Analiza dynamiki wybranych chorób zakaźnych w Europie na podstawie danych ECDC

Bartosz Wolny, Kacper Kortas, Paweł Kamiński, Sebastian Kaca

Uniwersytet Pomorski, Instytut Nauk Ścisłych i Technicznych

Email: bartosz.wolny@office.upsl.edu.pl, kacper.kortas@office.upsl.edu.pl, pawel.kaminski@office.upsl.edu.pl, sebastian.kaca@office.upsl.edu.pl

Abstrakt

Artykuł przedstawia analizę dynamiki czterech wybranych chorób zakaźnych: odry, krztuśca, salmonellozy oraz gruźlicy w krajach europejskich. Badanie oparto na rocznych danych epidemiologicznych pochodzących z Europejskiego Centrum ds. Zapobiegania i Kontroli Chorób (ECDC). Wykorzystano metody eksploracyjnej analizy danych oraz elementy modelowania statystycznego w celu identyfikacji trendów czasowych, zróżnicowania geograficznego oraz potencjalnych czynników wpływających na obserwowane zmiany zapadalności. Uzyskane wyniki zestawiono krytycznie z doniesieniami literaturowymi, wskazując zarówno spójności, jak i rozbieżności. Artykuł omawia również ograniczenia zastosowanego podejścia oraz możliwe kierunki dalszych badań.

Słowa kluczowe: epidemiologia cyfrowa, ECDC, uczenie maszynowe, choroby zakaźne, regresja, predykcja.

1. Wprowadzenie

Choroby zakaźne pozostają istotnym wyzwaniem zdrowia publicznego, mimo postępu medycyny, rozwoju programów szczepień oraz poprawy warunków sanitarnych. Globalizacja,

migracje ludności oraz zmiany klimatyczne sprzyjają ponownemu pojawianiu się chorób uznawanych wcześniej za opanowane. W Europie systematyczny nadzór epidemiologiczny prowadzony przez ECDC umożliwia analizę długoterminowych trendów i porównań między krajami. Celem niniejszego artykułu jest ilościowa i jakościowa analiza dynamiki czterech chorób zakaźnych o odmiennym mechanizmie transmisji i znaczeniu epidemiologicznym: salmonellozy (choroba przenoszona drogą pokarmową), gruźlicy (przewlekła choroba bakteryjna), oraz odry i krztuśca (wysoce zakaźne choroby wirusowe i bakteryjne).
Eliminacja odry w Regionie Europejskim WHO pozostaje jednym z kluczowych celów zdrowia publicznego, jednak od 2017 roku obserwuje się ponowny wzrost liczby zachorowań, głównie na skutek luk w wyszczepialności populacji [1]. Analiza obejmuje zarówno elementy eksploracyjne, jak i modelowe, co pozwala na pełniejsze zrozumienie badanych zjawisk.

2. Dane i metody

2.1. Źródło danych

Podstawę analizy stanowiły raporty miesięczne i roczne z *ECDC Surveillance Atlas of Infectious Diseases* dla 31 krajów Europy, obejmujące różne okresy, zależne od dostępnych danych. Zmienną celu zdefiniowano jako współczynnik zapadalności na 100 000 mieszkańców. Dane uzupełniono o roczne wskaźniki z Eurostatu: PKB per capita, gęstość zaludnienia oraz poziom wyszczepialności.

2.2. Przetwarzanie danych

Dane poddano wstępnemu czyszczeniu, obejmującemu usuwanie braków, standaryzację nazw krajów oraz ujednolicenie zakresów czasowych. Następnie obliczono podstawowe statystyki opisowe oraz przygotowano szeregi czasowe dla każdej choroby.

W procesie inżynierii cech utworzono zmienne opóźnione czasowo (Lag1, Lag2) oraz średnie kroczące (MA3), aby uwzględnić autokorelacje zjawisk epidemiologicznych.

W celu rzetelnej oceny zdolności predykcyjnych modeli oraz odwzorowania rzeczywistych warunków prognozowania, zgromadzony materiał badawczy podzielono zgodne z kryterium chronologicznym. Lata wcześniejsze zdefiniowano jako zbiór treningowy, a dane z lat 2022-2024 posłużyły jako odseparowany zbiór testowy, co pozwoliło wyeliminować ryzyko wycieku informacji z przyszłości. Na etapie wstępnego przetwarzania

zaadresowano również problem nieciągłości w szeregach czasowych. Luki w danych historycznych uzupełniono metodą interpolacji liniowej, natomiast dla brakujących obserwacji w końcowych latach okresu badawczego zastosowano metodę przeniesienia ostatniej obserwacji, aby uniknąć wprowadzania sztucznych trendów.

