Bugajski_Krzysztof_COVID

https://colab.research.google.com/drive/1vPzKKPQ5SVUqB_ZV6P8cIg2DH1WOXg7_?usp=sharing

1. Zapoznanie się z danymi

COVID-19 Data Hub to zbiór danych na temat przebiegu pandemii koronawirusa zawierający dane z całego świata. Dane dzielą się na kilka kategorii: Identifiers, Epidemiological variables, Policy measures, Government response, Administrative areas, Coordinates, ISO codes, External keys. Przedstawiają one przebieg przypadków zachorowań, śmierci, ozdrowień, szczepień a także różnych działań podjętych przez rządy danego kraju w czasie trwania pandemii. Link do zbioru danych: covid19datahub.io

2. Cel analizy

Celem analizy było sprawdzenie czy na podstawie zgromadzonych danych można dokonać globalnej i lokalnej (kraj) predykcji liczby zachorowań i liczby śmierci spowodowanej chorobą COVID-19. Dodatkowo, analiza miała na celu identyfikację kluczowych czynników, które miały największy wpływ na zmienność liczby zachorowań i zgonów w czasie oraz wyciągnięcia wniosków na temat przebiegu pandemii COVID-19.

3. Analiza lokalna a analiza globalna

Rozpatrywanie tego problemu globalnie bez stosowania indywidualnego podejścia do różnych krajów ze względu na ich sposób raportowania, dokładność i dostępność danych jest niemożliwe. Ponadto należy zwrócić uwagę na takie cechy kraju jak: jakość służby zdrowia, społeczeństwo i jego mentalność, klimat a także poziom rozwoju gospodarczego, które mocno wpływały na przebieg pandemii w państwach.

4. Wybór kraju do analizy

Kolumny w tabeli zostały pogrupowane według kategorii opisanych w dokumentacji danych. Następnie dla każdej kategorii została policzona ilość brakujących danych oraz ich procent. Jako istotne kategorie zostały wybrane: Epidemiological_variables, Policy_measures oraz Government_response, ostatnia z nich nie była jednak brana pod uwagę w liczeniu średniej ponieważ jest to kategoria wartości oceniających starania rządu policzona na podstawie danych z kategorii Policy_measures.

category	country	Identifiers		Epidemiological_variables		Policy_measures		Government_response		Administrative_areas		Coordinates		ISO_codes		External_Keys		mean_important_nan
metric		nan	nan_percentage	nan	nan_percentage	nan	nan_percentage	nan	nan_percentage	nan	nan_percentage	nan	nan_percentage	nan	nan_percentage	nan	nan_percentage	
0	Lithuania	0	0.0	0	0.00	294	2.55	84	2.55	1648	50.0	0	0.0	0	0.0	1648	33.33	1.275
1	Peru	0	0.0	761	10.22	109	1.15	82	3.03	1354	50.0	0	0.0	0	0.0	2031	50.00	5.685
2	Chile	0	0.0	1293	11.03	448	3.00	128	3.00	2132	50.0	0	0.0	0	0.0	2132	33.33	7.015
3	Estonia	0	0.0	2946	26.70	196	1.40	56	1.40	2006	50.0	0	0.0	0	0.0	2006	33.33	14.050
4	Portugal	0	0.0	2559	28.72	0	0.00	0	0.00	1620	50.0	0	0.0	0	0.0	1620	33.33	14.360
5	Belgium	0	0.0	2519	18.26	2506	14.27	716	14.27	2508	50.0	0	0.0	0	0.0	2508	33.33	16.265
6	Japan	0	0.0	3071	23.60	1792	10.82	512	10.82	2366	50.0	0	0.0	0	0.0	2366	33.33	17.210
7	Slovenia	0	0.0	3309	33.54	238	1.90	68	1.90	1794	50.0	0	0.0	0	0.0	1794	33.33	17.720
8	Denmark	0	0.0	3667	28.13	1274	7.68	364	7.68	2370	50.0	0	0.0	0	0.0	2370	33.33	17.905
9	Singapore	0	0.0	2859	35.56	56	0.55	16	0.55	1462	50.0	0	0.0	0	0.0	1462	33.33	18.055

Rysunek 1 Tabela zawierająca ilość brakujących danych według kategorii

Następnie na dla wartości mean_important_counts oraz ilości danych w zbiorze dla danego kraju każdemu kraju została przydzielona ranga dla tych 2 wartości, kolejnym krokiem było policzenie wartości średniej tych 2 zmiennych porządkowych i wybranie najlepszego kraju

