Supplementary Material

Machine Learning Methods for "Wicked" Problems: Exploring the Complex Drivers of Modern Slavery

Rosa Lavelle-Hill
The Alan Turing Institute, UK
University of Tübingen, Germany
rosa.lavelle-hill@uni-tuebingen.de

Gavin Smith
University of Nottingham, UK

Anjali Mazumder
The Alan Turing Institute, UK

Todd Landman
University of Nottingham, UK

James Goulding
University of Nottingham, UK

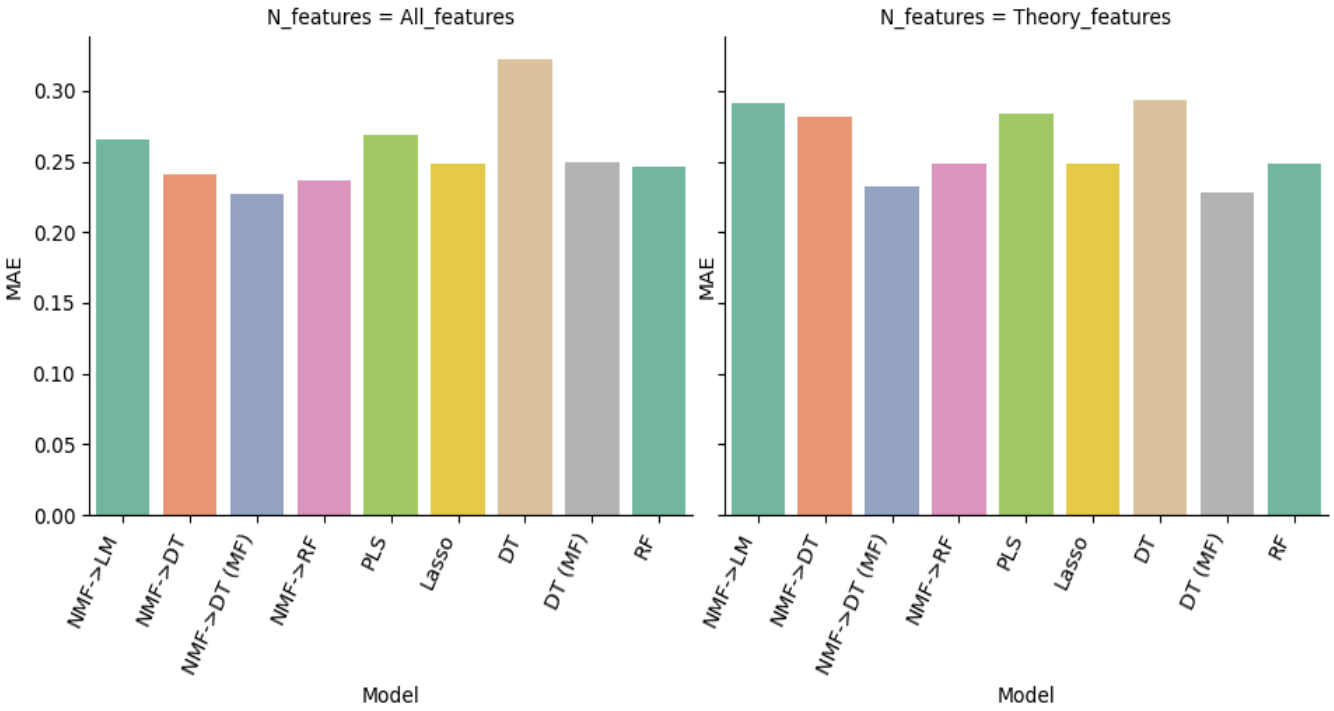
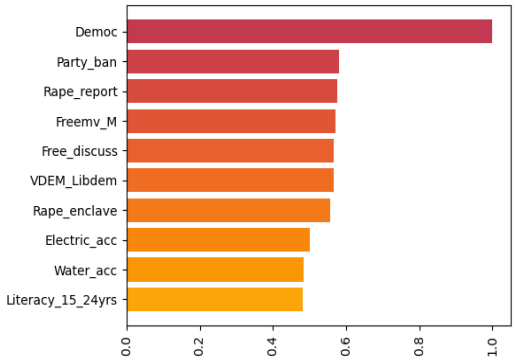
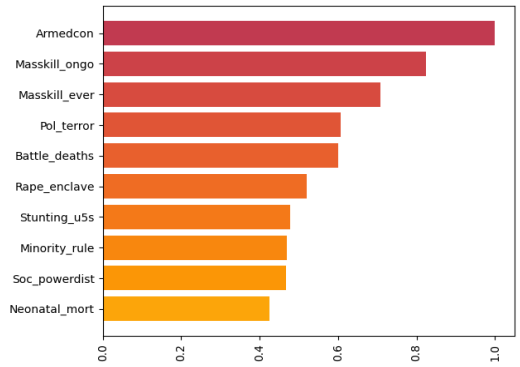


