Studium magisterskie

Kierunek: Metody Ilościowe w Ekonomii i Systemy Informacyjne

Imie i nazwisko autora: Oskar Furmańczuk

Nr albumu: 81794

Determinanty przebiegu choroby COVID-19.

Praca licencjacka napisana
w Katedrze Matematyki i Ekonomii Matematycznej
pod kierunkiem naukowym
dr hab. Michała Ramszy

Warszawa 2021

Spis treści

1 Wprowdzenie

- przegląd litertury dotyczącej koronawirusa - artykuły poświęcone badaniom wpływu czynników biologicznych na przebieg choroby
- związek pracy z ekonomią

2 Przegląd literatury

2.1 Rasa jako determinanta przebiegu choroby COVID-19 u pacjentów z chorobą współistniejącą

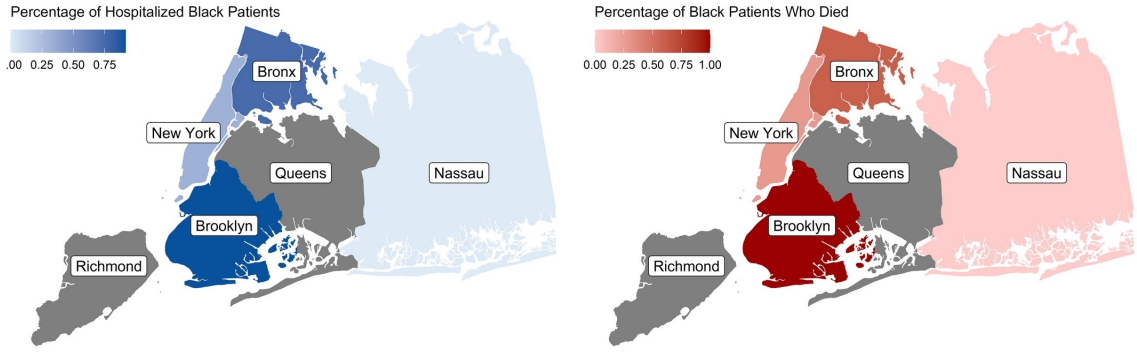
Jednym z projektów prowadzonych na rzecz opisu przebiegu COVID-19 u pacjentów pochodzących z różnych grup rasowych jest „Racial disparities in patients with coronavirus disease 2019 infection and gynecologic malignancy". Badanie to zostało przeprowadzone w Stanach Zjednoczonych na podstawie danych zebranych przez osiem systemów nowojorskich szpitali w okresie od 1 marca 2020 do 20 maja 202. Celem tego badania było sprawdzenie, czy istnieją rozbieżności na tle rasowym wśród pacjentek z rakiem ginekologicznym, u których stwierdzono obecność COVID-19.Wszystkie pacjentki biorące udział w badaniu miały ukończone 18 lat. Porównano charakterystykę wyjściową i kliniczną, zbadano różnice we wskaźnikach hospitalizacji i śmiertelności oraz wpływ rasy i innych czynników socjoekonomicznych i zdrowotnych na wyniki związane z COVID-19.

Charakterystyka pacjentów obejmowała wiek, podawaną przez siebie rasę i pochodzenie etniczne, hrabstwo zamieszkania, status zatrudnienia, status pracownika zasadniczego, status ubezpieczenia, status mieszkaniowy, medyczne choroby współistniejące, ciężkość zakażenia COVID-19, typ raka, stadium rozpoznania, aktualny stan choroby nowotworowej i ostatnie leczenie przeciwnowotworowe. Charakterystyka kliniczna COVID-19 obejmowała objawy COVID-19, parametry życiowe przy początkowym stadium choroby, powikłania szpitalne z powodu COVID-19 i konieczność stosowania dodatkowego tlenu, w tym inwazyjnej wentylacji mechanicznej.

Rasę sklasyfikowano jako 2 grupy: czarnoskórzy versus nie-czarnoskórzy (biali plus pozostali). Zgrupowano pacjentów, którzy identyfikowali się jako Azjaci, Amerykańscy Indianie lub Rdzenni mieszkańcy Alaski, a także rdzenni Hawajczycy lub mieszkańcy Wysp Pacyfiku do innej grupy ze względu na niską liczbę w każdej kategorii. Biorąc pod uwagę wysoki odsetek białych w grupie innej niż czarna, rasy następnie podzielono na trzy grupy - czarnoskórzy, biali oraz pozostali. Łącznie 193 pacjentów - 67 czarnoskórych oraz 126 pozostałych.

Statystyki opisowe zostały obliczone dla cech demograficznych, socjoekonomicznych, zdrowotnych, związanych z nowotworem oraz związanych z COVID-19 u pacjentów rasy czarnej i innej niż czarna. Zmienne ciągłe zostały opisane jako mediany i przedziały międzykwartylowe (IQR) i zostały porównane między grupami za pomocą testu sumy rang Wilcoxona. Zmienne kategoryczne przedstawiono jako częstości i proporcje, a następnie zostały porównane między grupami za pomocą testu chi kwadrat. Wskaźniki hospitalizacji i śmiertelności obliczono wśród pacjentów rasy czarnej i innej niż czarna w populacji ogólnej lub w subpopulacji stratyfikowanej przez inne zmienne kategoryczne, a następnie porównano je za pomocą testu chi kwadrat.

Analiza wyników wykazała, iż nad 70% pacjentów rasy czarnej w tym badaniu wymagało hospitalizacji z powodu zakażenia COVID-19, w porównaniu z zaledwie 46% pacjentów rasy innej niż czarna. Oprócz rasy i wieku, zły stan sprawności i większa liczba chorób współistniejących wiązały się ze zwiększonym prawdopodobieństwem przyjęcia do szpitala. W szczególności, pacjenci rasy czarnej w wieku poniżej 65 lat prawie 5 razy częściej wymagali hospitalizacji z powodu COVID-19 w porównaniu z pacjentami rasy czarnej w tym samym wieku. Rasa czarna nie była związana ze zwiększoną śmiertelnością z powodu COVID-19 przed lub po skorygowaniu o cechy kliniczne i socjoekonomiczne. Badanie wykazało, że pacjenci rasy czarnej są od 2 do 3 razy bardziej narażeni na konieczność hospitalizacji niż pacjenci rasy białej po skorygowaniu czynników zakłócających, w tym wieku, płci, chorób współistniejących i dochodów, a prawdopodobieństwo ich śmierci z powodu zakażenia COVID-19 jest ponad 5 razy większe.



