Jak różne czynniki wpływają na długość życia: Analiza wielowymiarowa

Marcin Fedorowicz 266852

5 czerwca 2023

1 Wprowadzenie

Problemem wybranym do badań jest zależność przewidywanej długości życia od innych czynników. Inspiracją do tego rodzaju rozważań była obserwacja szybko rozwijającego się świata oraz idących za tym problemów. Analiza zbioru danych może wykazać, czy poziom zanieczyszczenia powietrza oraz na przykład choroby psychiczne mają rzeczywisty wpływ na długość życia człowieka. Celem projektu jest zbadanie zależności przewidywanej długości życia od następujących parametrów:

- % społeczenstwa z problemami z używkami [6],
- liczba ludności [10],
- GDP per capita [9],
- średnie deklarowane szczęście [3],
- % społeczeństwa używającego internetu [8],
- powierzchnia kraju [7],
- mediana wieku [4],
- liczba przypadków zaburzeń psychicznych [1],
- zanieczyszczenie powietrza [5],
- gęstość zaludnienia [10],
- liczba samobójstw [2]

oraz stworzenie modelu estymującego przewidywaną długość życia na podstawie wybranych parametrów.

2 Zbiór danych i jego przetwarzanie

2.1 Zbiór danych

Zbiór danych został pozyskany ze strony internetowej ourworldindata.org za pomocą narzędzia Webdriver z biblioteki Selenium oraz BeautifulSoup z biblioteki Bs4. Zawiera on informacje na temat wyników badań przewidywanej przy narodzeniu długości życia, % społeczenstwa z problemami z używkami, liczby ludności, GDP per capita, średniego deklarowanego szczęścia, % społeczeństwa używającego internetu, powierzchni kraju, mediany wieku, liczby przypadków zaburzeń psychicznych, zanieczyszczenia powietrza, gęstości zaludnienia, liczby samobójstw. Wszystkie dane pochodzą z lat 2010-2017, gdzie dane z przedziału 2010-2016 służą jako zestawy treningowe, a te pochodzące z roku 2017 służą do testów jakości wytrenowanego modelu.

2.2 Przetwarzanie wstępne

Do konstrukcji modelu wybrane zostały dane z lat 2010-2016. Z powodu brakujących danych dotyczących niektórych krajów za pomocą biblioteki Pandas usunięta została część danych w celu pozostawienia jedynie krajów z pełnym zestawem czynników kluczowych.

2.3 Analiza eksploracyjna

Tabela 1 i 2 przedstawia wycinek danych po wstępnym ich przetworzeniu. Do wykonaniu analizy danych użyto biblioteki matplotlib.pyplot.

Rok	Kraj	Problemy z używkami	Liczba ludności	GDP per capita	Poziom szczęścia	Dostęp do internetu
2010	Afghanistan	1.00	28189672.00	1628.00	3.78	4.00
2010	Albania	2.20	2913402.00	9223.00	5.51	45.00
2010	Algeria	0.90	35856348.00	12588.00	5.60	12.50
2010	Angola	1.40	23364196.00	7521.00	4.36	2.80
2010	Argentina	2.90	41100124.00	18980.00	6.47	45.00

Tabela 1: Wycinek danych - część 1

Powierzchnia	Długość życia	Mediana wieku	Choroby psychiczne	Zanieczyszczenie powietrza	Gęstość zaludnienia	Liczba samobójstw
652860.00	60.90	14.40	16.88	65.25	43.22	1245.00
27400.00	77.90	32.30	10.85	21.28	106.33	231.00
2381741.00	73.80	25.10	13.94	33.64	15.05	1089.00
1246700.00	56.70	16.00	13.20	33.79	18.74	1629.00
2736690.00	75.70	28.90	12.60	16.86	15.02	3495.00

Tabela 2: Wycinek danych - część 2

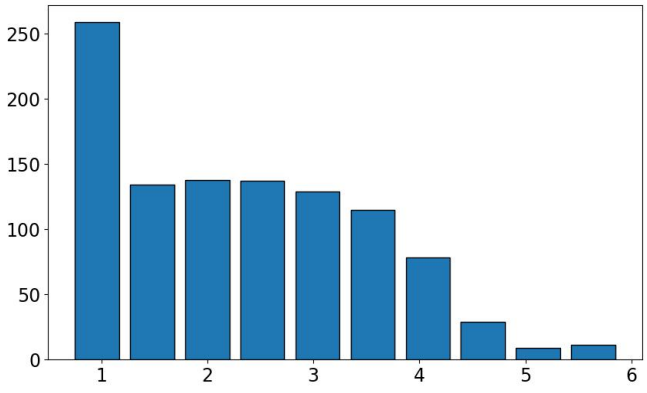
2.3.1 Analiza poszczególnych parametrów

1. Problemy z używkami

Wartość min.	Wartość maks.	Średnia	Wart. Unikalne
0.7	5.9	2.30	53

Tabela 3: Statystyka problemów z używkami

Rozkład danych



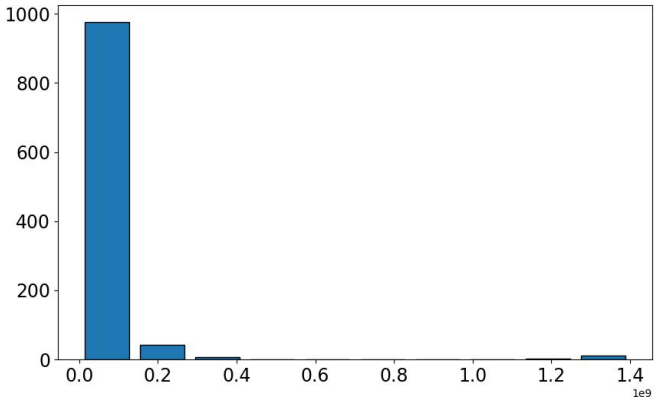
Rysunek 1: Histogram problemów z używkami

