Przewidywanie liczby ludności w Polsce za pomocą modeli regresji liniowej na podstawie zmiennych demograficznych

Krzysztof Kulka
272667@student.pwr.edu.pl
MSiD Lab Wtorek 9.15 NP

May 16, 2024

Spis treści

1 Wstęp

Problemem projektu jest analiza możliwości modelów regresji liniowej do przewidywania liczby ludności w Polsce. W tym celu wykorzystane zostaną dane historyczne dotyczące demografi, oraz innych czynników wpływających na liczebność populacji. Przedstawiona analiza ma na celu rozstrzygnięcie czy model regresji liniowej jest odpowiedni do przewidywania liczby ludności w Polsce, oraz jakie modele sprawdzają się do tego najlepiej. Analizie zostaną poddane następujące czynniki:

- Historyczna liczba ludności
- Imigracja do kraju
- Wskaźnik dzietności
- Oczekiwana długość życia
- Urbanizacja
- Wskaźnik zmiany populacji na przestrzeni ostatnich 5 lat

Zbadane zaś zostaną następujące modele regresji:

- Regresja liniowa
- Regresja typu Ridge
- Regresja Lasso
- Regresja Elastic Net
- Regresja Bayesian Ridge

2 Zbiór danych i jego analiza

2.1 Opis zbioru danych

Zbiór danych zawiera informacje na temat historycznej liczby ludności, imigracji do kraju, wskazniku dzietnosci, oczekiwanej długości życia w momencie urodzenia na przestrzeni lat 1960-2023. Dane zostały pobrane z serwisu internetowego World Bank[4] Dane dotyczą około 260 krajów. Dodatkowo informacje na temat urbanizacji zostały pobrane z serwisu internetowego Zintegrowana Platforma Edukacyjna Ministerstwa Edukacji Narodowej[5], a wskaźnik zmiany populacji na przestrzeni ostatnich 5 lat został obliczony na podstawie danych historycznych.

2.2 Przykładowe dane

"Country Name"	"Country Code"	"Indicator Name"	"Indicator Code"	"1960"	"1961"	...
"Aruba"	"ABW"	"Fertility rate, total (births per woman)"	"SP.DYN.TFRT.IN"	"4.82"	"4.655"	...
"Africa Eastern and Southern"	"AFE"	"Fertility rate, total (births per woman)"	"SP.DYN.TFRT.IN"	"6.72412501084242"	"6.74275210020318"	...
"Afghanistan"	"AFG"	"Fertility rate, total (births per woman)"	"SP.DYN.TFRT.IN"	"7.282"	"7.284"	...
"Africa Western and Central"	"AFW"	"Fertility rate, total (births per woman)"	"SP.DYN.TFRT.IN"	"6.45844789624312"	"6.47151755185967"	...
...						

Table 1: Wycinek danych z zbioru dzietności na kobietę

"Country Name"	"Country Code"	"Indicator Name"	"Indicator Code"	"1960"	"1961"	...
"Aruba"	"ABW"	"Population, total"	"SP.POP.TOTL"	"54608"	"55811"	...
"Africa Eastern and Southern"	"AFE"	"Population, total"	"SP.POP.TOTL"	"130692579"	"134169237"	...
"Afghanistan"	"AFG"	"Population, total"	"SP.POP.TOTL"	"8622466"	"8790140"	...
"Africa Western and Central"	"AFW"	"Population, total"	"SP.POP.TOTL"	"97256290"	"99314028"	...
...						

Table 2: Wycinek danych z zbioru liczby ludności

2.3 Obróbka danych

2.3.1 Pozyskanie danych

Najważniejszym krokiem w obróbce danych było wyizolowanie danych dotyczących Polski, oraz usunięcie kolumn, które nie były istotne dla analizy takicj jak kod kraju, nazwa wskaźnika i tym podobne. Dodatkowo, z racji tego że dane dotyczące imigracji rejestrowane co pięć lat, skorzystano z interpolacji liniowej, aby uzupełnić brakujące dane. Największe wyzwanie pojawiło się z danymi dotyczącymi urbanizacji. Jedyne z zaufanego oficjalnego źródła były w formie obrazka .jpg. Aby więc uzyskać dane, i móc zamienić je w data frame, użyto następujących kroków:

- Scraping obrazka za pomocą skryptu pythonowego
- Następnie przekazanie zescrapowanego obrazka do zewnętrznego programu WebPlotDigitizer[3]
- Dalej, z racji niedoskonałości otrzymanego wyniku (m.in. potraktowanie lat jako liczb rzeczywistych, a nie całkowitych), dane zostały poprawione za pomocą kolejnego skryptu napisanego w pythonie.
- Finalnie, ponieważ dane sięgały jedynie 2012 roku, dodano za pomocą skryptu dane z lat 2013-2022 korzystając ze witryny Gegrafia24.pl[1].

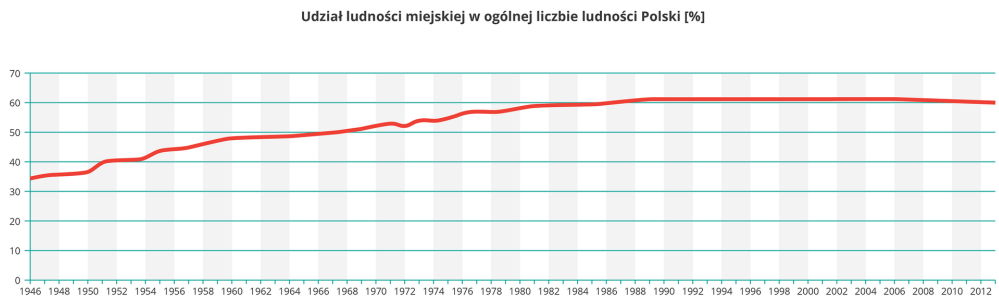

Figure 1: Zescrapowany graf urbanizacji w Polsce

2.4 Dane po obróbce

Po wstępnej obróbce danych, tak prezentują się zmienne demograficzne po ich normalizacji:

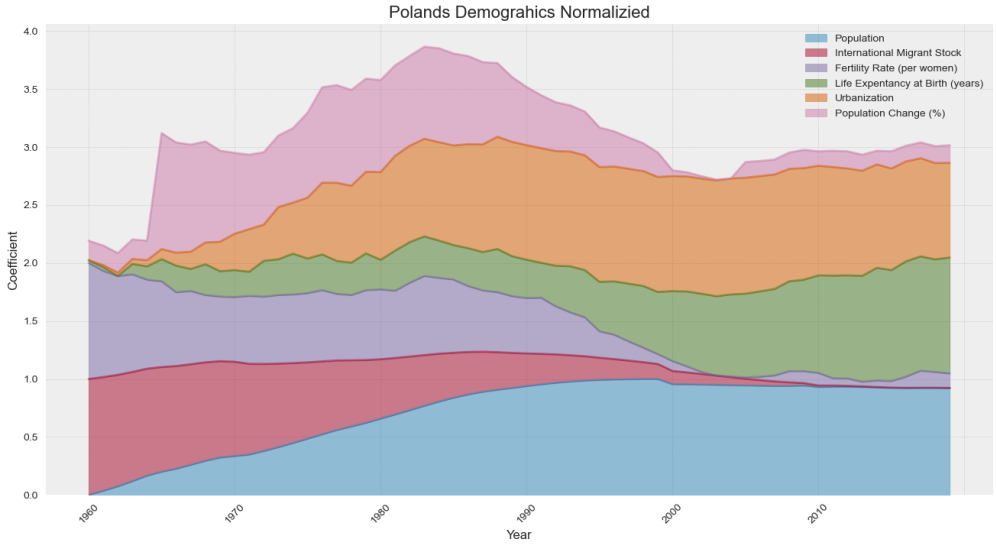

Figure 2: Dane po obróbce

2.5 Analiza danych

Do analizy eksploracyjnej danych wykorzystano bibliotekę pandas-profiling[6]

