Uczenie Maszynowe - Projekt Covid-19

Mikołaj Marchut

3 czerwca 2024

Spis treści

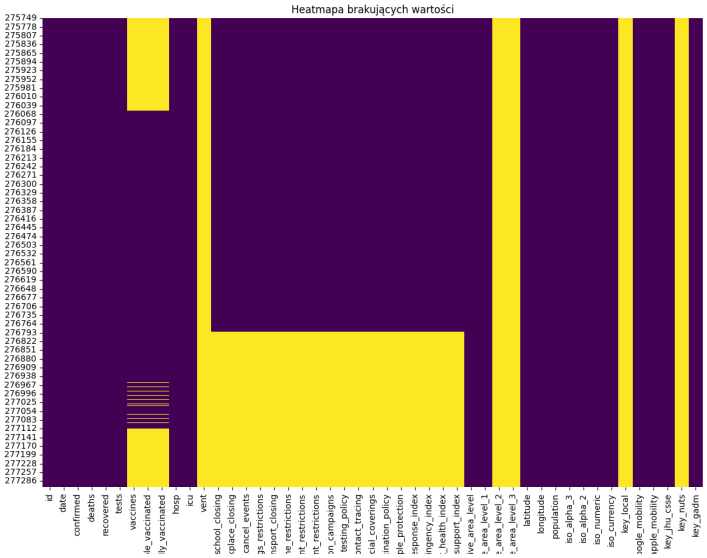
Dane Covidowe

Zapoznano się z globalnymi danymi związanymi z pandemią Covid-19 z platformy Covid-19. Platforma ta oferuje doskonały dostęp do obszernej ilości danych związanych z pandemią, co umożliwia analizę i śledzenie jej rozwoju na całym świecie. Dane te są skrupulatnie zebrane z raportów rządowych poszczególnych krajów, które zgłaszały informacje dotyczące pandemii. Platforma Covid-19 Data Hub wyróżnia się bardzo dobrą dokumentacją, która szczegółowo objaśnia każdą zmienną, co znacząco ułatwia ich interpretację i wykorzystanie w analizach. Dokumentacja zawiera opis metodologii zbierania danych, definicje zmiennych. Dane są również łatwe w użyciu dzięki możliwości instalacji dedykowanego pakietu, który umożliwia bezproblemowy import danych bezpośrednio do środowiska analitycznego. To sprawia, że proces pobierania, przetwarzania i analizy danych jest szybki i efektywny.

Wybór kraju do analizy

Wartości brakujące

Zaimportowano dane do notatnika i wybrano kraj Włochy do analizy (raport pisano po przeprowadzeniu analizy, w raporcie na początku będzie przeprowadzona analiza a następnie wybór zostanie argumentowany). Zbadano wartości brakujące dla tego kraju i stwierdzono, że Włochy posiadają w swoich raportach realatywnie mało braujących danych. Co najważniejsze nie mają brakujących danych dla liczby przypadków oraz liczby śmierci. Posiadają brakujące wartości dotyczące szczepień na początku pandemii (co jest logiczne) i na końcu. Posiadają też brakujące dane dla działań politycznych - około 1/3 obserwacji i wszystko znajduje się pod koniec pandemii i może być to spowodowane mniejszą dokładnością raportowania gdy pandemia wygasała. Pozostałe brakujące dane nie wydają się kluczowe dla przeprowadzenia algorytmu uczenia maszynowego, więc zostały pominięte.

