AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE
WYDZIAŁ GEOLOGII, GEOFIZYKI I OCHRONY ŚRODOWISKA

KATEDRA GEOINFORMATYKI I INFORMATYKI STOSOWANEJ

Projekt COVID 19

Prognozowanie i analiza możliwych zakażeń i zgonów podczas pandemii SARS-CoV-2

Autor: Maria Zalewska
Kierunek studiów: Inżynieria i Analiza Danych

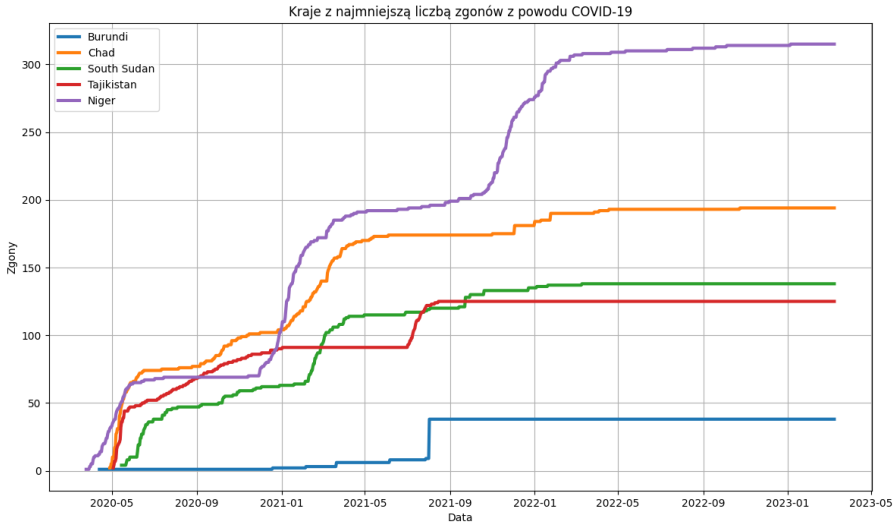
Kraków, 2024

Spis treści

1 Dane

Celem projektu będzie analiza danych dotyczących epidemii COVID-19 oraz próba wykonania predykcji liczby zachorowań i zgonów zarówno na skalę globalną, jak i lokalną (dla wybranego kraju). Dane posiadają zmienne identyfikacyjne (data i kod kraju), zmienne epidemiologiczne dotyczące między innymi zgonów, wyzdrowień, wykonanych testów czy wykonanych szczepień. Dodatkowymi informacjami są środki, jakimi państwa starały się zapobiec rozprzestrzenianiu się pandemii, zapisane jako flagi. W zbiorze znajdują się zarówno współrzędne obszaru, z którego brane były dane, jak i różnorodne kody ISO oraz identyfikatory do identyfikacji obszaru administracyjnego. Ostatni typ zmiennych został w większości usunięty z bazy danych dla większej przejrzystości zmiennych związanych bezpośrednio z chorobą.

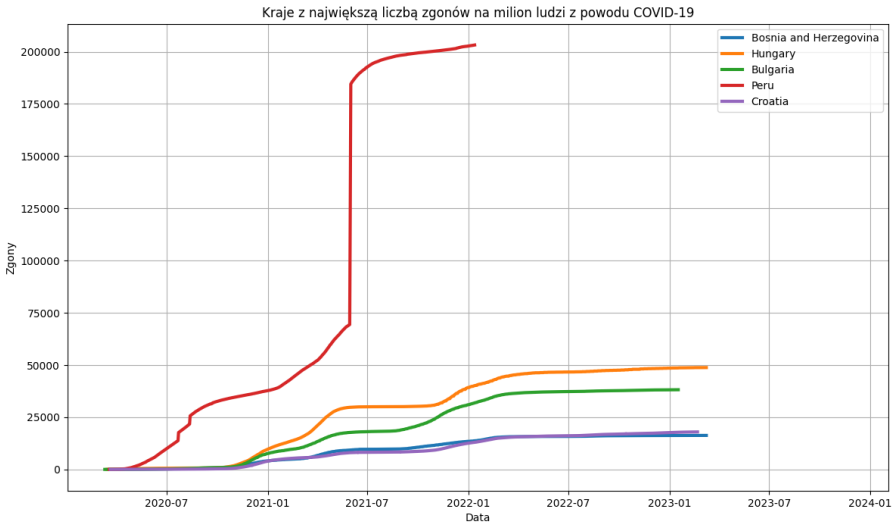
Początkowym krokiem będzie sprawdzenie jak pandemia rozprzestrzeniała się w różnych krajach i które z nich radziły sobie z nią najlepiej, bazując na zdobytych informacjach i źródłach dostarczonych przez te państwa. Poniżej przedstawione zostały wykresy zgonów w okresie czasu, a obok nich znajdują się tabele z wartościami ostatecznej liczby zgonów w kraju na milion mieszkańców.