2.6 Populacja w Polsce

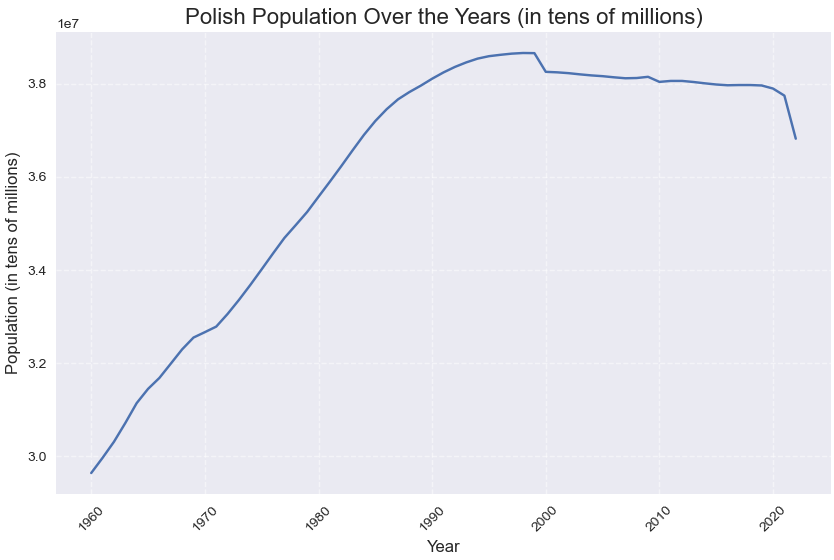

Figure 3: Wizualizacja liczby ludności w Polsce na przestrzeni lat

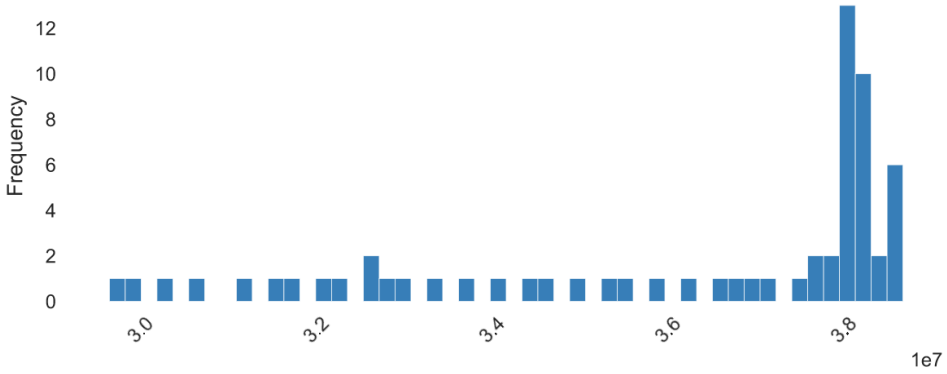

Figure 4: Histogram liczby ludności w Polsce

Minimum	Maximum	Mediana
29637450	38663481	37899070

Table 3: Statystyki liczby ludności w Polsce

2.7 Imigracja do Polski

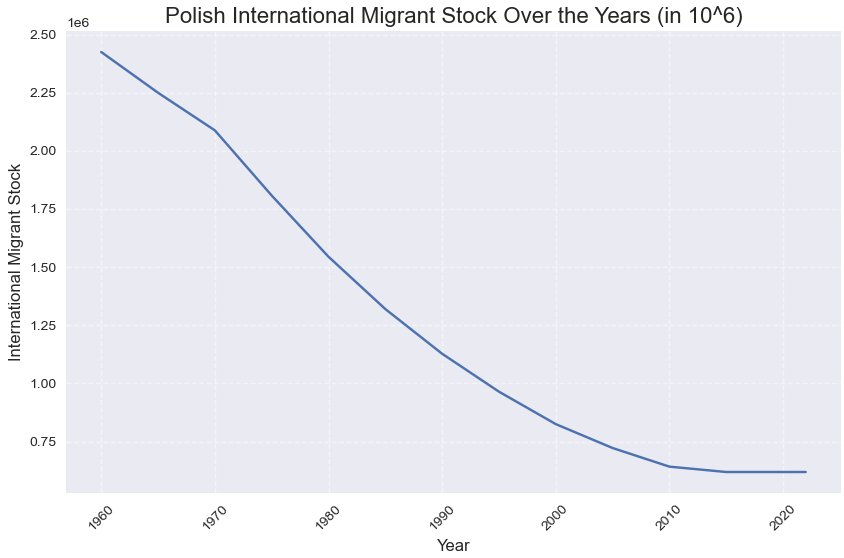

Figure 5: Wizualizacja imigracji do Polski na przestrzeni lat

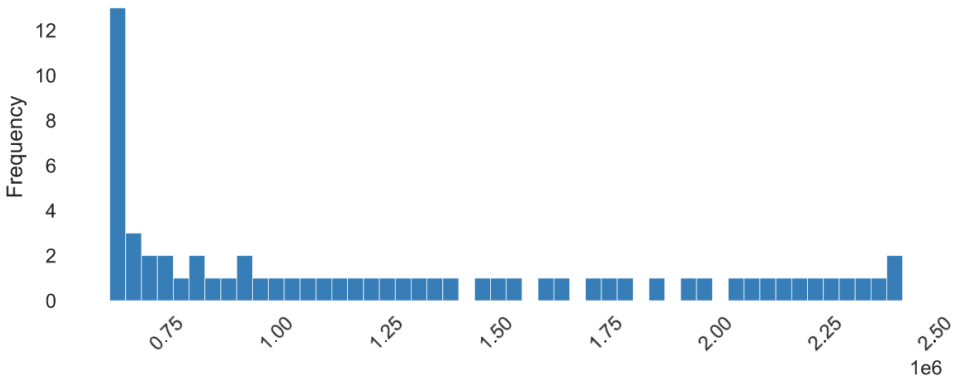

Figure 6: Histogram imigracji do Polski na przestrzeni lat

Minimum	Maximum	Mediana
619403	2424881	1095161

Table 4: Statystyki imigracji do Polski

2.8 Współczynnik dzietności

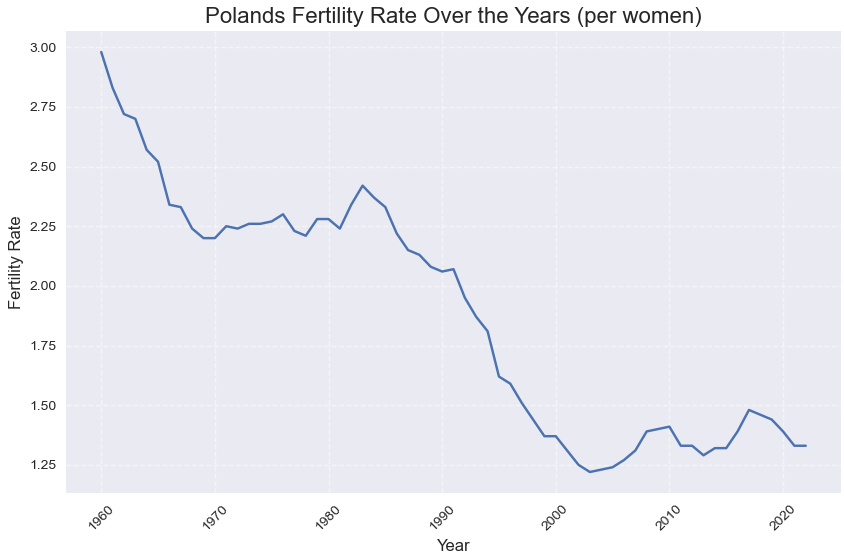

Figure 7: Wizualizacja współczynnika dzietności Polski na przestrzeni lat

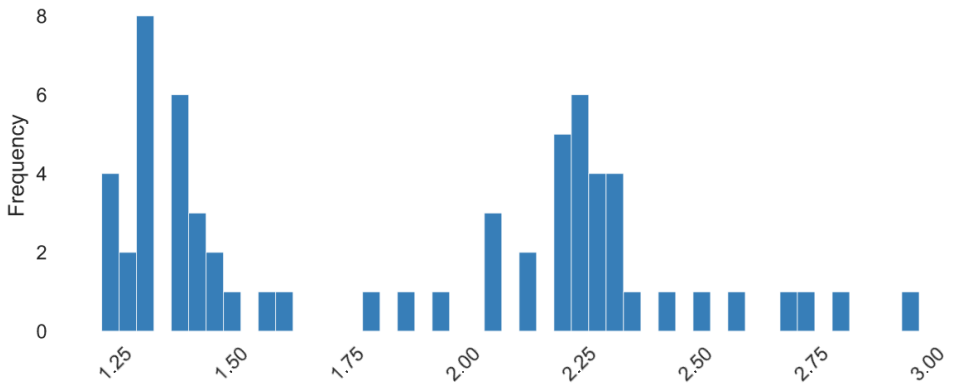

Figure 8: Histogram współczynnika dzietności na przestrzeni lat

Minimum	Maximum	Mediana
1.22	2.98	2.06

Table 5: Statystyki współczynnika dzietności

2.9 Oczekiwana długość życia

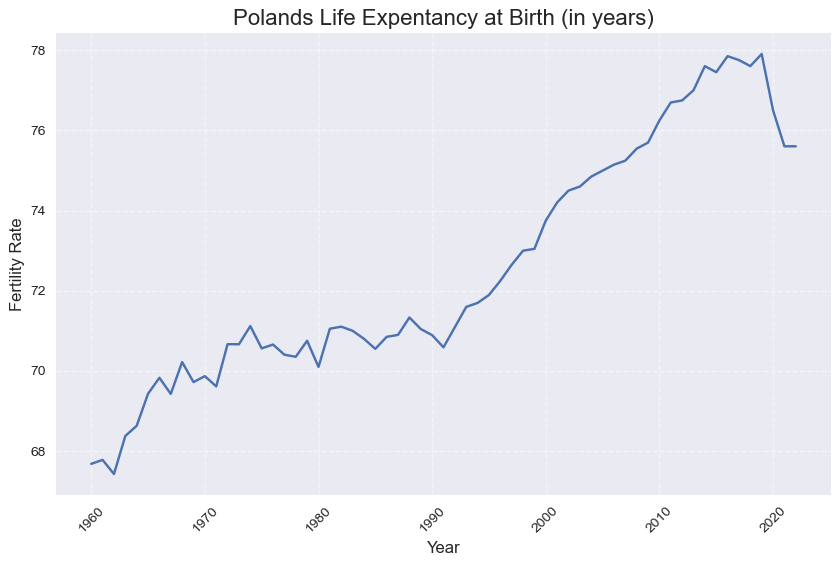

