Analiza skupień państw i narodów w kontekście gospodarczo-politycznym

Maria Nowicka[1] i Zuzanna Tabisz[1]

Politechnika Poznańska, Wydział Informatyki i Telekomunikacji, Informatyka

Streszczenie Niniejsza praca prezentuje kompleksową analizę skupień państw i narodów w kontekście gospodarczo-politycznym, opierając się na danych Banku Światowego z 2022 roku. Dane podzielono na osiem kluczowych kategorii tematycznych: ekonomia, środowisko, edukacja, finanse, zdrowie, sektor prywatny, sektor publiczny oraz czynniki społeczne, co pozwoliło na wielowymiarowe spojrzenie na zróżnicowanie krajów. W analizie wykorzystano trzy metody klastrowania: algorytm K-średnich, grupowanie hierarchiczne oraz DBSCAN, które oceniono za pomocą szeregu metryk ewaluacyjnych, takich jak współczynnik sylwetki, odległości wewnątrz- i międzyklastrowe, stosunek punktów szumu (dla DBSCAN) oraz znormalizowana informacja wzajemna (NMI). Wyniki badania wskazują, że metoda K-średnich charakteryzuje się najwyższą spójnością i separacją klastrów, szczególnie w przypadku danych dotyczących edukacji i sektora publicznego, gdzie uzyskano wyraźne i dobrze zdefiniowane grupy krajów. Z kolei algorytm DBSCAN, dzięki swojej zdolności do wykrywania punktów odstających, okazał się skuteczny w identyfikacji nietypowych struktur danych, jednak generowane przez niego klastry były często mniej intuicyjne i trudniejsze do interpretacji w porównaniu z pozostałymi metodami.

1 Wprowadzenie

Analiza skupień państw w kontekście gospodarczo-politycznym ma na celu identyfikację grup krajów o podobnych cechach w wymiarach: ekonomicznym, edukacyjnym, zdrowotnym, geograficznym oraz społecznym. Ideą przewodnią pracy jest pogłębienie zrozumienia różnic i podobieństw między krajami w wybranych wymiarach rozwojowych oraz zidentyfikowanie naturalnych grup państw na podstawie ich charakterystyk.

W badaniu wykorzystano dane pochodzące z Banku Światowego[1]. Zależności i wzorce identyfikowano przy zastosowaniu metod analizy skupień, w tym metody K-średnich (*K-means*), DBSCAN oraz analizy hierarchicznej. Proces analizy obejmował wstępne oczyszczanie danych, eliminację brakujących wartości oraz standaryzację cech, co umożliwiło uzyskanie wiarygodnych wyników. Dodatkowo, zastosowano techniki redukcji wymiarowości, takie jak t-SNE oraz TruncatedSVD, w celu wizualizacji i interpretacji złożonych zależności między krajami. Optymalizacja parametrów, takich jak liczba skupień czy liczba przedziałów dyskretyzacji, została przeprowadzona z użyciem metryk, takich jak

współczynnik sylwetki (silhouette score), co pozwoliło na wybór najbardziej adekwatnych modeli.

Projekt ma na celu nie tylko identyfikację wzorców i zależności między państwami, ale także stworzenie interaktywnych wizualizacji, które ułatwiają eksplorację danych. Wyniki analizy mogą znaleźć zastosowanie w badaniach porównawczych, polityce gospodarczej oraz planowaniu strategicznym na poziomie międzynarodowym. Przeprowadzone analizy pozwalają na lepsze zrozumienie globalnych dynamik rozwojowych.

2 Opis zbiorów danych

W ramach niniejszej analizy wykorzystano dane pochodzące z bazy *World Development Indicators (WDI)* udostępnianej przez Bank Światowy. Jest to obszerna kolekcja wskaźników rozwoju społeczno-ekonomicznego i środowiskowego, obejmująca dane zebrane z oficjalnych źródeł międzynarodowych. Baza ta zawiera około 1 600 wskaźników dla 217 gospodarek oraz ponad 40 grup krajowych, z danymi sięgającymi nawet 50 lat wstecz [1].

2.1 Źródło i charakterystyka danych

Dla celów analizy wybrano dane na rok 2022, aby zapewnić aktualność informacji. Wybór ten podyktowany był również dostępnością pełnych zestawów danych dla tego okresu, gdyż dla lat późniejszych występowały większe braki w danych.

Dane zostały następnie pogrupowane według głównych kategorii tematycznych, zgodnie z filtrami dostępnymi w bazie WDI:

- 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)

Każda z tych kategorii stanowi odrębny zbiór danych, który będzie analizowany indywidualnie w dalszej części pracy.

2.2 Rozmiar i struktura danych

Dla każdego z wyżej wymienionych obszarów tematycznych pobrano dane obejmujące:

- Liczbę krajów: Zgodnie z danymi dostępnymi w bazie WDI, dla roku 2022 dostępne są dane dla 214 gospodarek.

– Liczbę wskaźników: Liczba wskaźników różni się w zależności od kategorii tematycznej. Oryginalne dane zawierają cztery kolumny kategoryczne: rok, kod określający czas, pełna nazwa gospodarki oraz jej skrótowy kod (Time, Time Code, Country Name, Country Code). Pozostałe kolumny zawierają wartości wskaźników i mają charakter numeryczny. Ich liczba waha się od 59 do 361, w zależności od tematyki.

Każdy z tych zbiorów danych zawiera informacje dla poszczególnych krajów, umożliwiając porównania międzynarodowe w obrębie wybranych kategorii tematycznych.

2.3 Wstępne przetwarzanie danych

Wstępne przetwarzanie danych miało na celu przygotowanie zbiorów danych do analizy skupień poprzez oczyszczenie, normalizację i wybór najbardziej istotnych cech. Proces rozpoczął się od wczytania danych z plików CSV, zawierających informacje dla 214 krajów w ośmiu kategoriach tematycznych: ekonomicznej, środowiskowej, edukacyjnej, finansowej, zdrowotnej, sektora prywatnego, sektora publicznego oraz społecznej. Dane zostały zmapowane na kody krajów ISO 3166-1 alpha-2 przy użyciu pliku indeksu, co zapewniło spójność identyfikatorów między zbiorami.

Każdy zbiór danych został oczyszczony z kolumn zawierających ponad 50% brakujących wartości, co znacząco zmniejszyło liczbę cech, np. z 361 do 318 w przypadku danych ekonomicznych (wb_econ) czy z 160 do 117 dla danych edukacyjnych (wb_edu). Brakujące wartości w kolumnach numerycznych uzupełniono medianą, eliminując braki danych (0% brakujących wartości po imputacji). Dla cech o wartościach przekraczających 1000 zastosowano transformację logarytmiczną, aby ograniczyć wpływ skrajnych wartości, co było szczególnie istotne w zbiorach finansowych i ekonomicznych o wysokiej wariancji (np. wariancja przed skalowaniem dla cechy „Net primary income (current LCU)" w wb_econ wynosiła 2,47e+24).

Dane poddano dyskretyzacji za pomocą metody `KBinsDiscretizer`, stosując strategię kwantylową i kodowanie porządkowe [11]. Liczba przedziałów (n_bins) została wybrana eksperymentalnie przez przetestowanie wartości 3, 5, 7 i 10, a optymalna wartość była określana na podstawie maksymalnego wyniku `Silhouette Score` uzyskanego podczas grupowania [8].

Następnie przeprowadzono skalowanie `MinMaxScaler`, normalizując wartości do przedziału [0,1] [12]. Usunięto kolumny o niskiej wariancji, z mniej niż dwoma unikalnymi wartościami lub wysoko skorelowane (korelacja powyżej 0,95), co pozwoliło zredukować redundancję. Dla każdego zbioru danych wybrano 10 cech o najwyższej wariancji po skalowaniu, np. w danych ekonomicznych wybrano cechy takie jak „Agriculture, forestry, and fishing, value added (% of GDP)" czy „GDP per capita (current US$)".

Proces ten zapewnił, że dane są kompletne, znormalizowane i gotowe do dalszej analizy skupień, minimalizując wpływ szumu i nadmiarowych informacji.

Po wstępnym przetwarzaniu i analizie danych dokonano selekcji najistotniejszych cech dla każdego zbioru danych. Poniżej przedstawiono wybrane cechy wraz z ich krótkim opisem.

Polityka gospodarcza i zadłużenie (Economic Policy and Debt)

Wybrane cechy obejmują roczny wzrost PKB, PKB per capita w stałych i bieżących dolarach oraz transfery osobiste w dolarach jako procent PKB (remittances). Uwzględniono również PKB w parytecie siły nabywczej oraz dochód narodowy brutto per capita. Dodatkowo, cechy zawierają wskaźnik poziomu cen w relacji do kursu wymiany oraz saldo dochodów netto z zagranicy.

Środowisko (Environment)

W tym zbiorze wybrano wskaźniki dotyczące powierzchni chronionych terenów lądowych, udziału populacji wiejskiej i jej liczebności oraz rocznego wzrostu populacji miejskiej. Uwzględniono również dane o powierzchni lasów, gęstości zaludnienia oraz areału gruntów rolnych. Znalazły się tu też całkowite emisje dwutlenku węgla oraz emisje per capita z wyłączeniem LULUCF.

Edukacja (Education)

Cechy obejmują liczbę uczniów w edukacji podstawowej oraz wskaźniki zapisu do szkół podstawowych i średnich. Uwzględniono długość obowiązkowej i przedszkolnej edukacji oraz procent nauczycielek wśród kadry podstawowej. Dodatkowo analizowano wydatki rządowe na edukację jako procent PKB oraz odsetek dzieci w wieku szkolnym, które nie uczęszczają do szkoły.

Sektor finansowy (Financial Sector)

Wybrane wskaźniki to deflator PKB oraz wskaźniki inflacji – zarówno deflatora, jak i cen konsumenckich. Zawiera także indeks cen konsumpcyjnych i kursy wymiany walut lokalnych względem dolara. Ponadto, uwzględniono liczbę oddziałów banków komercyjnych na 100 tys. dorosłych, kredyty dla sektora publicznego i prywatnego oraz wskaźnik płynności banków.

Zdrowie (Health)

Wśród cech znalazły się wskaźniki szczepień przeciwko odrze i DPT dla dzieci w wieku 12–23 miesięcy oraz liczba zgonów matek i dzieci w wieku 5–9 lat. Uwzględniono także wskaźnik płodności wśród nastolatek oraz śmiertelność dorosłych mężczyzn. Dodatkowo analizowano udział wydatków zdrowotnych ponoszonych bezpośrednio przez pacjentów oraz odsetek populacji korzystającej z podstawowej sanitacji.

Sektor prywatny i handel (Private Sector and Trade)

W zbiorze tym wybrano dane dotyczące importu i eksportu towarów względem krajów o wysokich i średnich dochodach, z podziałem na regiony. Ujęto udział importu i eksportu w wartości całkowitej towarów oraz ich podział według krajów z regionów Azji Wschodniej i Pacyfiku oraz innych regionów.

