Raport

Weronika Łoś

June 2023

Spis treści

1 Wstęp

W ramach kursu Metody Systemowe i Decyzyjne podjęłam się projektu dotyczącego dostępu do antykoncepcji oraz wpływu różnych czynników na wskaźnik dzietności. Celem mojej pracy jest dokładne zbadanie korelacji między czynnikami, takimi jak dostęp do edukacji i antykoncepcji, a współczynnikiem dzietności w wybranych krajach na przestrzeni lat . Poprzez przeprowadzenie analizy, dążę do wyciągnięcia trafnych wniosków na temat tych zależności.

2 Dane

W ramach projektu wybrałam dwa zbiory danych w formacie CSV, do których miałam dostęp dzięki stronie Our World in Data. Linki do stron, z których pobierałam zbiory umieściłam na końcu dokumentu.

Pierwszy zbiór zawiera informacje dotyczące stosowanych metod antykoncepcji, przedstawione w czterech kolumnach. Pierwsza kolumna zawiera nazwę kraju, którego dotyczą dane, np. "Polska". Druga kolumna zawiera kod państwa,

np. "POL". Trzecia kolumna zawiera informację o roku, z którego pochodzą dane. Ostatnia kolumna prezentuje procentowy udział zamężnych kobiet w wieku 15-49, które korzystają - lub których partnerzy korzystają - z dowolnej z nowoczesnych lub tradycyjnych metod antykoncepcji.

Nowoczesne metody antykoncepcji obejmują sterylizację kobiet i mężczyzn, doustne tabletki antykoncepcyjne, wkładki wewnątrzmaciczne (IUD), prezerwatywy męskie, zastrzyki, implanty (takie jak Norplant), metody barierowe dopochwowe, prezerwatywy dla kobiet oraz antykoncepcję awaryjną. Tradycyjne metody antykoncepcji obejmują okresową abstynencję, odstawienie i inne tradycyjne metody.

Przykład danych z pierwszego zbioru:

Entity	Code	Year	Contraceptive prevalence(%)
Afghanistan	AFG	2000	5.3
Colombia	COL	2015	81

Drugi zbiór danych zawiera informacje dotyczące wskaźnika dzietności oraz średniej liczby lat nauki, uporządkowane w siedmiu kolumnach. Kolejno przedstawione są informacje dotyczące: kraju, kodu kraju, roku, wskaźnika dzietności, średniej długości edukacji, wielkości populacji w danym roku oraz kontynentu.

Przykład danych z drugiego zbioru:

Entity	Code	Year	Fertility rate	Total years of schooling	Population	Continent
Netherlands	NLD	1985	1.51	10.22	14515705	Europe
Costa Rica	CRI	2010	1.92	8.43	4622250	North America

3 Metody analizy

Aby dokonać analizy podanych zbiorów danych, znalezienia korelacji między nimi oraz ich wizualizacji, zastosowałam następujące metody:

- Przedstawienie i porównanie danych na wykresie
- Obliczenie współczynników korelacji
- Stworzenie modelu klasyfikacji i obliczenie jego dokładności

Wybrałam powyższe metody, ponieważ umożliwią mi one badanie zależności między danymi zawartymi w wybranych zbiorach. Przedstawienie danych na wykresie pozwoli mi porównać zmiany wartości wskaźnika dzietności w różnych krajach na przestrzeni lat, co dostarczy informacji na temat szybkości spadku tego wskaźnika oraz np. momentu, w którym zaczynał maleć w poszczególnych

krajach. Grupowanie danych i tworzenie wykresów pozwoli mi również zidentyfikować różnice w wartości średniego wskaźnika dzietności oraz dostępu do antykoncepcji w danym roku na poszczególnych kontynentach.

Na podstawie obliczonego współczynnika korelacji będę w stanie określić, czy istnieje silna korelacja między dostępem do antykoncepcji a wskaźnikiem dzietności, czy jest ona nieznacząca. Podobną zależność zbadam dla związku pomiędzy średnią długością edukacji szkolnej w latach a wskaźnikiem dzietności. Do powyższych analiz skorzystam z bibliotek pandas, matplotlib oraz seaborn.

