UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU
Wydział Matematyki i Informatyki
Zakład Statystyki Matematycznej i Analizy Danych

Kierunek studiów: Analiza i Przetwarzanie Danych

Nr albumu: 414342

Kacper Misiun

Analiza poziomu rozwoju społeczno-gospodarczego gmin w Województwie Zachodniopomorskim w latach 2016-2019 metodami analizy przestrzennej

Analysis of the level of socio-economic development of communes in the West Pomeranian Voivodeship in 2016-2019 using spatial analysis methods

Praca magisterska napisana

pod kierunkiem

prof. UAM dra hab. Łukasza Smagi

Poznań, 2021

Streszczenie

Praca ma na celu przedstawienie zróżnicowania przestrzennego poziomu rozwoju gmin w Województwie Zachodniopomorskim w latach 2016-2019. Przedstawiona została konstrukcja syntetycznego wskaźnika poziomu rozwoju społeczno-gospodarczego oraz jego aspektów społeczno-gospodarczych. Następnie wykorzystano metody analizy skupień (K-średnich oraz DBSCAN) w celu grupowania obserwacji ze względu na poziom rozwoju społeczno-gospodarczego. Korzystając ze stworzonych dzięki analizie skupień klas przeprowadzono klasyfikacje za pomocą metody lasów losowych oraz boostingu gradientowego (*xgboost*), dzięki czemu zaobserwowano, które z czynników oraz aspektów poziomu rozwoju wykazują się największą istotnością. Kolejnym krokiem było zastosowanie jednoczynnikowej analizy wariancji ANOVA oraz testów post-hoc do sprawdzenia czy występują istotne różnice między poszczególnymi poziomami rozwoju. Przeprowadzono parametryczną procedurę jak i nieparametryczną (test Kruskala-Wallisa). Sporządzono mapy przedstawiające poziomy rozwoju gmin w województwie Zachodniopomorskim w latach 2016-2019. Badanie przeprowadzono z użyciem środowiska R.

Słowa kluczowe: analiza skupień, klasyfikacja, analiza wariancji, środowisko R

Abstract

The master thesis aims to present the spatial differentiation of the level of development of municipalities in the West Pomeranian Voivodeship in the years 2016-2019. The construction of a synthetic indicator of the level of socio-economic development and its aspects. Then, cluster analysis methods (K-means and DBSCAN) were used to group the observations due to the level of socio-economic development. Using the classes created by the cluster analysis, classifications were carried out using the random forest method and gradient boosting (*xgboost*) to observe which factors and aspects of the level of development showed the greatest significance. The next step was to use one-way ANOVA and post-hoc tests to see if there were significant differences between the different levels of development. A parametric and non-parametric procedure (Kruskal-Wallis test) was performed. Maps were drawn up showing the levels of development of municipalities in the West Pomeranian Voivodeship in 2016-2019. The study was conducted using the R environment.

Keywords: cluster analysis, classification, analysis of variance, R environment

Spis treści

Streszczenie 3
Abstract 4

1 Badanie rozwoju społeczno-gospodarczego 7
1.1 Rozwój społeczno gospodarczy 7
1.2 Opis danych 9
1.3 Konstrukcja syntetycznego miernika poziomu rozwoju społeczno-gospodarczego 12

2 Metody analizy danych 15
2.1 Analiza skupień - pojęcie, K-średnich, DBSCAN 15
2.2 Klasyfikacja - definicja, lasy losowe, xgboost 19
2.3 Analiza wariancji - jednoczynnikowa ANOVA, test Kruskala-Wallisa, testy post-hoc 22

3 Analiza Danych 25
3.1 Wyliczenie syntetycznego miernika rozwoju 25
3.2 Analiza skupień 28
3.3 Klasyfikacja 39
3.4 Analiza wariancji 52
3.5 Wyniki dla ostatecznych grup -mapy 60

4 Podsumowanie 67

Bibliografia 69

Rozdział 1

Badanie rozwoju społeczno-gospodarczego

1.1 Rozwój społeczno gospodarczy

Rozwój społeczno-gospodarczy to wieloaspektowe i złożone zjawisko, które nie jest łatwe do zdefiniowania. Wzrost gospodarczy jest często utożsamiany z rozwojem, zwłaszcza w literaturze anglosaskiej, gdzie pojęcie *growth* i *development* jest stosowane zamiennie i odnosi się do wzrostu wskaźników makroekonomicznych (Borys (1999)). Zazwyczaj jednak w literaturze przedmiotu rozdziela się te dwa pojęcia z powodów, które trafnie ujął Z. Hull (Hull (2007)):

> „. . . to, co stanowi o treści rozwoju i określa charakter i formy jego realizacji pojmowane jest odmiennie: jedni sprowadzają go do przyrostu ilości materialnych dóbr i usług, wzrostu poziomu konsumpcji, ułatwień codziennego życia, zwiększania sfery wolności w życiu społecznym itp., natomiast inni kładą nacisk na kształtowanie nowych jakości życia, wypracowywanie nowych form i struktur życia społecznego, nowych form współbycia i współżycia w przyrodzie. . . ”

Dlatego też przyjęło się, że wzrost oznacza ilościowe zwiększenie produkcji dóbr i usług w danej ustalonej przestrzeni i czasie. Natomiast rozwój jest pojęciem szerszym

i oprócz zmian ilościowych obejmuje zmiany jakościowe. Rozwój społeczno-gospodarczy zawiera w sobie zmiany strukturalne oraz wiążące się z nimi zmiany instytucji i stosunków ekonomicznych (Pająk i in. (2016)).

Tak rozumiany rozwój wymaga odpowiednich, uszczegółowionych metod oraz narzędzi badawczych. Nie ma jednak konsensusu wśród badaczy zajmujących się tym zagadnieniem i proponowane są różne mierniki, dzięki którym można badać rozwój społeczno-gospodarczy. W badaniu przyjęto (za Opałło (1972)) ogólny podział mierników na dwa rodzaje:

- podstawowe, bazujące na wartościach bezwzględnych zjawiska i procesy ekonomiczne oraz społeczne (np. liczba ludności, dochód z podatku PIT w danej jednostce przestrzennej)
- relatywne, opisujące stosunek wartości bezwzględnych względem siebie w danej jednostce przestrzennej (np. dochody z podatku PIT na 1 mieszkańca).

Zgodnie z tą klasyfikacją w badaniu korzystano z mierników relatywnych. Poszczególne mierniki wykorzystane w badaniu opisano w podrozdziale 1.2.

Rozwój społeczno-gospodarczy istnieje w danej jednostce przestrzennej, w określonym regionie. W niniejszej pracy, kierowano się podziałem terytorialno-administracyjnym zgodnie z podejściem A. Hettnera wyrażającym się w poglądzie, że (za Pająk i in. (2016)):

> . . . Określone przez naturę regiony nie istnieją, ponieważ podział taki zawsze wynika z ustaleń człowieka, jego aktywności gospodarczej, a nie uwarunkowań geograficznych

Dlatego w badaniu kierowano się podziałem terytorialno-administracyjnym kraju. Zgodnie z powyższym za podstawowy region badań rozwoju przyjęto gminę. Natomiast obszar badania ograniczono do województwa Zachodniopomorskiego.

Zdecydowano się na województwo Zachodniopomorskie, ponieważ zgodnie z istniejącymi opracowaniami jest to województwo o dużej dysproporcji rozwoju. Na poziomie gminnym rozwojem społeczno-gospodarczym w województwie Zachodniopomorskim

zajmował się Czyżycki (2006) w pracy *Rozwój społeczno-gospodarczy gmin województwa Zachodniopomorskiego*. Konkluzją z jego pracy jest teza o dużym wewnętrznym zróżnicowaniu gmin w województwie. Za przyczynę tego stanu rzeczy stwierdzono brak opracowania zrównoważonej strategii rozwoju gmin oraz problem niegospodarności terenów po państwowych gospodarstwach rolnych (tak zwanych PGRów) położonych zwłaszcza w centralnej części województwa. Badanie te przeprowadzono ponad dekadę temu. Bardziej aktualne opracowanie dotyczy zróżnicowania poziomu ekonomicznego powiatów w województwie Zachodniopomorskim jak podają Adamczyk i in. (2012). W badaniu stwierdzono między innymi, że w latach 2004-2010 zwiększyła się skala rozwarstwienia ekonomicznego poziomu rozwoju gmin oraz że relatywnie lepszą sytuacją gospodarczą charakteryzują się powiaty w aglomeracji szczecińskiej.

Szersza jest literatura przedmiotu poziomu rozwoju społeczno-gospodarczego gmin na poziomie krajowym. Zagadnienie to poruszano między innymi w pracach Perdał (2018) oraz Churski i in. (2020). Jednak badanie na poziomie krajowym a regionalnym-wojewódzkim może doprowadzić do innych wniosków z uwagi na inne odniesienie do wzorca rozwoju. Niniejszą pracę różni od powyższych także modyfikacja zmiennych diagnostycznych. Przedstawiono różne metody wyznaczania liczby skupień/grup za pomocą analizy skupień, zastosowano algorytmy klasyfikacji (lasy losowe, xgboost) oraz analizy wariancji za pomocą języka R.

1.2 Opis danych

Dane statystyczne użyte do konstrukcji wskaźników pochodzą z Banku Danych Lokalnych Głównego Urzędu Statystycznego. Dane te pobrano dla gmin w województwie zachodniopomorskim w latach 2016-2019. Zostały wybrane tak, żeby powstałe na ich podstawie wskaźniki opisywały wybrane aspekty rozwoju społeczno-gospodarczego:

- kapitał ludzki (KL)
- kapitał społeczny (KS)
- kapitał materialny (KM)
- kapitał finansowy (KF)

- innowacje techniczne i organizacyjne (IT)

Dane wyrażone w liczbach absolutnych (bezwzględnych) z BDL GUS przekształcono na liczby względne poprzez uwzględnienie liczby ludności, powierzchni gminy lub wyrażenie danej zmiennej w % (np. Dochody z podatku PIT na **1 mieszkańca** [zł]). W przypadku, gdy w Banku Danych Lokalnych była możliwość wyboru kwartału, w którym zarejestrowano daną zmienną to zawsze wybierano stan na IV kwartał - 31 grudnia. Natomiast w przypadku, gdy dane dla gminy były podzielone na wieś i miasto to decydowano się na średnią z tych dwóch wartości.

Podstawą do wyboru zmiennych był artykuł Doktora Perdała "Zastosowanie analizy skupień i lasów losowych w klasyfikacji gmin w Polsce na skali poziomu rozwoju społeczno-gospodarczego" (Perdał, 2018). Dokonano modyfikacji kierując się dostępnością wszystkich zmiennych w badanym okresie, biorąc pod uwagę kryteria merytoryczne oraz dążąc do redundancji danych - w tym celu posłużono się macierzą korelacji (Rysunek 1.1).

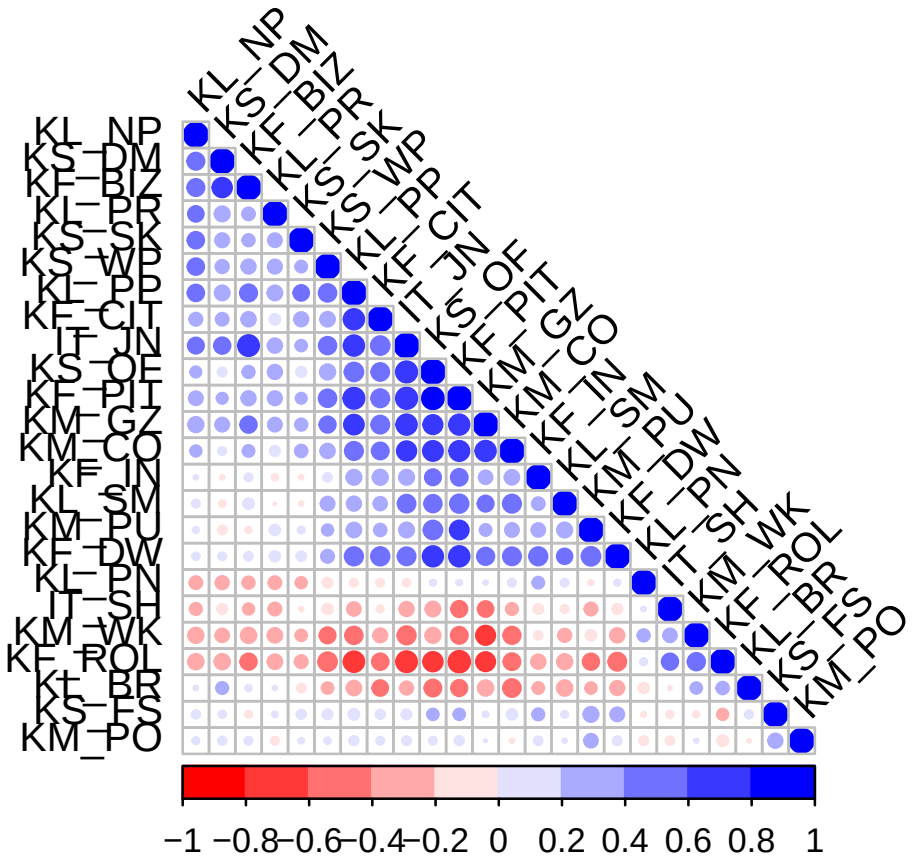
Zdecydowano się na współczynnik skolaryzacji netto dla szkół podstawowych mając na uwadze ciągłość danych, ponieważ reforma edukacyjna z 2019 roku zlikwidowała gimnazja. Dodano wskaźnik podmioty gospodarcze w sekcjach J-N (usługi, specjaliści, informatyka) na 1000 mieszkańców, w celu zwiększenia liczby wskaźników stanowiących o innowacyjności danej jednostki przestrzennej. W ten sposób powstały 24 zmienne diagnostyczne, które wykorzystano przy konstrukcji syntetycznego wskaźnika poziomu rozwoju dla lat 2016-2019 (Tabela 1.1).

Dane obejmują lata 2016-2019, w tym czasie doszło do kilku ważnych zmian administracyjnych w województwie zachodniopomorskim, które miały wpływ podczas procesu przetwarzania danych. W badanym okresie liczba gmin z 114 zmniejszyła się do 113. Stało się tak, ponieważ w 2019 roku zlikwidowano gminę Ostrowice. Przyczyną było bankructwo gminy- jest to pierwszy tego typu przypadek w Polsce. Teren gminy Ostrowice został włączony do dwóch sąsiednich gmin, Drawska Pomorskiego i gminy Złocieniec.(*Gmina Ostrowice zostanie zniesiona - Ministerstwo Spraw Wewnętrznych i Administracji - Portal Gov.pl* 2021) Z powodu problemów finansowych w gminie Ostrowice

Tablica 1.1: *Zmienne diagnostyczne, wykorzystane do konstrukcji syntetycznego wskaźnika rozwoju*

Skrot	Nazwa.wskaznika	Typ
KL_NP	ludność w wieku nieprod. na 100 w wieku prod.	Destymulanta
KL_PN	przyrost naturalny na 1000 ludności w ‰	Stymulanta
KL_SM	saldo migracji ogółem na 1000 ludności	Stymulanta
KL_PR	przychodnie na 10 tys. ludności	Stymulanta
KL_BR	udział bezrobotnych zarejestrowanych w liczbie ludności w wieku produkcyjnym	Destymulanta
KL_PP	pracujący na 1000 osób w wieku produkcyjnym	Stymulanta
KS_FS	fundacje, stowarz., organizacje na 1000 lub 10 tys. osób	Stymulanta
KS_OF	osoby fizyczne prow. działalność gos. na 1000 ludności	Stymulanta
KS_WP	udział przedstawicieli władz publicznych, wyższych urzędników, kierowników oraz specjalistów w ogóle radnych [%]	Stymulanta
KS_SK	współczynnik skolaryzacji netto szkoły podstawowe	Stymulanta
KS_DM	liczba dodatków mieszkaniowych na 1000 mieszkańców - wskaźnik ubóstwa	Destymulanta
KM_GZ	udział osób korzystających z instalacji gazowej w ogóle populacji [%]	Stymulanta
KM_PO	obszary prawnie chronione jako % powierzchni gminy – wskaźnik posiadanych walorów środowiska	Stymulanta
KM_WK	różnica pomiędzy odsetkiem ludności korzystającej z wodociągu i z kanalizacji wg lokalizacji	Destymulanta
KM_PU	przeciętna powierzchnia użytkowa mieszkania na 1 osobę	Stymulanta
KM_CO	% mieszkań posiadających centralne ogrzewanie	Stymulanta
KF_IN	wydatki majątkowe inwestycyjne budżetów gmin i miast na prawach powiatu na 1 mieszkańca [zł/os.]	Stymulanta
KF_PIT	dochody z podatku PIT na 1 mieszkańca [zł]	Stymulanta
KF_CIT	dochody z podatku CIT na 1 mieszkańca [zł]	Stymulanta
KF_ROL	dochody z podatku rolnego na 1 mieszkańca [zł]	Stymulanta
KF_DW	dochody własne per capita [zł]	Stymulanta
KF_BIZ	instytucje otoczenia biznesu na 10 tys. podmiotów gospodarki narodowej	Stymulanta
IT_SH	spółki handlowe z udziałem kapitału zagranicznego na 1000 podmiotów gospodarczych	Stymulanta
IT_JN	podmioty gospodarcze w sekcjach J-N (usługi, specjalisci, informatyka) na 1000 mieszkańców	Stymulanta

* Zródło: Opracowanie własne

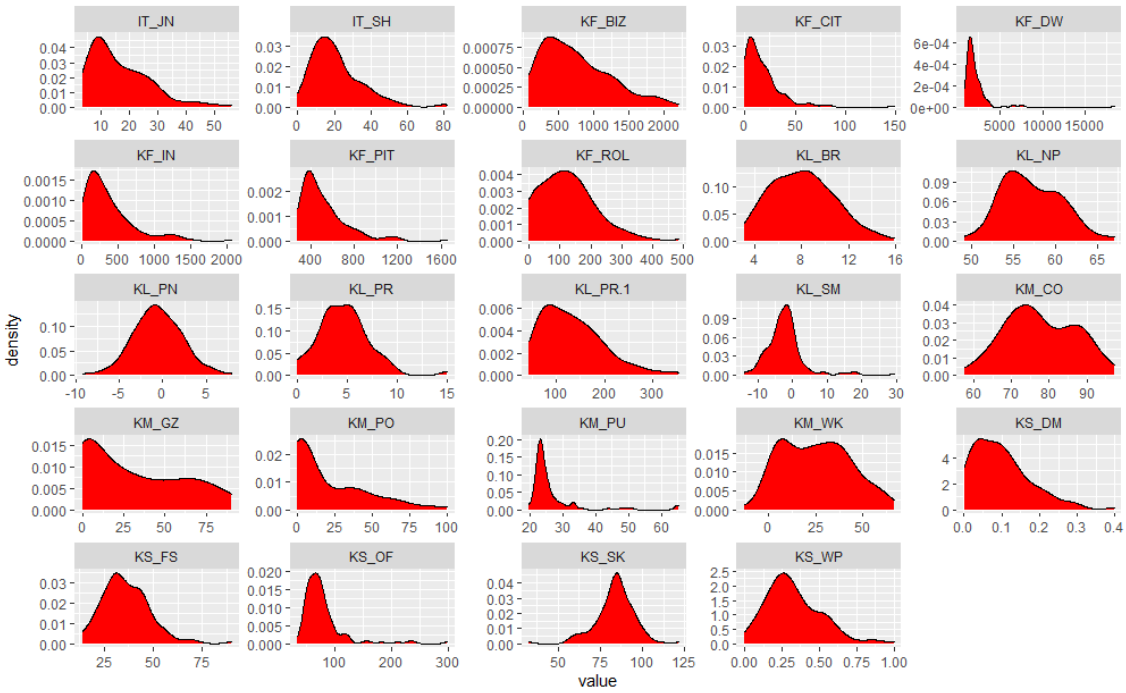


Rysunek 1.1: *Korelogram czynników diagnostycznych*

w latach 2017-2018 rada gminy nie istniała, a do zarządzania gminą został powołany przedstawiciel rządu. Dlatego zdecydowano, że wskaźnik udziału przedstawicieli władz publicznych, wyższych urzędników, kierowników oraz specjalistów w ogóle radnych dla gminy Ostrowice w latach 2017-2018 będzie wynosił 0. W badanym okresie gmina Mielno w 2017 przekształciła się z gminy wiejskiej na wiejsko-miejską, nie wpłynęło to jednak znacząco na proces badawczy.