Sektor publiczny (Public Sector)

Cechy dotyczą jakości rządzenia, w tym wskaźników takich jak wolność wypowiedzi, jakość regulacji, stabilność polityczna, praworządność oraz kontrola korupcji. Uwzględniono także wyniki wskaźników wydajności statystycznej na różnych poziomach oraz liczbę źródeł danych.

Ochrona socjalna i rynek pracy (Social Protection and Labor)

Zawiera wskaźniki zatrudnienia i bezrobocia, z podziałem na płeć i grupy wiekowe, a także udział kobiet w sile roboczej i relację ich aktywności do mężczyzn. Dodatkowo ujęto dane o uchodźcach, osobach samozatrudnionych oraz odsetek młodzieży nieuczestniczącej w edukacji ani zatrudnieniu, rozdzielone według płci.

3 Opis zaproponowanych eksperymentów

W przeprowadzonych eksperymentach zaproponowano zestaw metod grupowania danych, które umożliwiają analizę i łączenie krajów na podstawie różnorodnych wskaźników. Głównym celem było opracowanie podejścia do wstępnego przetwarzania danych, redukcji wymiarowości oraz przeprowadzenia skutecznej analizy skupień, z metodą bazową opartą na algorytmie K-średnich, którą starano się ulepszyć poprzez optymalizację parametrów i wizualizację wyników.

Dane wejściowe, pochodzące z różnych zbiorów tematycznych (np. ekonomicznych, zdrowotnych i edukacyjnych), zostały poddane wstępnemu czyszczeniu według metodologii przedstawionej w poprzednim rozdziale. Metoda bazowa K-średnich była stosowana na danych zredukowanych do dwóch wymiarów za pomocą t-SNE[5], co pozwalało na wizualizację klastrów. Pozostałe metody jakich użyto, to grupowanie hierarchiczne[4] oraz DBSCAN[3]. Metoda K-średnich[2] została wybrana jako metoda bazowa ze względu na jej prostotę, efektywność obliczeniową oraz szerokie zastosowanie w analizie danych.

Metoda K-Means opiera się na iteracyjnym przypisywaniu punktów do najbliższych centroidów i aktualizacji ich pozycji. Liczba klastrów (k) została wybrana na podstawie analizy wariancji wyjaśnionej przez metodę TruncatedSVD[6], gdzie liczba klastrów odpowiada liczbie komponentów objaśniających co najmniej 90% wariancji danych. Grupowanie hierarchiczne wykorzystuje metodę Warda z metryką euklidesową, tworząc hierarchiczną strukturę klastrów, która jest wizualizowana w postaci dendrogramu[4]. DBSCAN, w odróżnieniu od poprzednich metod, nie wymaga określenia liczby klastrów z góry, lecz automatycznie dostosowuje parametr eps na podstawie percentyla odległości do najbliższych sąsiadów[7], co pozwala na wykrywanie punktów odstających jako szum.

Ewaluacja eksperymentów opierała się na trzech miarach: współczynniku sylwetki[8], który ocenia spójność i separację klastrów, odległości wewnątrzklastrowej, wskazującej na zwartość klastrów, oraz odległości międzyklastrowej, mierzącej rozdzielność grup. Wyniki grupowania dla różnych zbiorów danych porównywano za pomocą znormalizowanej informacji wzajemnej (NMI)[9], co pozwoliło ocenić zgodność podziałów. Wizual, w tym interaktywne wykresy klastrów[10], oraz wykresy wariancji wyjaśnionej, wspierały interpretację wyników, umożliwiając identyfikację optymalnych parametrów i cech charakterystycznych dla każdego zbioru danych. Całość procesu była realizowana w sposób iteracyjny, z naciskiem na poprawę metody bazowej poprzez dostosowanie liczby przedziałów i klastrów oraz wybór istotnych cech.

4 Wyniki i dyskusja

Zgodnie z założeniami przedstawionymi we wstępie, celem projektu była identyfikacja naturalnych grup państw w oparciu o ich cechy gospodarcze, społeczne i polityczne. Analiza została przeprowadzona na danych pozyskanych z Banku Światowego, które po oczyszczeniu i standaryzacji zostały poddane analizie skupień przy użyciu trzech różnych metod: K-średnich (k-means), DBSCAN oraz grupowaniu hierarchicznemu.

Założenia projektu zostały w dużym stopniu zrealizowane: udało się wykryć sensowne skupienia państw, stworzyć przejrzyste wizualizacje oraz zidentyfikować kluczowe cechy różnicujące grupy krajów. Analizy potwierdziły, że złożone relacje między zmiennymi ekonomicznymi i społecznymi mogą być skutecznie odkrywane przy użyciu klasycznych metod eksploracji danych. Poniżej przedstawiona jest analiza klastrów utworzonych po zastosowaniu poszczeglnych metod.

4.1 Zbiór danych: Ekonomia

K-średnich

- **Klaster 0**: Kraje o zróżnicowanym poziomie rozwoju, często małe gospodarki wyspiarskie (np. Bahamy, Barbados, Cypr) oraz niektóre bogate w zasoby (np. Kuwejt, Wenezuela). Charakteryzują się umiarkowanym wzrostem PKB i wysokim PKB per capita, co sugeruje stabilność lub zależność od specyficznych sektorów (turystyka, ropa).
- **Klaster 1**: Kraje o niskim poziomie rozwoju (np. Afganistan, Burkina Faso, Haiti), z niskim PKB per capita i wysokim udziałem przekazów pieniężnych w PKB, wskazującym na zależność od wsparcia zewnętrznego.
- **Klaster 2**: Kraje o średnim poziomie rozwoju (np. Brazylia, Chiny, Polska), z umiarkowanym PKB per capita i zróżnicowanym wzrostem gospodarczym, często gospodarki wschodzące.
- **Klaster 3**: Wysoko rozwinięte gospodarki (np. USA, Niemcy, Japonia), z wysokim PKB per capita (PPP) i stabilnymi wskaźnikami.
- **Klaster 4**: Kraje rozwijające się (np. Indie, Bangladesz, Nigeria), z niższym PKB per capita i większym znaczeniem przekazów pieniężnych.

– **Klaster 5**: Kraje afrykańskie i niektóre azjatyckie (np. Angola, Etiopia, Laos), z niskim PKB per capita i wysoką zmiennością wzrostu.

Wnioski: Klastry odzwierciedlają globalne zróżnicowanie gospodarcze. Algorytm K-średnuch grupuje kraje boate, średni bogate oraz wyspy w osobnych klastrach. Wysoki udział przekazów pieniężnych w klastrach z krajami o nikim poziomie rozwoju i w klastrach z krajami rozwijającymi się wskazuje na migrację zarobkową. Klaster 3, zbierający wysoko rozwinięte gospodarki, jest najbardziej homogeniczny pod względem stabilności gospodarczej.

Grupowanie hierarchiczne

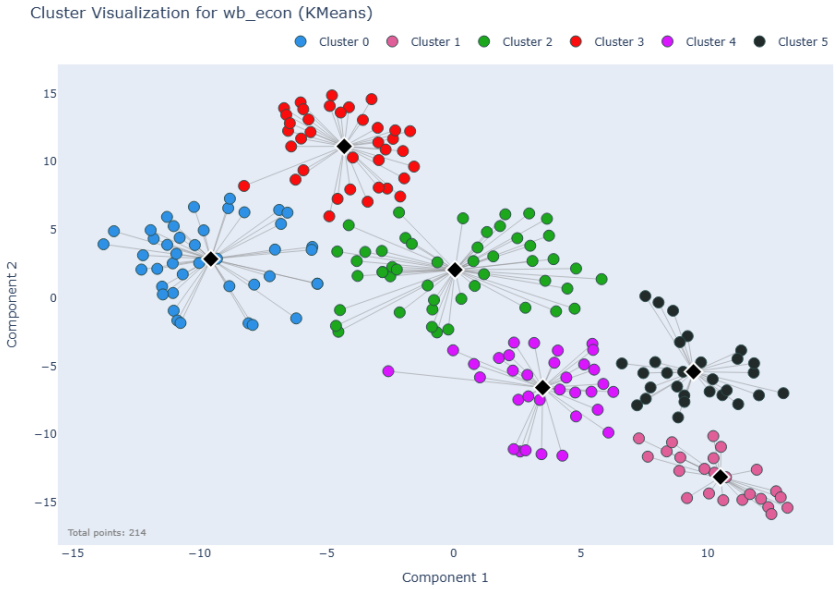
– **Klaster 0**: Kraje o najniższym poziomie rozwoju (np. Afganistan, Burundi, Haiti), z niskim PKB per capita i wysokim udziałem przekazów pieniężnych.
– **Klaster 1**: Kraje rozwijające się (np. Indie, Nigeria, Wietnam), z umiarkowanym PKB per capita i zmiennością wzrostu.
– **Klaster 2**: Kraje o średnim poziomie rozwoju (np. Brazylia, Chiny, Polska), z rozwiniętym sektorem przemysłowym/usługowym.
– **Klaster 3**: Kraje o zróżnicowanym profilu (np. Albania, Kuba, Wenezuela), często z niestabilnością gospodarczą.
– **Klaster 4**: Małe kraje wyspiarskie (np. Bahamy, Malediwy, Seszele), z wysokim PKB per capita opartym na turystyce/usługach.
– **Klaster 5**: Wysoko rozwinięte gospodarki (np. USA, Niemcy, Japonia), z wysokim PKB per capita i stabilnym wzrostem.

Wnioski: Hierarchiczne grupowanie lepiej odróżnia kraje o najniższym poziomie rozwoju od rozwijających się. Klaster 4, z małymi krajami wyspiarskimi, jest bardziej spójny, podkreślając unikalny profil małych gospodarek.

DBSCAN

– **Noise**: Kraje z unikalnymi profilami (np. Botswana, Brunei, Wenezuela), często z bogactwem zasobów lub niestabilnością.
– **Klaster 0**: Kraje rozwijające się (np. Indie, Nigeria, Wietnam), z niskim/umiarkowanym PKB per capita i wysokim udziałem przekazów pieniężnych.
– **Klaster 1**: Kraje o średnim poziomie rozwoju (np. Brazylia, Chiny, Polska), z rozwiniętym sektorem przemysłowym.
– **Klaster 2**: Małe kraje wyspiarskie i niektóre rozwinięte (np. Bahamy, Kuwejt, Portugalia), z wysokim PKB per capita.
– **Klaster 3**: Wysoko rozwinięte gospodarki (np. USA, Niemcy, Japonia), z wysokim PKB per capita i stabilnym wzrostem.
– **Klaster 4**: Kraje azjatyckie o niższym PKB per capita (np. Bangladesz, Nepal, Filipiny), z wysokim udziałem przekazów.
– **Klaster 5**: Kraje o specyficznym profilu (np. Kuba, Korea Północna, Monako), z nietypowymi strukturami gospodarczymi.
– **Klaster 6**: Kraje o najniższym poziomie rozwoju (np. Afganistan, Haiti, Togo), z bardzo niskim PKB per capita.

