Raport

Illya Nikonenko 272673

May 19, 2024

1 Wprowadzenie

Ten raport opisuje mój projekt dotyczący jakości powietrza i działania podjęte przeze mnie w celu jego wykonania. Projekt ten pokazuje wpływ różnych czynników na wartość zanieczyszczenia powietrza w różnych krajach świata, stopień ich wpływu, a także przewiduje zanieczyszczenie dla dowolnych danych. Całość projektu znajduje się w moim repozytorium na GitHub.

2 Dane

2.1 Znajdowanie i scrapowanie danych

Jako źródeł danych użyłem Wikipedia oraz Kaggle. Scrapowanie danych z tych stron bardzo się różniło:
- Kaggle:
W przypadku Kaggle wystarczyło utworzyć konto na tej stronie, by mieć dostęp do API, oraz pobrać bibliotekę kaggle w Python, która zawiera wszystkie niezbędne komendy do ściągnięcia baz danych w formacie .csv.
- Wikipedia:
Za pomocą kodu trzeba było przeszukać stronę by znaleźć potrzebną tabelę z danymi, a następnie wyciągać z niej dane komórka po komórce i samodzielnie zapisywać te dane do pliku .csv.
Cały proces zbierania danych jest zawarty w pliku upload_and_preprocess_data.py, a dane - w plikach .csv.

2.2 Preprocessing

Preprocessing danych w moim przypadku polegał na uzupełnieniu brakujących danych w kolumnach liczbowych. Uzupełniałem je przy użyciu wartości średniej ze wszystkich danych w kolumnie. Nie jest to najbardziej dokładna metoda, jednak tutaj była wystarczająco dobra i niezbędna, aby uniknąć komplikacji i błędów w późniejszej pracy z danymi. Proces ten znajduje się w preprocess_data.py.

2.3 Uzupełnienie i złączenie

Następnym etapem pracy z danymi było ich uzupełnienie. Używałem danych z lat 1990-2017, i nie dla każdego roku wszystkie dane były dostępne. Uzupełniałem je przy użyciu progresji arytmetycznej, na przykład dla danych:
1990 - 2.0
1993 - 3.5
tworzyłem dane:
1991 - 2.5
1992 - 3.0
Powtarzałem ten proces dla każdego brakującego roku w każdej bazie danych, a następnie złączyłem już uzupełnione dane w jeden DataFrame dla wygodnej pracy z nimi. Cały ten proces zawarty jest w process_data.py.

2.4 Wizualizacja danych

W tym momencie, dane są już gotowe, aczkolwiek nie wiadomo, czy mają one jakikolwiek związek między sobą. W tym momencie dane wyglądają w następujący sposób:

Country	Code	Year	Population	Pollution	CO2	GDP	Vehicles	VPC	Area
Afghanistan	AFG	1990	10694796	65.486794	0.191745	356.0	800000	0.074803	652864.0
Afghanistan	AFG	1991	11677927	65.409973	0.167682	356.0	800000	0.068505	652864.0
Afghanistan	AFG	1992	12661058	65.333153	0.095958	356.0	800000	0.063186	652864.0
Afghanistan	AFG	1993	13644189	65.256332	0.084721	356.0	800000	0.058633	652864.0
Afghanistan	AFG	1994	14627320	65.179512	0.075546	356.0	800000	0.054692	652864.0

Table 1: Dane dla Afganistanu w latach 1990-1994

Jak widać, są to dane bardzo zróżnicowane, więc warto znaleźć zależności między nimi. Do tego posłużymy się wykresami pokazującymi zależności wartości zanieczyszczenia od innych wartości (tworzenie tych wykresów zawarte jest w draw_functions.py.):

Uwaga: poniższe dane skupiają się na 92.5% wszystkich danych, aby uniknąć anomalnie dużych lub małych wartości.