1.3 Konstrukcja syntetycznego miernika poziomu rozwoju społeczno-gospodarczego

Do konstrukcji syntetycznych wskaźników rozwoju konieczne było przekształcenie zmiennych tak by były porównywalne. Dokonując wyboru metody pozwalającej na ujednolicenie rzędu wielkości zmiennych oraz pozbycia się mian kierowano się wykresem



Rysunek 1.2: *Wykres estymatorów gęstości zmiennych diagnostycznych (2016)*

gęstości zmiennych. Ustalono, że większość zmiennych diagnostycznych cechuje się asymetrią prawostronną rozkładu (Rysunek 1.2).

Mając na uwadze to że większość zmiennych nie charakteryzowało się rozkładem normalnym, zdecydowano się na metodę unitaryzacji zerowanej (za Churski i in. (2020)). Metoda ta wymaga następującej formuły dla:

- stymulanty

$$Z_{ij} = \frac{X_{ij} - \min_i X_{ij}}{\max_i X_{ij} - \min_i X_{ij}}$$

- destymulanty

$$Z_{ij} = \frac{\max_i X_{ij} - X_{ij}}{\max_i X_{ij} - \min_i X_{ij}}$$

Polega to na tym, że każda minimalna wartość danej zmiennej jest przekształcana na 0, maksymalna wartość jest przekształcana na 1, a każda inna wartość jest przekształcana

na ułamek z zakresu od 0 do 1. Dzięki zastosowaniu normalizacji zerowanej mamy pewność że wszystkie poddane normalizacji zmienne (także te ujemne) będą nieujemne i znajdowały się w tym samym przedziale (od 0 do 1).

Następnym krokiem jest konstrukcja syntetycznego wskaźnika rozwoju na bazie zunitaryzowanych wskaźników z wykorzystaniem metody wzorca rozwoju. W metodzie tej transformowano miarę niepodobieństwa Braya-Curtisa na miarę podobieństwa do wzorca, gdzie jako wzorzec przyjęto hipotetyczną jednostkę przestrzenną, która przyjmowała wartość maksymalną dla wszystkich zmiennych diagnostycznych (za Perdał (2018))

$$d_{kj}^{BC} = 1 - \frac{\sum_{j=1}^{m} |Z_{ij} - Z_{kj}|}{\sum_{j=1}^{m} |Z_{ij} + Z_{kj}|}$$

gdzie: Z_{ij} - zunitaryzowana wartość wskaźnika j dla gminy i

k - gmina "wzorzec"

$j \in \{1, 2, \ldots, m\}$ - numer wskaźnika

Wskaźniki syntetyczne przyjmują wartość z przedziału [0,1], im większa wartość wskaźnika syntetycznego rozwoju tym wyższy poziom rozwoju społeczno-gospodarczego. Syntetyczny wskaźnik poziomu rozwoju obejmuje wszystkie 24 zmienne diagnostyczne (Patrz Tabela 1.1). Wskaźniki syntetyczne wyróżnionych aspektów poziomu rozwoju społeczno-gospodarczego (tj. kapitał ludzki, materialny), obejmują określone w Tabeli 1.1 wskaźniki diagnostyczne.

Rozdział 2

Metody analizy danych

2.1 Analiza skupień - pojęcie, K-średnich, DBSCAN

Analiza skupień jest ważnym elementem statystycznej analizy danych. Pojęcie to pierwszy raz użyto w 1939 roku przez Roberta Choate Tryona w pracy *cluster analysis* (za Migdał-Najman i in. (2013)). Analiza skupień polega na grupowaniu danych w grupy zwane tez skupieniami, w zależności od tego jak ściśle są ze sobą powiązane. Dąży się do tego żeby dane w jednym skupieniu charakteryzowały się wysokim poziomem podobieństwa, natomiast dane między skupieniami miały minimalny poziom podobieństwa. Obecnie istnieje wiele metod i technik analizy skupień, w niniejszej pracy opisano metodę K-średnich oraz metodę DBSCAN.

Metoda K-średnich należy do jednych z najczęściej stosowanych algorytmów analizy skupień. W grupowaniu K-średnich każde skupienie jest reprezentowane przez swój środek (tj. centroid), który odpowiada średniej z obserwacji przypisanych do skupienia. Istnieje kilka wariantów algorytmu dla tej metody, do najbardziej znanych należą (za Morissette i in. (2013)):

- Forgy/Lloyd
- MacQueen
- Hartigan-Wong

Standardowym oraz użytym w niniejszym badaniu algorytmem jest algorytm Hartigana-Wonga (1979), który definiuje całkowitą zmienność w obrębie skupienia jako sumę kwadratów odległości euklidesowych między elementami, a odpowiadającym im centroidem (za *K-means Cluster Analysis · UC Business Analytics R Programming Guide* (2021)).

$$W(C_k) = \sum_{X_i \in C_k} (X_i - \mu_k)^2$$

gdzie: X_i - obserwacja należąca do skupienia C_k

μ_k - średnia wartość obserwacji przypisanych do skupienia C_k

Każda obserwacja (X_i) zostaje przypisana do określonego skupienia w taki sposób, że suma kwadratów odległości obserwacji do przypisywanego im centroidu (μ_k) jest minimalna. Całkowita suma kwadratów w obrębie skupienia jest definiowana wzorem:

$$\sum_{k=1}^{K} W(C_k) = \sum_{k=1}^{K} \sum_{X_i \in C_k} (X_i - \mu_k)^2$$

Im mniejsza wartość całkowitej sumy kwadratów tym lepiej.

Pierwszym krokiem przy wykorzystaniu algorytmu K-średnich jest wskazanie liczby skupień (k), które zostaną wygenerowane w ostatecznym rozwiązaniu. Następnie algorytm losowo wybiera k obserwacji ze zbioru danych, które mają służyć jako początkowe centroidy dla skupień. Każda z pozostałych obserwacji jest przypisywana do najbliższego centroidu, gdzie najbliższy jest definiowany za pomocą odległości euklidesowej:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}$$

między obserwacją a średnią skupień (przypisanie obserwacji do skupień). W kolejnym kroku algorytm oblicza średnią wartość dla każdego skupienia (koryguje centroidy). Wszystkie obserwacje są ponownie przypisywane przy użyciu skorygowanych centroidów. Następnie powtarza się przypisanie obserwacji do skupień i korygowanie centroidów do momentu, aż przypisania do skupień przestaną się zmieniać. Oznacza to, że skupienia utworzone w bieżącej iteracji są takie same, jak te uzyskane w poprzedniej iteracji.

Pierwszym krokiem w powyższym algorytmie jest wskazanie przez analityka liczby klastrów/skupień. Przy wyborze optymalnej liczby klastrów oprócz wiedzy na temat charakteru danych warto kierować się jedną z metod służących do doboru liczby skupień. W badaniu podczas wyboru optymalnej liczby skupień porównywano wyniki uzyskane trzema metodami:

- Metodą Łokcia (*Elbow Method*)
- Metodą Średniego Zarysu (*Average Silhouette Method*)
- Statystyczną Metodą Luki (*Gap Statistic Method*)

Metoda Łokcia polega na obliczeniu dla każdego k całkowitej sumy kwadratów wewnątrz klastra, a następnie przedstawienie na wykresie. Można wyrazić ją wzorem:

$$min(\sum_{k=1}^{K} W(C_k))$$

Optymalna liczba klas występuje w miejscu zgięcia (kolana), czyli w gdzie k i liczba klas $k >$ charakteryzuje się niską rozpiętością całkowitej sumy kwadratów wewnątrz skupienia.

Metoda Średniego Zarysu polega na wyliczeniu zarysu dla każdej obserwacji a następnie współczynnika zarysu na podstawie średniej wartości zarysu ($\overline{S}_k$). Zarys wyraża się wzorem:

$$S_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

gdzie: a_i - średnia miara niepodobieństwa miedzy obserwacją a wszystkimi obserwacjami wewnątrz tego samego skupienia

b_i - najmniejsza średnia odległość do najbliższego skupienia, do którego nie należy a_i

Współczynnik zarysu wyraża się wzorem:

$$SC = max_k \overline{S}_k$$

gdzie: $\overline{S}_k$ - średnia wartość zarysu dla całego zbioru o k skupieniach

Statystyczną Metodą Luki została opracowana w 2001 roku przez R. Tibshirani, G. Walther, i T. Hastie (Tibshirani i in. (2001)). Wyraża się wzorem:

$$Gap_n(k) = E_n^* log(W_k) - log(W_k)$$

gdzie: E_n^* - oznacza wartość oczekiwaną przy wielkości próby n z rozkładu referencyjnego i jest zdefiniowane za pomocą metody bootstrap (B) poprzez wygenerowanie B kopii referencyjnych zbiorów danych i przez obliczenie średniej

W_k - macierz kowariancji wewnątrzgrupowych wtedy, gdy każda zmienna oryginalna zastąpiona została zmienną wygenerowaną z rozkładu jednostajnego nad odcinkiem, który jest rozstępem próby zmiennej oryginalnej (za Korzeniewski (2014))

Wybieramy liczbę skupień jako najmniejsze k takie, że:

$$Gap(k) \geq Gap(k+1) - S_{k+1.}$$

DBSCAN (ang. *Density-Based Spatial Clustering of Applications with Noise*) należy do algorytmów gęstościowych, czyli takich które biorą pod uwagę fakt, że skupienia są gęstymi grupami obserwacji. Uproszczając chodzi o to, że jeśli konkretna obserwacja należy do skupienia to powinna znajdować się w pobliżu wielu innych obserwacji w tym skupieniu.

Najpierw wybieramy dwa parametry, maksymalny promień sąsiedztwa epsilon (ϵ) i liczbę minimalnej liczby obserwacji, dla których możemy powiedzieć że tworzy skupienie (*minPts). Następnie zaczynamy od wybrania dowolnej obserwacji w naszym zbiorze danych. Jeśli w odległości ϵ od tej obserwacji znajduje się więcej obserwacji niż wyznaczona minimalna liczba obserwacji dla skupienia, to uważamy obserwacje za część skupienia Następnie rozszerzamy to skupienie, sprawdzając wszystkie nowe obserwacje i sprawdzając, czy one również mają więcej niż ustaloną minimalną liczbę obserwacji w odległości ϵ. Gdy nie ma obserwacji spełniających nasze warunki wewnątrz skupienia to wybieramy nową dowolną obserwacje i powtarzamy algorytm. Obserwacje, które w odległości ϵ mają mniej niż minimalną liczbę obserwacji i nie są częścią żadnego

skupienia uważamy za szum i nie przydzielamy do żadnego skupienia. Sąsiedztwo ϵ obserwacji p, oznaczone przez N wyraża się wzorem (za Ester i in. (1996)):

$$N_{Eps}(p) = \{q \in D | dist(p,q) \leq \epsilon\}$$

2.2 Klasyfikacja - definicja, lasy losowe, xgboost

Klasyfikacja to proces określenia przydziału obserwacji do jednej z predefiniowanych klas. Każdą obserwację można oznaczyć parą (x, y) gdzie x reprezentuje zbiór mierzonych cech, które mają mieć wpływ na etykietę klasy y. Problem klasyfikacji występuje, gdy trenujemy model w celu przewidzenia klasy y do jakiej należy obserwacja na podstawie cech obserwacji x. Elementem systemu klasyfikacji nadzorowanej jest próba ucząca (ang. *learning sample*) składająca się z n niezależnych par zmiennych $(x_1, y_1), \ldots, (x_n, y_n)$ (za Krzyśko i in. (2008)). Drugim elementem jest klasyfikator, który jest funkcją określona na przestrzeni wartości cech o wartościach w zbiorze etykiet skonstruowaną na bazie próby uczącej. Klasyfikator jest funkcją:

$$d : X \to Y$$

Gdy obserwujemy nowy wektor X to prognozą etykiety Y jest $d(X)$ Kolejnym elementem jest ocena skuteczności działania klasyfikatora. Ocenę wartości błędu klasyfikatora przeprowadza się na zbiorze testowych przykładów, dla których rzeczywista przynależność do klas jest znana i które nie są częścią zbioru uczącego. Miarą jakości klasyfikatora d jest rzeczywisty poziom błędu (za Krzyśko i in. (2008)):

$$e(d) = P(d(X) \neq Y)$$

Błąd ten jest jednak niemożliwy do zmierzenia, ponieważ zazwyczaj rozkład pary (X, Y) nie jest znany. Dlatego opracowano metody oszacowania rzeczywistego poziomu błędu,

które oparte są obliczaniu jakości klasyfikatora $\hat{d}$ za pomocą warunkowego prawdopodobieństwie błędu, które ma postać:

$$e(\hat{d}) = P(\hat{d}(X)) \neq Y|L_n)$$

gdzie: $e(\hat{d})$ aktualny poziom błędu

L_n - losowa próba ucząca

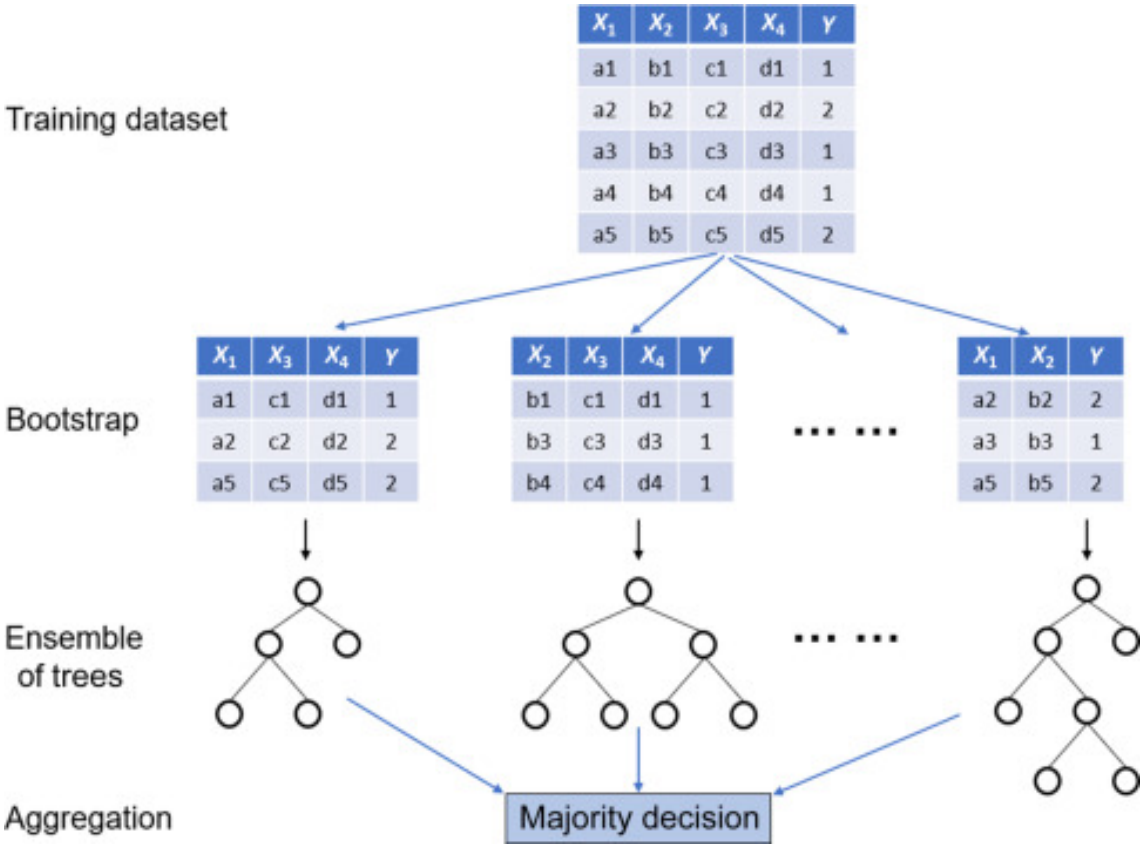
$L_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ jest ciągiem niezależnych par losowych o identycznym rozkładzie prawdopodobieństwa takim, jak rozkład pary (X, Y)

Lasy losowe to metoda klasyfikacji nadzorowanej opracowana przez Leo Breimana w 2001 roku (za Perdał (2018)). W założeniu metoda ma na celu modyfikacje klasycznych drzew klasyfikacyjnych poprzez dodanie przypadkowości za pomocą metody bootstrap. Bootstrap to metoda która symuluje tworzenie wielu nowych zestawów danych poprzez pobieranie próbek z oryginalnych danych ze zwracaniem (patrz Rysunek 2.1)

Dzięki tej modyfikacji lasy losowe w przeciwieństwie do drzew klasyfikacyjnych nie są tak wrażliwe na zmiany w danych uczących, a więc są bardziej odporne na błąd przeuczenia modelu. Algorytm lasów losowych można przedstawić następująco (za Grzybowska i in. (2015)):

1. Tworzymy N pseudo-próbek takich że $S \in \{S_1, S_2, \ldots, S_n\}$ poprzez pobieranie próbek ze zwracaniem
2. Dla każdego $i \in \{1, 2, \ldots, N\}$ dokonywana jest losowa selekcja atrybutów ze zbioru uczącego X_* i budowany jest klasyfikator $h_i = Learn(S_i; X_*)$
3. Wynik uzyskiwany jest za pomocą reguły większościowej (ang. *majority~vote*)

Metoda XGBoost czyli boosting gradientowy została opracowana w 1999 roku przez J. Fridmana (za Grzybowska i in. (2015)). Konstruowana jest rodzina drzew, tak że następne drzewo tworzone jest na podstawie poprzedniego w kierunku wektora gradientu, dzięki czemu minimalizowana jest funkcja straty. Algorytm XGBoost konstruowany w następujący sposób (za Grzybowska i in. (2015)):



Rysunek 2.1: *Schemat algorytmu lasów losowych. Źródło:https://www.sciencedirect.com/topics/engineering/rando forest*

Niech $Y = Learn(h(X))$ oraz $b_m(X)$ będą predyktorami Y

$$h(X) = \sum_{m=1}^{M} b_m(x) = \sum_{m=1}^{M} \beta_m b(x;\theta_m)$$

1. Podstawiamy:

$$h_o(x) \leftarrow 0$$

2. Dla $m \in \{1,2,\ldots,M\}$ i funkcji straty $L()$:

$$r(x,y) \leftarrow -\frac{d}{dh} L(h_{m-1}(x), y)$$

$$b_m \leftarrow argmin_b \sum_{(x,y)} (b(x) - r(x,y))^2$$

$$h_m(X) \leftarrow h_{m-1}(x) + v * b_m(x)$$

W algorytmie tym wykorzystuje się fakt, że do zmiany $h(x)$ na $h(x) + vb(x)$ dla małych wartości v możemy użyć wyrażenia:

$$\sum_{(x,y)} L(h_{m-1}(x) + b_m(x), y) \approx \sum_{(x,y)} L(h_{m-1}(x), y) + \sum_{(x,y)} \frac{d}{dh} L(h_m - 1(x)y) b_m(x)$$

2.3 Analiza wariancji - jednoczynnikowa ANOVA, test Kruskala-Wallisa, testy post-hoc

Analiza wariancji to statystyczne narzędzie, które ma na celu porównanie wartości średnich zmiennej zależnej w analizowanych grupach wydzielonych ze względu na wartość zmiennych niezależnych. Twórca analizy wariancji R. Fisher (za *Hypothesis Testing - Analysis of Variance (ANOVA)* (2021)) wykazał, że większa różnica w średniej wartości zmiennej zależnej pomiędzy porównywanymi grupami przekłada się na większą wariancję międzygrupową. Idea testu polega na rozłożeniu wariancji całkowitej na sumę wariancji międzygrupowych i wariancję wewnątrzgrupową (patrz Rysunek 2.2).