Wnioski: DBSCAN wyróżnia więcej klastrów, wskazując na heterogeniczność danych. Punkty szumu podkreślają unikalne cechy krajów, takich jak Brunei czy Wenezuela. Algorytm ten stworzył również nietypowe pod względem czynników ekonomicznych połączenia jak zgrupowanie Chorwacji oraz Rwandy w jednym klastrze.



Rysunek 1. Wizualizacja klastrów dla metody K-średnich - dane ekonomiczne

4.2 Zbiór danych: Środowisko

K-średnich

- **Klaster 0**: Kraje o dużej populacji i intensywnym rolnictwie (np. Indie, Chiny, Nigeria), z wysoką gęstością zaludnienia i emisjami CH4/N2O.
- **Klaster 1**: Małe kraje wyspiarskie i wysoko rozwinięte (np. Bermudy, Kuwejt, Singapur), z niską gęstością zaludnienia i ograniczonymi emisjami.
- **Klaster 2**: Kraje o zróżnicowanym profilu (np. Botswana, Islandia, Urugwaj), z większym udziałem terenów wiejskich.
- **Klaster 3**: Rozwinięte kraje europejskie i azjatyckie (np. Niemcy, Japonia, Francja), z umiarkowanymi emisjami i wysoką urbanizacją.
- **Klaster 4**: Małe gospodarki wyspiarskie (np. Malediwy, Seszele), z niskim wpływem na środowisko, ale wysoką wrażliwością na zmiany klimatyczne.

– **Klaster 5**: Kraje z zasobami naturalnymi (np. Brazylia, Rosja, Australia), z wysokimi emisjami CO2 i zróżnicowaną urbanizacją.

Wnioski: Klastry odzwierciedlają zróżnicowanie środowiskowe. Widozny jest podział na państwa wspiarskie. Kraje rozwinięte mają umiarkowane emisje dzięki technologiom i regulacjom (np. kraje europejskie). Polska zgrupowana została razem z Kanadą, Australią i Rosją, co może mieć związek z uwzględnianiem nie tylko emisji CO2 oraz CH4/N2O, ale również procentu terenu porkrytego lasami oraz wykorzystywanego rolniczo.

Grupowanie hierarchiczne

– **Klaster 0**: Małe kraje wyspiarskie i niektóre rozwinięte (np. Singapur, Kuwejt, Barbados), z niską gęstością i emisjami.
– **Klaster 1**: Kraje rozwinięte i o średnim poziomie rozwoju (np. Niemcy, Japonia, Serbia), z umiarkowanymi emisjami i wysoką urbanizacją.
– **Klaster 2**: Kraje o zróżnicowanym profilu (np. Botswana, Irak, Urugwaj), z emisjami związanymi z zasobami.
– **Klaster 3**: Kraje z zasobami naturalnymi (np. Brazylia, Rosja, Australia), z wysokimi emisjami CO2.
– **Klaster 4**: Kraje o dużej populacji i intensywnym rolnictwie (np. Indie, Nigeria, Bangladesz), z wysoką gęstością i emisjami CH4/N2O.
– **Klaster 5**: Kraje o średnim poziomie rozwoju (np. Chiny, Turcja, Indonezja), z wysokimi emisjami i dynamiczną urbanizacją.

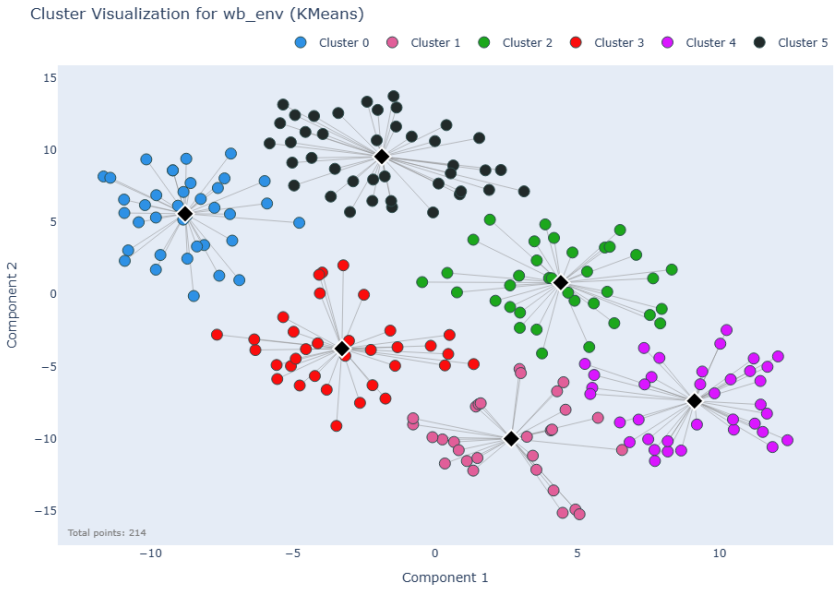
Wnioski: Hierarchiczne grupowanie lepiej oddziela kraje rolnicze od zasobowych. Klaster 0 podkreśla niską presję środowiskową małych krajów. Zarówno grupowanie k-średnich jak i hierarchiczne wyraźie wydzieliło kraje rozwinięte przemysłowo.

DBSCAN

– **Noise**: Kraje o nietypowych profilach (np. Indonezja, Islandia, Maroko), z unikalnymi kombinacjami emisji i gęstości.
– **Klaster 0**: Kraje rozwijające się (np. Indie, Nigeria, Bangladesz), z wysoką gęstością i intensywnym rolnictwem.
– **Klaster 1**: Kraje rozwinięte i o średnim poziomie rozwoju (np. Niemcy, Japonia, Turcja), z umiarkowanymi emisjami.
– **Klaster 2**: Kraje bogate w zasoby (np. Algieria, Iran, Kazachstan), z wysokimi emisjami CO2 per capita.
– **Klaster 3**: Małe kraje wyspiarskie (np. Dominika, Samoa), z minimalnymi emisjami.
– **Klaster 4**: Kraje o rozwiniętej infrastrukturze (np. Chorwacja, Seszele), z umiarkowanym wpływem na środowisko.
– **Klaster 5**: Kraje Bliskiego Wschodu (np. Bahrajn, Katar), z wysoką urbanizacją.
– **Klaster 6**: Małe kraje wyspiarskie (np. Barbados, Malediwy), z wysoką wrażliwością na zmiany klimatyczne.

– **Klaster 7**: Kraje z zasobami naturalnymi (np. Brazylia, Rosja, Kanada), z wysokimi emisjami CO2.
– **Klaster 8**: Terytoria zależne (np. Bermudy, Portoryko), z niskim wpływem środowiskowym.

Wnioski: DBSCAN generuje więcej klastrów, odzwierciedlając złożoność danych. Punkty szumu wskazują na nietypowe profile krajów. Zgrupowanie razem takich krajów jak Polska + Curaçao oraz Indie + Wyspy Owcze sugeruje, że to grupowanie może nie odzwierciedlać rzeczywistości.



Rysunek 2. Wizualizacja klastrów dla metody K-średnich - dane środowiskowe

4.3 Zbiór danych: Edukacja

K-średnich

– **Klaster 0**: Kraje o zróżnicowanym poziomie edukacji (np. Bahrain, Rosja, Serbia), z wysokimi zapisami, ale zmiennym dostępem.
– **Klaster 1**: Kraje o średnim poziomie edukacji (np. Chiny, Egipt, Senegal), z wyzwaniami w dostępie.
– **Klaster 2**: Wysoko rozwinięte kraje (np. Australia, Niemcy, Japonia), z wysokimi zapisami i wydatkami na edukację.

- **Klaster 3**: Kraje z problemami w dostępie (np. Afganistan, Somalia, Jemen), z wysokim odsetkiem dzieci poza szkołą.
- **Klaster 4**: Kraje afrykańskie i wyspiarskie (np. Angola, Haiti, Uganda), z niskim dostępem i wydatkami.
- **Klaster 5**: Kraje rozwijające się (np. Indie, Meksyk, Turcja), z umiarkowanym poziomem zapisów.
- **Klaster 6**: Kraje o niskim poziomie edukacji (np. Bangladesz, Etiopia, Pakistan), z krótkim okresem obowiązkowej edukacji.

Wnioski: Klastry odzwierciedlają globalne nierówności w edukacji. K-średnich łączy kraje o podobnym poziomie rozwoju. Kraje rozwinięte inwestują najwięcej, kraje rozwijające się potrzebują większych inwestycji.

Grupowanie hierarchiczne

- **Klaster 0**: Kraje rozwinięte (np. Australia, Japonia, Niemcy), z wysokimi zapisami i wydatkami na edukację.
- **Klaster 1**: Małe kraje (np. Bahrajn, Malta, Seszele), z wysokim poziomem edukacji, ale ograniczonymi danymi.
- **Klaster 2**: Kraje rozwijające się (np. Indie, Meksyk, Turcja), z umiarkowanym poziomem zapisów.
- **Klaster 3**: Kraje o średnim poziomie edukacji (np. Chiny, Egipt, Rosja), z wyzwaniami w dostępie.
- **Klaster 4**: Kraje z problemami w dostępie (np. Afganistan, Somalia, Jemen), z wysokim odsetkiem dzieci poza szkołą.
- **Klaster 5**: Kraje afrykańskie i wyspiarskie (np. Angola, Nigeria, Uganda), z niskim dostępem.
- **Klaster 6**: Kraje o niskim poziomie edukacji (np. Bangladesz, Etiopia, Pakistan), z krótkim okresem edukacji.

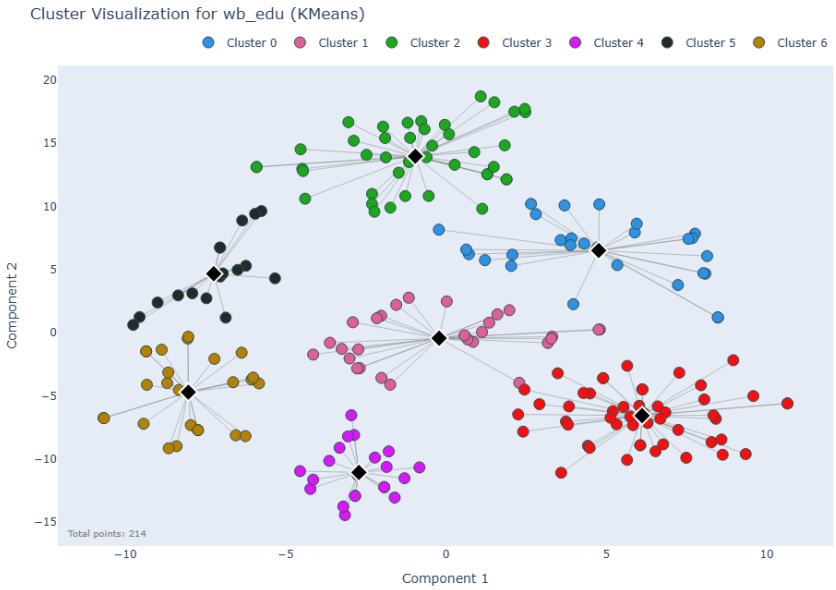
Wnioski: Hierarchiczne grupowanie precyzyjniej oddziela kraje rozwinięte od małych krajów. Klastry 4, 5 i 6 wskazują na nierówności w edukacji.