2. Liczba ludności

Wartość min.	Wartość maks.	Średnia	Wart. Unikalne
318346	1401889700	47347187.70	1039

Tabela 4: Statystyka liczby ludności

Rozkład danych



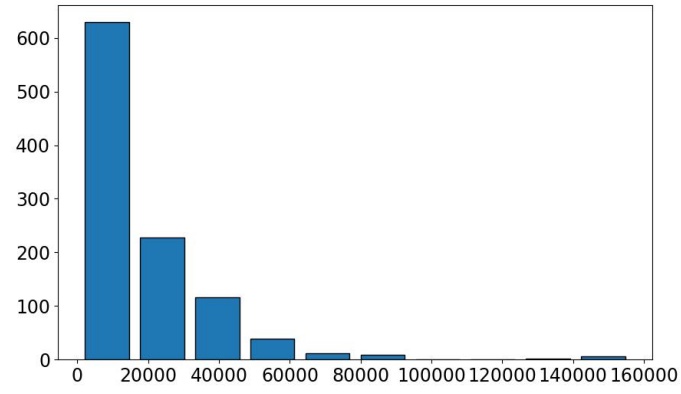
Rysunek 2: Histogram liczby ludności

3. GDP per capita

Wartość min.	Wartość maks.	Średnia	Wart. Unikalne
561	156299	17453.20	1014

Tabela 5: Statystyka GDP per capita

Rozkład danych



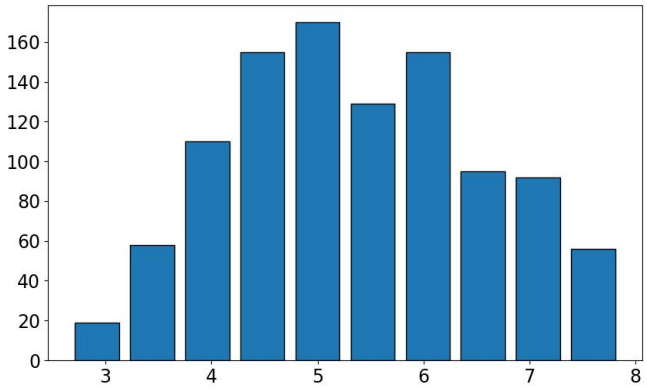
Rysunek 3: Histogram GDP per capita

4. Deklarowane szczęście

Wartość min.	Wartość maks.	Średnia	Wart. Unikalne
2.66	7.86	5.37	389

Tabela 6: Statystyka deklarowanego szczęścia

Rozkład danych



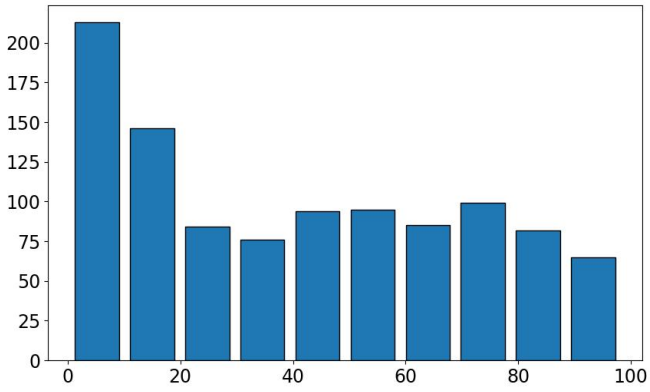
Rysunek 4: Histogram deklarowanego szczęścia

5. Dostęp do internetu

Wartość min.	Wartość maks.	Średnia	Wart. Unikalne
0.30	98.20	41.22	593

Tabela 7: Statystyka dostępu do internetu

Rozkład danych



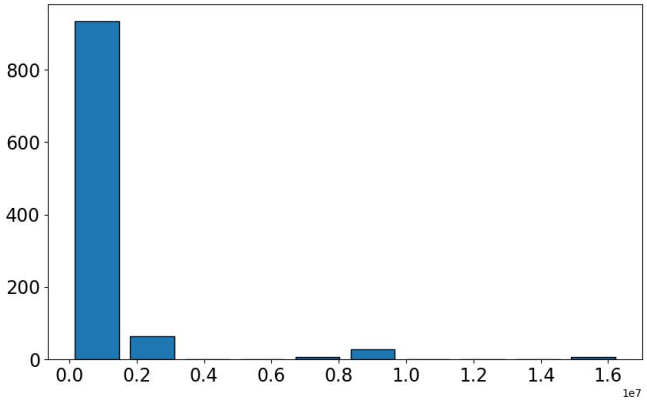
Rysunek 5: Histogram dostępu do internetu

6. Powierzchnia kraju

Wartość min.	Wartość maks.	Średnia	Wart. Unikalne
320.00	16376870.00	840848.82	149