Kluczowym elementem przygotowania danych była inżynieria cech, ukierunkowana na wydobycie istotnych zależności temporalnych i kontekstowych. Wygenerowano zestaw zmiennych opóźnionych pierwszego i drugiego rzędu (`Incidence_Lag1`, `Incidence_Lag2`) w celu uchwycenia autokorelacji zjawiska, a także zmienną średniej ruchomej z trzech lat (`Incidence_MA3`), służącą do wygładzania krótkoterminowego szumu informacyjnego. Zbiór predyktorów wzbogacono o zewnętrzne dane kontekstowe, integrując wskaźniki poziomu wyszczepialności populacji (`Vaccination_Coverage_Pct`) oraz zmienne demograficzne określające udział dzieci w wieku 0-14 lat w strukturze populacji (`Pop_Structure_0_14_Pct`).

2.3. Metody analizy

Zastosowano eksploracyjną analizę danych (EDA), obejmującą wizualizację trendów rocznych, porównania międzynarodowe oraz identyfikację anomalii (epizodów epidemicznych). Ze względu na roczny charakter danych zrezygnowano z analizy sezonowości miesięcznej na rzecz badania cykliczności wieloletniej. W ramach statystyki opisowej przeanalizowano miary tendencji centralnej (średnia, mediana) oraz rozproszenia, co pozwoliło na wykrycie silnej prawostronnej skośności rozkładu zachorowalności oraz potwierdzenie wysokiej stabilności poziomu wyszczepienia. Przeprowadzono również audyt kompletności danych, który wykazał, że braki występują wyłącznie w zmiennych dotyczących liczby przypadków i zapadalności, co wynika prawdopodobnie ze specyfiki raportowania.

2.4. Modelowanie

Zastosowano podejście porównawcze, testując modele o różnej złożoności:

1. Modele liniowe: Regresja liniowa, Ridge, Lasso.
2. Modele zespołowe: Random Forest, Gradient Boosting.
3. Szeregi czasowe: ARIMA(1,1,1) do prognozy trendów na lata 2025-2027.

Model ARIMA(1,1,1) wybrano jako standardowy punkt odniesienia w analizie szeregów czasowych. Parametry (1,1,1) oznaczają: uwzględnienie jednej poprzedniej wartości (p=1), jednokrotne różnicowanie w celu usunięcia trendu (d=1) oraz uwzględnienie jednego poprzedniego błędu prognozy (q=1). Zastosowanie tego samego modelu dla wszystkich chorób umożliwiło obiektywne porównanie prognoz.

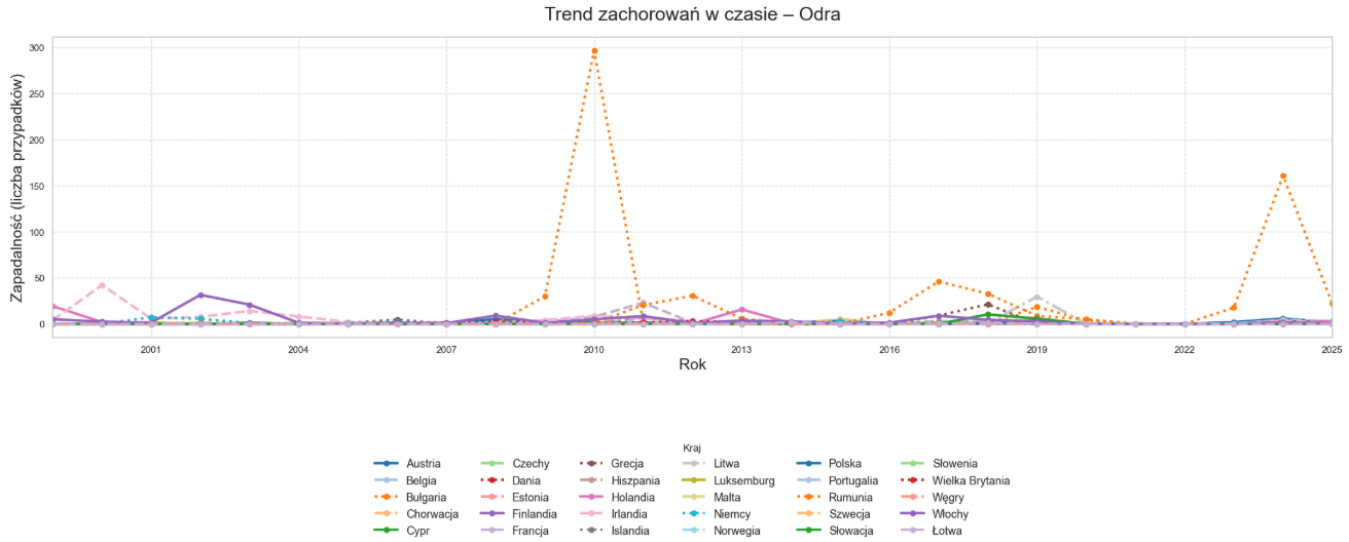
Podziału danych dokonano metodą temporal split: lata wcześniejsze stanowiły zbiór treningowy, a ostatnie 2-3 lata (zależnie od dostępności dla kraju) zbiór testowy, aby uniknąć wycieku danych. Jako metryki oceny przyjęto RMSE (Root Mean Square Error) oraz współczynnik determinacji R^2.