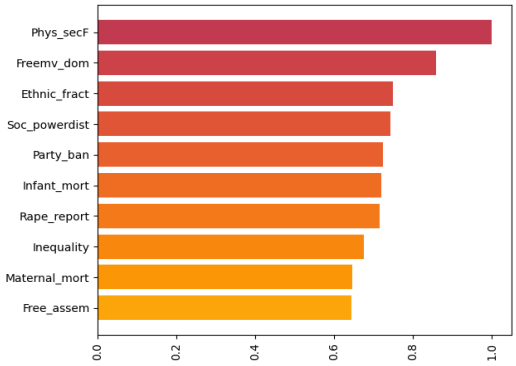
Figure 1. Graphs to visually compare the pipeline performances. NMF = Non Negative Matrix Factorisation; DT = Decision Tree; MF = Max features as a parameter available to the decision tree; RF = Random Forest; LM = Linear (regression) Model; PLS = Partial Least Squares regression.



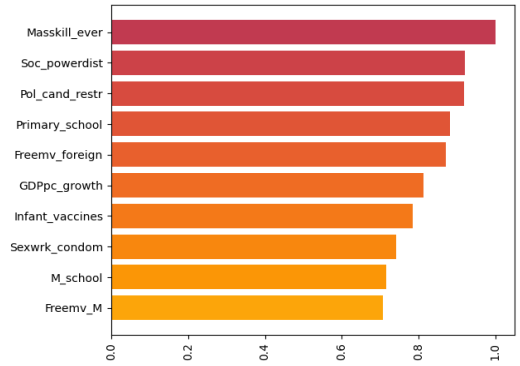
(a) **Component 1: Democratic Rule**
The top loading variable is Democracy. Other high loading variables *directly* related to democracy include bans for political parties, freedom of discussion, and liberal democracy.



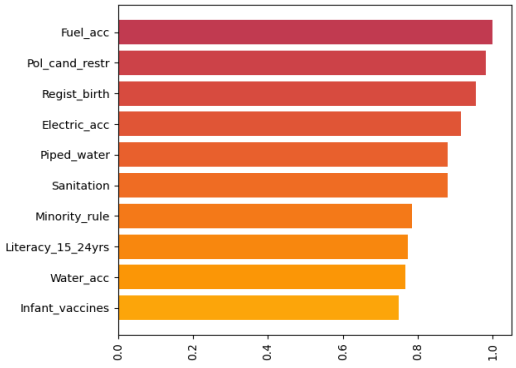
(b) **Component 2: Armed Conflict**
The high loading variables are armed conflict, an ongoing (or ever has been a) mass killing, political terror, battle deaths, and enclaves of rape (i.e. in refugee or military camps).



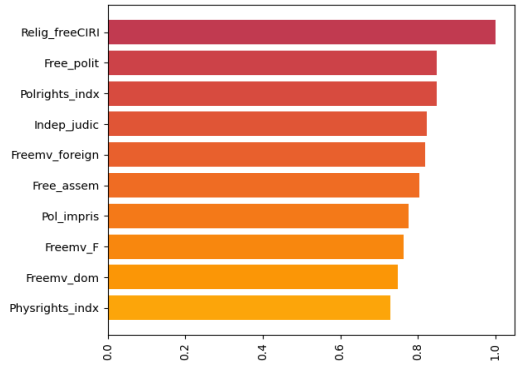
(c) **Component 3: Physical Security of Women**
The top loading variable represents the physical security of women. Other relevant variables include the unequal distribution of power, and the laws/protection around reporting rape.



(d) **Component 4: Social Inequality and Discrimination**
The two highest loading variables depicts whether there has ever been a mass killing, such as a genocide, before; and the unequal distribution of power between social groups.