Wykres 1: Odsetki hospitalizacji (po lewej) i dane dotyczące śmiertelności (po prawej) zilustrowane dla pacjentów rasy czarnej w badanych dzielnicach Nowego Yorku (USA)

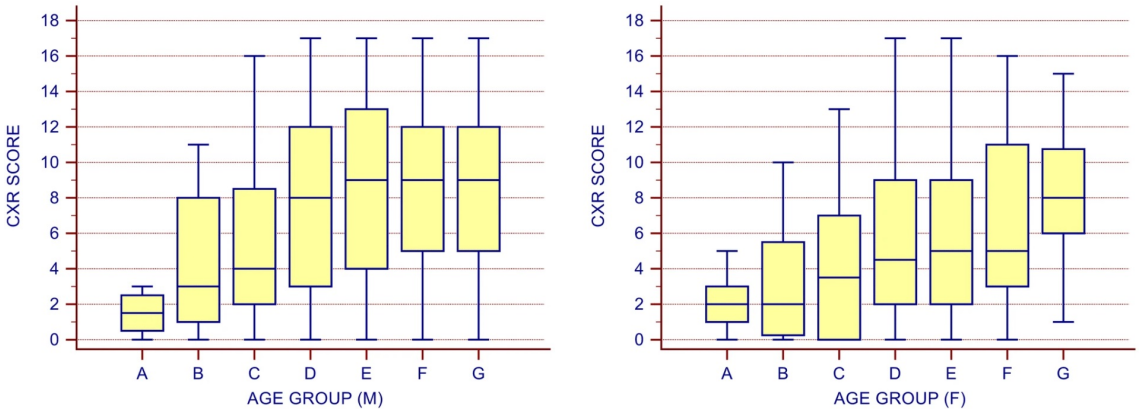
Autorzy badania wskazują za przyczynę gorszych rokowań czarnoskórych czynniki takie jak ograniczony dostęp do świadczeń opieki zdrowotnej, strukturalne i społeczne uwarunkowania opieki medycznej, rasizm i dyskryminację. Ponadto, Afroamerykanie są bardziej narażeni na współistniejące choroby medyczne, o których wiadomo, że są czynnikami ryzyka ciężkiego zakażenia COVID-19, w tym nadciśnienie tętnicze, cukrzycę, choroby nerek i układu oddechowego. W tej grupie badawczej większy odsetek czarnych pacjentów w wieku poniżej 65 lat miał więcej niż 3 współistniejące choroby i charakteryzował się większą częstością występowania nadciśnienia tętniczego, otyłości i cukrzycy w porównaniu z pacjentami rasy innej niż czarna w tej grupie wiekowej. Czarnoskórzy pacjenci mieli również częściej zamieszkiwali obszary poniżej granicy ubóstwa. [**?**]

2.2 Wpływ wieku oraz płci na przebieg COVID-19

Jednym z badań które zgłębia przebieg infekcji COVID-19 u pacjentów różnej płci z podziałem na grupy wiekowe jest Żadiographic severity index in COVID-19 pneumonia: relationship to age and sex in 783 Italian patients". Badnie zostało przeprowadzone na grupie 783 (532 mężczyzn i 251 kobiet) włoskich pacjentów z laboratoryjnie stwierdzonym COVID-19. Osoby poniżej 20 roku życia zostali wykluczeni z badania. Pozostali pacjenci zostali podzieleni na 7 grup wiekowych: 20-29 lat (grupa A), 30-39 lat (grupa B), 40-49 lat (grupa C), 50-59 lat (grupa D), 60-69 lat (grupa E), 70-79 lat (grupa F) oraz powyżej 80 lat (grupa G). Do interpretacji stanu pacjenta posłużono się osiemnastopoziomowym wskaźnikiem oceny sprawności płuc uzyskiwanym przez analizę prześwietleń klatki piersiowej (dalej: CXR).

Mediana wieku wynosiła 65 lat (zakres międzykwartylowy, 55-74 lata). Spośród włączonych pacjentów, 10 (1,3%) było z grupy A, 29 (3,7%) z grupy B, 80 (10,2%) z grupy C, 168 (21,5%) z grupy D, 196 (25%) z grupy E, 210 (26,8%) z grupy F i 90 (11,5%) z grupy G. Dla każdej grupy, test Manna-Whitneya U został użyty do porównania wyników CXR mężczyzn i kobiet. Korelacja rang Spearmana została zastosowana do oceny zależności pomiędzy wynikiem CXR a wiekiem. Test Kruskala-Wallisa zastosowano również w celu określenia, czy istnieją istotne różnice w punktacji CXR pomiędzy grupami wiekowymi. Wartości $p \leq 0{,}05$ uznano za istotne statystycznie.

Wynik CXR był istotnie wyższy u mężczyzn niż u kobiet tylko w grupach D, E i F ($p < 0,020$). Stwierdzono istotną korelację między wynikiem CXR a wiekiem zarówno u mężczyzn, jak i u kobiet ($rho = 0,205$, $p < 0,0001$ dla mężczyzn; $rho = 0,310$, $p < 0,0001$). U mężczyzn wynik CXR w grupach D, E, F i G był znamiennie wyższy niż w grupach A, B i C (ryc. 3). U kobiet wynik CXR w grupach E, F i G był znamiennie wyższy niż w grupach A i B (ryc. 3). Wynik CXR w grupie G był również znamiennie wyższy niż w grupach C, D i E. Wynik CXR w grupie F był znamiennie wyższy niż w grupie C.



Wykres 2: Rozkład wyników badania RTG klatki piersiowej (CXR) w zależności od grupy wiekowej u mężczyzn (M) i kobiet (F)

Podsumowując, autorzy badania wskazują na istotny wpływ wieku na śmiertelność przy infekcji COVID-19. Stwierdzono, że mężczyźni w wieku 50 lat lub starsi i kobiety w wieku 80 lat lub starsze wykazywali najwyższe ryzyko rozwoju ciężkiej choroby płuc. [**?**]