Hipoteza zerowa w teście analizy wariancji ma postać:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_L,$$

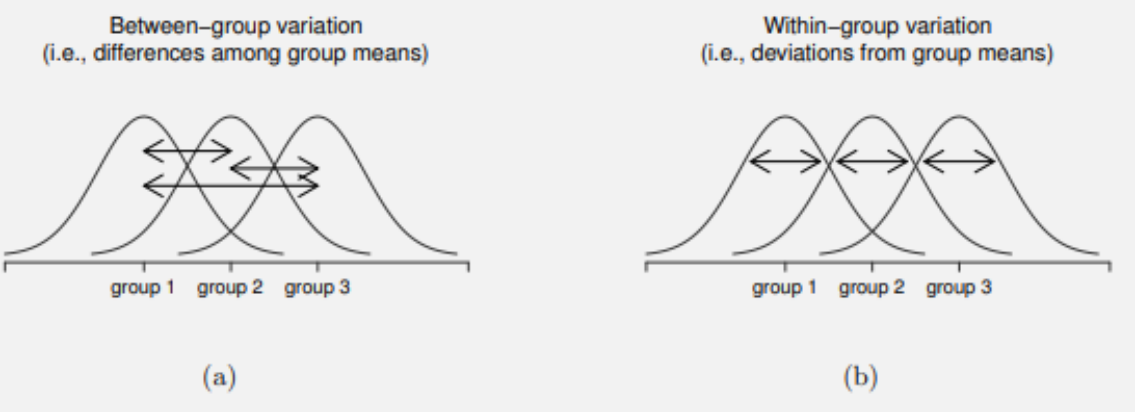
gdzie μ_l jest wartością oczekiwaną w l-tej grupie, $l = 1, 2, \ldots, L$. Hipoteza alternatywna jest zaprzeczeniem hipotezy zerowej.

Jednoczynnikowa analiza wariancji (jednoczynnikowa *ANOVA*) rozpatruje czy istnieją różnice w średnich wartościach ciągłej zmiennej zależnej w podgrupach wydzielonych ze względu na wartość zmiennej niezależnej. Statystyka testowa przyjmuje postać (za *Hypothesis Testing - Analysis of Variance (ANOVA)* (2021)):

$$F = \frac{[\sum_{j=1}^{K} (\overline{x}_j - \overline{x})^2 n_j] / (k-1)}{[\sum_{i=1}^{n_j} \sum_{j=1}^{K} (x_i j - \overline{x}_j)^2] / (n-k)} \sim F(k-1, n-k)$$

gdzie: n_j - liczebności grup

$\overline{x}_j$ - średnia w j-tej grupie



Rysunek 2.2: *Wariancje międzygrupowa (a) i wewnątrzgrupowa (b) Źródło:https://stats.libretexts.org*

$\overline{x}$ - średnia ogólna zmiennej zależnej

Powyższy rozkład jest prawdziwy przy prawdziwości hipotezy zerowej.

W przypadku przyjęcia hipotezy alternatywnej dla dokładniejszej analizy niezbędne są dodatkowe testy- tak zwane testy post-hoc (porównania wielokrotne). W badaniu w celu stwierdzenia jakie grupy istotnie różnią się między sobą wykonano test HSD Tukeya.

Test Kruskala-Wallisa został opracowany w 1952 roku. Jest to test nieparametryczny (nie wymaga założeń odnośnie rozkładu populacji) będący alternatywą dla jednoczynnikowej analizy wariancji, która zakłada normalność cechy oraz homoskedastyczność wariancji. Zakłada jednak, że rozkłady w grupach mają te same kształty. Jest to test rangowy będący uogólnieniem testu Wilcoxona, obejmujący więcej niż dwie grupy, w którym testuje się równość median.

Statystyka testowa wyraża się wzorem (za *Test ANOVA Kruskala-Wallisa | Statystyka – Porady | Analizy | Opracowania | Obliczenia | Pomoc statystyczna* (2021)):

$$T = \frac{12}{n(n+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(n+1)$$

gdzie: R_i - suma rang w i-tej próbie

n_i - liczebność i-tej próby

n - łączna liczebność wszystkich grup

Rozdział 3

Analiza Danych

3.1 Wyliczenie syntetycznego miernika rozwoju

Analizę danych rozpoczęto od wczytania wstępnie oczyszczonych danych w formacie *xlsx* za pomocą pakietu *openxlsx* (Schauberger i in., 2021). Konstrukcje syntetycznego wskaźnika rozwoju rozpoczęto od wykonania funkcji unitaryzacji zerowanej (ang. *min-max normalization*) dla wskaźników będących stymulantami lub destymulantami, za pomocą następujących instrukcji.

```
#stymulanta
minmaxS <- function(x, na.rm = FALSE) {
  return((x- min(x)) /(max(x)-min(x)))
}
#destymulanta
minmaxD <- function(x, na.rm = FALSE) {
  return((max(x)- x) /(max(x)-min(x)))
}
```

Następnym krokiem było zastosowanie przygotowanej funkcji unitaryzacji zerowanej na danych dla odpowiednich wskaźników przy wykorzystaniu funkcji *as.data.frame*,

lapply z pakietu *base* (R Core Team, 2021) oraz *select* z pakietu *dplyr* (Wickham i in., 2021b).

```
library(dplyr)
#>
#> Dołączanie pakietu: 'dplyr'
#> Następujący obiekt został zakryty z 'package:kableExtra':
#>
#>     group_rows
#> Następujące obiekty zostały zakryte z 'package:stats':
#>
#>     filter, lag
#> Następujące obiekty zostały zakryte z 'package:base':
#>
#>     intersect, setdiff, setequal, union
# wybieram i normalizuje stymulanty
gus2016_normS <- as.data.frame(lapply(select(Gus_2016, 4:6, 8:12,
    14:15, 17:26), minmaxS))

# wybieram i normalizuje destymulanty
gus2016_normD <- as.data.frame(lapply(select(Gus_2016, 3, 7,
    13, 16), minmaxD))

# łącze i dodaje TERYT, nazwa gminy
gus2016_norm <- cbind(gus2016_normS, gus2016_normD)
```

Następnie za pomocą następujących instrukcji skonstruowano i zastosowano funkcję opartą na mierze niepodobieństwa Braya-Curtisa (patrz Podrozdział 1.3) wyliczającą syntetyczny wskaźnik poziomu rozwoju oraz jego aspekty (patrz Podrozdział 1.2).

Tablica 3.1: *Ramka danych dla wskaźnika syntetycznego poziomu rozwoju społeczno-gospodarczego i jego aspektów na przykładowych obserwacjach*

	Rozwoj	K_Ludzki	K_Spoleczny	K_Materialny	K_Finansowy	Innowacje
Białogard (1)	0.5271442	0.6075089	0.6007241	0.5272582	0.3664491	0.5127502
Białogard (2)	0.3540260	0.4667454	0.5075338	0.2969588	0.1483996	0.2475569
Karlino (3)	0.5041937	0.5668628	0.4540023	0.4760860	0.4708909	0.5907579
Tychowo (3)	0.3803655	0.4953640	0.5466270	0.2702022	0.2395358	0.1676706
Bierzwnik (2)	0.4312479	0.4574030	0.6481029	0.4503713	0.1981006	0.2673995
Choszczno (3)	0.5403547	0.5786519	0.5921272	0.6861926	0.3077595	0.4828032

* Zródło: Opracowanie własne

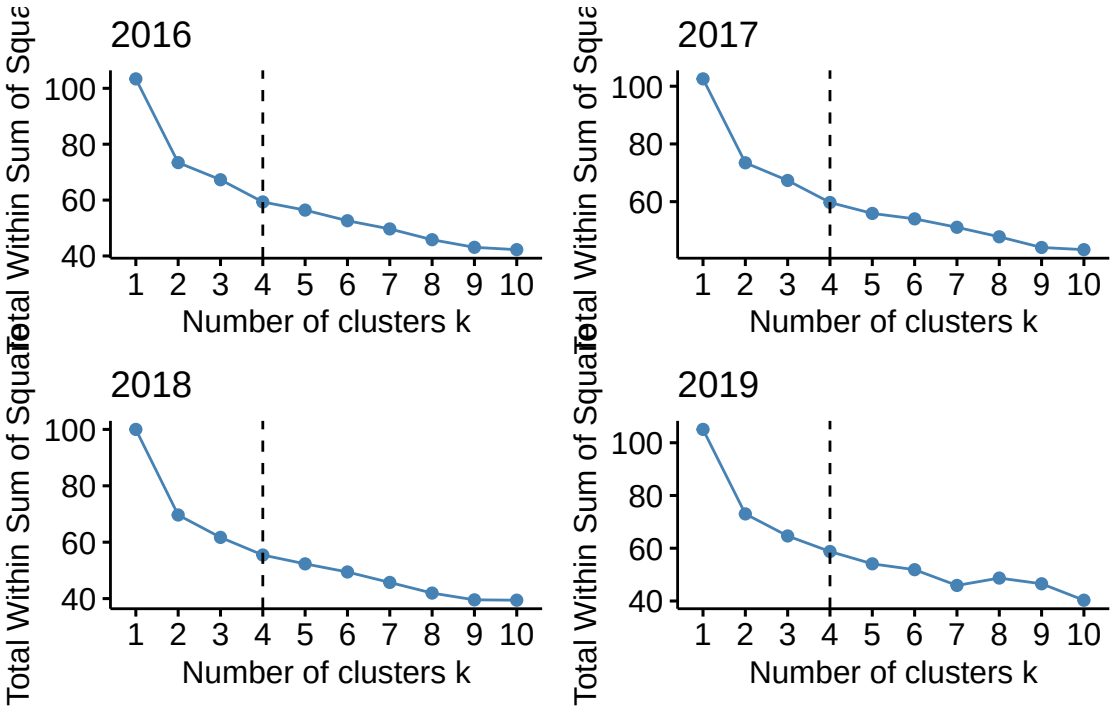
```
library(dplyr)

wskaznik <- function(x) {

  x_abs <- as.data.frame(lapply(x, function(y) abs(y - max(y))))

  x_sum <- as.data.frame(lapply(x, function(y) y + max(y)))

  return(1 - rowSums(x_abs) / rowSums(x_sum))

}


gus2016_norm <- gus2016_norm %>%

  select(order(colnames(gus2016_norm))) %>%

  mutate(Rozwoj = wskaznik(.),

         K_Ludzki = wskaznik(select(., 9:14)),

         K_Spoleczny = wskaznik(select(., 20:24)),

         K_Materialny = wskaznik(select(., 15:19)),

         K_Finansowy = wskaznik(select(., 3:8)),

         Innowacje = wskaznik(select(., 1:2)))
```

Powstała ramka danych zawierająca wskaźnik syntetycznego poziomu rozwoju społeczno-gospodarczego i jego aspektów dla gmin w województwie Zachodniopomorskim w 2016 roku przedstawiała się następująco (patrz Tablica 3.1). Analogicznie przeprowadzono unitaryzacje oraz obliczono syntetyczny wskaźnik rozwoju społeczno-gospodarczego oraz jego aspekty dla danych za rok 2017-2019.

3.2 Analiza skupień

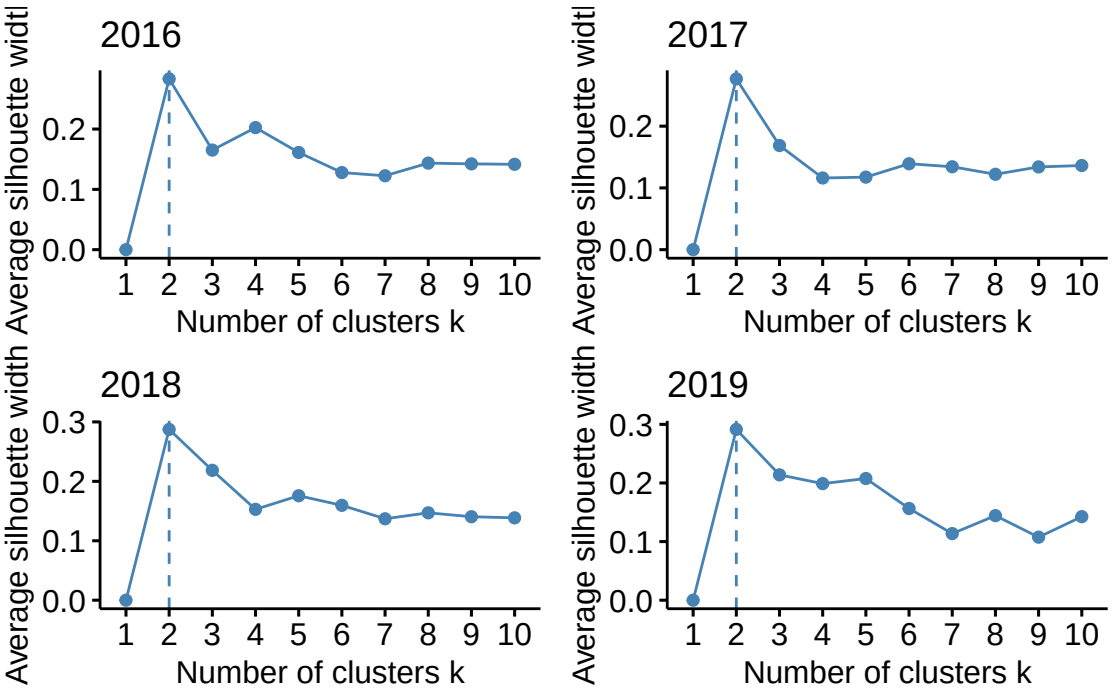
Analizę skupień metodą K-średnich rozpoczęto od znalezienia optymalnej liczby skupień. Skorzystano z pakietu *factoextra* (Kassambara i in. (2020)) i następującymi instrukcjami wykorzystując funkcję *fviz_nbclust*. Wykresy przedstawiające optymalną liczbę skupień wykonano dla każdego rocznika z przedziału 2016-2019 z pomocą pakietu *ggplot2* (Wickham i in. (2021a)) oraz połączono je za pomocą funkcji *grid.arrange* z pakietu *gridExtra* (Auguie (2017))

```
# szukanie optymalnej liczby skupień Elbow Method
set.seed(123)
p1 <- factoextra::fviz_nbclust(gus2016_norm, kmeans, method = "wss",
    k.max = 10) + ggplot2::geom_vline(xintercept = 4, linetype = 2) +
    ggplot2::ggtitle("2016")
p2 <- factoextra::fviz_nbclust(gus2017_norm, kmeans, method = "wss",
    k.max = 10) + ggplot2::geom_vline(xintercept = 4, linetype = 2) +
    ggplot2::ggtitle("2017")
p3 <- factoextra::fviz_nbclust(gus2018_norm, kmeans, method = "wss",
    k.max = 10) + ggplot2::geom_vline(xintercept = 4, linetype = 2) +
    ggplot2::ggtitle("2018")
p4 <- factoextra::fviz_nbclust(gus2019_norm, kmeans, method = "wss",
    k.max = 10) + ggplot2::geom_vline(xintercept = 4, linetype = 2) +
    ggplot2::ggtitle("2019")

gridExtra::grid.arrange(p1, p2, p3, p4)
```



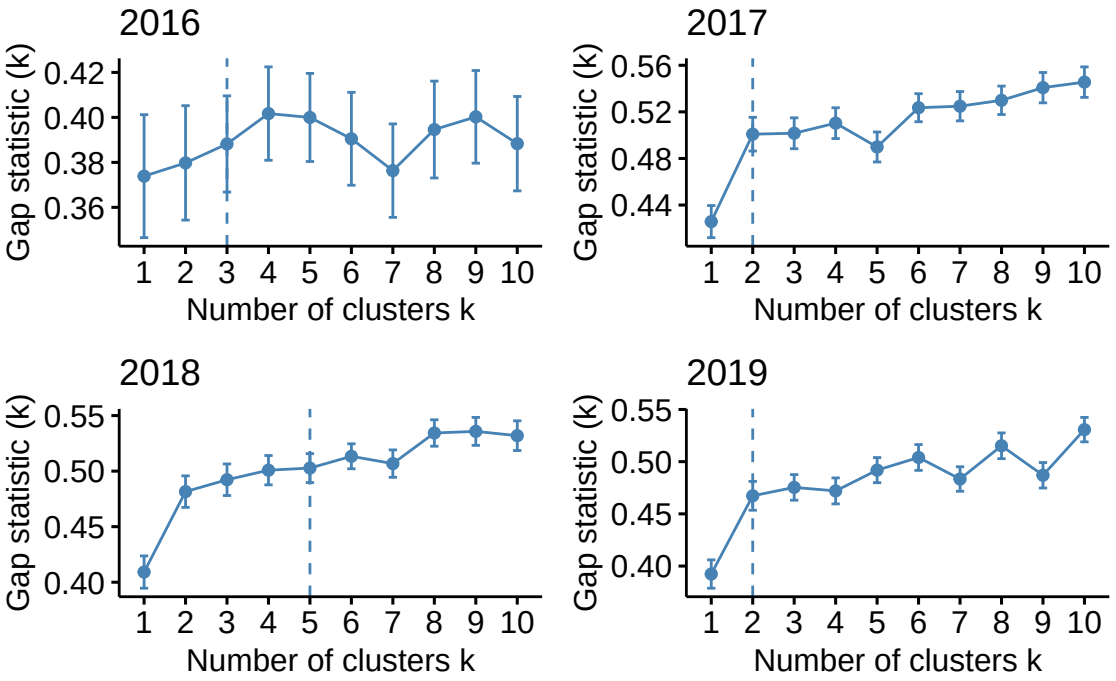
W przypadku metody Łokcia, optymalną liczbę skupień wyznacza miejsce zgięcia (patrz Podrozdział 2.1). W tym przypadku dla danych za rok 2016 optymalna liczba skupień (k) to 4. Metoda Łokcia dla innych roczników wyznaczała taką samą optymalną liczbę skupień (4).

```
# szukanie optymalnej liczby skupień Average Silhouette
# Method
p1 <- factoextra::fviz_nbclust(gus2016_norm, kmeans, method = "silhouette") +
    ggplot2::ggtitle("2016")  #idealnie 2, do 5 jest ok
p2 <- factoextra::fviz_nbclust(gus2017_norm, kmeans, method = "silhouette") +
    ggplot2::ggtitle("2017")  #idealnie 2, do 5 jest ok
p3 <- factoextra::fviz_nbclust(gus2018_norm, kmeans, method = "silhouette") +
    ggplot2::ggtitle("2018")  #idealnie 2, do 5 jest ok
p4 <- factoextra::fviz_nbclust(gus2019_norm, kmeans, method = "silhouette") +
    ggplot2::ggtitle("2019")

gridExtra::grid.arrange(p1, p2, p3, p4)
```



W przypadku metody średniego zarysu (Average Silhouette Method) im większa wartość średniego zarysu dla danej liczby skupień k tym bardziej jest ona optymalna. Jak widać optymalna liczba skupień we wszystkich latach to 2, jednak warto zauważyć, że średnia wartość zarysu dla 3 skupień jest niewiele mniejsza, zwłaszcza dla lat 2017-2019.

```
# szukanie optymalnej liczby skupień Gap Statistic Method
p1 <- factoextra::fviz_nbclust(gus2016_base, kmeans, method = "gap_stat") +
    ggplot2::ggtitle("2016")
p2 <- factoextra::fviz_nbclust(gus2017_norm, kmeans, method = "gap_stat") +
    ggplot2::ggtitle("2017")
p3 <- factoextra::fviz_nbclust(gus2018_norm, kmeans, method = "gap_stat") +
    ggplot2::ggtitle("2018")
p4 <- factoextra::fviz_nbclust(gus2019_norm, kmeans, method = "gap_stat") +
    ggplot2::ggtitle("2019")

gridExtra::grid.arrange(p1, p2, p3, p4)
```

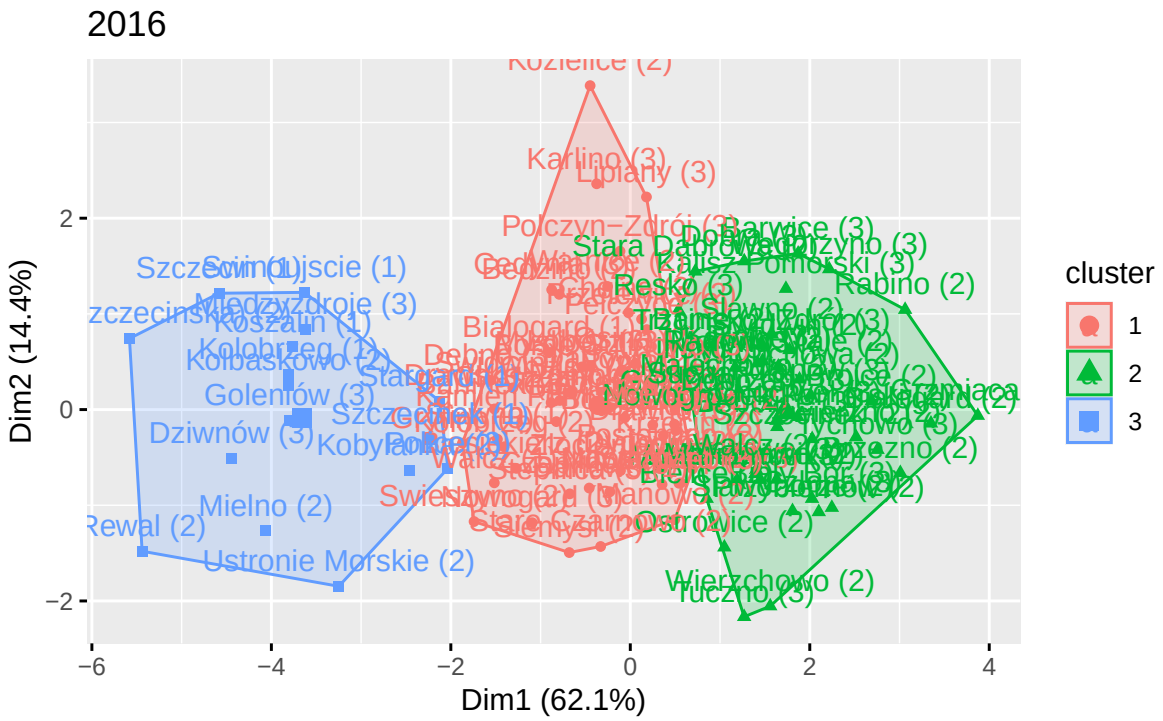


Dla statystycznej metody luki (Gap Statistic Method) optymalna liczba skupień k to taka, dla której wartość statystyki luki jest większa niż $k+1$ minus odchylenie standardowe. Co ciekawe dla różnych roczników statystyka ta przyjmuje różne wartości. Dla 2016 roku optymalna liczba skupień to 3, dla 2017, 2019 roku optymalną liczbą skupień to 2, natomiast dla 2018 roku optymalna liczba skupień to 5.