3. Wyniki

3.1. Analiza trendów i sezonowości

3.1.1. Odra

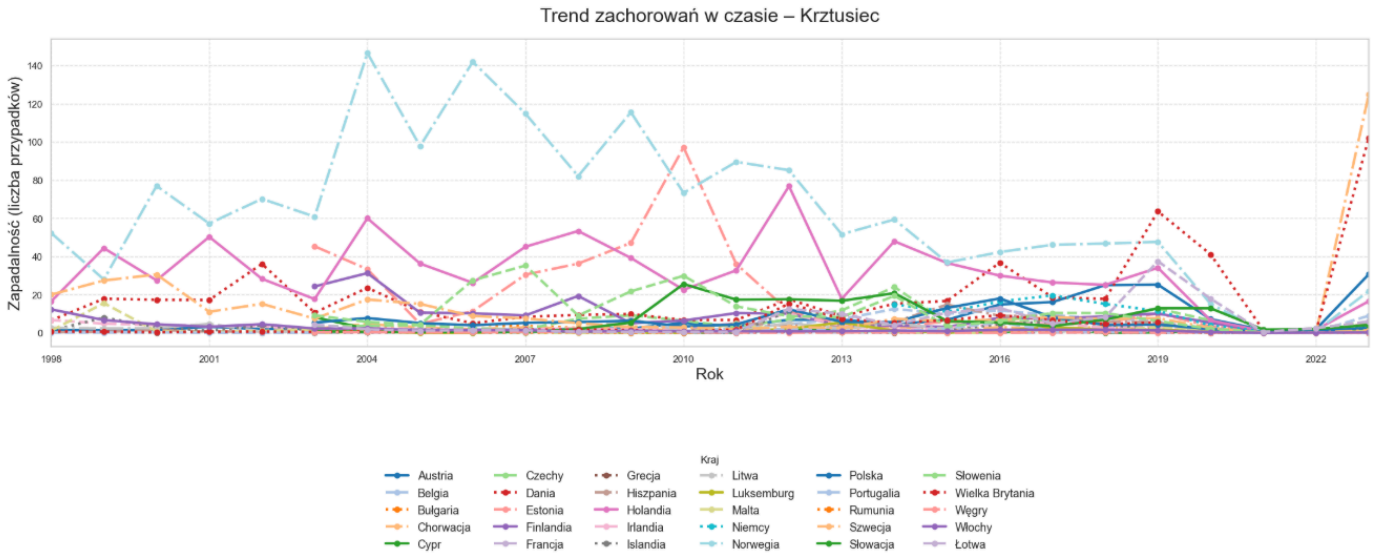
Analiza danych dotyczących odry wskazuje na wyraźne fluktuacje zapadalności w czasie, z okresami gwałtownych wzrostów w niektórych krajach. Wykres rocznych wskaźników zapadalności (Rys. 1) pokazuje epizody nawrotów choroby, co może być związane z obniżeniem wyszczepialności. Nawet niewielkie spadki poziomu szczepień mogą prowadzić do rozległych ognisk epidemicznych, ponieważ zapobieganie trwałej transmisji odry wymaga utrzymania odporności populacyjnej na poziomie około 95% [2].



Rys. 1. Trend zachorowań w czasie na odrę w krajach Europy. Źródło: opracowanie własne.