2.3 Zmienne meteorologiczne jako determinanty częstotliwości zachorowań na COVID-19

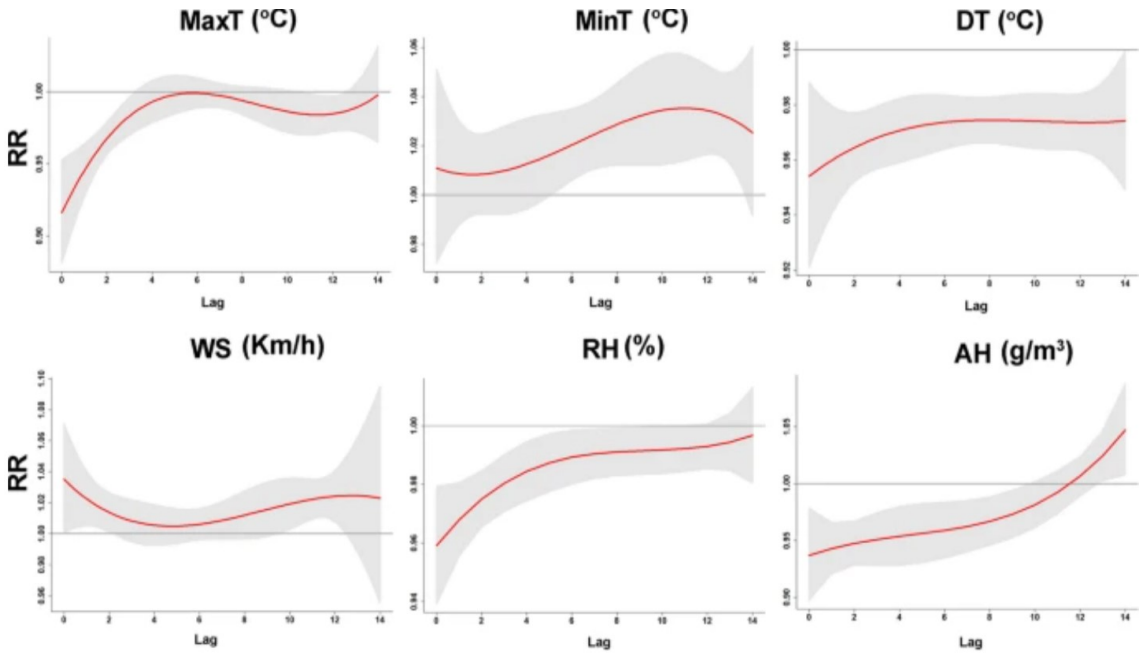
Opublikowany w wrześniu 2020 artykuł „Effect of meteorological factors on COVID-19 cases in Bangladesh" koncertuje się na opisie wpływu warunków pogodowych na możliwość zarażenia się koronawirusem. Badanie na temat którego powstał wspomniany artykuł przeprowadzono na terytorium Bangladeszu. Kraj znajduję się w strefie klimatu zwrotnikowego, a na jego terytorium występują 3 główne pory roku: pre-monsun,monsun, post-mosum. Ze względu na złożone warunki klimatyczne i dużą gęstość zaludnienia obszar ten jest uznawany za wysoce wrażliwy na zmiany klimatu.

Do przeprowadzenia badania zebrano dane z 43 stacji meteorologicznych z których wyodrębniono informacje na temat kilkunastu parametrów atmosferycznych (maksymalna temperatura dobowa (MaxT), minimalna temperatura dobowa (MinT), siła wiatru (WS), wilgotność relatywna (RH), wilgotność absolutna (AH) itp.) dla odpowiadającym im regionów. Na podstawie danych o przypadkach zachorowani na COVID-19 z odpowiadających im obszarów ustalono wskaźnik relatywnego ryzyka (RR), który oceniał prawdopodobieństwo zarażenia się SARS-CoV-2 w określonym dniu. Ze względu na od 2 do 14 dniowy okres inkubacji wirusa przyjęto maksymalnie 14 dniowe opóźnienie między obserwacjami meteorologicznymi, a poszczególnymi przypadkami zgłoszonych zachorowań.

Koherencja transformaty falkowej (WTC) i częściowa koherencja falkowa (PWC) zostały wykorzystane w tym badaniu do uzyskania wyrazów rozdzielczości czasowej i częstotliwościowej zmiennych klimatycznych i przypadków COVID-19 w Bangladeszu. WTC kwantyfikuje wielkość kowariancji między dwoma szeregami czasowymi, która waha się od 0 do 1 ($0 \leq R2 \leq 1$). 0 odnosi się do całkowitego braku spójności, podczas gdy 1 odnosi się do doskonałej spójności. Zakres ten definiuje się jako kwadrat widma krzyżowego znormalizowanego przez wygładzone indywidualne widmo mocy.

RH wykazało silny pozytywny znaczący związek z przypadkami COVID-19 w Singapurze. Oznacza to, że maksymalne RH (71,4 $\pm$ 4%) w maju sprzyjało rozprzestrzenianiu się COVID-19. WS wykazał znaczący związek z potwierdzonymi przypadkami COVID-19 w Bangladeszu. Natomiast AP wykazywał stosunkowo silną odwrotną zależność z pozytywnymi przypadkami zakażenia COVID-19 w Bangladeszu w początkowej fazie epidemii SARS-CoV-2. Ogólnie rzecz biorąc, wysokie wartości MinT, RH i AH wraz z niskimi WS w maju nasiliły rozprzestrzenianie się SARS-CoV-2.

Przedstawiono zależność opóźnionej odpowiedzi RR na wzrost o 1 jednostkę wszystkich wskaźników meteorologicznych w różnych dniach opóźnienia (do 14 dni). Największe RR dla MaxT wyniosło 1,00 (95% CI 0,99-1,01) w opóźnieniu 6-dniowym, a najmniejsze 0,92 (95% CI 0,88-0,95) bez opóźnienia. Największe RR dla MinT wyniosło 1,04 (95% CI 1,01-1,06) w opóźnieniu 11-dniowym, a najmniejsze 1,01 (95% CI 0,99-1,02) w opóźnieniu 2-dniowym.



Wykres 3: Pojedyncze efekty MaxT, MinT, DT, WS, RH i AH. Laboratorium Y oznacza wartość ryzyka względnego (RR)

Podsumowując, zaobserwowano znaczący wpływ pomiędzy COVID-19 a zmiennymi meteorologicznymi. Temperatura, może odgrywać istotną rolę w procesach życiowych człowieka oraz w zakresie ograniczania i kontroli epidemii. Ponadto, oddnotowano wpływ siły wiatru na propagowanie epidemii. Wiatr może wpływać na czas zawieszenia

koronawirusa i jego rozprzestrzenianie się. Badanie wykazało, że wskaźnik RR wyraźnie wzrósł, w czasie gdy WS podwyższył się o ponad 21 km/h. Stężenie koronawirusa może być rozcieńczone przez podwyższony WS, co może stanowić prawdopodobne wyjaśnienie tego wyniku. Można wnioskować, iż temperatura i prędkość wiatru wykazują silny związek z wybuchem i przebiegiem epidemii COVID-19 w Bangladeszu. [**?**]

3 Opis zbioru danych

- liczebność oraz źródło zbioru danych
- opisy zmiennych
- rozkład zmiennych
- charakterystyki zmiennych (np. średnie, odchylenia std)
- charakterystyka brakujących danych
- problemy wynikające z brakujących danych

Do przeprowadzenia badania zostały użyte 3 zbiory danych:

- dane na temat poszczególnych przypadków zachorowań na COVID-19 w USA - pochodzą z zasobów Centers for Disease Control and Prevention
- wskaźniki atmosferyczne dla poszczególnych części USA - pochodzą z zasobów National Centers for Environmental Information
- dane demograficzne dla poszczególnych części USA - pochodzą z portalu ArcGis

3.1 Zbiór danych dotyczący poszczególnych przypadków zachorowań na COVID-19

Zbiór danych pochodzi z zastrzeżonych zasobów Centers for Disease Control and Prevention (CDC). Zbiór ten został udostępniony po bezpośredniej prośbie skierowanej do jego właścicieli i za ich pozwoleniem został wykorzystany w tej pracy. Zebrane są w nim podstawowe dane medyczne osób ze stwierdzonym COVID-19 zarejestrowanych przez amerykańską służbę zdrowia z okresu od pierwszego potwierdzonego przypadku zachorowania (styczeń 2020) do końca lutego 2021 roku - łącznie 20,5 mln obserwacji. Zmienne, które zostały użyte w tej pracy zostały scharakteryzowane w tabeli 1 .

Zmienna	Opis	Typ zmiennej
race_ethnicity_combined	Rasa lub grupa etnicna	wielomianowa nieuporządkowana
cdc_case_earliest_dt	Najwcześniejsza data związana z obserwacją	data
sex	Płeć	dwumianowa
hosp_yn	Hospitalizacja pacjenta	dwumianowa
icu_yn	Pobyt na oddziale intensywnej terapi	dwumianowa
hc_work_yn	Status pracownika służby zdrowia	dwumianowa
pna_yn	Obecne zapalenie płuc	dwumianowa
abxchest_yn	Nieprawidłowości wykryte przez RTG klatki piersiowej	dwumianowa
acuterespdistress_yn	Obecność zespołu ostrej niewydolności oddechowej	dwumianowa
mechvent_yn	Użycie mechanicznej wentylacji (intubacja)	dwumianowa
fever_yn	Występowanie gorączki	dwumianowa
sfever_yn	Subiektywne poczucie gorączki	dwumianowa
chills_yn	Występowanie dreszczy	dwumianowa
myalgia_yn	Występowanie bóli mięśniowych	dwumianowa
runnose_yn	Występowanie kataru	dwumianowa
sthroat_yn	Występowanie bólu gardła	dwumianowa
cough_yn	Występowanie kaszlu	dwumianowa
sob_yn	Występowanie trudności z oddechaniem	dwumianowa
nauseavomit_yn	Występowanie nudności	dwumianowa
headache_yn	Występowanie bólu głowy	dwumianowa
abdom_yn	Występowanie bólu brzucha	dwumianowa
diarrhea_yn	Występowanie biegunki	dwumianowa
medcond_yn	Występowanie współistniejących schorzeń	dwumianowa
county_fips_code	Kod pocztowy	licznikowa
res_county	Zamieszkane hrabstwo	wielomianowa nieuporządkowana
res_state	Zamieszkany stan	wielomianowa nieuporządkowana
age_group	Grupa wiekowa	wielomianowa uporządkowana
death_yn	Zgon pacjenta	dwumianowa

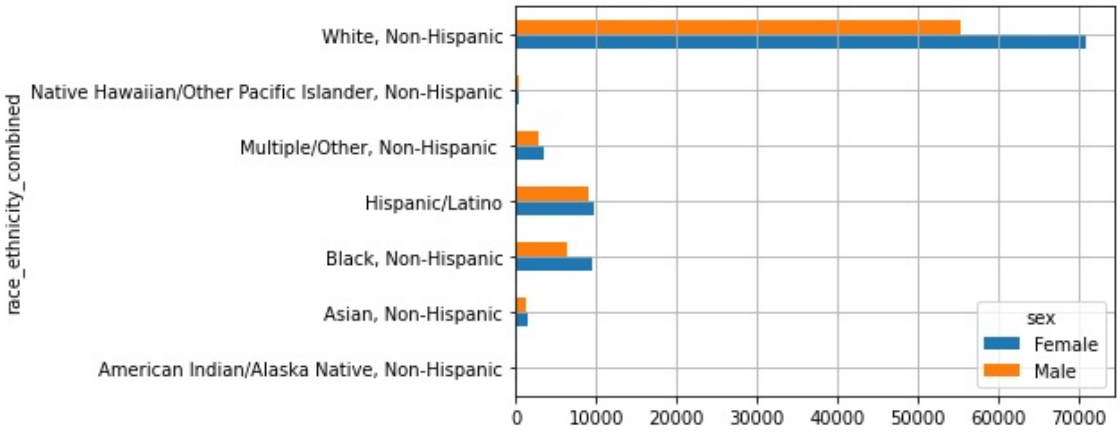
Tabela 1: Zmienne ze zbioru CDC użyte w dalszej części pracy.

W zbiorze danych występują 3 typy brakujących danych. Najczęściej występującym jest *Unknown* i jak autor zbioru danych opisuje kodowany był w przypadku udzielenia dokładnie takiej odpowiedzi do pytania szpitalnej ankiety. Kolejnym typem jest *Missing* i występuję w przypadkach nie udzielenia jakiejkolwiek odpowiedzi. Najrzadziej występującym typem wskazującym brak danych jest *NaN* i zakodowany został w miejscach logicznej nieścisłości danych oraz w przypadkach błędów na poziomie zapisu danych w centralnej bazie.

Pomimo bardzo dużej ilości obserwacji wadą zbioru okazał się niski udział w pełni opisanych przypadków (większość zmiennych zawiera mniej niż 25% uzupełnionych danych). Obliczone częstości występowania oraz korelacje brakujących danych pomiędzy zmiennymi z uwzględnieniem oraz bez uwzględnienia typów braków danych nie wskazała przyczyny dużego natężenia braku danych. Uznano również, że lokalizacja brakujących danych jest niedeterministyczna. Ze względu na jakościową naturę zmiennych oraz wysoki udział brakujących danych nie możliwe było przeprowadzenie imputacji brakujących danych. Znikoma wartość informacja niesiona przez brak danych przyczyniłaby się do zaburzenia interpretacji wyników dalszej części pracy dlatego podjęto decyzję o odrzuceniu wszystkich niepełnych obserwacji. Dalsza część pracy odnosi się do wyselekcjonowanych tym sposobem 171 147 pełnych obserwacji.

