Analiza skupień państw i narodów w kontekście gospodarczo-politycznym

Maria Nowicka, Zuzanna Tabisz

Agenda

1. Wprowadzenie
2. Dane
3. Kroki analizy:

- Przetwarzanie wstępne danych
- Podział na przedziały: K-bins
- Analiza SVD
- t-SNE
- Klastrowanie
- Ewaluacja i Wyniki

Wprowadzenie

Wprowadzenie

- Cel projektu: Zgrupowanie krajów na podstawie cech ekonomicznych, środowiskowych, edukacyjnych itp.

- Metody: Przetwarzanie danych, K-bins, SVD, t-SNE, K-średnich, ewaluacja

- Znaczenie: Identyfikacja podobieństw między krajami dla lepszego zrozumienia globalnych wzorców

Dane

Dane

Źródło danych:

- World Bank – World Development Indicators (WDI):
- https://databank.worldbank.org/source/world-development-indicators

Zbiory danych obejmujące globalne wskaźniki rozwoju społeczno-ekonomicznego i środowiskowego. Dane pogrupowane według głównych kategorii:

- **Polityka gospodarcza i zadłużenie (Economic Policy & Debt)**
- **Edukacja (Education)**
- **Środowisko (Environment)**
- **Sektor finansowy (Financial Sector)**
- **Zdrowie (Health)**
- **Sektor prywatny i handel (Private Sector & Trade)**
- **Sektor publiczny (Public Sector)**
- **Ochrona socjalna i rynek pracy (Social Protection & Labor)**

Dane

Format danych:

- Pliki CSV
- Kody krajów w formacie ISO 3166-1 alpha-3
- Dane liczbowe (wskaźniki numeryczne)

Zakres danych:

- wb_econ — Wskaźniki ekonomiczne (np. PKB per capita, wzrost sektora rolnictwa)
- wb_env — Dane środowiskowe (np. emisje CH_4, dostęp do energii)
- wb_edu — Edukacja (np. wskaźniki szkolnictwa)
- wb_fin — Finanse (np. usługi finansowe, transfery międzynarodowe)
- wb_health — Zdrowie (np. dane demograficzne, wydatki na zdrowie)
- wb_private — Sektor prywatny i handel
- wb_public — Sektor publiczny
- wb_soc — Ochrona socjalna i rynek pracy

Przetwarzanie wstępne danych

Przetwarzanie wstępne danych

Kroki:

- Usunięcie kolumn o stałej wartości (nunique ≤ 1)
- Usunięcie wysoko skorelowanych cech (korelacja > 0.95)
- Konwersja danych na wartości numeryczne
- Wypełnianie brakujących danych medianą
- Wybór cech na podstawie wariancji po skalowaniu (MinMaxScaler)

K-bins

K-bins

Metoda:

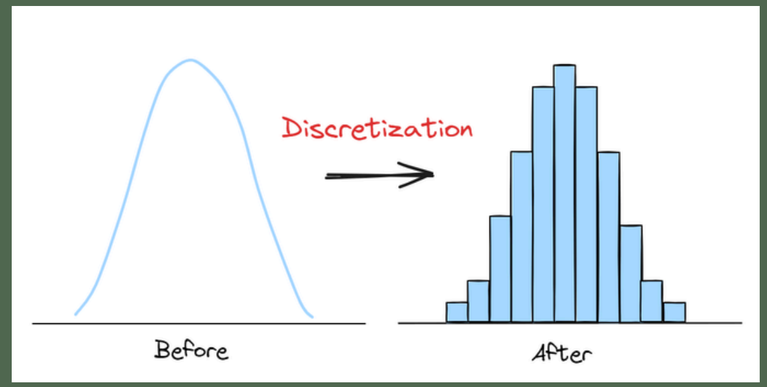
KBinsDiscretizer (n_bins=3, 5, 7, 10)

Strategia:

Quantile (równe przedziały liczebności)

Optymalizacja:

Liczba przedziałów wybrana na podstawie silhouette score.



Analiza SVD

Analiza SVD

Cel:
Oszacowanie optymalnej liczby klastrów.

Metoda:
TruncatedSVD (do 10 komponentów)

Działanie:
- Redukcja wymiarowości danych.
- Sugerowanie potencjalnej liczby klastrów.

Kryterium wyboru:
Liczba komponentów wyjaśniających ≥90% wariancji.

Wynik analizy:
- Wykres wartości singularnych.
- Poszukiwanie punktu załamania (ang. elbow), w którym wartości gwałtownie spadają.
- Pomaga to oszacować odpowiednią liczbę klastrów (k).

Dodatkowe uwagi:
- Im więcej wariancji wyjaśniają komponenty, tym lepiej zachowana struktura danych.
- Punkt załamania wskazuje, gdzie dodatkowe komponenty przestają wnosić znaczącą informację.

t-SNE

Redukcja wymiarowości za pomocą t-SNE

Jak działa t-SNE:

- Analizuje podobieństwa między punktami w danych wielowymiarowych — dla każdego punktu szacuje prawdopodobieństwo bycia "sąsiadem" innych punktów.
- Stara się odwzorować te relacje w przestrzeni 2D lub 3D, tak aby punkty, które były blisko w oryginalnych danych, pozostały blisko również po redukcji.
- Optymalizuje układ punktów poprzez minimalizację różnicy między rozkładem odległości w danych wysokowymiarowych a niskowymiarowych.
- Proces odbywa się iteracyjnie, z użyciem metod numerycznych (np. gradient descent).

Kluczowe parametry:

- Perplexity – liczba sąsiadów branych pod uwagę (kontroluje lokalność analizy).
- Learning rate – szybkość dostosowywania układu punktów.
- Liczba iteracji – liczba kroków optymalizacji prowadzących do ustabilizowania układu.

Wyniki

Ocena jakości

NMI (Normalized Mutual Information)

- Mierzy, jak dobrze klastry odpowiadają znanym grupom (np. kontynentom, systemom politycznym).
- Skala: 0 = brak zgodności, 1 = pełna zgodność.

Odległość wewnątrz klastra (Within-Cluster, WC Distance)

- Średnia odległość punktów od środka klastra.
- Im niższa wartość, tym punkty są bliżej siebie — zwarte klastry.

Odległość między klastrami (Between-Cluster, BC Distance)

- Odległość między środkami różnych klastrów.
- Im wyższa wartość, tym klastry są bardziej oddzielone.

Wskaźnik sylwetki (Silhouette Score)

- Zakres od -1 do 1.
- Wyższa wartość oznacza, że punkty dobrze pasują do swojego klastra i są daleko od innych – dobre klastry.

