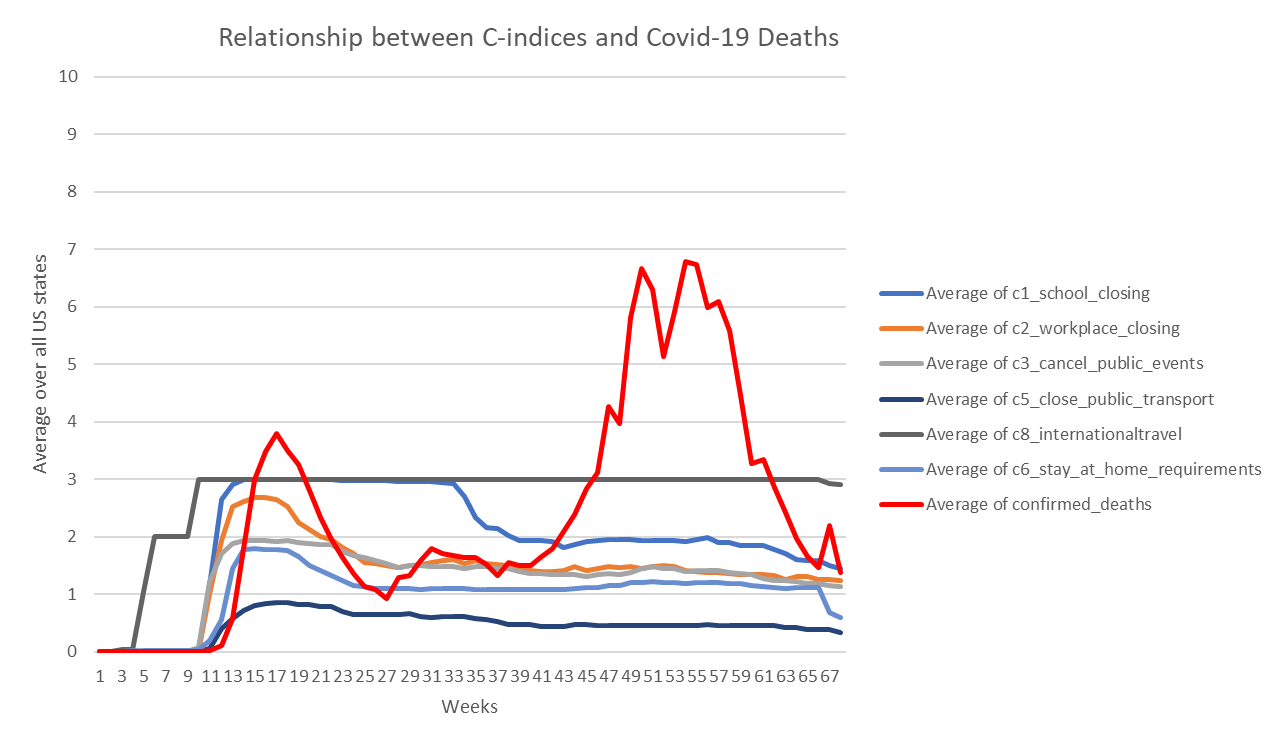
confirmed\_cases = low where Support = 12, confidence = 50 => #of actionrules = 17