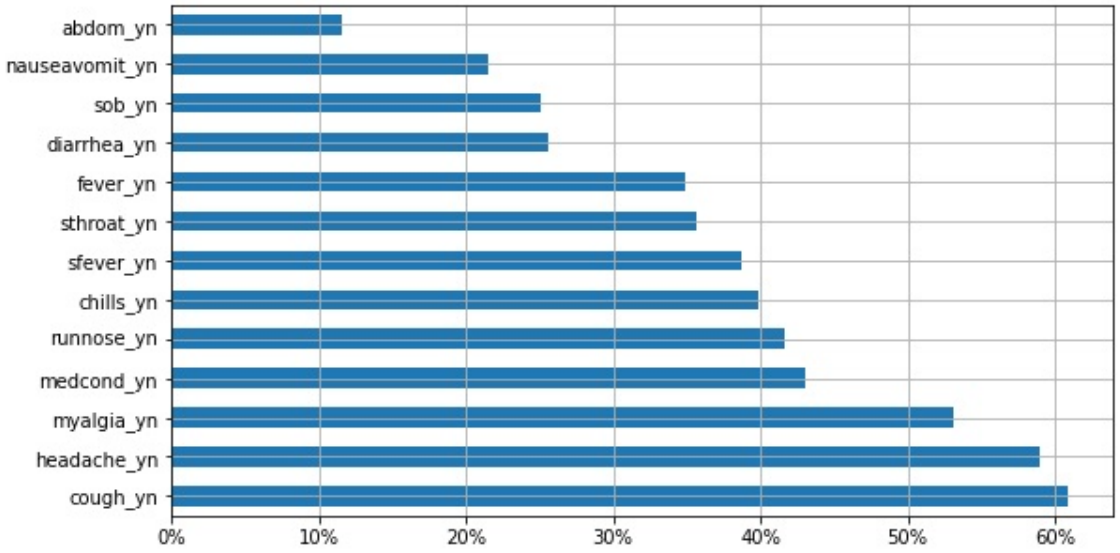
W zbiorze udostępnionym przez CDC pacjenci zostali podzieleni na 8 grup wiekowych o rozpiętości 10 lat, a dla najstarszych stworzono grupę 80+. Zmienna *race_ethnicity_combined* przyjmuje wartości: *American Indian/Alaska Native, Non-Hispanic* dla natywnych mieszkańców Ameryki Północnej, *Asian, Non-Hispanic* dla rasy żółtej, *Black, Non-Hispanic* dla rasy czarnej, *Hispanic/Latino* dla mniejszości latynoskiej, *Native Hawaiian/Other Pacific Islander, Non-Hispanic* dla natywnych mieszkańców wysp Pacyfiku, *White, Non-Hispanic* dla rasy białej oraz *Multiple/Other, Non-Hispanic* dla pozostałych. Wartości zmiennej *sex* zostały przekodowane na 1 dla mężczyzn oraz 0 dla kobiet. Wartości dla pozostałych zmiennych dwumianowych zostały przekodowane na 1 dla wartości *Yes* oraz na 0 dla *No*.

Najliczniejszą grupą w zbiorze okazali się być przedstawiciele rasy białej i składają się na 73.8% wszystkich obserwacji. Natywni mieszkańcy wysp Pacyfiku (0,38%) oraz natywni mieszkańcy Ameryki Północnej (0,11%) stanowią najmniej liczne grupy etniczne w opisywanym zbiorze danych. Udział procentowy kobiet jest porównywalny do udziału mężczyzn i wynosi odpowiednio 44,2% dla mężczyzn oraz 55,8% dla kobiet (Wykres: 4)



Wykres 4: Liczebność poszczególnych grup etniczno-rasowych z podziałem na płeć

Najczęstszymi symptomami COVID-19 był kaszel (60,9%), ból głowy (59,0%) oraz bóle mięśniowe (53.1%). Najrzadziej pacjenci informowali o dolegliwościach związanych z bólem brzucha (11,5%) oraz nudnościami (21,5%) W opisanym zbiorze wskaźnik hospitalizacji wyniósł 6,72%, a wskaźnik śmiertelności 1,28%.



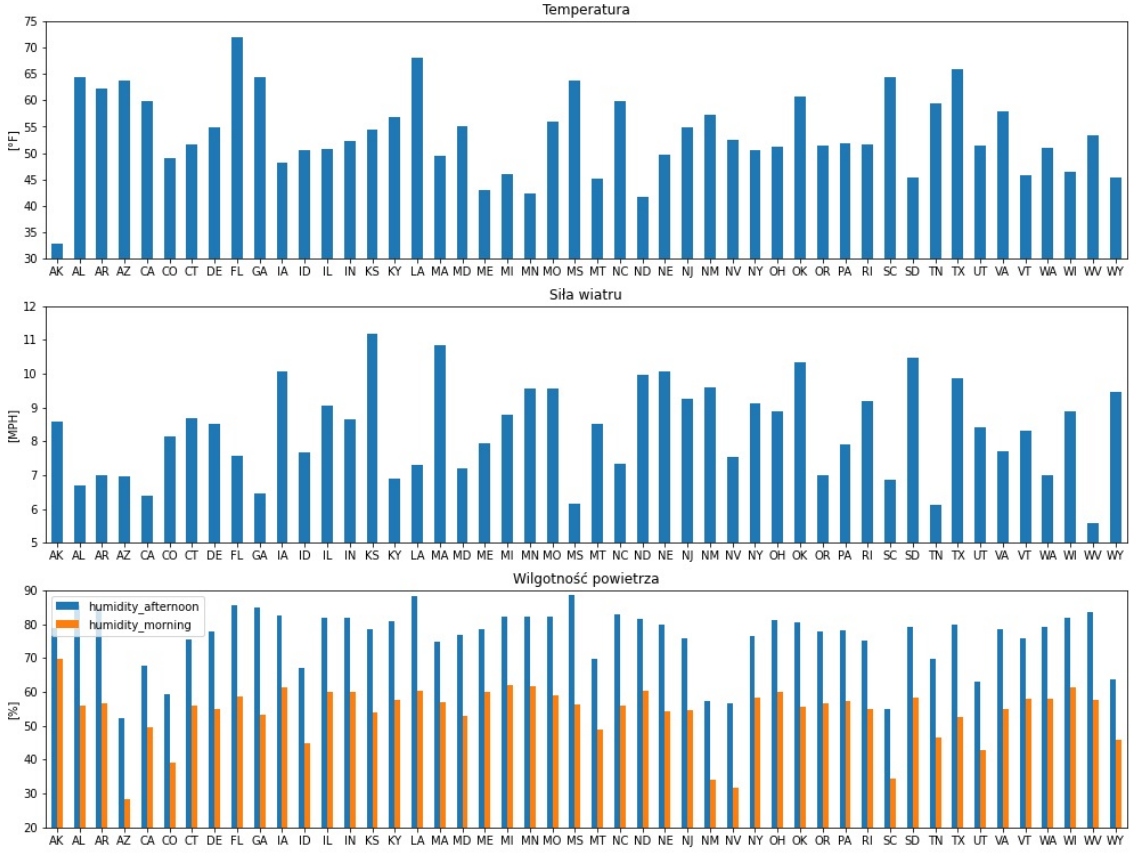
Wykres 5: Częstość występowania poszczególnych objawów COVID-19