Rysunek 1: Kraje z najmniejszą liczbą zgonów

Burundi	3.4
Czad	12.5
Sudan Południowy	12.6
Tadżykistan	13.7
Niger	14.0

Rysunek 2: Najmniejsze wartości zgonów na milion osób



Rysunek 3: Kraje z największą liczbą zgonów

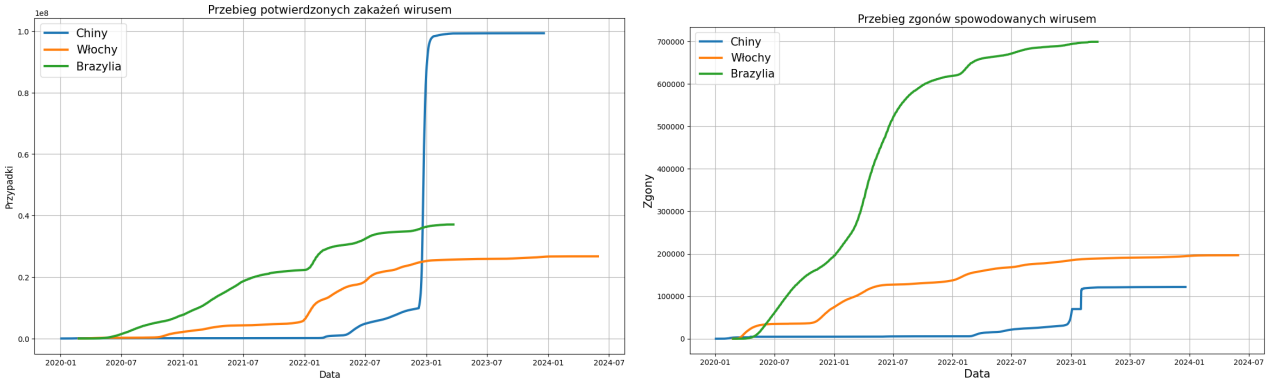
Bośnia i Hercegowina	4897.8
Węgry	4988.2
Bułgaria	5430.6
Peru	6351.9
Chorwacja	8955.8

Rysunek 4: Największe wartości zgonów na milion osób

Z początkowej analizy można wywnioskować, że zebrane dane nie zawsze będą odzwierciedlały rzeczywistość. Przykład oczywistego zakłamania liczby zgonów ukazuje jednodniowy skok wartości dla Peru z około siedemdziesięciu tysięcy do aż stu osiemdziesięciu tysięcy. Porównując przebieg linii dla innych państw i brak tak gwałtownej zmiany, wniosek o zatajaniu informacji przez peruwiański rząd jest niepodważalny, co oznacza, że inne państwa mogą posiadać podobne problemy w bazie danych.

Następnym krokiem w analizie jest wybranie kraju, na podstawie którego dokonana będzie predykcja lokalna. Z tego powodu sprawdzone zostały braki danych w kolumnach dotyczących potwierdzonych przypadków choroby i związanych z nią zgonów. Pośród państw, które nie miały braków w tych kolumnach, były Litwa, Mariany Północne, Chiny, Brazylia, Włochy, Republika Czeska, Costa Atlantica. Spomiędzy siedmiu do dalszej analizy wybrane zostały trzy, ze względu na ich istotność na arenie międzynarodowej. Dla Brazylii, Włoch i Chin stworzone zostały wykresy przebiegu pandemii – liczby potwierdzonych zakażeń oraz liczby zgonów od czasu.

Na poniższych wykresach zauważyć można, podobnie jak w przypadku danych z Peru, że Chiny posiadają skokowy przebieg zarówno w liczebności zakażeń, jak i zgonów. Może świadczyć to o ukrywaniu rzeczywistej sytuacji w państwie, co następnie wpłynęłoby negatywnie na wyniki kolejnych kroków predykcji. Do dalszych analiz wykorzystana zostanie baza danych zawierająca informacje na temat Brazylii, ze względu na najmniejszą liczebność wartości brakujących w tym zbiorze.