3.1.2. Krztusiec

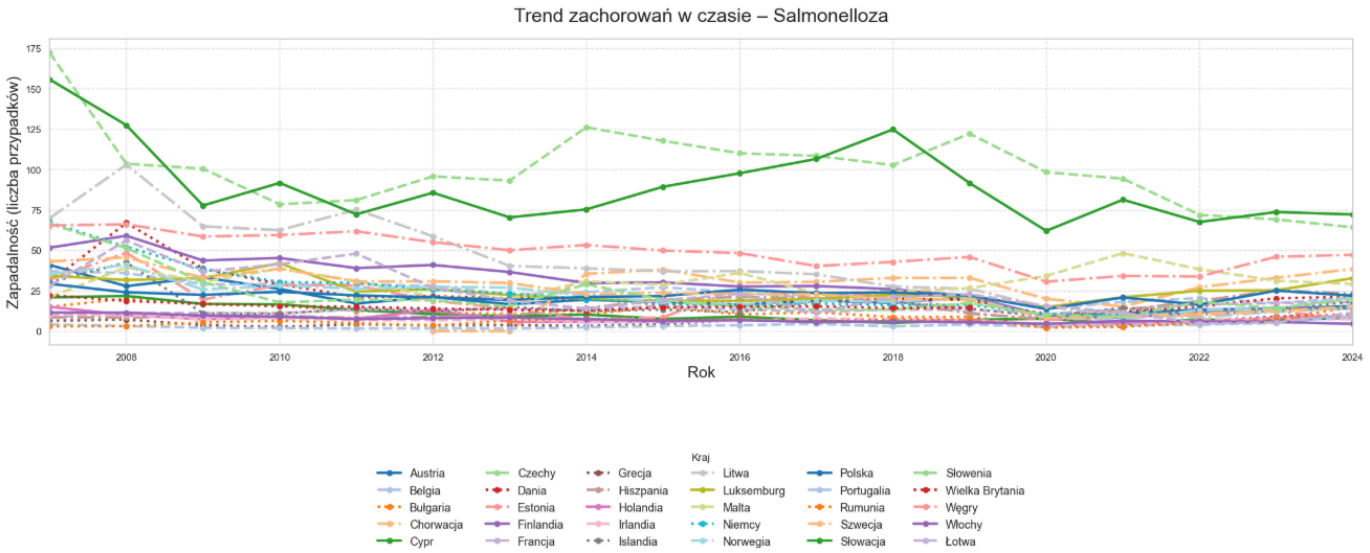
W przypadku krztuśca obserwuje się trend falowy, charakterystyczny dla chorób o cyklicznej dynamice epidemiologicznej. Rysunek 2 ilustruje różnice między krajami, sugerując wpływ lokalnych strategii szczepień oraz systemów raportowania.



Rys. 2. Trend zachorowań w czasie na krztusiec w krajach Europy. Źródło: opracowanie własne.

3.1.3. Salmonelloza

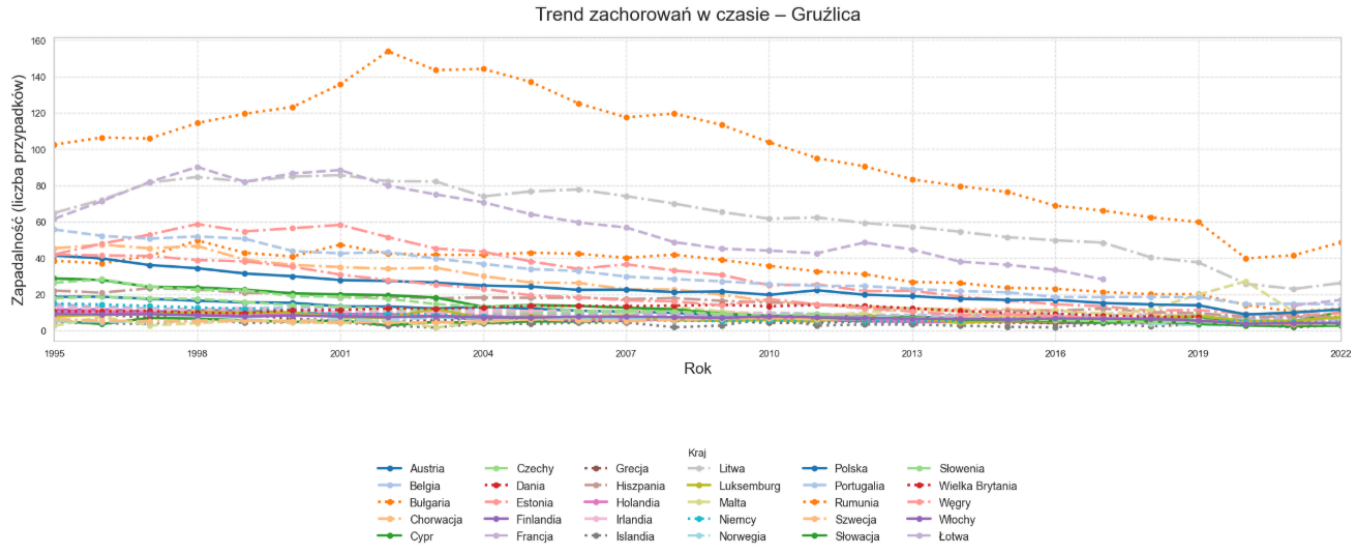
Salmonelloza wykazuje względnie stabilny, choć zróżnicowany przestrzennie poziom zapadalności. Na wykresie 3 widoczna jest tendencja spadkowa w części krajów, co może świadczyć o poprawie standardów bezpieczeństwa żywności.