(e) **Component 5: Access to Resources**
High loading relevant variables include access to fuel, electricity, water, and sanitation. Registered births, poor literacy, and a lack of infant vaccines also denote poor access to infrastructure/services.



(f) **Component 6: Religious and Political Freedoms**
High loading variables include religious and political freedoms, political and physical rights, independent judicial system, foreign and domestic freedom of movement, freedom to assemble, political imprisonment, and freedom of movement for women.

Figure 2. The six components from the best model. On the y axis are the original raw variables, and the bars show their loading value onto the new component (normalised to be between 0 and 1). The full variable names, a short description, and their source can be found in the Supplementary Materials. Note that here only the top 10 loading variables are shown for simplicity. The themes the components represent were decided after conversations with domain experts.

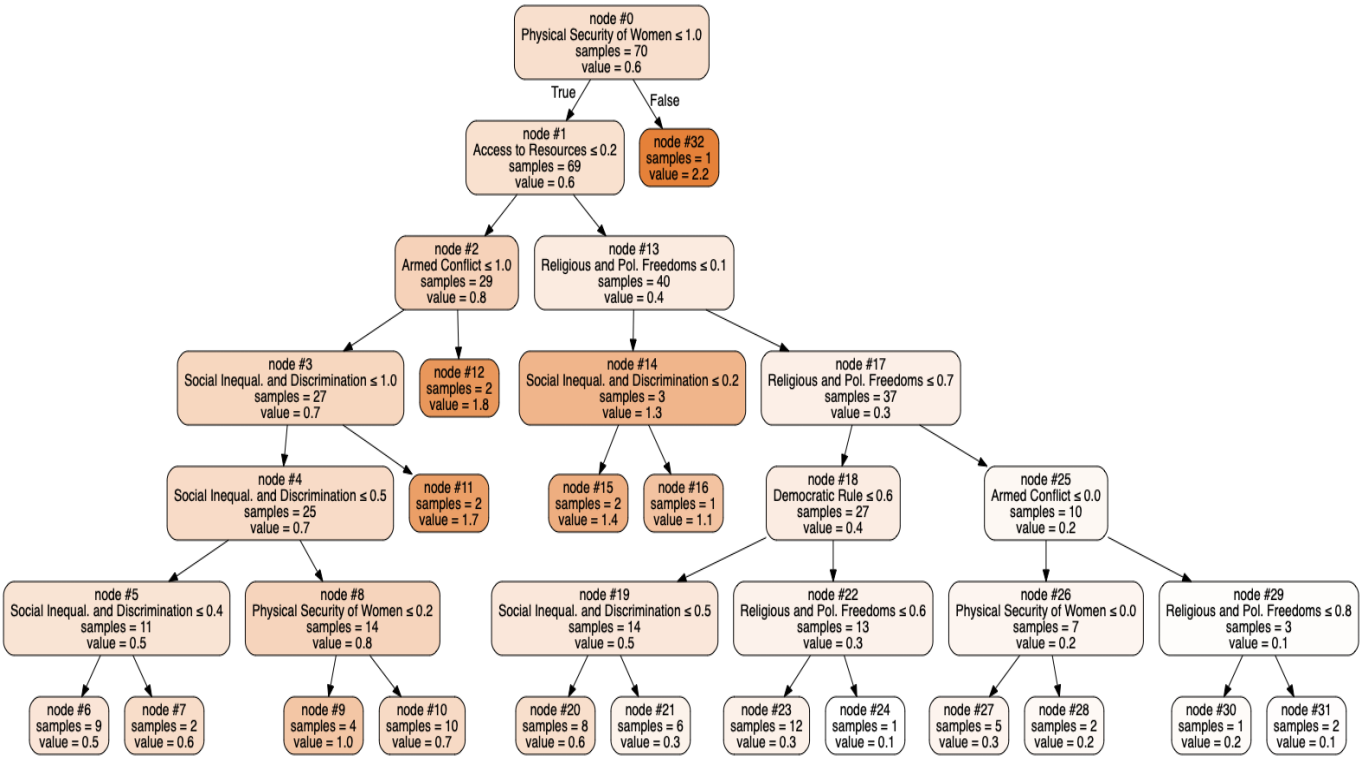
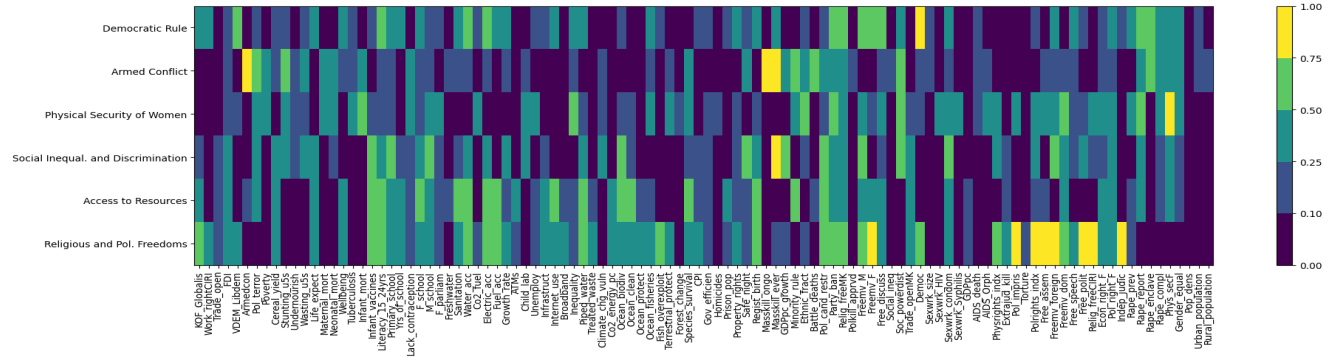
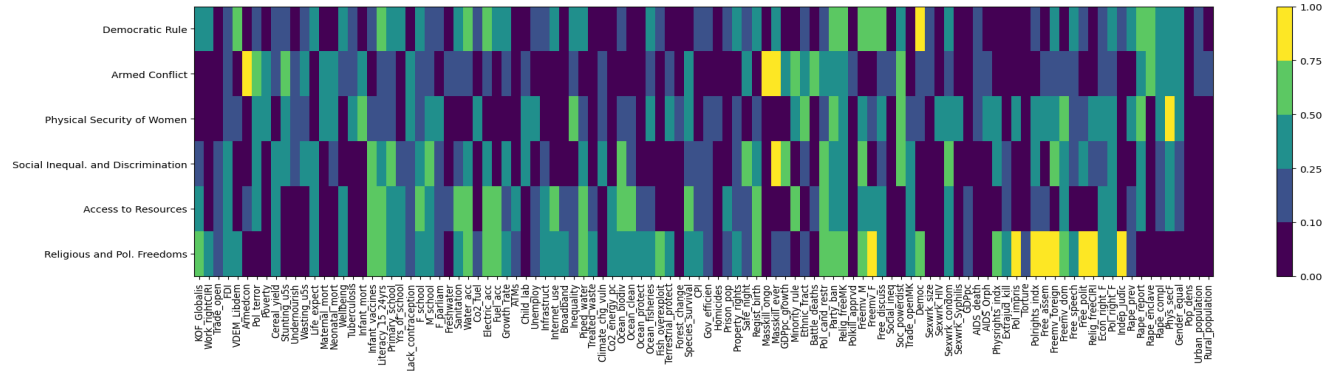