Tabela 8: Statystyka powierzchni krajów

Rozkład danych



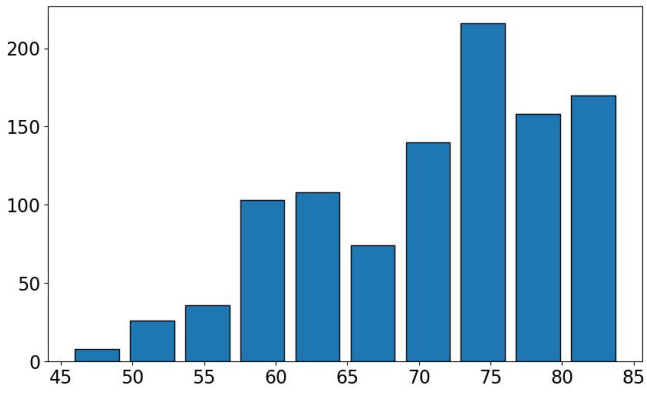
Rysunek 6: Histogram powierzchni krajów

7. Przewidywana długość życia

Wartość min.	Wartość maks.	Średnia	Wart. Unikalne
45.60	84.1	70.97	301

Tabela 9: Statystyka przewidywanej długości życia

Rozkład danych



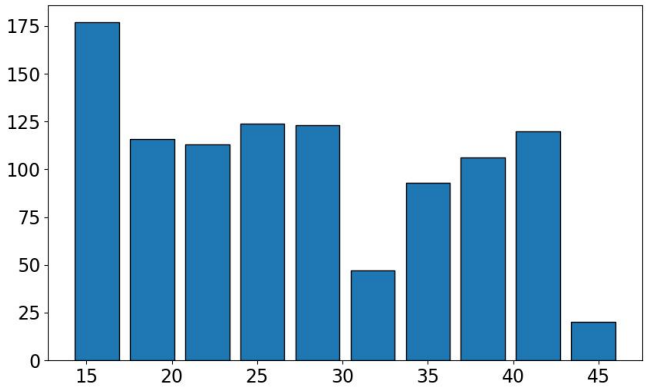
Rysunek 7: Histogram przewidywanej długości życia

8. Mediana wieku

Wartość min.	Wartość maks.	Średnia	Wart. Unikalne
14.00	46.30	27.75	292

Tabela 10: Statystyka mediany wieku

Rozkład danych



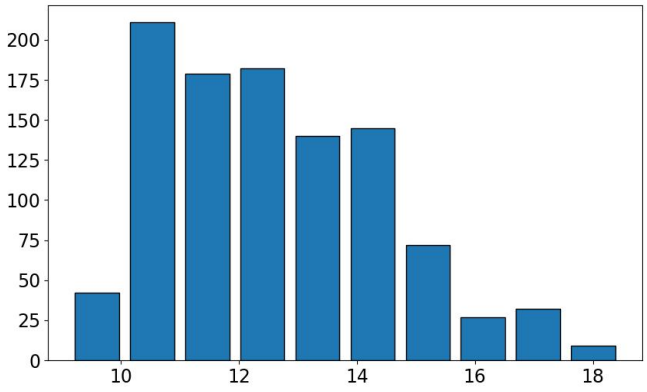
Rysunek 8: Histogram mediany wieku

9. Choroby psychiczne

Wartość min.	Wartość maks.	Średnia	Wart. Unikalne
9.13	18.47	12.66	494

Tabela 11: Statystyka chorób psychicznych

Rozkład danych



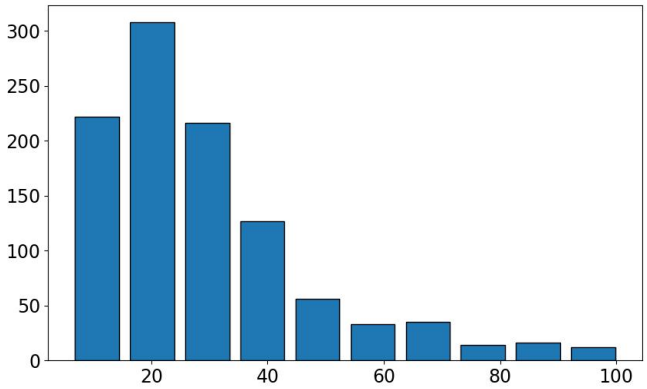
Rysunek 9: Histogram chorób psychicznych

10. Zanieczyszczenie powietrza

Wartość min.	Wartość maks.	Średnia	Wart. Unikalne
5.89	100.78	29.44	930

Tabela 12: Statystyka zanieczyszczenia powietrza

Rozkład danych



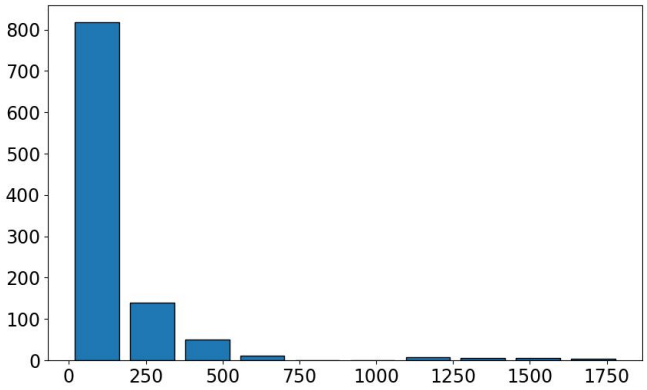
Rysunek 10: Histogram zanieczyszczenia powietrza

11. Zagęszczenie ludności

Wartość min.	Wartość maks.	Średnia	Wart. Unikalne
1.74	1795.76	138.41	1002

Tabela 13: Statystyka zagęszczenia ludności

Rozkład danych



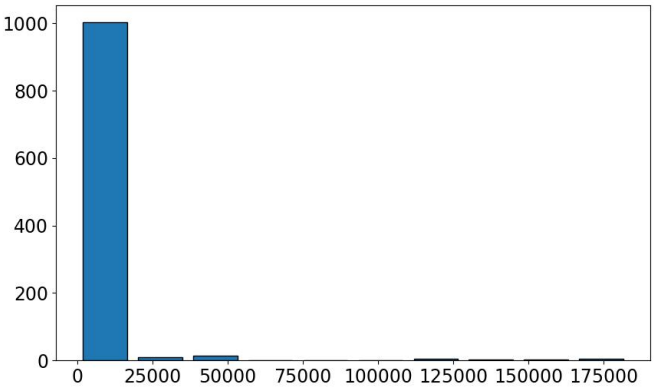
Rysunek 11: Histogram zagęszczenia ludności