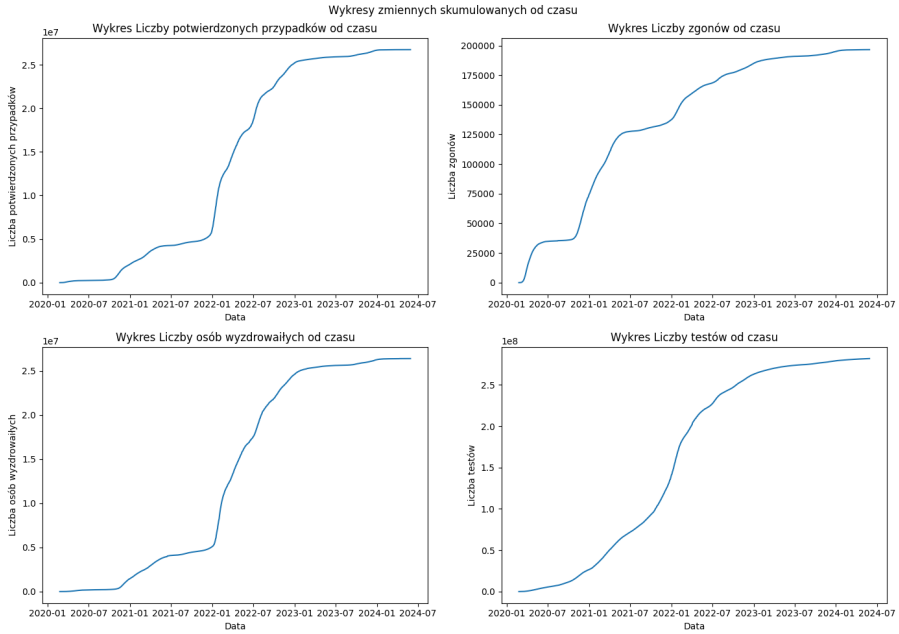


Rysunek 1: *Brakujące wartości dla Włoch*

Usunięto dane dotyczące obszarów administracyjnych, współrzędnych, kodów ISO, kluczy zewnętrznych oraz populacji, gdzyż stwierdzono, że nie potrzeba dogłębnej analizy by uznać je za zbędnę, a usunięcię ich ułatwi dalszą analizę.

Analiza zmiennych

Wykonano wykresy liczby zachorowań, liczby zgonów, liczby osób wyzdrowiałych oraz liczby testów i uznano, że nie są liniowe dla Włoch. Jednak jest to logiczne, ponieważ jak wiadomo pandemia w różnych etapach miała różną siłę, stąd nagłe wzrosty w poszczególnych okresach. Co najważniejsze (ponieważ te zmienne wydają się być dosyć istotne) dane są ciągłe. Przydatne w dalszym etapie wydaję się być rozbicie wartości skumulowanych na przypadki dzienne. Zauważono też, że zmienne te są przedstawione w sposób skumulowany.



Rysunek 2: *Wykresy zmiennych skumulowanych*

Wykonano, również wykresy rozkładu (histogramy) i wykresy boxplot dla zmiennych dotyczących liczby osób hospitalizowanych i liczby osób na oddziale intensywnej terapii i stwierdzono, że histogramy dla ilości osób hospitalizowanych oraz na odziale intensywnej terapii nie są równomierne, co pokazują też wykresy boxplot na których widać dużo wartości odstających. Te wartości mogą wnieść jednak cenne informacje do modelu.

Argumentacja wyboru

Po wstępnej analizie stwierdzono, że Włochy będą dobrym krajem do przeprowadzenia predykcji liczby zachorowań na Covid-19 i śmierci spowodowanych tym wirusem. Decyzję tą uargumentowano małą liczbą danych brakujących, ciągłością danych oraz dobrym raportowaniem informacji dotyczących pandemii przez ten kraj. Jak wiadomo z informacji kraj ten na początku pandemii borykał się z dużym problemami z tym wirusem i w późniejszym czasie podjął kroki aby zapobiec takim wydarzeniom w przyszłości.