Figure 9: Wizualizacja oczekiwanej długości życia na przestrzeni lat

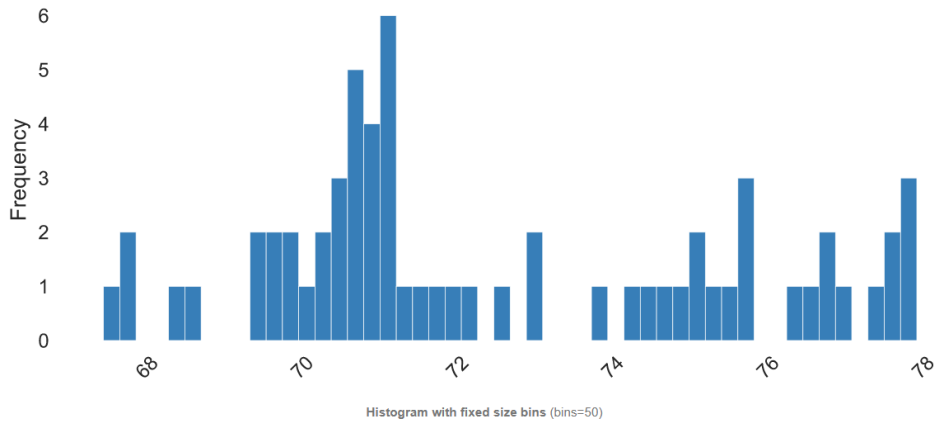

Figure 10: Histogram oczekiwanej długości życia na przestrzeni lat

Minimum	Maximum	Mediana
67.42	77.9	71.1

Table 6: Statystyki oczekiwanej długości życia

2.10 Urbanizacja

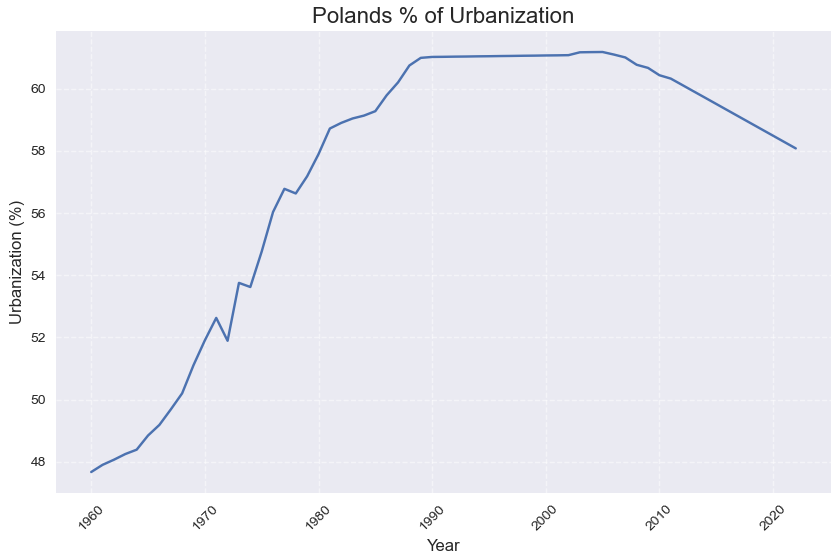

Figure 11: Wizualizacja urbanizacji na przestrzeni lat

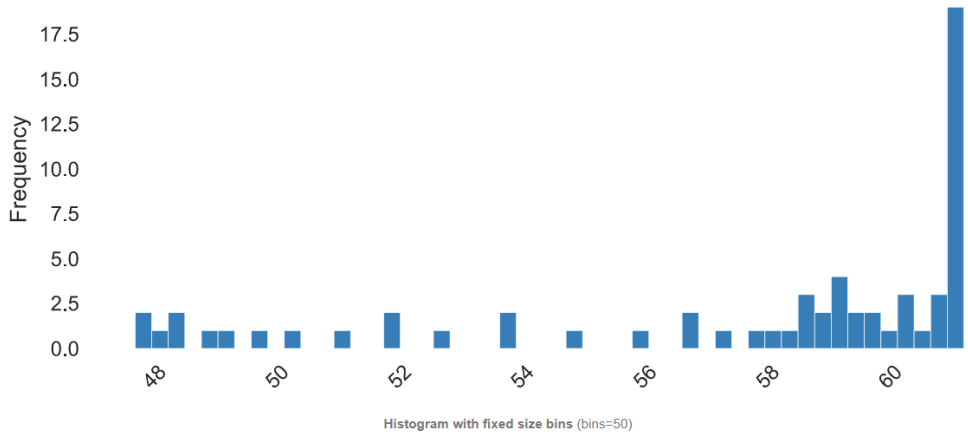

Figure 12: Wizualizacja urbanizacji na przestrzeni lat

Minimum	Maximum	Mediana
47.66	61.18	59.27

Table 7: Statystyki urbanizacji

2.11 Wskaźnik zmiany populacji

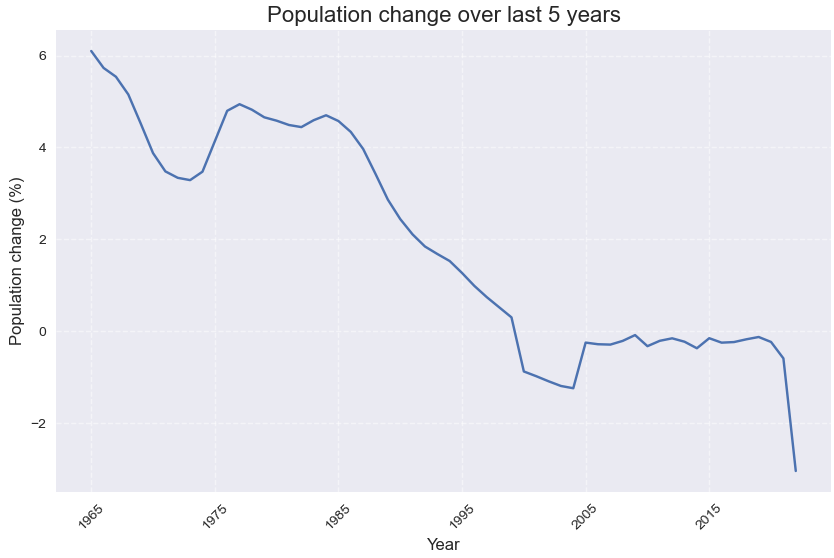

Figure 13: Wizualizacja wskaźnika zmiany populacji na przestrzeni lat

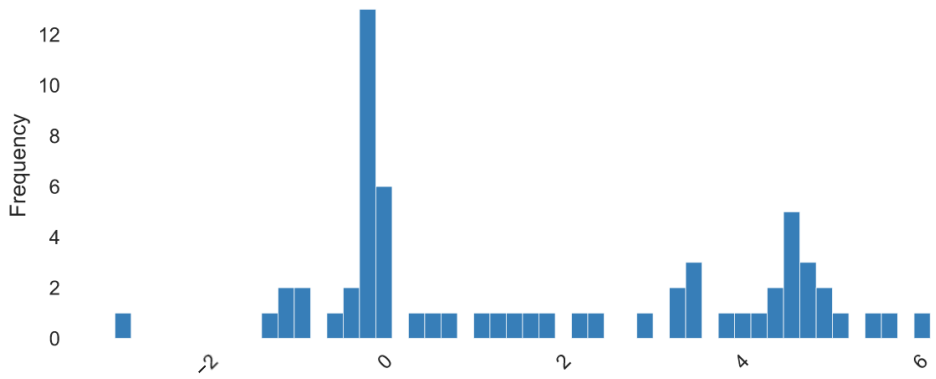

Figure 14: Histogram zmiany polskiej populacji na przestrzeni ostatnich 5 lat

Minimum	Maximum	Mediana
-3.03	6.09	0.98

Table 8: Statystyki wskaźnika zmiany populacji

2.12 Korelacja

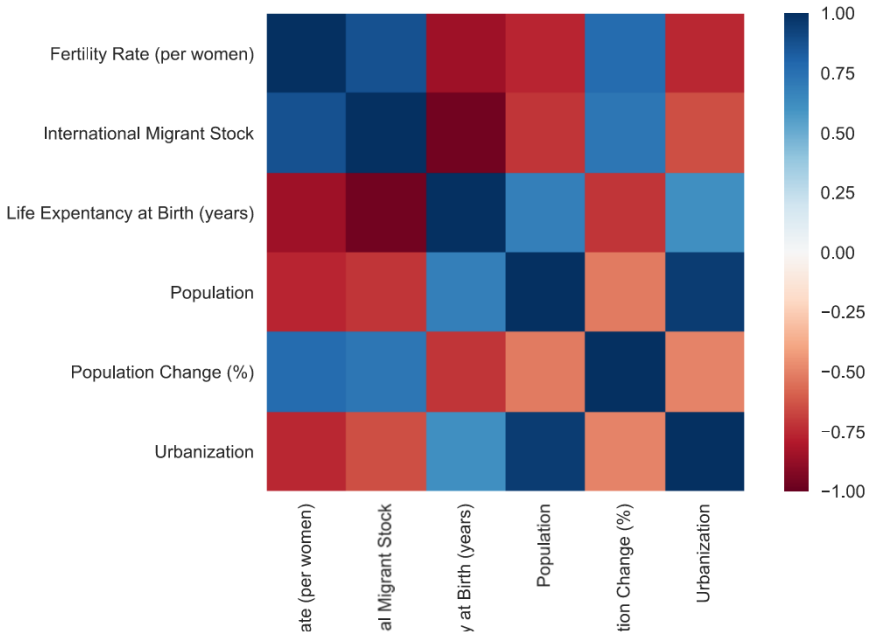

Figure 15: Macierz korelacji

