Zofia Grzebita
Maciej Pazdecki
Gabriela Rydwelska

PROBLEMATYKA:

Wykorzystując dane godzinowe PM2.5 z Ambasady USA w Pekinie, a także dane meteorologiczne z międzynarodowego lotniska w Pekinie przewidujemy poziom zanieczyszczenia powietrza. Prognoza ta pomaga zrozumieć aktualną jakość powietrza, a także potencjalne zagrożenie dla zdrowia publicznego. Pozwala ona także na podjęcie odpowiednich środków ostrożności, a także ostrzeżeniem przed zbyt wysokim poziomem zanieczyszczenia i zapobieganie niepożądanym skutkom złej jakości powietrza.

PRZYGOTOWANIE DANYCH:

Prace nad przygotowaniem danych rozpoczynamy od wczytania pliku csv i przechowywanych w obiekcie danych, typu DataFrame. W kolejnym kroku odczytujemy nazwy kolumn z pliku:

- *Numer wiersza:* Numer identyfikacyjny danego wiersza w tabeli.
- *Rok:* Rok, do którego odnoszą się dane w danym wierszu
- *Miesiąc:* Miesiąc, do którego odnoszą się dane w danym wierszu.
- *Dzień:* Dzień, do którego odnoszą się dane w danym wierszu.
- *Godzina:* Godzina, do której odnoszą się dane w danym wierszu.
- *PM2,5:*Stężenie pyłu PM2,5 (cząstki o średnicy mniejszej niż 2,5 mikrometra) w jednostce ug/m3.
- *DEWP:* Punkt rosy.
- *TEMP:* temperatura w stopniach Celsjusza
- *PRES:* (hPa) Ciśnienie
- *cbwd:* symbole wiatru, skrótowe (SE - południowy wschód, NW - północny zachód, NE - północny wschód).
- *Iws:* Skumulowana prędkość wiatru w metrach na sekundę (m/s).
- *Is:* Skumulowane godziny opadów śniegu.
- *Ir:* Skumulowane godziny opadów deszczu.

Kolejny etap to czyszczenie danych, składa się na niego:

- Wypełnienie brakujących wartości w kolumnach: 'year', 'month', 'day'.
- Uzupełnienie wartości w kolumnach 'hour', 'DEWP', 'TEMP', 'PRES'.
- Usuwane zostają wiersze, gdzie występuje wartość NaN, w kolumnie PM2.5
- Zostaje utworzona kolumna ‘Timestamp’, która powstaje na podstawie kolumn 'year', 'month', 'day', 'hour', co pozwala nam na utworzenie obiektu - dataframe tworzony jest obiekt typu datetime, który jest odpowiedzialny za datę i godzinę pomiaru.
- Zostają także usunięte zbędne kolumny, które zawierają ‘'No', 'year', 'month', 'day', 'hour'.