12. Samobójstwa

Wartość min.	Wartość maks.	Średnia	Wart. Unikalne
25	183284	4669.92	870

Tabela 14: Statystyka samobójstw

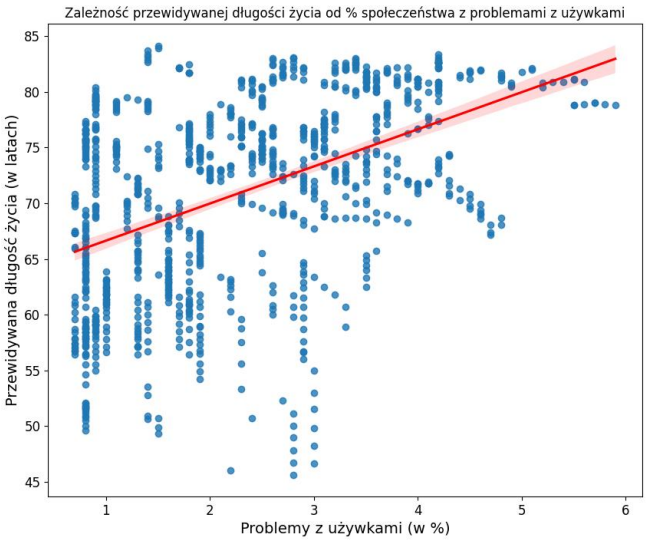
Rozkład danych



Rysunek 12: Histogram samobójstw

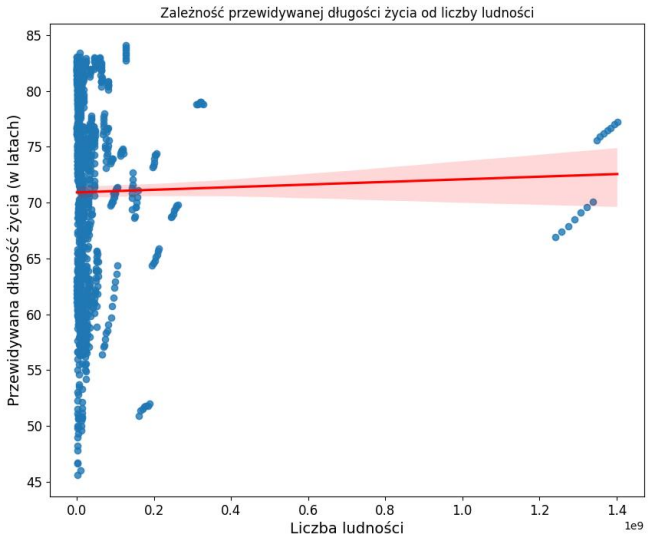
2.3.2 Zależności między danymi

1. Problemy z używkami



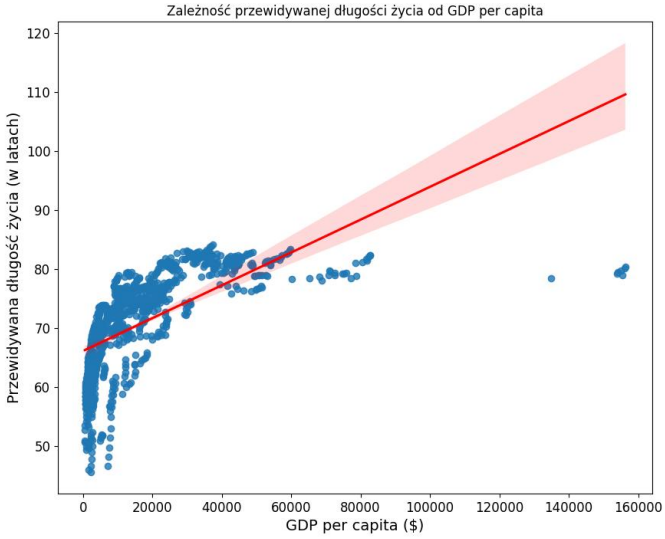
Rysunek 13: Zależność przewidywanej długości życia od % społeczeństwa z problemami z używkami

2. Liczba ludności



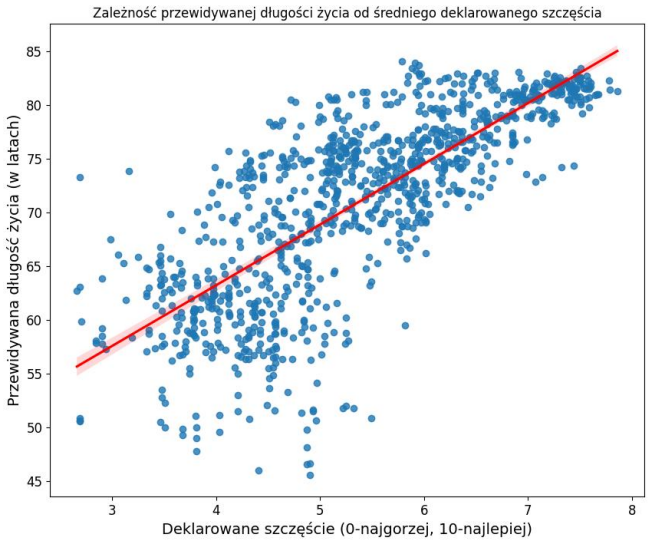
Rysunek 14: Zależność przewidywanej długości życia od liczby ludności