Rysunek 5: Wykresy przebiegu pandemii w Brazylii, Chinach i we Włoszech

2 Analiza lokalna dla Brazylii

2.1 Wartości brakujące

Zestaw danych dla Brazylii posiada całkowity brak informacji w kolumnach *hosp* (liczba hospitalizowanych pacjentów), *icu* (liczba hospitalizowanych pacjentów w intensywnej terapii), *vent* (liczba pacjentów wymagających inwazyjnej wentylacji). Są to jednak dane, które nie będą opisywały rozwoju pandemii w państwie, a jedynie jej skutki zdrowotne dla mieszkańców. Zauważono równocześnie, że w całym zbiorze danych globalnych kolumny te posiadają bardzo mało danych konkretnych, co może świadczyć o tym, że większość krajów nie podawała do informacji publicznej tych informacji, a Brazylia nie jest wyjątkiem i braki te nie wpłyną znacząco na predykcję. Braki około 30% danych występują również w kolumnach *vaccines*, *people_ vaccinated* oraz *people_ fully_ vaccinated*, które dostarczają informacji o szczepieniach przeciwko wirusowi. Szybka analiza globalnych danych pozwoliła na wniosek, że braki te są związane z niedostępnością szczepionek na rynku do początku roku 2021. Wartości brakujące w tych kolumnach zostaną zamienione na wartości zero, tak samo jak braki w kolumnie *recovered*, znajdujące się na początku pandemii oraz *tests*, które pomimo bardzo szybkiego wzrostu (sto osiemdziesiąt wykonanych testów w jeden dzień) mogą być spowodowane brakiem testów w sklepach przez pierwsze dni bądź brakiem zainteresowania obywateli państwa do zgłaszania wykonanych testów rządowi ze względu na brak takich wymogów.

2.2 Regresja liniowa

Za pomocą pętli wybrane zostały różne zmienne mogące mieć wpływ na zmienną objaśnianą, jaką były *deaths* oraz *confirmed*. Dla tych zmiennych wykonane zostały modele regresji liniowej, co miało skutkować znalezieniem zależności pomiędzy badanymi zmiennymi. Efekty tej operacji wraz z wybranymi metrykami znajdują się w poniższych tabelach. Wyniki posortowane są pod względem jak najmniejszej wartości dla błędu średniokwadratowego i jak największej dla współczynnika determinacji R-kwadrat.

Warto zauważyć, że w przypadku, gdy zmienną objaśnianą jest liczebność potwierdzonych przypadków, możliwości zmiennych objaśniających jest więcej niż w przypadku liczby zgonów. Dzieje się tak, ponieważ zmienne epidemiologiczne posiadają błędną korelację z liczebnością zgonów, gdy w rzeczywistości współczynniki regresji liniowej były równe zero. Świadomym wyborem było usunięcie tych zmiennych w przypadku liczby zgonów. Oczywistym wnioskiem jest na przykład, że im większa liczba testów zostanie zrobiona na osobach zmarłych, narażonych przed śmiercią na działanie wirusa na przykład w szpitalu, tym większej ilości osób zarażonych, lecz umierających z innego powodu, również ich zgon zostanie uznany na poczet pandemii.

Wartości MSE w obu przypadkach są wysokie, co sugeruje, że model nie będzie dokładnie przewidywał wartości zmiennej objaśnianej na podstawie takich wartości objaśniających. Ponadto, wartość R^2 dla większości zmiennych jest wysoka, co za to sugeruje, że zmienne objaśniające wyjaśniają zmiany w liczbach zgonów oraz liczbach potwierdzonych zakażeń, jednak wyniki mogą być błędne przez inne czynniki i zmienne, które nie zostały uwzględnione w modelach. Warto również zwrócić uwagę na typ danych, które są używane w modelach regresji, gdy większość z nich jest zmiennymi kategorycznymi osiągającymi wartości od zera do maksymalnie pięciu. Może być to trudne do interpretacji i może nie przynieść wartościowych wyników.