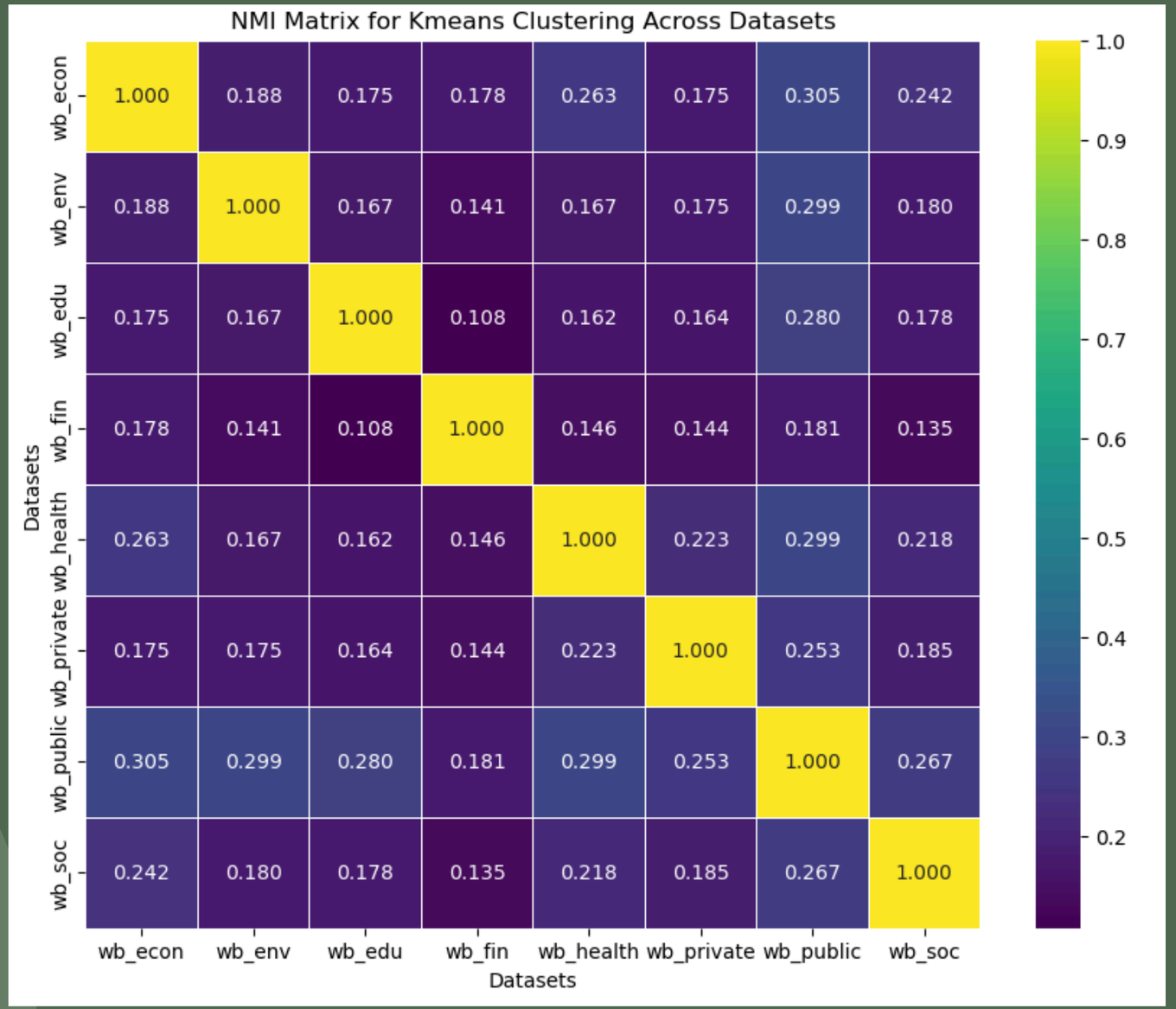
Ocena jakości

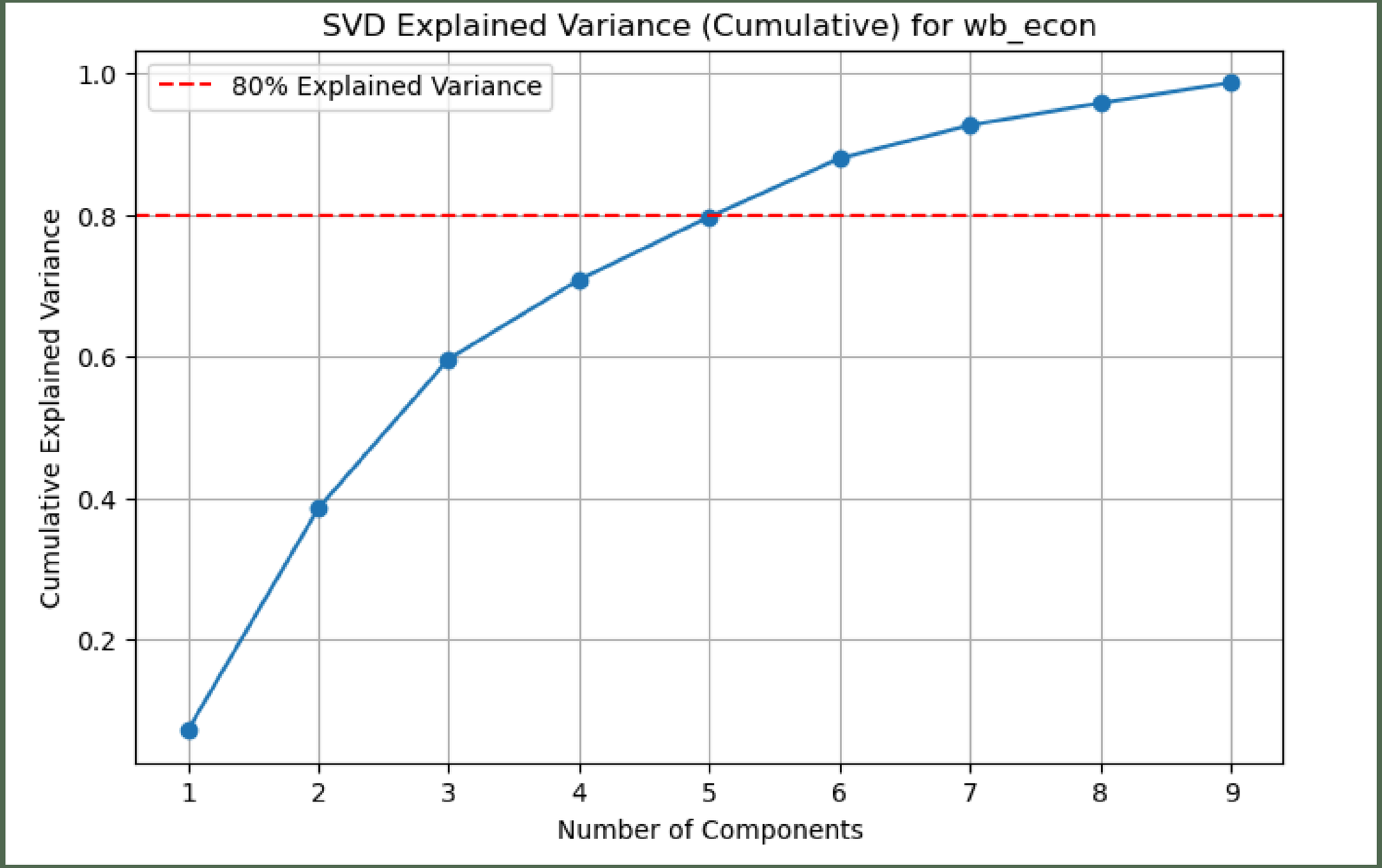
Clustering Evaluation Metrics

	Dataset	Method	Silhouette Score	WC Distance	BC Distance	Noise Ratio
0	wb_econ	kmeans	0.4515	2.9829	14.9581	nan
1	wb_env	kmeans	0.4318	3.0283	13.1011	nan
2	wb_edu	kmeans	0.4990	2.7731	13.5684	nan
3	wb_fin	kmeans	0.4259	3.7664	15.8665	nan
4	wb_health	kmeans	0.4423	3.0321	13.5871	nan
5	wb_private	kmeans	0.4375	3.5069	15.8773	nan
6	wb_public	kmeans	0.5795	2.9679	13.1595	nan
7	wb_soc	kmeans	0.5019	5.1836	19.3842	nan
8	wb_econ	hierarchical	0.4232	3.0859	14.8449	nan
9	wb_env	hierarchical	0.3659	3.4120	12.6976	nan
10	wb_edu	hierarchical	0.4604	2.9488	13.3220	nan
11	wb_fin	hierarchical	0.4167	3.8103	16.0483	nan
12	wb_health	hierarchical	0.4298	3.0793	13.7001	nan
13	wb_private	hierarchical	0.4076	3.6331	16.1503	nan
14	wb_public	hierarchical	0.5725	3.0079	13.0394	nan
15	wb_soc	hierarchical	0.4743	5.3993	18.8798	nan
16	wb_econ	dbscan	0.2810	2.9452	14.4606	0.1075
17	wb_env	dbscan	0.1909	3.4805	11.6778	0.0794
18	wb_edu	dbscan	0.3799	1.6132	12.1837	0.0888
19	wb_fin	dbscan	0.1976	3.2939	15.1592	0.0748
20	wb_health	dbscan	0.3385	1.8647	11.9149	0.0794
21	wb_private	dbscan	0.3131	3.4737	14.6742	0.1028
22	wb_public	dbscan	0.3574	1.8953	10.1262	0.1075
23	wb_soc	dbscan	0.3945	15.0625	76.0878	0.0561