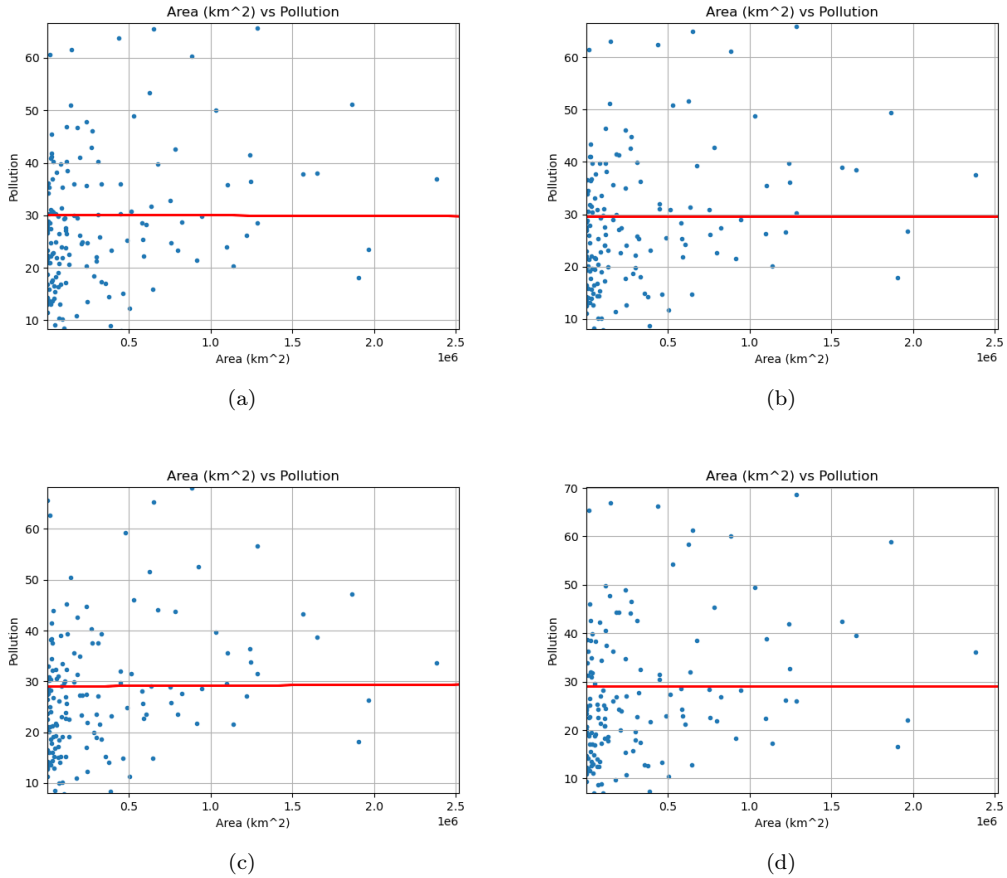

Figure 1: Rozmiar kraju vs Zanieczyszczenie (a) 1990 (b) 2000 (c) 2010 (d) 2015

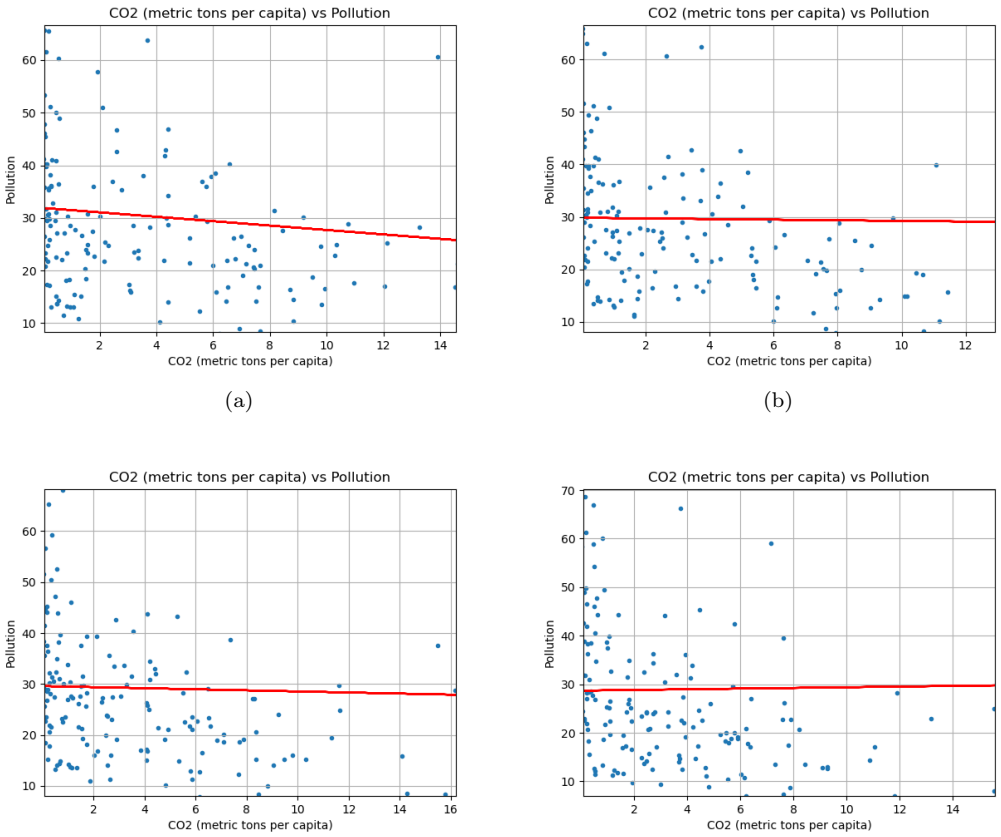

Figure 2: CO2 vs Zanieczyszczenie (a) 1990 (b) 2000 (c) 2010 (d) 2015

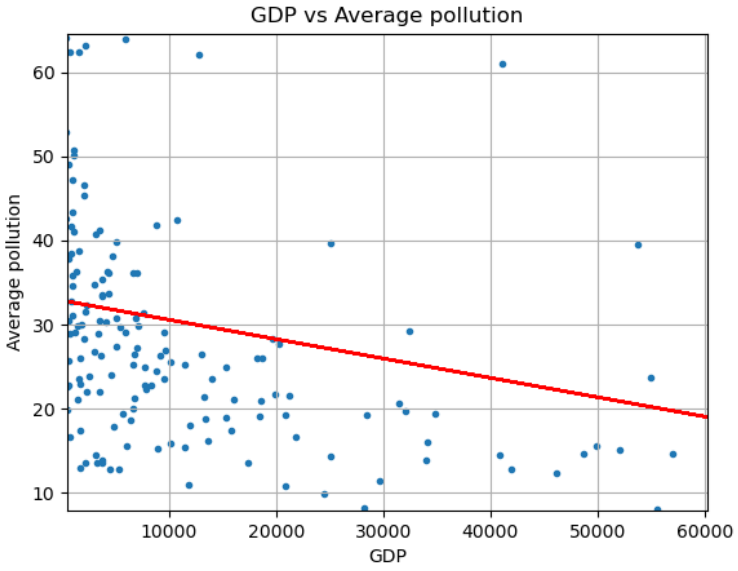

Figure 3: GDP vs Średnia zanieczyszczenia kraju przez lata

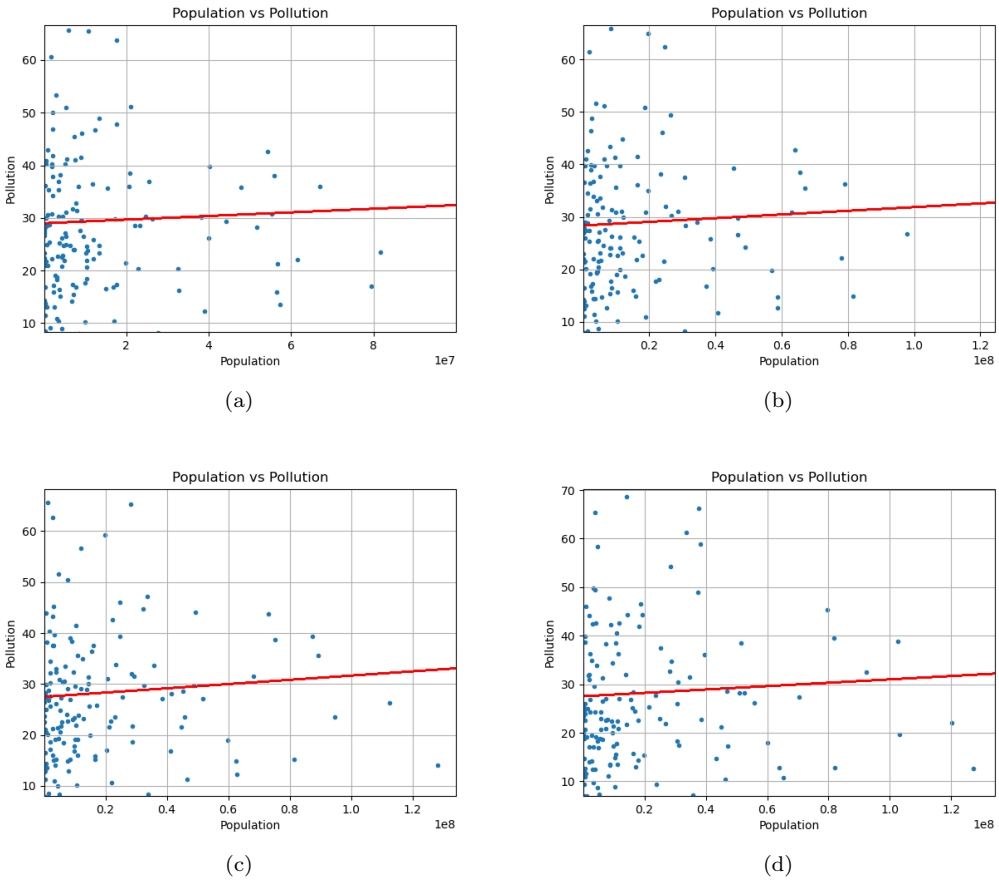

Figure 4: Populacja vs Zanieczyszczenie (a) 1990 (b) 2000 (c) 2010 (d) 2015