DBSCAN

- **Noise**: Kraje o unikalnych systemach edukacji (np. Botswana, Indonezja, Rosja), z nietypowymi zapisami lub wydatkami.
- **Klaster 0**: Kraje z problemami w dostępie (np. Afganistan, Somalia, Jemen), z wysokim odsetkiem dzieci poza szkołą.
- **Klaster 1**: Małe kraje (np. Bahrajn, Katar), z wysokim poziomem edukacji.
- **Klaster 2**: Kraje afrykańskie i wyspiarskie (np. Angola, Nigeria, Uganda), z niskim dostępem.
- **Klaster 3**: Kraje o średnim poziomie rozwoju (np. Algieria, Niemcy, Włochy), z umiarkowanymi zapisami.
- **Klaster 4**: Kraje rozwijające się (np. Armenia, Filipiny, Sri Lanka), z zróżnicowanym dostępem.
- **Klaster 5**: Kraje rozwinięte (np. Australia, Japonia, Szwecja), z wysokimi zapisami.

– **Klaster 6**: Kraje afrykańskie (np. Bangladesz, Kamerun), z wysokim odsetkiem dzieci poza szkołą.
– **Klaster 7**: Małe kraje (np. Belize, Malediwy), z wysokim poziomem edukacji.
– **Klaster 8**: Kraje rozwijające się (np. Indie, Meksyk, Tunezja), z umiarkowanym finansowaniem.
– **Klastry 9-18**: Mniejsze klastry o specyficznych cechach, np. Białoruś, Chiny, Dania.

Wnioski: DBSCAN generuje wiele klastrów, wskazując na heterogeniczność danych. Punkty szumu podkreślają unikalne systemy edukacyjne.



Rysunek 3. Wizualizacja klastrów dla metody K-średnich - dane edukacyjne

4.4 Zbiór danych: Sektor finansowy

K-średnich

– **Klaster 0**: Małe kraje i terytoria zależne (np. Bermudy, Monako, Singapur), z niską inflacją i wysoką płynnością banków.
– **Klaster 1**: Kraje z problemami finansowymi (np. Argentyna, Sudan, Zimbabwe), z wysoką inflacją.

- **Klaster 2**: Kraje rozwijające się (np. Bangladesz, Kenia, Tanzania), z umiarkowaną inflacją i ograniczonym dostępem do usług bankowych.
- **Klaster 3**: Kraje rozwinięte i wschodzące (np. Australia, Korea Południowa, Malezja), z rozwiniętym sektorem bankowym.
- **Klaster 4**: Kraje zróżnicowane finansowo (np. Japonia, Rosja, Wenezuela), z wahaniami kursów walut.
- **Klaster 5**: Kraje europejskie i azjatyckie (np. Niemcy, Chiny, Filipiny), z rozwiniętym sektorem finansowym.

Wnioski: Klastry wskazują na zróżnicowanie w stabilności finansowej. Małe kraje pełnią rolę centrów finansowych, a klaster 1 wskazuje na kraje z problemami finansowymi.

Grupowanie hierarchiczne

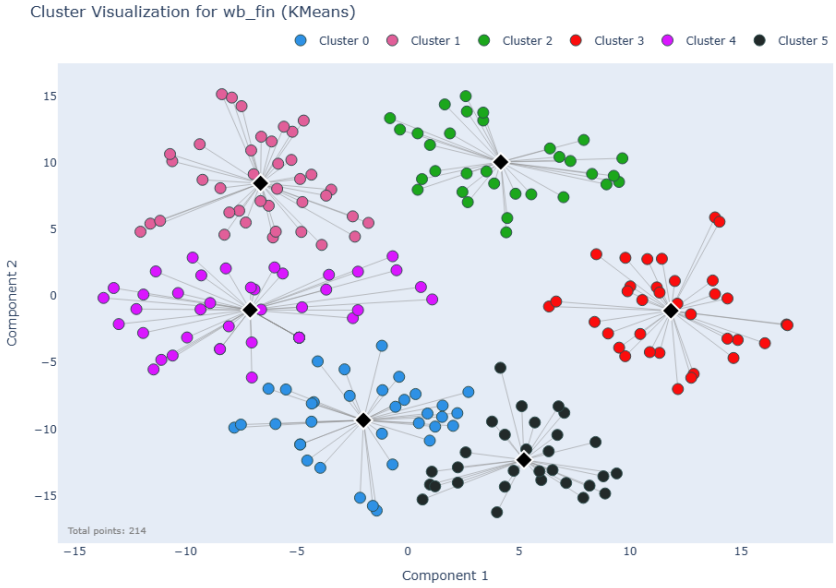
- **Klaster 0**: Kraje rozwinięte europejskie i azjatyckie (np. Niemcy, Japonia, Filipiny), z niską inflacją i rozwiniętym sektorem bankowym.
- **Klaster 1**: Małe kraje i terytoria zależne (np. Bermudy, Singapur, Monako), z wysoką płynnością bankową.
- **Klaster 2**: Kraje z problemami finansowymi (np. Argentyna, Sudan, Zimbabwe), z wysoką inflacją.
- **Klaster 3**: Kraje zróżnicowane finansowo (np. Japonia, Rosja, Wenezuela), z wahaniami kursów walut.
- **Klaster 4**: Kraje rozwinięte i wschodzące (np. Australia, Korea Południowa, Malezja), z stabilnymi wskaźnikami.
- **Klaster 5**: Kraje rozwijające się (np. Bangladesz, Kenia, Tanzania), z umiarkowaną inflacją.

Wnioski: Hierarchiczne grupowanie lepiej rozdziela kraje rozwinięte od małych krajów (o małym obszarze). Klaster 2 ułatwia identyfikację krajów wymagających reform. Grupy są bardzo podobne do grup stworzonych przez algorytm K-średnich.

DBSCAN

- **Noise**: Kraje o nietypowych profilach (np. Arabia Saudyjska, Singapur, Wietnam), z wysoką inflacją lub unikalnymi systemami bankowymi.
- **Klaster 0**: Kraje rozwijające się (np. Albania, Indonezja, Turcja), z umiarkowaną inflacją.
- **Klaster 1**: Kraje z problemami finansowymi (np. Argentyna, Sudan, Zimbabwe), z wysoką inflacją.
- **Klaster 2**: Małe kraje i terytoria zależne (np. Bermudy, Monako, Kuba), z wysoką płynnością bankową.
- **Klaster 3**: Kraje rozwinięte i wschodzące (np. Niemcy, Malezja, USA), z rozwiniętym sektorem bankowym.
- **Klastry 4-10**: Mniejsze klastry o specyficznych cechach, np. Estonia, Chiny, Kolumbia.

Wnioski: DBSCAN lepiej identyfikuje nietypowe profile finansowe (np. Singapur, Zimbabwe). Występuje duża liczba małych klastrów oraz jedna grupa dominująca krajów bogatych lub średniozamożnych.



Rysunek 4. Wizualizacja klastrów dla metody K-średnich - dane finansowe

4.5 Zbiór danych: Zdrowie

K-średnich

- **Klaster 0**: Kraje o średnim poziomie opieki zdrowotnej (np. Brazylia, Meksyk, Panama), z umiarkowanymi wskaźnikami szczepień.
- **Klaster 1**: Kraje rozwinięte (np. Japonia, Niemcy, USA), z wysokimi wskaźnikami szczepień i dostępem do opieki.
- **Klaster 2**: Kraje rozwijające się (np. Angola, Etiopia, Somalia), z niskimi wskaźnikami szczepień i wysoką śmiertelnością.
- **Klaster 3**: Kraje o zróżnicowanym poziomie opieki (np. Azerbejdżan, Wietnam), z umiarkowanym dostępem.
- **Klaster 4**: Kraje z wysoką urbanizacją i dostępem do wody pitnej (np. Egipt, Tajlandia, Turcja), ale zróżnicowanymi wskaźnikami.
- **Klaster 5**: Kraje afrykańskie i wyspiarskie (np. Botswana, Zambia, Papua Nowa Gwinea), z problemami w dostępie.

Wnioski: Kraje rozwinięte mają najlepsze systemy zdrowotne. Klastry 2 i 5 wskazują na potrzebę poprawy opieki w krajach rozwijających się.

Grupowanie hierarchiczne

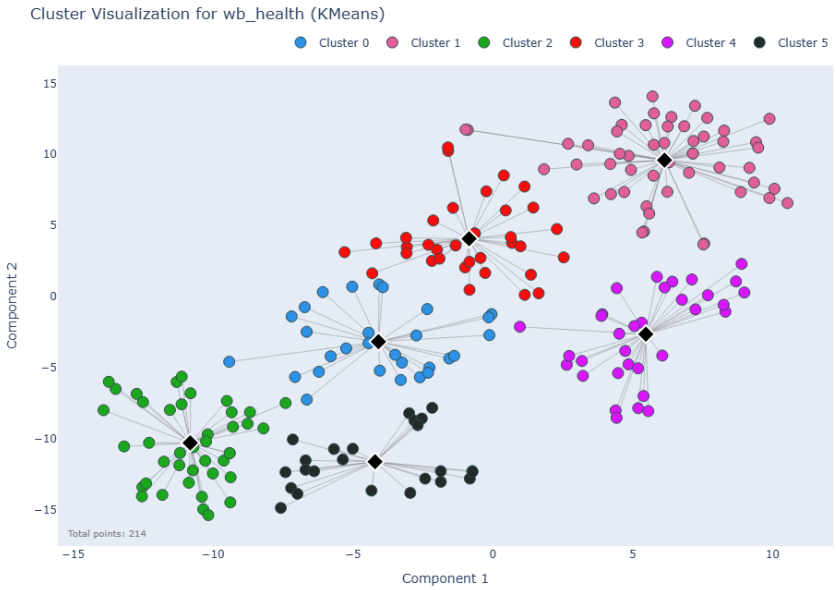
- **Klaster 0**: Kraje afrykańskie i wyspiarskie (np. Botswana, Zambia, Papua Nowa Gwinea), z problemami w dostępie.
- **Klaster 1**: Kraje rozwijające się (np. Angola, Etiopia, Somalia), z niskimi wskaźnikami szczepień.
- **Klaster 2**: Kraje rozwinięte (np. Japonia, Niemcy, USA), z wysokimi wskaźnikami szczepień.
- **Klaster 3**: Kraje o zróżnicowanym poziomie opieki (np. Azerbejdżan, Wietnam), z umiarkowanym dostępem.
- **Klaster 4**: Kraje o średnim poziomie opieki (np. Brazylia, Meksyk, Panama), z umiarkowanymi wskaźnikami.
- **Klaster 5**: Kraje z wysoką urbanizacją i dostępem do wody pitnej (np. Egipt, Tajlandia, Turcja), z zróżnicowanymi wskaźnikami.