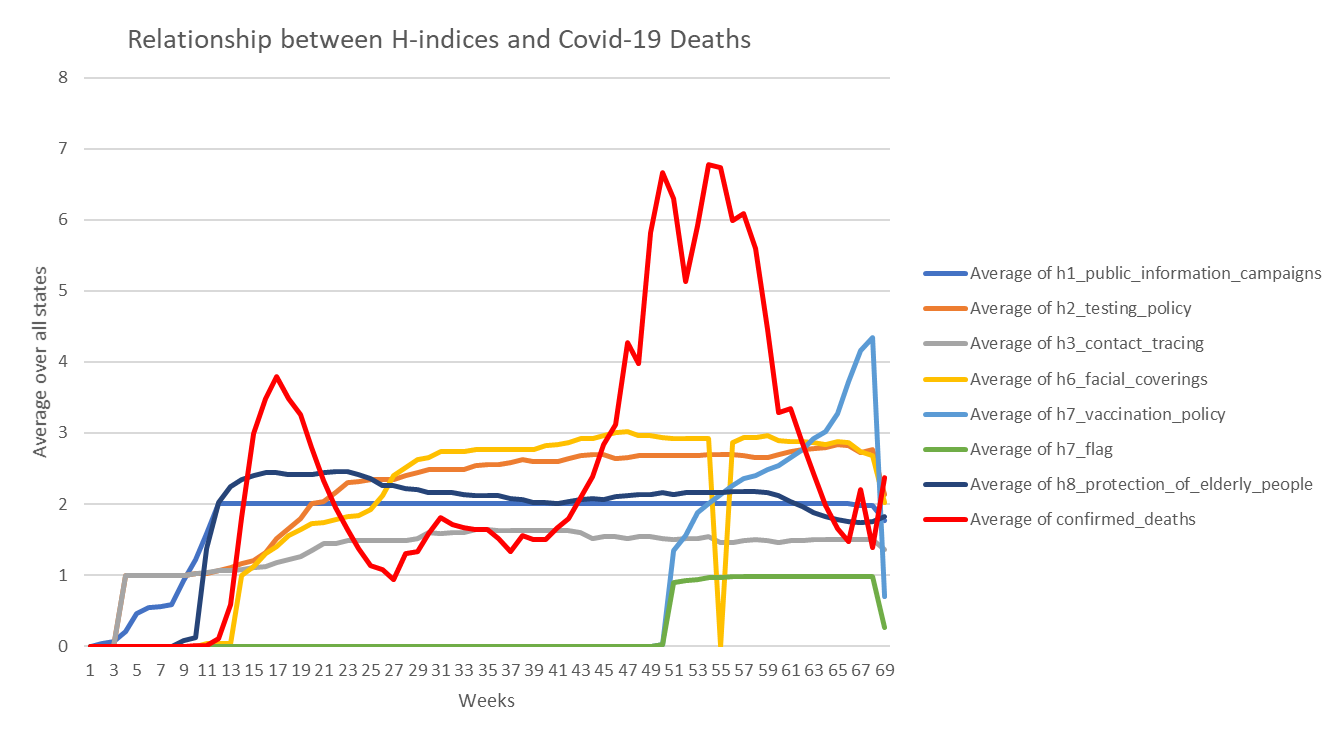
By studying all the rules we have noticed the trend that change of each flexible attribute values from higher level to lower results in changing the target variable to higher level to low. The downside of using these four indices that (i) the support and confidence is not enough high and (ii) we found that there are a few spurious rules where the trend is opposite.

#### Experiments using H and C Index

To overcome the above mentioned challenges of using government indices, we have used a more detailed level of response indices for C - containment and closure policies and H - health system policies.

At first, we have studied the relationships between these indices and target and noticed the similar co-relations as above 4 indices.





Now, we have generated action rules with each of the indices. The best combinations of support and confidence is as follows:

C-index | target = confirmed\_deaths where Support = 10, confidence = 90 => #of asso Rules = 248

C-index | confirmed\_deaths=low where Support = 10, confidence = 60 => #of actionules = 6

C-index | confirmed\_cases=low where Support = 5 (very low), confidence = 70 => #of actionules = 3

C-index | confirmed\_cases=low where Support = 12, confidence = 60 => #of actionules = 8

C-index | confirmed\_cases=low where Support = 10, confidence = 60 => #of actionules = 13

and

H-index | target = confirmed\_deaths where Support = 10, confidence = 70 => #of asso Rules = 40

H-index | confirmed\_deaths=low where Support = 5(very low), confidence = 70 => #of actionules =3