Ostatecznie kierując się powyższymi statystykami, sprawdzając jak rozkładają się obserwacje dla różnej liczby skupień i kierując się wiedzą merytoryczną z zakresu badań nad poziomem rozwoju oraz praktyką badawczą, sugerującą nieparzystą i ograniczoną liczbę skupień zdecydowano się na dalsze badania dla 3 skupień. Natomiast warto zaznaczyć, że liczby skupień 4 lub 5 wydają się też dość optymalne. W badaniach Pana Doktora Perdała (Perdał (2018)) za optymalna liczbę skupień określono 5.

Przydzielanie do skupień metodą K-średnich wykonano za pomocą następujących instrukcji korzystając z funkcji *kmeans* oraz *fviz_cluster* z pakietu *factoExtra* (**R-factoExtra**). W funkcji *fviz_cluster* obserwacje są reprezentowane przez punkty na wykresie, przy użyciu dwóch pierwszych składowych głównych.

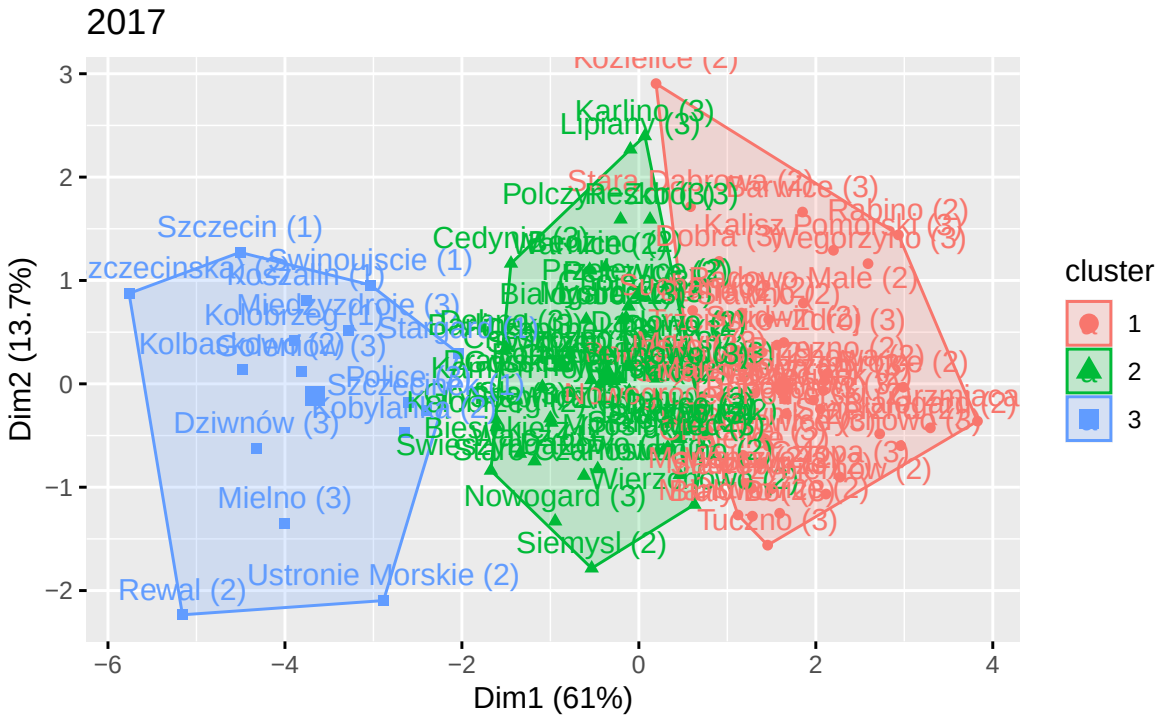
```
k1 <- kmeans(gus2016_base$Rozwoj, centers = 3, nstart = 10)
factoextra::fviz_cluster(k1, data = gus2016_base) + ggplot2::ggtitle("2016")
```



Interpretacja skupień dla obserwacji z 2016 roku przedstawia się następująco (dokładane przedstawienie poziomów rozwoju dla poszczególnych obserwacji w Podrozdziale (3.5)):

- skupienie o numeracji 1 przedstawia obserwacje o średnim poziomie rozwoju
- skupienie o numeracji 2 przedstawia obserwacje o niskim poziomie rozwoju
- skupienie o numeracji 3 przedstawia obserwacje o wysokim poziomie rozwoju

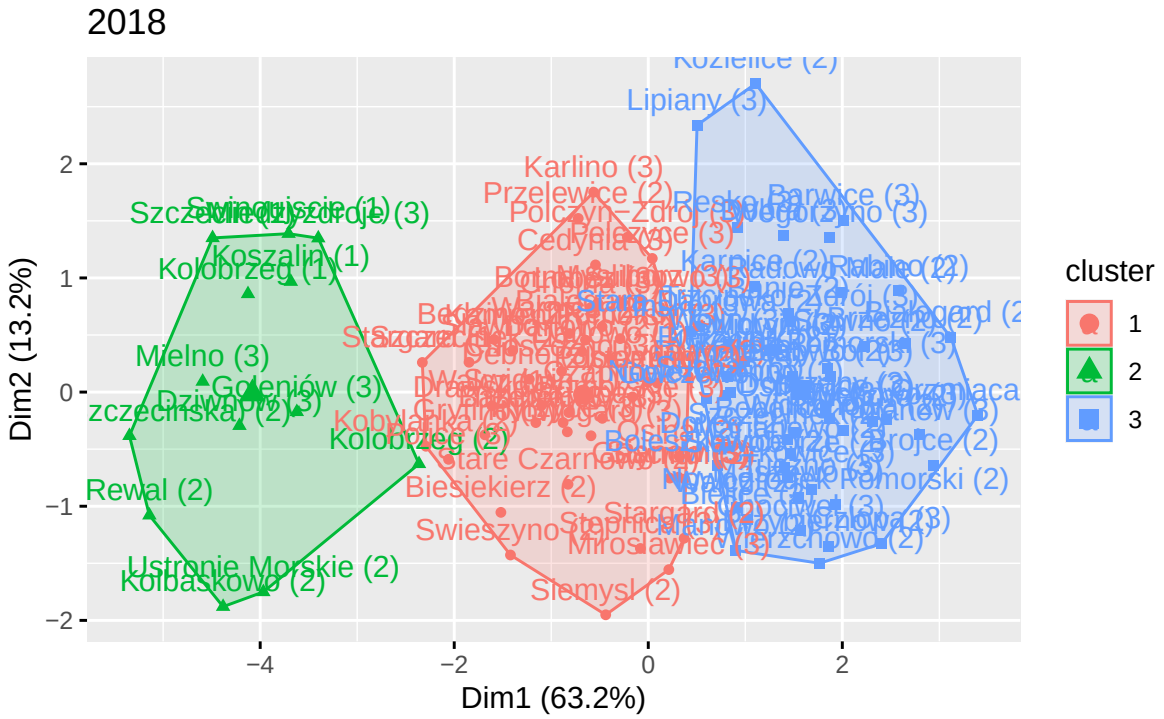
```
k2 <- kmeans(gus2017_base$Rozwoj, centers = 3, nstart = 10)
factoextra::fviz_cluster(k2, data = gus2017_base) + ggplot2::ggtitle("2017")
```



Interpretacja skupień dla obserwacji z 2017 roku przedstawia się następująco:

- skupienie o numeracji 1 przedstawia obserwacje o średnim poziomie rozwoju
- skupienie o numeracji 2 przedstawia obserwacje o niskim poziomie rozwoju
- skupienie o numeracji 3 przedstawia obserwacje o wysokim poziomie rozwoju

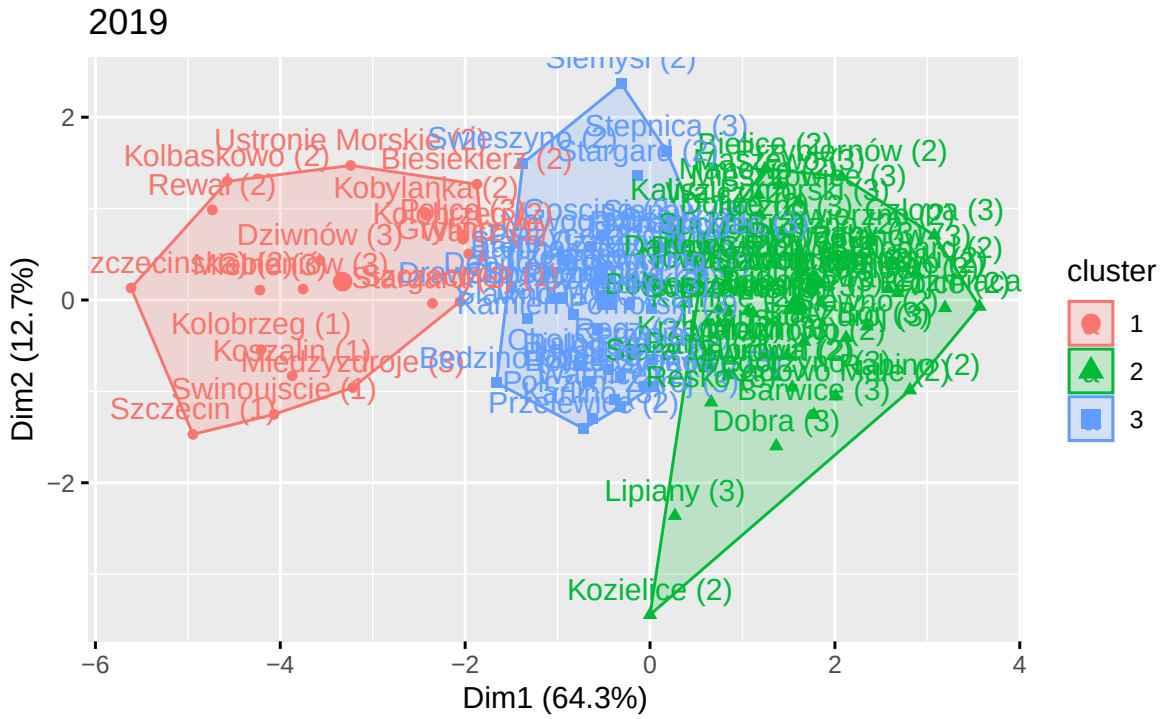
```
k3 <- kmeans(gus2018_base$Rozwoj, centers = 3, nstart = 10)

factoextra::fviz_cluster(k3, data = gus2018_base) + ggplot2::ggtitle("2018")
```



Interpretacja skupień dla obserwacji z 2018 roku przedstawia się następująco:

- skupienie o numeracji 1 przedstawia obserwacje o średnim poziomie rozwoju
- skupienie o numeracji 2 przedstawia obserwacje o wysokim poziomie rozwoju
- skupienie o numeracji 3 przedstawia obserwacje o niskim poziomie rozwoju

```
k4 <- kmeans(gus2019_base$Rozwoj, centers = 3, nstart = 10)

factoextra::fviz_cluster(k4, data = gus2019_base) + ggplot2::ggtitle("2019")
```



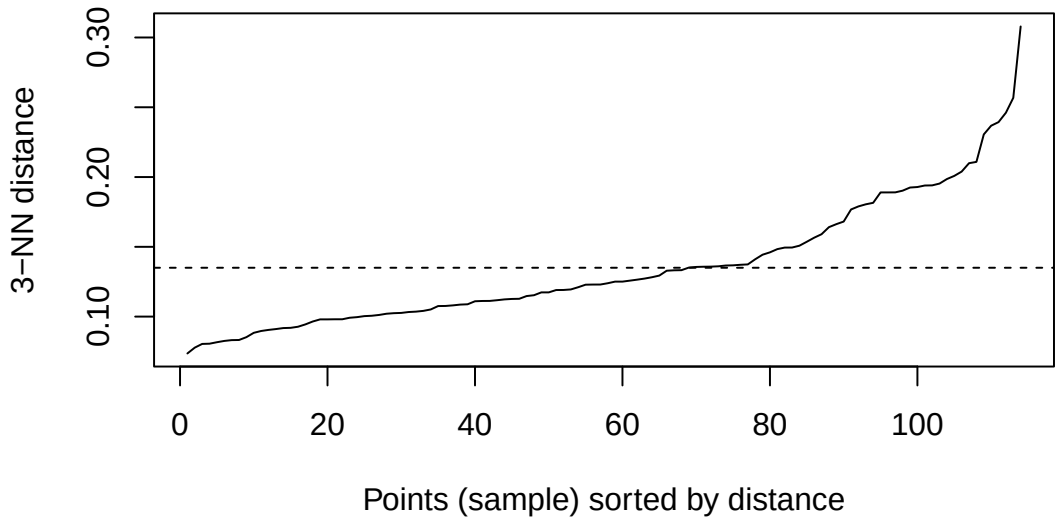
Interpretacja skupień dla obserwacji z 2019 roku przedstawia się następująco:

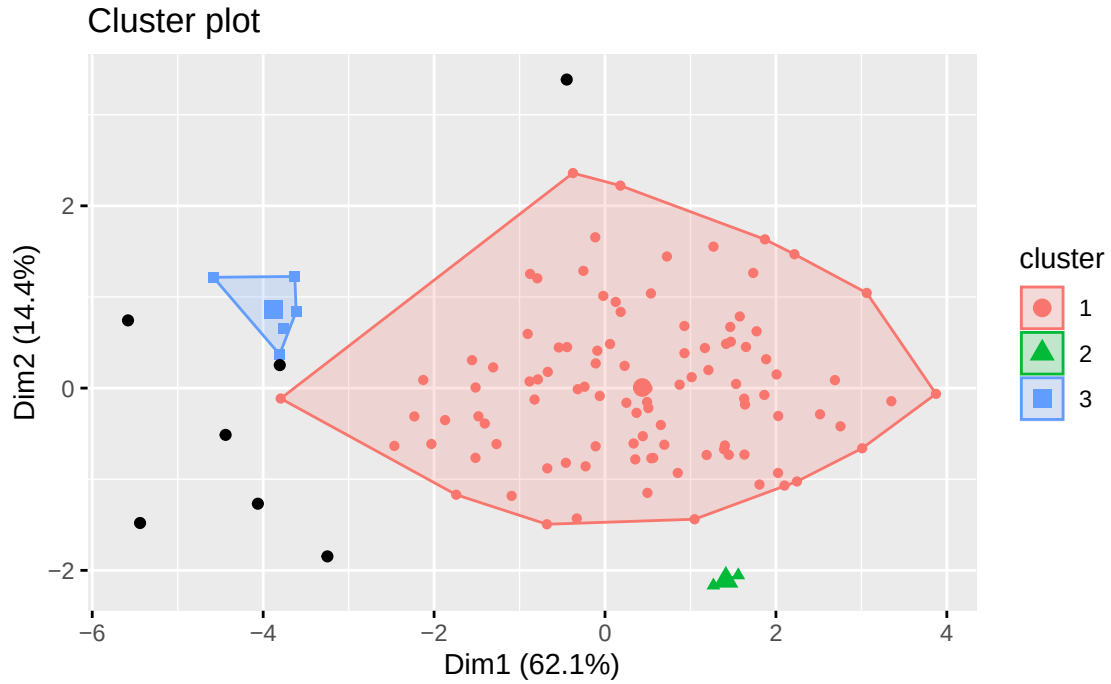
- skupienie o numeracji 1 przedstawia obserwacje o wysokim poziomie rozwoju
- skupienie o numeracji 2 przedstawia obserwacje o niskim poziomie rozwoju
- skupienie o numeracji 3 przedstawia obserwacje o średnim poziomie rozwoju

Jak widać na powyższych wykresach, liczba skupień 3 dla metody K-średnich wygląda zadowalająco. Także, metoda K-średnich dla podziału obserwacji ze względu na syntetyczny wskaźnik poziomu rozwoju wydaje się optymalna.

W celu porównania wyników grupowania skupień za pomocą innej metody niż K-średnich zdecydowano się na wykonanie analizy DBSCAN. Analizę skupień metodą DBSCAN wykonano za pomocą pakietu *dbscan* (Hahsler i in. (2021)). Rozpoczęto od oszacowania wartości ϵ za pomocą funkcji *kNNdistplot*, która polega na obliczeniu i wyrysowaniu K-najbliższych odległości. Następnie analityk określa moment zgięcia (kolana), który odpowiada optymalnemu parametrowi ϵ.

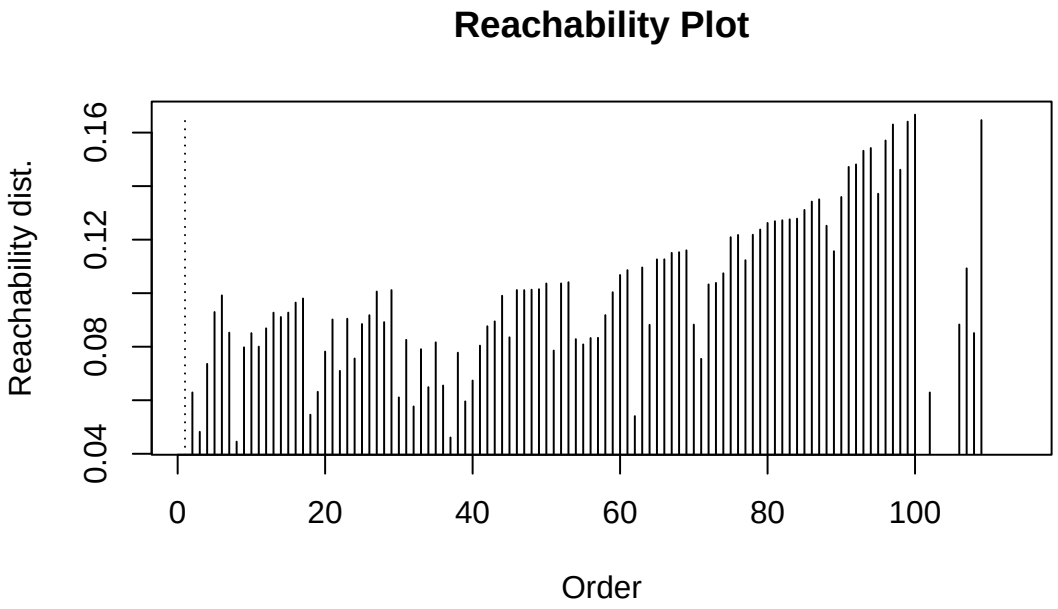
```
dbscan::kNNdistplot(gus2016_base, k = 3)
abline(h = 0.135, lty = 2)
set.seed(12345)
model.dbscan <- fpc::dbscan(gus2016_base, eps = 0.175, MinPts = 2)
factoextra::fviz_cluster(model.dbscan, gus2016_base, geom = "point")
```





Jak widać powyżej dla badanych obserwacji metoda DBSCAN nie daje satysfakcjonujących wyników z uwagi na przydzielenie wielu obserwacji do szumu, co w przypadku, gdy obserwacjami są gminy jest niezasadne merytorycznie. Dlatego pewnym rozwinięciem metody DBSCAN jest metoda OPTICS, w której efekty grupowania są podobne jak w DBSCAN, z tą różnicą, że metoda ta nie zostawia obserwacji bez przypisanego skupienia (szumu).