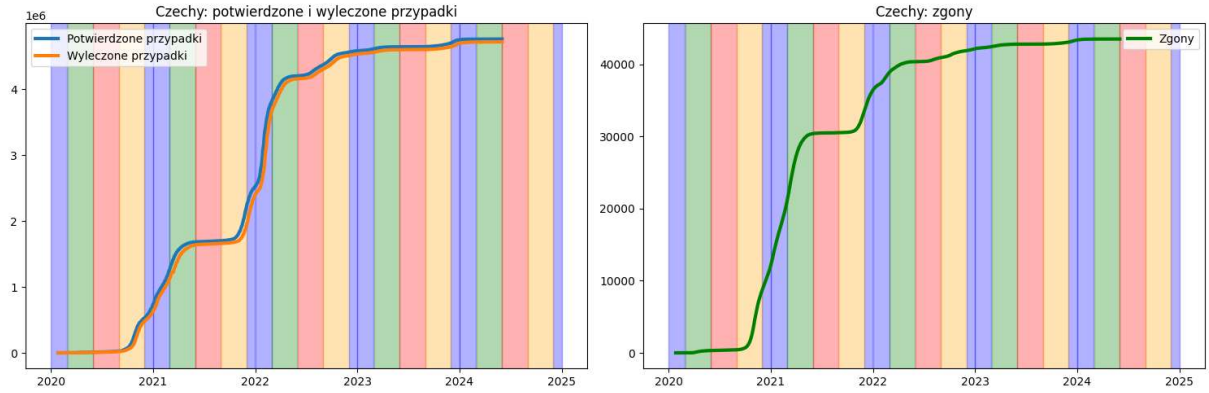
	country	mean_important_nan	counts	rank_important_nan	rank_counts	avg_rank
0	Czech Republic	20.230	1588	14	8	11.0
1	Italy	25.630	1557	30	24	27.0
2	Ireland	31.410	1593	49	7	28.0
3	Malaysia	32.675	1610	54	2	28.0
4	Canada	35.435	1612	71	1	36.0
5	Belgium	16.265	1254	6	84	45.0
6	France	20.585	1266	16	76	46.0
7	Australia	23.050	1288	22	73	47.5
8	Argentina	35.470	1488	72	25	48.5
9	United States	36.030	1573	76	21	48.5

Rysunek 2 Kraje uporządkowane według ilości danych oraz braków

Kraj który zajmował najlepsze miejsce po uwzględnieniu tych 2 rankingów to Czechy, dlatego został on poddany dalszej analizie.

5. EDA – Czechy

Zapoznano się z przebiegiem pandemii w Czechach, stworzono wykresy przedstawiające przebieg ilości zachorowań, ozdrowień oraz zgonów. Na wykresach kolorem niebieskim zaznaczono miesiące zimowe, zielonym wiosnę, czerwonym lato a pomarańczowym jesień. Można dostrzec 2 wyraźne fale nowych zachorowań koronawirusa jakie miały miejsce od jesieni 2020 roku do wiosny 2021 oraz od jesieni 2021 do wiosny 2022 roku.

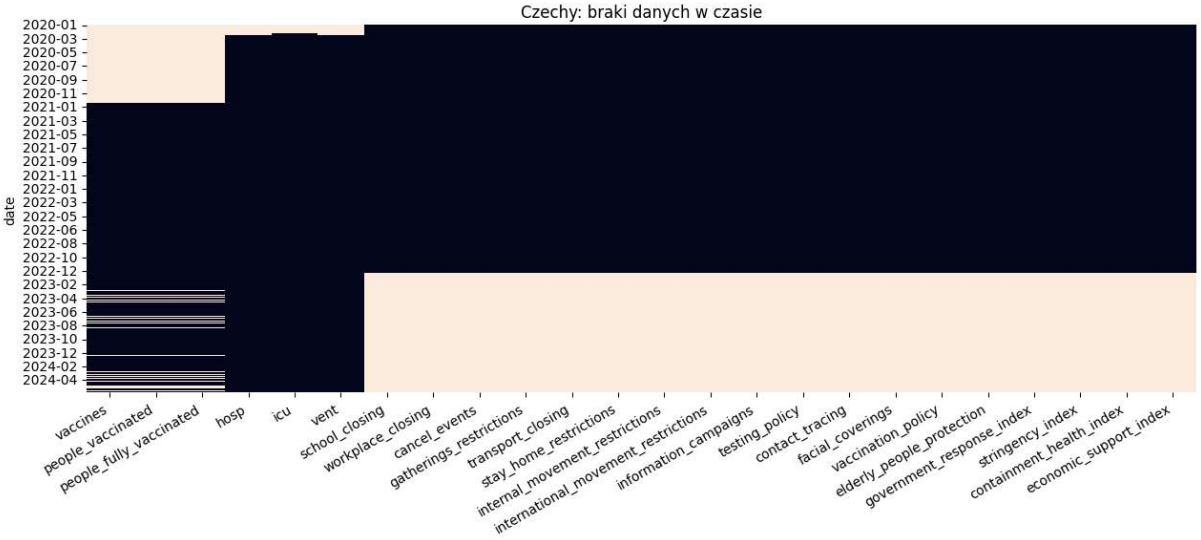


Rysunek 3 Zachorowania, ozdrowienia i zgony w Czechach

W trakcie całej pandemii w Czechach:

- 4 759 716 potwierdzonych przypadków zachorowań
- 4 716 076 wyleczonych przypadków ozdrowień
- 43 523 zgonów
- Śmiertelność wyniosła 0.92%
- W pełni zaszczepiło się 6 893 505 osób, co stanowi 64.85% populacji Czech
- Przeprowadzono 56 308 447 testów, co przy populacji 10 629 928 osób daje około 5 testów na jedną osobę

Zidentyfikowano brakujące wartości w istotnych dla tej analizy zmiennych został utworzony wykres prezentujący ich przebieg w czasie.