3.2 Zbiory danych zawierający wskaźniki atmosferyczne dla poszczególnych części USA

Kolejne zbiory danych jakie użyto w tej pracy pochodzą z upublicznionych zasobów National Centers For Environmental Information. Zawierają one dane na temat najważniejszych parametrów metrologicznych obserwowanych przez największe miasta USA. Udostępnione dane zawierają średnie wartości zagregowane na poziomie miesięcznym i rocznym. Korzystając z tego źródła do tej pracy użyto trzech poniższych tabel:

- *Normal Daily Maximum Temperature, °F* - średnie dobowe temperatury obliczone na podstawie obserwacji zebranych w latach 1971-2000 wyrażone w stopniach Fahrenheita.
- *Wind - Average Speed (MPH)* - średnie prędkości wiatru obliczone bez uwzględnienia jego kierunku wyrażone w milach na godzinę.
- *Average Relative Humidity (Percent) - Morning (M) and Afternoon (A)* - średnie wilgotności powietrza rano i wieczorem obliczone wyrażone w procentach.

Ze względu na fakt, iż dla niektórych hrabstw nie odnotowano żadnych obserwacji postanowiono zgrupować dane na poziomie stanów i wyliczyć dla nich średnie wartości. W zestawieniu rocznym najgorętszym stanem okazała się Floryda osiągając średnią temperaturę 72,1 °F (22,3 °C), najchłodniejszym zaś Alaska ze średnią temperaturą 32,8 °F (0,4 °C). Największa średnia roczna prędkość wiatru przypada na stan Kansas (11,2 mph) i Massachusetts (10,9 mph), najmniejsza zaś w stanie West Virginia (5,6 mph).



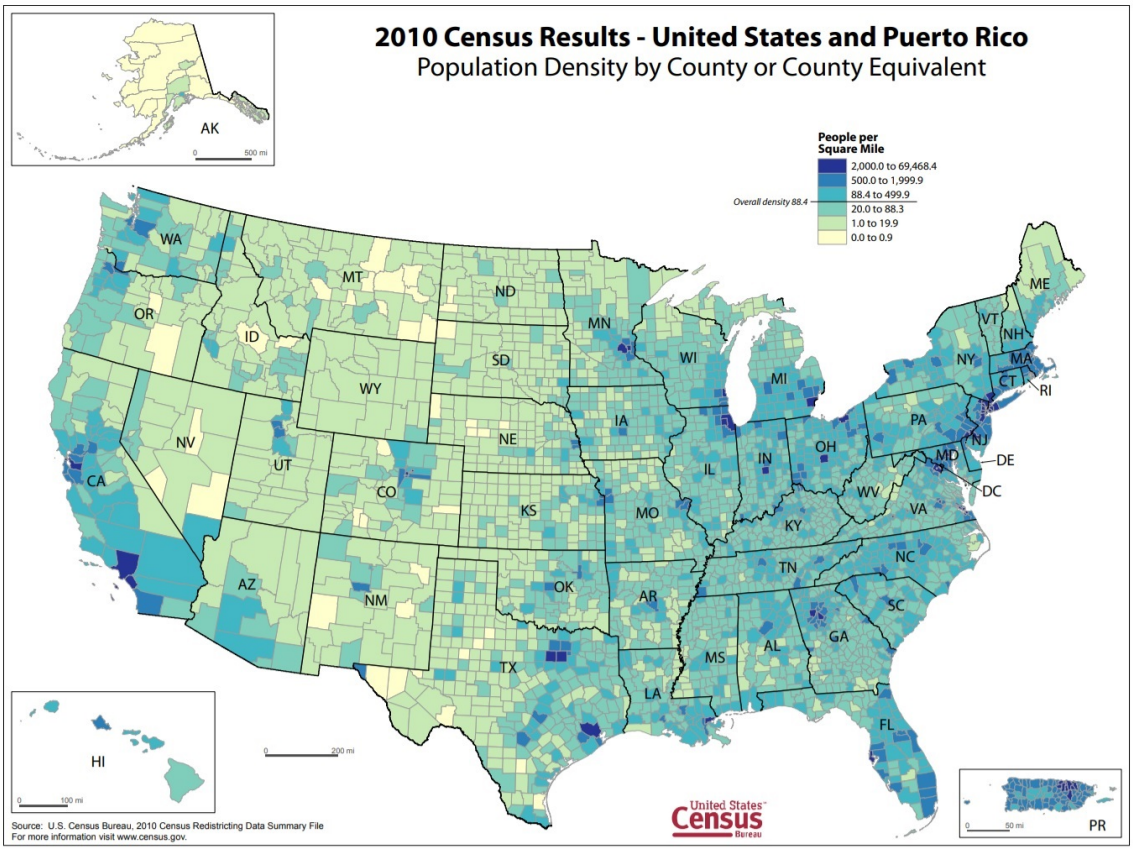
Wykres 6: Wybrane wskaźniki atmosferyczne zagregowane na poziomie rocznym zgrupowane dla poszczególnych stanów.

3.3 Zbiór danych zawierający dane demograficzne

Ostatnim zbiorem jaki został wykorzystany w tej pracy jest zbiór *USA Counties* pochodzący z ArcGIS Hub. Autorem zbioru jest ArcGIS Data and Maps (poprzednio Esri Data & Maps), czyli sam zespół portalu ArcGIS. Oprócz danych demograficznych w zbiorze znaleźć można dane geograficzne oraz wskaźniki związane z rolnictwem. Dane demograficzne datowane są na 2010 rok i pochodzą z amerykańskiego spisu powszechnego. Wszystkie zmienne zostały określone na poziomie hrabstw.

Ze względu na tematykę pracy postanowiono wykorzystać z opisanego zbioru tylko jedną zmienną, czyli *POP_SQMI* - zagęszczenie ludności na mile kwadratową. Analiza zbioru wykazała, że najmniej rzadziej zaludnione hrabstwa w USA to Lake and Peninsula, North Slope, Denali, Yakutat oraz Loving z gęstością zaludnienia na poziomie 0,1 osoby na mile kwadratowej. Wszystkie wymienione hrabstwa oprócz Loving (stan Te-

xas) zlokalizowane są na obszarze Alaski. Najgęściej zaludnione hrabstwa to New York (73032,2 osoby na mile kwadratową), Kings (38512,3 osoby na mile kwadratową) oraz Bronx (34919,1 osoby na mile kwadratową) - wszystkie zlokalizowane w stanie New York. Gęstość zaludnienia Stanów Zjednoczonych, obliczona przez zsumowanie populacji wszystkich hrabstw i podzielenie przez sumę ich powierzchni, wynosi 91,1 osoby na mile kwadratową.