Rys. 3. Trend zachorowań w czasie na salmonellozę w krajach Europy. Źródło: opracowanie własne.

3.1.4 Gruźlica

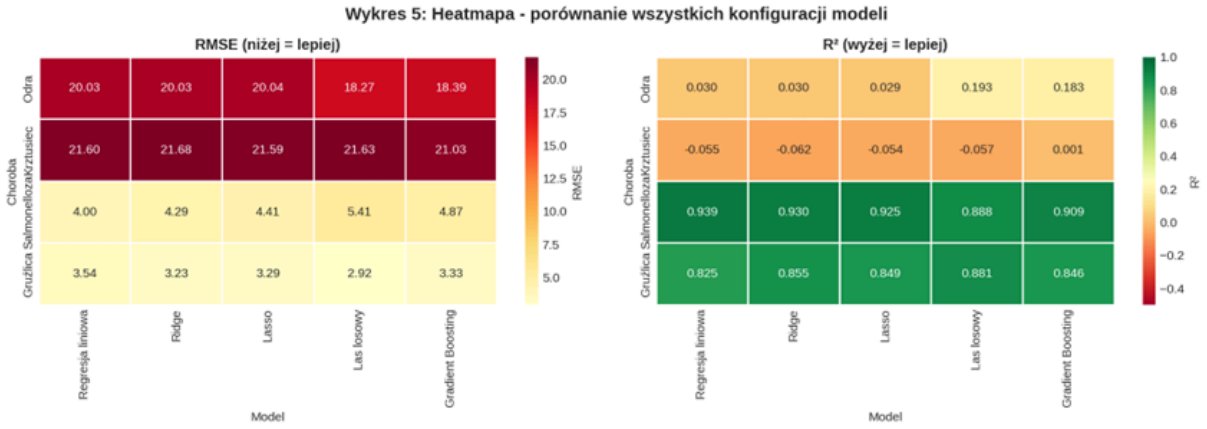
Dane dotyczące gruźlicy wskazują na długoterminowy trend spadkowy w większości krajów Europy (Rys. 4). Jednocześnie utrzymujące się różnice regionalne podkreślają znaczenie czynników społeczno-ekonomicznych.



Rys. 4. Trend zachorowań w czasie na gruźlicę w krajach Europy. Źródło: opracowanie własne.

3.2. Skuteczność modeli predykcyjnych

Wyniki walidacji (Wykres 1) wskazują na drastyczne różnice w przewidywalności poszczególnych chorób.