Z racji ręcznego dobioru danych i ich selekcji, analiza wykazała bardzo dużą korelację (oraz antykorelacje) pomiędzy wybranymi współczynnikami. Pojawia się bardzo zaskakująca, przecząca logice korelacja, pomiędzy współczynnikiem dzietności a populacją wynosząca -0.763. Prawdopodobnie wynika ona z tego że przez większość badanego okresu, współczynnik był nadal na bardzo wysokim

poziomie, więc mimo że malał, to populacja stale się zwiększała. Analogicznie zaskakuje negatywna korelacja migracji z populacją, co także dziwi, ponieważ zgodnie z intuicją, imigracja powinna zwiększać populację. Prawdopodobnie wynika to z faktu, że dane dotyczące imigracji są niewielkie, co powoduje że słabo, choć wciąż, przekładają się na polską populację. Jeśli zaś chodzi o spodziewane korelacje, należy szczególnie zwrócić uwagę:

- Oczekiwana długość życia z populacją: 0.683
- Urbanizacja z populacją: 0.949
- Urbanizacja z oczekiwaną długością życia: 0.611

Korelacje te dobrze wróżą dla modeli regresji, ponieważ są one na tyle silne, że powinny pozwolić na skuteczne przewidywanie liczby ludności w Polsce.