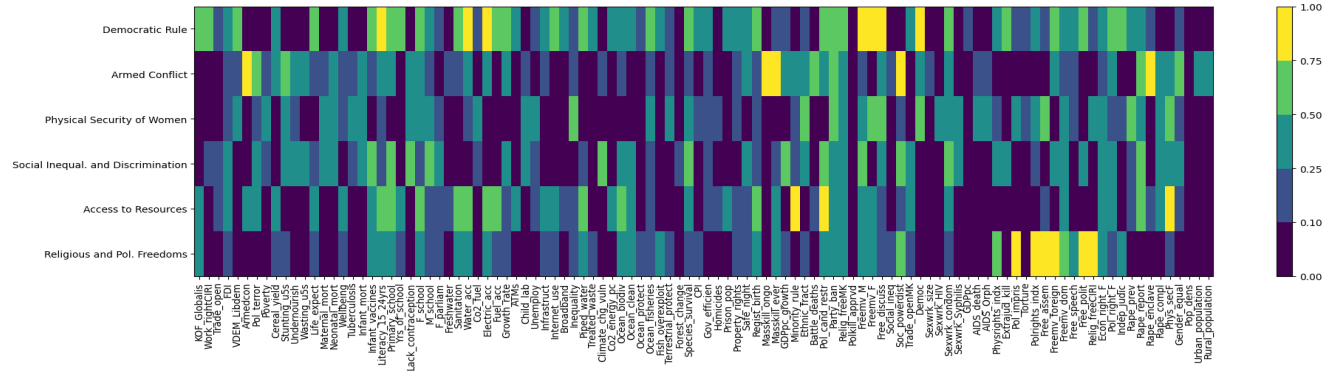
Figure 3. The best decision tree model predicting prevalence of modern slavery using the six components from the NMF. 'Samples' represents the number of country-year data points at each node (out of 70). The 'value' refers to the predicted prevalence (slavery as a percentage of the population) for that node.