Wnioski: Hierarchiczne grupowanie lepiej oddziela kraje afrykańskie od krajów o średnim poziomie opieki.

DBSCAN

- **Noise**: Kraje o nietypowych systemach zdrowotnych (np. Brazylia, Tajlandia, Syria), z unikalnymi wskaźnikami.
- **Klaster 0**: Kraje rozwijające się (np. Angola, Nigeria, Jemen), z niskimi wskaźnikami szczepień.
- **Klaster 1**: Kraje o średnim poziomie opieki (np. Bułgaria, Panama, Serbia), z umiarkowanymi wskaźnikami.
- **Klaster 2**: Kraje rozwinięte (np. Japonia, Szwecja, Singapur), z wysokimi wskaźnikami szczepień.
- **Klastry 3-12**: Mniejsze klastry o specyficznych cechach, np. Białoruś, Chiny, Belize.

Wnioski: DBSCAN generuje więcej klastrów. Występuje jeden dominujący klaster z krajami o średnim poziomie opieki oraz szum z nietypowymi cechami w systemach zdrowotnych tych krajów.



Rysunek 5. Wizualizacja klastrów dla metody K-średnich - dane zdrowotne

4.6 Zbiór danych: Sektor prywatny i handel

K-średnich

- **Klaster 0**: Kraje rozwinięte (np. Niemcy, Japonia, USA), z wysokim udziałem eksportu/importu do gospodarek rozwiniętych.
- **Klaster 1**: Kraje rozwijające się (np. Kenia, Nigeria, Uganda), z ograniczonym handlem międzynarodowym.
- **Klaster 2**: Kraje azjatyckie i pacyficzne (np. Chiny, Singapur, Wietnam), z silnym handlem regionalnym.
- **Klaster 3**: Małe kraje wyspiarskie (np. Bahamy, Seszele), z wysokim udziałem importu z gospodarek rozwiniętych.
- **Klaster 4**: Kraje o zróżnicowanym profilu handlowym (np. Burkina Faso, Nepal, Syria), z ograniczonym handlem.
- **Klaster 5**: Kraje o średnim poziomie rozwoju (np. Indie, Turcja, Kolumbia), z rozwiniętym handlem.
- **Klaster 6**: Kraje o wysokim udziale handlu z gospodarkami rozwiniętymi (np. Brazylia, Rosja, USA).

Wnioski: Klastry odzwierciedlają globalne wzorce handlowe. Kraje rozwinięte dominują w handlu, a kraje rozwijające się mają ograniczony udział.

Grupowanie hierarchiczne

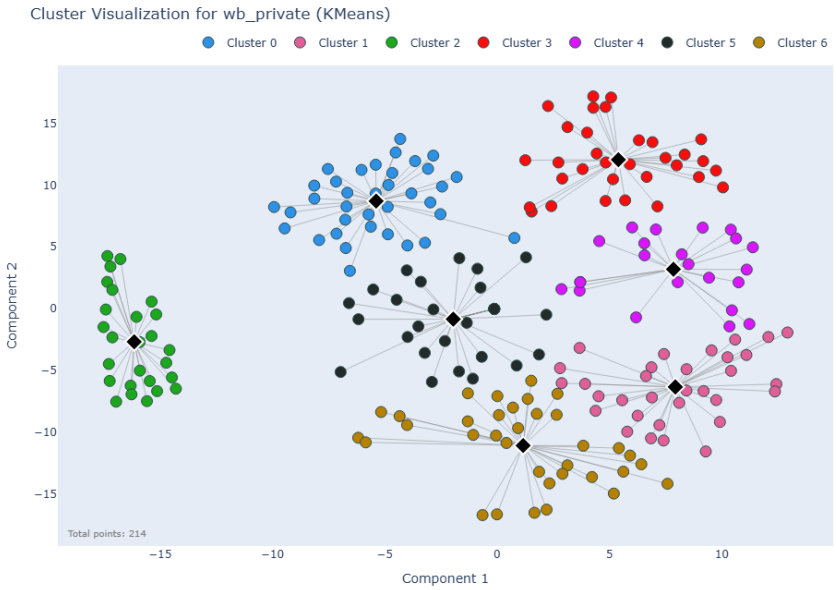
- **Klaster 0**: Kraje rozwijające się (np. Kenia, Nigeria, Uganda), z ograniczonym handlem.
- **Klaster 1**: Kraje o wysokim udziale handlu z gospodarkami rozwiniętymi (np. Brazylia, Rosja, USA).
- **Klaster 2**: Kraje o zróżnicowanym profilu handlowym (np. Burkina Faso, Nepal, Syria).
- **Klaster 3**: Małe kraje wyspiarskie (np. Bahamy, Seszele), z wysokim udziałem importu.
- **Klaster 4**: Kraje azjatyckie i pacyficzne (np. Chiny, Singapur, Wietnam), z silnym handlem regionalnym.
- **Klaster 5**: Kraje rozwinięte (np. Niemcy, Japonia, USA), z wysokim udziałem eksportu/importu.
- **Klaster 6**: Kraje o średnim poziomie rozwoju (np. Indie, Turcja, Kolumbia), z rozwiniętym handlem.

Wnioski: Hierarchiczne grupowanie lepiej wyróżnia kraje azjatyckie z silnym handlem regionalnym. Klaster 3 podkreśla zależność małych krajów od importu.

DBSCAN

- **Noise**: Kraje o nietypowych wzorcach handlowych (np. Bangladesz, Etiopia, Kuba).
- **Klaster 0**: Kraje o wysokim udziale handlu z gospodarkami rozwiniętymi (np. Brazylia, Rosja, USA).
- **Klaster 1**: Małe kraje wyspiarskie (np. Barbados, Seszele), z wysokim udziałem importu.
- **Klaster 2**: Kraje o średnim poziomie rozwoju (np. Izrael, Turcja, Ukraina), z rozwiniętym handlem.
- **Klastry 3-7**: Mniejsze klastry o specyficznych wzorcach, np. Chiny, Singapur, Austria.

Wnioski: DBSCAN lepiej wyróżnia nietypowe wzorce handlowe. Klaster 1 jest bardziej precyzyjny w identyfikacji małych krajów od poprzednich metod.



Rysunek 6. Wizualizacja klastrów dla metody K-średnich - dane publiczne

4.7 Zbiór danych: Sektor publiczny

K-średnich

- **Klaster 0**: Kraje o niskich wartościach wskaźników zarządzania publicznego (np. Afganistan, Somalia, Syria), z niskimi wynikami w stabilności politycznej i kontroli korupcji.
- **Klaster 1**: Małe kraje i terytoria zależne (np. Bermudy, Monako, Seszele), z wysokimi wartościami wskaźników zarządzania, szczególnie w regulacjach jakościowych i praworządności.
- **Klaster 2**: Kraje rozwinięte (np. Australia, Niemcy, Japonia), z wysokimi wynikami w stabilności politycznej, kontroli korupcji.
- **Klaster 3**: Kraje o średnich wartościach wskaźników zarządzania (np. Brazylia, Indie, Turcja), z umiarkowanymi wynikami w badanych czynnikach.

Wnioski: Klastry odzwierciedlają zróżnicowanie w poziomie zarządzania publicznego. Podział obejmuje 4 dobrze oddzielone klastry, z wyraźną koncentracją krajów wyspiarskich, wysoko rozwiniętych, średnio rozwiniętych i słabo rozwiniętych w osobnych grupach.

Grupowanie hierarchiczne

- **Klaster 0**: Kraje o niskich wartościach wskaźników zarządzania publicznego (np. Afganistan, Somalia, Syria), z niskimi wynikami w stabilności politycznej i praworządności.

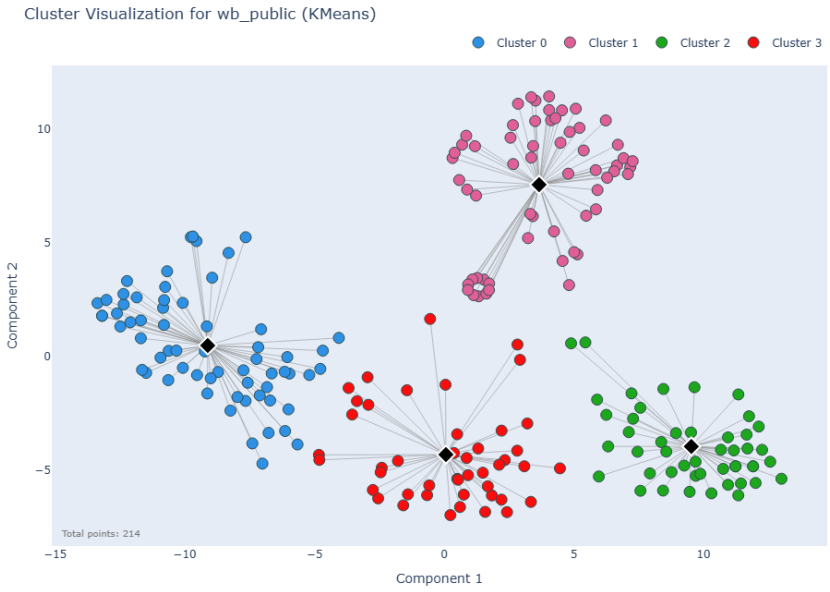
- **Klaster 1**: Małe kraje i terytoria zależne (np. Bermudy, Monako, Seszele), z wysokimi wartościami regulacji jakościowych i kontroli korupcji.
- **Klaster 2**: Kraje rozwinięte (np. Australia, Niemcy, Japonia), z wysokimi wynikami w zdolnościach statystycznych i głosie społecznym.
- **Klaster 3**: Kraje o średnich wartościach wskaźników zarządzania (np. Brazylia, Indie, Turcja), z umiarkowanymi wynikami w stabilności.

Wnioski: Wyniki są zbliżone do K-średnich, zachowując podział i bliskość klastrów. Klaster 1 obejmuje także niektóre kraje rozwijające się (np. Rwanda). Podział odzwierciedla poziom rozwoju krajów.

DBSCAN

- **Noise**: Kraje o nietypowych profilach zarządzania publicznego (np. Singapur, Arabia Saudyjska, Kuba), z unikalnymi kombinacjami wskaźników, np. z wysokimi wskaźnikami oceny jakości regulacji przy niskim wskaźniku możliwości wyrażania głosu przez społeczeństwo.
- **Klaster 0**: Kraje o niskich wartościach wskaźników zarządzania (np. Afganistan, Somalia, Syria), z niskimi wynikami w stabilności i kontroli korupcji.
- **Klaster 1**: Kraje o średnich wartościach wskaźników zarządzania (np. Brazylia, Indie, Turcja), z umiarkowanymi wynikami w udokumentowaniu publicznych danych i praworządności.
- **Klaster 2**: Małe kraje i terytoria zależne (np. Bermudy, Monako, Seszele), z wysokimi wartościami regulacji.
- **Klaster 3**: Kraje rozwinięte (np. Niemcy, Japonia, USA), z wysokimi wynikami w stabilności politycznej i głosie społecznym.
- **Klastry 4-9**: Mniejsze klastry o specyficznych cechach, np. Bahrajn, Bangladesz.