Inżynieria cech

Wartości brakujące

Pozbyto się kolumny ‘id‘ zawierającej informację o unikalnym identyfikatorze jednostki geograficznej gdyż stwierdzono, że nie wnosi ona istotnej informacji oraz kolumny ‘vent‘ zawierającej informację o liczbie pacjentów wymagających inwazyjnej wentylacji na dzień, ponieważ posiadała ona tylko wartości brakujące. Dla danych dotyczących szczepień (liczba podanych dawek, liczba ludzi którzy otrzymali conjamniej jedną dawkę, liczbą ludzi którzy otrzymali pełną dawkę) stwierdzono, że najlepszym podjeściem będzie wypełnienie brakujących wartości używjąc ostatnich dostępnych wartości, a wartości na początku uzupełnić zerami jako, że są to dane skumulowane. Warte zauważenia jest to że te dane już wcześniej ulegały stagnacji (porównanie z heatmapą wartości brakujących). I dla liczby osób zaszczepionych istotny jest tak naprawdę fragment od początku roku 2021 do początku roku 2022. Dane dotyczące działań politycznych mających na celu ograniczenie rozprzestrzeniania się pandemii przyjmują brakujące wartości w dokładnie tym samym momencie. Na wykresie widać, że w okresie tym liczba zachorowań oraz zgonów ustabilizowała się co świadczy o wygaszaniu się pandemii wirusa. Z informacji prasowych wynika też, że od tego okresu restrykcje były już zniesione co sugerować może, że nie było to już raportowane. Zdecydowano się więc na usunięcie tych rekordów ze zbioru danych, ze względu na łagodniejszy charaketer przebiegu pandemii (można też powiedzieć zakończeniu pandemii) w tym okresie.

Rysunek 3: *Wykresy liczby zgonów i liczby zachorowań z zaznaczonym moemntem występowania brakujących danych dla działań politycznych*

Wartości skumulowane

Zamieniono kolumny z wartościami skumulowanymi na wartości odpowiadające danemu dniu. Ma to na celu uwzględnienie trendów i wzorców, co pozwala na lepsze działanie modeli liniowych na danych w takiej formie.

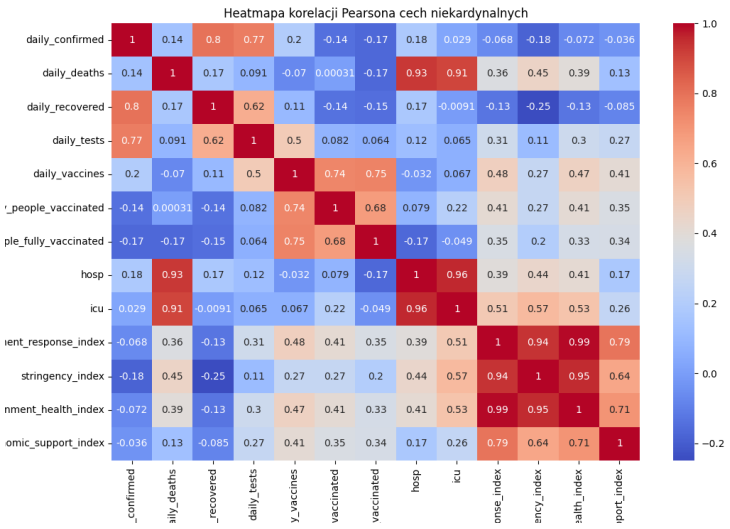
Cechy kardynalne

Cechy `information_campaigns` oraz `contact_tracing` przyjmują stałe wartości, więc stwierdzono że można je usunąć, ponieważ nie wniosą nic do modelu. Dokonano zmiany ujemnych wartości cech kardynalnych na wartości dodatnie, przekształcając je na wartości o pół mniejsze niż wartość bezwzględna pierwotnej wartości. Ujemne wartości reprezentowały wolę polityków u władzy, lecz niekoniecznie odzwierciedlały rzeczywisty stan wdrożonych restrykcji. Przyjęto więc, że restrykcje o poziom niższe są oficjalnie wprowadzone, a do tych wartości dodano pół, aby zaznaczyć, że surowsze restrykcje mogą być planowane lub rozważane. Dzięki temu zabiegowi, dane lepiej oddają potencjalne działania rządów i umożliwiają bardziej precyzyjną analizę wpływu polityk na rozwój sytuacji pandemicznej.