Grupowanie metodą K-średnich

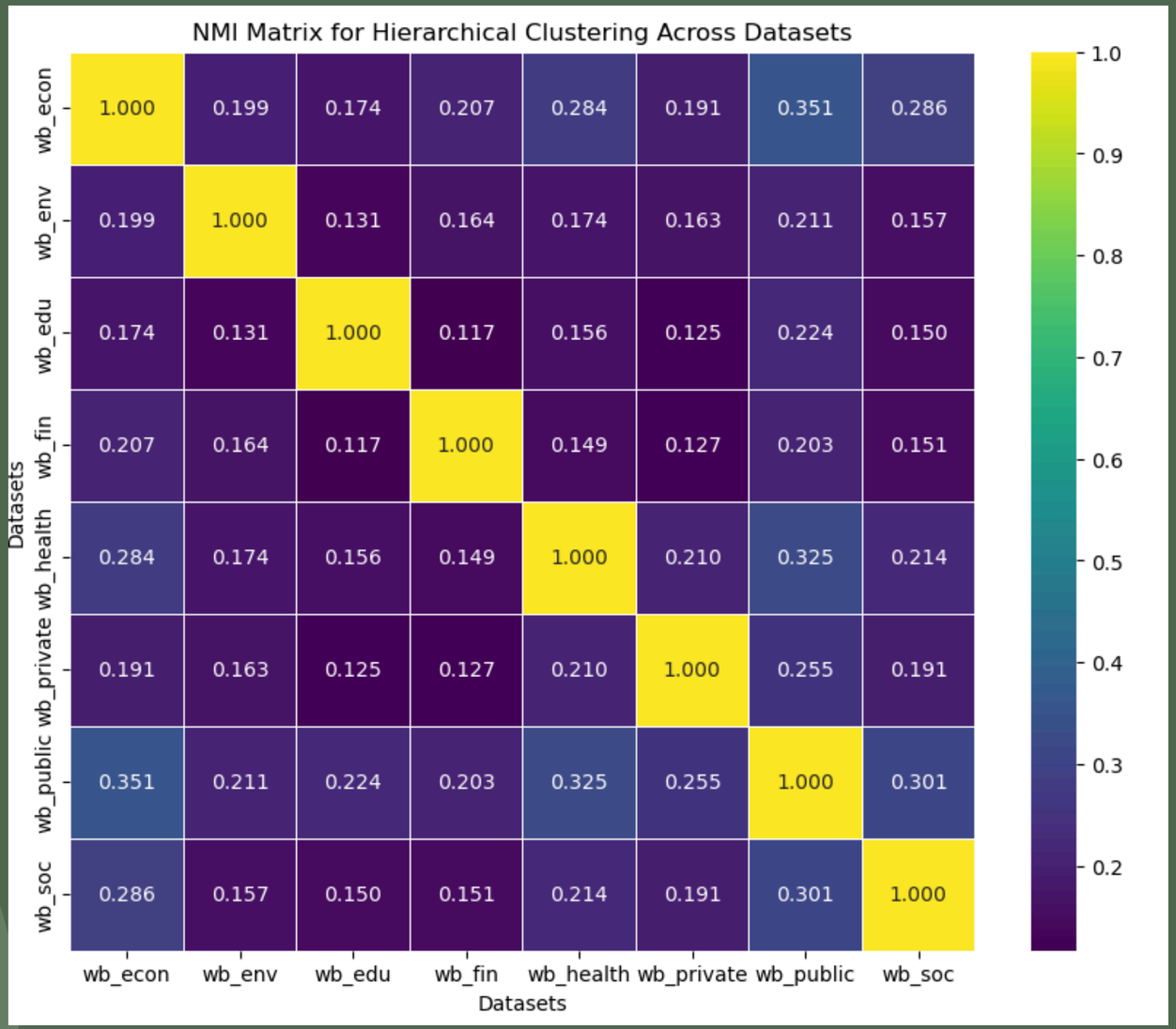
K-średnich


NMI Matrix for Kmeans Clustering Across Datasets
wb_econ 1.000 0.188 0.175 0.178 0.263 0.175 0.305 0.242
wb_env 0.188 1.000 0.167 0.141 0.167 0.175 0.299 0.180
wb_edu 0.175 0.167 1.000 0.108 0.162 0.164 0.280 0.178
wb_fin 0.178 0.141 0.108 1.000 0.146 0.144 0.181 0.135
wb_health 0.263 0.167 0.162 0.146 1.000 0.223 0.299 0.218
wb_private 0.175 0.175 0.164 0.144 0.223 1.000 0.253 0.185
wb_public 0.305 0.299 0.280 0.181 0.299 0.253 1.000 0.267
wb_soc 0.242 0.180 0.178 0.135 0.218 0.185 0.267 1.000
Datasets
wb_econ wb_env wb_edu wb_fin wb_health wb_private wb_public wb_soc
Datasets
1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2


SVD Explained Variance (Cumulative) for wb_econ
80% Explained Variance
Cumulative Explained Variance