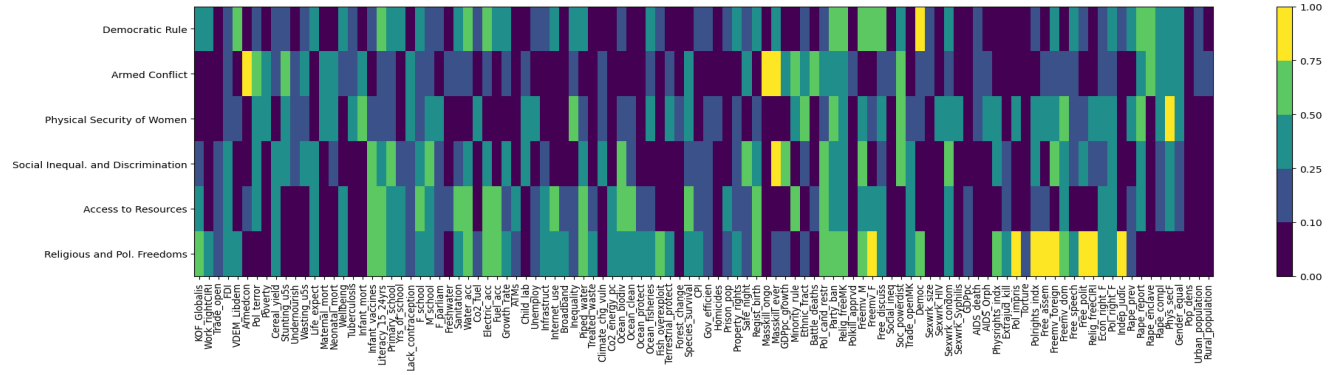
(a) Best Model



(b) Rashomon 2



(c) Rashomon 3



(d) Rashomon 4

Figure 4. The normalised variable loadings onto the components (H matrix from the NMF) for the Rashomon models with the same pipeline as the best model (all the features and the model class NMF->DT (MF)). Some differences in loadings occur due variances in the NMF paramterisation (different random seeds being used for coordinate descent, and different constants that multiply the regularisation terms (alpha)) yet the overarching themes remain stable.

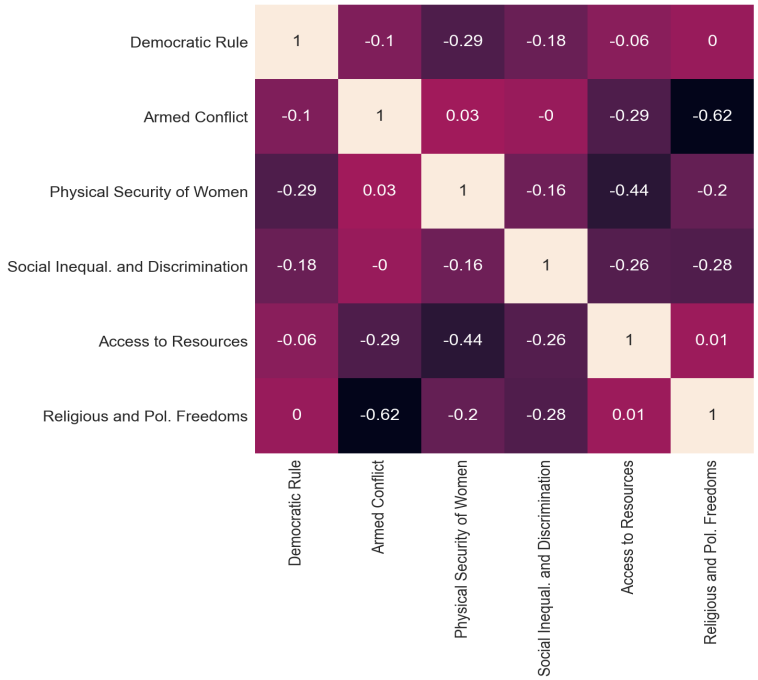


Figure 5. A correlation matrix showing the Pearson r correlations between the components. Here a high value for Physical Security of Women, depicts less security.