Rysunek 4 Brakujące dane w Czechach w trakcie pandemii

Stwierdzono, że brakujące wartości na początku pandemii dla zmiennych związanych ze szczepionkami wynikają z tego iż wtedy jeszcze nie było szczepionki na COVID-19. Analogicznie dla danych reprezentujących przypadki z ciężkim stanem zdrowia osób chorych na COVID stwierdzono, że przez ten zdecydowanie krótszy od danych dotyczących szczepionek okres również nie było takich przypadków z racji początkowego stanu pandemii. Przyjęto, że te wszystkie początkowe braki zostaną uzupełnione wartościami zerowymi.

Ponadto przyjęto, że braki w danych reprezentujących reakcje rządu po 1 stycznia 2023 roku wynikają z tego, że pandemia wtedy już skończyła się w Czechach a wszystkie obostrzenia zostały zniesione, w tym okresie również ilość nowych przypadków jest zdecydowanie mniejsza w odniesieniu do całej pandemii. Zdecydowano się odrzucić wszystkie dane po tej dacie i analizować przebieg pandemii na podstawie danych do 31 grudnia 2022.

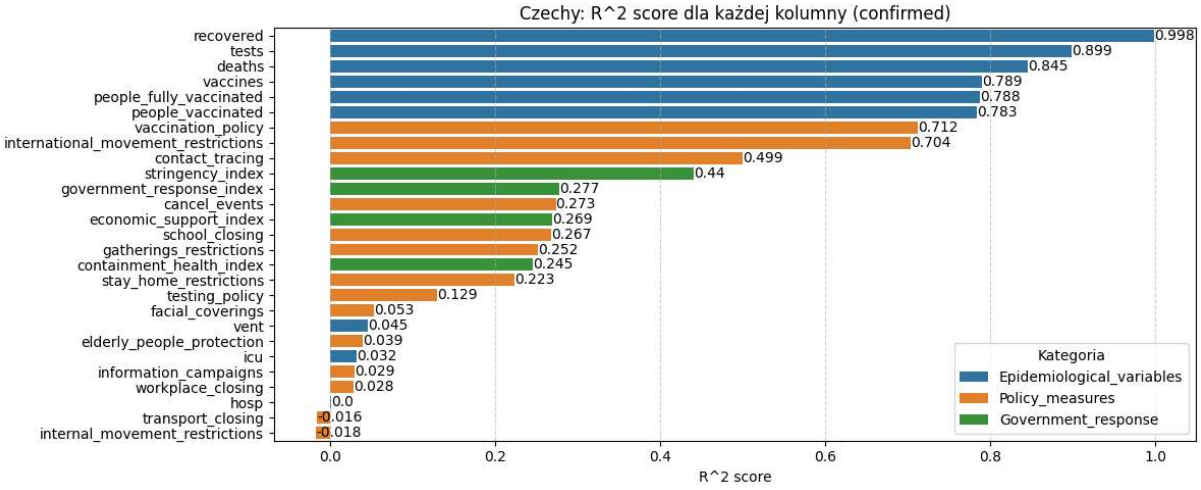
Braki w pewnych momentach w zmiennych dotyczących szczepień wynikają z tego, że w danych dniach nie było szczepionych nowych osób. Jedyną brakującą wartość jaka została po odrzuceniu danych z okresu po pandemii uzupełniono taką wartością jaka występuje przed i po tym dniu, taką prostą operację można wykonać wieloma metodami uzupełniania danych.

date	confirmed	deaths	recovered	tests	vaccines	people_vaccinated	people_fully_vaccinated	hosp	icu	vent
2022-12-22	4576953.0	42083.0	4525521.0	56276145.0	13459494.0	6977614.0	6893420.0	836.0	60.0	15.0
2022-12-23	4577460.0	42094.0	4526613.0	56280993.0	13459518.0	6977625.0	6893437.0	757.0	55.0	14.0
2022-12-24	4577558.0	42099.0	4527545.0	56281919.0	NaN	NaN	NaN	594.0	53.0	15.0
2022-12-25	4577674.0	42104.0	4528381.0	56283166.0	13459518.0	6977625.0	6893437.0	603.0	57.0	15.0
2022-12-26	4577790.0	42108.0	4529132.0	56284402.0	13459519.0	6977626.0	6893438.0	627.0	64.0	17.0