2.13 Eliminacja danych odstających

Podczas przeglądania grafów reprezentujących zebrane dane, nie trudno było zauważyć że dane po 2019 roku są znacznie odstające od reszty. Oczywistym jest że w 2020 roku, z racji pandemii, wiele wskaźników uległo zmianie, co sprawia że dane z tego roku są nieprzydatne do analizy. Z tego powodu, dane z 2020 roku wzwyż zostały usunięte.

3 Dobór metryk oceny

Zanim przystąpimy do analizy modeli, należy zdefiniować metryki, które pozwolą nam ocenić ich skuteczność. Dobór odpowiednich metryk jest kluczowy, ponieważ pozwala na obiektywną ocenę modeli, oraz porównanie ich ze sobą. W przypadku regresji, najczęściej stosowanymi metrykami są:

- Mean Squared Error (MSE) - średni błąd kwadratowy
- Mean Absolute Percentage Error (MAPE) - średni błąd procentowy bezwzględny
- R2 Score - współczynnik determinacji

3.1 Mean Squared Error

MSE jest jedną z najczęściej stosowanych metryk w regresji. Oblicza ona średni błąd kwadratowy pomiędzy wartościami przewidywanymi przez model, a wartościami rzeczywistymi.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{1}$$

3.2 Mean Absolute Percentage Error

MAPE jest metryką, która mierzy średni błąd procentowy bezwzględny pomiędzy wartościami przewidywanymi przez model, a wartościami rzeczywistymi.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y_i}|}{y_i} \tag{2}$$

3.3 R2 Score

R2 Score jest metryką, która mierzy jak dobrze model przewiduje dane w porównaniu do średniej wartości. Wartość R2 Score może przyjmować wartości od $-\infty$ do 1, gdzie 1 oznacza idealne dopasowanie modelu.

$$R2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

3.4 Analiza metryk

Metryki zostały wybrane tak, aby pokrywały one różne aspekty oceny modeli. MSE pozwala na ocenę jak dobrze model przewiduje wartości, MAPE pozwala na ocenę jak dobrze model przewiduje wartości w procentach, a R2 Score pozwala na ocenę jak dobrze model przewiduje wartości w porównaniu do średniej wartości. Nie ulega jednak wątpliwości że najlepszym wyborem jest MAPE, który najlepiej nadaje się do predykcji dotyczących populacji[7], a więc jest najbardziej adekwatny do naszego problemu.

4 Analiza modeli

4.1 Dla podziału 80/20

Podział 80/20 oznacza to że model uczyć będzie się na danych z lat 1960-2008, a testowany będzie na danych z lat 2009-2019.