3. GDP per capita



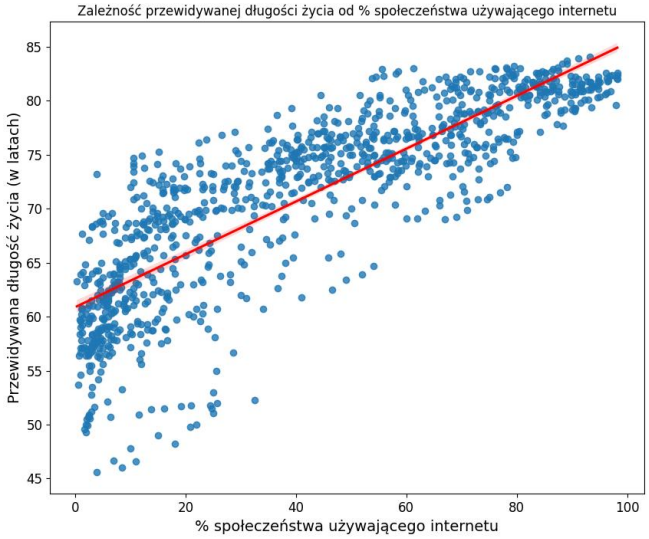
Rysunek 15: Zależność przewidywanej długości życia od GDP per capita

4. Deklarowane szczęście



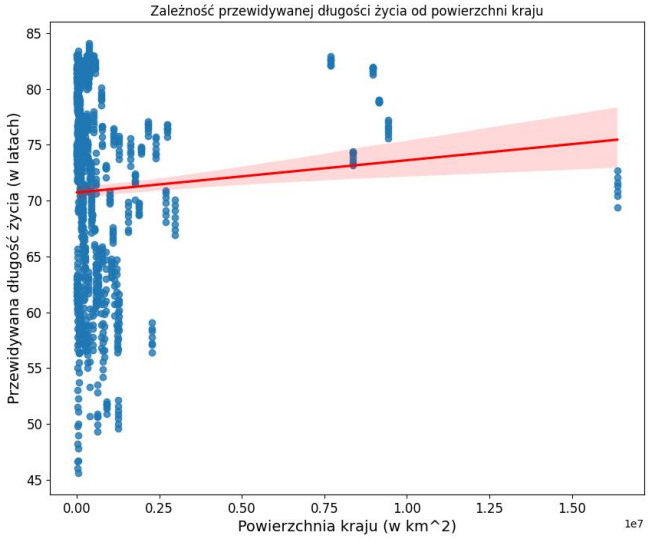
Rysunek 16: Zależność przewidywanej długości życia od średniego deklarowanego szczęścia

5. Dostęp do internetu



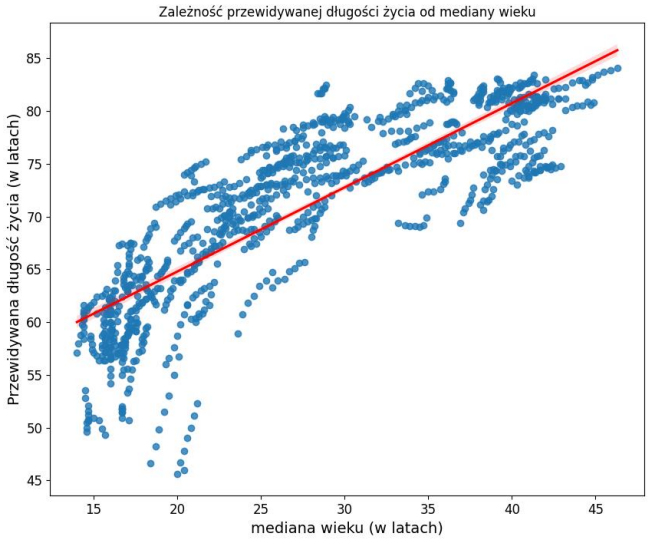
Rysunek 17: Zależność przewidywanej długości życia od % społeczeństwa używającego internetu

6. Powierzchnia kraju



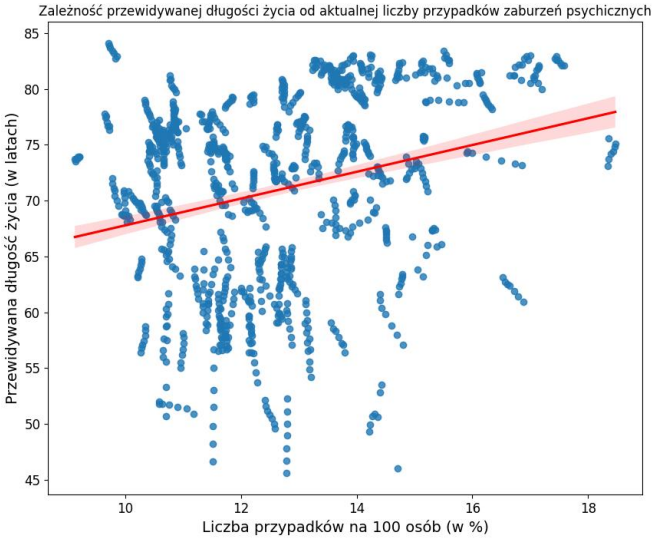
Rysunek 18: Zależność przewidywanej długości życia od powierzchni kraju

7. Mediana wieku



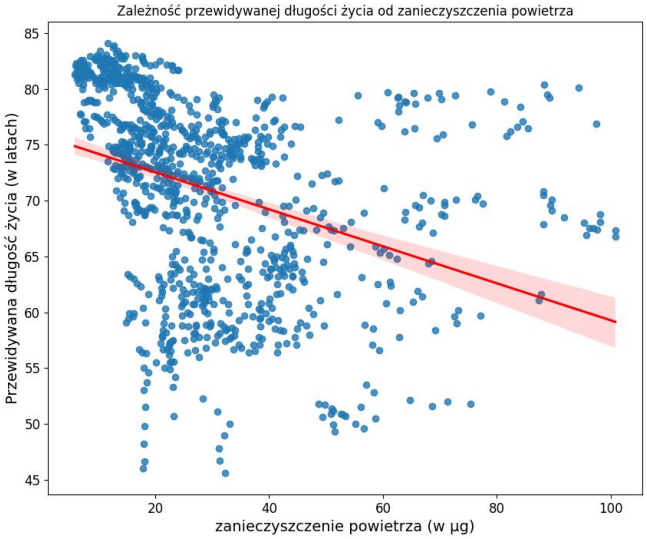
Rysunek 19: Zależność przewidywanej długości życia od mediany wieku