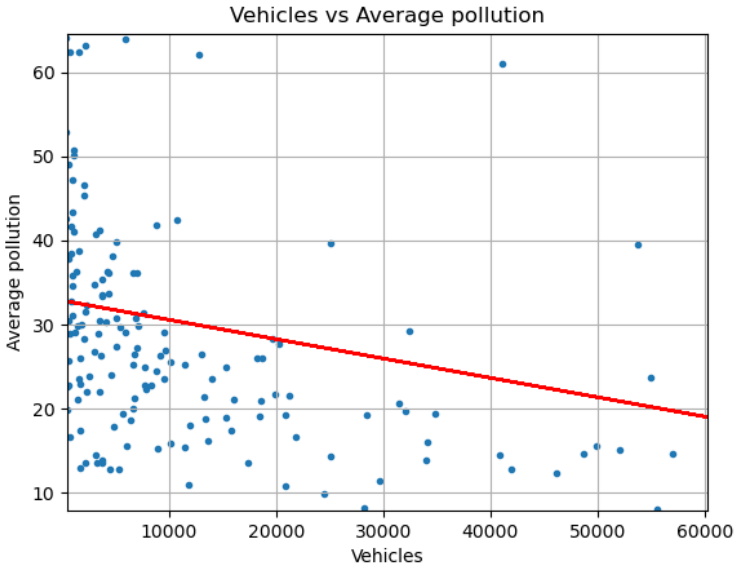

Figure 5: Ilość pojazdów vs Średnia zanieczyszczenia kraju przez lata

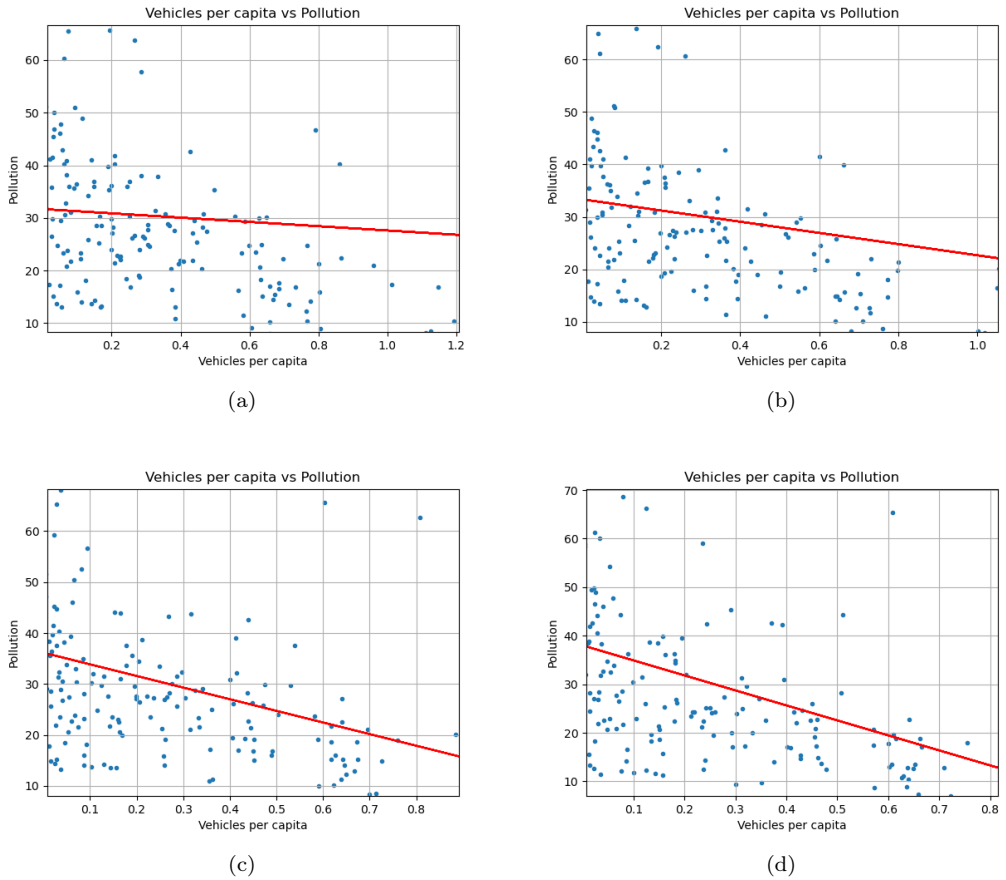

Figure 6: Pojazdy na osobę vs Zanieczyszczenie (a) 1990 (b) 2000 (c) 2010 (d) 2015

3 Modele

Gdy mamy już gotowe dane, czas na utworzenie modeli, które będą przewidywać wartość zanieczyszczenia powietrza dla zadanych wartości innych czynników. Utworzyłem cztery modele, każdy z których działa w różny sposób, w celu ich porównania. Parametry tych modeli były dopasowywane przez GridSearch (zawarty w grid_search.py) lub ręcznie, gdy użycie tej metody wiązało się ze zbyt dużymi komplikacjami implementacyjnymi. Dane testowe to 10% ze wszystkich danych wylosowane przy użyciu random_state=42.

3.1 Model regresji liniowej

Model regresji liniowej jest narzędziem statystycznym służącym do modelowania i analizy relacji między jedną zmienną zależną a jedną lub więcej zmiennymi niezależnymi. Celem regresji liniowej jest znalezienie najlepiej dopasowanej płaszczyzny do danych, która minimalizuje różnicę między przewidywanymi a rzeczywistymi wartościami zmiennej zależnej.
Oto wyniki modelu regresji liniowej dla moich danych:

Linear Regression - Mean Absolute Error: 9.708893087800012
Linear Regression - Mean Squared Error: 206.10542179037637
Linear Regression - R2 Score: 0.33200742956879403