4.1.1 Regresja liniowa

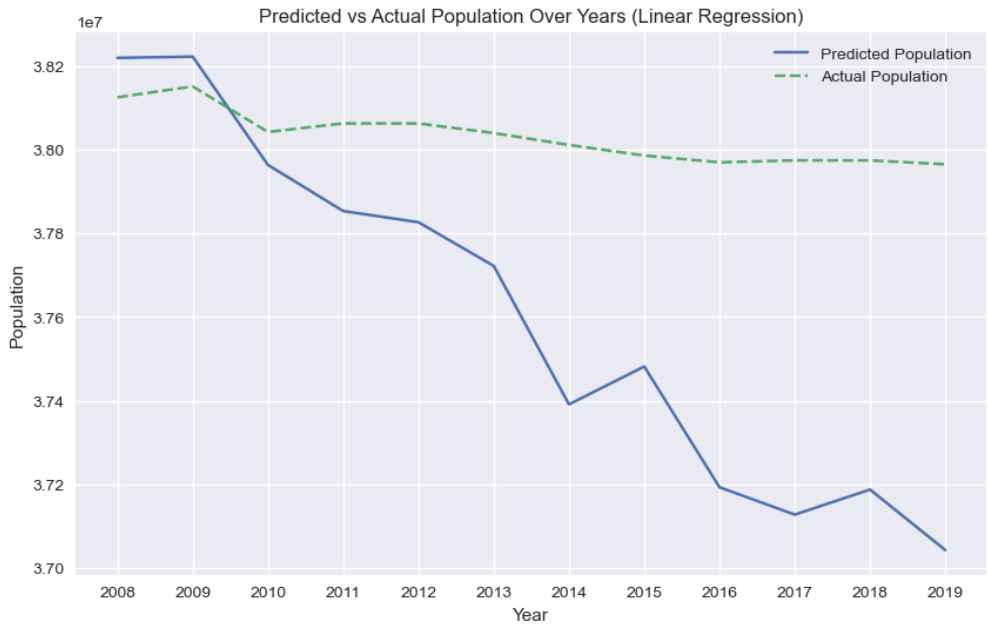

Figure 16: Regresja liniowa

Model	MSE	MAPE	R2 Score
Linear Regression	3.04e+11	0.012	-84.972

Table 9: Metryki dla regresji liniowej i podziału 80/20

4.1.2 Regresja Ridge

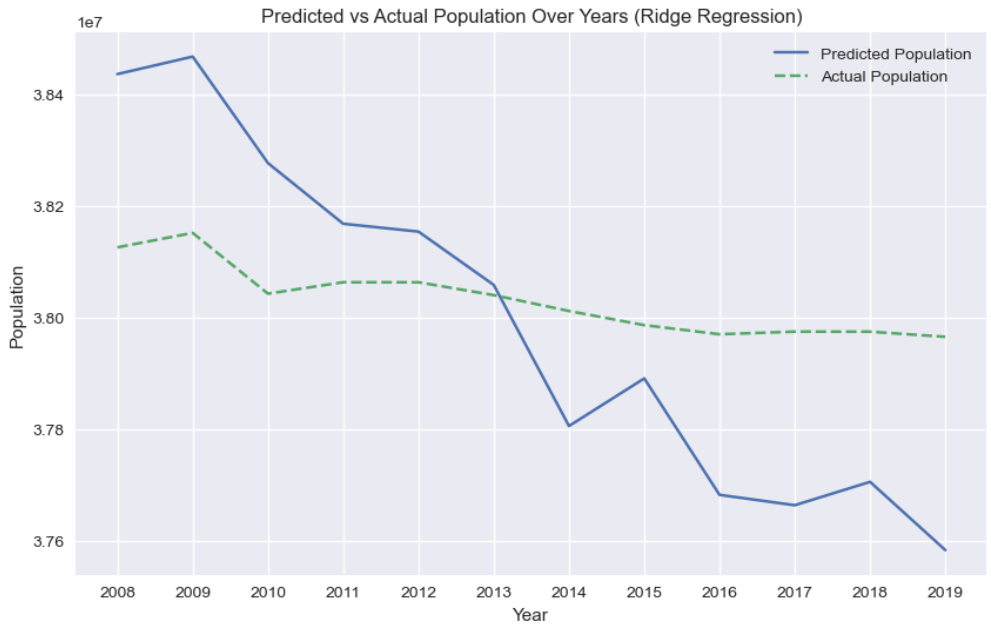

Figure 17: Regresja Ridge

Model	MSE	MAPE	R2 Score
Lasso Regression	6.00e+10	0.006	-15.926

Table 10: Metryki dla regresji Ridge i podziału 80/20

4.1.3 Regresja Lasso

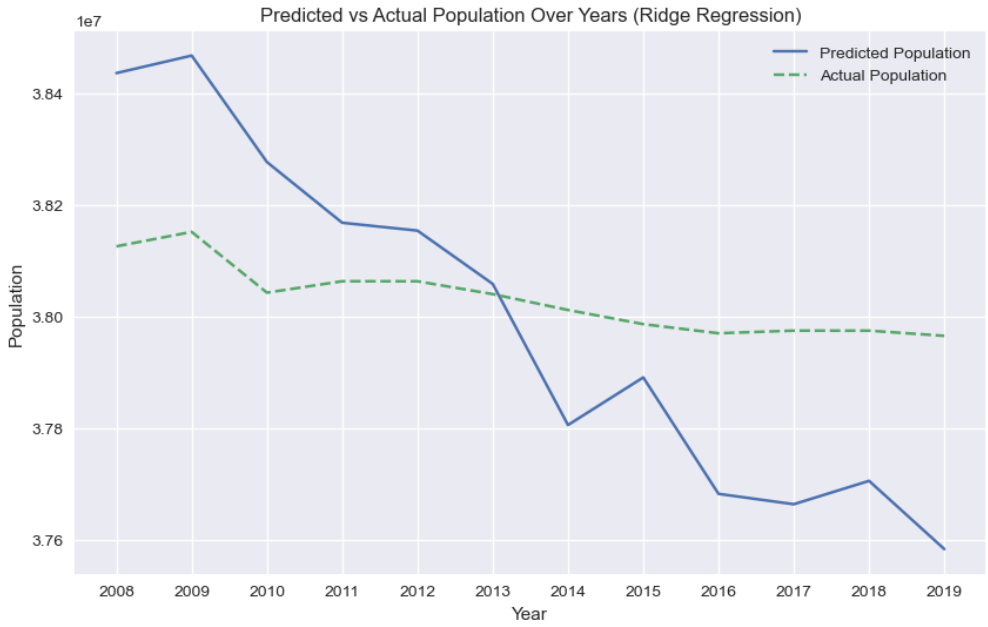

Figure 18: Regresja Lasso

Model	MSE	MAPE	R2 Score
Lasso Regression	3.04e+11	0.011	-84.952

Table 11: Metryki dla regresji Lasso i podziału 80/20

4.1.4 Regresja Elastic Net

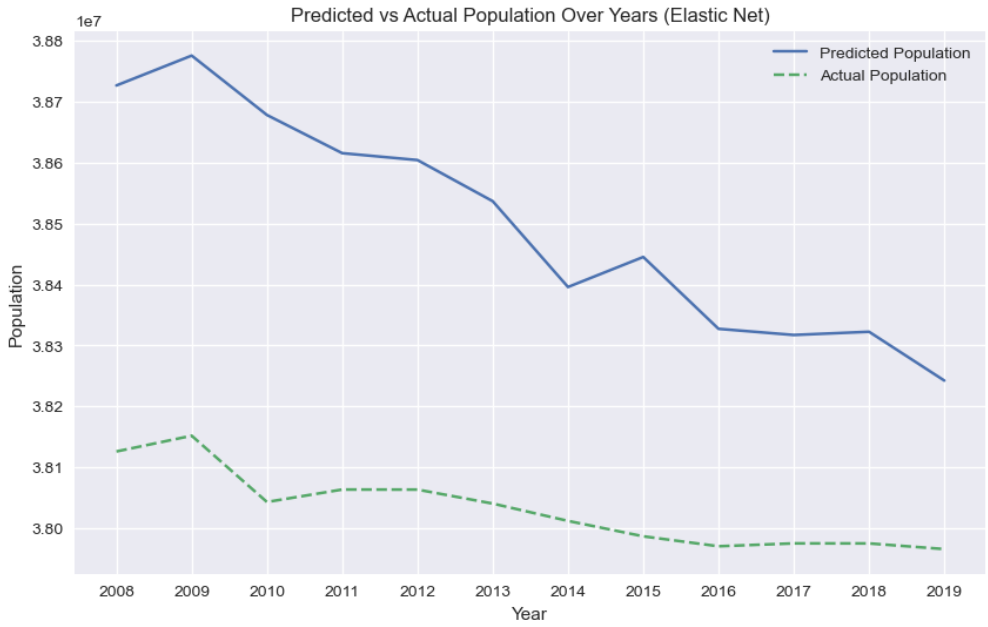

Figure 19: Regresja Elastic Net

Model	MSE	MAPE	R2 Score
Elastic Net	2.33e+11	0.012	-64.851

Table 12: Metryki dla regresji Elastic Net i podziału 80/20

4.1.5 Regresja Bayesian Ridge

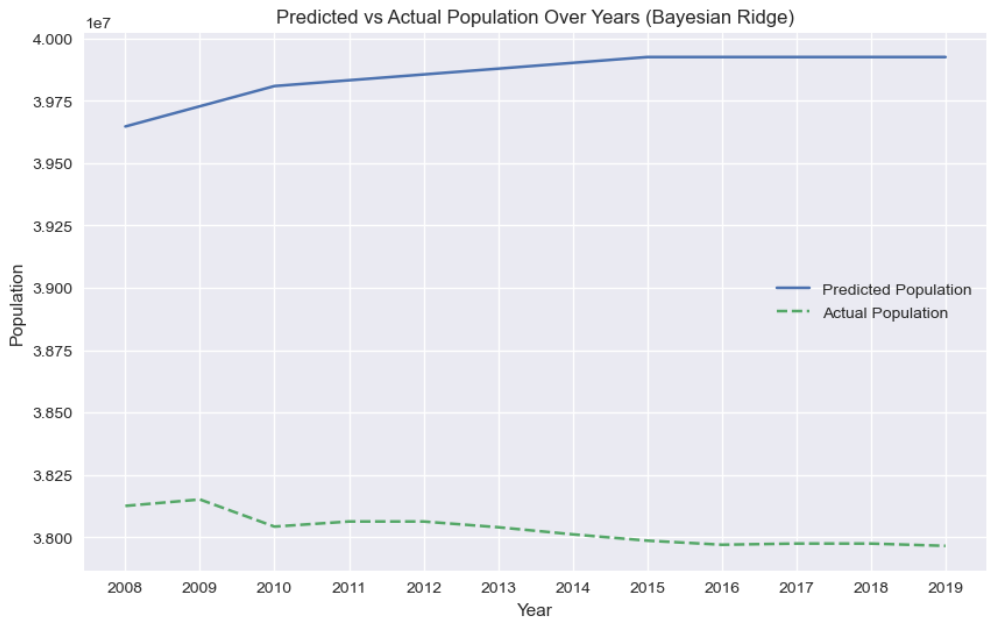

Figure 20: Regresja Bayesian Ridge

Model	MSE	MAPE	R2 Score
Bayesian Ridge Regression	3.35e+12	0.048	-946.190

Table 13: Metryki dla regresji Bayesian Ridge i podziału 80/20

4.1.6 Podsumowanie dla podziału 80/20

Model	MSE	MAPE	R2 Score
Linear Regression	3.04e+11	0.012	-84.972
Ridge Regression	6.00e+10	0.006	-15.926
Lasso Regression	3.04e+11	0.011	-84.952
Elastic Net	2.33e+11	0.012	-64.851
Bayesian Ridge	3.35e+12	0.048	-946.190

Table 14: Wyniki dla podziału 80/20

Metryka MSE jest bardzo wysoka, co jest spowodowane tym że dane do przewidzenia są bardzo wysokie (rzędzu 10^7), co sprawia że błędy są również bardzo wysokie. Dodatkowo, model otrzymał niewielkie ilości danych szkoleniowych, co sprawia że jest on niewystarczająco dopasowany do danych testowych. MAPE, która jest główną metryką, daje jednak nadzieje na to że model może być użyteczny, ponieważ wartość błędu procentowego dla Ridge wynosi zaledwie 0.006.