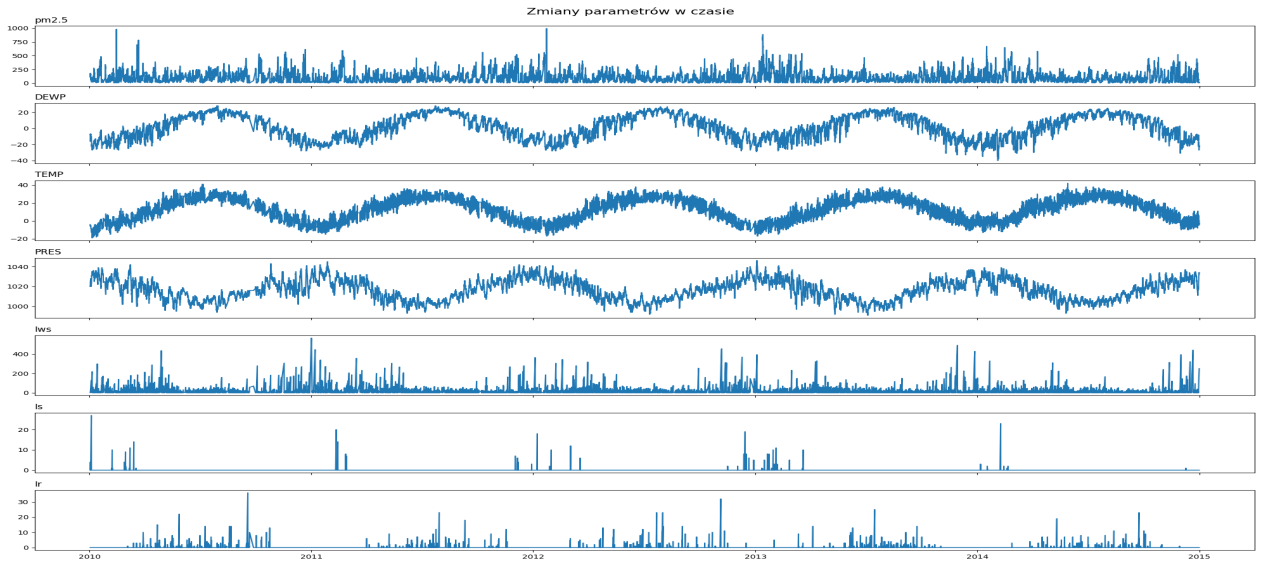
W kolejnej części przygotowanego kodu, występuje kodowanie zmiennych kategorycznych, przy zastosowaniu Label Encoding, który pozwala na zmianę zmiennych kategorycznych na numeryczne, co jest niezbędne do ujednolicenia naszych danych. Dodatkowo zostaje utworzona kolumna ‘'pm_1h_after', która zawiera wartości PM2.5 z następnego zapisu. Kolumna 'pm_1h_after' staje się wartością wychodzącą Y, która jest skutkiem wartości które osiągały atrybuty w poprzedniej godzinie. Jednym z ostatnich kroków jest normalizacja danych numerycznych, przy użyciu MinMaxScaler, co pozwala na skalowanie danych do zakresu [0,1]. Zostaje jeszcze przygotowanie danych do modelowania. Dzieje się to przez podzielenie danych na zbiór treningowy i testowy, dodatkowo wyodrębnione zostają właściwe kolumny, które są atrybutami (X) i zmiennymi docelowymi (Y). Podjęliśmy decyzję o innym podziale danych na testowe niż za pomocą funkcji *train_test_split* która zakłada losowość pobierania danych. Nam zależy na nauce dla każdego okresu.