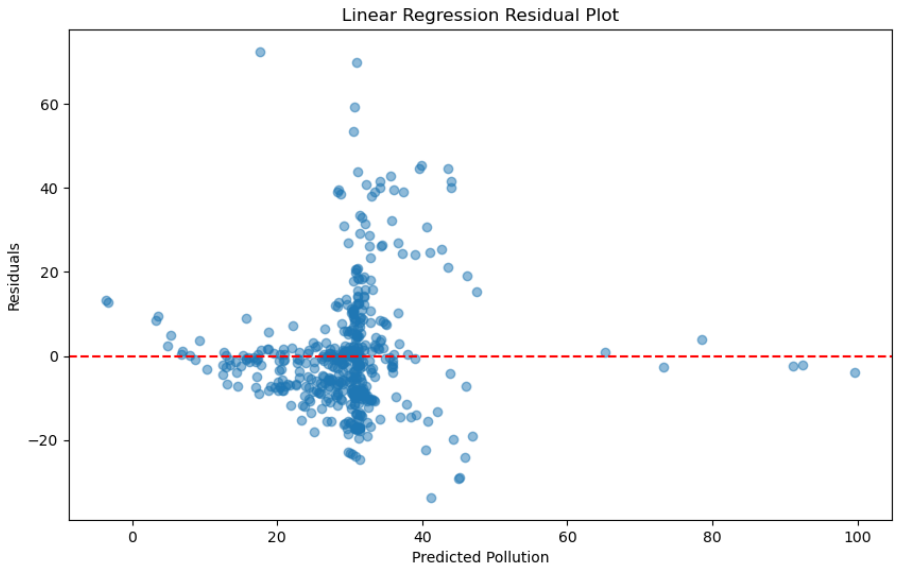

Figure 7: Wartość przewidziana a wartość błędu (Regresja liniowa)

Jak widać, ten model nie sprawdził się za dobrze, co może wskazywać na nieliniowe zależności pomiędzy wartościami zanieczyszczeń a resztą czynników.

3.2 Model Random Forest

Model Random Forest to zaawansowany algorytm uczenia maszynowego, który wykorzystuje wiele drzew decyzyjnych do przewidywania wyników. Random Forest jest stosowany zarówno w problemach klasyfikacji, jak i regresji.
Dla moich danych, model Random Forest otrzymał następujące wyniki:

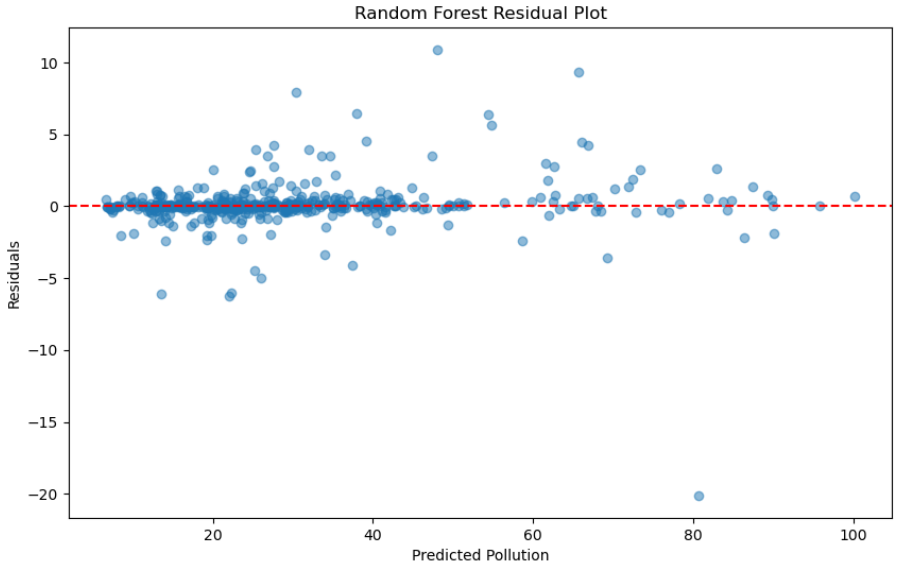

Figure 8: Wartość przewidziana a wartość błędu (Random Forest)

Random Forest - Mean Absolute Error: 0.6781001721493567
Random Forest - Mean Squared Error: 2.8051916338895153
Random Forest - R2 Score: 0.9909083072449214

W przeciwieństwie do poprzedniego modelu, ten niemal idealnie przewidywał wartości zanieczyszczenia prawie w każdym przypadku testowym.