Zastosowanie modelu klasyfikacji pozwoli na przewidzenie, czy dany kraj ma wysoki czy niski wskaźnik dzietności na podstawie dostępnych danych dotyczących antykoncepcji i edukacji, tutaj użyję KNeighborsClassifier z biblioteki sklearn, który implementuje algorytm k- najbliższych sąsiadów.

4 Eksploracja danych

Podzieliłam eksplorację danych na kroki:

1. Przygotowanie danych do analizy

   Aby wczytać dane z dwóch plików źródłowych, skorzystałam z funkcji read_csv z biblioteki pandas i przypisałam je do dwóch zmiennych. Następnie, połączyłam te zbiory danych, aby móc wykorzystać i porównywać dane z obu zbiorów w dalszej analizie. Do tego celu użyłam funkcji merge również z biblioteki pandas, wykonując scalanie na podstawie kolumny zawierającej informację o kraju i roku otrzymując tym samym plik merged.csv.

   Fragment pliku merged.csv:

Lp.	Entity	Code	Year	C.p.	F. rate	Total years of schooling	Population	Cont.
2	Afghanistan	AFG	2005	13.6	6.83	3.32	24411196	Asia

   Gdzie C.p. - Contraceptive prevalence, F. rate - Fertility rate,
   Cont. - Continent

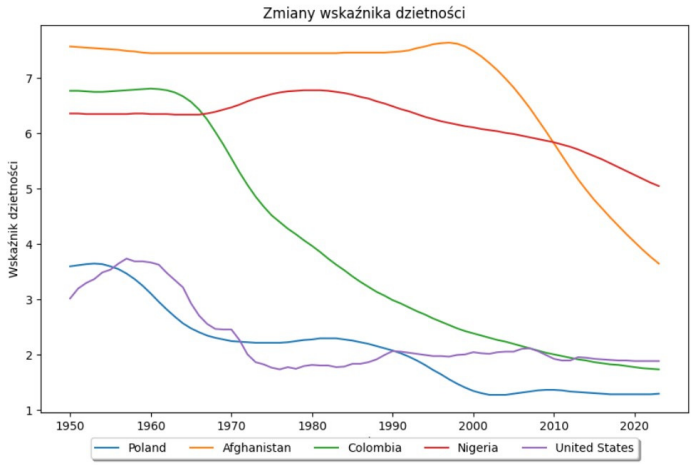
   W celu wykonania punktu nr 5 przeprowadziłam operację filtracji danych, aby wybrać tylko te dotyczące roku 2015. Następnie, grupowałam te dane według kontynentów, obliczając jednocześnie średni współczynnik dzietności w roku 2015 biorąc pod uwagę wszystkie kraje na danym kontynencie. W efekcie uzyskałam plik continent_data.csv zawierający dane dotyczące sześciu kontynentów.

   Fragment pliku continent_data.csv:

Continent	Year	Contraceptive prevalence(%)	Fertility rate	Population
Oceania	2015	29.3	3.91	612670