Rysunek 5 Tabela przedstawiająca brakującą wartość w danych na temat szczepień

5. Model regresji liniowej

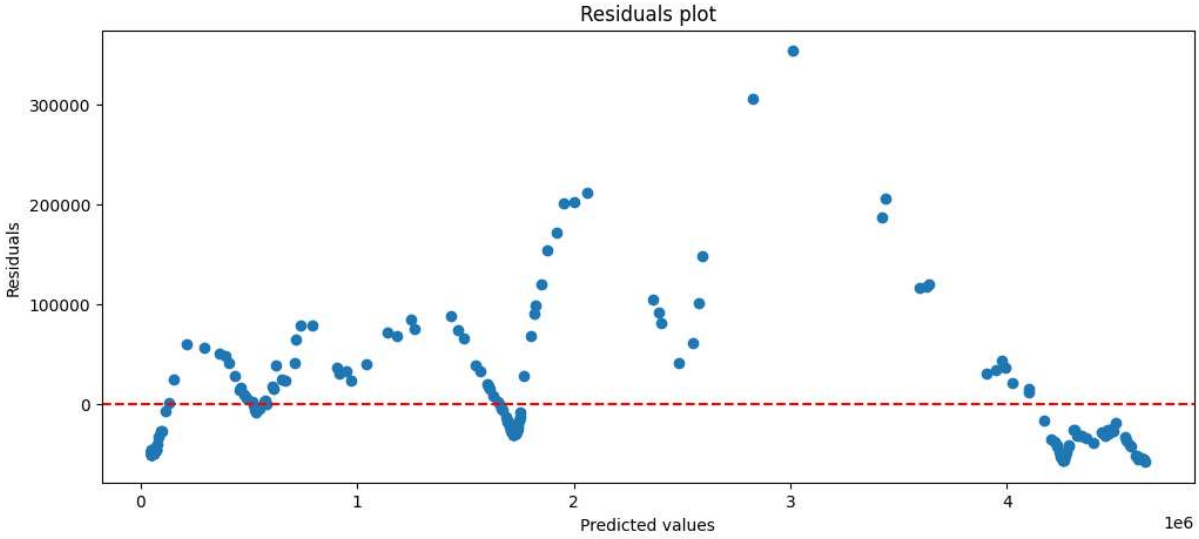
Dla zmiennej zależnej ‘confirmed’ stworzono wiele modeli regresji liniowej, osobny dla każdej zmiennej i sprawdzono ich współczynniki R^2.



Rysunek 6 Współczynniki R2 dla modelu dla zmiennej confirmed

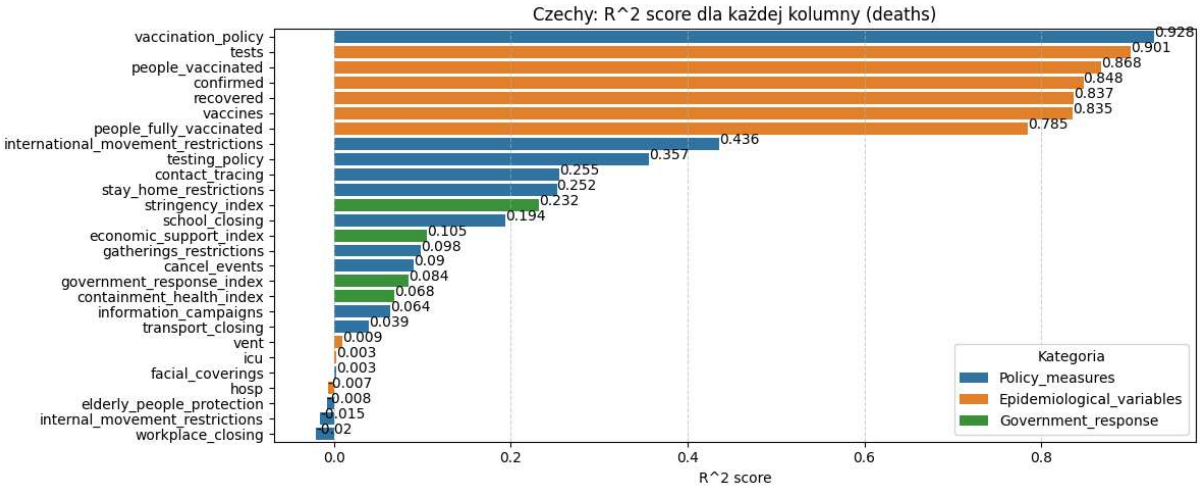
Największe R^2 osiąga zmienna "recovered" jednak przewidywanie wartości zmiennej zachorowań na podstawie wartości ozdrowień to błędne rozumowanie, gdyż liczba ozdrowień jest wynikiem przypadków choroby, która jest zależna od liczby zachorowań, a nie bezpośrednim czynnikiem wpływającym na przyszłe zachorowania. Ponadto śmiertelność w tym zbiorze wynosi tylko 1.2%, co oznacza, że liczba osób które wyzdrowiały jest bardzo zbliżona do osób, które zachorowały

Reszty nie są rozmieszczone losowo i nie spełniają założeń homoskedastyczności, tworzą one wyraźny wzór.