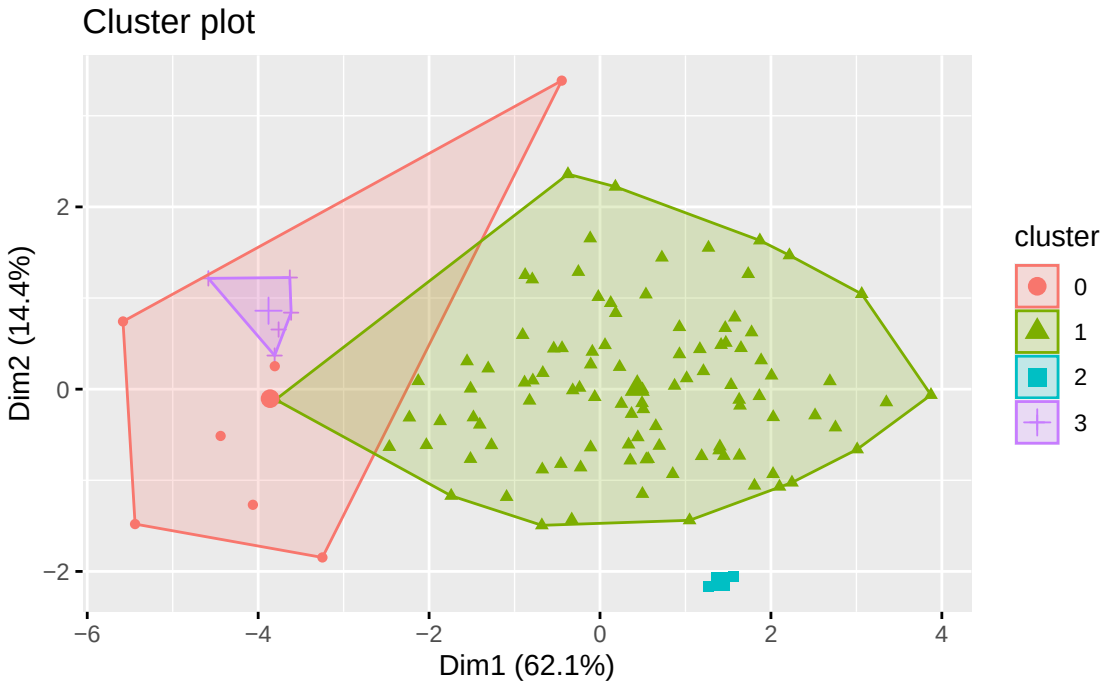
```
# Model optics, Reachability plot
model.optics <- dbscan::optics(gus2016_base, eps = 0.175, minPts = 2)
plot(model.optics)
```



Doliny na wykresie dostępności informują o istnieniu skupień, a szerokość doliny informuje o licznie obiektów przydzielonych do różnych skupień.

```
# optics skupienia
(result.optics <- dbscan::extractDBSCAN(model.optics, eps_cl = 0.4))
#> OPTICS ordering/clustering for 114 objects.
#> Parameters: minPts = 2, eps = 0.175, eps_cl = 0.4, xi = NA
#> The clustering contains 3 cluster(s) and 7 noise points.
```

```
#>
#>   0   1   2   3
#>   7 100   2   5
#>
#> Available fields: order, reachdist, coredist, predecessor, minPts, eps,
#>                   eps_cl, xi, cluster
factoextra::fviz_cluster(list(data = gus2016_base, cluster = result.optics$cluster),
    geom = "point")
```

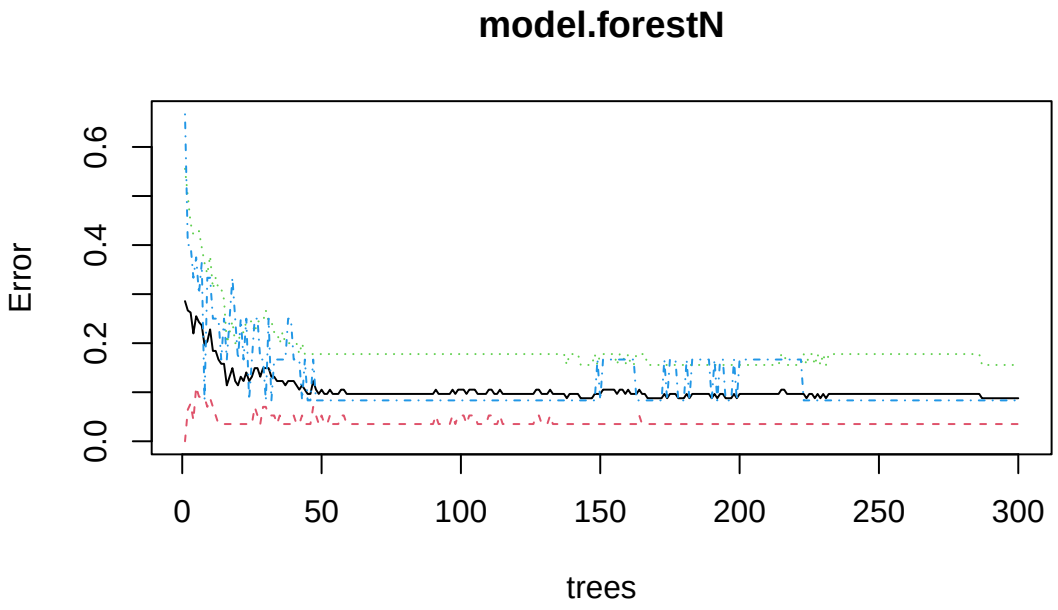


Niestety metoda ta wydaje się niezadowalająca. Znacząca większość obiektów została przypisana do jednego skupienia. Pozostałe 3 skupienia zawierają niewielką liczbę obserwacji. Biorąc pod uwagę powyższe wyniki grupowania za pomocą różnych metod analizy skupień, zdecydowano w dalszym badaniu wykorzystać wyniki uzyskane za pomocą metody K-średnich. Podział na 3 skupienia za pomocą analizy skupień pozwolił na określenie trzech poziomów rozwoju (wysoki, średni, niski).

3.3 Klasyfikacja

Klasyfikacje metodą lasów losowych wykonano w programie R następującymi instrukcjami wykorzystując funkcję *randomForest*, z pakietu *randomForest* (Breiman i in., 2018), która tworzy model lasów losowych.

```
model.forestN <- randomForest::randomForest(formula = Poziom_Rozwoju ~
    ., data = gus2016_norm, ntree = 300, importance = TRUE, proximity = TRUE)
plot(model.forestN)
```

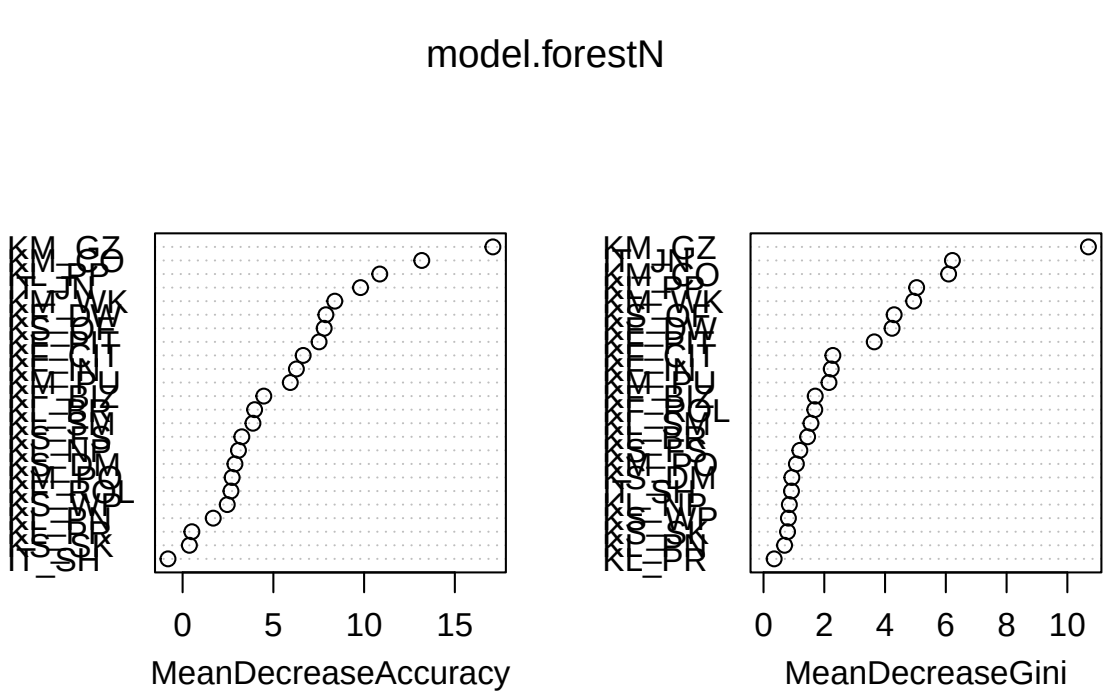


Powyższy wykres przedstawia błąd OOB (ang. *out-of-bag*) - kolor czarny, oraz krzywe wskaźnika błędów błędnej klasyfikacji - kolor niebieski, zielony oraz czerwony. Widać, ze powyżej ok. 150 drzew błąd OOB nie zmniejsza się.

Wykres średniej dokładności spadku (ang. *The Mean Decrease Accuracy*) wyraża, jak dużą dokładność traci model przez wykluczenie każdej zmiennej. Im bardziej ucierpi dokładność, tym ważniejsza jest zmienna dla pomyślnej klasyfikacji. Zmienne są prezentowane według malejącego znaczenia. Średni spadek współczynnika Giniego (ang. *The mean decrease in Gini*) jest miarą tego, jak każda zmienna przyczynia się do jednorodności węzłów i liści w powstałym lesie losowym. Im większa wartość średniej dokładności

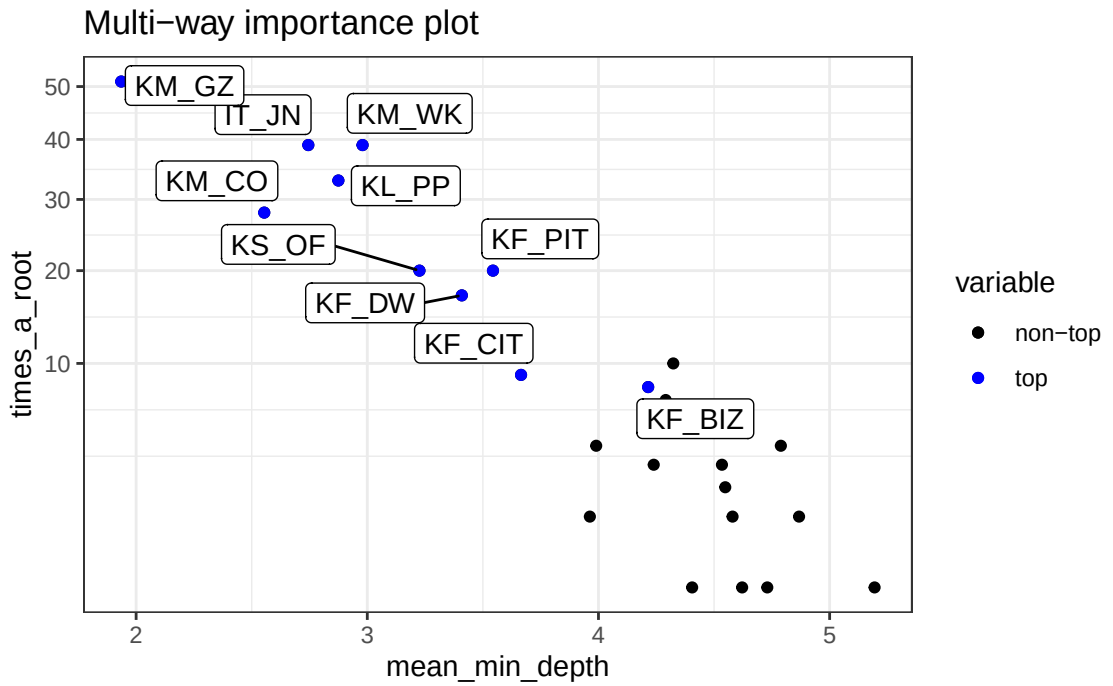
spadku lub średniego spadku wyniku Giniego, tym większe znaczenie zmiennej w modelu. Wykres średniej dokładności spadku oraz średniego spadku współczynnika Giniego wykonano następującymi instrukcji za pomocą funkcji *varImpPlot*.

```
randomForest::varImpPlot(model.forestN, sort = TRUE)
```



Innym narzędziem przedstawiającym istotność czynników stanowiących model lasów losowych jest *plot_multi_way_importance* z pakietu *randomForestExplainer* (Paluszynska i in., 2020)

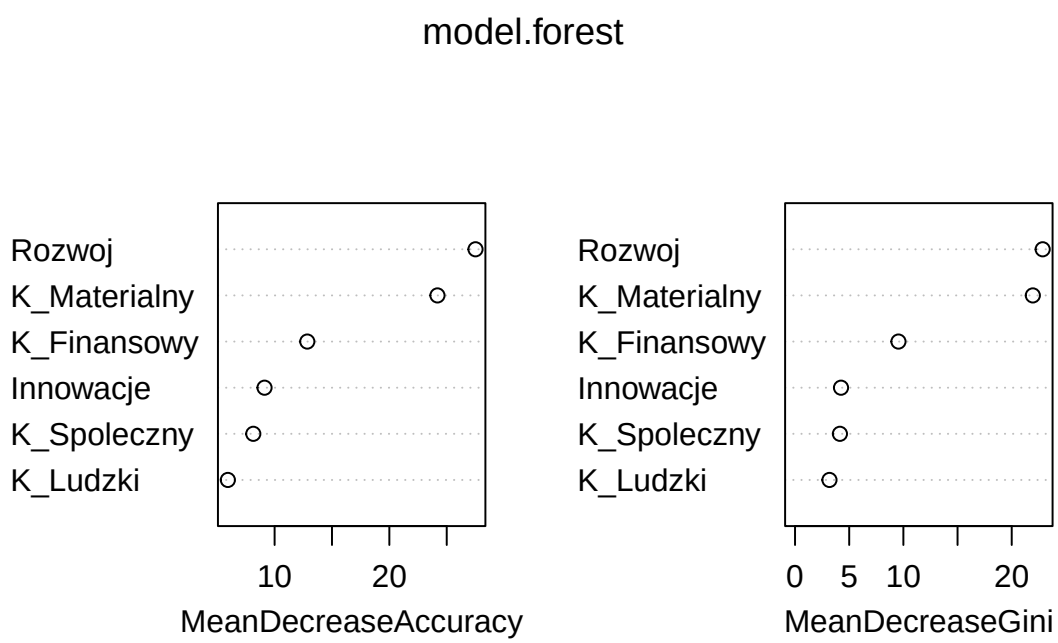
```
randomForestExplainer::plot_multi_way_importance(model.forestN,
    no_of_labels = 10)
```



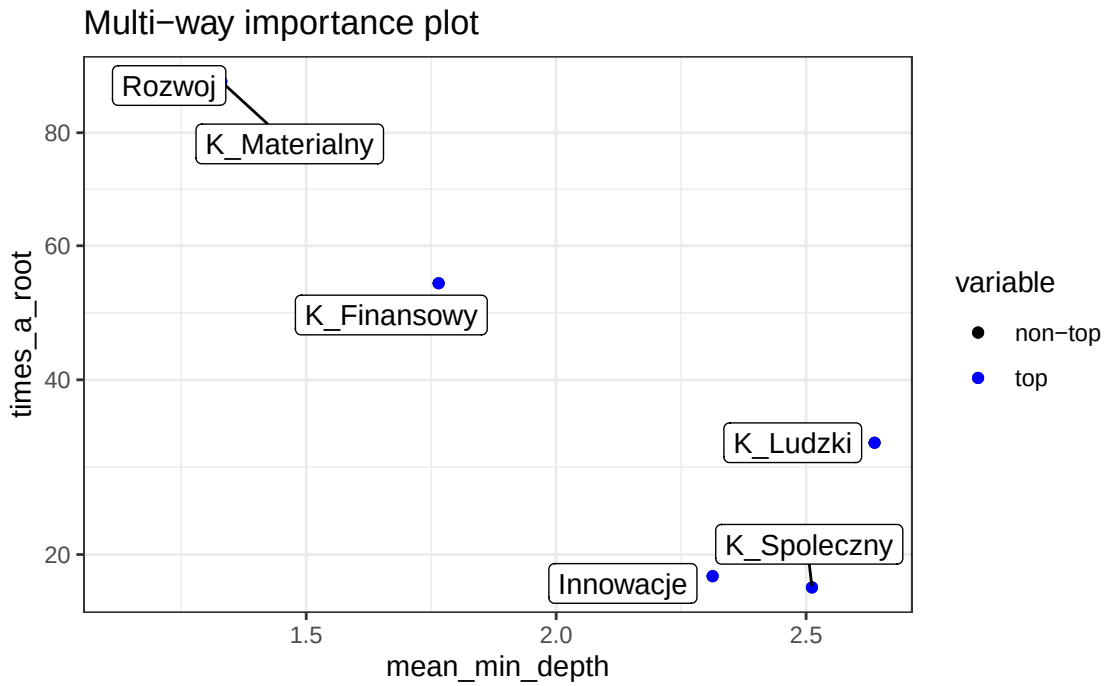
Jak widać wysoki wskaźnik średniej dokładności spadku oraz średnim spadkiem Giniego charakteryzuje się wskaźnik określający dostępność do gazu (*KM_GZ* - patrz Tabela 1.1), a kolejnym wskaźnikiem jest dostępność do ogrzewania centralnego (*KM_OG*). Oba wskaźniki, prócz tej samej kategorii aspektu rozwoju (Kapitał Materialny) są charakterystyczne dla obszarów o zwartej zabudowie. Model więc w sposób pośredni poinformował nas, że wyższy poziom rozwoju będzie występował na terenach o zwartej zabudowie, czyli uogólniając w miastach. Następnym istotnym i wiele mówiącym wskaźnikiem są podmioty gospodarcze w sekcjach J-N (*IT_JN*), czyli uogólniając podmioty dostarczające wysoko wykwalifikowanych usług (finanse, nauka, informatyka). Co ciekawe drugi wskaźnik z aspektu Innowacji- udział spółek handlowych z udziałem kapitału zagranicznego charakteryzuje się niską istotnością. Co prowadzi do ciekawych wniosków, jednak może być uwarunkowane lokalną charakterystyką województwa Zachodniopomorskiego. Wbrew intuicji przyrost naturalny (*KL_PR*) ma niewielkie znaczenie, prawdopodobnie z uwagi na jego niewielki rozrzut wśród gmin oraz duże migracje wewnętrzne. Sugeruje to, że możliwą poprawę modelu uzyska się włączając do modelu przyrost rzeczywisty zamiast przyrostu naturalnego.