Eksploracyjna analiza danych

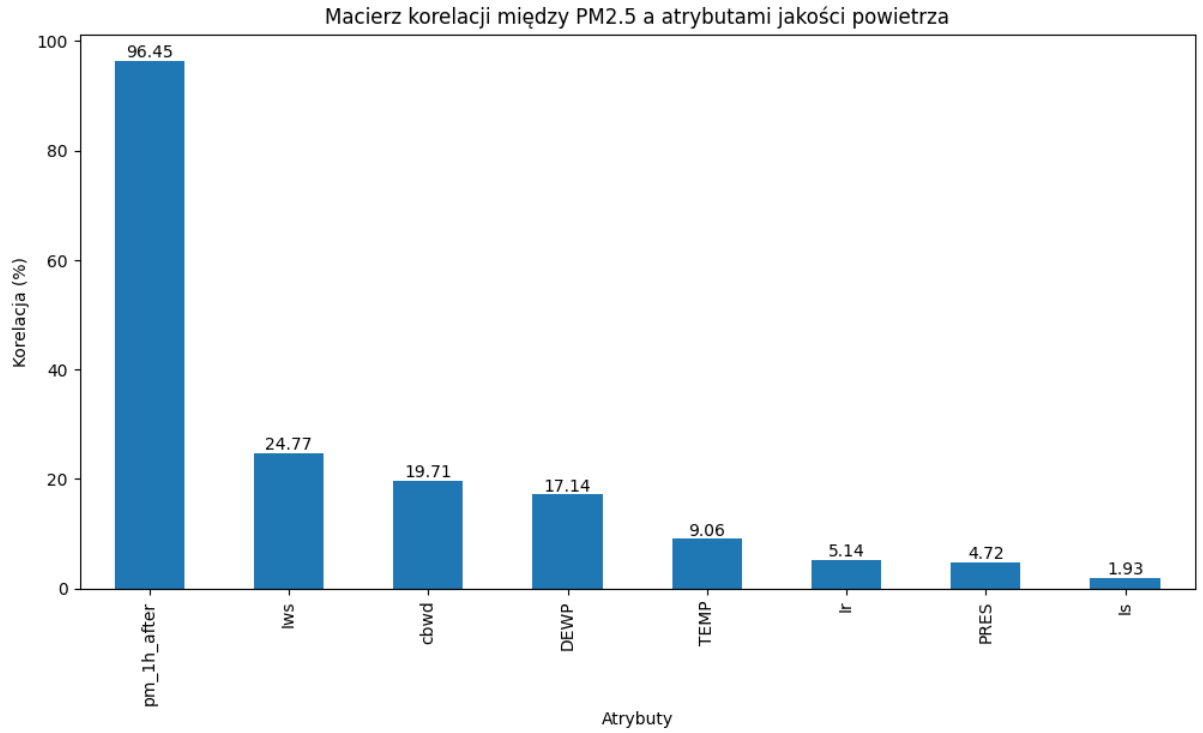
W analizie danych wykorzystaliśmy następujące kolumny: pm2.5 - Stężenie pyłu osadowego, DEWP - punkt rosy, TEMP - temperatura, PRES - ciśnienie, cbwd - kierunek wiatru, Iws - prędkość wiatru, Is - opady deszu, Ir - opady śniegu, Timestamp - kolumna powstała z kolumn YEAR, MONTH, DAY, HOUR, pm_1h_after - prognoza jakości powietrza na jedną godzinę w przód.

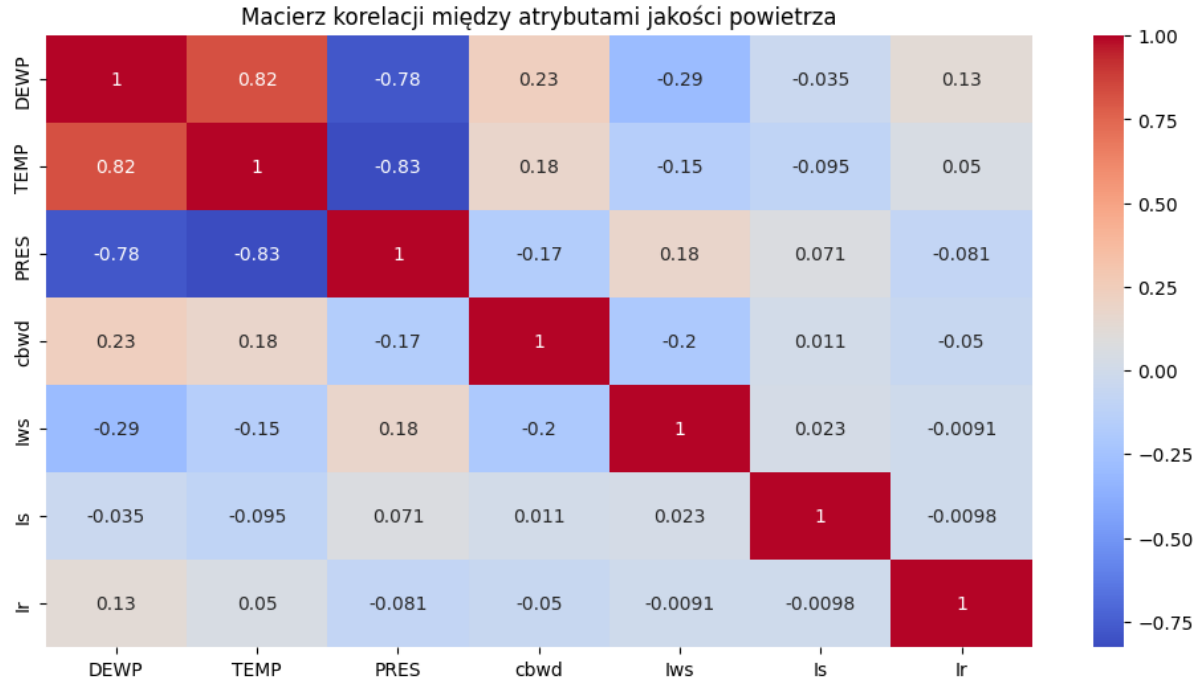
Obserwacje na podstawie poniższego wykresu