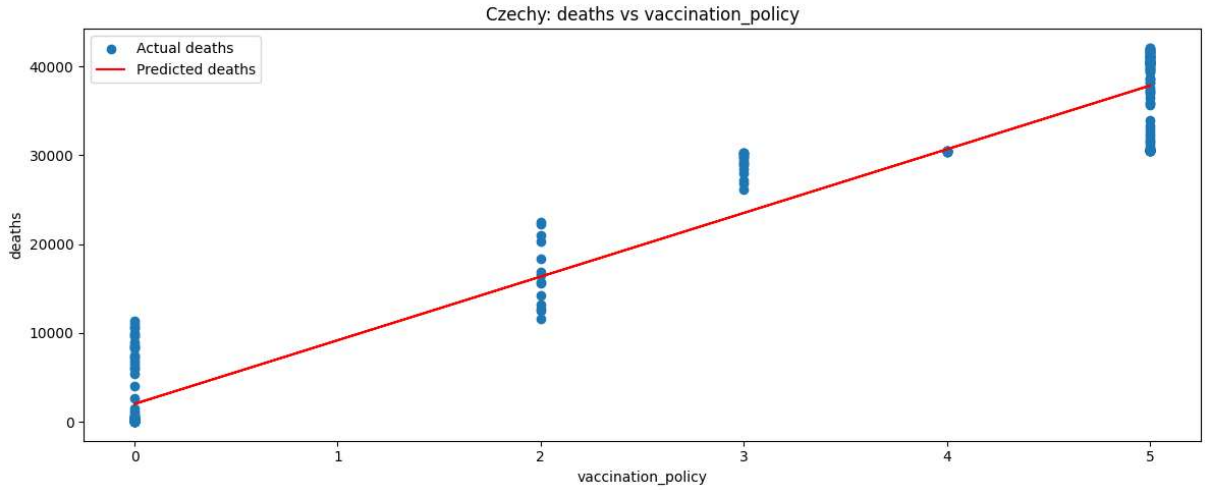
Rysunek 7 Reszty rezyduów dla zmiennej recovered

Takie same modele zostały wykonane również dla zmiennej ‘deaths’



Rysunek 8 Współczynniki R2 dla modelu dla zmiennej deaths

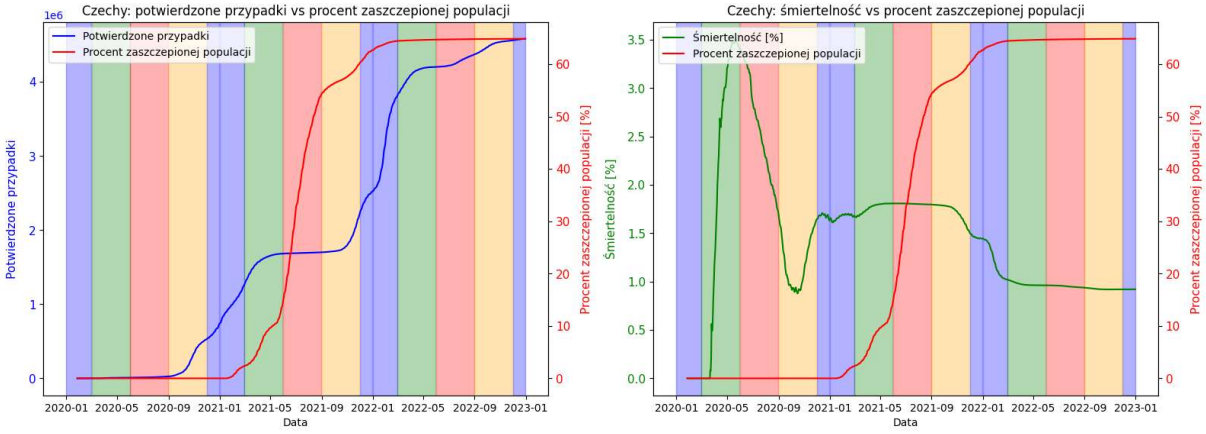
Najlepszą wartość R^2 osiągnął model dla zmiennej vaccination_policy, jednak ze względu na sposób w jaki zapisane są dane w tym zbiorze oraz jego małą kardynalność wynoszącą 5 model ten nie wydaje się być najlepszym rozwiązaniem do przewidywania wartości zgonów.



Rysunek 9 Model zmiennej deaths od vaccination_policy

Model ten przedstawia niepoprawne rozumowanie, jego wyraz wolny 7163.83 to współczynnik nachylenia linii regresji, który pokazuje, jak bardzo zmieni się liczba zgonów przy jednostkowej zmianie wartości zmiennej vaccination_policy. Oznacza to, że lepszy wskaźnik vaccination_policy (szersza dostępność szczepionek) oznacza większą liczbę zgonów. Można spodziewać się, że tempo wzrostu maleje wraz z większymi wartościami vaccination_policy po tym jak ułożone są dane, jednak model regresji liniowej dla tych danych nie oddaje tej zależności.