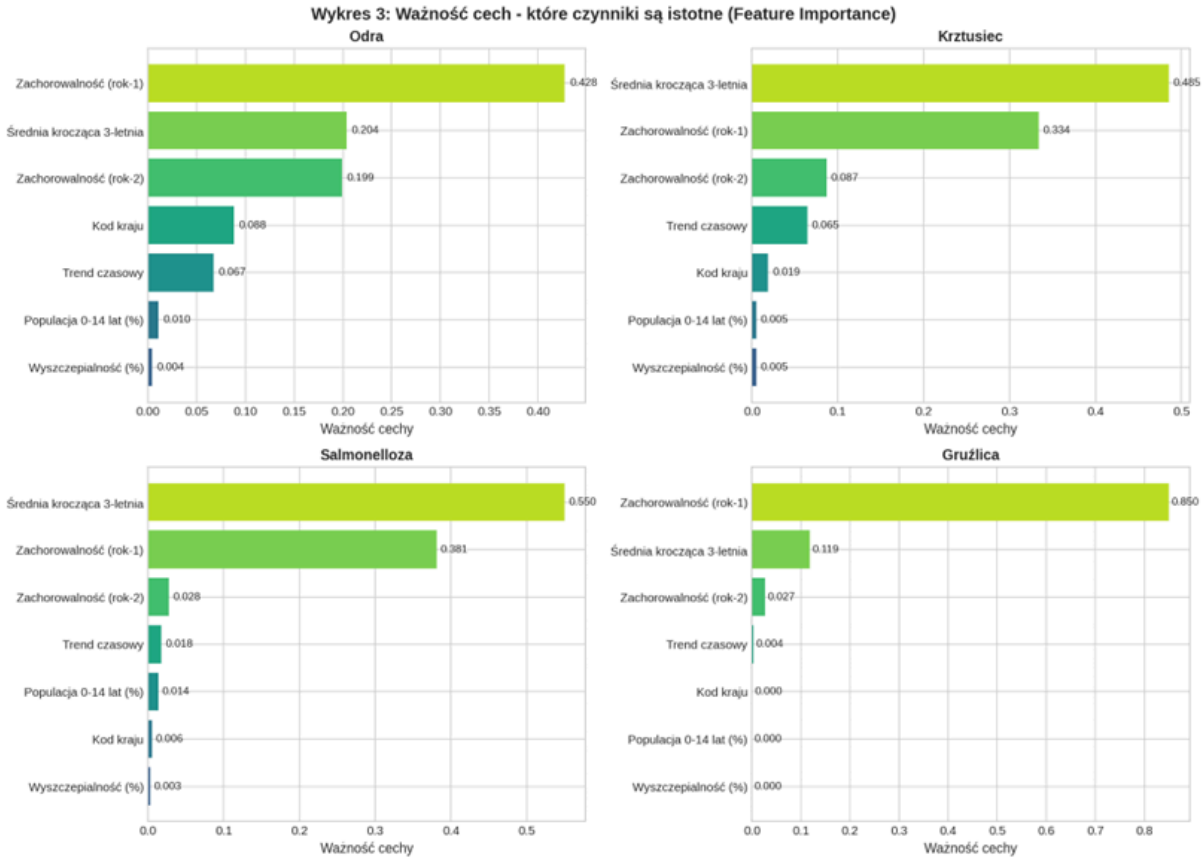


Wykres 1. Heatmapa RMSE i R² dla wszystkich kombinacji model × choroba. Źródło: opracowanie własne.

Dla salmonellozy, prosta regresja liniowa okazała się wystarczająca ($R^2 \approx 0,94$). Wynika to ze stabilności epidemiologicznej i braku nagłych interwencji zaburzających trend. Dla gruźlicy, las losowy ($R^2 \approx 0,88$) skutecznie odwzorował powolne zmiany, wychwytując interakcje między czynnikami socjoekonomicznymi a trendem spadkowym. Badając odrę, mimo zastosowania lasu losowego, uzyskano niski poziom dopasowania ($R^2 = 0,18$) oraz wysoki błąd RMSE. Wynika to z niestabilnego, wzrostowego trendu oraz silnego wpływu zdarzeń incydentalnych, takich jak ogniska epidemiczne czy zmiany poziomu wyszczepialności. Model ma trudności z generalizacją, a obecność wartości odstających znacząco obniża jakość predykcji. Silnie losowy i nieliniowy charakter transmisji odry, często objawiający się lokalnymi, gwałtownymi wybuchami epidemii, stanowi istotne wyzwanie dla klasycznych modeli prognostycznych [3]. W przypadku krztuśca żaden z zastosowanych modeli nie osiągnął zadowalających wyników ($R^2 \approx 0$), nawet przy użyciu Gradient Boostingu. Choroba charakteryzuje się wyraźnym trendem cyklicznym, prawdopodobnie związanym z wygasaniem odporności poszczepiennej i okresowymi falami zachorowań. Nieregularność i zmienność danych uniemożliwiają wiarygodne prognozowanie przy użyciu standardowych modeli regresyjnych.

3.3. Analiza ważności cech

Analiza modelu lasu losowego wykazała, że najważniejszym predyktorem dla wszystkich chorób jest `Incidence_Lag1` (zachorowalność w roku poprzednim), co potwierdza silną autokorelację zjawisk epidemiologicznych.