Figure 6. Missing Data for all 106 variables before data was imputed using CART.

Table 1
A comparison of 2018 prevalence estimates made by our model compared those made by the WFF's model as part of the 2018 GSI. Missing data represents the percentage of missing independent variable data for our model before imputations were made.

Country	% Missing Data	Our Best Model Estimate	GSI Estimate
Albania	5	0.1	0.69
Algeria	2	0.61	0.27
American Samoa	94	0.5	0.26
Andorra	83	0.18	0.08
Angola	4	0.69	0.72
Antigua and Barbuda	82	0.18	0.33
Aruba	92	0.5	0.3
Australia	4	0.18	0.06
Austria	11	0.26	0.17
Azerbaijan	8	0.28	0.45
Bahamas	78	0.18	0.38
Bahrain	39	0.61	0.19
Bangladesh	0	0.69	0.37
Barbados	53	0.18	0.27
Belarus	9	0.18	1.09
Belgium	7	0.18	0.2
Belize	75	0.18	0.34
Benin	3	0.5	0.55
Bermuda	94	0.69	0.4
Bhutan	74	0.69	1.13
Bolivia	10	0.5	0.21
Bosnia and Herzegovina	8	0.28	0.34
Brazil	1	0.18	0.18
British Virgin Islands	96	0.18	0.29
Brunei Darussalam	62	0.61	1.09
Bulgaria	3	0.1	0.45
Burkina Faso	10	0.5	0.45
Burundi	8	0.69	4.0
Cabo Verde	25	0.5	0.41
Canada	7	0.18	0.05
Cayman Islands	93	0.18	0.21
Central African Republic	17	0.5	2.23
Chad	8	0.5	1.2
China	7	0.28	0.28
Comoros	76	0.28	0.55
Congo Republic	7	0.69	0.8
Costa Rica	2	0.18	0.13
Cote d'Ivoire	8	0.69	0.59
Croatia	6	0.18	0.6
Cuba	39	0.61	0.38
Curacao	94	0.61	0.4
Cyprus	8	0.18	0.42
Denmark	6	0.18	0.16
Djibouti	36	0.28	0.71
Dominica	85	0.5	0.34
Dominican Republic	0	0.28	0.4
Ecuador	2	0.18	0.24

Table 1 continued from previous page

Country	% Missing Data	Our Best Model Estimate	GSI Estimate
El Salvador	0	0.61	0.25
Equatorial Guinea	40	0.28	0.64
Eritrea	45	0.28	9.3
Estonia	2	0.18	0.36
Eswatini	8	0.61	0.88
Faeroe Islands	93	0.18	0.77
Fiji	80	0.18	0.52
Finland	4	0.18	0.17
France	5	0.18	0.2
French Polynesia	96	0.18	0.4
Gabon	2	0.61	0.48
Gambia	5	0.64	0.58
Germany	3	0.18	0.2
Gibraltar	96	0.18	0.4
Greece	5	0.18	0.79
Greenland	94	0.18	0.2
Grenada	83	0.18	0.41
Guam	94	0.18	0.29
Guinea-Bissau	36	0.5	0.75
Guinea	2	0.69	0.78
Guyana	7	0.5	0.26
Hong Kong	92	0.61	0.14
Iceland	25	0.18	0.21
Iran	7	0.61	1.62
Iraq	11	0.61	0.48
Ireland	3	0.18	0.17
Isle of Man	95	0.15	0.06
Israel	6	0.08	0.39
Italy	4	0.18	0.24
Jamaica	3	0.18	0.26
Japan	5	0.26	0.03
Kazakhstan	7	0.61	0.42
Kenya	2	0.5	0.69
Kiribati	83	0.69	0.52
Kuwait	10	0.61	0.15
Kyrgyz Republic	6	0.61	0.41
Laos	12	0.69	0.94
Lesotho	6	0.5	0.42
Liberia	2	0.5	0.74
Libya	40	0.61	0.77
Liechtenstein	85	0.18	0.17
Lithuania	6	0.18	0.58
Luxembourg	27	0.18	0.15
Macao	92	0.18	0.29
Macedonia	11	0.18	0.87
Madagascar	1	0.69	0.75
Malaysia	2	0.61	0.69
Maldives	77	0.28	0.64
Mali	7	0.5	0.36
Malta	74	0.18	0.37
Marshall Islands	85	0.5	0.33