Analiza korelacji

Zmienne niekardynalne

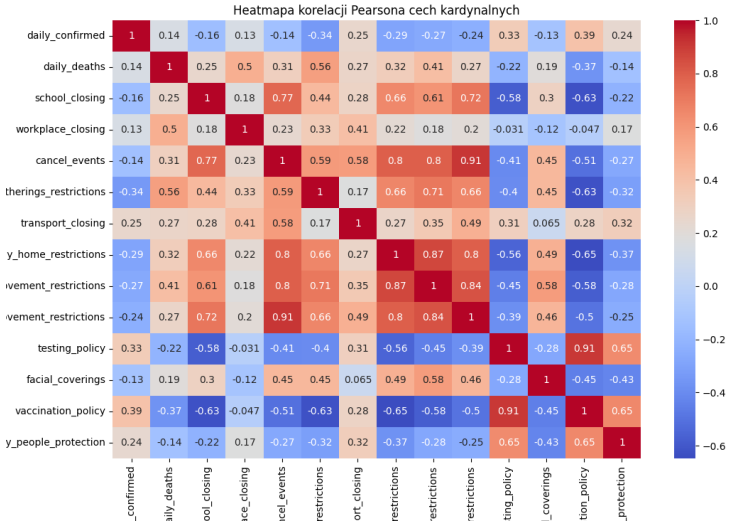
Przeprowadzono analizę korelacji liniowej Pearsona zmiennych niekardynalnych i zaobserwowano następujące wnioski. Liczba dziennych przypadków zachorowań jest silnie dodatnio skorelowana z dzienną liczbą testów oraz z dzienną liczbą osób wyzdrowiałych. W przypadku tej drugiej zmiennej to raczej liczba przypadków ma wpływ na liczbę osób wyzdrowiałych, natomiast w przypadku testów można przypuszczać, że im większa liczba testów, tym większa liczba zachorowań. W przypadku dziennej liczby zgonów, mają one bardzo silną dodatnią korelację z liczbą pacjentów hospitalizowanych oraz liczbą pacjentów na oddziale intensywnej terapii w danym dniu. Dane dotyczące wszystkich trzech cech powiązanych z testami są silnie dodatnio skorelowane ze sobą. Cechy liczbowe dotyczące działań politycznych są bardzo silnie skorelowane między sobą, najprawdopodobniej są obliczane na podstawie podobnych zmiennych lub są od siebie zależne.



Rysunek 4: *Korelacja Pearsona zmiennych niekardynalnych*

Zmienne kardynalne

W przypadku korelacji cech kardynalnych z liczbą zachorowań oraz liczbą zgonów nie zaobserowowano wniosków wnoszących wartościową informację. Umiarkowana korelacja jest pomiędzy zamykaniem miejsc pracy i restrykcjami dotyczącymi spotykania się z ilościa zgonów na covid co raczej nie jest porządanym efektem. Do tego zauważono słabą ujemną korelację pomiędzy restrykcjami dotyczącymi spotykania się, restrykcjami dotyczącymi zostawania w domu, restrykcjami dotyczącymi przemieszczania z ilościa zachorowań, co oznacza że gdy te restrykcje są wprowadzane to ilość zachorowań może maleć.