1.0
0.8
0.6
0.4
0.2
1
2
3
4
5
6
7
8
9
Number of Components

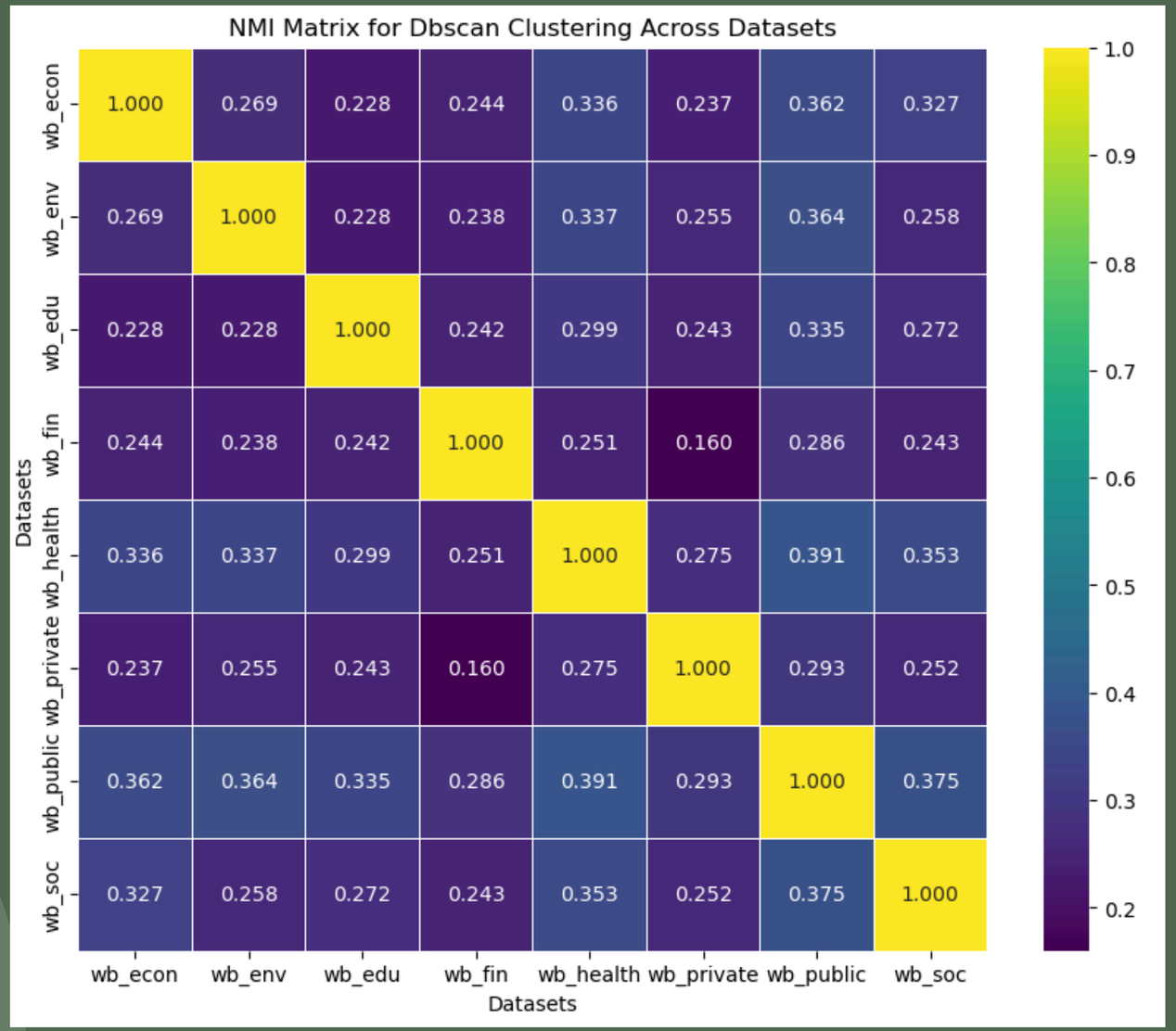
Grupowanie hierarchiczne

Grupowanie hierarchiczne



DBSCAN

DBSCAN


NMI Matrix for Dbscan Clustering Across Datasets
wb_econ 1.000 0.269 0.228 0.244 0.336 0.237 0.362 0.327
wb_env 0.269 1.000 0.228 0.238 0.337 0.255 0.364 0.258
wb_edu 0.228 0.228 1.000 0.242 0.299 0.243 0.335 0.272
wb_fin 0.244 0.238 0.242 1.000 0.251 0.160 0.286 0.243
wb_health 0.336 0.337 0.299 0.251 1.000 0.275 0.391 0.353
wb_private 0.237 0.255 0.243 0.160 0.275 1.000 0.293 0.252
wb_public 0.362 0.364 0.335 0.286 0.391 0.293 1.000 0.375
wb_soc 0.327 0.258 0.272 0.243 0.353 0.252 0.375 1.000
Datasets
wb_econ wb_env wb_edu wb_fin wb_health wb_private wb_public wb_soc
Datasets
1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2