Najbardziej istotną obserwacją jest sezonowość danych i widoczne pory roku - zwiększone stężenie pyłu osadowego w okresie zimowym, prawdopodobnie ze względu na ogrzewanie mieszkań. Występuje korelacja ujemna pomiędzy ciśnieniem, a temperaturą powietrza.



Nie powinno nas dziwić, że najsilniej skorelowaną cechą jest prognozowana jakość powietrza na jedną godzinę do przodu. Istotna korelacja występuje również z prędkością wiatru, kierunkiem wiatru, punktem rosy i temperaturą. Zdecydowaliśmy o pozostawieniu wszystkich przedstawionych cech, ponieważ ich pozostawienie nie wydłużyło znacząco działania modeli.





PROGNOZOWANIE

Do prognozowania stężenia pyłu osadowego w powietrzu wybraliśmy następujące modele, dzięki którym uzyskaliśmy najlepsze wyniki prognoz na danych testowych. Te modele to:

Regresja liniowa

Zaletą tego modelu jest prostota w jego interpretacji oraz niski stopień skomplikowania, co przekłada się na mniejsze zużycie wymaganych zasobów obliczeniowych. Z zauważonych wad możemy wymienić jego liniową zależność pomiędzy zmiennymi, która nie uwzględnia nieliniowych interacji.

LASSO

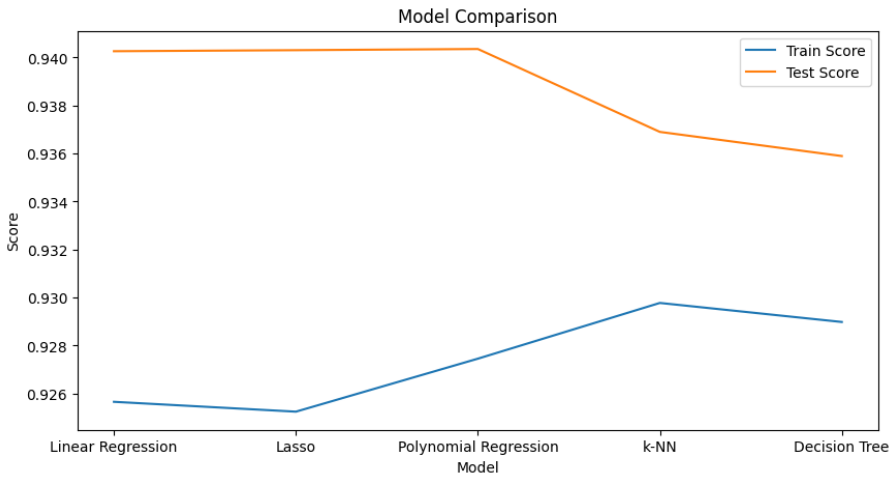
Zaletą tego modelu jest możliwość redukcji zbioru zmiennych do istotnych dla naszej predykcji. Redukuje to problem związany z wielowymiarowością, poprawia interpretowalność modelu, a także zwiększa jego wydajność. Z zauważonych wad możemy wymienić dużą zależność od parametru pm_1h_after

Regresja Wielomianowa

Zaletą tego modelu jest zdolność do dopasowania bardziej odstających od siebie danych, uwzględniając nieliniowe interakcje między zmiennymi oraz pozwala on na większą elastyczność w modelowaniu różnorodnych zjawisk. Jest on rozszerzeniem regresji liniowej. Z zauważonych wad możemy wymienić zwiększoną złożoność modelu, co może prowadzić do overfittingu i trudności z interpretacją wyników.

