AKADEMIA GÓRNICZO-HUTNICZA
im. Stanisława Staszica w Krakowie

WYDZIAŁ ZARZĄDZANIA

Przetwarzanie i analiza danych w języku python

Prowadzący: prof. dr hab. inż. Oleksandr Petrov

Autor: Dawid Ciochoń
Projekt zaliczeniowy: Wykorzystanie analizy skupień do oceny rozwoju społeczno-ekonomicznego państw Unii Europejskiej

Kraków, 03.2021

Wstęp

Na przestrzeni ostatnich 25 lat mogliśmy zaobserwować znaczne zmiany w rozwoju nie tylko technicznym, ale także społecznym czy ekonomicznym świata. Dzięki globalizacji, państwa mogą ze sobą ściślej współpracować, a gospodarki słabiej rozwinięte mogą inspirować się sposobem działania krajów wysoko rozwiniętych. Warto również dodać, że kraje, które były do tej pory uznawane za rozwijające się, teraz są jednymi z głównych graczy na arenie politycznej i ekonomicznej na świecie. Jednak państwa mogą rozwijać się dzięki ludziom, którzy je zamieszkują. Bo tak naprawdę poziom rozwoju danego państwa zależy od tego, na ile rozwinięte jest jego społeczeństwo. A od czego zależy poziom rozwoju społecznego? Które czynniki najmocniej decydują o przynależności kraju do słabiej lub lepiej rozwiniętej grupy? I czy słuszna jest teoria, że państwa Europy Wschodniej wciąż są za państwami Europy Zachodniej pod względem ekonomicznym? Te pytania zostaną poddane weryfikacji w dalszej części pracy.

Mając na uwadze ilość czynników, które mają wpływ na rozwój społeczny (wielowymiarowość danych), celem pracy będzie klasyfikacja państw Unii Europejskiej w grupy obrazujące kraje, które cechują się podobnymi wartościami zmiennych przyjętych do analizy za pomocą analizy skupień.

Metodyka badań

Analiza skupień jest narzędziem analizy danych służącym do grupowania *n* obiektów, opisanych za pomocą wektora *p* cech, w *K* niepustych, rozłącznych i możliwie "jednorodnych" grup - skupień. Obiekty należące do danego skupienia powinny być "podobne" do siebie, a obiekty należące do różnych skupień powinny być z kolei możliwie mocno "niepodobne" do siebie[1]. Głównym celem tej analizy jest wykrycie w zbiorze danych, tzw. "naturalnych" skupień, czyli skupień, które dają się w sensowny sposób interpretować.[2]

Wspomniany wyżej podział zbioru obiektów spełnia trzy naturalne wymagania:
1) każde skupienie powinno zawierać przynajmniej jeden obiekt;
2) każdy obiekt musi należeć do pewnego skupienia;
3) każdy obiekt musi należeć do dokładnie jednego skupienia.

Analiza skupień często jest nazywana uczeniem nienadzorowanym. Brakuje tu bowiem informacji o przynależności obiektów do klas, jak również nie wiadomo, ile tych klas powinno naprawdę być.

[1] Petrov O. i in, „Introduction to data mining", Wydawnictwo AGH, Kraków 2019
[2] Krzyśko M. i in., "Systemy uczace się", Wydawnictwo Naukowo-Techniczne, Warszawa 2008

Do najpopularniejszych metod grupowania danych zalicza się metody: hierarchiczne, kombinatoryczne (nazywane też podziałowymi), gęstościowe, gridowe oraz metody korzystające z modeli[3]. Do dalszej analizy zostaną wykorzystane, jednak, tylko metody hierarchiczne i podziałowe.

Metody hierarchiczne

Metody hierarchiczne polegają na sukcesywnym łączeniu bądź dzieleniu obserwacji. W wyniku takiego postępowania otrzymuje się drzewo-podobną strukturę nazywaną dendrogramem.

Techniki aglomeracyjne rozpoczynają się od zbioru obserwacji, z których każda traktowana jest jako oddzielne skupienie. Skupienia są łączone ze sobą zgodnie ze zmniejszającym się stopniem podobieństwa do chwili, aż powstanie jedno skupienie. Natomiast techniki rozdrobnieniowe rozpoczyna się od jedynego skupienia, które podlega sukcesywnym podziałom zgodnie ze wzrastającym stopniem podobieństwa wewnątrzgrupowego. Jednak techniki te są stosowane w praktyce rzadziej niż aglomeracyjne.