8. Choroby psychiczne



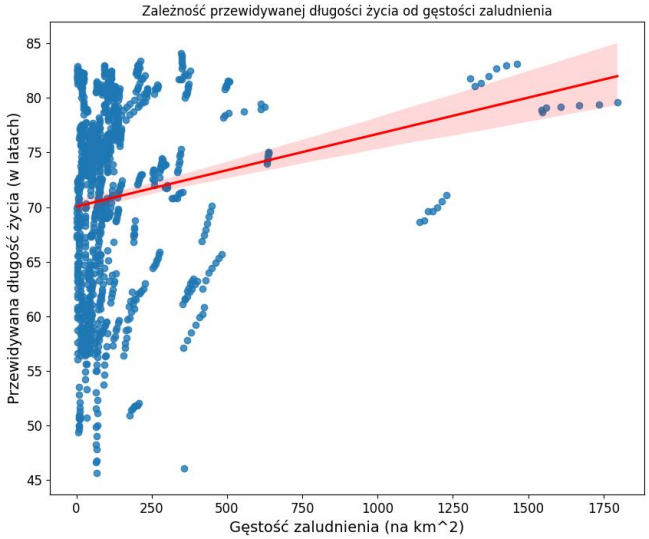
Rysunek 20: Zależność przewidywanej długości życia od aktualnej liczby przypadków zaburzeń psychicznych

9. Zanieczyszczenie powietrza



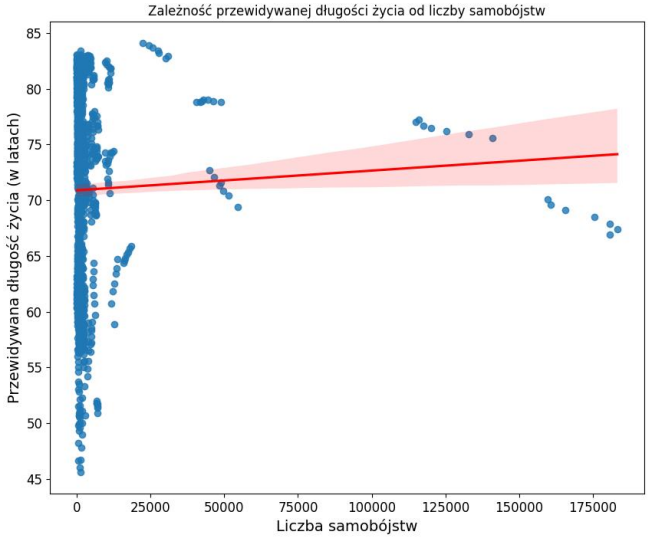
Rysunek 21: Zależność przewidywanej długości życia od zanieczyszczenia powietrza

10. Zagęszczenie ludności



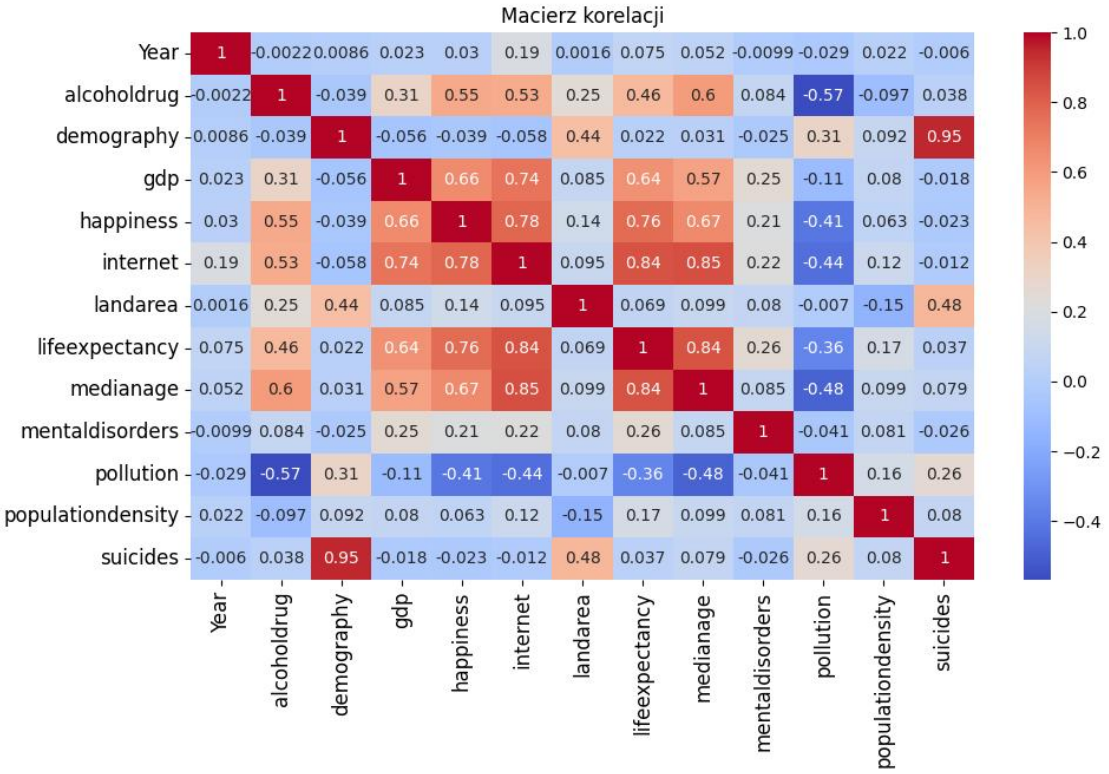
Rysunek 22: Zależność przewidywanej długości życia od gęstości zaludnienia