- Najlepszym modelem okazała się regresja Ridge, która osiągnęła najniższe wartości MSE, MAPE oraz najwyższe R2 Score.

- Najgorszym modelem okazała się regresja Bayesian Ridge, która osiągnęła najwyższe wartości MSE, MAPE oraz niespotykanie niski R2 Score.
- Regresja Lasso oraz regresja liniowa osiągnęły bardzo zbliżone wyniki, co sprawia że nie można jednoznacznie stwierdzić która z nich jest lepsza.
- Regresja Elastic Net osiągnęła wyniki zbliżone, choć gorsze, od regresji Ridge, ale lepsze R2 i MSE niż Lasso i Liniowa, co sprawia że również jest ona dobrym modelem.

4.2 Dla podziału 60/40

Podział 60/40 oznacza to że model uczyć będzie się na danych z lat 1960-1995, a testowany będzie na danych z lat 1996-2019.

4.2.1 Regresja liniowa

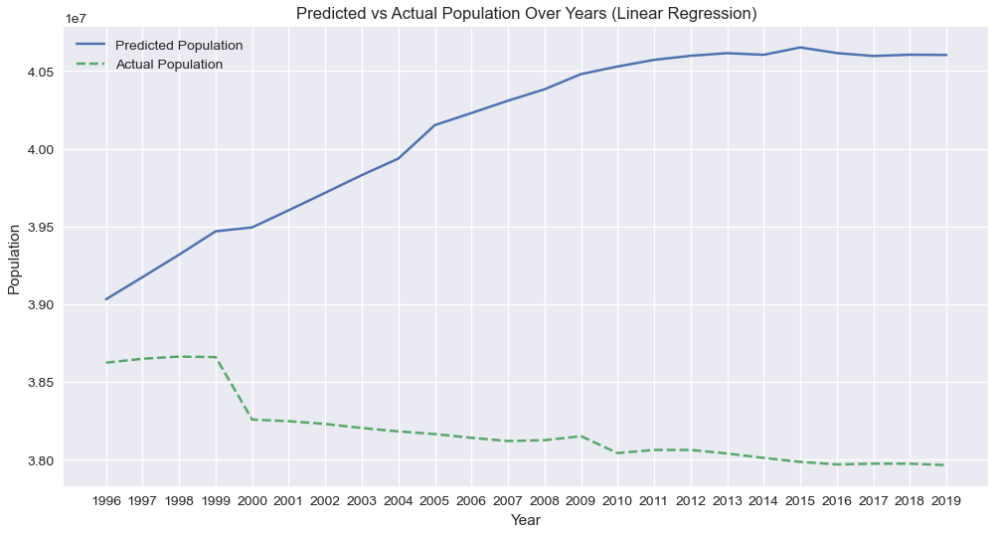

Figure 21: Regresja liniowa

Model	MSE	MAPE	R2 Score
Linear Regression	4.31e+12	0.051	-84.631

Table 15: Metryki dla regresji liniowej i podziału 60/40

4.2.2 Regresja Ridge

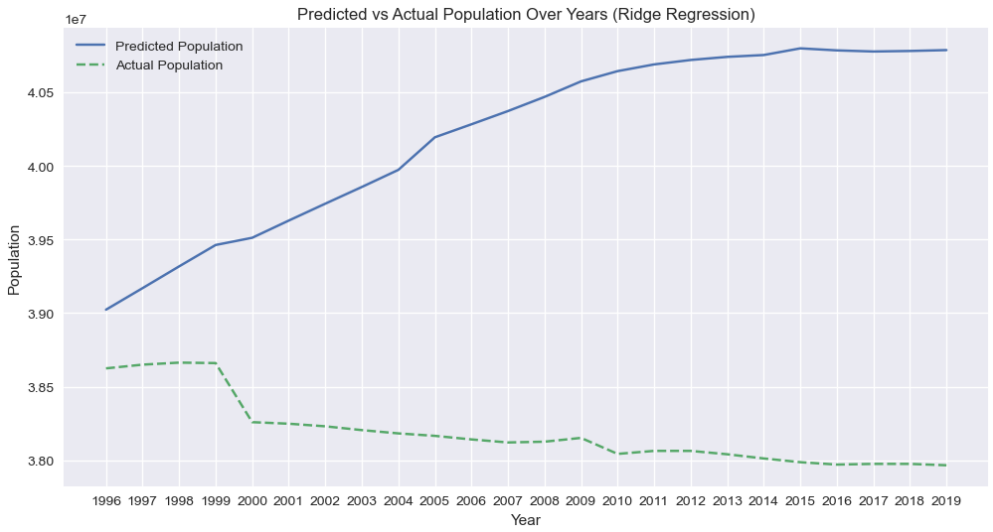

Figure 22: Regresja Ridge

Model	MSE	MAPE	R2 Score
Ridge Regression	4.73e+12	0.053	-92.887

Table 16: Metryki dla regresji Ridge i podziału 60/40

4.2.3 Regresja Lasso

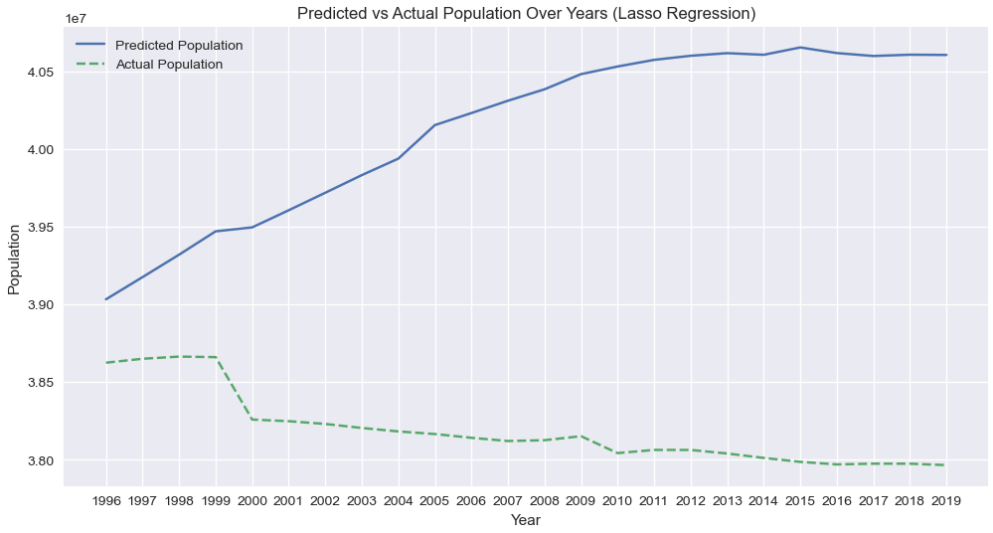

Figure 23: Regresja Lasso

Model	MSE	MAPE	R2 Score
Lasso Regression	4.31e+12	0.050	-84.638

Table 17: Metryki dla regresji Lasso i podziału 60/40

4.2.4 Regresja Elastic Net

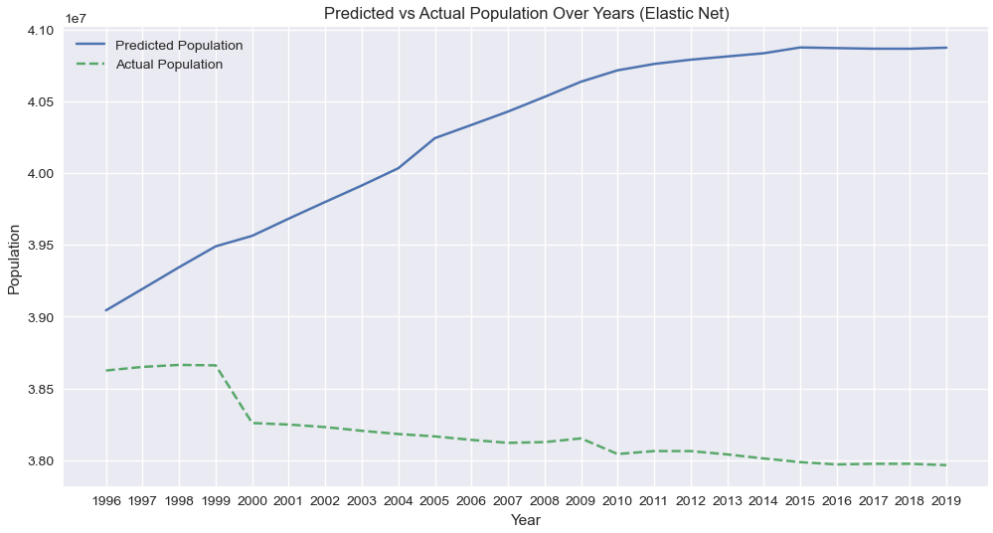

Figure 24: Regresja Elastic Net

Model	MSE	MAPE	R2 Score
Elastic Net	4.99e+12	0.054	-98.200

Table 18: Metryki dla regresji Elastic Net i podziału 60/40

4.2.5 Regresja Bayesian Ridge

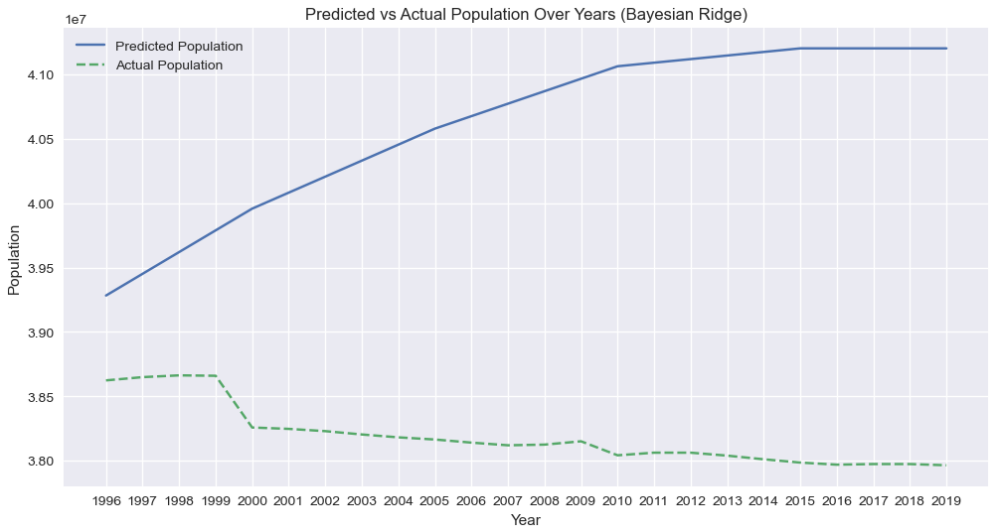

Figure 25: Regresja Bayesian Ridge

Model	MSE	MAPE	R2 Score
Bayesian Ridge	6.55e+12	0.063	-129.007

Table 19: Metryki dla regresji Bayesian Ridge i podziału 60/40

4.2.6 Podsumowanie dla podziału 60/40

Model	MSE	MAPE	R2 Score
Linear Regression	4.31e+12	0.051	-84.631
Ridge Regression	4.73e+12	0.053	-92.887
Lasso Regression	4.31e+12	0.050	-84.638
Elastic Net	4.99e+12	0.054	-98.200
Bayesian Ridge	6.55e+12	0.063	-129.007

Table 20: Wyniki dla podziału 60/40

Jak widać, wyniki dla podziału 60/40 są znacznie gorsze niż dla podziału 80/20. Wynika to z faktu że model otrzymał mniej danych szkoleniowych, co sprawia że jest on niewystarczająco dopasowany do danych testowych. Split 80/20 okazał się zdecydowanie lepszy, co sprawia że jest on bardziej odpowiedni do analizy, i zapewnia dobry bilans między szkoleniem i testowaniem modelu.