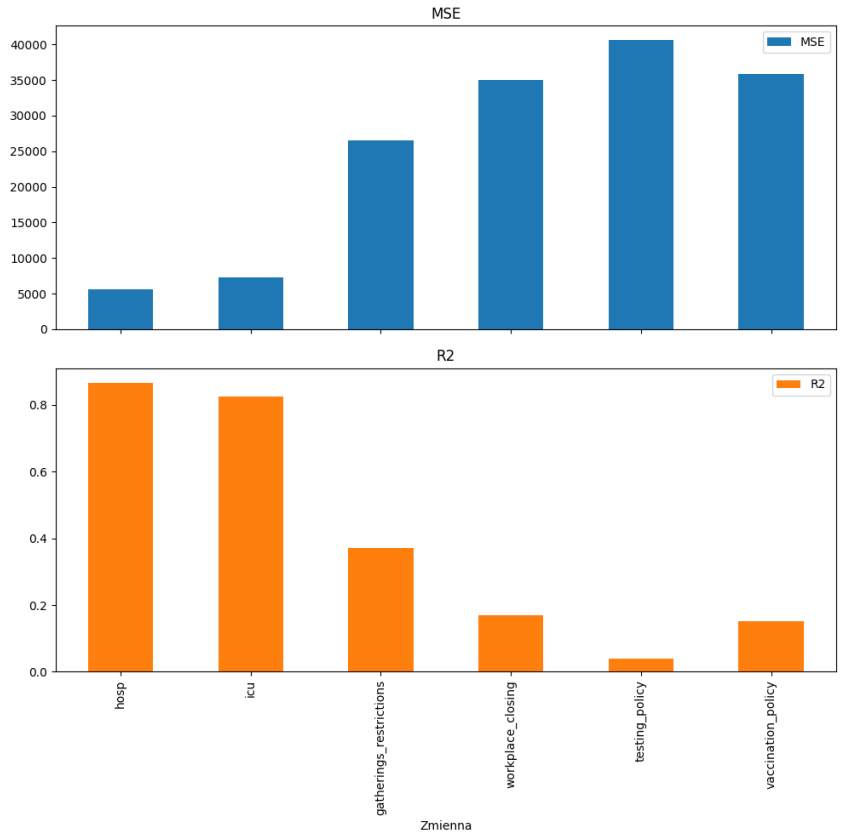


Rysunek 5: *Korelacja Pearsona zmiennych kardynalnych*

Model regresji liniowej

Liczba zgonów

Dla modelu regresji liniowej dla przewidywania liczby zgonów wybrano zmienne `hosp`, `icu`, gdyż miały bardzo silną korelację liniową ze zmienną `daily_deaths` oraz zmienne `gatherings_restrictions`, `workplace_closing` (miały one umiarkowaną korelację dodatnią ze zmienną `daily_deaths` i stwierdzono, że ciekawe będzie sprawdzenie tego mimo że korelacja nie sugeruje zależności jakiej oczekiwano, ponieważ wzrost tych zmiennych powinien powodować spadek liczby zgonów) oraz `testing_policy`, `vaccination_policy`. Najmniejszy błąd średniokwadratowy (MSE) miała zmienna `hosp`, czyli liczba pacjentów hospitalizowanych, zaraz po niej była zmienna `icu` oznaczająca liczbę pacjentów na oddziale intensywnej terapii. Pozostałe zmienne miały wysoką wartość błędu średniokwadratowego. Jeżeli chodzi o wartość współczynnika R^2, najwyższą wartość osiągnięta została dla zmiennej `hosp`, a zaraz po niej `icu`. Pozostałe zmienne miały dużo gorsze wartości. Oznacza to, że dla zmiennej `hosp` model najlepiej dopasowuje się do danych.



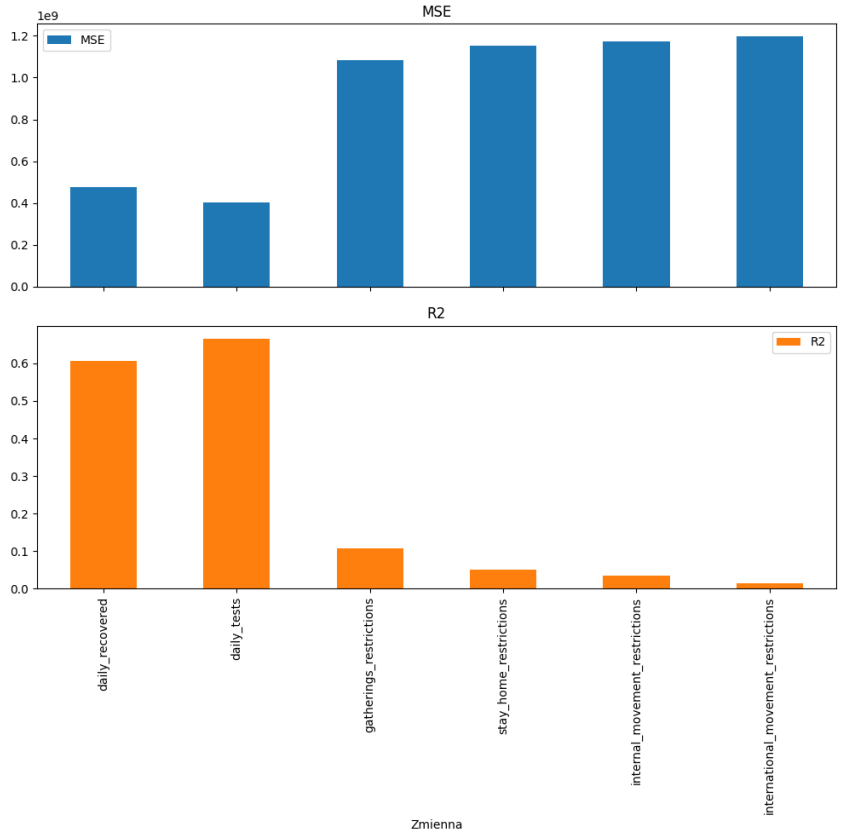
Rysunek 6: *Wykres metryk dla regresji liniowe - liczba zgonówj*