Sprawdzono, także istotność syntetycznego wskaźnika poziomu rozwoju społeczno-gospodarczego oraz jego poszczególnych aspektów syntetycznego.

```
randomForest::varImpPlot(model.forest, sort = TRUE)
```



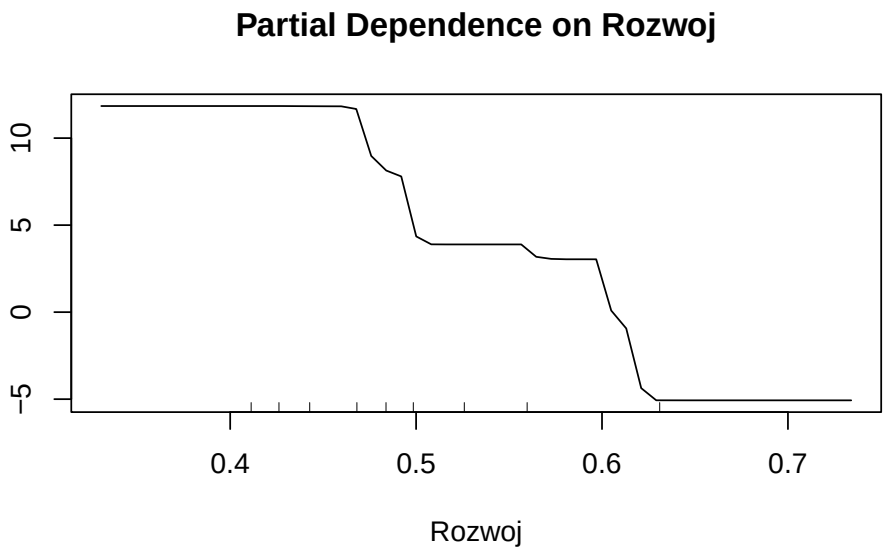

```
randomForestExplainer::plot_multi_way_importance(model.forest,
    no_of_labels = 10)
```



Jak widać najistotniejszymi kategoriami dla modelu jest rozwój społeczno-gospodarczy oraz kapitał materialny. Istotny jest również kapitał finansowy. Natomiast Kapitał ludzki, społeczny oraz innowacje cechują się dość niską istotnością. Na niską istotność innowacji jako aspektu przekłada się niewielka liczba wskaźników innowacji, które brały udział w konstrukcji syntetycznego wskaźnika rozwoju. Wiemy natomiast, że biorąc pod uwagę wskaźniki, które przekładają się na syntetyczny poziom rozwoju to wskaźnik będący częścią aspektu innowacje miał bardzo duże znaczenie (*IT_JN*).

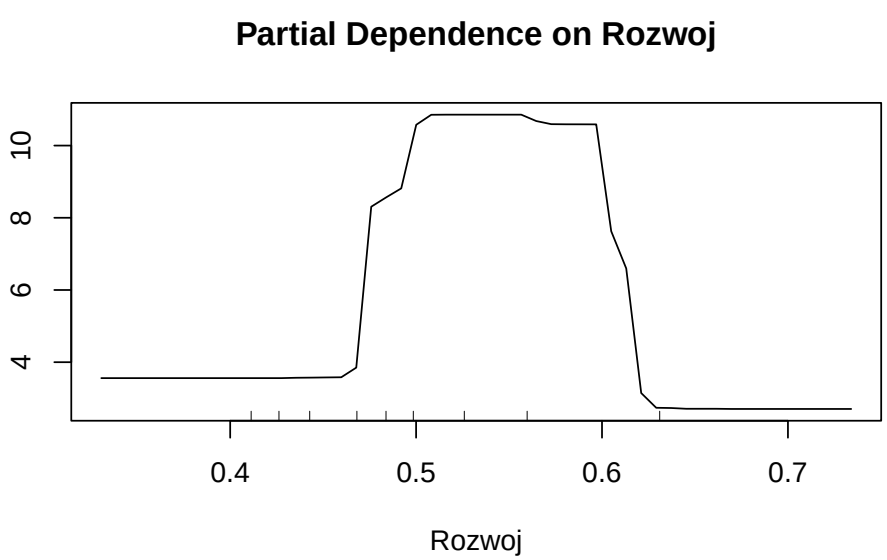
Za pomocą poniższych instrukcji korzystając z funkcji *partialPlot* z pakietu *randomForest* wygenerowano wykres częściowej zależności, który daje graficzne przedstawienie wpływu zmiennej na prawdopodobieństwo przydzielenia do danego skupienia.

```
randomForest::partialPlot(model.forest, gus2016_base, Rozwoj,
    "niski")
```



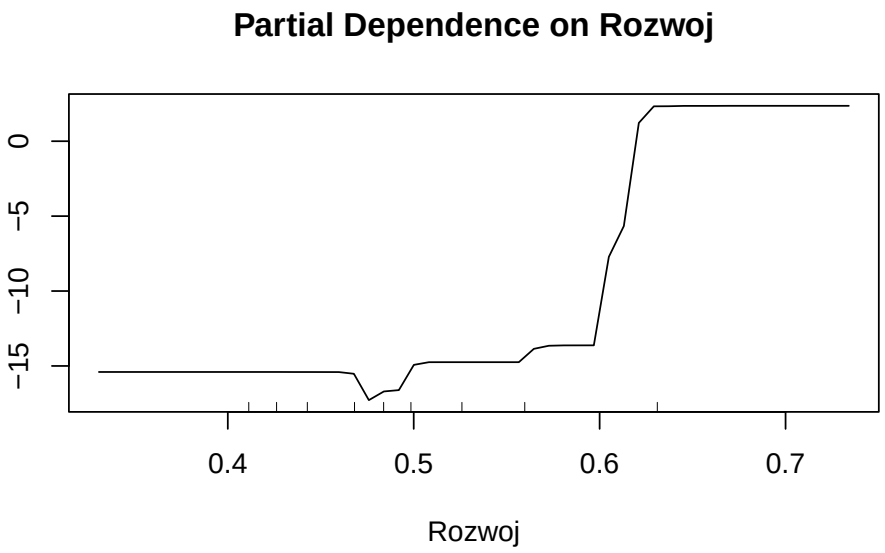
Wykres częściowej zależności pokazuje, że gminy o niskim poziomie rozwoju charakteryzują się syntetycznym wskaźnikiem poziomu rozwoju społeczno-gospodarczego na poziomie poniżej 0.6, przy czym większość poniżej 0.45.

```
randomForest::partialPlot(model.forest, gus2016_base, Rozwoj,
    "średni")
```



Zgodnie z powyższym wykresem, większość gmin (obserwacji) zaklasyfikowanych jako gminy o średnim poziomie rozwoju to gminy, których syntetyczny wskaźnik rozwoju zawiera się w przedziale $\{0.45, \ldots, 0.6\}$

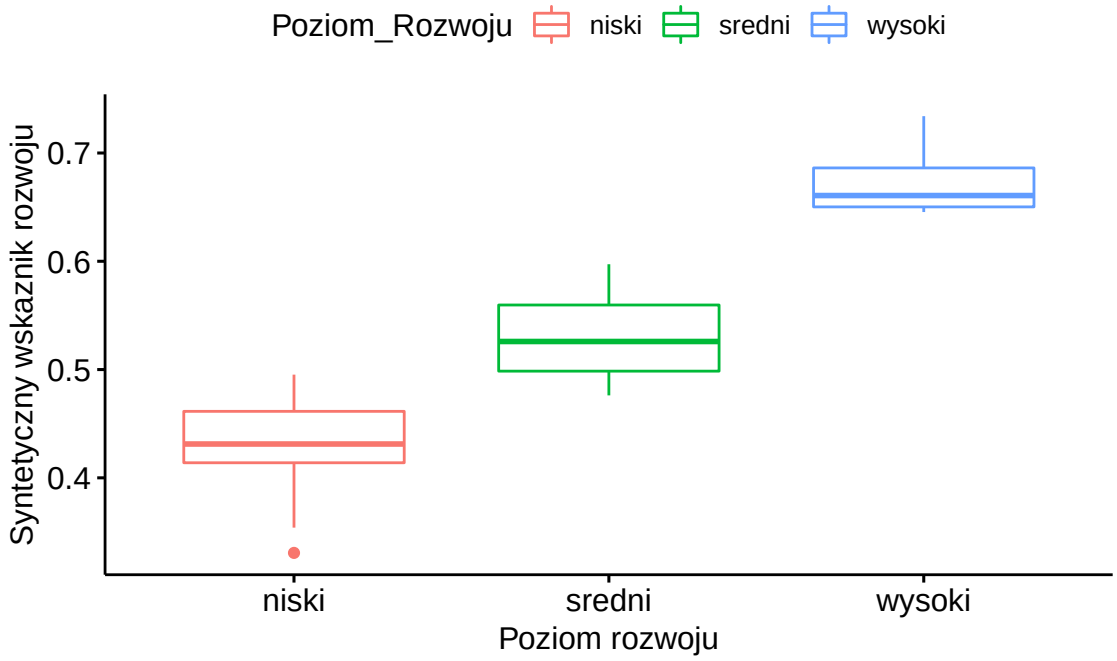
```
randomForest::partialPlot(model.forest, gus2016_base, Rozwoj,
    "wysoki")
```



Większość gmin zakwalifikowanych jako wysoko rozwinięte charakteryzuje się syntetycznym wskaźnikiem poziomu rozwoju powyżej 0.6.

Podobnie przedstawia się wykres pudełkowy ukazujący poziom rozwoju w zależności od syntetycznego wskaźnika poziomu rozwoju społeczno-gospodarczego. Wykres pudełkowy stworzono za pomocą funkcji *ggboxplot* z pakietu *ggpubr* (Kassambara, 2020).

```
ggpubr::ggboxplot(gus2016_base, x = "Poziom_Rozwoju", y = "Rozwoj",
          color = "Poziom_Rozwoju",
          xlab = "Poziom rozwoju", ylab = "Syntetyczny wskaźnik rozwoju")
```



Klasyfikację metodą xgboost wykonano w programie R za pomocą pakietu *xgboost* (Chen i in., 2021). Poniższe instrukcje mają na celu stworzenie modelu xgboost.

```
dtrain <- list(as.matrix(gus2016_norm[, 1:24]),
               label = as.numeric(gus2016_norm$Poziom_Rozwoju) - 1)


(xgb.fit <- xgboost::xgboost(data = dtrain[[1]],
                    label = dtrain[[2]],
                    eta = 0.1, # niższa wartość przeciwdziała przeuczeniu
                    max_depth = 10,
                    nround = 100,
                    subsample = 0.5,
                    eval_metric = 'mlogloss',
                    objective = 'multi:softprob',
                    num_class = 3,
                    nthread = 4))
#> [1]  train-mlogloss:0.997885
#> [2]  train-mlogloss:0.909257
```

```
#> [3]  train-mlogloss:0.838608
#> [4]  train-mlogloss:0.773251
#> [5]  train-mlogloss:0.718040
#> [6]  train-mlogloss:0.668381
#> [7]  train-mlogloss:0.617961
#> [8]  train-mlogloss:0.572172
#> [9]  train-mlogloss:0.530559
#> [10] train-mlogloss:0.494900
#> [11] train-mlogloss:0.460779
#> [12] train-mlogloss:0.430935
#> [13] train-mlogloss:0.404206
#> [14] train-mlogloss:0.380869
#> [15] train-mlogloss:0.356333
#> [16] train-mlogloss:0.336796
#> [17] train-mlogloss:0.318126
#> [18] train-mlogloss:0.298931
#> [19] train-mlogloss:0.282705
#> [20] train-mlogloss:0.266052
#> [21] train-mlogloss:0.253279
#> [22] train-mlogloss:0.240274
#> [23] train-mlogloss:0.229947
#> [24] train-mlogloss:0.218088
#> [25] train-mlogloss:0.209529
#> [26] train-mlogloss:0.199292
#> [27] train-mlogloss:0.189607
#> [28] train-mlogloss:0.181717
#> [29] train-mlogloss:0.173295
#> [30] train-mlogloss:0.166231
#> [31] train-mlogloss:0.159693
#> [32] train-mlogloss:0.153206
```

```
#> [33] train-mlogloss:0.147119
#> [34] train-mlogloss:0.141460
#> [35] train-mlogloss:0.136236
#> [36] train-mlogloss:0.132112
#> [37] train-mlogloss:0.127626
#> [38] train-mlogloss:0.123212
#> [39] train-mlogloss:0.118937
#> [40] train-mlogloss:0.114967
#> [41] train-mlogloss:0.111629
#> [42] train-mlogloss:0.107682
#> [43] train-mlogloss:0.104072
#> [44] train-mlogloss:0.100303
#> [45] train-mlogloss:0.097642
#> [46] train-mlogloss:0.095017
#> [47] train-mlogloss:0.092613
#> [48] train-mlogloss:0.089290
#> [49] train-mlogloss:0.087161
#> [50] train-mlogloss:0.084854
#> [51] train-mlogloss:0.082858
#> [52] train-mlogloss:0.080699
#> [53] train-mlogloss:0.078807
#> [54] train-mlogloss:0.077011
#> [55] train-mlogloss:0.075521
#> [56] train-mlogloss:0.073813
#> [57] train-mlogloss:0.072472
#> [58] train-mlogloss:0.070746
#> [59] train-mlogloss:0.069476
#> [60] train-mlogloss:0.068129
#> [61] train-mlogloss:0.067116
#> [62] train-mlogloss:0.065451
```

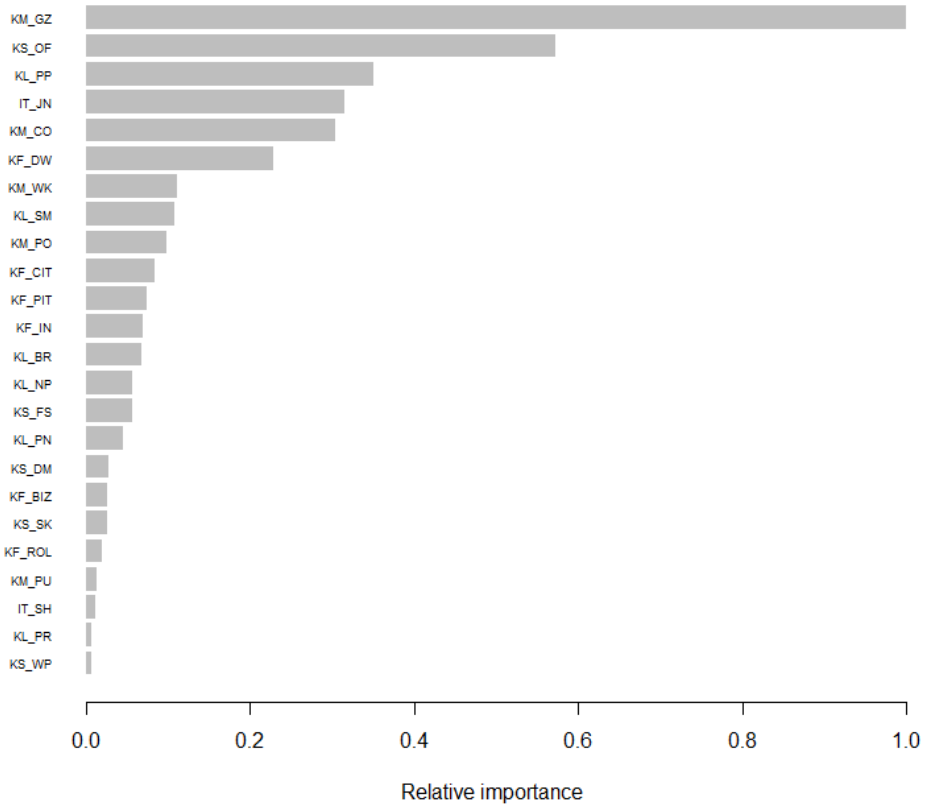
```
#> [63] train-mlogloss:0.063980
#> [64] train-mlogloss:0.062357
#> [65] train-mlogloss:0.060823
#> [66] train-mlogloss:0.059931
#> [67] train-mlogloss:0.059013
#> [68] train-mlogloss:0.058509
#> [69] train-mlogloss:0.057438
#> [70] train-mlogloss:0.056224
#> [71] train-mlogloss:0.055519
#> [72] train-mlogloss:0.054804
#> [73] train-mlogloss:0.054045
#> [74] train-mlogloss:0.053553
#> [75] train-mlogloss:0.052829
#> [76] train-mlogloss:0.052280
#> [77] train-mlogloss:0.051682
#> [78] train-mlogloss:0.050844
#> [79] train-mlogloss:0.050293
#> [80] train-mlogloss:0.049280
#> [81] train-mlogloss:0.049056
#> [82] train-mlogloss:0.048635
#> [83] train-mlogloss:0.048103
#> [84] train-mlogloss:0.047474
#> [85] train-mlogloss:0.047179
#> [86] train-mlogloss:0.046598
#> [87] train-mlogloss:0.046149
#> [88] train-mlogloss:0.045753
#> [89] train-mlogloss:0.045068
#> [90] train-mlogloss:0.044401
#> [91] train-mlogloss:0.043755
#> [92] train-mlogloss:0.043506
```

```
#> [93] train-mlogloss:0.042797
#> [94] train-mlogloss:0.042242
#> [95] train-mlogloss:0.042111
#> [96] train-mlogloss:0.041620
#> [97] train-mlogloss:0.041395
#> [98] train-mlogloss:0.041277
#> [99] train-mlogloss:0.041062
#> [100]    train-mlogloss:0.040747
#> ##### xgb.Booster
#> raw: 207.5 Kb
#> call:
#>   xgb.train(params = params, data = dtrain, nrounds = nrounds,
#>     watchlist = watchlist, verbose = verbose, print_every_n = print_every_n,
#>     early_stopping_rounds = early_stopping_rounds, maximize = maximize,
#>     save_period = save_period, save_name = save_name, xgb_model = xgb_model,
#>     callbacks = callbacks, eta = 0.1, max_depth = 10, subsample = 0.5,
#>     eval_metric = "mlogloss", objective = "multi:softprob", num_class = 3,
#>     nthread = 4)
#> params (as set within xgb.train):
#>   eta = "0.1", max_depth = "10", subsample = "0.5", eval_metric = "mlogloss", objectiv
#> xgb.attributes:
#>   niter
#> callbacks:
#>   cb.print.evaluation(period = print_every_n)
#>   cb.evaluation.log()
#> # of features: 24
#> niter: 100
#> nfeatures : 24
#> evaluation_log:
#>     iter train_mlogloss
```

```
#>         1       0.997885
#>         2       0.909257
#> ---
#>        99       0.041062
#>       100       0.040747
```

Wykres istotności zmiennych dla modelu xgboost wykonano za pomocą poniższych instrukcji.

```
xgboost::xgb.plot.importance(xgboost::xgb.importance(colnames(gus2016_norm),
    model = xgb.fit))
```

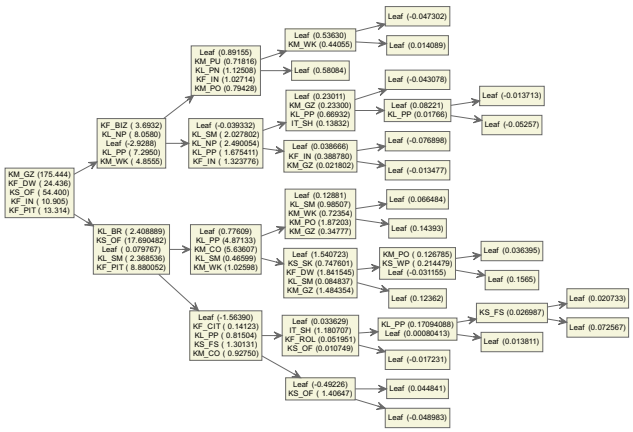


Zgodnie z powyższym wykresem najistotniejszym czynnikiem jest *KM_GZ*, czyli podobnie jak przy zastosowaniu przy klasyfikacji metodą lasów losowych. Kolejnym

istotnym czynnikiem jest udział osób fizycznych prowadzących działalność gospodarczą (KS_OF), wskaźnik ten informuje nas o przedsiębiorczości mieszkańców danej gminy. Istotny jest też udział osób pracujących w ogóle osób w wieku produkcyjnym oraz dochody własne gminy per capita. Co ciekawe wskaźnik mający w założeniu przedstawiać jakość rządzenia władzy lokalnej, objawiający się w udziałem specjalistów, kierowników w ogóle radnych nie ma istotnego znaczenia na poziom rozwoju. Niezwykle istotną role zwartej tkanki przestrzennej na poziom rozwoju społeczno-gospodarczego potwierdza, oprócz wskaźników *KM_GZ*, *KM_CO* niska istotność dochodów z podatku rolnego na mieszkańca (*KF_ROL*).