Wnioski: DBSCAN identyfikuje 10 bardzo nierównomiernych klastrów, z trzema dominującymi (kilkadziesiąt krajów) i kilkoma mniejszymi grupami. Algorytm precyzyjnie wyróżnia nietypowe profile zarządzania, a klaster 0 dokładniej określa kraje o niskich wartościach wskaźników.



Rysunek 7. Wizualizacja klastrów dla metody K-średnich - dane publiczne

4.8 Zbiór danych: Ochrona socjalna i rynek pracy

K-średnich

- **Klaster 0**: Kraje o zróżnicowanym poziomie zatrudnienia (np. Argentyna, Włochy, Zambia), z umiarkowanym udziałem kobiet.
- **Klaster 1**: Kraje rozwijające się (np. Egipt, Nepal, Syria), z wysokim samozatrudnieniem i bezrobociem młodzieży.
- **Klaster 2**: Kraje rozwinięte (np. Japonia, Niemcy, USA), z wysoką partycypacją i niskim bezrobociem.
- **Klaster 3**: Kraje o niskim poziomie zatrudnienia (np. Bangladesz, Etiopia, Nigeria), z wysokim bezrobociem młodzieży.

Wnioski: Klastry odzwierciedlają nierówności na rynku pracy. Kraje rozwinięte mają wysoką partycypację kobiet na rynk pracy, a rozwijające się wymagają wsparcia. Wyraźnie zgrupowane są kraje zamożne.

Grupowanie hierarchiczne

- **Klaster 0**: Kraje rozwinięte (np. Japonia, Niemcy, USA), z wysoką partycypacją i niskim bezrobociem.
- **Klaster 1**: Kraje o niskim poziomie zatrudnienia (np. Bangladesz, Etiopia, Nigeria), z wysokim bezrobociem młodzieży.

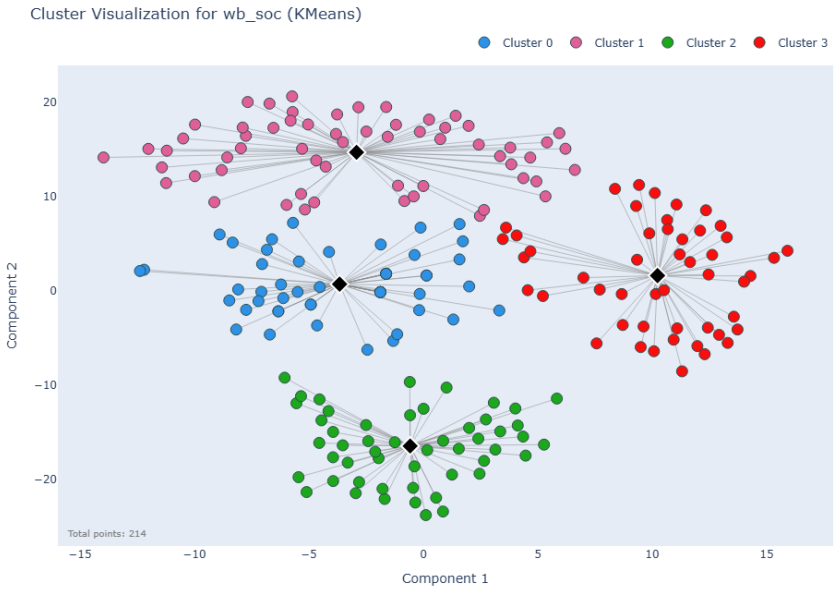
- **Klaster 2**: Kraje o zróżnicowanym poziomie zatrudnienia (np. Argentyna, Włochy, Zambia), z umiarkowanym udziałem kobiet.
- **Klaster 3**: Kraje rozwijające się (np. Egipt, Nepal, Syria), z wysokim samozatrudnieniem.

Wnioski: Hierarchiczne grupowanie lepiej oddziela kraje rozwinięte od zróżnicowanych. Klaster 1 grupuje kraje rozwijające się mierzące się z wyzwaniami społecznymi.

DBSCAN

- **Noise**: Kraje o nietypowych profilach rynku pracy (np. Malezja, Zambia, Ukraina).
- **Klaster 0**: Kraje rozwijające się (np. Egipt, Syria, Turcja), z wysokim bezrobociem młodzieży.
- **Klaster 1**: Terytoria zależne (np. Bermudy, Monako), z ograniczonymi danymi.
- **Klaster 2**: Małe kraje wyspiarskie (np. Antigua i Barbuda, Seszele), z wysokim samozatrudnieniem.
- **Klastry 3-9**: Mniejsze klastry o specyficznych cechach, np. Australia, Bangladesz, Botswana.

Wnioski: DBSCAN generuje bardzo dużo klastrów, rozbija grupy na bardo mało liczne, lecz zachowuje odrębność krajów wyspiarskich.



Rysunek 8. Wizualizacja klastrów dla metody K-średnich - dane socjalne

4.9 Ogólne wnioski

- **Zróżnicowanie globalne**: Wszystkie metody potwierdzają podziały między krajami rozwiniętymi, rozwijającymi się i najsłabiej rozwiniętymi. Kraje rozwinięte (np. USA, Niemcy, Japonia) konsekwentnie grupowane są razem.
- **Korelacje między wymiarami**: Kraje o wysokim poziomie rozwoju gospodarczego osiągają dobre wyniki w edukacji, zdrowiu i pod względem czynników środowiskowych, co wskazuje na silną współzależność.
- **Specyfika małych krajów**: Małe kraje wyspiarskie i terytoria zależne (np. Bermudy, Seszele) często tworzą osobne klastry, co podkreśla ich unikalny profil.

Porównanie metod:

- **K-średnich**: Skutecznie identyfikuje naturalne grupy, szczególnie w dużych zbiorach danych. Stabilne, ale może łączyć kraje o zróżnicowanych cechach.
- **Hierarchiczne grupowanie**: Daje bardziej jasne podziały, lepiej oddzielając małe kraje i kraje o specyficznych profilach, ale mniej stabilne przy dużej liczbie klastrów.
- **DBSCAN**: Wyróżnia nietypowe przypadki (punkty szumu) i generuje więcej klastrów, co odzwierciedla heterogeniczność danych, ale utrudnia ogólne wnioski.

W celu oceny jakości przeprowadzonego klastrowania zastosowano cztery główne metryki ewaluacyjne dla każdego ze zbiorów danych:

- **Współczynnik sylwetki (Silhouette Score)**: określa stopień spójności wewnętrznej oraz separacji klastrów. Im wyższa jego wartość (w przedziale $[-1, 1]$), tym lepiej ukształtowane są klastry.
- **Odległość wewnątrzklastrowa (Within-Cluster Distance)**: średnia odległość pomiędzy punktami należącymi do tego samego klastra. Mniejsza wartość wskazuje na bardziej zwarte skupiska.
- **Odległość międzyklastrowa (Between-Cluster Distance)**: średnia odległość pomiędzy centroidami różnych klastrów. Wyższe wartości sugerują lepsze rozdzielenie klastrów.
- **Stosunek punktów szumu (Noise Ratio)**: istotny jedynie w przypadku algorytmu DBSCAN; wskazuje, jaka część próbek została oznaczona jako odstające.

Dodatkowo, w celu porównania spójności rezultatów pomiędzy zastosowanymi algorytmami (K-średnich, klasteryzacja hierarchiczna, DBSCAN), wykorzystano metrykę znormalizowanej informacji wzajemnej (NMI). Pozwala ona ocenić zgodność dwóch różnych podziałów tych samych danych, niezależnie od liczby oraz etykiet klastrów.

Tabela 1. Porównanie metryk ewaluacyjnych dla różnych metod klastrowania i zbiorów danych

Zbiór danych	**Metoda**	**Silhouette Score**	**WC Distance**	**BC Distance**	**Noise Ratio**
wb_ econ	K-średnich	0.4317	3.0175	13.3852	–
wb_ env	K-średnich	0.4322	3.1656	13.1438	–
wb_ edu	K-średnich	0.4964	2.6883	13.7044	–
wb_ fin	K-średnich	0.4163	3.8012	15.8267	–
wb_ health	K-średnich	0.4423	3.0321	13.5871	–
wb_ private	K-średnich	0.4450	3.4361	15.1041	–
wb_ public	K-średnich	0.5795	2.9679	13.1595	–
wb_ soc	K-średnich	0.4844	3.7239	16.2441	–
wb_ econ	Hierarchiczne	0.4012	5.2393	20.5031	–
wb_ env	Hierarchiczne	0.3957	3.2668	13.4432	–
wb_ edu	Hierarchiczne	0.4898	2.7226	13.9064	–
wb_ fin	Hierarchiczne	0.4230	3.9133	15.5494	–
wb_ health	Hierarchiczne	0.4298	3.0793	13.7001	–
wb_ private	Hierarchiczne	0.4197	3.5625	15.0255	–
wb_ public	Hierarchiczne	0.5725	3.0079	13.0394	–
wb_ soc	Hierarchiczne	0.4743	5.3993	18.8798	–
wb_ econ	DBSCAN	0.2005	4.6279	17.9054	0.0748
wb_ env	DBSCAN	0.2439	2.9548	12.9752	0.0701
wb_ edu	DBSCAN	0.4007	1.5726	11.9046	0.0841
wb_ fin	DBSCAN	0.2682	4.3135	15.0243	0.1308
wb_ health	DBSCAN	0.3385	1.8647	11.9149	0.0794
wb_ private	DBSCAN	0.2483	3.2999	12.1518	0.0421
wb_ public	DBSCAN	0.3574	1.8953	10.1262	0.1075
wb_ soc	DBSCAN	0.3130	28.2115	127.1226	0.0888

Współczynnik sylwetki

Analiza współczynnika sylwetki pokazuje, że metoda K-średnich uzyskała najwyższe wyniki spośród rozważanych metod klastrowania. Szczególnie wyróżniają się zbiory danych publicznych ($s = 0{,}5795$), edukacyjnych ($s = 0{,}4964$) oraz socjalnych ($s = 0{,}4844$), co potwierdza wyraźną separację klastrów w tych dziedzinach. Nieco niższe wartości osiągnęło grupowanie hierarchiczne, co może być wynikiem większej liczby utworzonych klastrów lub mniej precyzyjnego przypisania punktów granicznych. DBSCAN, jako metoda oparta na gęstości, systematycznie osiągał niższe wartości ($s \approx 0{,}20$–$0{,}40$), co może wynikać z obecności punktów szumu i nieregularnych kształtów klastrów.