5 Wnioski

Po przeanalizowaniu wielu różnych modeli regresji, na róznych podziałach danych, można stwierdzić że regresja Ridge jest najlepszym modelem do przewidywania liczby ludności w Polsce. Model ten osiągnął najniższe wartości MSE, MAPE oraz najwyższe R2 Score, co sprawia że jest on najbardziej

adekwatny do naszego problemu. Warto szczególnie zwrócić uwagę na wartośćć MAPE, która wynosi zaledwie 0.006, co oznacza że model przewiduje wartości zaledwie o 0.6% odchylające się od wartości rzeczywistych. Kategoryzuje to ten model w kategorię bardzo dobrych[2], co sprawia że jest on najbardziej odpowiedni do przewidywania liczby ludności w Polsce.

Dodatkowo, projekt ten udowania że zmienne demograficzne, takie jak dzietność, oczekiwana długość życia, urbanizacja, oraz imigracja, mają duży wpływ na liczebność populacji, do tego stopnia, że na ich podstawie można z dużą dokładnością przewidzieć liczebność populacji w przyszłości. Oczywiście, jak pokazał przykład pandemii COVID-19, zdarzenia losowe mogą znacząco wpłynąć na liczebność populacji, co sprawia że model ten nie jest idealny, ale nadal jest on bardzo dobrym narzędziem do przewidywania liczby ludności w Polsce, i może być użyty do analizy trendów demograficznych w przyszłości, zwłaszcza w sytuacjach stablilnych, gdzie nie występują znaczące zmiany w populacji.

Ciekawą i wartą uwagi obserwacją, bez wątpienia jest fakt, że najlepsze modele regresji liniowej, przewidywały populację niższą niż rzeczywista. W istocie może być to wytłumaczone faktem, że dane imigracyjne mogły być zaniżone, co przełożyło się na fakt że modele przewidywały zaniżoną populację.

References

[1] Geografia24, Urbanizacja w Polsce i w Europie. https://geografia24.pl/urbanizacja-w-polsce-i-w-europie/.

[2] Research Gate, Interpretation of typical MAPE values. https://www.researchgate.net/figure/nterpretation-of-typical-MAPE-values_tbl1_257812432.

[3] WebPlotDigitizer. https://automeris.io/.

[4] World Bank Data poland. https://www.worldbank.org/pl/country/poland.

[5] Zintegrowana Platforma Edukacyjna Ministerstwa Edukacji Narodowej urbanizacja w polsce. https://zpe.gov.pl/a/zroznicowanie-poziomu-urbanizacji-w-polsce/D19MUchJD.

[6] Fabian Clemente. YData-Profiling. https://github.com/ydataai/ydata-profiling.

[7] Aybeniz S. Aliyev Makrufa Sh. Hajirahimov. Development of a prediction model on demographic indicators based on machine learning methods: Azerbaijan example.