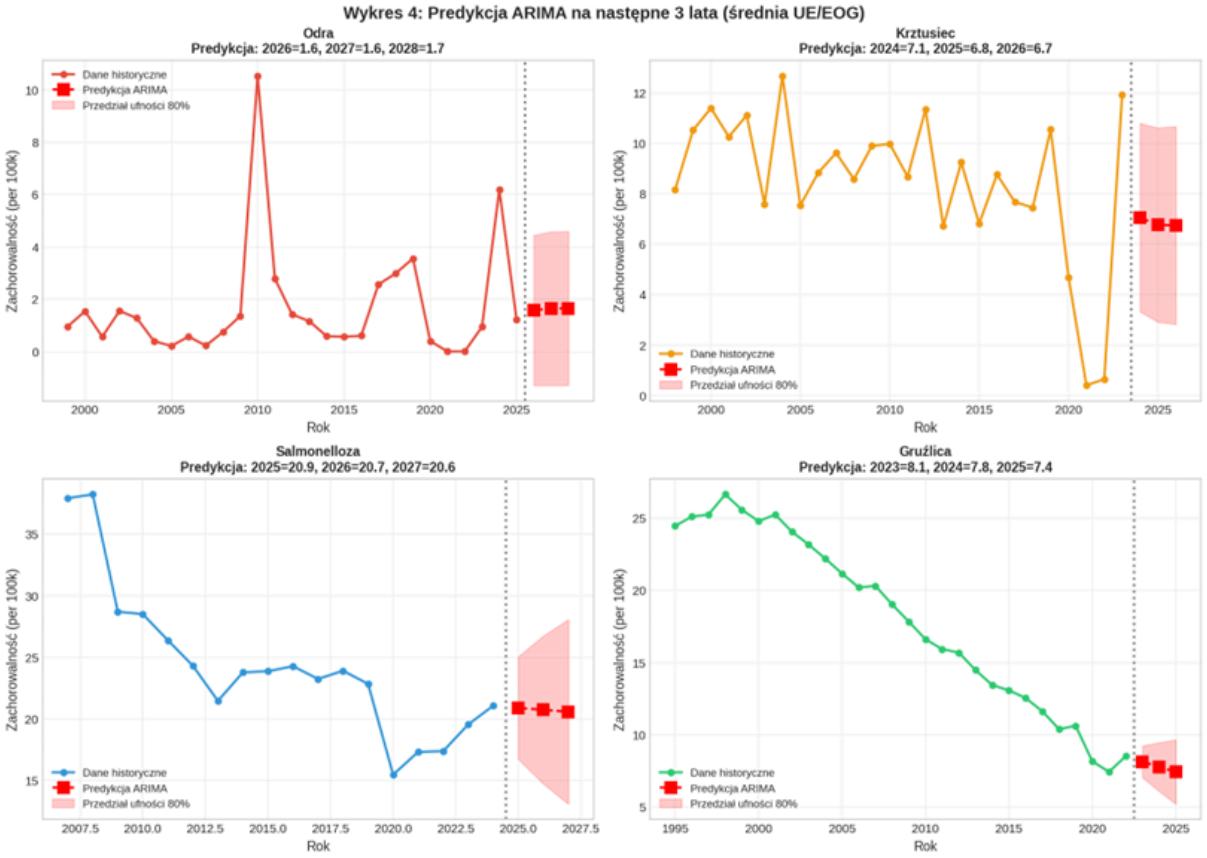


Wykres 2. Ważność cech. Źródło: opracowanie własne.

Analiza ważności zmiennych objaśniających w modelu lasu losowego jednoznacznie wskazała na dominującą rolę historycznych danych epidemiologicznych w procesie predykcji. Najistotniejszym czynnikiem dla wszystkich analizowanych jednostek chorobowych okazała się zmienna `Incidence_Lag1`, reprezentująca poziom zachorowalności w roku poprzednim. Wynik ten stanowi statystyczne potwierdzenie występowania tzw. efektu kaskadowego, w którym wysoka liczba aktywnych źródeł zakażenia w danym roku bezpośrednio przekłada się na zwiększoną transmisję patogenu w roku kolejnym $(t + 1)$. Drugim kluczowym parametrem, szczególnie w odniesieniu do chorób wirusowych takich jak odra, okazał się poziom wyszczepialności populacji (`Vaccination_Coverage_Pct`). Model wykrył istotną ujemną korelację między odsetkiem zaszczepionych a zapadalnością, identyfikując

jednocześnie nieliniowy “efekt progowy”. Wskazuje to, że spadek wyszczepialności nie skutkuje natychmiastowym wybuchem epidemii, lecz generuje ryzyko z opóźnieniem rzędu 1-3 lat, co odpowiada czasowi niezbędnemu na akumulację w populacji krytycznej liczby osób podatnych na infekcję.

3.4. Prognoza na lata 2025-2027 (ARIMA)



Wykres 3. Predykcja ARIMA na następne 3 lata. Źródło: opracowanie własne.

Model ARIMA przewiduje niepokojące trendy dla chorób zwalczanych drogą szczepień. Dla odry prognozowany jest dalszy wzrost zachorowań, co jest skorelowane z lukami w szczepieniach powstałymi w okresie pandemii COVID-19 (2020-2022). Jeśli chodzi o krztusiec, to model wskazuje na wzrostową fazę naturalnego cyklu 3-5 letniego. A dla salmonellozy i gruźlicy, przewidywana jest kontynuacja stabilnych trendów (odpowiednio stałego i spadkowego).

4. Dyskusja

Uzyskane wyniki są w dużej mierze zgodne z obserwacjami raportowanymi w literaturze. Wahania zapadalności na odrę i krztusiec potwierdzają znaczenie utrzymania wysokiego poziomu wyszczepialności populacyjnej. Spadek zachorowań na salmonellozę i gruźlicę koresponduje z wcześniejszymi badaniami wskazującymi na skuteczność interwencji systemowych. Ograniczeniem analizy jest wykorzystanie danych rocznych, które nie pozwalają na ocenę sezonowości. Ponadto zastosowane modele mają charakter uproszczony i nie uwzględniają czynników demograficznych ani polityki zdrowotnej poszczególnych krajów.

Wyniki badania podkreślają fundamentalną różnicę między chorobami endemicznymi o stabilnym przebiegu, a chorobami epidemicznymi. Wysokie wyniki R^2 dla salmonellozy i gruźlicy oznaczają, że obecne systemy kontroli są skuteczne i przewidywalne. Możliwe jest precyzyjne planowanie alokacji zasobów (leki, personel) na kolejny rok.