Wnioski

Czynniki ekonomiczne

K-średnich:

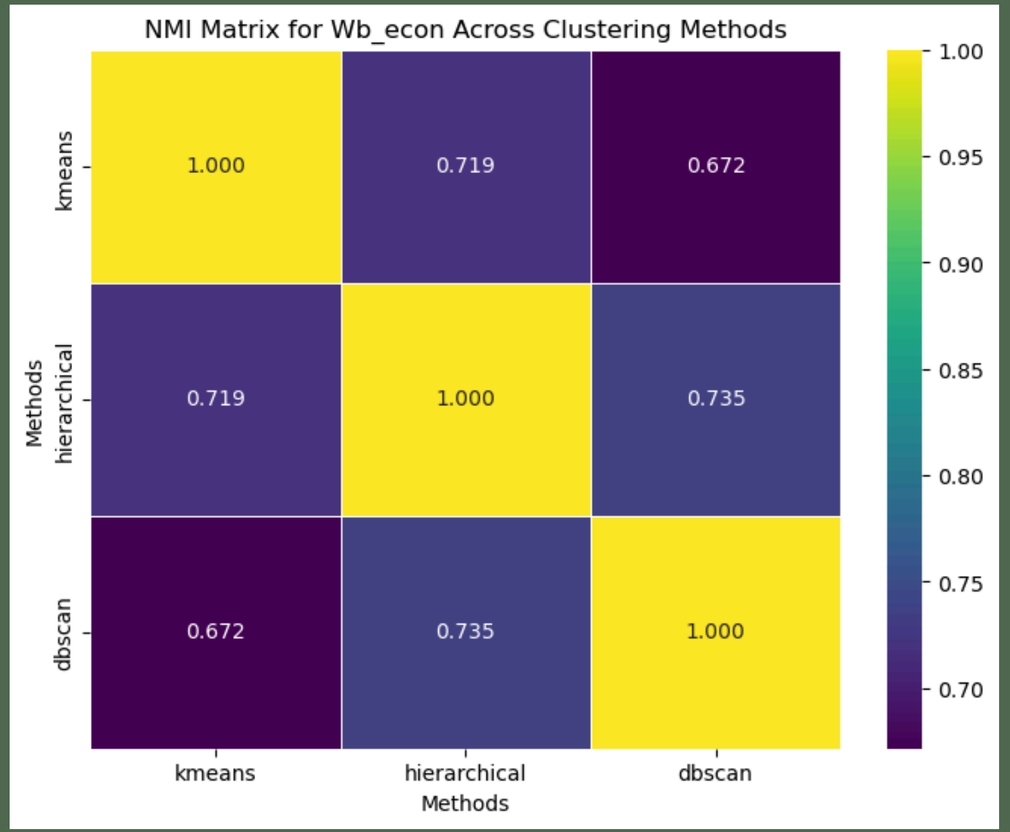
- Grupuje kraje bogate, średnio bogate oraz wyspy w osobnych klastrach.
- Mniej wyraźne podziały między grupami.

Hierarchiczne:

- Tworzy bardziej zróżnicowane klastry pod względem zamożności.
- Mniej wysp w jednym klastrze, lepsze rozróżnienie.

DBSCAN:

- Zwiększa liczbę klastrów, ale zaciera podziały.
- Nietypowe połączenia, np. Chorwacja + Rwanda,.



Czynniki środowiskowe

K-średnich:

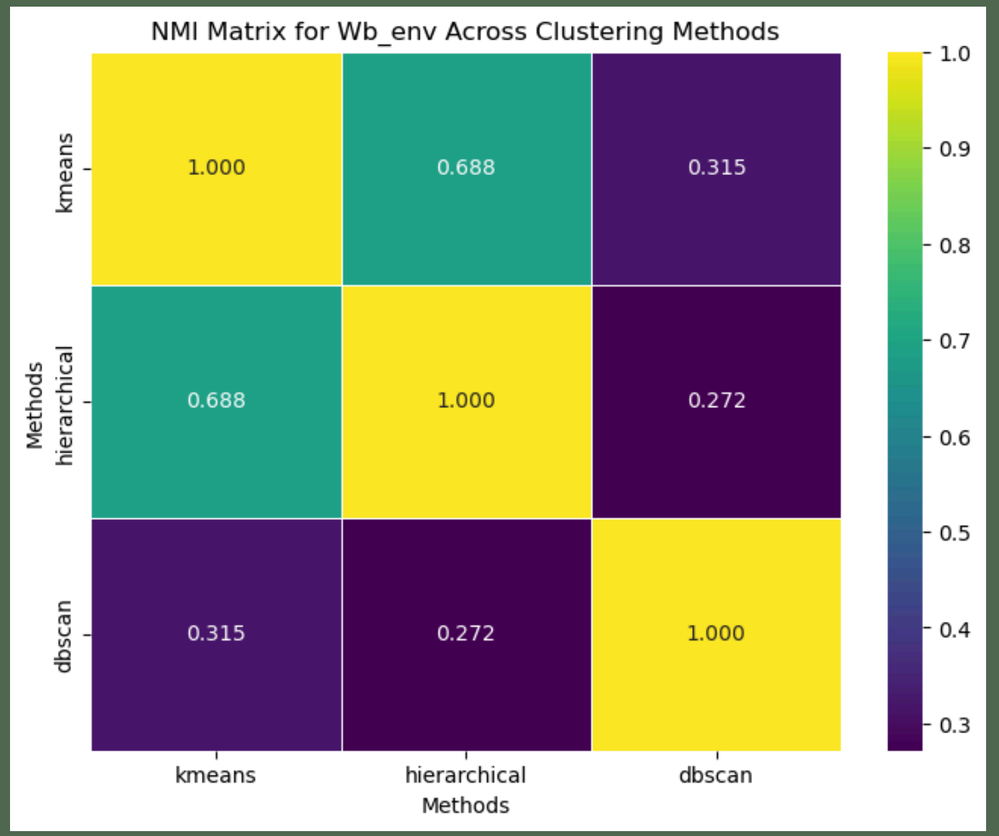
- Widoczny podział na państwa wyspiarskie.
- Polska zgrupowana razem z Kanadą, Australią i Rosją.