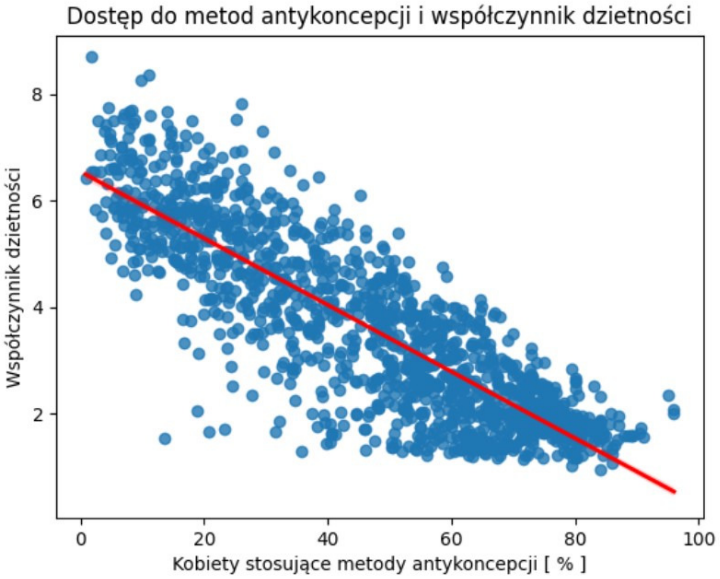
2. Stworzenie wykresów przedstawiających zmiany wskaźnika dzietności w 5 krajach na przestrzeni lat od 1950 do 2023.

   Żeby przedstawić dane na wykresie, skorzystałam z funkcji dostępnych w bibliotece matplotlib. Na wykresie uwzględniłam dane dla pięciu krajów: Polski, Afganistanu, Kolumbii, Nigerii i Stanów Zjednoczonych. Wybór tych konkretnych krajów wynikał z faktu, że reprezentują one różne części świata, charakteryzują się odmiennymi kulturami, religiami oraz zróżnicowanym poziomem rozwoju.



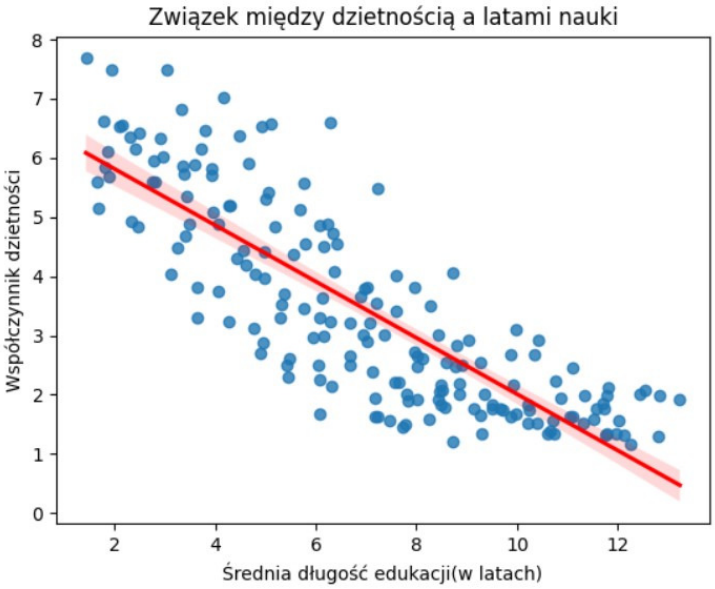
3. Stworzenie wykresu przedstawiającego związek między wskaźnikiem dzietności a stosowaniem antykoncepcji oraz obliczenie współczynnika korelacji.

   W celu zbadania zależności między dostępem do antykoncepcji a dzietnością skorzystałam z funkcji regplot dostępnej w bibliotece seaborn, która wyświetla linię regresji. Wybrałam odpowiednie kolumny dotyczące poziomu dostępu do antykoncepcji wyrażonego w procentach oraz poziomu dzietności ze scalonego wcześniej zbioru danych. Po przekształceniu wartości w tych kolumnach na typ float, przedstawiłam je na wykresie i obliczyłam współczynnik korelacji między tymi danymi, który wyniósł około -0.85.