Odległości wewnątrz- i międzyklastrowe

Pod względem zwartości klastrów (WC Distance) najlepsze wyniki uzyskano dla zbiorów dotyczących edukacji i wydatków publicznych przy użyciu K-średnich, gdzie odległości te wynosiły odpowiednio 2,6883 i 2,9679. Oznacza to, że punkty w obrębie jednego klastra są blisko siebie, co sprzyja interpretacji wyników. Z kolei największe wartości międzyklastrowe (BC Distance) zaobserwowano w zbiorach sektora socjalnego i finansowego ($> 15{,}0$), co wskazuje na dużą rozdzielność pomiędzy grupami krajów.

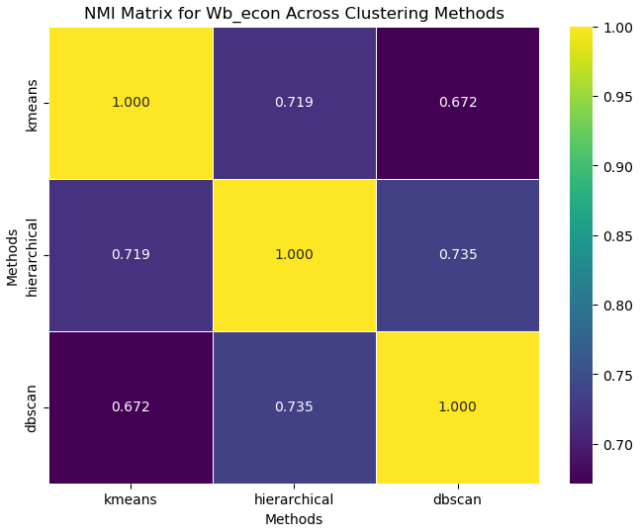
Dla DBSCAN odległości wewnątrzklastrowe były zróżnicowane – od niskich (np. edukacja: 1,5726) po bardzo wysokie (np. czynniki socjalne: 28,2115), co odzwierciedla różny charakter klastrów w zależności od struktury danych i parametrów. Duże wartości odległości wewnątrz- i międzyklastrowych dla zbioru danych dotyczących czynników socjalnych mogą wynikać z dużej różnorodności wybranych cech, takich jak migracja netto, zatrudnienie, bezrobocie z podziałem na płeć czy odsetek młodzieży nieuczestniczącej w edukacji i pracy, które charakteryzują się wysoką zmiennością między krajami. Te cechy, obejmujące zarówno wskaźniki ekonomiczne (np. udział siły roboczej), jak i społeczne (np. uchodźcy, migracja), mogą tworzyć klastry o nieregularnych kształtach, co utrudnia uzyskanie niskich odległości wewnątrzklastrowych, szczególnie w przypadku metody DBSCAN.

Znormalizowana informacja wzajemna (NMI)

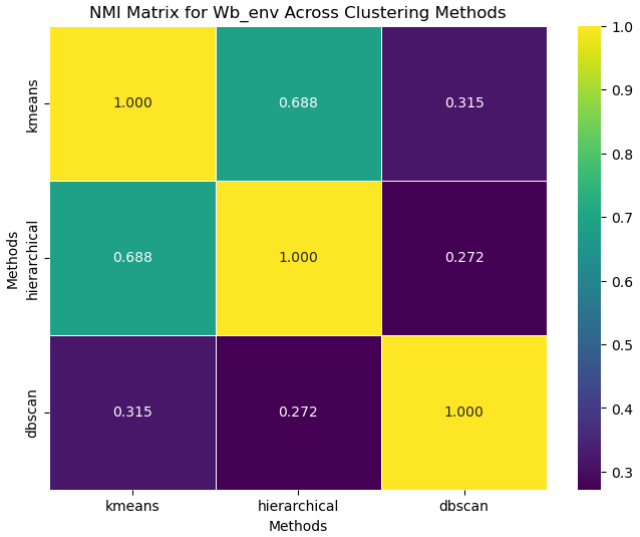
Zgodność wyników klastrowania między metodami oceniono za pomocą znormalizowanej informacji wzajemnej (NMI). Najwyższe wartości NMI między metodą K-średnich a klastrowaniem hierarchicznym uzyskano dla zbiorów z danymi edukacyjnymi (0,877), sektora publicznego (0,886) oraz danych finansowych (0,907), co wskazuje na podobne podziały krajów i potwierdza wyraźną strukturę klastrową w tych danych. Niższe wartości NMI dla DBSCAN, szczególnie w zbiorach z danymi środowiskowymi (0,315 dla K-średnich, 0,272 dla hierarchicznego) i finansowymi (0,473 dla K-średnich, 0,468 dla hierarchicznego), sugerują, że DBSCAN generuje bardziej zróżnicowane klastry, prawdopodobnie

z powodu lokalnych gęstości danych i większej liczby punktów oznaczonych jako szum.

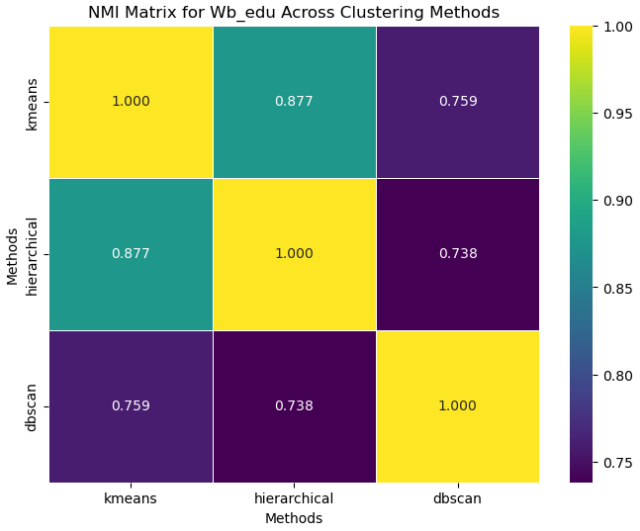
- Wysoka zgodność dla danych edukacyjnych, z sektora publicznego oraz finansowych wynika z cech takich jak wydatki na edukację, jakość rządzenia czy inflacja, które mają stosunkowo jednolitą strukturę między krajami.
- Niskie NMI dla DBSCAN w przypadku danych środowiskowych może być efektem dużej zmienności cech, takich jak emisje CO_2 czy gęstość zaludnienia, które tworzą nieregularne klastry.
- W danych finansowych DBSCAN wykazuje niższą zgodność, prawdopodobnie z powodu wrażliwości na skrajne wartości w cechach finansowych, takich jak kredyty czy kursy walut.



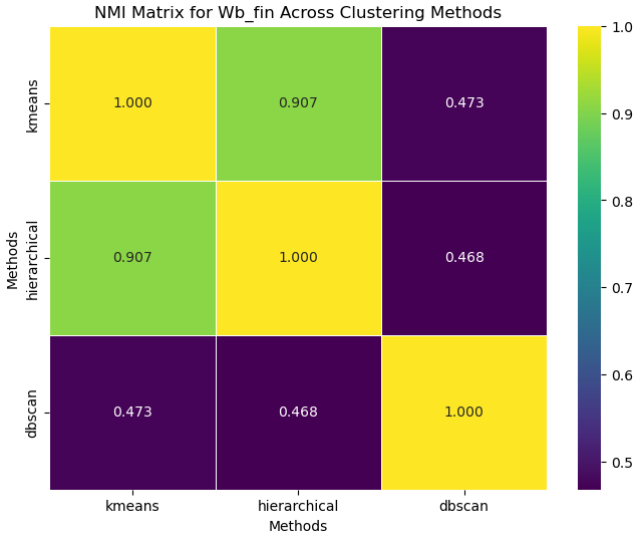
Rysunek 9. Macierz NMI dla danych ekonomicznych dla poszczególnych metod klastrowania



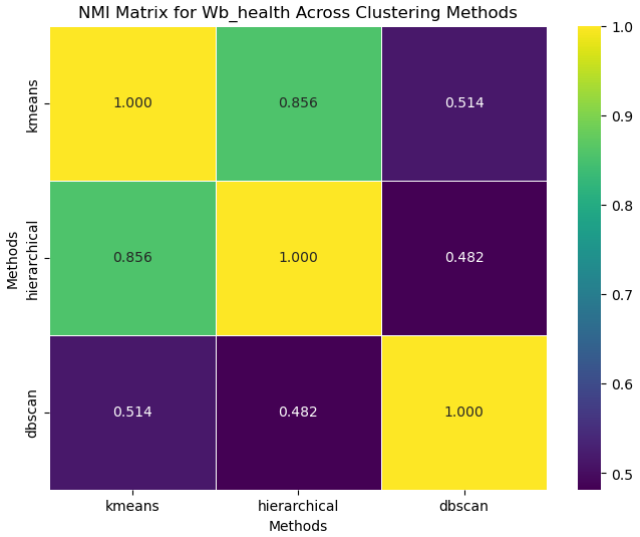
Rysunek 10. Macierz NMI dla danych środowiskowych dla poszczególnych metod klastrowania



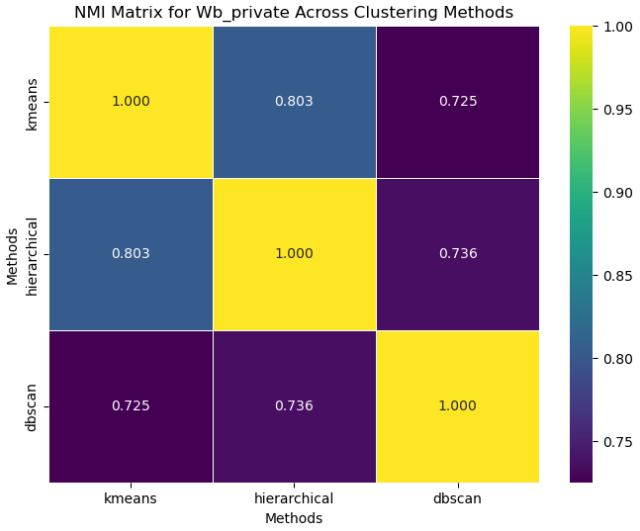
Rysunek 11. Macierz NMI dla danych edukacyjnych dla poszczególnych metod klastrowania



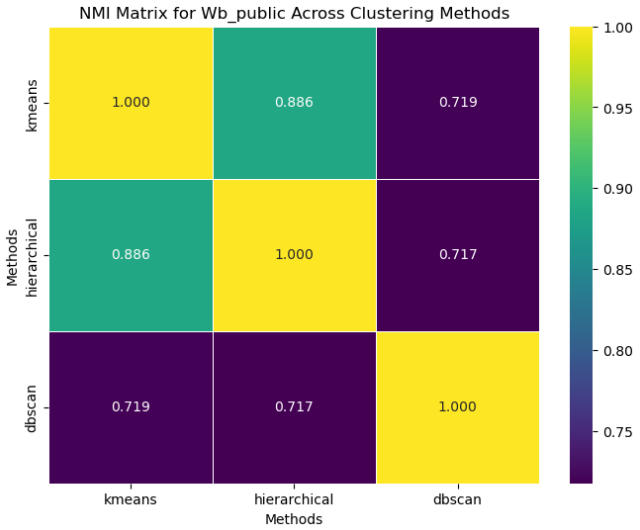
Rysunek 12. Macierz NMI dla danych finansowych dla poszczególnych metod klastrowania