Lepszym pomysłem wydaje się być stworzenie modelu, który wyjaśnia śmiertelność w zależności od ilości osób zaszczepionych (procenta zaszczepionej populacji), aby to sprawdzić stworzono najpierw 2 wykresy przedstawiające ilość zachorowań z procentem zaszczepionej populacji w czasie oraz śmiertelność wraz z procentem zaszczepionej populacji w czasie

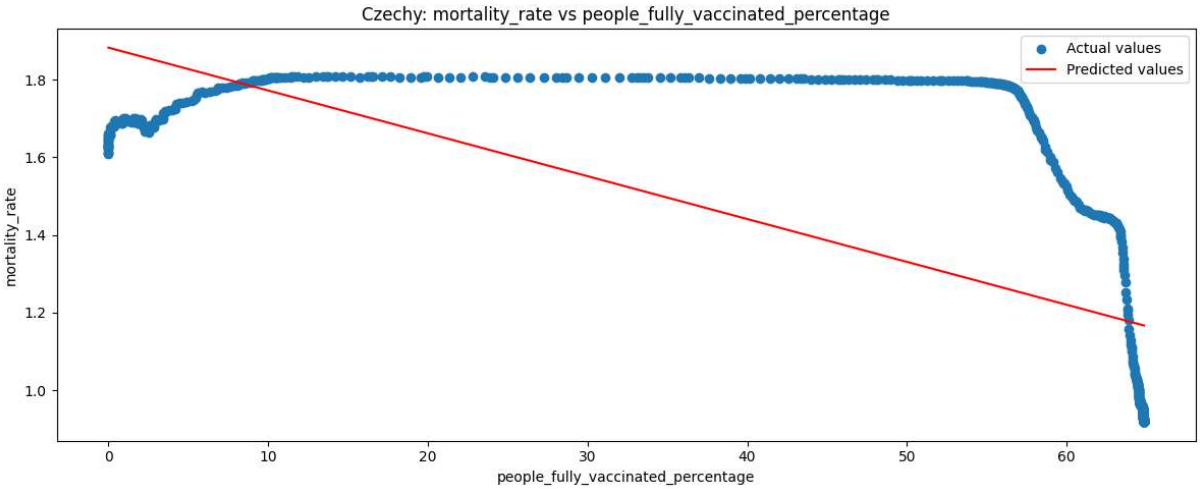


Rysunek 10 Zachorowania, śmiertelność oraz procent zaszczepionej populacji w czasie

Jak można zauważyć procent śmiertelności w lepszy sposób pozwala dostrzec co się działo w trakcie pandemii i jak liczba zaszczepionych osób miała na to wpływ. Widać, że w trakcie 2 fali koronawirusa śmiertelność była zdecydowanie mniejsza. Warto zwrócić też uwagę na to, że w trakcie 1 fali wirus niejako zaczynał zarażanie ludzi od zera, natomiast w 2 fali duża część osób była już wtedy chora, nosicielami itp. przez co wirus miał dużo łatwiejszy początek fali 2. Ponadto zdecydowana większość osób zaszczepiła się w lecie 2021 roku, obserwowano wtedy bardzo mało nowych zakażeń i stałą śmiertelność, najprawdopodobniej głównie wśród grup szczególnie wrażliwych na tę chorobę. Dobrym pomysłem byłoby zamodelowanie tempa rozprzestrzeniania

się wirusa w 1 i 2 fali przyjmując 1 falę jako tą kiedy nie było szczepionek a 2 jako tą kiedy większość społeczeństwa była zaszczepiona, niemniej wymagałoby to dokładnych danych i informacji na temat tego jak rozprzestrzenia się wirus

Następnie stworzono model regresji liniowej śmiertelności od procenta w pełni zaszczepionej populacji. Dla całego zbioru danych model osiągnął bardzo słaby wynik R^2 wynoszący 0.14, aby uniknąć uczenie modelu na danych w których nie było jeszcze prowadzonych szczepień podjęto decyzję o ograniczeniu zakresu danych od momentu kiedy wykonano pierwsze szczepienia



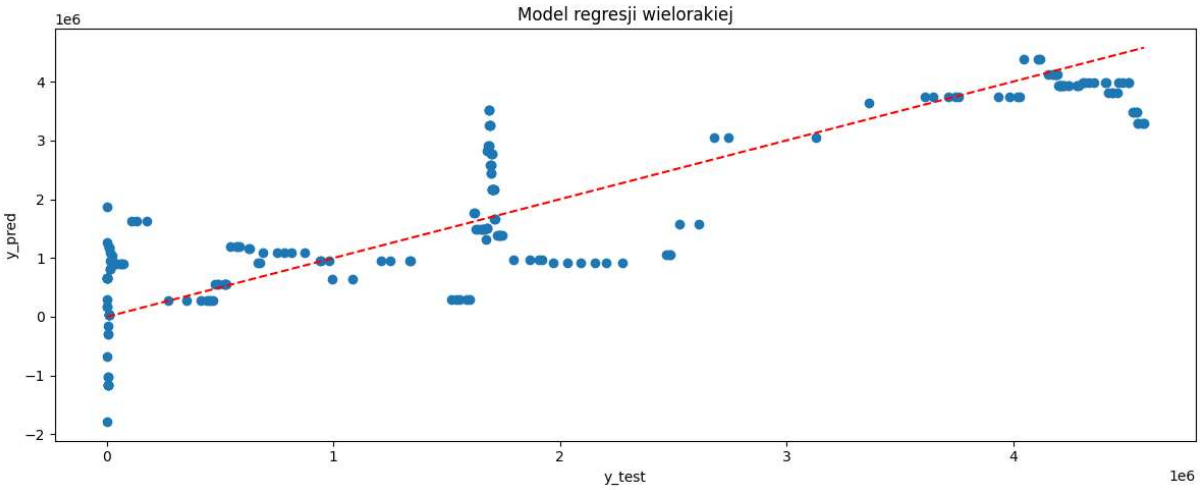
Rysunek 11 Model śmiertelności od procenta zaszczepionej populacji