Za pomocą poniższych instrukcji za pomocą funkcji *xgb.plot.multi.trees* można przedstawić zespół drzew decyzyjnych jako zbiorczą jednostkę.

```
xgboost::xgb.plot.multi.trees(colnames(gus2016_norm), model = xgb.fit)
#> Column 2 ['No'] of item 2 is missing in item 1. Use fill=TRUE to fill with NA (NULL fo
```



\begin{center}

3.4 Analiza wariancji

Jednoczynnikową analizę wariancji (ang. *one-way ANOVA*) wykonujemy w programie R następującymi instrukcjami wykorzystując funkcję *aov* z pakietu *stats* (**R-stats**). Funkcja *summary* służy do podsumowania modelu analizy wariancji.

```
# ANOVA
ANOVA <- stats::aov(gus2016_base$Rozwoj ~ gus2016_base$Poziom_Rozwoju)
summary(ANOVA)
#>                              Df Sum Sq Mean Sq F value Pr(>F)
#> gus2016_base$Poziom_Rozwoju   2 0.6485  0.3242     269 <2e-16 ***
#> Residuals                   111 0.1338  0.0012
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ponieważ p-wartość jest mniejsza niż poziom istotności 0.001 to mamy podstawy, żeby odrzucić hipotezę zerową. Średnie grupowe różnią się znacząco między sobą. Możemy stwierdzić, że poziom rozwoju ma istotny wpływ na syntetyczny wskaźnik poziomu rozwoju społeczno-gospodarczego oraz jego aspekty (oznaczone jako reszty z modelu - ang. *Residuals*)

W jednoczynnikowej analizie wariancji istotna p-wartość wskazuje, że niektóre średnie grupowe są różne, ale nie wiemy, które pary grup są różne. Możliwe jest przeprowadzenie wielokrotnych porównań parami, aby określić, czy średnia różnica między poszczególnymi parami grupy jest statystycznie istotna.

W tym celu wykonano test Tukey-a (jeden z testów post hoc) za pomocą funkcji *TukeyHSD* z pakietu *stats*

```
# testy post-hoc
TukeyHSD(ANOVA)
#>   Tukey multiple comparisons of means
#>     95% family-wise confidence level
#>
#> Fit: stats::aov(formula = gus2016_base$Rozwoj ~ gus2016_base$Poziom_Rozwoju)
#>
#> $`gus2016_base$Poziom_Rozwoju`
#>                     diff        lwr       upr p adj
```

```
#> średni-niski  0.09563711 0.07919022 0.1120840     0
#> wysoki-niski  0.23961273 0.21341741 0.2658080     0
#> wysoki-średni 0.14397562 0.11717976 0.1707715     0
```

Zgodnie z powyższym wszystkie pary istotnie się różnią, ponieważ p-wartość dla każdej z pary wynosi 0. Przy czym największa różnica w średnich jest między poziomami wysoki-niski, co świadczy dobrze o podziale. Różnica między wysokim a średnim poziomem jest większa niż między średnim a niskim. Może to świadczyć o postępującej polaryzacji rozwoju, żeby to sprawdzić przeprowadzono testy ANOVA i post-hoc Tukeya dla innych roczników.

```
# testy post-hoc
ANOVA <- stats::aov(gus2017_base$Rozwoj ~ gus2017_base$Poziom_Rozwoju)
summary(ANOVA)
#>                              Df Sum Sq Mean Sq F value Pr(>F)
#> gus2017_base$Poziom_Rozwoju   2 0.6119 0.30597   253.8 <2e-16 ***
#> Residuals                   111 0.1338 0.00121
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
TukeyHSD(ANOVA)
#>   Tukey multiple comparisons of means
#>     95% family-wise confidence level
#>
#> Fit: stats::aov(formula = gus2017_base$Rozwoj ~ gus2017_base$Poziom_Rozwoju)
#>
#> $`gus2017_base$Poziom_Rozwoju`
#>                     diff        lwr       upr p adj
#> średni-niski  0.08358656 0.06686513 0.1003080     0
#> wysoki-niski  0.21762703 0.19409653 0.2411575     0
#> wysoki-średni 0.13404047 0.11002986 0.1580511     0
```

```
# testy post-hoc
ANOVA <- stats::aov(gus2018_base$Rozwoj ~ gus2018_base$Poziom_Rozwoju)
summary(ANOVA)
#>                               Df Sum Sq Mean Sq F value Pr(>F)
#> gus2018_base$Poziom_Rozwoju    2 0.6045 0.30227   285.4 <2e-16 ***
#> Residuals                    111 0.1175 0.00106
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
TukeyHSD(ANOVA)
#>   Tukey multiple comparisons of means
#>     95% family-wise confidence level
#>
#> Fit: stats::aov(formula = gus2018_base$Rozwoj ~ gus2018_base$Poziom_Rozwoju)
#>
#> $`gus2018_base$Poziom_Rozwoju`
#>                       diff        lwr       upr p adj
#> średni-niski  0.09075572 0.07527941 0.1062320     0
#> wysoki-niski  0.22656744 0.20276805 0.2503668     0
#> wysoki-średni 0.13581172 0.11147047 0.1601530     0
```

```
# testy post-hoc
ANOVA <- stats::aov(gus2019_base$Rozwoj ~ gus2019_base$Poziom_Rozwoju)
summary(ANOVA)
#>                               Df Sum Sq Mean Sq F value Pr(>F)
#> gus2019_base$Poziom_Rozwoju    2 0.6541  0.3270     275 <2e-16 ***
#> Residuals                    110 0.1308  0.0012
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
TukeyHSD(ANOVA)
#>   Tukey multiple comparisons of means
#>     95% family-wise confidence level
```

```
#>
#> Fit: stats::aov(formula = gus2019_base$Rozwoj ~ gus2019_base$Poziom_Rozwoju)
#>
#> $`gus2019_base$Poziom_Rozwoju`
#>                     diff        lwr       upr p adj
#> średni-niski  0.09599066 0.07960284 0.1123785     0
#> wysoki-niski  0.23991372 0.21440405 0.2654234     0
#> wysoki-średni 0.14392306 0.11841339 0.1694327     0
```

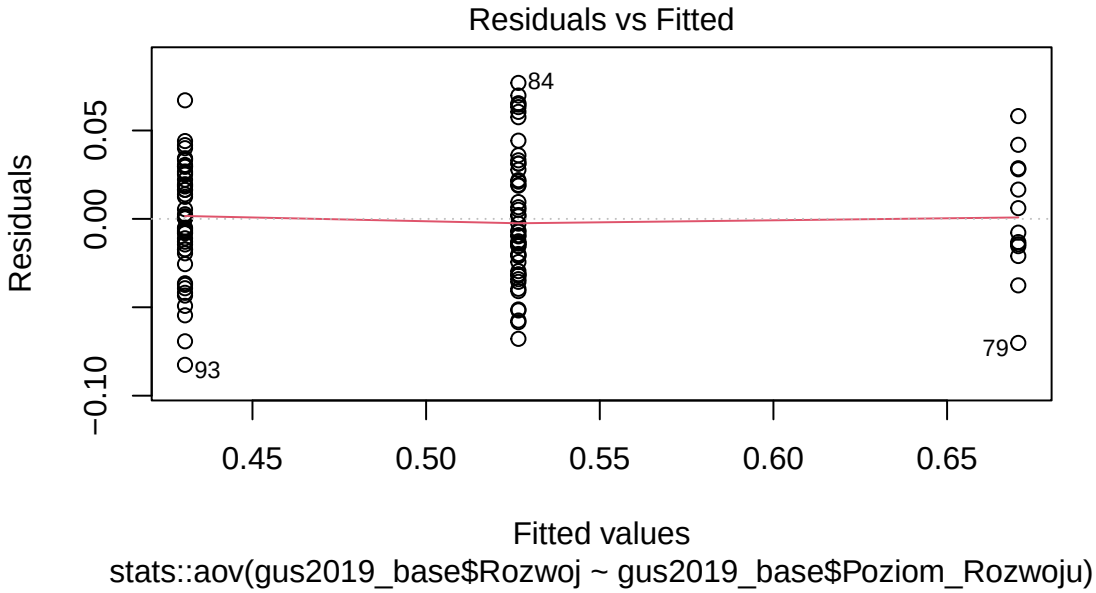
Powyższe testy prowadzą do ciekawych wniosków. W 2017 roku znacząco zmniejszyła się polaryzacja poziomu rozwoju w województwie Zachodniopomorskim. Następnie sukcesywnie co roku wzrastała i w 2019 roku osiągnęła wyższy poziom niż w 2016 roku. Można również użyć jednoczesnych testów ogólnych hipotez liniowych wielokrotnego porównywania średnich metodą Tukey-a za pomocą pakietu *multcomp* (Hothorn i in., 2021)

```
summary(multcomp::glht(ANOVA, lincft = mcp(group = "Tukey")))
#>
#>   Simultaneous Tests for General Linear Hypotheses
#>
#> Fit: stats::aov(formula = gus2019_base$Rozwoj ~ gus2019_base$Poziom_Rozwoju)
#>
#> Linear Hypotheses:
#>                                          Estimate Std. Error t value Pr(>|t|)
#> (Intercept) == 0                         0.430581   0.004877   88.28   <2e-16 ***
#> gus2019_base$Poziom_Rozwojuśredni == 0 0.095991   0.006898   13.92   <2e-16 ***
#> gus2019_base$Poziom_Rozwojuwysoki == 0 0.239914   0.010737   22.34   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> (Adjusted p values reported -- single-step method)
```

Także ten test odrzucił hipotezę zerową, p-wartość dla każdego przypadku jest mniejsza niż 0.001

Ponieważ ANOVA zakłada, że dane mają rozkład normalny, a wariancja między grupami jest jednorodna to należy to sprawdzić. Jednorodność wariancji między grupami sprawdzono dzięki wykresowi reszt w funkcji dopasowań za pomocą następującej instrukcji.

```
plot(ANOVA, 1)
```



Zgodnie z powyższym wykresem nie ma ewidentnych zależności między resztami a średnią z każdej grupy. Można więc założyć jednorodność wariancji.

Innym sposobem na sprawdzenie jednorodności wariancji jest test Levene, który można wykonać za pomocą funkcji *leveneTest* z pakietu *car* (Fox i in., 2021).

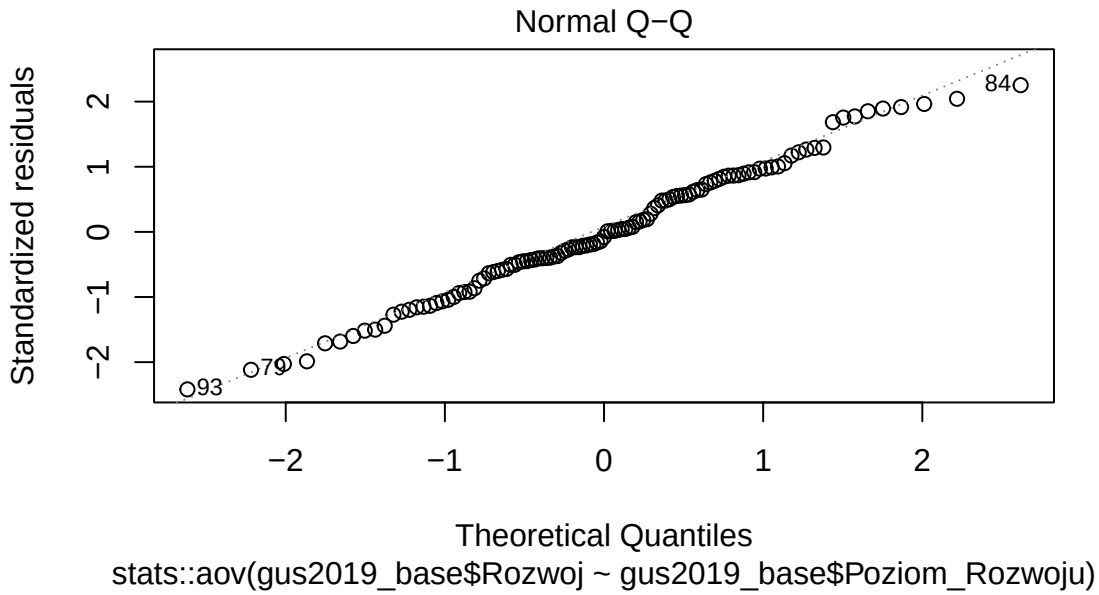
```
car::leveneTest(gus2016_base$Rozwoj ~ gus2016_base$Poziom_Rozwoju)
#> Levene's Test for Homogeneity of Variance (center = median)
#>        Df F value Pr(>F)
```

```
#> group   2  0.4087 0.6655
#>       111
```

p-wartość tego testu jest większa niż poziom istotności $0,05$. Oznacza to, że nie ma dowodów sugerujących, że wariancja między grupami jest istotnie różna. Można założyć jednorodność wariancji w różnych grupach.

Normalność rozkładu można sprawdzić za pomocą wykresu normalności reszt (wykres kwantyl-kwantyl zastosowany do reszt), za pomocą poniższej instrukcji.

```
plot(ANOVA, 2)
```



Reszty w przybliżeniu przebiegają wzdłuż linii prostej (45° linii odniesienia), co sugeruje, że można założyć normalność danych.

Powyższe założenie można potwierdzić stosując test Shapiro-Wilka na resztach ANOVA za pomocą poniższych instrukcji, wykorzystując funkcje *shapiro.test* z pakietu *stats*. Test ten nie wykazuje żadnych oznak naruszenia normalności.

```
# test shapiro-wilka na resztach w celu sprawdzenia
# normalności
ANOVA_reszty <- residuals(object = ANOVA)
stats::shapiro.test(x = ANOVA_reszty)
#>
#>  Shapiro-Wilk normality test
#>
#> data:  ANOVA_reszty
#> W = 0.99041, p-value = 0.6155
```

Zgodnie z powyższym p-wartość jest większa niż 0.05 można więc założyć normalność danych. Potwierdzono, że dane mają rozkład normalny, a wariancja między grupami jest jednorodna.

Wykonano także nieparametryczny test sumy rang Kruskala-Wallisa, który nie wymaga założenia normalności oraz jednorodności wariancji między grupami. Test Kruskala-Wallisa wykonano za pomocą funkcji *kruskal.test* z pakietu *stats*:

```
# test Kruskala-Wallis
stats::kruskal.test(gus2016_base$Rozwoj ~ gus2016_base$Poziom_Rozwoju)
#>
#>  Kruskal-Wallis rank sum test
#>
#> data:  gus2016_base$Rozwoj by gus2016_base$Poziom_Rozwoju
#> Kruskal-Wallis chi-squared = 88.85, df = 2, p-value < 2.2e-16
```

Zgodnie z powyższym przyjmujemy hipotezę alternatywną informującą nas o tym, że nie wszystkie mediany grupy są równe. Potwierdza to wnioski płynące z testu ANOVA, dlatego też nie wykonano testów post-hoc dla testu sumy rang Kruskala-Wallisa.

3.5 Wyniki dla ostatecznych grup -mapy

W celu przedstawienia syntetycznego wskaźnika poziomu rozwoju społeczno-gospodarczego należało przygotować mapy, ukazujące rozmieszczenie przestrzenne rozwoju. W tym celu skorzystano z danych (plik w formacie *shapefile* - *shp*) dotyczących granic administracyjnych (gmin) z Głównego Urzędu Geodezji i Kartografii. Dane te obejmowały całą Polskę, dlatego korzystając z programu QGIS ograniczono je do województwa Zachodniopomorskiego. Następnie za pomocą biblioteki *sf* wczytano przygotowany plik *shp* do środowiska R za pomocą funkcji *read_sf*:

```
gminy2 <- sf::read_sf('zachpom_gminy.shp')
```

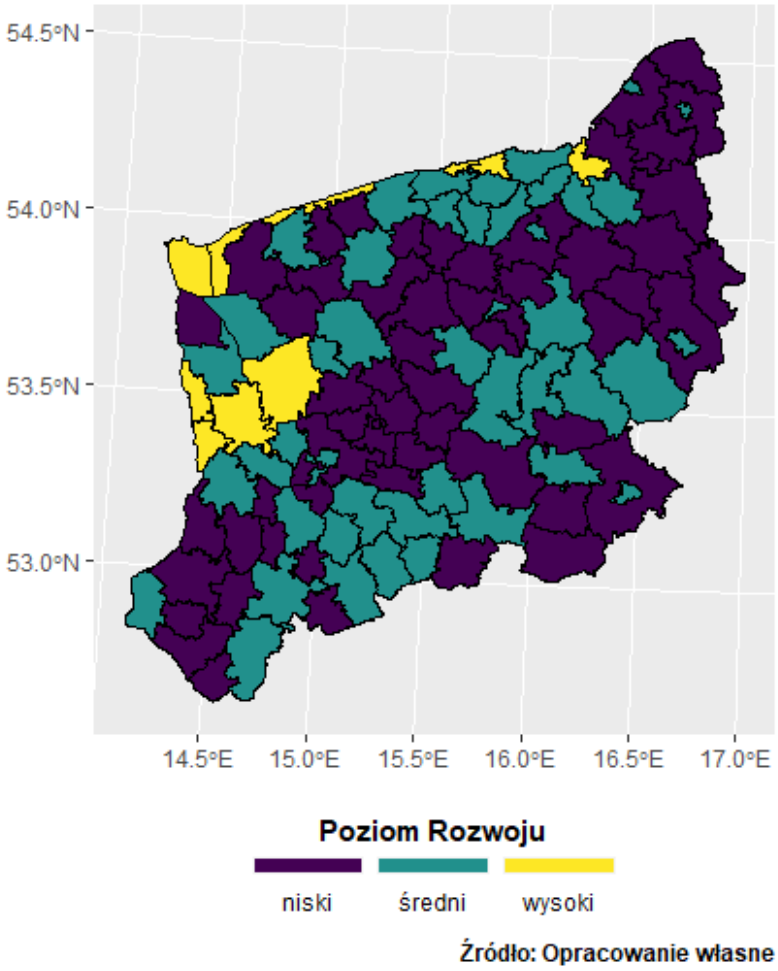
Następnie przygotowano dane dotyczące poziomu rozwoju oraz jego aspektów tak, żeby połączyć je z mapą. Zrobiono to za pomocą poniższych instrukcji.

```
data_map2016 <- select(Gus_2016, 1:2)
data_map2016$ID <- seq.int(nrow(data_map2016))  #dodaje ID

data_map2016_temp <- dplyr::select(gus2016_base, 1:7)
data_map2016_temp$ID <- seq.int(nrow(data_map2016_temp))
data_map2016 <- dplyr::right_join(data_map2016, data_map2016_temp,
    by = "ID")

names(gminy2)[1] <- "TERYT"  #zmiana nazwy

gminy2 <- dplyr::right_join(gminy2, data_map2016)
gminy2 <- dplyr::select(gminy2, -"ID")  #usuwam niepotrzebne kolumny
gminy2 <- ggplot2::fortify(gminy2)  #dataframe
```

Mając przygotowane dane wykonano mapę za pomocą poniższych instrukcji dla roku 2016. Analogicznie wykonano mapy dla innych roczników.