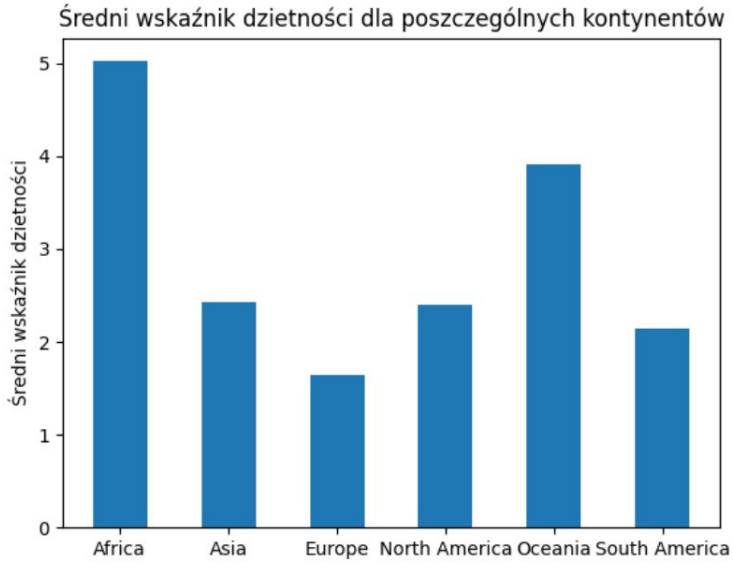
4. Stworzenie wykresu przedstawiającego związek między wskaźnikiem dzietności a liczbą lat edukacji oraz obliczenie współczynnika korelacji.

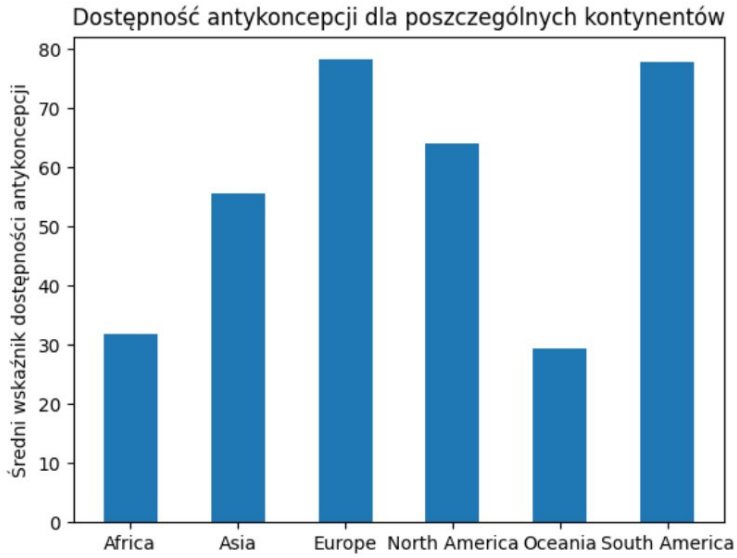
   Aby przedstawić na wykresie dane dotyczące związku pomiędzy edukacją a dzietnością ponownie użyłam funkcji regplot z biblioteki seaborn. Wybrałam odpowiednie kolumny, tym razem dotyczące średniej długości edukacji w latach oraz poziomu dzietności. Obliczony współczynnik korelacji wyniósł w tym przypadku około -0.83.



5. Porównanie średnich wskaźników dzietności oraz dostępności antykoncepcji dla poszczególnych kontynentów w roku 2015 i przedstawienie wyników na wykresach.

   W celu porównania danych dla poszczególnych kontynentów dla roku 2015 wykorzystałam przygotowany przeze mnie zbiór continent_data.csv. Dane przedstawiłam na wykresach słupkowych wykorzystując do tego funkcję bar z biblioteki matplotlib.


Średni wskaźnik dzietności dla poszczególnych kontynentów
Średni wskaźnik dzietności
0
1
2
3
4
5
Africa
Asia
Europe
North America
Oceania
South America