Zalety metod hierarchicznych:

- działają według jednej procedury
- wyniki klasyfikacji są przedstawione w postaci ciągu klasyfikacji (możliwość kontrolowania procesu klasyfikacji)
- wyniki klasyfikacji można przedstawić graficznie w formie dendrogramu, wskazującego na kolejność połączeń między klasami. Uzyskana hierarchia umożliwia dokładne określenie, jak są wzajemnie usytuowane poszczególne klasy oraz obiekty w nich zawarte;[4]

Metody hierarchiczne nie są wolne od wad. Najistotniejsze z nich to:

- Tracą swoją przejrzystość ze wzrostem liczby analizowanych obiektów;
- Nie ma możliwości przegrupowania obiektów, które w początkowych etapach analizy zostały nieprawidłowo sklasyfikowane;
- Rezultaty odzwierciedlają stopień, w jakim dane dopasowują się do struktury implikowanej przez wybrany algorytm ("łańcuch" lub zwarta "chmurka")

Jako metoda hierarchiczna wykorzystana zostanie aglomeracyjna Metoda minimalnej wariancji Warda. W metodzie tej optymalizuje się odległości między każdym obiektem a środkiem skupienia, do którego ten obiekt należy. Choć jest traktowana jako bardzo efektywna, zmierza do tworzenia skupień o zbliżonej (niewielkiej) liczebności. Jej główną zaletą jest to, że powstałe skupienia nie łączą się w łańcuchy.

[3] Wierzchoń S., Kłopotek M., "Algorytmy analizy skupień", Wydawnictwo WNT, Warszawa 2015
[4] Walesiak M., Gatnar E., "Statystyczna analiza danych z wykorzystaniem programu R", Wydawnictwo Naukowe PWN, Warszawa 2012

Opis danych i wybór zmiennych

Dane do projektu zostały pobrane ze strony [Human Development Reports](http://hdr.undp.org/en), na której można znaleźć raporty o rozwoju społecznym państw z całego świata. Chodzi tu nie tylko o wzrost gospodarczy, ale także o rozwój ludzi poprzez poprawę warunków życia oraz aktywne uczestnictwo w różnych procesach kształtujących ich życie. Ponieważ na omawianej stonie znajdują się dane, które są brane pod uwagę przy obliczaniu wskaźnika rozwoju społecznego (HDI), dlatego też jako zestaw cech do niniejszego opracowania zostało wybranych 10 zmiennych, które mają wpływ na kształtowanie się wspomnianego wyżej wskaźnika. Zmienne te pochodzą z 2015 roku. Analizie zostanie podanych 28 obiektów, którymi będą kraje członkowskie Unii Europejskiej. Jako cztery dodatkowe cechy diagnostyczne wybrano zmienne X_{11}, X_{12}, X_{13} i X_{14}, które zostały pobrane ze zbioru danych dotyczących sytuacji politycznej oraz życia społecznego, dostępnych na stronie [kaggle](https://www.kaggle.com/roshansharma/europe-datasets#pollution_2016.csv). Te zmienne pochodzą z 2016 roku, lecz na przestrzeni jednego roku nie nastąpiły znaczące zmiany, więc podjęto decyzję o uwzględnieniu obu grup zmiennych.

Na początku wybrano zestaw 12 zmiennych diagnostycznych:

X_1 - Oczekiwana długość życia (Life expectancy)
X_2 - Średnia liczba lat w szkole (Mean years of schooling)
X_3 - Dochód narodowy brutto na jedną osobę (Gross national income GNI per capita)
X_4 - Procentowy udział kobiet w parlamencie (Share of seats in parliament (% held by women))
X_5 - Liczba ludności (Total population (millions))
X_6 - Wydatki na publiczną opiekę zdrowotną jako procent z produktu krajowego brutto (Public health expenditure (% of GDP))
X_7 - Odsetek osób pracujących (powyżej 15 roku życia) (Employment to population ratio (% ages 15 an older))
X_8 - Odsetek osób mających dostęp do Internetu (Internet users)
X_9 - Wskaźnik aktywności zawodowej kobiet (Labour force participation rate (% ages 15 and older) Female)
X_{10} - Procent zarejestrowanego zanieczyszczenia (Percentage of reported Pollution)
X_{11} - Procent przestępstw (Percentage of Crime)
X_{12} - Procent osób z wysoką satysfakcją z wykonywanej pracy (Percentage of high job satisfaction)

Dla początkowego zbioru danych obliczono statystyki opisowe:

Tab. 1. Podstawowe statystyki opisowe

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
count	28.000000	28.000000	28.000000	28.000000	28.000000	28.000000	28.000000	28.000000	28.000000	28.000000	28.000000	28.000000
mean	79.457143	11.660714	33083.714286	26.100000	18.039286	6.442857	52.471429	79.017857	51.532143	12.996429	11.075000	27.142857
std	2.926055	0.966605	10714.241419	9.888227	23.194194	1.862581	5.316282	11.373721	5.307022	5.416195	4.488555	7.176114
min	73.500000	8.900000	16261.000000	10.100000	0.400000	3.300000	38.800000	55.800000	38.800000	4.600000	3.000000	14.000000
25%	77.375000	11.200000	25706.500000	19.375000	3.875000	4.975000	49.475000	71.000000	48.500000	9.275000	8.475000	22.975000
50%	80.750000	11.800000	29479.500000	25.250000	9.150000	6.350000	52.850000	79.650000	52.250000	13.150000	10.350000	26.900000
75%	81.300000	12.300000	41834.500000	35.200000	17.550000	7.750000	56.500000	87.800000	54.775000	15.225000	13.575000	29.600000
max	83.300000	13.300000	62471.000000	43.600000	80.700000	10.000000	59.900000	97.300000	60.900000	30.200000	25.000000	44.400000

Na podstawie wyznaczonych statystyk opisowych można już zauważyć zmienne charakteryzujące się małą zmiennością (poniżej 10%). Są to na pewno zmienne X_1 i X_2, o czym świadczy średnia (mean) oraz odchylenie standardowe (std), dlatego też zmienne te zostaną usunięte ze zbioru danych. Przeciwna sytuacja ma miejsce dla zmiennej X_5, której odchylenie standardowe jest większe od średniej, co świadczy o bardzo dużym zróżnicowaniu obserwacji. Pozostałe zmienne mają współczynnik zmienności przyjmujący zadowalające wartości.

Kolejnym krokiem do redukcji wstępnej liczby zmiennych jest analiza współczynników korelacji liniowej Pearsona. W przypadkach, w których korelacja będzie wynosiła więcej niż |0,9| jedna ze zmiennych będzie odrzucana, ponieważ zmienne mocno skorelowane niosą bardzo podobne informacje (współliniowość), więc nie ma potrzeby uwzględniać ich razem.

Tab. 2. Macierz korelacji

	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
X3	1.000000	0.616418	0.166371	0.641737	0.458486	0.789821	0.405529	-0.128558	0.158289	0.421783
X4	0.616418	1.000000	0.267281	0.719782	0.253989	0.549869	0.449043	-0.224871	0.104543	0.315581
X5	0.166371	0.267281	1.000000	0.352317	-0.065725	0.053193	-0.069522	0.141186	0.343516	-0.285449
X6	0.641737	0.719782	0.352317	1.000000	0.296140	0.612046	0.271130	-0.088941	0.237986	0.317965
X7	0.458486	0.253989	-0.065725	0.296140	1.000000	0.625455	0.781295	-0.244936	0.069534	0.598214
X8	0.789821	0.549869	0.053193	0.612046	0.625455	1.000000	0.585453	-0.243647	-0.046879	0.553342
X9	0.405529	0.449043	-0.069522	0.271130	0.781295	0.585453	1.000000	-0.534409	-0.042396	0.512468
X10	-0.128558	-0.224871	0.141186	-0.088941	-0.244936	-0.243647	-0.534409	1.000000	0.220809	-0.396417
X11	0.158289	0.104543	0.343516	0.237986	0.069534	-0.046879	-0.042396	0.220809	1.000000	-0.424570
X12	0.421783	0.315581	-0.285449	0.317965	0.598214	0.553342	0.512468	-0.396417	-0.424570	1.000000

Z macierzy korelacji wynika, że najsłabiej skorelowane z pozostałymi zmiennymi są X_5, X_{10} i X_{11}, więc odrzucam te zmienne i ponownie wypisuję macierz korelacji:

Tab. 3. Macierz korelacji po usunięciu zmiennych

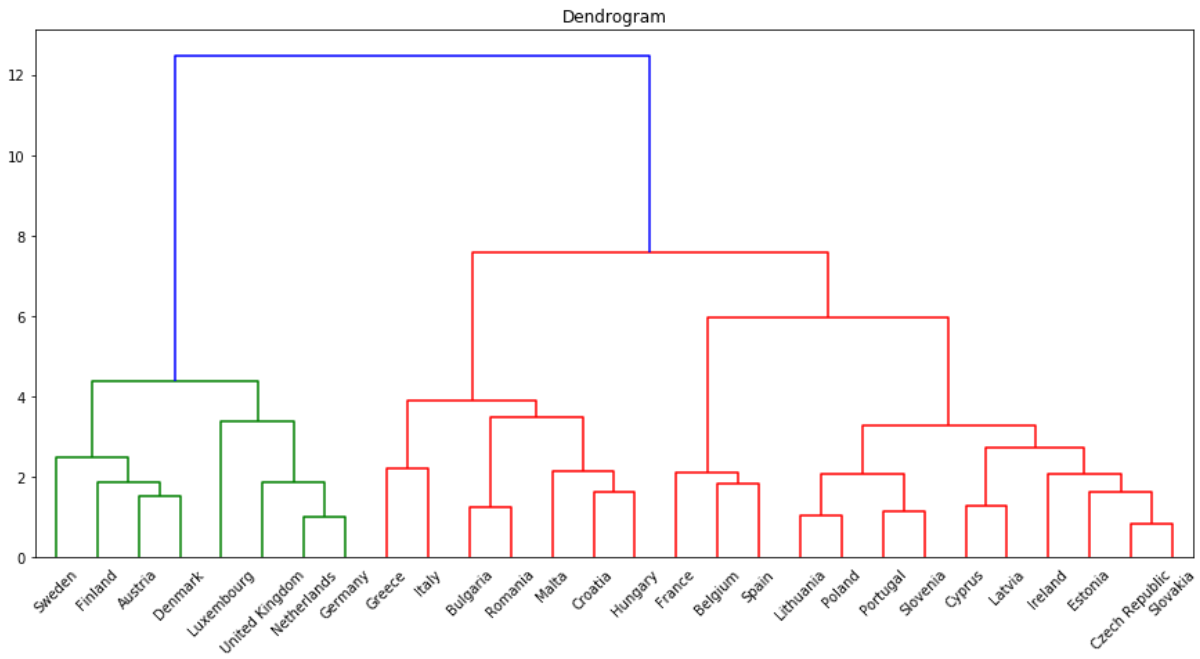
	X3	X4	X6	X7	X8	X9	X12
X3	1.000000	0.616418	0.641737	0.458486	0.789821	0.405529	0.421783
X4	0.616418	1.000000	0.719782	0.253989	0.549869	0.449043	0.315581
X6	0.641737	0.719782	1.000000	0.296140	0.612046	0.271130	0.317965
X7	0.458486	0.253989	0.296140	1.000000	0.625455	0.781295	0.598214
X8	0.789821	0.549869	0.612046	0.625455	1.000000	0.585453	0.553342
X9	0.405529	0.449043	0.271130	0.781295	0.585453	1.000000	0.512468
X12	0.421783	0.315581	0.317965	0.598214	0.553342	0.512468	1.000000

Tym razem najsłabsza korelacja występuje pomiędzy zmienną X_4 a zmienną X_7, jednak zdecydowano o pozostawieniu obu zmiennych.

Grupowanie hierarchiczne – metoda Warda

W grupowaniu hierarchicznym do wyznaczenia macierzy odległości użyto Euklidesową miarę odległości, która wynosi pierwiastek z sumy kwadratów różnic pomiędzy jedną zmienną dla różnych obiektów. Jest to w praktyce najczęściej stosowana miara odległości.