11. Samobójstwa



Rysunek 23: Zależność przewidywanej długości życia od liczby samobójstw

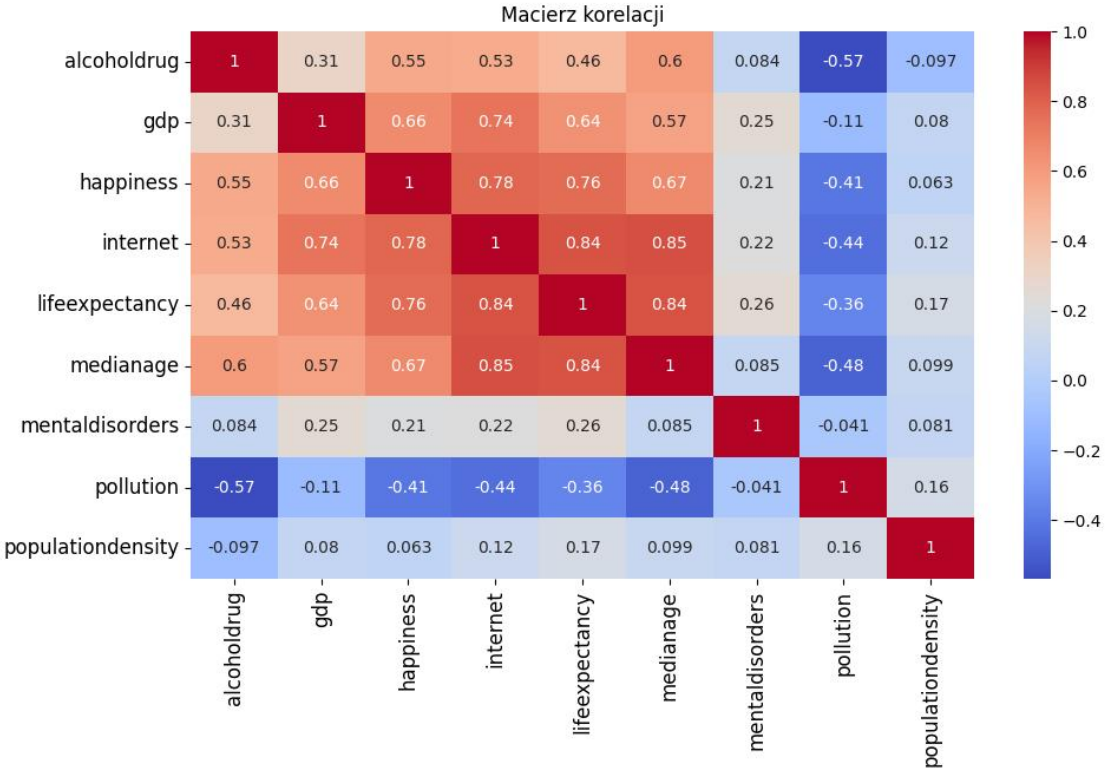
12. Macierz korelacji



Rysunek 24: Macierz korelacji zmiennych

2.4 Analiza wyników

Analiza wykazała wysoką zależność przewidywanej długości życia z parametrami takimi jak: problemy z używkami, GDP per capita, deklarowany poziom szczęścia, dostęp do internetu, mediana wieku, choroby psychiczne, zanieczyszczenie powietrza i gęstość zaludnienia. Zauważyć można również niewielką korelację przewidywanej długości życia od: roku badań, liczby ludności, powierzchni kraju i liczby samobójstw. Zależności te zostały zaprezentowane za pomocą macierzy korelacji na Rys. 24
W związku z małą istotnością części czynników zostaną one usunięte z tabeli czynników decyzyjnych. Parametry o wysokim znaczeniu dla dalszej części raportu można przedstawić ponownie za pomocą miacierzy korelacji na Rys. 25



Rysunek 25: Końcowa macierz korelacji zmiennych

3 Model

Wykorzystując przygotowane w poprzednich etapach dane możliwe jest podzielenie ich na dwie części, czynniki wpływające na estymowany parametr oraz jego faktyczną wartość. W tym przypadku danymi treningowymi dla modelu będą dane z lat 2010-2016, w sumie 1039 rekordów. Model bedzie uczył się przewidywać długość życia na podstawie wyodrębnionych czynników istotnych. Na potrzeby raportu wykorzystane zostały 3 modele z biblioteki sklearn: LinearRegression, RandomForestRegressor, MLPRegressor (sieć neuronowa). Każdy z modeli uczył się na ty samym zestawie danych oraz następnie był testowany na danych z roku 2017.