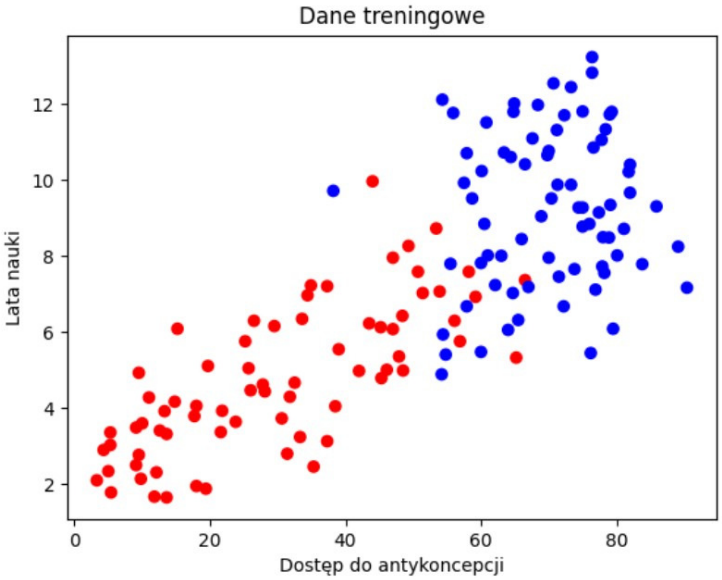

Dostępność antykoncepcji dla poszczególnych kontynentów
Średni wskaźnik dostępności antykoncepcji
0
10
20
30
40
50
60
70
80
Africa
Asia
Europe
North America
Oceania
South America

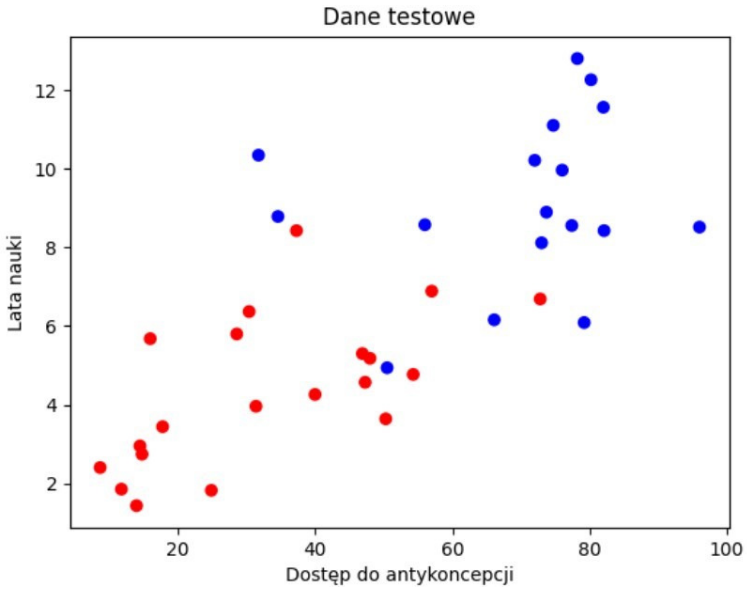
5 Model

Stworzyłam modelu klasyfikacji, który przewiduje, czy dany kraj ma wysoki czy niski wskaźnik dzietności, opierając się na danych dotyczących lat nauki i dostępności do antykoncepcji. Podczas tworzenia modelu nie jest uwzględniana informacja o kontynencie.

Aby zbudować model klasyfikacji do przewidywania wysokiego lub niskiego wskaźnika dzietności na podstawie dostępu do antykoncepcji oraz lat nauki użyłam metody KNN i KNeighborsClassifier z biblioteki sklearn. Po wyodrębnieniu potrzebnych kolumn, czyli 'Contraceptive prevalence', 'Fertility rate' i 'Total years of schooling' oraz konwersji na odpowiednie typy danych(float), dokonałam zamiany danych z kolumny 'Fertility rate' na etykiety "wysoki" oraz "niski". Na potrzeby analizy przyjęłam, że wskaźnik dzietności większy niż 3 uważam za wysoki, a mniejszy niż 3- niski.