Z kolei niskie wartości R^2 dla odry (0,18) i krztuśca (bliskie 0,00) nie świadczą o błędzie w sztuce modelowania, lecz o nieprzewidywalnej naturze ognisk epidemicznych. Odra charakteryzuje się dynamiką “boom-and-bust”, napędzaną przez nieliniowe efekty progowe odporności zbiorowej. Model regresyjny ma tendencję do wygładzania wyników, przez co niedoszacowuje szczytów epidemii. Natomiast nieprzewidywalność krztuśca wynika z biologii bakterii (zanik odporności po 5-10 latach) oraz znacznego niedodiagnozowania u dorosłych (“przewlekły kaszel”), co zaszumia dane wejściowe.

Analiza wykazała również, że pandemia COVID-19 zaburzyła historyczne wzorce, tworząc dług immunologiczny, który obecnie (2024-2026) zaczyna manifestować się wzrostem zachorowań na choroby, którym można zapobiegać poprzez szczepienia. Zakłócenia rutynowych programów szczepień w trakcie pandemii COVID-19 doprowadziły do powstania niebezpiecznych luk odpornościowych, zwiększając ryzyko występowania ognisk chorób możliwych do zapobiegania szczepieniami w okresie popandemicznym [4].

5. Wnioski

Analiza potwierdza, że dane ECDC stanowią wartościowe źródło do badań porównawczych chorób zakaźnych w Europie. Połączenie analizy eksploracyjnej z prostym modelowaniem umożliwia identyfikację kluczowych trendów i problemów

epidemiologicznych. Dalsze badania powinny obejmować dane o wyższej rozdzielczości czasowej oraz bardziej zaawansowane modele epidemiologiczne.

Przeprowadzone badanie dowodzi, że nie istnieje jeden uniwersalny model analityczny adekwatny dla wszystkich chorób zakaźnych. Choć algorytmy zespołowe, w szczególności las losowy, wykazały najwyższą ogólną skuteczność, ich użyteczność jest ściśle zdeterminowana przez specyfikę biologiczną danego patogenu i charakter transmisji. Wyraźna dychotomia wyników między chorobami o stabilnych trendach a tymi o dynamice epidemicznej wskazuje, że sukces predykcji zależy w równym stopniu od doboru metod, co od stabilności samych procesów epidemiologicznych.

Kluczowym czynnikiem wpływającym na przyszłą zapadalność okazała się autokorelacja czasowa, co potwierdza, że najsilniejszym predyktorem zachorowań są wartości historyczne. Z perspektywy zdrowia publicznego oznacza to, że szybkie i skuteczne wygaszanie ognisk w bieżącym okresie jest najefektywniejszą metodą redukcji obciążenia epidemiologicznego w latach kolejnych. Równie istotną rolę odgrywa monitoring poziomu wyszczepialności, szczególnie w kontekście odry. Wykazano, że obserwacja spadków wskaźników szczepień pozwala na prognozowanie ryzyka wybuchu epidemii z wyprzedzeniem od jednego do trzech lat, co daje służbom medycznym niezbędny czas na wdrożenie kampanii uzupełniających.

W świetle uzyskanych wyników, dla chorób charakteryzujących się niskim współczynnikiem determinacji (R^2), takich jak odra czy krztusiec, rekomenduje się modyfikację podejścia prognostycznego. Tradycyjne prognozy w ujęciu rocznym powinny zostać zastąpione przez systemy wczesnego ostrzegania operujące na danych o wyższej częstotliwości (tygodniowej lub miesięcznej). Pozwoli to na efektywniejsze wykrywanie nagłych zmian trendu oraz nieliniowych efektów progowych, które umykają modelom opartym na danych zagregowanych rocznie.

6. Bibliografia

[1] W. J. Moss, “Measles control and elimination in Europe,” *The Lancet Infectious Diseases*, vol. 20, no. 5, pp. 531–532, 2020.
[2] M. K. Patel *et al.*, “Impact of vaccination rates on measles outbreaks in Europe,” *Vaccine*, vol. 41, no. 3, pp. 456–463, 2023.

[3] X. Li *et al.*, “Spatiotemporal modeling of measles transmission using negative binomial regression,” *Epidemics*, vol. 38, Art. no. 100536, 2022.
[4] World Health Organization, *Measles and Rubella Surveillance Data: WHO European Region 2022–2024*, WHO, Copenhagen, 2024.
[5] European Centre for Disease Prevention and Control, *Measles surveillance data*, ECDC, Stockholm, 2024.
[6] Eurostat, *Population on 1 January by age and sex*, European Commission, Luxembourg, 2024.