Hierarchiczne:

- Mniej równomierne klastry niż w metodzie K-średnich.
- Wdzielenie krajów rozwiniętych przemysłowo

DBSCAN:

- Najwięcej klastrów, w tym 2 duże (np. Polska + Curaçao, Indie + Wyspy Owcze).
- Podziały mniej intuicyjne.



Czynniki edukacyjne

K-średnich:

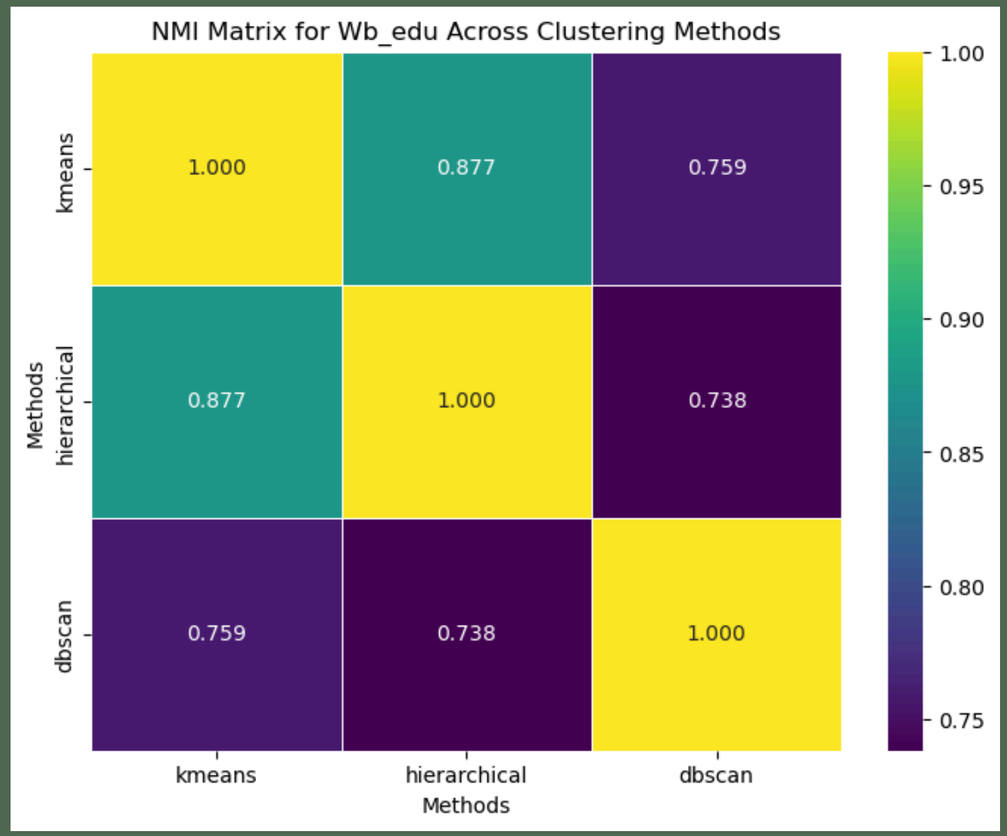
- 7 klastrów, łączy kraje o podobnym rozwoju (np. Bangladesz + Pakistan).
- Mniej oczywiste połączenia, np. Niemcy + Samoa. Nigeria + Monaco.

Hierarchiczne:

- Również 7 klastrów, ale z wyraźnym podziałem na wysoki i niski poziom edukacji.

DBSCAN:

- Aż 19 klastrów, rozbija kraje na małe grupy.
- Jeden duży klaster, np. USA + Afganistan.



Czynniki finansowe

K-średnich:

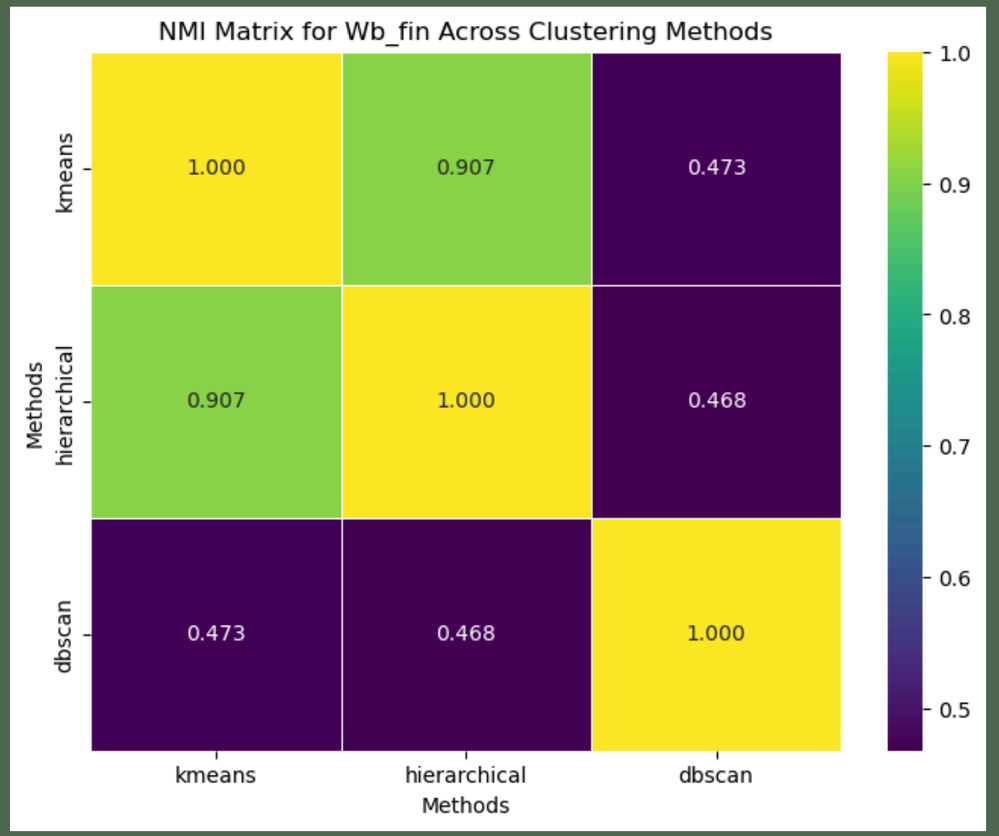
- 6 klastrów, z wyraźniejszymi granicami między nimi.
- Połączenie Arabii Saudyjskiej z Somalią

Hierarchiczne:

- 6 bardziej jednolitych klastrów (np. Iran + Sudan, Luksemburg + Monako).
- Bardzo duże podobieństwo do K-średnich

DBSCAN:

- 11 klastrów, w tym jeden dominujący (np. USA + Norwegia).
- Pozostałe klastry od kilku do kilkudziesięciu krajów.