Zwizualizowano rezydua na wykresie względem przewidywanych wartości punkty na nim są równomiernie rozproszone co może sugerować, że model jest dobrze dopasowany i założenie liniowości jest spełnione. Zwizualizowano także kwantyle rozkładu rezyduów modelu z kwantylami teoretycznego rozkładu normalnego. Punkty są relatywnie blisko prostej jednak na końcach znacząco się oddalają. Na histogramie rozkładu rezyduów widać, że jego kształt odbiega od rozkładu normalnego. Przeprowadzono także test Shapiro-Wilka, gdzie wartości statystyki: 0.8888800144195557 oraz p-value 2.6240815076405966e-11 sugerują, że rezydua tego modelu nie są normalnie rozproszone co oznacza, że rozkład rezyduów nie spełnia założenia dla regresji liniowej. Wykonano model regresji wielowymiarowej na bazie zmiennych użytych wcześniej w jednoparametrowej regresji liniowej. Porównując metryki MSE i R2 stwierdzono, że lepiej wypadł model wielowymiarowy, jednak poprawa ta nie jest bardzo duża.

Liczba zachorowań

Dla modelu regresji liniowej dla przewidywania liczby zgonów wybrano zmienne `daily_recovered` i `daily_tests`, gdyż miały bardzo silną korelację liniową ze zmienną `daily_confirmed` oraz zmienne `gatherings_restrictions`, `stay_home_restrictions`, `internal_movement_restri` `international_movement_restrictions`, ponieważ miały one umiarkowaną ujemną korelację ze zmienną `daily_confirmed`, co było pożądane ze względu na to, że gdy restrykcje rosną, zakażenia powinny maleć i na odwrót. Najmniejszy błąd średniokwadratowy (MSE) miała zmienna `daily_tests`, czyli liczba przeprowadzonych testów, zaraz po niej była zmienna `daily_recovered`, oznaczająca liczbę pacjentów wyzdrowiałych. Po-

zostałe zmienne miały wysoką wartość błędu średniokwadratowego. Jeżeli chodzi o wartość współczynnika R^2, najwyższą wartość osiągnięta została dla zmiennej `daily_tests`, a zaraz po niej `daily_recovered`. Pozostałe zmienne miały dużo gorsze wartości. Oznacza to, że dla zmiennej `daily_tests` model najlepiej dopasowuje się do danych. Jednak jak wiadomo, logiczne będzie, że gdy więcej osób zostanie przetestowanych, tym więcej będzie zarażonych.