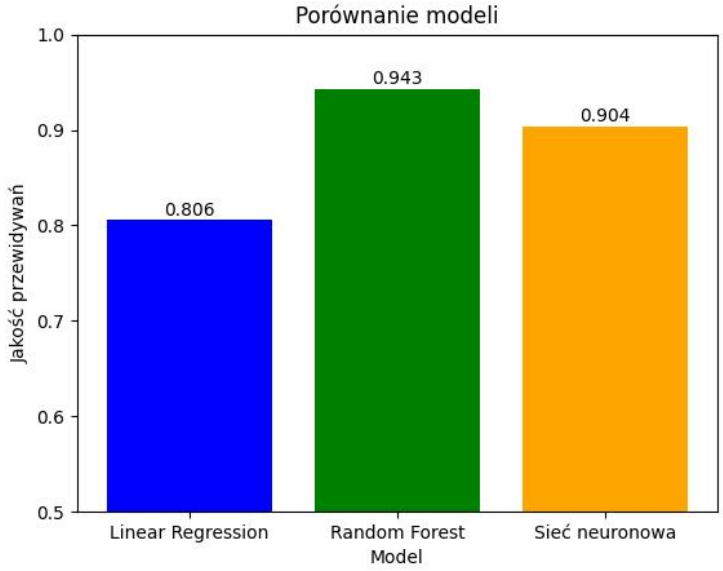
3.1 Ocena jakości modelu

Najlepszym modelem, którego przewidywania są najbardziej poprawne ($R^2 = 0.942$), jest model oparty na Random Forest Regressor. Powodem takiego wyniku mogą być jego właściwości, Random Forest Regressor wykorzystuje las losowy, który łączy wiele drzew decyzyjnych co pozwala na modelowanie wielu skomplikownaych zależności, jest on również wrażliwy na różnorodność danych co w przypadku wielu zmiennych decyzyjnych pomogło osiągnąć wysoki wynik estymacji tego modelu.

Drugim najlepszym modelem okazała się sieć neuronowa ($R^2 = 0.904$), jej niższy wynik można usprawiedliwiać względnie małą ilością iteracji (5000), natomiast dla kilku sprawdzanych ilości iteracji wyniki modelu były zbliżone. Sieci neuronowe mogą również zostać "przeuczone", co oznacza iż model oparty na sieci nuronowej po przejściu przez zbyt dużą ilość iteracji może pogorszyć zamiast poprawić wyniki swojego działania. Zjawisko takie udało mi się zauważyć dla liczby około 100000 iteracji co spowodowało spadek jakości modeli o kilka procent.

Modelem o najgorszej jakości okazał się model oparty na Linear Regression ($R^2 = 0.806$), fakt ten może być spowodowany charakterystyką owego modelu. Model regresji liniowej zakłada, że zależność między zmiennymi niezależnymi a zmienną docelową jest liniowa, co nie zawsze zgadza się z realiami przekazanych danych. Model ten jest również wrażliwy na wartości odstające, co również w pewnym stopniu wpłynęło na jego końcowy wynik.

Wyniki jakości wszystkich modeli zaprezentowane są na Rys. 26



Rysunek 26: Zestawienie wyników poprawności predykcji

4 Wnioski

Czynniki takie jak problemy z używkami, które odnoszą się do nadużywania substancji psychoaktywnych o dziwo przyczyniają się do wydłużenia długości życia. Istnieje również pozytywna korelacja między poziomem szczęścia a długością życia, co może wynikać z pozytywnego wpływu dobrego samopoczucia i zadowolenia na zdrowie ogólne.

Wyższy poziom PKB per capita jest związany z lepszymi warunkami życia, lepszym dostępem do opieki zdrowotnej i lepszymi warunkami środowiskowymi, co przekłada się na dłuższą długość życia. Silna korelacja między GDP a długością życia może wynikać z większej dostępności usług medycznych i lepszej jakości życia w krajach o wyższym PKB.

Czynniki takie jak zanieczyszczenie powietrza mają negatywny wpływ na zdrowie i mogą prowadzić do różnych chorób, takich jak problemy z oddychaniem, choroby sercowo-naczyniowe i nowotwory. Dlatego kraje z niższym poziomem zanieczyszczenia powietrza mogą mieć wyższą długość życia.

Mediana wieku jest pozytywnie skorelowana z długością życia, co sugeruje, że im wyższa mediana wieku w danej populacji, tym dłużej ludzie żyją. Wysoka gęstość zaludnienia może wiązać się z lepszymi warunkami życia, lepszym dostępem do zasobów i usług, co wpływa na zdrowie i długość życia.

4.1 Ulepszenia

W celu uzyskania lepszej jakości modeli można by było:

- zgromadzić większą ilość czynników istotnych,
- zgromadzić większą ilość danych.,
- zapewnić większą różnorodność danych treningowych.

Bibliografia

[1] Saloni Dattani, Hannah Ritchie i Max Roser. "Mental Health". W: *Our World in Data* (2021). URL: https://ourworldindata.org/mental-health.

[2] Saloni Dattani i in. "Suicides". W: *Our World in Data* (2023). URL: https://ourworldindata.org/suicide.

[3] Esteban Ortiz-Ospina i Max Roser. "Happiness and Life Satisfaction". W: *Our World in Data* (2013). URL: https://ourworldindata.org/happiness-and-life-satisfaction.

[4] Hannah Ritchie i Max Roser. "Age Structure". W: *Our World in Data* (2019). URL: https://ourworldindata.org/age-structure.

[5] Hannah Ritchie i Max Roser. "Air Pollution". W: *Our World in Data* (2017). URL: https://ourworldindata.org/air-pollution.

[6] Hannah Ritchie i Max Roser. "Drug Use". W: *Our World in Data* (2019). URL: https://ourworldindata.org/drug-use.

[7] Hannah Ritchie i Max Roser. "Land Use". W: *Our World in Data* (2013). URL: https://ourworldindata.org/land-use.

[8] Hannah Ritchie i in. "Internet". W: *Our World in Data* (2023). URL: https://ourworldindata.org/internet.

[9] Max Roser. "Economic Growth". W: *Our World in Data* (2013). URL: https://ourworldindata.org/economic-growth.

[10] Max Roser i in. "World Population Growth". W: *Our World in Data* (2013). URL: https://ourworldindata.org/world-population-growth.