H-index | confirmed\_deaths=low where Support = 14, confidence = 40(very low)=> #of actionules =3

H-index | confirmed\_cases=low where Support = 5 (very low), confidence = 70 => #of actionules = 3

H-index | confirmed\_cases=low where Support = 12, confidence = 50(low) => #of actionules = 3

The trend of result is same as 4 overall government indices, but there is no spurious rules. Few examples of rules are as follows:

**C-index | supp = 10, conf = 60:**

Rule 1: If attribute 'population\_density\_discretized' is 'low', attribute 'c2\_workplace\_closing' value '1.0' is changed to '0.0', attribute 'c1\_school\_closing' value '2.0' is changed to '0.0', then 'confirmed\_deaths\_discretized' value 'high' is changed to 'low' with support: 0.12176470588235294, confidence: 0.6017441860465116 and uplift: 0.12176470588235294.

Rule 2: If attribute 'population\_density\_discretized' is 'low', attribute 'c2\_workplace\_closing' value '1.0' is changed to '0.0', attribute 'c1\_school\_closing' value '2.0' is changed to '0.0', attribute 'c7\_movementrestrictions' value '1.0' is changed to '0.0', then 'confirmed\_deaths\_discretized' value 'high' is changed to 'low' with support: 0.10823529411764705, confidence: 0.6123128119800333 and uplift: 0.10823529411764705.

**H-index |** **supp = 7, conf = 50**:

Rule 1: If attribute 'population\_density\_discretized' is 'low', attribute 'h2\_testing\_policy' value '3.0' is changed to '1.0', attribute 'h6\_facial\_coverings' value '3.0' is changed to '0.0', then 'confirmed\_deaths\_discretized' value 'high' is changed to 'low' with support: 0.12058823529411765, confidence: 0.49160671462829736 and uplift: 0.12165608689518972.

Rule 2: If attribute 'population\_density\_discretized' is 'low', attribute 'h1\_public\_information\_campaigns' value '2.0' is changed to '0.0', attribute 'h7\_vaccination\_policy' value '2.0' is changed to '0.0', then 'confirmed\_deaths\_discretized' value 'high' is changed to 'low' with support: 0.075, confidence: 0.7919254658385093 and uplift: 0.075.

We observed that change of H or C indices from higher to lower values changes target values from high/medium to low.

## Conclusion

From the overall discussion we observed the following:

* None of the response indices are strongly co-related with target variables.
* Therefore, we could not generate any action rules with very high support and confidence.
* Overall, the rules generated with moderately high support and confidence values makes sense in real world, means change in flexible attributes positively change values of target variables.

All codes are available at github:

## References

[1] <https://drive.google.com/file/d/1CKx2t3vgxINwPPvY5uweeYLZSCdkjiM7/view?usp=sharing>

[2] <https://www.bsg.ox.ac.uk/sites/default/files/2021-05/BSG-WP-2020-034-v3.pdf>

[3] [OxCGRT](https://github.com/OxCGRT)/[covid-policy-tracker](https://github.com/OxCGRT/covid-policy-tracker)

[4] [GitHub - lukassykora/actionrules: Action Rules Mining](https://github.com/lukassykora/actionrules)

[5] [BSG-WP-2020-034-v3.pdf (ox.ac.uk)](https://www.bsg.ox.ac.uk/sites/default/files/2021-05/BSG-WP-2020-034-v3.pdf)

[6] [https://www.stat=ta.com/stat=tics/183588/population-density-in-the-federal-states-of-the-us/](https://www.statista.com/statistics/183588/population-density-in-the-federal-states-of-the-us/)

[7] [https://www.icip.iastate.edu/tables/population/urban-pct states%20we%20have%20used%202010%20urban%20population](https://www.icip.iastate.edu/tables/population/urban-pct%20states%20we%20have%20used%202010%20urban%20population)

[8] <https://www.hhs.gov/about/agencies/iea/regional-offices/index.html>