Table 1 continued from previous page

Country	% Missing Data	Our Best Model Estimate	GSI Estimate
Mauritius	6	0.5	0.1
Micronesia, Fed. Sts.	86	0.5	0.87
Moldova	11	0.18	0.55
Monaco	84	0.18	0.06
Montenegro	6	0.18	0.59
Mozambique	3	0.69	0.54
Namibia	4	0.5	0.33
Nauru	83	0.15	0.17
Netherlands	3	0.18	0.18
New Caledonia	96	0.69	0.67
New Zealand	5	0.18	0.06
Nicaragua	0	0.61	0.29
Niger	9	0.5	0.67
North Korea	52	0.28	10.46
Northern Mariana Islands	94	0.61	0.29
Norway	6	0.26	0.18
Oman	10	0.28	0.21
Palau	88	0.18	0.02
Palestine	91	0.61	0.53
Panama	0	0.18	0.21
Papua New Guinea	39	0.69	1.03
Paraguay	5	0.15	0.16
Peru	0	0.69	0.26
Portugal	2	0.18	0.25
Puerto Rico	92	0.61	0.4
Qatar	8	0.61	0.15
Rwanda	6	0.69	1.16
Saint-Martin	98	0.69	0.96
Samoa	82	0.28	1.09
San Marino	83	0.18	0.15
Sao Tome and Principe	77	0.5	0.4
Saudi Arabia	10	0.61	0.19
Senegal	0	0.5	0.29
Seychelles	78	0.18	0.18
Sierra Leone	4	0.5	0.5
Sint Maarten	95	0.61	0.21
Slovakia	10	0.18	0.29
Slovenia	6	0.18	0.22
Solomon Islands	77	0.69	1.0
Somalia	37	1.06	1.55
South Korea	12	0.26	0.19
South Sudan	41	0.5	2.05
Spain	2	0.18	0.23
St. Kitts and Nevis	83	0.18	0.42
St. Lucia	77	0.18	0.4
St. Vincent and the Grenadines	82	0.18	0.17
Sudan	34	0.61	1.2
Suriname	21	0.15	0.23
Sweden	8	0.18	0.16
Switzerland	10	0.18	0.17
Syria	44	0.61	0.73

Table 1 continued from previous page

Country	% Missing Data	Our Best Model Estimate	GSI Estimate
Tajikistan	6	0.28	0.45
Tanzania	6	0.69	0.62
Timor-Leste	42	0.5	0.77
Togo	8	0.5	0.68
Tonga	80	0.5	0.75
Trinidad and Tobago	13	0.18	0.3
Turkey	6	0.28	0.65
Turkmenistan	40	0.28	1.12
Turks and Caicos Islands	94	0.18	0.24
Tuvalu	86	0.5	0.64
United Arab Emirates	11	0.61	0.17
United Kingdom	7	0.18	0.21
United States	7	0.18	0.13
United States Virgin Islands	94	0.64	0.24
Uruguay	0	0.18	0.1
Uzbekistan	35	0.28	0.52
Vanuatu	78	0.5	0.41
Venezuela	12	0.61	0.56
Western Sahara	98	0.69	0.48
Yemen	5	0.61	0.31
Zambia	5	0.5	0.57
Zimbabwe	7	0.61	0.67

Model Parameters

The best performing model (MAE=0.227) used the following meta-parameters for the NMF and Decision Tree functions in the Scikit-learn python package: random state = 45, NMF K components = 6, NMF solver = 'cd', NMF tolerance = 0.005, NMF alpha = 2, NMF max iterations = 350, max depth of the tree = 6; tree max features = 0.3, tree minimum samples to split on = 3.

Variable Descriptions

The short variable descriptions, their sources, and whether or not the variable was selected based on theory or not can be found on the GitHub repository ml-slavery:
`https://github.com/ml-slavery/ml-slavery/blob/main/Data/Meta_Data/Variable_descriptions.csv`