Przeprowadziliśmy również testy na modelach k-NN oraz drzew decyzyjnych. Pomimo, że wyniki danych treningowych i testowych były bardziej zbliżone dla modeli k_NN oraz Drzew decyzyjnych to zdecydowaliśmy się wybrać wyżej opisane modele, ponieważ osiągnęły najlepsze wyniki dla danych testowych. Wyniki danych testowych dla modeli k-NN oraz Drzew decyzyjnych były mniej satysfakcjonujące.

Porównanie działania modeli na danych treningowych i testowych

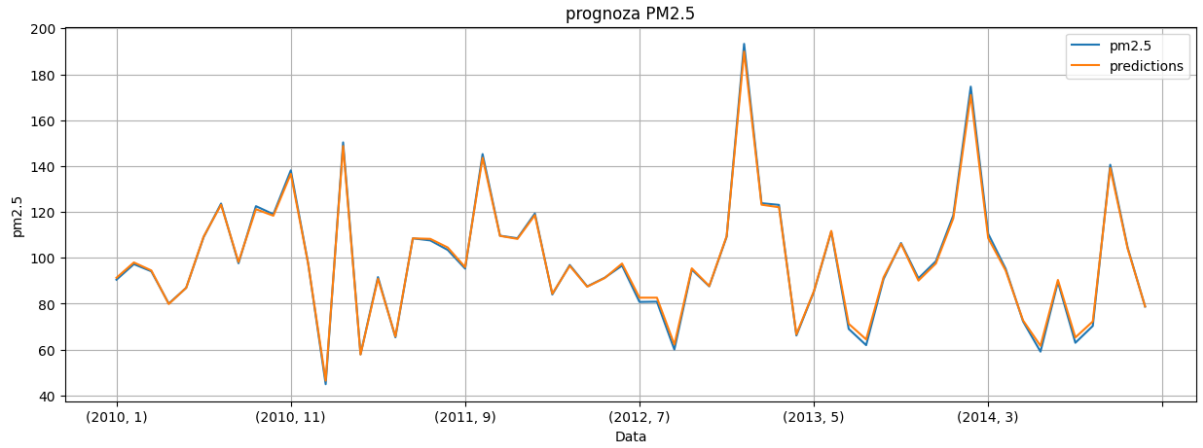


Ocena modeli

Wyniki zostały porównane zbiorowo, ze względu na ich duże podobieństwo. W załączonej poniżej tabeli można zobaczyć, że wyniki R^2 są relatywnie wysokie. Można zatem stwierdzić, że modele są całkiem skuteczne.
Przedstawiają się one następująco:

Model	**RMSE**	**R^2**
Regresja liniowa	21.000874	0.940264
LASSO	20.993305	0.940307
Regresja wielomianowa	20.984684	0.940356

Poniżej znajduje się wykres prognozy dla modelu Regresji Liniowej. Wszystkie modele miały zbliżone wyniki, więc nie ma potrzeby dodawania wykresu dla każdego z modeli.



Wnioski i rekomendacje

Każdy z wyżej przedstawionych modeli uzyskał satysfakcjonujące wyniki. Skuteczność modeli jest zależna od wielu czynników związanych z geolokalizacją. Położenie geograficzne miasta będzie miało znaczny wpływ na rozkład wyników. Było wiele możliwych taktyk do wyboru. Wybrana przez nas dotyczy przewidywania przyszłej wartości zanieczyszczeń za godzinę na podstawie danych z tej godziny - dlatego przyjęliśmy ja jako wartość przyszłą, tą którą chcemy przewidywać.

Dla poprawy wyników można zastosować większą ilość parametrów oraz naukę modelu na dłuższym okresie czasowym.