Czynniki zdrowotne

K-średnich:

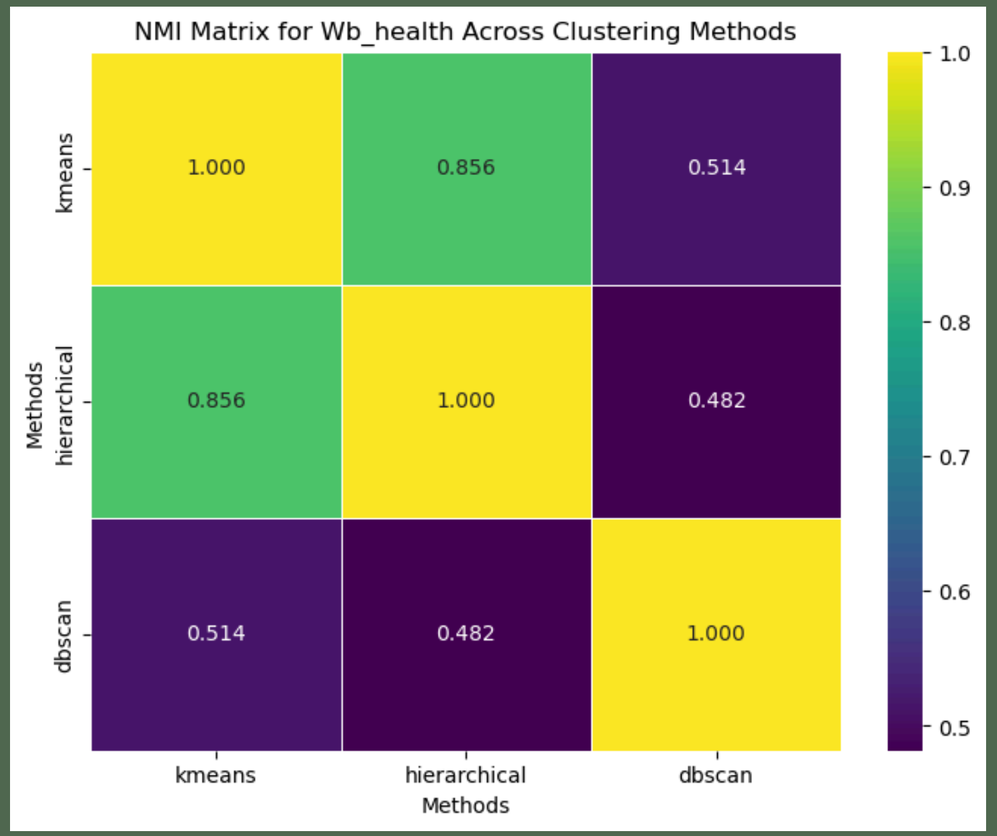
- Klastry o podobnej liczebności.
- Podziały takie jak Korea Płn. + Afganistan, Wielka Brytania + USA.

Hierarchiczne:

- Wyraźny rozdział na kraje wyspiarskie, średnio rozwinięte jak i dobrze rozwinięte.
- Duża podobieństwo do K-średnich.

DBSCAN:

- 12 klastrów, sensownie pogrupowanych.
- Dobra separacja między grupami.



Czynniki sektora prywatnego

K-średnich:

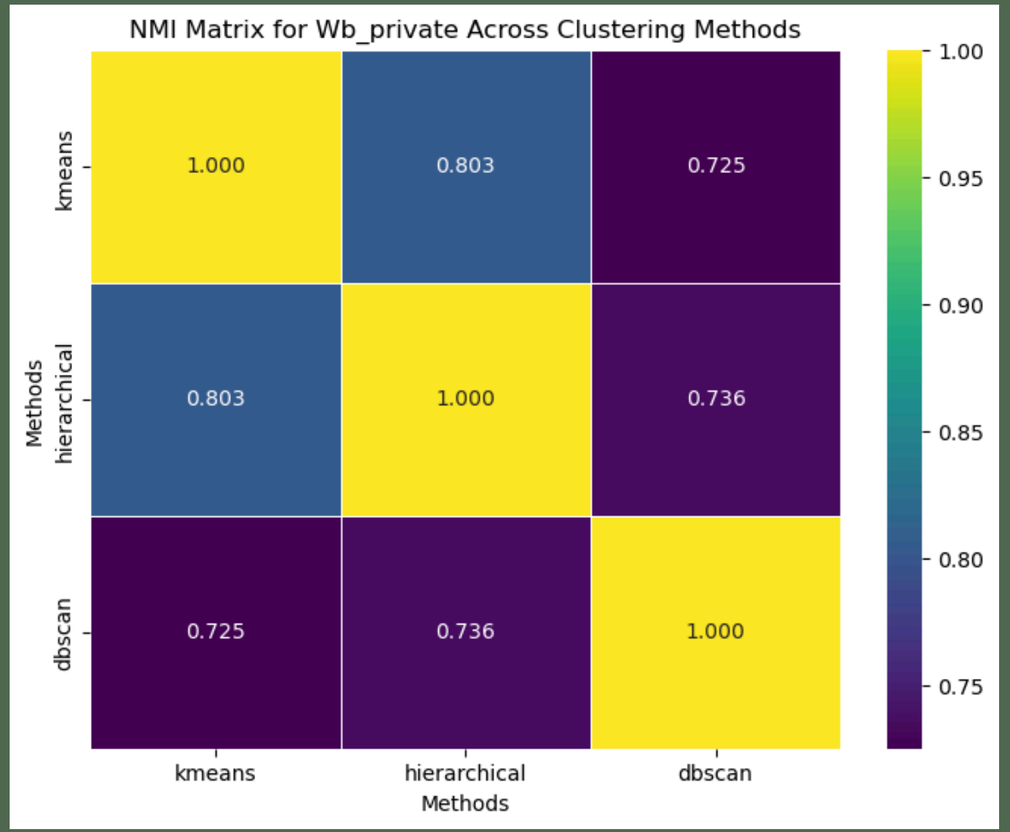
- Podobne klastry do metody Hierarchicznej, ale mniej wycentrowane.

Hierarchiczne:

- Rozbija duży klaster DBSCAN na mniejsze, bardziej jednolite grupy.
- Klastry bardziej zrównoważone liczebnie.

DBSCAN:

- Jeden dominujący klaster (np. Syria, Indie).
- 5 średnich klastrów z niepewnymi przypadkami.