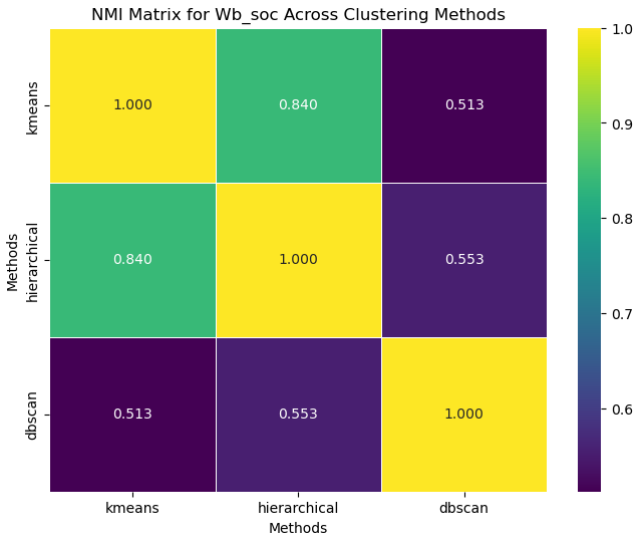
Rysunek 13. Macierz NMI dla danych zdrowotnych dla poszczególnych metod klastrowania



Rysunek 14. Macierz NMI dla danych sektora prywatnego dla poszczególnych metod klastrowania



Rysunek 15. Macierz NMI dla danych sektora publicznego dla poszczególnych metod klastrowania



Rysunek 16. Macierz NMI dla danych socjalnych dla poszczególnych metod klastrowania

Podsumowanie

- **K-średnich** zapewnia najlepszą spójność i separację klastrów w zbiorach o dobrze ukształtowanej strukturze (np. edukacja, sektor publiczny).
- **Grupowanie hierarchiczne** często daje porównywalne wyniki, szczególnie tam, gdzie klastry są nieregularne, ale spójne (np. sektor zdrowia).
- **DBSCAN** dobrze identyfikuje punkty odstające i nietypowe struktury, ale jego wyniki są trudniejsze do porównania z innymi metodami.
- Wysoka zgodność między K-średnich i klasteryzacją hierarchiczną potwierdza stabilność wyników w większości zbiorów, natomiast rozbieżności z DBSCAN wskazują na jego odmienną naturę analityczną.

4.10 Porównanie z innymi podejściami

Podejście przedstawione w tej pracy bazowane było na artykule N.Darapanei et al. [13]. Obie prace stosują klastrowanie K-średnich i hierarchiczne, co pozwala na identyfikację podobnych grup krajów, jednak dane wejściowe różnią się ze względu na cel analizy. W tej pracy analizowano szeroki zestaw surowych wskaźników Banku Światowego z 2022 roku, podzielonych na osiem kategorii (ekonomia, środowisko, edukacja, finanse, zdrowie, sektor prywatny, publiczny, społeczny), co zapewnia szczegółową analizę różnych wymiarów. Z kolei Darapaneni i współautorzy opierają się na danych obejmujących demografię, ekonomię, zdrowie, pogodę oraz dane epidemiologiczne COVID-19 dla 142 krajów, z naciskiem na wpływ pandemii i reakcje rządów mierzone indeksem stringencji. W

przeciwieństwie do tego, kolejne podejście przedstawione w artykule autorstwa N.Koutsoukis [14] korzysta z 15 syntetycznych indeksów (np. KOF Globalization Index), co upraszcza analizę, ale ogranicza elastyczność w porównaniu z surowymi danymi użytymi w obu pozostałych podejściach.

Metodologicznie, podejście tej pracy wyróżnia się zaawansowanym preprocessingiem, obejmującym imputację medianą, usuwanie wysoko skorelowanych cech, dyskretyzację oraz redukcję wymiarowości (t-SNE, TruncatedSVD), co zwiększa jakość klastrowania. Dodatkowo, stosuje się trzy metody klastrowania (K-średnich, hierarchiczne, DBSCAN) i formalne metryki oceny, takie jak Silhouette Score czy NMI, umożliwiając szczegółowe porównanie wyników. Darapaneni i współautorzy stosują PCA do redukcji wymiarowości, a ich analiza ogranicza się do K-średnich i klastrowania hierarchicznego, bez DBSCAN, a ocena klastrów opiera się głównie na wizualnych interpretacjach (np. wykresy łokciowe, dendrogramy), bez formalnych metryk. Koutsoukis, stosując prostsze klastrowanie hierarchiczne w oprogramowaniu Orange bez zaawansowanego preprocessingu, pozostaje w tyle pod względem technicznej złożoności.

Porównując wyniki, podejście tej pracy ujawnia zróżnicowane klastry w zależności od kategorii danych. Na przykład, w czynnikach ekonomicznych algorytm K-średnich grupuje kraje bogate, średnio zamożne i wyspiarskie, podczas gdy DBSCAN tworzy mniej intuicyjne połączenia, np. Chorwacja z Rwandą. W edukacji hierarchiczne grupowanie lepiej rozdziela kraje o wysokim i niskim poziomie edukacji, a DBSCAN generuje aż 19 klastrów, w tym nietypowe pary jak USA z Afganistanem. W zdrowiu i sektorze publicznym K-średnich i grupowanie hierarchiczne tworzą dobrze oddzielone klastry, odzwierciedlające poziom rozwoju, podczas gdy DBSCAN często generuje nierównomierne grupy z dominującymi klastrami. Wyniki te wskazują na wysoką zmienność w zależności od metody i kategorii danych, co potwierdza niskie NMI (0.1–0.4) między zbiorami danych.

Z kolei wyniki Darapaneni i współautorów pokazują cztery klastry przed pandemią, które po uwzględnieniu danych COVID-19 rozpadają się na siedem klastrów. Kluczowe różnice w grupach 1 i 3 (np. Belgia, Francja, USA w grupie 1; Argentyna, Brazylia w grupie 3) przypisano różnym strategiom rządowym. W przeciwieństwie do tej pracy, wyniki Darapaneni są bardziej specyficzne dla kontekstu pandemii i mniej zróżnicowane pod względem kategorii danych, ale oferują praktyczne wnioski dla polityki antykryzysowej.

Podobieństwa w wynikach obejmują zdolność obu podejść do identyfikacji zróżnicowania krajów w zależności od analizowanych wymiarów. Na przykład, podobnie jak w tej pracy, gdzie czynniki ekonomiczne i zdrowotne tworzą klastry odzwierciedlające poziom rozwoju, Darapaneni pokazują, że kraje o podobnych cechach przed pandemią (np. grupa 1) różnią się po jej wpływie z powodu działań rządowych. Koutsoukis, choć również identyfikuje zróżnicowanie (np. brak jednorodności w UE), oferuje bardziej ogólne wnioski bez szczegółowej analizy wpływu konkretnych czynników, jak w obu pozostałych pracach.

Różnice w wynikach wynikają głównie z zakresu danych i celów analizy. Podejście tej pracy, analizując osiem kategorii danych, ujawnia szersze spektrum wzorców klastrowych, ale mniej skupia się na praktycznych implikacjach, takich

jak strategie rządowe w pandemii, które są centralne dla Darapaneni. Dodatkowo, DBSCAN w tej pracy generuje bardziej zróżnicowane, czasem mniej intuicyjne klastry w porównaniu z bardziej spójnymi wynikami K-średnich i hierarchicznego grupowania u Darapaneni. Wyniki Koutsoukisa, oparte na indeksach, są mniej szczegółowe i nie uwzględniają dynamiki czasowej, jak w przypadku Darapaneni, ani różnorodności kategorii, jak w tej pracy.

4.11 Dalsze możliwości rozwoju projektu

W celu dalszego udoskonalenia projektu i pogłębienia analizy, można rozważyć następujące kierunki rozwoju:

- **Rozszerzenie zbioru danych o inne źródła** – Integracja danych pochodzących z organizacji międzynarodowych, takich jak *ONZ* (Organizacja Narodów Zjednoczonych) czy *OECD* (Organizacja Współpracy Gospodarczej i Rozwoju), pozwoliłaby na uwzględnienie dodatkowych wymiarów analizy, np. wskaźników dotyczących bezpieczeństwa, jakości rządzenia czy ustroju politycznego państw.
- **Optymalizacja i automatyzacja doboru metod klasteryzacji** – Obecny proces doboru parametru `eps` dla algorytmu `DBSCAN` może zostać zoptymalizowany, na przykład poprzez zastosowanie metod heurystycznych lub algorytmów do automatycznego wyboru optymalnych parametrów. Alternatywnie, możliwe jest wdrożenie mechanizmu automatycznego doboru najbardziej odpowiedniej metody klasteryzacji w zależności od charakterystyki analizowanego zbioru danych.

Repozytorium zawierające kod źródłowy oraz dodatkowe materiały do niniejszej pracy jest dostępne pod adresem: https://github.com/Catcuss/ED_Projekt.git.

Literatura

1. World Bank. *World Development Indicators*. Dostępne online: https://databank.worldbank.org, dostęp: 19.06.2025.
2. Scikit-learn. *KMeans clustering*. Dostępne online: KMeans documentation, dostęp: 19.06.2025.
3. Scikit-learn. *DBSCAN clustering*. Dostępne online: DBSCAN documentation, dostęp: 19.06.2025.
4. SciPy. *Hierarchical clustering*. Dostępne online: cluster.hierarchy, dostęp: 19.06.2025.
5. Scikit-learn. *t-SNE embedding*. Dostępne online: TSNE documentation, dostęp: 19.06.2025.
6. Scikit-learn. *Truncated SVD*. Dostępne online: TruncatedSVD documentation, dostęp: 19.06.2025.
7. Scikit-learn. *NearestNeighbors*. Dostępne online: NearestNeighbors documentation, dostęp: 19.06.2025.
8. Scikit-learn. *Silhouette Score*. Dostępne online: silhouette_score, dostęp: 19.06.2025.

9. Scikit-learn. *Normalized Mutual Information (NMI)*. Dostępne online: NMI documentation, dostęp: 19.06.2025.
10. Plotly. *Interactive graphing library*. Dostępne online: https://plotly.com/python/, dostęp: 19.06.2025.
11. Scikit-learn. *KBinsDiscretizer*. Dostępne online: KBinsDiscretizer documentation, dostęp: 19.06.2025.
12. Scikit-learn. *MinMaxScaler*. Dostępne online: MinMaxScaler documentation, dostęp: 19.06.2025.
13. Darapaneni, Narayana, et al. *Machine learning approach for clustering of countries to identify the best strategies to combat Covid-19*. In: 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), IEEE, 2021.
14. Koutsoukis, Nikitas-Spiros. *Global political economy clusters: the world as perceived through black-box data analysis of proxy country rankings and indicators*. Procedia Economics and Finance, vol. 33, pp. 18–45, 2015.