Wykres 7: Zagęszczenie ludności zwizualizowane na mapie Stanów Zjednoczonych.

3.4 Docelowy zbiór danych

Po uzyskaniu zbiorów z zewnętrznych zasobów oraz wstępnego oczyszczenia danych postanowiono stworzyć zbiór danych, który posłuży do budowy modelu XGBoost. Podstawą do budowy tego zasobu był zbiór danych związany z poszczególnymi obserwacjami zachorowań COVID-19 do którego dowiązano w określony sposób pozostałe zbiory zbiory.

Pierwszym krokiem było dowiązanie wskaźników atmosferycznych do zbioru danych związanego z obserwacjami COVID-19 . Dokonano tego przez zmapowanie miesiąca zachorowania i zamieszkanego stanu ze wskazaną zmienną meteorologiczną przypisaną dla tego stanu w tym terminie. W ten sposób otrzymano następujące zmienne:

- *avg_temp* - średnia dobowa temperatura wyrażona w stopniach Fahrenheita.
- *avg_wind* - średnia prędkość wiatru wyrażona w milach na godzinę.
- *avg_humidity_M* - średnia wilgotność powietrza rano wyrażona w procentach.
- *avg_humidity_A* - średnia wilgotność powietrza wieczorem wyrażona w procentach.

Kolejnym etapem było dowiązanie powstałego w poprzednim kroku zbioru ze zbiorem danych uzyskanym z portalu ArcGIS. Osiągnięto to przez zmapowanie nazwy zamieszkanego hrabstwa osoby ze stwierdzonym COVID-19 z gęstością zaludnienia w tym obszarze. W ten sposób uzyskano zmienną *pop_density*, która określa gęstość zaludnienia wyrażoną w liczbie osób na mile kwadratową.

Ostatnim krokiem było odrzucenie niepożądanych zmiennych z powstałego zbioru danych. Ten etap był konieczny gdyż niektóre zmienne nie miały bezpośredniego związku z wyjaśnieniem przedmiotu tej pracy. Odrzuconymi zmiennymi były: *county_fips_code*, *res_county*, *res_state*, *cdc_case_earliest_dt* oraz *date_month*.

4 Metody

- opis kolejności obliczeń/transformacji/budowy modelu
- opis użytego modelu (docelowo: xgboost) oraz jego zasadność względem niezbalansowanej próby
- opis metod walidacji oraz jakości modelu (F1-score, accuracy)
- opis sposobu interpretacji wpływu zmiennych objaśniających na zmienną objaśnianą (docelowo: SHAP)

4.1 Podział zbioru na treningowy, walidacyjny i testowy

W uczeniu maszynowym powszechnym zadaniem jest badanie i konstruowanie algorytmów, które potrafią uczyć się na podstawie danych i dokonywać predykcji na ich podstawie. [**?**] Model jest początkowo dopasowywany do zbioru danych treningowych, który jest zbiorem przykładów używanych do dopasowania parametrów (np. wag połączeń między neuronami w sztucznych sieciach neuronowych) modelu. Model (np. sieć neuronowa lub naiwny klasyfikator Bayesa) jest trenowany na zbiorze danych treningowych za pomocą metody uczenia nadzorowanego, na przykład za pomocą metod optymalizacji, takich jak opadanie gradientowe lub stochastyczne opadanie gradientowe. W praktyce, zbiór danych szkoleniowych często składa się z par wektora wejściowego (lub skalara) i odpowiadającego mu wektora wyjściowego (lub skalara), gdzie klucz odpowiedzi jest powszechnie oznaczany jako cel (lub etykieta). Bieżący model jest uruchamiany z zestawem danych treningowych i wytwarza wynik, który jest następnie porównywany z celem, dla każdego wektora wejściowego w zestawie danych treningowych. W oparciu o wynik porównania i określony algorytm uczenia się, parametry modelu są dostosowywane. Dopasowanie modelu może obejmować zarówno selekcję zmiennych, jak i estymację parametrów.

Następnie, dopasowany model jest używany do przewidywania odpowiedzi dla obserwacji w drugim zbiorze danych zwanym zbiorem danych walidacyjnych. Zbiór danych walidacyjnych zapewnia bezstronną ocenę dopasowania modelu na zbiorze danych treningowych podczas dostrajania hiperparametrów modelu (np. liczba jednostek ukrytych - warstw i szerokości warstw - w sieci neuronowej). Walidacyjne zbiory danych mogą być

użyte do regularyzacji poprzez wczesne zatrzymanie (zatrzymanie treningu, gdy błąd na walidacyjnym zbiorze danych wzrasta, ponieważ jest to oznaką przepasowania do zbioru treningowego). Ta prosta procedura jest skomplikowana w praktyce przez fakt, że błąd zbioru walidacyjnego może wahać się podczas treningu, tworząc wiele lokalnych minimów. Ta komplikacja doprowadziła do powstania wielu reguł ad hoc służących do określania, kiedy przepełnienie naprawdę się rozpoczęło.

Finalnie, testowy zbiór danych jest zbiorem danych używanym w celu zapewnienia bezstronnej oceny ostatecznego dopasowania modelu na treningowym zbiorze danych. Jeśli dane w testowym zbiorze danych nigdy nie były używane w trenowaniu modelu (na przykład w walidacji krzyżowej), testowy zbiór danych jest również nazywany zbiorem danych przejściowych (ang. holdout dataset). Termin źbiór walidacyjny"jest czasami używany zamiast terminu źbiór testowy"(np. jeżeli oryginalny zbiór danych został podzielony tylko na dwa podzbiory, zbiór testowy może być określany jako zbiór walidacyjny).[**?**]

Docelowy zbiór danych użyty w tej pracy został losowo podzielony na treningowy, walidacyjny i testowy w stosunku 3:1:1. Etap ten umożliwił w kolejnych krokach dobranie właściwych parametrów oraz zbudowanie optymalnego modelu.

4.2 Budowa modelu klasyfikacyjnego XGBoost

XGBoost to biblioteka oparta na licencji otwartego kodu źródłowego, która dostarcza regularyzujący framework do wspomagania gradientowego dla C++, Javy, Pythona, R, Julii, Perla i Scali. XGBoost jest kompatybilny z systemami Linux, Windows i macOS. Działa na pojedynczej maszynie, jak również na frameworkach przetwarzania rozproszonego Apache Hadoop, Apache Spark i Apache Flink.