Czynniki sektora publicznego

K-średnich:

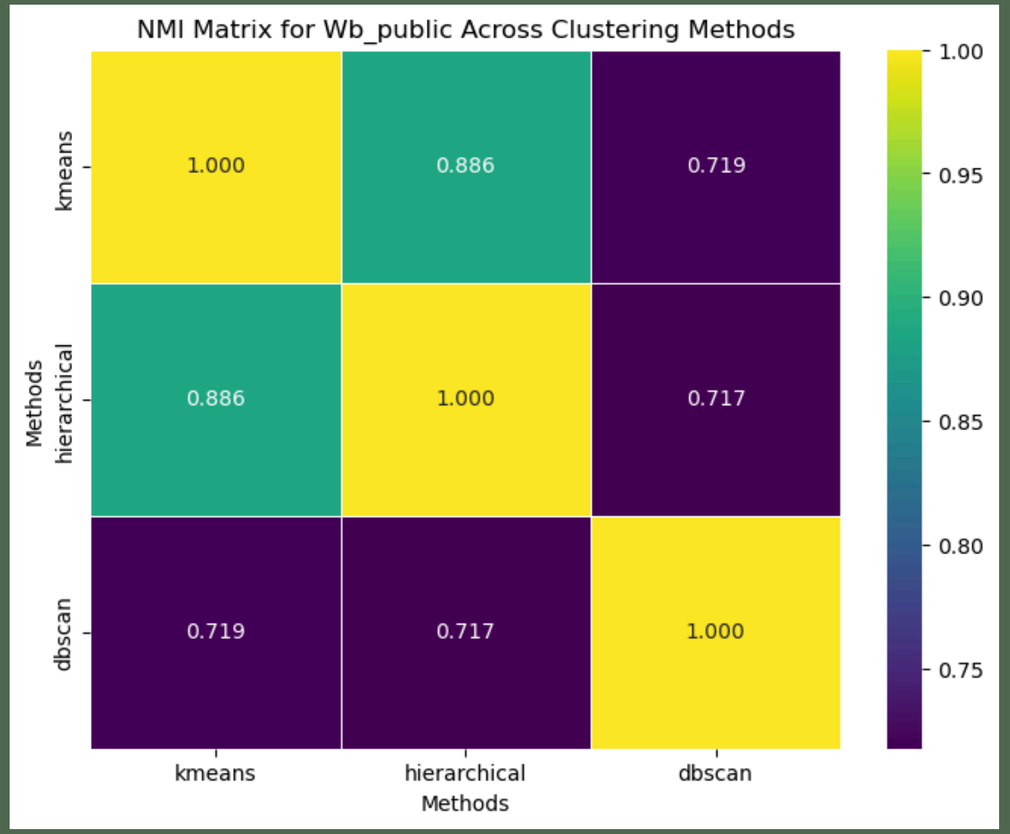
- 4 dobrze oddzielone klastry, dobrze skoncentrowane.
- Bliskość krajów wyspiarskich, krajów wysoko rozwiniętych, średnio rozwiniętych i słabo rozwiniętych w osobnych klastrach.

Hierarchiczne:

- Zachowuje podział i bliskość K-średnich.
- Podział pod względem poziomu rozwoju.

DBSCAN:

- 10 klastrów, bardzo nierównomiernych.
- Trzy dominujące klastry (kilkadziesiąt krajów) i mniejsze grupki.



Czynniki społeczne

K-średnich:

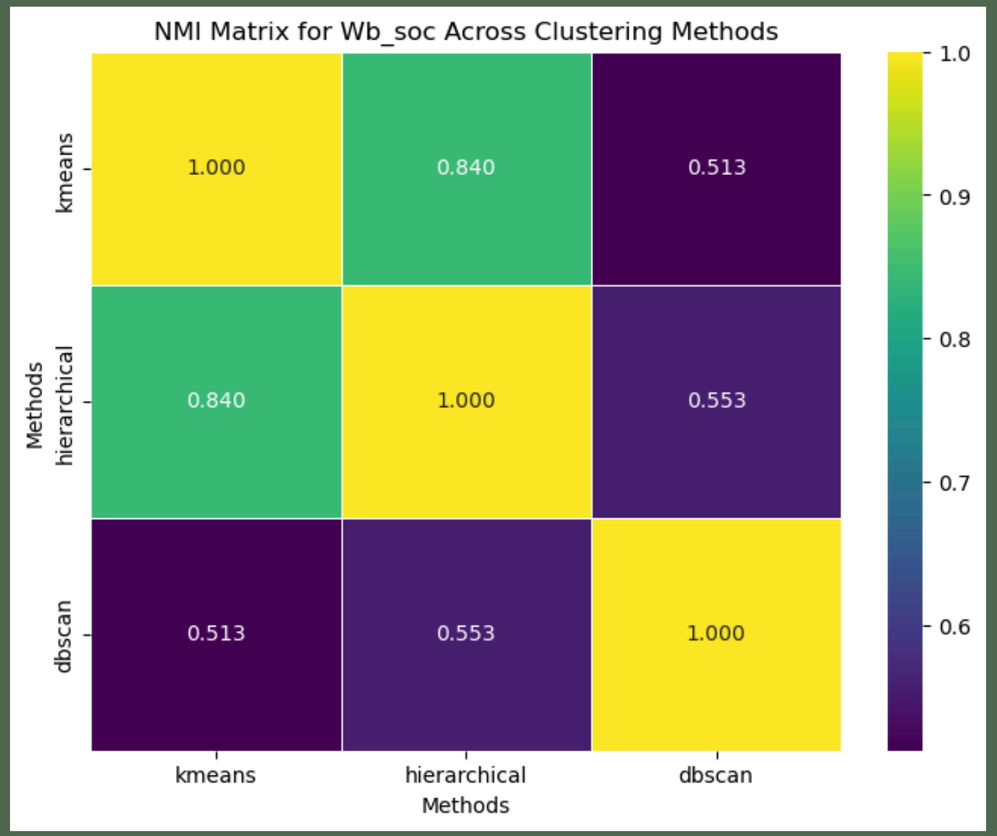
- Klastry podobne do metody Hierarchicznej, ale bardziej zrównoważone liczebnie.
- Bardzo rozproszone klastry, wyraźnie zgrupowane kraje zamożne.

Hierarchiczne:

- 4 rozproszone klastry, z dominującym klastrem krajów rozwijających się.
- Tuvalu w klastrze z Nauru i Włochami.

DBSCAN:

- 10 klastrów, w tym 2 dominujące (kraje bogate i biedne).
- Najmniejszy klaster: tylko Tuvalu.



Dziękujemy