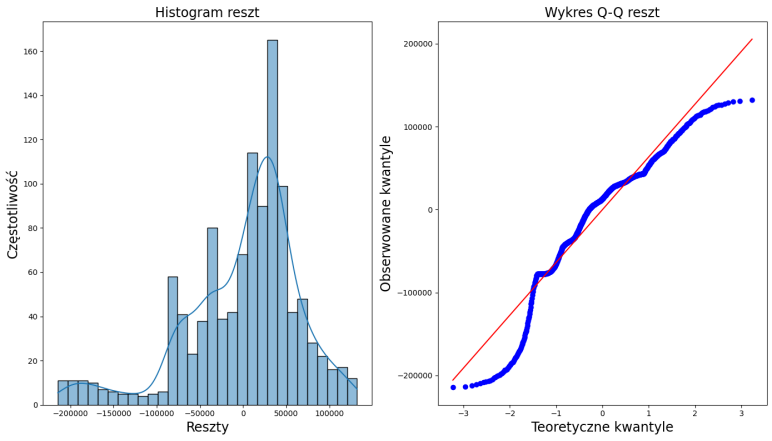
Rysunek 6: Zbiorcze wyniki modeli regresji liniowej, gdzie liczebność zgonów jest zmienną objaśnianą.

Zmienna	Współczynnik	Odcięta	MSE	R^2
vaccination_policy	115839.2	77257.4	4314461041.9	0.9
contact_tracing	-229259.3	568281.6	23136894020.3	0.6
school_closing	-174218.5	742445.9	23733274118.4	0.6
elderly_people_protection	-399397.1	631482.3	25068134912.4	0.6
transport_closing	-239314.6	632790.9	27725728203.1	0.6
gatherings_restrictions	-114607.1	679904.6	30364331340.9	0.5
testing_policy	204218.5	-6374.7	32686008250.2	0.5
stay_home_restrictions	-165426.0	626592.6	36937451371.7	0.4
workplace_closing	-138209.0	703638.1	41855922378.9	0.4
facial_coverings	-151785.3	906808.0	43707924575.9	0.3
cancel_events	-137469.1	652912.3	51787948933.8	0.2

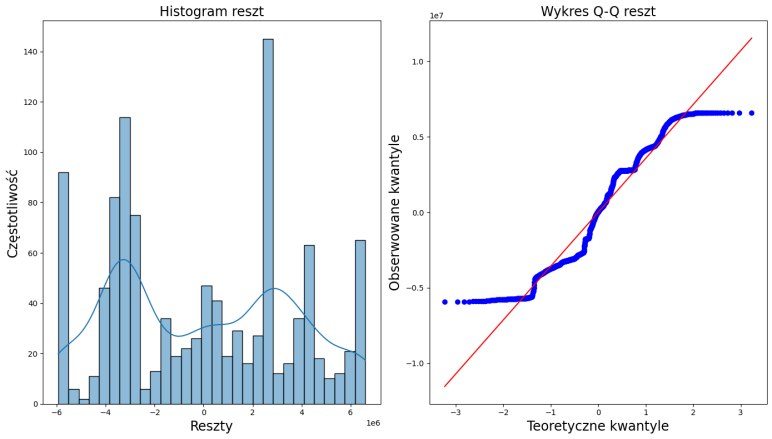
Rysunek 7: Zbiorcze wyniki modeli regresji liniowej, gdzie liczebność zachorowań jest zmienną objaśnianą.

Zmienna	Współczynnik	Odcięta	MSE	R^2
vaccines	0.1	5912221.4	11293243652131.5	0.9
tests	0.5	-2775523.6	13443814368491.6	0.9
people_vaccinated	0.1	4751378.6	14431368124701.2	0.9
people_fully_vaccinated	0.2	6310508.9	16407673029688.8	0.9
vaccination_policy	5572926.9	1713758.2	22406052344771.4	0.9
school_closing	-9667073.4	35922516.1	35946391436363.7	0.8
gatherings_restrictions	-6572917.7	32893046.6	49119027879068.8	0.7
elderly_people_protection	-20639145.7	29048064.3	56294899752885.6	0.7
transport_closing	-12786399.0	29447886.7	56556255904035.6	0.7
stay_home_restrictions	-9529118.1	29881648.3	69932614585422.3	0.6
contact_tracing	-10691124.3	25152124.9	71797799764519.5	0.6
testing_policy	10678262.0	-4189202.1	74591136315281.1	0.5
workplace_closing	-8270983.7	34902937.4	80211646646962.7	0.5
facial_coverings	-9032736.8	46906609.5	87685789711143.5	0.5
cancel_events	-9155408.4	33283105.3	104654096190361.4	0.4

Jako najlepsza zmienna objaśniająca *deaths* wybrana została zmienna *vaccination_policy*. Jeśli reszty modelu mają rozkład normalny, histogramy na wykresach 8 i 9 powinny przypominać krzywą dzwonową, a na wykresie Q-Q punkty powinny układać się wzdłuż linii prostej. Dla zmiennej objaśniającej rozkład rezyduów jest bliski normalnemu. Widoczne są jednak odchylenia, które nie pozwalają uznać tego modelu za poprawny, co zgadza się z poprzednimi wątpliwościami co do metryk modelu. Dla zmiennej objaśnianej *confirmed* użyta została zmienna objaśniająca *tests*, ponieważ jest ona logicznie potrzebna do predykcji zakażeń, lecz nie jest ona również dobrym wyborem. Założenia regresji liniowej nie są spełnione.