Rysunek 7: *Wykres metryk dla regresji liniowe - liczba zachorowańj*

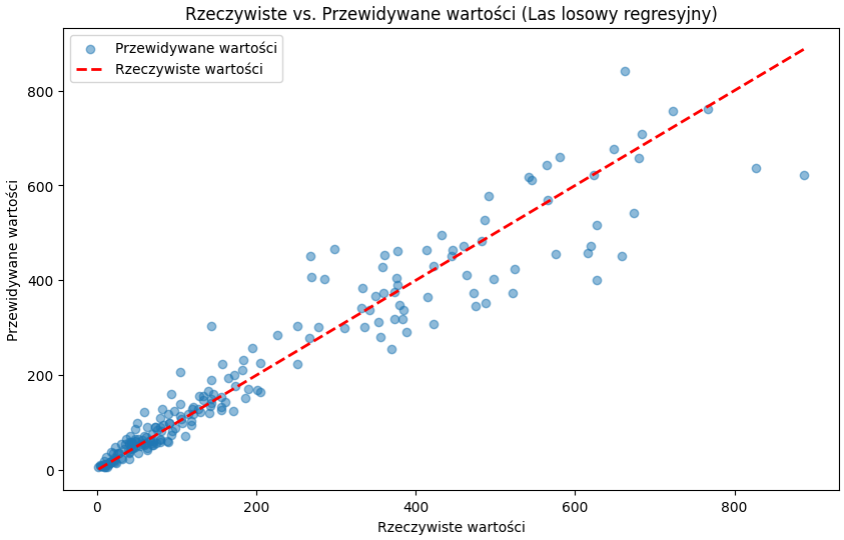
Tutaj rónież wykonano wykres rezyduów względem przewidywanych wartości, gdzie wartości były rozłożone równomiernie. Jednak ponownie wynik testu Shapiro-Wilk: statystyka: 0.9323372840881348, p-value=2.9834893666702555e-08 sugeruje, że rozkład rezyduów znacznie odbiega od normalności. W przypadku danych dotyczących liczby zachorowań wyniki metryk MSE i R2 dla regresji wielowymiarowej jest imponująco dobry co może świadczyć o przetrenowaniu danych. Dla regresji liniowej jednoparametrowej wartości tych metryk nie były za dobre. Przetrenowanie dla regresji wielowymiarowej może oznaczać, że za dużo zmiennych zostało podanych do modelu i jest on zbyt złożony.

Model regresji nieliniowej

Liczba zgonów

Dla modeli przewidujących liczbę śmierci porównano algorytmy SVR, drzew regresyjnych i losowego lasu regresyjnego. Jako zmiennych użyto wcześniej używanych do re-

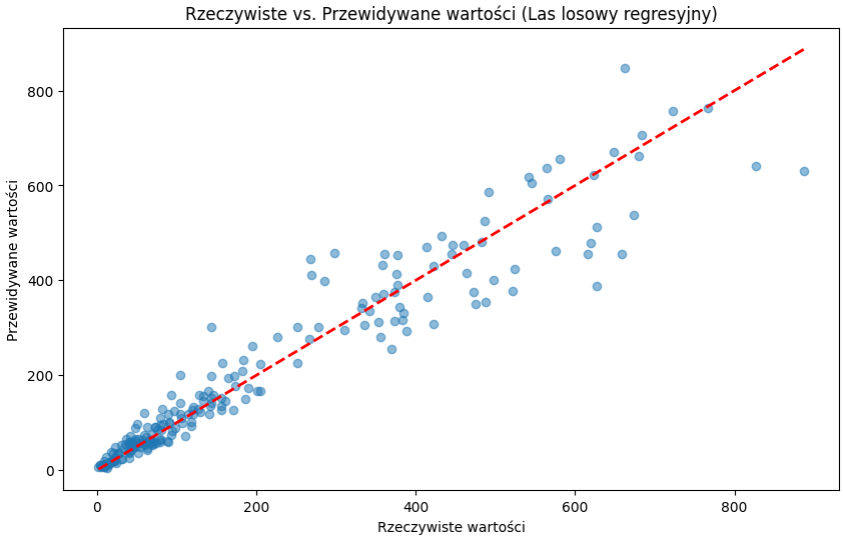
gresji liniowej zmiennych oraz `daily_confirmed` ze względu na możliwy wpływ ilości zachorowań na ilość śmierci, `government_response_index`, `stringency_index` oraz `containment_health_index`, ponieważ zmienne te miały umiarkowaną dodatnią korelację ze zmienną dotyczącą liczby śmierci. Dodano również `vaccination_policy`, `daily_people_vaccinated`, `daily_people_fully_vaccinated` ze względu na możliwy wpływ polityki szczepień oraz ilości osób zaszczepionych na liczbę zgonów. Najlepsze metryki MSE i R^2 wykazał model losowego lasu regresyjnego, co widoczne jest też na wykresie porównującym wartości rzeczywiste z przewidywanymi. Najgorzej wypadł algorytm SVR, który posiadał bardzo duży błąd średniokwadratowy.



