Przewidywanie wskaźnika zanieczyszczenia powietrza AQI, na podstawie różnych typów zanieczyszczeń

Tymoteusz Lango
272742@student.pwr.edu.pl
MSiD Lab Wtorek 7:30 P

Czerwiec 9 2024

Spis treści

1 Wprowadzenie **3**
1.1 Opis problemu 3
1.2 Cel 3

2 Zbiór danych **3**
2.1 Źródło i pozyskanie danych 3
2.2 Wstępne przetwarzanie 3
2.3 Mapa korelacji 5

3 Eksperymenty **6**
3.1 Przygotowanie danych 6
3.2 Badanie długości historii na dokładność modelu 6
3.3 Dane 7

4 Wnioski **9**

1 Wprowadzenie

1.1 Opis problemu

Rozpatrywanym problemem jest zbadanie wpływu różnych rozajów zanieczyszczeń na ogólną jakość powietrza. Analiza pozwoli na przewidywanie jakości powietrza i

1.2 Cel

Celem niniejszego projektu jest stworzenie modelu predykcyjnego, który na podstawie danych dotyczących zanieczyszczeń powietrza różnymi typami zanieczyszczeń, będzie w stanie dokładnie przewidzieć poziomy zanieczyszczeń powietrza we Wrocławiu. Parametrami wybrarnimi do analizy są:

- PM10 (pył zawieszony o średnicy 10 mikrometrów lub mniejszej)
- PM2.5 (pył zawieszony o średnicy 2.5 mikrometra lub mniejszej)
- O3 (ozon)
- CO (tlenek węgla)
- NO2 (dwutlenek azotu)
- SO2 (dwutlenek siarki)

Są to podstawowe czynniki brane pod uwagę podczas obliczania indeksu jakości powietrza AQI.

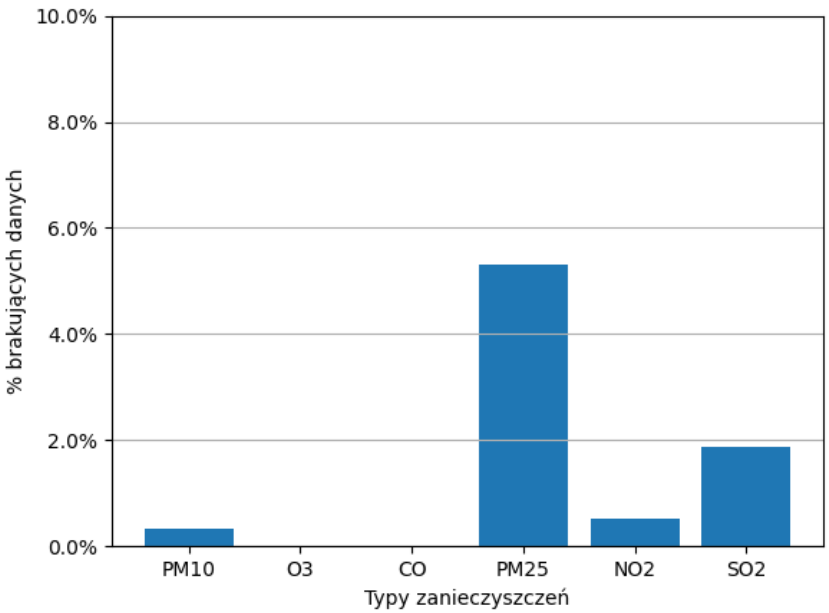
2 Zbiór danych

2.1 Źródło i pozyskanie danych

Dane wykorzystane do rozwiązania problemu zostały pobrane za pomocą publicznego banku danych pomiarowych udostępnionego przez Główny Inspektorat Ochrony Środowiska na stronie gios.gov.pl. Dane zostały pobrane wykorzystując moduł BeautifulSoup w formacie xlsx.