Przykładowy model śmiertelności od ilości zaszczepionych uzyskuje ujemny współczynnik przy zmiennej 'people_fully_vaccinated_percentage' co potwierdza ideę, że szczepienia zmniejszają śmiertelność populacji. Niemniej jednak ze względu na to, że zdecydowana większość osób w pełni zaszczepionych zaszczepiła się przed połową jesieni 2021, czyli przed 2 falą zachorowań na wykresie obserwujemy gwałtowny spadek i linia predykcji nie pokrywa się z faktycznymi wartościami.

6. Regresja wieloraka

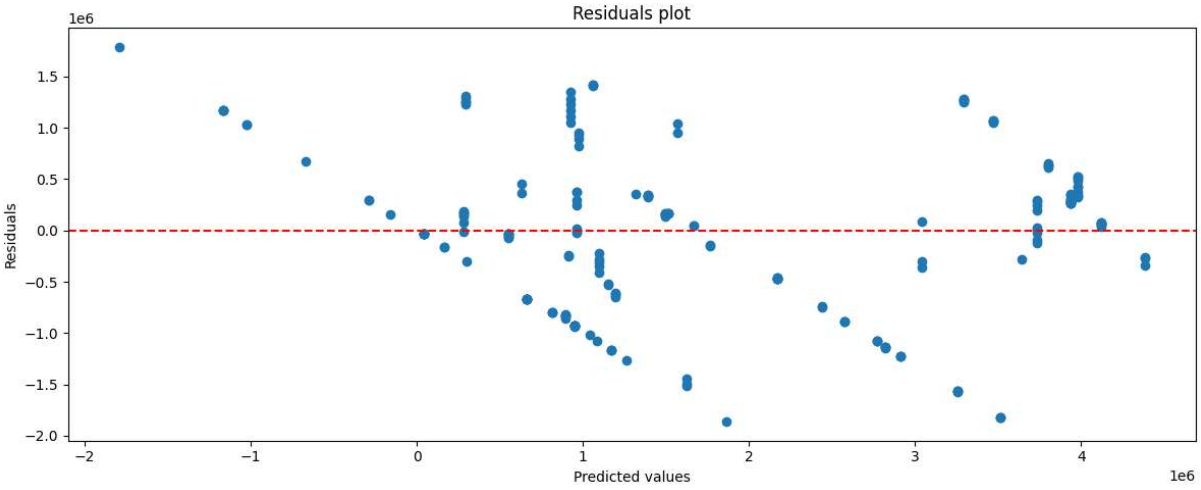
Na wcześniejszych wykresach dotyczących wartości R^2 zauważono, że zmienne z kategorii Policy_measures prezentują bardzo rozbieżne wartości R^2, aby uwzględnić wszystkie działania mające na celu ograniczanie pandemii zdecydowana się do dalszych analiz wykorzystać indeksy z kategorii Government_response, które nie tylko zawierają w sobie informacje o skali lockdownu ale też o innych czynnikach takich jak np. to jak bardzo taki lockdown był respektowany przez ludzi.

Ponadto zidentyfikowane, że zmienna 'economic_support_index' tylko nieznacznie podnosi jakość modelu i dalszą część modelowania można by przeprowadzić bez jej uwzględniania i osiągnąć bardzo zbliżone wyniki, jednak ze względu na spójność zdecydowano się ją zostawić w dalszych analizach aby zastosować wszystkie indeksy dostępne w zbiorze danych.



Rysunek 12 Model zmiennej 'confirmed' od zmiennych z kategorii 'Government_response'