```
library(ggplot2)
mapa_2016 <- ggplot()+
  geom_sf(aes(fill=Poziom_Rozwoju),color='transparent',data=gminy2)+
  geom_sf(fill='transparent',color='white',data=gminy2)+
  scale_fill_viridis_d(name='Poziom Rozwoju',
                       guide=guide_legend(
                         direction='horizontal',
                         title.position='top',
                         title.hjust = .5,
                         label.hjust = .5,
                         label.position = 'bottom',
                         keywidth = 3,
                         keyheight = .5
                       ))+
  labs(caption=c('Źródło: Opracowanie własne'))+
  theme_gray()+
  theme(title=element_text(face='bold'),
        legend.position = 'bottom')
```

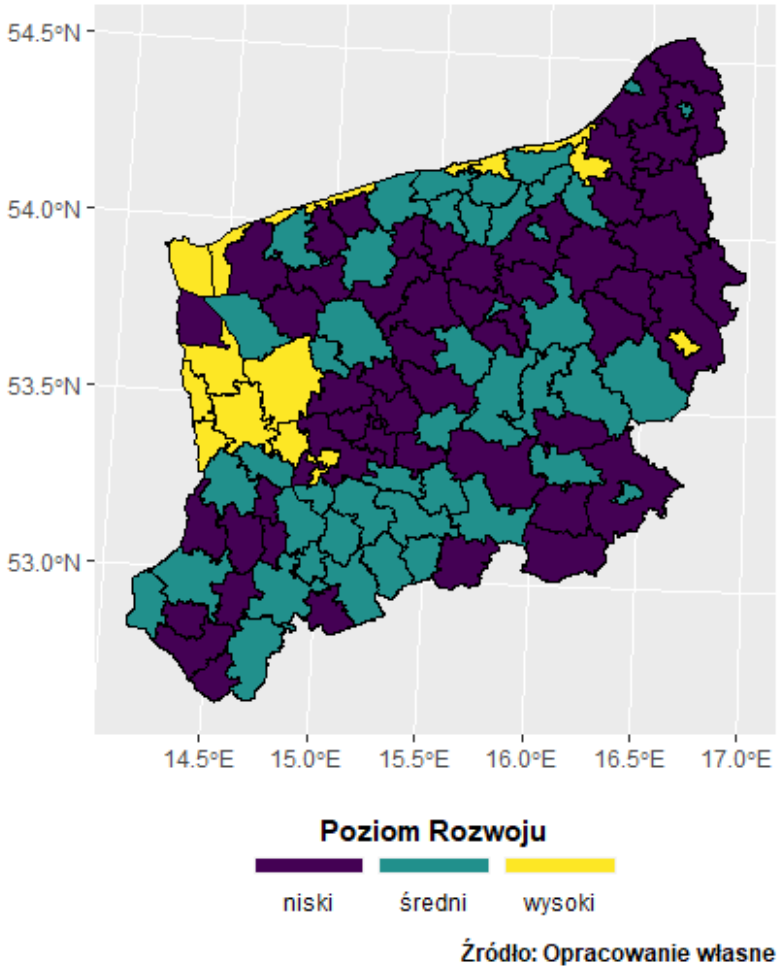
W 2016 roku (patrz Rysunek 3.1) wysokim poziomem rozwoju charakteryzował się Szczecin wraz z pobliskim gminami- Dobrą Szczecińską Kołbaskowem i Goleniowem oraz miejscowości nabrzeżne- Świnoujście, Kołobrzeg, Mielno, Ustronie Morskie, Międzyzdroje, Dziwnów, Rewal. Drugim ośrodkiem miejskim nie należącym ani do aglomeracji Szczecińskiej ani nie będący gminą nadmorską, a charakteryzujący się wysokim poziomem rozwoju jest Koszalin. Koszalin oraz Kołobrzeg są gminami na północy województwa, których wysoki poziom rozwoju „dyfuzyjnieżozchodzi się na powiat, tworząc razem z Białogardem (większy ośrodek miejski, zakwalifikowany jako gmina o średnim poziomie rozwoju) pas gmin o średnim poziomie rozwoju. Scharakteryzować można pas gmin o niskim poziomie rozwoju ciągnący się pojezierzem zachodniopomorskim oraz gminy na wschód i południe od Koszalina.



Rysunek 3.1: *Poziom rozwoju gmin w województwie Zachodniopomorskim (2016)*

W 2017 roku (patrz Rysunek 3.2) widać przyrost gmin charakteryzujących się wysokim poziomem rozwoju w obrębie aglomeracji szczecińskiej. W porównaniu do roku poprzedniego do kategorii tej dołączył Stargard, Police, Kobylanka. Koszalin oraz miejscowości nadmorskie także zostały przydzielone do wysokiego poziomu rozwoju. Natomiast ciekawym przypadkiem jest miasto Szczecinek, które w 2017 roku osiągnęło wysoki poziom rozwoju społeczno-gospodarczego, a okola je gmina wiejska o takiej samej nazwie charakteryzująca się niskim poziomem rozwoju. Gminy w powiecie stargardzkim (Dolice, Suchań) odznaczają się średnim poziomem rozwoju. Generalizując jest to rok „lepszy"dla gmin w zachodniej części województwa.

W 2018 (patrz Rysunek 3.3) roku Police oraz Kobylanka przydzielone zostały do gmin o średnim poziomie rozwoju. Można powiedzieć, że aglomeracja Szczecińska „osłabła“,

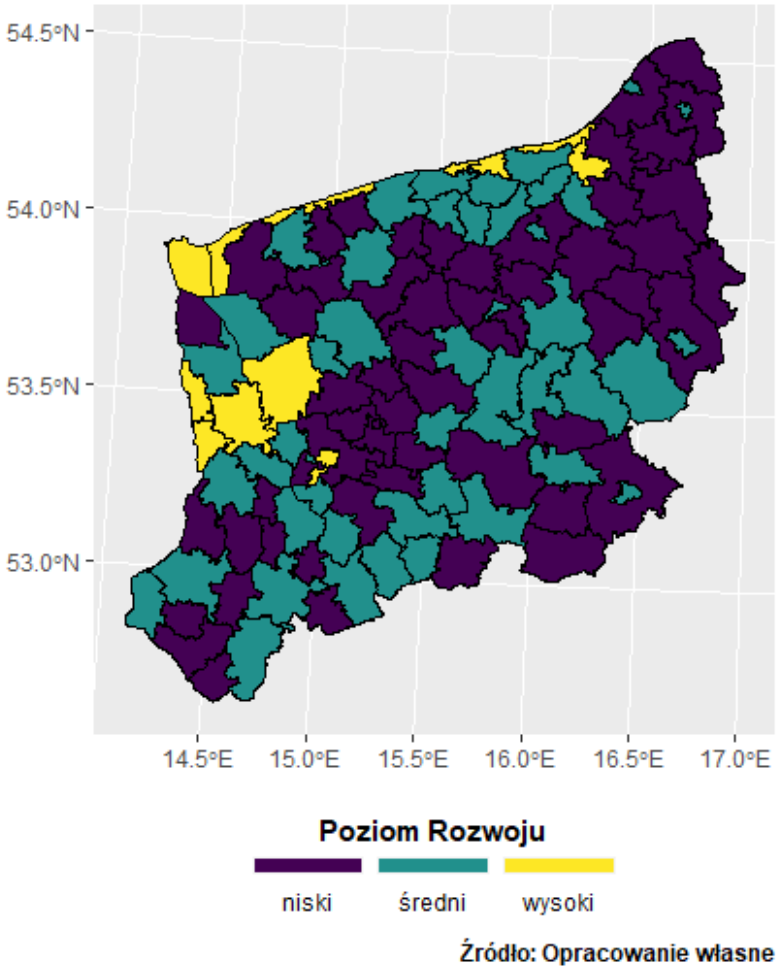


Rysunek 3.2: *Poziom rozwoju gmin w województwie Zachodniopomorskim (2017)*

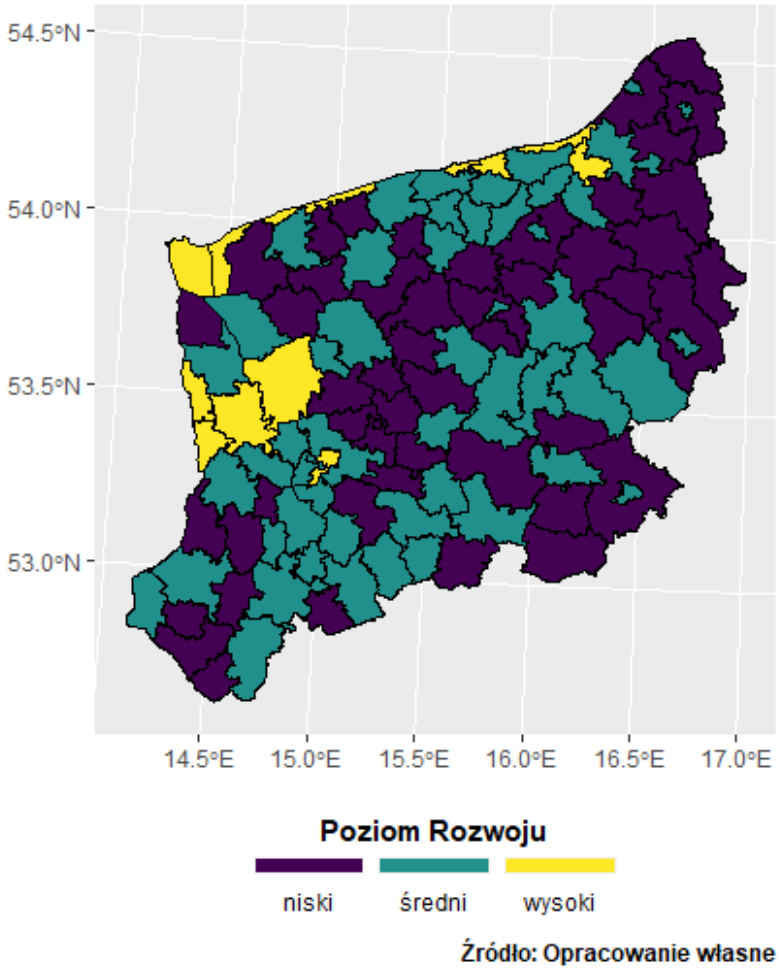
jednak Stargard pozostał gminą o wysokim poziomie rozwoju. Warto zaznaczyć, że jest to największy ośrodek miejski w aglomeracji Szczecińskiej oprócz Szczecina. Miasto Szczecinek tak jak w 2016 roku zostało sklasyfikowane jako gmina o średnim poziomie rozwoju. Dolice, Lipiany i Suchań ponownie „wróciły" do gmin o niskim poziomie rozwoju.

Brak zmian wśród gmin o wysokim poziomie rozwoju w 2019 roku (patrz Rysunek 3.4) w stosunku do roku poprzedniego. Widać pozytywny wpływ Stargardu, który pozostał jako gmina o wysokim poziomie rozwoju i wpłynął na okalającej miasto gminę wiejską o tej samej nazwie. Gmina miejsko-wiejska Sianów na wschód od Koszalina zaklasyfikowała się jako gmina o średnim poziomie rozwoju.

```
#> tweaking randomForest
```



Rysunek 3.3: *Poziom rozwoju gmin w województwie Zachodniopomorskim (2018)*



Rysunek 3.4: *Poziom rozwoju gmin w województwie Zachodniopomorskim (2019)*

Rozdział 4

Podsumowanie

W pracy potwierdzono, że metoda k-średnich jest dobrym narzędziem dla stosunkowo niewielkiej liczby obserwacji (114). Przedstawiono pomocne metody wyliczenia optymalnej liczby skupień. W pracy zdecydowano się na 3 skupienia, jednak opierając się na metodach wyboru optymalnej liczby skupień ciekawe byłyby dalsze badania oparte na 4 lub 5 skupieniach.

Badanie poziomu istotności czynników podczas klasyfikacji pokazało jak istotna dla poziomu rozwoju jest zwarta tkanka przestrzenna, która ujęta została w takich czynnikach jak *udział osób korzystających z instalacji gazowej w ogóle populacji [%]*, *% mieszkań posiadających centralne ogrzewanie*, *różnica pomiędzy odsetkiem ludności korzystającej z wodociągu i z kanalizacji* oraz *dochody z podatku rolnego na 1 mieszkańca [zł]*. Prócz powyższych istotny wpływ na poziom rozwoju mają „twarde"wskaźniki dotyczące życia gospodarczego takie jak *osoby fizyczne prowadzące działalność gospodarczą na 1000 ludności*, *podmioty gospodarcze w sekcjach J-N (usługi, specjaliści, informatyka) na 1000 mieszkańców*, *pracujący na 1000 osób w wieku produkcyjnym*. Zgodnie z badaniami dominującym aspektem poziomu rozwoju jest kapitał materialny, a za nim kapitał finansowy. Warto zauważyć duże znaczenie innowacji na rozwój, bowiem mimo stosunkowo małego wpływu na poziom rozwoju to był on wyższy lub porównywalny od kapitału społecznego oraz ludzkiego i to pomimo jedynie dwóch wskaźników jakie przekładały się na poziom innowacji. Pokazało to, że słusznym okazało się zawarcie do

danych dotyczących podmiotów gospodarczych w sekcjach J-N (usługi, specjaliści, informatyka) na 1000 mieszkańców do wskaźników zaproponowanych przez Pana Doktora Perdała (Perdał (2018)).

Dalszych badań wymaga dobór danych do konstrukcji syntetycznego wskaźnika rozwoju, bowiem jak zauważono w badaniu istotny wpływ na poziom rozwoju ma jedynie część z dobranych wskaźników. Należałoby zwłaszcza przyjrzeć się innym możliwym wskaźnikom mogącym składać się na poziom innowacyjności. Możliwe, że część otrzymanych wyników dotyczących istotności badanych wskaźników wynika ze specyfikacji województwa Zachodniopomorskiego, dlatego w dalszych badaniach należałoby objąć cały kraj.

Jednoczynnikowa analiza wariancji pozwoliła prześledzić polaryzacje poziomu rozwoju gmin w województwie Zachodniopomorskim. Na przestrzeni lat 2016-2019 ustalono, że różnica między średnim poziomem rozwoju i wysokim jest znacząco wyższa niż między średnim poziomem rozwoju i niskim. Świadczy to o rosnącej dysproporcji w poziomie rozwoju gmin. Dzięki temu ustalono, że między rokiem 2016, a 2017 różnica między poszczególnymi poziomami rozwoju znacząco spadła. Natomiast w roku 2018, 2019 polaryzacja poziomu rozwoju wzrastała. Różnica poziomów rozwoju w 2019 roku była wyższa niż w 2016.

Zaobserwowano także jak niezwykle wszechstronne jest środowisko R. Dzięki mnogości pakietów, pozwala zarówno na transformacje, eksploracje, analizę danych jak i również daje narzędzia do wizualizacji przestrzennej. Czyni to środowisko R użytecznym narzędziem do badań nad poziomem rozwoju społeczno-gospodarczym.

Bibliografia

Adamczyk, A i S Franek (2012). Badanie zróżnicowania rozwoju ekonomicznego powiatów województwa zachodniopomorskiego. PL. *Zeszyty Naukowe Uniwersytetu Szczecińskiego. Ekonomiczne Problemy Usług* (nr 98 Przedsiębiorczość szansą rozwoju regionu. T. 2, Kształtowanie przedsiębiorczości), 405–417. (Term. wiz. 08. 09. 2021).

Auguie, B (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3. https://CRAN.R-project.org/package=gridExtra.

Borys, T (1999). *Wskaźniki ekorozwoju*. Polish. OCLC: 749760853. Białystok: Wydaw. Ekonomia i Środowisko.

Breiman, L, A Cutler, A Liaw i M Wiener (2018). *randomForest: Breiman and Cutler's Random Forests for Classification and Regression*. R package version 4.6-14. https://www.stat.berkeley.edu/~breiman/%20RandomForests/.

Chen, T, T He, M Benesty, V Khotilovich, Y Tang, H Cho, K Chen, R Mitchell, I Cano, T Zhou, M Li, J Xie, M Lin, Y Geng i Y Li (2021). *xgboost: Extreme Gradient Boosting*. R package version 1.4.1.1. https://github.com/dmlc/xgboost.

Churski, P, B Konecka-Szydłowska, R Perdał i T Herodowicz (2020). *Teoretyczny i praktyczny wymiar polityki rozwoju zorientowanej terytorialnie = (Theoretical and practical dimension of place-base territorial policy)*. Streszczenie w języku angielskim. OCLC: 1253419417. Warszawa: Polska Akademia Nauk. Komitet Przestrzennego Zagospodarowania Kraju.

Czyżycki, R (2006). Rozwój społeczno-gospodarczy gmin województwa zachodniopomorskiego. pl. (Term. wiz. 08. 09. 2021).

Ester, M, HP Kriegel, J Sander i X Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. W: *Proceedings of the Second International*

Conference on Knowledge Discovery and Data Mining. KDD'96. Portland, Oregon: AAAI Press, pp.226–231. (Term. wiz. 13. 09. 2021).

Fox, J, S Weisberg i B Price (2021). *car: Companion to Applied Regression*. R package version 3.0-11. https://CRAN.R-project.org/package=car.

Gmina Ostrowice zostanie zniesiona - Ministerstwo Spraw Wewnętrznych i Administracji - Portal Gov.pl (2021). pl-PL. https://www.gov.pl/web/mswia/gmina-ostrowice-zostanie-zniesiona (term. wiz. 27. 08. 2021).

Grzybowska, U i M Karwański (2015). Szacowanie parametrów ryzyka kredytowego przy użyciu rodzin klasyfikatorów. PL. *Studia Ekonomiczne* **248**, 107–120. (Term. wiz. 14. 09. 2021).

Hahsler, M i M Piekenbrock (2021). *dbscan: Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms*. R package version 1.1-8. https://github.com/mhahsler/dbscan.

Hothorn, T, F Bretz i P Westfall (2021). *multcomp: Simultaneous Inference in General Parametric Models*. R package version 1.4-17. https://CRAN.R-project.org/package=multcomp.

Hull, Z (2007). Czy idea sustainable development ukazuje nową wizję rozwoju cywilizacyjnego? PL. *Problemy Ekorozwoju* **Vol. 2**(nr 1), 49–57. (Term. wiz. 07. 09. 2021).

Hypothesis Testing - Analysis of Variance (ANOVA) (2021). https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_hypothesistesting-anova/bs704_hypothesistesting-anova_print.html (term. wiz. 15. 09. 2021).

K-means Cluster Analysis · UC Business Analytics R Programming Guide (2021). https://uc-r.github.io/kmeans_clustering (term. wiz. 13. 09. 2021).

Kassambara, A (2020). *ggpubr: ggplot2 Based Publication Ready Plots*. R package version 0.4.0. https://rpkgs.datanovia.com/ggpubr/.

Kassambara, A i F Mundt (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.7. http://www.sthda.com/english/rpkgs/factoextra.

Korzeniewski, J (2014). Index of the Choice of the Number of Clusters. pl. *Przegląd Statystyczny. Statistical Review* **61**(2). (Term. wiz. 13. 09. 2021).

Krzyśko, M, W Wołyński, T Górecki i M Skorzybut (2008). *Systemy uczące się. Rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości*.

Migdał-Najman, K i K Najman (2013). Analiza porównawcza wybranych metod analizy skupień w grupowaniu jednostek o złożonej strukturze grupowej. PL. *Zarządzanie i Finanse* (R. 11, nr 3, cz. 2), 179–194. (Term. wiz. 13. 09. 2021).

Morissette, L i S Chartier (2013). The k-means clustering technique: General considerations and implementation in Mathematica. W:

Opałło, M (1972). *Mierniki rozwoju regionów*. Warszawa: Państwowe Wydawnictwo Ekonomiczne. https://books.google.pl/books?id=azgDAAAAMAAJ.

Pająk, K, P Dahlke i O Kvilinskyi (2016). Determinanty rozwoju regionalnego - współczesne odniesienie. PL. *Roczniki Ekonomiczne Kujawsko-Pomorskiej Szkoły Wyższej w Bydgoszczy* (nr 9), 109–122. (Term. wiz. 07. 09. 2021).

Paluszynska, A, P Biecek i Y Jiang (2020). *randomForestExplainer: Explaining and Visualizing Random Forests in Terms of Variable Importance*. R package version 0.10.1. https://github.com/ModelOriented/%20randomForestExplainer.

Perdał, R (2018). ZASTOSOWANIE ANALIZY SKUPIEŃ I LASÓW LOSOWYCH W KLASYFIKACJI GMIN W POLSCE NA SKALI POZIOMU ROZWOJU SPOŁECZNO-GOSPODARCZEGO. *Metody Ilościowe w Badaniach Ekonomicznych* **19**(3), 263–273. (Term. wiz. 26. 08. 2021).

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Schauberger, P i A Walker (2021). *openxlsx: Read, Write and Edit xlsx Files*. R package version 4.2.4. https://CRAN.R-project.org/package=openxlsx.

Test ANOVA Kruskala-Wallisa | Statystyka – Porady | Analizy | Opracowania | Obliczenia | Pomoc statystyczna (2021). https://www.statystyka.az.pl/test-anova-kruskala-wallisa.php (term. wiz. 15. 09. 2021).

Tibshirani, R, G Walther i T Hastie (2001). Estimating the number of clusters in a data set via the gap statistic. en. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**(2), 411–423. (Term. wiz. 13. 09. 2021).

Wickham, H, W Chang, L Henry, TL Pedersen, K Takahashi, C Wilke, K Woo, H Yutani i D Dunnington (2021a). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. R package version 3.3.5. https://CRAN.R-project.org/package=ggplot2.

Wickham, H, R François, L Henry i K Müller (2021b). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.7. https://CRAN.R-project.org/package=dplyr.