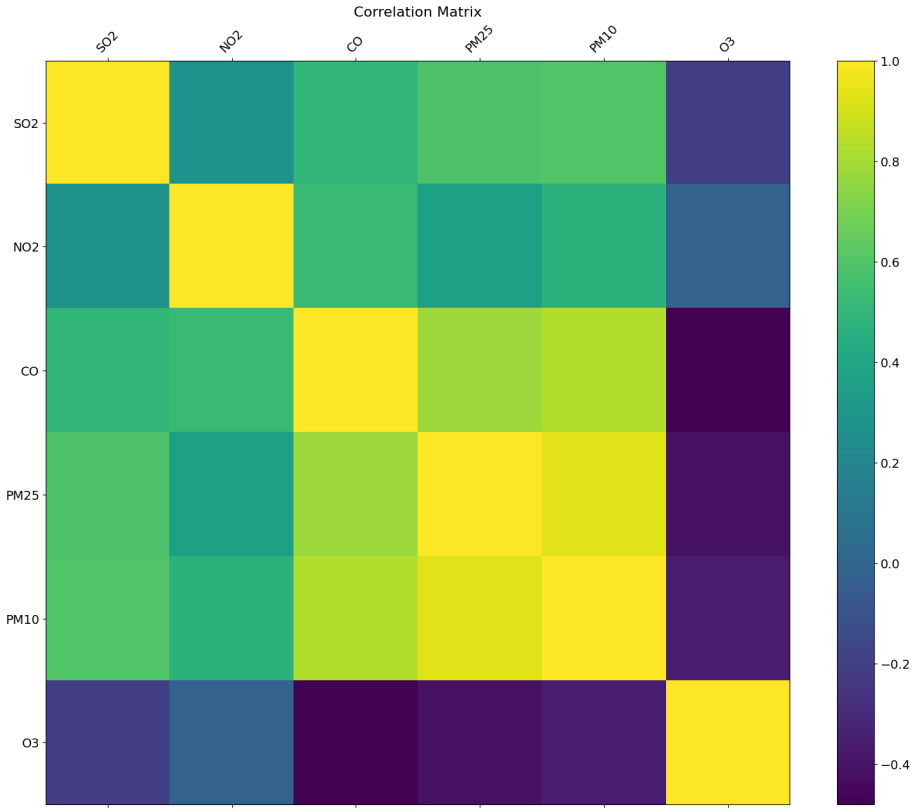
2.2 Wstępne przetwarzanie

Dane zawierają w sobie pomiary z każdego dnia lub godziny w zależności od typu zanieczyszczenia. Wartości dla typów PM10 i PM2.5 są mierzone w uśrednieniu do całego dnia, natomiast pozostałe czynniki były mierzone co godzinę. Biorąc pod uwagę braki czujników w latach 2005 - 2014 oraz częste braki w danych, zbiór został ograniczony do lat 2015 - 2022 r. Braki pojawiające się w tym zbiorze są znikome a procent brakujących danych prezentuje się następująco:



Rysunek 1: % brakujących wartości dla poszczególnych typów zanieczyszczeń

2.3 Mapa korelacji



Rysunek 2: Mapa korelacji

Analizując mapę korelacji parametrów możemy zauważyć że ozon (O3) i tlenek węgla (CO) mają największy wpływ na końcowy wynik jakości powietrza.

3 Eksperymenty

3.1 Przygotowanie danych

Przygotowując dane do dalszej pracy z modelem konieczne było obliczenie AQI dla każdego z parametrów, co opisuje się następującym wzorem:

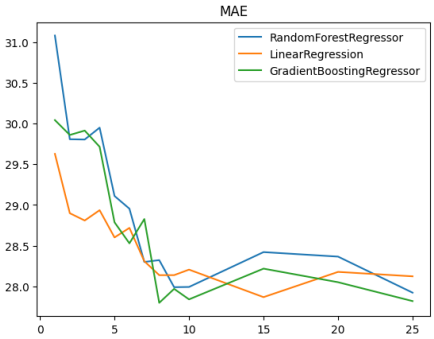
$$I_p = \frac{I_{Hi} - I_{Lo}}{BP_{Hi} - BP_{Lo}}(C_p - BP_{Lo}) + I_{Lo} \tag{1}$$

Gdzie

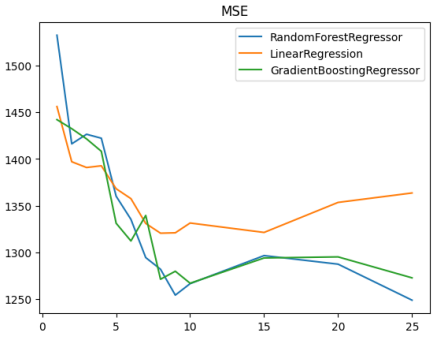
- Ip = indeks dla zanieczyszczenia, Cp = obcięte stężenie zanieczyszczenia
- BP_Hi = punkt załamania stężenia, który jest większy lub równy C_p
- BP_Lo = punkt załamania stężenia, który jest mniejszy lub równy C_p
- I_Hi = wartość AQI odpowiadająca BP_Hi
- I_Lo = wartość AQI odpowiadająca BP_Lo

3.2 Badanie długości historii na dokładność modelu

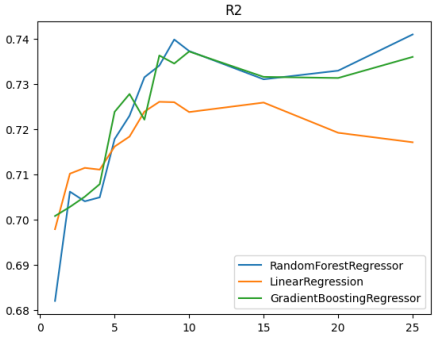
W celu osiągnięcia najlepszej dokładności modelu, zostały wytrenowane modele regresji liniowej, Random Forest Regressor, Gradient Boosting Regressor biorące pod uwagę od 1 do 25 dni wstecz. Wszystkie modele zostały wytrenowane jako modele autoregresywne, tj. patrzące na poprzednie pomiary w celu przewidzenia kolejnych.