Następnie wykonałam podział na dane treningowe oraz testowe. Ziarno losowości ma wartość 42, natomiast parametr test_size został przeze mnie ustawiony na 0.2, co oznacza, że 20% danych zostało przypisanych do zbioru testowego, a 80% stanowi zbiór treningowy. Następnie wykonałam dopasowanie danych treningowych do modelu oraz predykcję dla danych testowych. Po obliczeniu dokładności(accuracy) modelu, otrzymałam dokładność równą w przybliżeniu 0.84. Aby lepiej zobrazować dane treningowe oraz testowe, wykonałam wykresy, na których kolorem niebieskim oznaczono niski wskaźnik dzietności (mniejszy niż 3), a czerwonym wysoki (większy niż 3).





6 Wnioski

Na podstawie analizy danych, która obejmowała dokładne przyjrzenie się zmianom wskaźnika dzietności w wybranych krajach na przestrzeni lat, zbadanie zależności między dostępem do antykoncepcji a wskaźnikiem dzietności oraz zastosowanie modelu klasyfikacji, można wysunąć kilka wniosków:

1. Wnioski dotyczące eksploracji danych

   (a) Zmiany wskaźnika dzietności
   Na wykresie przedstawiającym zmiany wskaźnika dzietności w pięciu krajach można zauważyć różnice w trendach. Kraje takie jak Polska, czy Stany Zjednoczone już w roku 1950 charakteryzowały się znacznie mniejszym współczynnikiem dzietności niż Nigeria, Kolumbia, czy Afganistan. Można zauważyć, że w Polsce, Stanach Zjednoczonych oraz Kolumbii współczynnik ten około roku 2000 stabilizuje się, a w Nigerii i Afganistanie nadal jest wysoki i ciągle spada. Obrazuje to między innymi różnice w rozwoju tych krajów.

   (b) Zależność między dostępem do antykoncepcji a wskaźnikiem dzietności
   Analiza danych pokazuje, że istnieje negatywna korelacja(około -0.85) między dostępem do antykoncepcji a wskaźnikiem dzietności. Wyższy poziom dostępności do antykoncepcji często wiąże się z niższym wskaźnikiem dzietności. To sugeruje, że dostęp do skutecznych me-

tod antykoncepcji może wpływać na kontrolowanie wzrostu populacji i zmniejszenie liczby dzieci na rodzinę.

(c) Zależność między długością edukacji a wskaźnikiem dzietności
Tutaj również istnieje wysoka negatywna (około -0.83) korelacja między danymi. Dłuższa średnia edukacja często wiąże się z niższym wskaźnikiem dzietności. Wobec tego można wnioskać, że poziom edukacji w danym państwie ma znaczący wpływ na współczynnik dzietności, a tym samym na przyrost naturalny w tym państwie.

(d) Porównanie danych dla kontynentów
Na podstawie analizy wykresów możemy stwierdzić, że współczynnik dzietności oraz dostęp do metod antykoncepcji różni się nie tylko przy podziale na kraje, ale również przy podziale na kontynenty. Widać istotne różnice między kontyntami bardziej rozwiniętymi, takimi jak Europa, a kontynentami rozwijającymi się, takimi jak Afryka.

2. Wnioski dotyczące modelu klasyfikacji
Model klasyfikacji został wykorzystany do przewidywania, czy dany kraj ma wysoki czy niski wskaźnik dzietności na podstawie danych dotyczących dostępu do antykoncepcji i lat nauki.

   Można użyć takiego modelu, aby przewidzieć, czy w danym kraju należy skoncentrować się na zwiększaniu dostępu do antykoncepcji lub przekazywaniu większych funduszy na rozwój oświaty w celu zmniejszenia wskaźnika dzietności.

   W przypadku tej konretnej analizy, otrzymany wynik (dopasowanie około 0.84) pozwala stwierdzić, że na podstawie danych dotyczących lat nauki i antykoncepcji jesteśmy w stanie z dużym prawdopodobieństwem przewidzieć, czy dany kraj cechuje się wysokim czy niskim współczynnikiem dzietności.

7 Źródła i linki

Linki do danych, z których korzystałam:

- Zbiór pierwszy
- Zbiór drugi