Rysunek 8: Rozkład rezyduów dla zmiennej objaśnianej *death* oraz zmiennej objaśniającej *vaccination_policy*



Rysunek 9: Rozkład rezyduów dla zmiennej objaśnianej *confirmed* oraz zmiennej objaśniającej *tests*

2.3 Regresja wielowymiarowa

Do stworzenia regresji wielowymiarowej potrzebne było stworzenie macierzy korelacji pomiędzy zmiennymi oraz wykorzystanie współczynnika VIF. Problemem okazała się zbyt duża ilość zmiennych przedstawionych w taki sposób (dane sumaryczne bądź kategoryczne), że większość zmiennych posiadała dużą korelację Spearmana ze zmiennymi objaśnianymi, przez co modele regresji mogą być nadmiernie dopasowane i trudno zrozumiałe, które dokładnie dane są ważne dla modelu.

Rysunek 10: Prównanie metryk dla regresji liniowej i regresji wielowymiarowej

Dane dla modelu regresji liniowej		
Zmienna objaśniana	*deaths*	*confirmed*
MSE	4314461041.9	11293243652131.5
R^2	0.9	0.9
Dane dla modelu regresji wielowymiarowej		
Zmienna objaśniana	*deaths*	*confirmed*
MSE	634488831.7	2012631966552.1
R^2	0.9	0.9

Do regresji wielowymiarowej dla zmiennej objaśnianej *deaths* użyte zostały zmienne *tests, vaccination_policy, elderly_people_protection*, a dla danych *confirmed* kolumny *tests, vaccination_policy, gatherings_restrictions, facial_coverings*. Chociaż metryki błędu średniokwadratowego są nadal zbyt duże, aby uznać model za odpowiedni, jego wartości w metodzie regresji wielowymiarowej są mniejsze nawet o 10 razy.

2.4 Algorytm SVR

Metoda *Support Vector Regression (SVR)* polega na znalezieniu funkcji, która ma odchylenie od rzeczywistych wartości mieszczące się w granicach wartości ϵ dla wszystkich punktów w zbiorze treningowym. Do tego zadania metoda ta używa funkcji kosztów, ignorując błędy, które są poniżej ϵ, co uniemożliwia powstanie nadmiernego dopasowania. Co więcej, SVR może używać *kernel trick*, aby modelować nieliniowe relacje między zmiennymi.

W przypadku danych dla Brazylii do sprawdzenia regresji dla zmiennej *deaths* oraz *confirmed* użyte zostały zmienne objaśniające *recovered, tests, vaccination_policy, people_vaccinated*, aby uniknąć korzystania ze zmiennych kategorycznych i sprawdzić, czy wtedy model będzie lepiej dopasowany, stosując najpierw normalizację dla tych danych. W wyniku zwrócone wartości reprezentują się w następujący sposób:

Rysunek 11: Wyniki algorytmu SVR

Zmienna objaśniana	*deaths*	*confirmed*
MSE	92422573113.9	167116693594378.3
R^2	-0.42	-0.06

Wyniki nie są jednak poprawne. Ujemne R^2 znaczy o całkowitym błędzie modelu i braku jego użyteczności.

2.5 Algorytm drzewa regresyjnego

Algorytm drzewa regresyjnego polega na tworzeniu struktury drzewa decyzyjnego poprzez dzielenie zbioru danych ze względu na wartości zmiennych. Każdy węzeł drzewa odpowiada testowi na wartości cechy, a każda gałąź wynikowi tego testu. Testy te tworzone są w taki sposób, aby zminimalizować sumę kwadratów błędów.

Jako zmienne objaśniające użyte zostały te same zmienne, co dla metody SVR, jednak w algorytmie tym nie ma potrzeby normalizacji wartości. Metoda ta posiada lepsze metryki niż metoda SVR oraz lepsze niż prosta regresja liniowa czy nawet wielowymiarowa.