3.3 Model prostej sieci neuronowej

Prosta sieć neuronowa to podstawowy model sztucznej sieci neuronowej, który naśladuje działanie fragmentów systemu biologicznych układów nerwowych. Stworzyłem ten model o 64 neuronach w pierwszej warstwie, 32 w drugiej i 16 w trzeciej, z 1 neuronem w warstwie wyjściowej.
Wyniki tego modelu dla moich danych to:

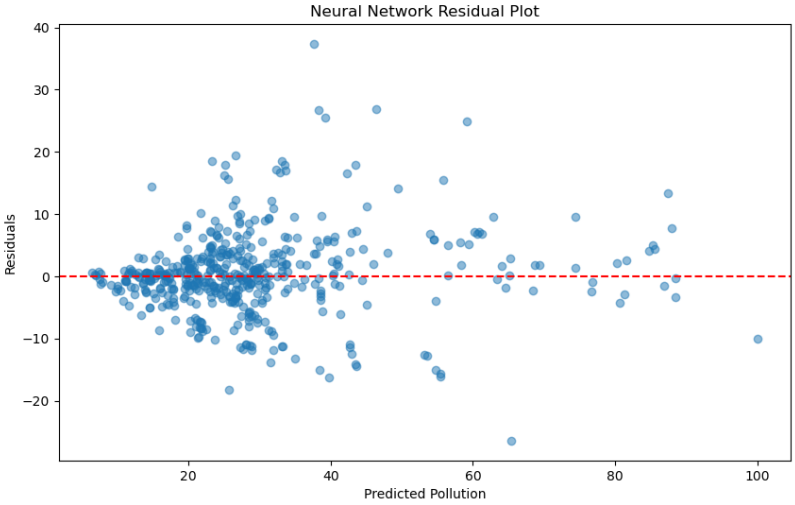

Figure 9: Wartość przewidziana a wartość błędu (Sieć neuronowa)

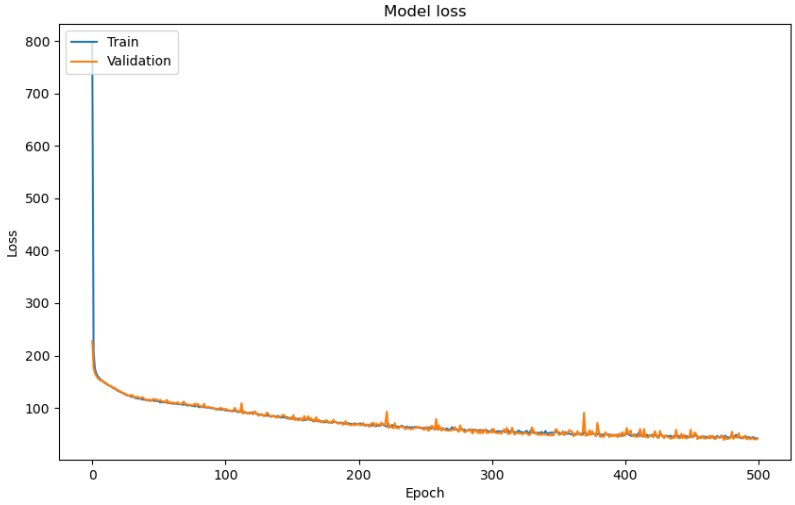

Figure 10: Zmiana straty w miarę postępu (Sieć neuronowa)

Neural Network - Mean Absolute Error: 4.463869042494863
Neural Network - Mean Squared Error: 44.381364281765826
Neural Network - R2 Score: 0.8561589435722222

Model ten nie potrafił przewidywać wartości zanieczyszczenia tak dobrze, jak robił to model Random Forest, ale i tak pokazał dosyć dobre wyniki.