Rysunek 8: *Wykres wartości przewidywanych w porównaniu z rzeczywistymi dla losowego lasu regresyjnego - liczba zgonów*

Liczba zachorowań

Dla tych samych algorytmów utworzono modele przewidujące liczbę zachorowań. Poza wcześniejszymi zmiennymi używanymi do przewidywania liczby zachorowań użyto dodatkowo zmiennych odnoszących się do obowiązujących restrykcji, gdyż stwierdzono, że mogą one wnieść cenną informację do modelu ze względu, że wprowadzane restrykcję miały na celu zmniejszenie liczby zachorowań. Dodano również zmienne dotyczące polityki szczepień i liczby szczepień, ponieważ szczepienia również były prowadzone w celu zmniejszenia liczby zachorowań. Ponownie najlepsze wyniki osiągnął algorytm Losowego lasu regresyjnego i ponownie najgorszy okazał się algorytm SVR.



Rysunek 9: *Wykres wartości przewidywanych w porównaniu z rzeczywistymi dla losowego lasu regresyjnego - liczba zachorowań*

Podsumowanie

Stwierdzono, że algorytmy oparte na regresji wydają się być dobrym podejściem do zadanego problemu, czyli próby przewidzenia liczby zgonów i zachorowań na Covid-19, ale nie idealnym. Przede wszystkim dane te nie są liniowe i są bardzo złożone, czego modele regresyjne mogą nie wychwycić. Po drugie dane bardzo zmieniają się w czasie (z doświadczenia wiadomo, że pandemia miała inny przebieg w okresie letnim niż w okresie zimowym). Poza tym w tym przypadku dane mają często związek z tym co działo się kilka dni wcześniej a brak jest szerszych wzorców. Modele regresji nieliniowej, jak chociażby Losowy las regresyjny poradziły sobie lepiej od modeli liniowych. Uznano, że dobrą alternatywą mogą być modele oparte o algorytmu szeregów czasowych ze względu na charakter danych. Stwierdzono, że w celu polepszenia wyników można zmienić model, chociażby jak podano wcześniej na oparty o algorytmy szeregów czasowych, można równiez przeprowadzić tuning hyperparametrów. Przydatne też mogłoby się okazać rozpatrywanie danych ze względu na okres pandemii, ponieważ jak zauważono na wykresach liniowych liczby zachorowań i zgonów parametry te miały okresy nagłego wzrostu oraz okresy stagnacji. Globalne rozpatrywanie tego problemu wydaje się niezwykle problematyczne, ze względu na bardzo dużą złożoność zagadnienia. Każdy kraj prowadził inną politykę dotyczącą obostrzeń, nie wszystkie kraje raportowały dane w sposób rzetelny, jaki i poszczególne kraje raportowały dane w różniący się sposób. Te zależności powodują, że skonstruowanie dobrego modelu który globalnie przewidywał by liczbę zgonów lub liczbę zachorowań mogłoby się okazać bardzo trudne. Jak już wcześniej wspomniano, zarówno liczba zgonów, jak i liczba zachorowań, miały nieregularny przebieg i były zależne od fazy pandemii. Zobserwowano, że liczba zgonów wykazuje silną liniową zależność od liczby pacjentów hospitalizowanych oraz osób przebywających na oddziałach intensywnej terapii. W przypadku liczby zachorowań, zaobserwowano silną zależność od liczby przeprowadzonych testów. To potwierdza, że intensyfikacja działań testowych może prowadzić do wykrycia większej liczby przypadków zachorowań, co ma kluczowe znaczenie dla skutecznej kontroli i zarządzania pandemią. Te odkrycia podkreślają ważność monito-

rowania i odpowiedniej reakcji na zmiany w dynamice pandemii, aby skutecznie zarządzać jej skutkami i ograniczyć wpływ na społeczeństwo oraz systemy opieki zdrowotnej.