Rysunek 12: Wyniki algorytmu drzewa regresyjnego

Zmienna objaśniana	*deaths*	*confirmed*
MSE	1133846.3	3564770715.5
R^2	0.99	0.99

2.6 Algorytm lasu losowego

Metoda lasu losowego jest rozszerzeniem algorytmu drzewa regresyjnego. Jest to wiele drzew decyzyjnych, połączonych w taki sposób, aby uzyskać bardziej stabilne i dokładne przewidywania. Każde drzewo jest trenowane na innym podzbiorze danych wybranym losowo. Końcowy wynik predykcji jest średnią przewidywań ze wszystkich drzew.

Rysunek 13: Wyniki algorytmu losowego lasu regresyjnego

Zmienna objaśniana	*deaths*	*confirmed*
MSE	492176.9	2596407900.7
R^2	0.99	0.99

3 Wnioski

W przeprowadzonej analizie zastosowano różne algorytmy predykcyjne oparte na regresji, takie jak regresja liniowa, regresja wielowymiarowa, SVR, drzewa regresji oraz las losowy regresji. Każdy z tych modeli został zastosowany do przewidywania liczby zgonów (*deaths*) oraz liczby potwierdzonych przypadków (*confirmed*) COVID-19. Z poniższego podsumowania wynika, że najlepiej z tym problemem poradził sobie algorytm regresji lasu losowego, osiągając wartość MSE dla zmiennej objaśnianej *deaths* około 8696 razy mniejszą niż dla zwykłej regresji. Wszystkie modele (poza modelem SVR) posiadają duży współczynnik determinacji, co może być spowodowane błędną korelacją związaną z zapisem danych. Sposób podania wartości jako wartości sumarycznych może prowadzić do nieprawdziwych zależności, na przykład wraz ze wzrostem szczepień rosły przypadki zakażeń ze względu na upływ czasu, a niekoniecznie ze względu na pomoc szczepionek w poszerzaniu choroby. Aby polepszyć wyniki, można rozważyć zastosowanie bardziej zaawansowanych technik modelowania, takich jak uczenie głębokie, które może lepiej radzić sobie ze złożonością danych oraz nieliniowymi zależnościami między nimi.

Wnioski na temat przyrostów zgonów oraz przypadków zakażeń zdają się zależeć od ilości szczepionek czy wykonanych testów. Problemem okazało się porównanie tych zmiennych ze środkami ostrożności podjętymi przez rząd. Prawdopodobnie dalsza analiza mogłaby pomóc w lepszym zrozumieniu tych zależności, lecz dla tak przygotowanych danych skomplikowane okazało się wywnioskowanie tego z liczb.

Rysunek 14: Porównanie różnych rodzajów regresji

	deaths	*confirmed*	zmienne objaśniające
Regresja liniowa			*deaths: vaccination_policy* *confirmed:tests*
MSE	4314461041.9	13443814368491.6	
R^2	0.9	0.9	
Regresja wielowymiarowa			*deaths: tests, vaccination_policy, elderly_people_protection* *confirmed: tests, vaccination_policy,* *gatherings_restrictions, facial_covering*
MSE	634488831.7	2012631966552.1	
R^2	0.9	0.9	
Algorytm SVR			*deaths: recovered, tests, vaccination_policy, people_vaccinated* *confirmed: recovered, tests, vaccination_policy,* *people_vaccinated*
MSE	92422573113.9	167116693594378.3	
R^2	-0.4	-0.1	
Algorytm drzew regresji			*deaths: recovered, tests, vaccination_policy, people_vaccinated* *confirmed: recovered, tests, vaccination_policy,* *people_vaccinated*
MSE	1133846.3	3564770715.5	
R^2	0.9	0.9	
Algorytm lasu losowego regresji			*deaths: recovered, tests, vaccination_policy, people_vaccinated* *confirmed: recovered, tests, vaccination_policy,* *people_vaccinated*
MSE	492176.9	2596407900.7	
R^2	0.9	0.9	

Ze względu na dużą ilość brakujących wartości oraz różne sposoby dokumnetacji przebiegu pandemii dla różnych państw analiza globalna byłaby obarczona większymi błędami. Jednym z rozwiązań tego problemu mogłoby się okazać pogrupowanie danych pod względem państw, których sposób dostarczania informacji był podobny, bądź chociaż pogrupowanie ze względów geograficznych, aby sprawdzić, w jakich kierunkach rozchodziła się fala zakażeń i które z państw sąsiadujących ze sobą radziły sobie lepiej.