Rysunek 3: Średni błąd bezwzględny w zależności od długości historii analizowanej przez model



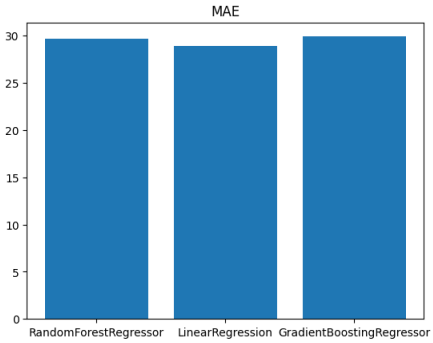
Rysunek 4: Błąd średniokwadratowy w zależności od długości historii analizowanej przez model



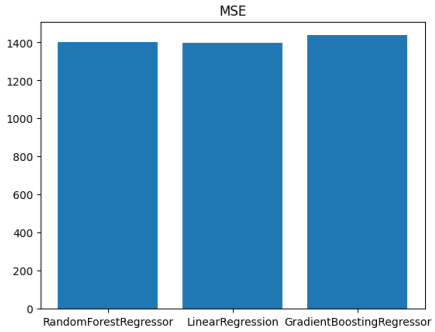
Rysunek 5: Współczynik determinacji w zależności od długości historii analizowanej przez model

3.3 Dane

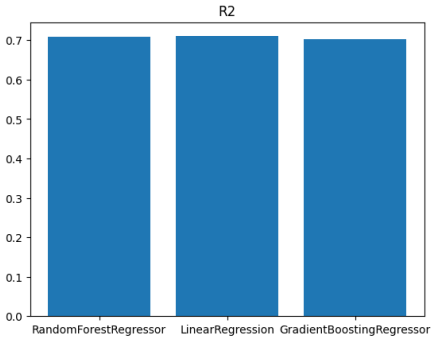
Patrząc na metryki jakości modeli, które zostały wykorzystane do porównania to: błąd średniokwadratowy (MSE), średni błąd bezwzględny (MAE) oraz współczynnik determinacji R2. Na danych wykresach bierzemy pod uwagę 2 dni historyczne i można zauważyć, że model regresji liniowej wypada najlepiej.



Rysunek 6: Średni błąd bezwzględny dla 2 dni historycznych

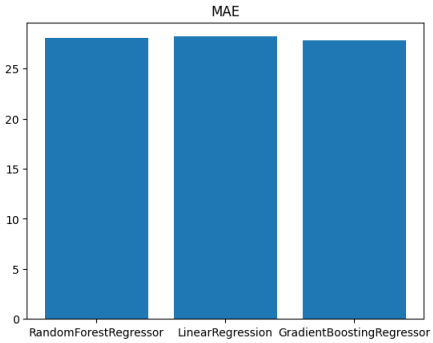


Rysunek 7: Błąd średniokwadratowy dla 2 dni historycznych

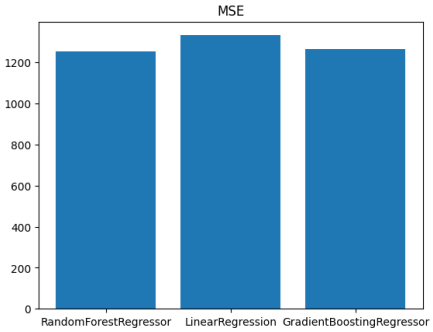


Rysunek 8: Współczynik determinacji dla 2 dni historycznych

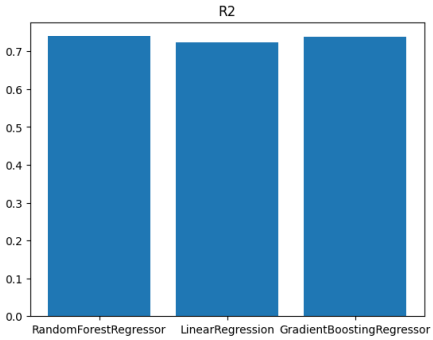
Jednak patrząc na dane dla 10 dni historycznych najlepiej wypada model lasów losowych (Random Forest Regressor) co pokazuje, że ilość dni branych pod uwagę przy przewidywaniu wartości zanieczyszczenia ma wpływ na dokładność modeli.



Rysunek 9: Średni błąd bezwzględny dla 10 dni historycznych



Rysunek 10: Błąd średniokwadratowy dla 10 dni historycznych



Rysunek 11: Współczynik determinacji dla 10 dni historycznych

4 Wnioski

Przewidywanie zanieczyszczeń powietrza jednynie na podstawie historii pomiarów jest trudnym zadaniem. Najlepsze wyniki uzyskane w tej pracy miały błąd wartości bezwzględnej równy 28 na zbiorze testowym, co wynosi ponad 5% zakresu wartości współczynnika AQI. Model może więc służyć do ogólnej estymacji trendu, ale jego wyniki są obarczone niepomijalnym błędem.