3.4 Model SVM (Support Vector Machine)

Model SVM (Support Vector Machine) to zaawansowana technika uczenia maszynowego stosowana głównie w problemach klasyfikacji, ale również wykorzystywana w regresji i wykrywaniu anomalii. Musiałem ręcznie wybrać kernel dla tego modelu (wybrałem 'rbf'), ponieważ gdy przekazałem wszystkie możliwe kernele do GridSearch wraz z innymi wartościami parametrów, obliczenia zajmowały zbyt dużo czasu.
Wyniki otrzymane przez SVM dla moich danych:

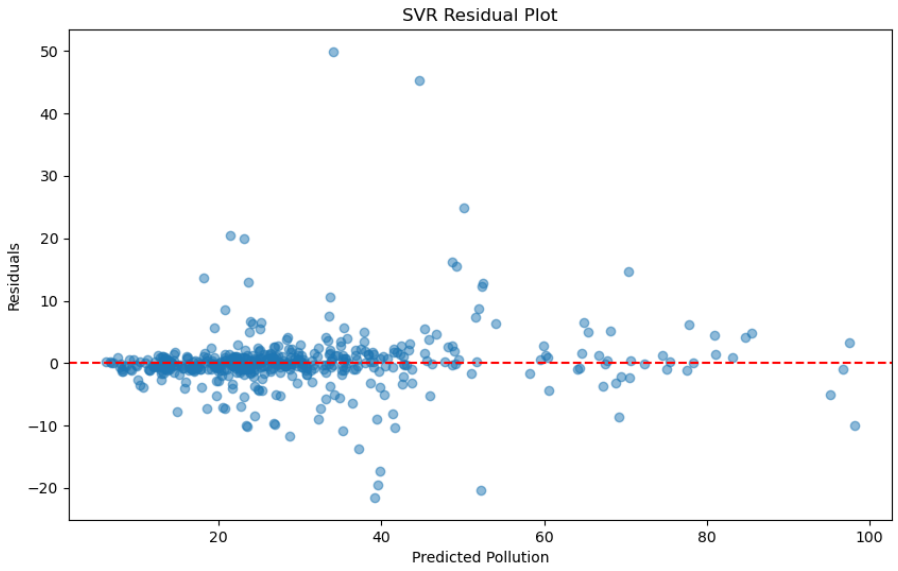

Figure 11: Wartość przewidziana a wartość błędu (SVM)

SVM - Mean Absolute Error: 2.413290442652552
SVM - Mean Squared Error: 26.386026510106557
SVM - R2 Score: 0.914482260976722

SVM okazał się bardziej dokładny niż sieć neuronowa, jednak daleko mu do Random Forest.

4 Wnioski

Przeprowadzona analiza i modele predykcyjne dostarczyły istotnych wniosków na temat jakości powietrza w różnych krajach oraz skuteczności zastosowanych metod predykcji.

4.1 Jakość danych

Dane zebrane z Kaggle i Wikipedii po odpowiednim przetworzeniu okazały się wystarczająco kompletne i spójne do przeprowadzenia analiz. Proces uzupełniania brakujących danych metodą średniej oraz progresji arytmetycznej pozwolił na stworzenie solidnej bazy danych do modelowania.

4.2 Porównanie modeli

R2 Score, od najgorszego do najlepszego:

Regresja liniowa: 0.332
Sieć neuronowa: 0.856
SVM: 0.914
Random Forest: 0.991

Najgorzej wypadł model regresji liniowej, co wskazuje na nieliniowość zależności między danymi. Najlepszy wynik miał Random Forest, co oznacza, że ten model jest najbardziej skuteczny w przewidywaniu zanieczyszczenia powietrza. Sieć neuronowa i SVM miały niezłe wyniki, ale nie są to najlepsze modele to tego typu zadań.

4.3 Rekomendacje i przyszłe kierunki

W dalszej pracy warto rozważyć doskonalenie modelu Random Forest oraz eksplorację innych zaawansowanych technik uczenia maszynowego. Warto również zwiększyć zakres i szczegółowość danych, co może poprawić dokładność modeli i pozwolić na bardziej precyzyjne przewidywanie zanieczyszczeń w przyszłości.

5 Podsumowanie

Projekt ten dostarczył cennych informacji na temat zanieczyszczeń powietrza i wykazał, że model Random Forest jest niezwykle efektywnym narzędziem do ich predykcji. Wyniki te mogą stanowić podstawę do dalszych badań i wdrożenia praktycznych rozwiązań w zakresie monitorowania i poprawy jakości powietrza.