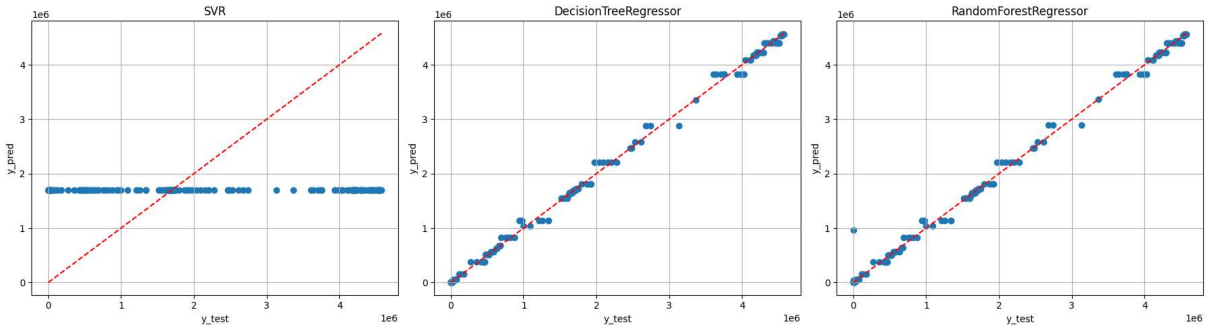
Udało się osiągnąć dość dobry model dla regresji liniowej o wartości R^2 0.76, świadczy to o tym, że na podstawie zmiennych z kategorii 'Government_response' można całkiem dobrze przewidywać ilość nowych zachorowań. Wartość R^2 nie jest na tyle wysoka aby świadczyć o przeuczeniu się modelu. Linia predykcji dość dobrze pokrywa się z punktami na wykresie.



Rysunek 13 Reszty rezyduów dla modelu zmiennej 'confirmed' od zmiennych z kategorii 'Government_response'

Rezydua tutaj nie są idealnie rozproszone wokół wartości 0 i daje się dostrzec w nich pewien wzór, jednak są one jednymi z najlepszych jakie udało się osiągnąć wśród wszystkich modeli w trakcie tej analizy.

Kolejnym krokiem w tej analizie było sprawdzenie działania algorytmów SVR, drzew regresyjnych i losowego lasu regresyjnego. Zauważono, że model stworzony w oparciu o algorytm SVR kompletnie nie radzi sobie z tymi danymi i osiągnął on ujemne R^2. Modele drzewa regresyjnego i losowego lasu regresyjnego wykazują bardzo wysoką skuteczność dla tego przypadku na podobnym poziomie, oba osiągnęły R^2 powyżej progu 0.99.



Rysunek 14 Modele zmiennej 'confirmed' od zmiennych z kategorii 'Government_response'

Bardzo zbliżone wyniki i takie samie wnioski osiągnięto analizując modele zmiennej ‘deaths’ od zmiennych z kategorii ‘Government_response’.

7. Wnioski

Modelowanie rozprzestrzenia się lub śmiertelności chorób to bardzo trudne zagadnienie wymagające nie tylko wiedzy o modelach predykcji ale również specyfiki konkretnego wirusa. Tempo rozprzestrzeniania się wirusa zależy od bardzo wielu złożonych i nieraz trudnych do opisania czynników takich jak: zjadliwość, okres inkubacji, sposób zakażenia, zachowania społeczeństwa, poziomu higieny, odporności zbiorowej populacji, gęstości zaludnienia, klimatu, szybkości i skuteczności reakcji podjętych przez władze, reakcji społeczeństwa na zagrożenie. W zasadzie można by stwierdzić, że każda z tych zmiennych oddziałuje w pewien sposób na inną.

Modele oparte na regresji są dobrym rozwiązaniem dla tego problemu, należy jednak zwrócić uwagę na to jaki jest związek przyczynowo skutkowy między zmiennymi. Dobrym przykładem jest na przykład liczba wykonanych testów, która osiągnęła współczynnik R^2 na poziomie 0.9 będąc zmienną wyjaśniającą dla ilości potwierdzonej przypadków. Jednak to liczba testów zależy od liczby zachorowań, ponieważ testowano głównie ludzi, którzy zgłosili się do lekarza z objawami choroby, czyli w większości przypadków zachorowali już na COVID-19.

Istotnym spostrzeżeniem jakie udało się zaobserwować w trakcie jest analizy jest fakt, że zaszczepienie większości w populacji jest w stanie obniżyć śmiertelność na tę chorobę. Mimo iż w 2 fali ilość zakażeń była większa, co mogłoby świadczyć o nieskuteczności szczepionek w tym zakresie należy zwrócić uwagę na to, że w trakcie 2 fali wirus ten był już zdecydowanie bardziej rozprzestrzeniony w społeczeństwie i szybciej dochodziło do zakażeń.

Warto również jeszcze raz zwrócić uwagę na pierwszy wykres tj. Rysunek 3 i końcówkę jesieni i początek zimy w roku 2023, gdzie można dostrzec małe wzniesienie na wykresie oznaczające to, że w tym okresie również miał wzrost zakażeń, jednak w porównaniu do 2 fal koronawirusa ta ilość wydaje się być marginalna a COVID-19 było już wtedy chorobą endemiczną.

Ze względu na to, że w trakcie trwania całej pandemii można wyróżnić 2 rodzaje okresów tj. fala (głównie chłodne miesiące), gdzie ilość zachorowań jest ogromna oraz okres małej ilości zachorowań (głównie ciepłe miesiące) dobrym sposobem na modelowanie przebiegu pandemii byłoby podzielenie danych na takie okresy i modelowanie ich osobno, lub też uwzględnić warunki klimatyczne w danych miesiącach.