Klasyfikator XGBoost posiada wiele hiperparametrów, które mogłyby posłużyć do strojenia modelu. Najważniejsze z nich zostały użyte do stworzenia optymalnego modelu dla tej pracy:

- *learning_rate* - Tempo uczenia modelu. Zmniejszanie wielkości kroków w aktualizacji zapobiega przeszacowaniu modelu.

- *max_depth* - Maksymalna głębokość drzewa. Zwiększenie tej wartości sprawi, że model będzie bardziej złożony i prawdopodobnie obszerniejszy.

- *min_child_weight* - Minimalna suma wagi instancji w poprzedniku. Jeśli krok podziału drzewa skutkuje powstaniem węzła liścia z sumą wagi instancji mniejszą niż min_child_weight, wówczas proces tworzenia rezygnuje z dalszego podziału.
- *scale_pos_weight* - Współczynnik zbalansowania zbioru treningowego. Kontrola równowagi wag dodatnich i ujemnych, przydatna w przypadku niezrównoważonych klas.
- *gamma* - Minimalna redukcja strat wymagana do utworzenia kolejnej partycji na węźle liścia drzewa. Im większa wartość gamma, tym bardziej konserwatywny algorytm.

By zapewnić wysoką wydajność procesowania autorzy biblioteki zaparzyli ją w wysoce wydajne mechanizmy. Biblioteka XGBoost obsługuję między innymi trzy formy wspomagania gradientowego:

- Algorytm Gradient Boosting zwany również maszyną gradient boostingową z uwzględnieniem współczynnika uczenia.
- Stochastyczny Gradient Boosting z podpróbkowaniem na poziomie wiersza, kolumny i kolumny na poziom rozdzielenia.
- Regularyzowany Gradient Boosting z regularyzacją zarówno L1 jak i L2.

Implementacja algorytmu została zaprojektowana pod kątem efektywności czasu obliczeń i zasobów pamięci. Niektóre kluczowe cechy implementacji algorytmu obejmują:

- Implementacja uwzględniająca rzadkość danych z automatyczną obsługą brakujących wartości danych.
- Struktura blokowa wspierająca paralelizację konstrukcji drzewa.
- Inkrementalne trenowanie modelu, dzięki któremu można dalej wzmacniać już dopasowany model na nowych danych.

Oprócz poprzednio wymienionych mechanizmów algorytm klasyfikatora XGBoost umożliwia korzystanie z niezbalansowanych zbiorów danych. Dla użytkownika klasyfikatora ta funcjonalność udostępniona jest przez hiperparametr *scale_pos_weight*. Ze względu na niezbalansowany zbiór wykorzystany w tej pracy (zmienna objaśniana *de-*

ath_yn zawiera tylko 1,3% pozytywnych wartości) mechanizm ten okazał się niezbędny. Jest to rzadko spotykana funkcjonalność zaimplementowana w modelach uczenia maszynowego. Ze względu na ten opisany mechanizm, wysoką wydajność obliczeń i zasobów oraz dużą popularność w komercjalnych zastosowaniach klasyfikator XGBoost został wybrany na docelowy model uczenia maszynowego dla tej pracy.

4.3 Interpretacja wskaźników jakości modelu

4.3.1 F1-score

W analizie statystycznej klasyfikacji binarnej, F1-score lub F1-measure jest miarą dokładności testu. Jest on obliczany na podstawie miary precyzji (ang. precision) i czułości (ang. recall), gdzie precyzja jest liczbą prawdziwie pozytywnych wyników podzielonych przez liczbę wszystkich pozytywnych wyników, w tym tych, które nie zostały prawidłowo zidentyfikowane, a czułość jest liczbą prawdziwie pozytywnych wyników podzielonych przez liczbę wszystkich próbek, które powinny być zidentyfikowane jako pozytywne. Precyzja jest również znana jako pozytywna wartość predykcyjna, a wycofanie jest również znane jako czułość w diagnostycznej klasyfikacji binarnej. Wynik F1-score jest średnią harmoniczną precyzji i czułości.

Najwyższą możliwą wartością F1-score jest 1.0, wskazując idealną precyzję i czułość, a najniższą możliwą wartością jest 0, jeśli precyzja lub czułość wynosi zero. Wynik F1-score jest również znany jako współczynnik Sørensena-Dice'a lub współczynnik podobieństwa Dice'a (DSC).

$$F_1 = \frac{2}{recall^{-1} + precision^{-1}} = 2 * \frac{precision * recall}{precision + recall}$$

F1-score został użyty do wyznaczenia najlepszych hiperparametrów. Po podstawieniu kolejnych kombinacji hiperparametrów do budowanego modelu dla danych treningowych obliczana była wartość F1-score. Za najlepsze hiperparametry uznano te dla, których model przetestowany na danych walidacyjnych osiągnął najwyższą wartość F1-score.

Ze względu na szerokie spektrum własności oraz celność tego wskaźnika F1-score został użyty również do przedstawienia jakości docelowego modelu. Wykonano to przez przetestowanie modelu na danych testowych i późniejszym obliczeniu wartości F1-score.

4.3.2 Precyzja

W rozpoznawaniu wzorców, wyszukiwaniu informacji i klasyfikacji (uczenie maszynowe), precyzja (ang. precision) - zwana również pozytywną wartością predykcyjną - jest frakcją odpowiednich instancji wśród odzyskanych instancji. Zarówno precyzja, jak i recall są zatem oparte na trafności.

W przypadku zadań klasyfikacyjnych terminy *prawdziwie dodatni*, "prawdziwie ujemny", "fałszywie dodatni"i "fałszywie ujemny"porównują wyniki testowanego klasyfikatora z wiarygodną oceną zewnętrzną. Terminy pozytywny i negatywny odnoszą się do przewidywań klasyfikatora (czasami znanych jako oczekiwania), a terminy prawdziwy i fałszywy odnoszą się do tego, czy przewidywania te odpowiadają zewnętrznemu osądowi (czasami nazywanemu jako obserwacja).

5 Wyniki i dyskusja

- przedstawienie wzkaźników jakości modelu
- przstawienie wyników SHAP oraz ich interpretacja
- porównanie wniosków uzyskanych z opracowanego modelu z wnioskami z przywołanej literatury naukowej

6 Zakończenie

- podsumowanie pracy

7 Literatura

A Dodatek: Ważne rzeczy do dodania

Spis tablic

Spis rysunków

Streszczenie

Tutaj zamieszczają Państwo streszczenie pracy. Streszczenie powinno być długości około pół strony.