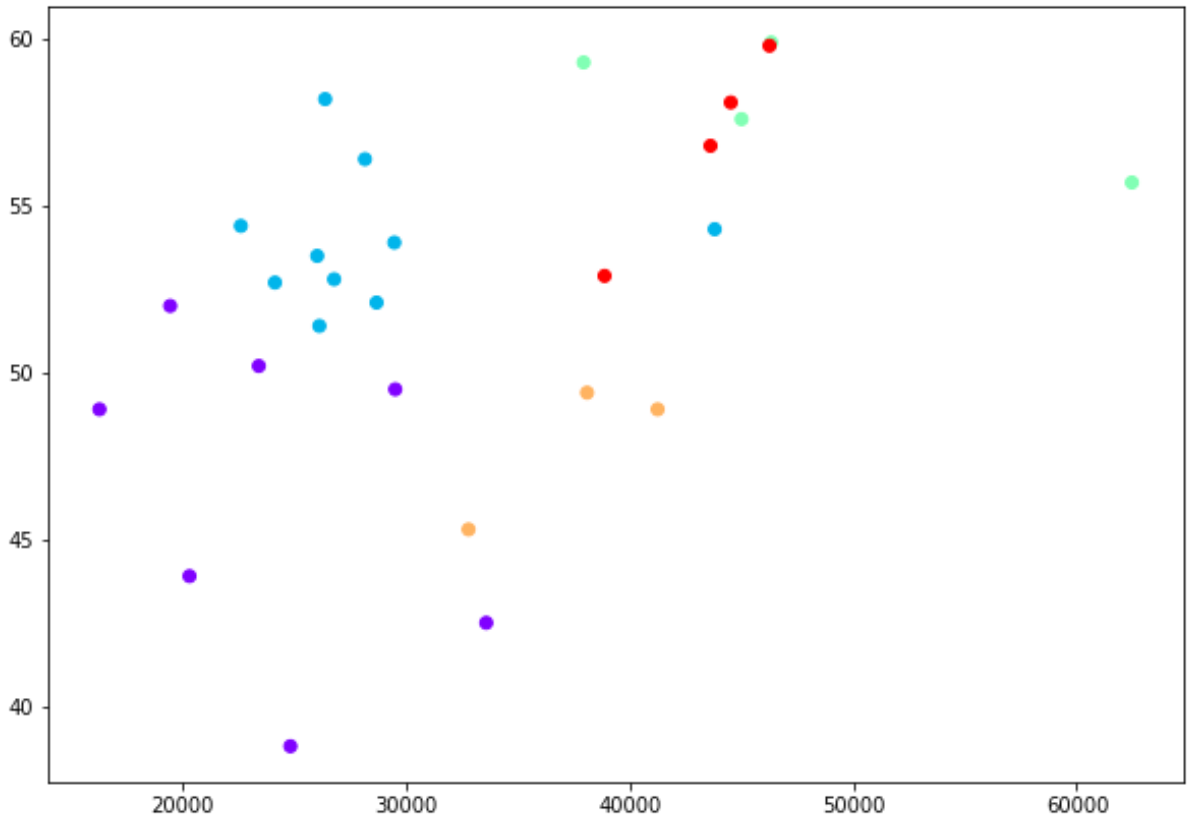
W pierwszym kroku dokonano standaryzacji zmiennych, a następnie za pomocą metody *dendrogram* z biblioteki *scipy.cluster* narysowano dendrogram.



Rys. 1. Dendrogram

Wnioski

Analizując otrzymany wykres można dojść do wniosku, że odpowiednią liczbą klastrów będzie 5. Dla wybranej liczby klastrów można narysować wykres rozrzutu, będzie przedstawiał poszczególne grupy w zależności od dwóch zmiennych – X_3 oraz X_6 (X_3 - Dochód narodowy brutto na jedną osobę, X_6 - Wydatki na publiczną opiekę zdrowotną jako procent z produktu krajowego brutto).



Rys. 2. Wykres rozrzutu dla grup odczytanych na podstawie dendrogramu

Punkty oznaczone na fioletowo wskazują na kraje, które charakteryzują się niskim dochodem narodowym brutto per capita oraz niskimi i średnimi wydatkami na publiczną opieke zdrowotną. Grupa oznaczona na pomarańczowo wskazuje na kraje o podobnych wydatkach na publiczną opieke zdrowotną, jednak o trochę wyższym dochodzie narodowym brutto na osobę. Najwyższe wydatki na oba czynniki ponoszą państwa należące do grupy punktów czerwonych oraz zielonych – można przypuszczać, że są to najlepiej rozwinięte kraje.

Po podziale na 5 grup otrzymano nastepujące grupy:

- Austria, Dania, Finlandia, Szwecja – kolor czerwony

- Holandia, Niemcy, Wielka Brytania, Luksemburg – kolor zielony
- Belgia, Francja, Hiszpania – pomarańczowy kolor
- Bułgaria, Chorwacja, Grecja, Malta, Rumunia, Węgry, Włochy – fioletowy kolor
- Cypr, Czechy, Estonia, Irlandia, Litwa, Łotwa, Polska, Portugalia, Słowacja, Słowenia – niebieski kolor

Podział taki zdaje się być sensowny, gdyż w obrębie grup pozostały państwa, które są na zbliżonym poziomie rozwoju gospodarczego oraz społecznego. Łączy je również bliskość geograficzna, która umożliwia im ściślejszą współpracę. Polska, co nie może dziwić jest w grupie państw słabiej rozwiniętych, w których pomimo że wydatki na opiekę zdrowotną są dość wysokie, to kraje te charakteryzują się wciąż niskim dochodem narodowym brutto na osobę.

Bibliografia

- Krzyśko M. i in., “Systemy uczace się”, Wydawnictwo Naukowo-Techniczne, Warszawa 2008
- Petrov O. i in, „Introduction to data mining”, Wydawnictwo AGH, Kraków 2019
- Walesiak M., Gatnar E., “Statystyczna analiza danych z wykorzystaniem programu R”, Wydawnictwo Naukowe PWN, Warszawa 2012
- Wierzchoń S., Kłopotek M., “Algorytmy analizy skupień”, Wydawnictwo WNT, Warszawa 2015